8th AIAA/USAF/NASA/ISSMO
Symposium on Multidisciplinary
Analysis and Optimization
6-8 Sept. 2000 Long Beach, CA

A00-40160

AIAA-2000-4921

METAMODELING SAMPLING CRITERIA IN A GLOBAL OPTIMIZATION FRAMEWORK

Michael J. Sasena *, Panos Y. Papalambros †, Pierre Goovaerts ‡
University of Michigan, Ann Arbor, MI 48109

Abstract

The use of approximate models or metamodeling has lead to new areas of research in the optimization of computer simulations. Metamodeling approaches have advantages over traditional techniques when dealing with the noisy responses and/or high computational cost characteristic of many computer simulations, most notably those in MDO. While a number of methods in the literature discuss how to exploit the benefits of metamodeling approaches, one particular algorithm, Efficient Global Optimization (EGO), is the focus of this paper. Specifically, we look at the criteria used by the algorithm to select additional points to add to the data set used in fitting the metamodel. In addition to modifications to the original criterion, three criteria originally proposed for use in infill sampling in the field of geostatistics are explored. The impact of these criteria on the search strategy of EGO is examined through several analytical examples. In addition, several enhancements to EGO are explored. Finally, a case study is presented using a computer simulation to predict the fuel economy of hybrid vehicles.

1. Introduction

As computer simulations become more complex and computationally expensive, the number of function evaluations required for optimization must be carefully limited. To that end, researchers have explored the use of metamodels, namely, simpler approximate models based on data derived from running the original simulations. After the initial expense of collecting a data sample, approximations can be constructed from a wide variety of mathematical forms. In an optimization framework the new model can replace the original one, thus reducing the computational burden of evaluating numerous designs.

One particular algorithm used in this way is the *Efficient Global Optimization* (EGO) algorithm developed by Schonlau, Welch and Jones [7]. EGO takes an initial, small data sample within the design space and fits a kriging approximation model [6]. The approximation is then used to select the set of points to add, the so-called infill samples, that are most beneficial according to some criteria. The sample is then updated, the model is refit, and the process of choosing new points continues until the improvement expected from sampling additional points has become sufficiently small. The results presented here add only one infill sample at a time, but the more general case applies.

This paper explores the strategy used for adding new data points to the sample. EGO uses a *generalized expected improvement* function whereby points that have either low objective function value or high uncertainly (i.e., model inaccuracy) are given precedence. A single parameter, g, determines the balance between them. Researchers in the field of geostatistics have proposed three additional criteria for choosing the location of additional sampling sites in the field. No studies to date have shown how these alternative criteria could impact an optimization algorithm such as EGO. In the paper, we investigate the influence that the parameter g and the choice of sampling criteria have on EGO.

This paper is organized as follows. First, the expected improvement function is described based on Schonlau's thesis [10]. Next, three alternative infill sampling criteria proposed by Watson and Barnes [11] are reviewed. Four analytical examples are then presented to demonstrate the differences between the criteria. Three small discussions follow highlighting various aspects of the EGO algorithm that impact its interaction with the sampling criteria. Finally, a simulation-based example is presented and general conclusions are summarized.

2. Generalized Expected Improvement

Let $f^n_{\min}$ be the minimum sampled value of the function $y = f(\boldsymbol{x})$ after n evaluations where $\boldsymbol{x}$ is a vector of input values. The response is treated as a realization of a random function $Y(\boldsymbol{x})$ that is assumed to have a Gaussian distribution with variance s^2. For simplicity, we will leave out the dependence on $\boldsymbol{x}$, notating the value $y(\boldsymbol{x})$ as y. Now we can define the improvement over the current $f^n_{\min}$ as

---

* Graduate Student, Department of Mechanical Engineering and Applied Mechanics.
† Professor, Department of Mechanical Engineering and Applied Mechanics, AIAA member.
‡ Assistant Professor, Department of Civil and Environmental Engineering.

$$I = \max\{0, f^n_{\min} - y\}. \quad (1)$$

The expected improvement is then defined as

$$E(I) = \int_{-\infty}^{f^n_{\min}} (f^n_{\min} - y)\phi(y)dy, \quad (2)$$

where $\phi(\ )$ is the Gaussian probability density function representing uncertainty about y. Using a generalized least squares technique (kriging) to predict $\hat{y}$, the expected improvement can be computed as

$$E(I) = \begin{cases} (f^n_{\min} - \hat{y})\Phi\left(\frac{f^n_{\min} - \hat{y}}{s}\right) + s\phi\left(\frac{f^n_{\min} - \hat{y}}{s}\right) & \text{if } s > 0 \\ 0 & \text{if } s = 0 \end{cases}, \quad (3)$$

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

Inspection of the expected improvement function reveals two important trends. The first term in Equation (3) is the improvement over $f^n_{\min}$ multiplied by the probability of actually achieving that amount of improvement. The second term is the square root of the variance multiplied by the probability that y will be smaller than $f^n_{\min}$. Therefore, the expected improvement function will be large for points $\mathbf{x}$ where the predicted value $\hat{y}$ is likely smaller than $f^n_{\min}$ and/or where there is high uncertainty in the value of the prediction itself. As the variance vanishes at the sampled data points, the expected improvement will also vanish there.

One can force the expected improvement function to search more globally by introducing a parameter, g. Indeed, as g increases, the EGO algorithm tends to search more globally around the design space. Defining the generalized improvement as

$$I^g = \max\left\{0, (f^n_{\min} - y)^g\right\}, \quad (4)$$

where g is a non-negative integer, one can define the generalized expected improvement function using a recursive formula. The derivation is found in Schonlau [10] and the resulting formula is

$$E(I^g) = s^g \sum_{k=0}^{g} (-1)^k \left(\frac{g!}{k!(g-k)!}\right) (f'^n_{\min})^{g-k} T_k, \quad (5)$$

where $f'^n_{\min} = \frac{f^n_{\min} - \hat{y}}{s}$ and

$$T_k = -\phi(f'^n_{\min})(f'^n_{\min})^{k-1} + (k-1)T_{k-2} \quad (6)$$

starting with $T_0 = \Phi(f'^n_{\min})$ and $T_1 = -\phi(f'^n_{\min})$. Unless otherwise stated, it can be assumed throughout this paper that the default value of $g = 1$ is used.

To demonstrate the impact of the parameter g, we show in Figure 1 a one-dimensional example for the ordinary Expected Improvement (EI) function, that is, with $g = 1$. The w-shaped dashed line is the true objective function we wish to model, the solid line is the kriging approximation conditional to the sample points shown as circles. The light colored function at the bottom is the EI function, normalized to make better comparisons across infill criteria.

One can see that expected improvement rises significantly in only two areas. The one on the left is a region where the data sampling is sparse, and thus the model uncertainty is high. Indeed, the kriging approximation is least similar to the true function in this region. The region on the right is where the expectation of finding a better objective function value is high. Again, the plot justifies this interpretation.

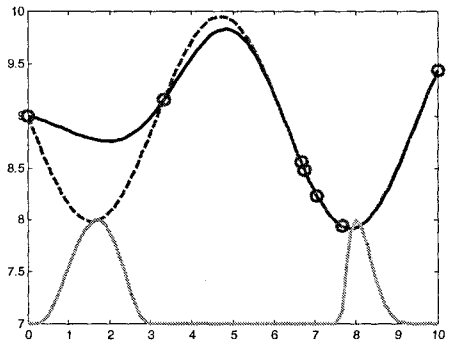

Figure 1: EI function for g = 1

In contrast, with $g = 5$ the EI function rises more in areas where the uncertainty is high than in areas where the probability of improving on $f^n_{\min}$ is high. In Figure 2, the same test function is shown with the same set of sample points. This time, however, the EI function shows interest only for the area on the left, where model uncertainty is high. For this iteration, the region on the right has become seemingly unimportant. This illustrates that increasing the value of g shifts emphasis towards searching the design space more globally.

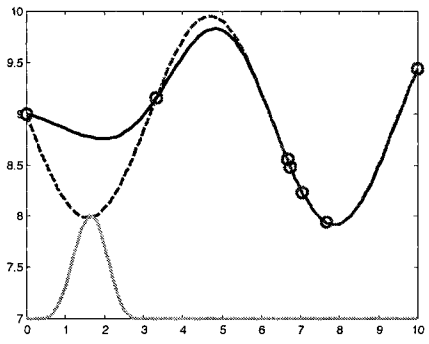

Figure 2: EI function for g = 5

3. Alternative Sampling Criteria

Watson and Barnes describe possible criteria to determine a set of locations for further sampling once an initial set of data, $\boldsymbol{S}$, has been collected [11]. Each criterion attempts to solve a problem with a different objective, namely, (i) locate the "threshold-bounded" extreme, (ii) locate the regional extreme, or (iii) minimize surprises. They are abbreviated here as WB1, WB2, and WB3 respectively. The mathematical formulation for each of these criteria is given below.

3.1 Locating Threshold-Bounded Extremes

The goal of this formulation is to locate points which maximize the probability that at least one of the infill samples exceeds some specified threshold. While this threshold was originally understood in the context of contamination testing, one can easily expand the thought to design optimization by using $f^n_{\min}$ as the threshold. In this way, the merit function rewards points that are likely to yield results better than the current best sampled point. The mathematical formulation appears as the cdf of the z-score statistic

$$\text{WB1} = \Phi\left(\frac{f^n_{min} - \hat{y}}{s}\right). \tag{7}$$

WB1 can be interpreted as the probability that $\hat{y}$ is no greater than $f^n_{\min}$ and it is maximized. Notice that this formulation is in fact the EI function with $g = 0$, and is thus extremely local in its search. One must therefore be fairly confident that the model has found the region of the optimum for this criteria to be successful.

The behavior of this function is illustrated in Figure 3 for the same example as above. Again, there are two peaks, but the one on the left is much smaller than previously. This reflects the fact that the criterion places less importance on searching globally to improve model accuracy than on locally improving the region where predictions are most likely to improve upon $f^n_{\min}$.

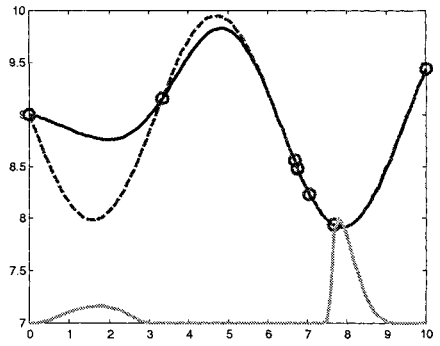

Figure 3: Threshold-bounded criteria

3.2 Locating the Regional Extreme

The second criterion attempts to minimize the expected value of the smallest observation once the infill samples have been added. The resulting formula is remarkably similar to the EI function.

$$\text{WB2} = \begin{cases} \hat{y} + (f^n_{\min} - \hat{y})\Phi(f'^n_{\min}) + s\phi(f'^n_{\min}) & \text{if } s > 0 \\ 0 & \text{if } s = 0 \end{cases} \tag{8}$$

The only difference between Equations (3) and (8) is the additional first term, the predicted value at the location of interest. In this sense, this criterion gives slightly more merit to locations likely to result in better objective function values than does the EI function.

In Figure 4, the differences between this criterion and the EI function are made clear. While they strike a similar balance between the local and global searching, the WB2 criterion is smoother, because it does not return to zero at the sampled points. Rather, it adds the predicted value of the objective function and maintains continuity in this example. This appealing trait is worthy of note, as it may help in locating the maximum of the criterion.

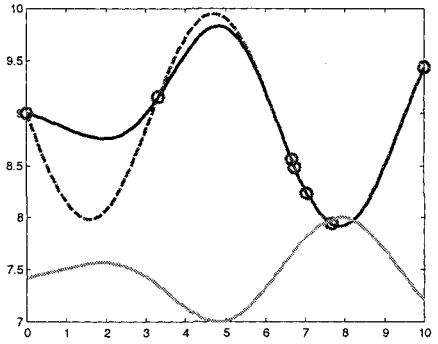

Figure 4: Regional extreme criteria

3.3 Minimizing Surprises

The third criterion attempts to minimize the maximum probability that a true value deviates significantly from its predicted value after the infill samples have been added. The term "significantly" is quantified by t, the tolerance level. The mathematical formula is

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{v}} \{\Pr[|Y(\boldsymbol{v}) - \hat{y}(\boldsymbol{v})| > t | \boldsymbol{S}]\}, \tag{9}$$

where $\boldsymbol{x}$ is the candidate infill sample point of interest, $\boldsymbol{v}$ is a generic location in the design space, $Y(\boldsymbol{v})$ is the random variable at the unobserved location $\boldsymbol{v}$, and $\hat{y}(\boldsymbol{v})$ is the predicted value there. The $\boldsymbol{S}$ appearing on the far right side indicates that the probability of the error exceeding t is conditional to the sample set, $\boldsymbol{S}$. Watson and Barnes note that a simplification is possible.

Because the probability is an increasing function of the conditional variance, one may use the formula

$$\text{WB3} = \min_{x} \max_{v} \{\text{Var}[Y(v)|S \text{ and } x]\}. \quad (10)$$

Notice that the variance is conditional to both the sample set, S, and the candidate infill samples, x. One can compute the variance of $Y(v)$ in this problem because the updated variance (i.e., the variance of the model once the candidate infill samples have been added) is a function of the locations of the x only, not their values.

This minimax problem within the original design optimization problem adds significantly to the total run time. Thus it is best suited for problems where the objective function is extremely expensive to calculate, and function evaluations must be rigorously conserved.

There are several other difficulties associated with this criterion. Consider Figure 5. Using the same sample, it was necessary to use a log transformation as WB3 varied by four orders of magnitude. Notice also that the criterion returns to its minimum value at each of the sampled values. This is as expected, because the variance would not be improved anywhere in the design space if one were to just resample an existing point.

Another characteristic worth noting is the somewhat flat distribution of WB3. That is, the peaks are roughly the same height due to the minimax behavior of the criterion. Watson and Barnes observe that WB3 would be flat in spots when there exist some distance beyond which two points are uncorrelated. If there are areas in the design space which are not "covered", adding infill samples far from this area will not improve the maximum variance in the uncovered area. This is the cause of the relatively flat region on the left side of Figure 5. Worse yet, if the design space would not be covered even with the additional infill samples, then the maximum variance remains constant regardless of where infill samples are taken and the WB3 function becomes flat. These characteristics could cause serious difficulties in computing the maximum of WB3.

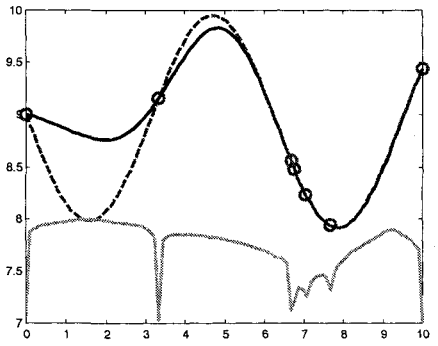

Figure 5: Minimizing surprises criterion

4. Analytical Examples

In order to assess what effect the various criteria have on the search progress of EGO, four analytical examples are shown: a one-dimensional, a two-dimensional, and two constrained two-dimensional functions. Test problems with a small number of variables were chosen for better visualization. Readers are directed to Schonlau [10] to see how EGO works with problems of six to ten design variables. Problems with a large number of variables may be solved using metamodels in a decomposition strategy [3]. We begin this section by describing each of the functions used in the testing, then describe the comparison methods and report our findings.

Example 1: One-dimensional function

The first test was created specifically for this work, comprising a one-dimensional function with two local minima, one just slightly better than the other. This is intended to be the simplest solvable problem that could still fool some algorithms. The multimodal nature of the function would defy local optimizers, and the fact that the local minimum is only slightly worse than the global minimum is intended to cause difficulties for EGO.

Example 1 is indeed the same function used in the illustrations of the previous two sections. The function,

$$f = -\sin(x) - \exp(x/100) + 10, \quad (11)$$

is defined over the range $x = [0,10]$ and is shown in Figure 6. The local minimum of 7.9841 occurs at $x = 1.5810$, shown as a circle. The global minimum of 7.9182 occurs at $x = 7.8648$, shown as an asterisk.

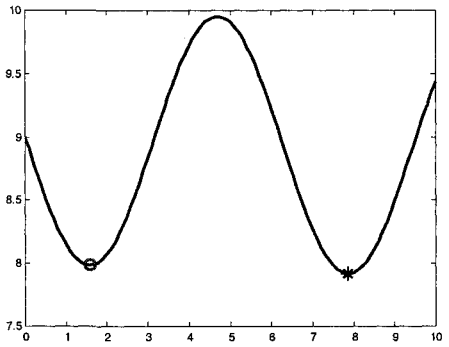

Figure 6: One-dimensional function

Example 2: Branin function

The second example is the Branin test function, taken from Dixon and Szegö [2]. The function

$$f = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10 \quad (12)$$

contains three local minima at $\boldsymbol{x}$ = {3.1416, 2.2750}, $\boldsymbol{x}$ = {9.4248, 2.4750}and $\boldsymbol{x}$ = {-3.1416, 12.2750} shown as asterisks in Figure 7. They have identical function values of 0.3979. This problem should test EGO's ability to find local optima of equal value.

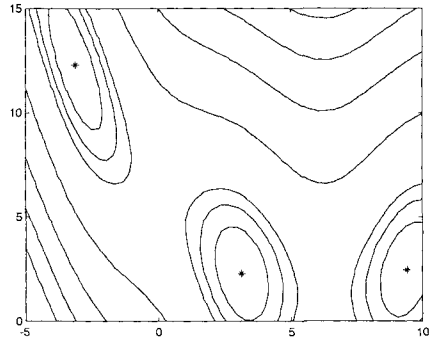

Figure 7: Contour plot of Branin function

Example 3: Two-dimensional function

A sinusoidal constraint is placed on a multimodal function in two dimensions for the third example. The function takes the form

$$f = 2 + 0.01(x_2 - x_1^2)^2 + (1-x_1)^2 + 2(2-x_2)^2 + 7\sin(0.5x_1)\sin(0.7x_1x_2) \quad (13)$$

with the constraint function

$$g = -\sin(x_1 - x_2 - \pi/8). \quad (14)$$

The function is defined over the range $x_i = [0,5]$, $i = 1, 2$ and the constraint is active at the true solution. Because of the multimodal behavior of the function, several local minima appear along the constraint boundary, but the global solution has a value of -1.1743 at the point $\mathbf{x}$ = {2.7450, 2.3523}. Figure 8 below shows the contour plot of the objective function and constraint boundary (the diagonal lines). The hash marks indicate the infeasible side of the constraint. The unconstrained and constrained minima are shown as the circle and the asterisk respectively.

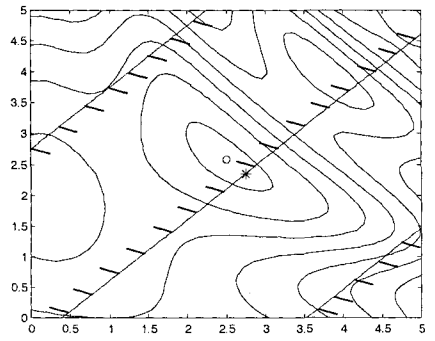

Figure 8: Contour plot of Two-dimensional function

Example 4: Gomez #3 function

The Gomez #3 test function, taken from Gomez [5], is a difficult example because of a constraint that cuts the feasible design space into several small islands. Its objective and constraint function are

$$f = \left(4 - 2.1x_1^2 + \frac{1}{3}x_1^4\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2 \quad (15)$$

$$g = -\sin(4\pi x_1) + 2\sin^2(2\pi x_2). \quad (16)$$

In Figure 9, the objective function is shown in solid lines, and the islands of feasible design space lie inside the dashed circles. There is obviously a local optimum within each feasible island, but the global optimum lies at $\boldsymbol{x}$ = {0.1093, -0.6234} with a value of -0.9711, shown as an asterisk.

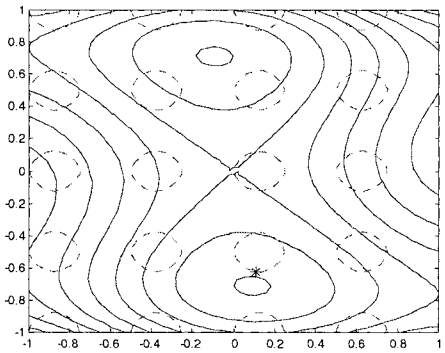

Figure 9: Contour plot of Gomez #3 function

4.1 Comparison Metrics

Comparisons of the infill sampling criteria are made difficult because there is no rigorous convergence criterion for EGO. A stopping rule proposed by Schonlau [10] suggests one should stop searching once the ratio of the expected improvement to the current best sample value becomes sufficiently small, say 0.01. However, this stopping rule no longer has meaning once the alternative infill sampling criteria are employed. Thus, each test is stopped after 100 function calls have been made. Then, instead of looking only at the number of function evaluations required to convergence, several other comparisons are made. In all cases, lower values are better.

- *Function calls to* $x_{1\%}$: The number of function calls required before a point is sampled within a box the size of +/- 1% of the design space range centered around the true solution.
- *Function calls to* $f_{1\%}$: The number of function calls required before a feasible point is sampled that has a function value within 1% of the true solution.
- *Accuracy at* x_*: Euclidean distance from the best sample point to the nearest global solution (x_*).

- *RMS error*: The overall modeling error. After 100 evaluations, the metamodel is compared to the true function on a 50 by 50 gridded set of locations (100 locations in the case of Ex1). The resulting errors at the N comparison points are summarized by the Root Mean Squared error (RMS), calculated as

$$RMS = \frac{1}{N}\sqrt{\sum_{i=1}^{N} (\hat{y}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))^2}. \qquad (17)$$

The first two of these comparisons are more traditional measures of how efficiently the algorithm finds the solution. The third comparison measures how accurately EGO finds the solution. The last metric checks how accurately the final metamodel approximates the design space. This would show how reliably the designer could reuse the metamodel for solving similar design problems quickly and efficiently. Both of the last two metrics have been normalized in the tables as percentages of the largest values in each row to facilitate comparisons for a given test function.

4.2 Initial Testing Results

The two tables below show the initial testing results for the analytical examples. Table 1 shows the $x_{1\%}$ comparison metric, Table 2 shows the $f_{1\%}$ comparison. The columns are for the different infill criteria. The number in parenthesis below EI indicates the value used for g in Equations 5 and 6. The last three columns are for the three criteria proposed by Watson and Barnes (WB). Each row is for a different test function.

The '+' symbol refers to cases where the criterion was not met within the 100 function call limit. Because all three minima are global, the numbers for the Branin function (Ex2) in Table 1 refers to the first time the criteria lead to a function call within $x_{1\%}$ of any of the three minima.

Table 1: Function calls to $x_{1\%}$ (+ indicates >100)

	EI (1)	EI (2)	EI (5)	EI (10)	WB 1	WB 2	WB 3
Ex1	10	7	10	10	8	8	+
Ex2	24	24	24	24	34	24	+
Ex3	+	+	+	+	91	29	+
Ex4	62	71	93	+	78	60	+

Table 2: Function calls to $f_{1\%}$ (+ indicates >100)

	EI (1)	EI (2)	EI (5)	EI (10)	WB 1	WB 2	WB 3
Ex1	7	7	7	8	8	8	7
Ex2	24	24	56	+	60	+	+
Ex3	35	+	+	36	88	28	+
Ex4	26	25	25	23	37	25	+

From the tables we can see two obvious shortcomings. First, most of the criteria performed poorly on the first of the constrained examples (Ex3), while WB2 performed fairly well on both constrained examples. This shall be discussed more in Section 7. Second, the minimizing surprises criterion (WB3) failed to sample a point within the $x_{1\%}$ box for any of the functions before the 100 function call limit was reached. A simple explanation for this failure is that WB3 seeks only to sample points that reduce the overall uncertainty in the surrogate models. There is no attempt to locate even local minima. Thus, there is no reason to expect it will sample a point within the $x_{1\%}$ limits except by chance. It is clear from these tables that the criteria that try to balance improving model accuracy with locating minima perform much better at sampling points of interest.

The next comparison made is the x_* metric shown in Table 3. As WB3 never sampled points within the $x_{1\%}$ bounds, these tests were left out of the table. The WB1 criteria finds the solution with the highest accuracy on all four examples, which is likely due to the fact that all the emphasis is put on locating a regional extreme. This is however somewhat in contradiction with the fact that the local accuracy does not systematically deteriorate as value of g in the EI function increases.

Table 3: Normalized distance to optimum

	EI (1)	EI (2)	EI (5)	EI (10)	WB 1	WB 2
Ex1	0.2	29.4	100	100	0.2	29.4
Ex2	18.5	38.0	9.8	100	0.4	17.8
Ex3	40.5	29.8	100	11.8	9.2	18.5
Ex4	100	23.7	43.3	8.3	5.7	23.7

The last metric looks at the reliability of the kriging models as global surrogates. One would expect that the minimizing surprises criterion results in models with lowest RMS errors. Table 4 presents the results.

Table 4: Normalized RMS errors after 100 samples

	EI (1)	EI (2)	EI (5)	EI (10)	WB 1	WB 2	WB 3
Ex1	12.3	1.4	10.0	0.6	100	75.4	0.08
Ex2	26.7	16.5	42.0	38.7	5.3	29.0	100
Ex3	19.1	36.0	19.3	31.1	47.3	46.5	100
Ex4	19.3	29.0	27.5	16.1	27.7	72.7	100

Evidently, the trend is not as one would expect. For Ex1 and Ex4, the criteria which search more globally (either EI with a high g value, or WB3) tend to have lower RMS errors. However, this is not in general the case. More baffling still is the fact that the most "local" of the methods, WB1, yields the lowest RMS error for Ex2.

In Figure 10 we show the contour of the true Branin function with circles indicating the initial 21 point design and the x's marking the points selected by the WB3 criterion for the remaining 79 points. Notice the unexpected clustering in several locations. If the criterion of minimizing the maximum variance had successfully been implemented, one would have expected the infill sample points to be spread out more evenly.

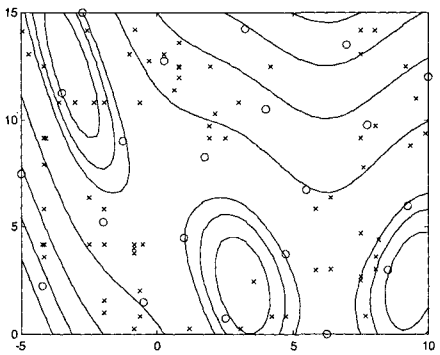

Figure 10: Branin example using WB3

One possible explanation for this unexpected behavior is the failure of the internal optimization algorithm to find the solution to the difficult minimax problem at each iteration. Currently, the DIRECT algorithm [8] is being used for 30 iterations. Allowing DIRECT to search for more iterations could produce better results. However, the overhead associated with the problem is already staggering in some cases, and increasing the iterations on the internal optimization problem is not a favorable alternative. Another possible explanation is that the process of fitting the surrogate model is failing at some (or all) iterations. This issue is examined further in the next section.

5. Model Fitting

One of the reasons the results in the Section 4 were difficult to interpret was the occasional failure of the Maximum Likelihood Estimation procedure to accurately estimate the parameters of the covariance model, which effects the kriging prediction. For example, after 70 function evaluations, the kriging fit to Branin function was evaluated visually and found to be quite poor. With an inaccurate surrogate model, the infill sampling stage became misguided and points were added in regions the infill criteria would not really find interesting with the true function. As a result, interpreting the impact of the infill criteria was obscured.

Other model parameters were tried which led to a surrogate model that matched the true Branin function much more accurately. Oddly, the likelihood value (the measure which assesses the goodness of the model) of the poor fit had a better value than the likelihood value of the good fitting model. Previous studies using the Branin function reported that the normal distribution assumption inherent to the present application of kriging was acceptable [10]. Assuming therefore that the Branin function itself is not the cause of modeling error, we speculate that the likelihood function may not be providing a good measure of model accuracy.

A test was performed with the initial thirty data points sampled from the Branin function using the regular EI criterion. The sample was used in a cross-validation fitting process to find better values of the kriging parameters whereby the RMS error of cross-validation was minimized. The resulting kriging parameters performed much better even though the likelihood value was much worse. This suggests that cross-validation at each iteration could provide more reliable metamodels than the traditional MLE technique. The trade-off is the increased computational cost incurred.

Armed with more reliable surrogate models, another test was performed to compare the impact of the infill criteria. The model parameters were held fixed at the better values across infill criteria for all iterations to eliminate the influence of model accuracy. The comparison metrics were once again checked after 100 function evaluations. One may think the comparisons measuring the model accuracy might be invalid. considering that the . However, the kriging model is an interpolating model. As such, it will go through the data samples exactly, but will have some error elsewhere, regardless of the fitting parameters. Thus, the RMS metric will show how globally the data samples are spread.

One additional comparison made here ($3x_{1\%}$) is the number of function calls required before at least one point has been sampled within the $x_{1\%}$ tolerance for all three of the Branin function's global minima.

Table 5: Modified Branin example (+ indicates >100)

	EI (1)	EI (2)	EI (5)	EI (10)	WB 1	WB 2	WB 3
$x_{1\%}$	24	24	24	24	22	25	+
$3x_{1\%}$	28	28	30	+	31	+	+
$f_{1\%}$	24	24	30	+	25	26	+
x_*	59.2	35.5	17.9	100	1.3	23.6	N/A
RMS	33.4	40.9	31.0	98.3	41.5	100	24.0

There are some interesting results that come about. Although normalizing the RMS metric hides this information, the RMS values went down usually by two orders of magnitude compared to the initial results. Indeed the better kriging model parameters improved the overall accuracy significantly. It also appears that the improved models at each iteration have allowed the infill criteria to make more informed choices of sample points. As a result, the number of function calls required to reach the $x_{1\%}$ and $f_{1\%}$ limits is fewer than initially. Also, more of the criteria actually found all three minima within the $x_{1\%}$ and $f_{1\%}$ limits before 100 function calls.

One may notice a slightly larger number of function evaluations for the $x_{1\%}$ and $f_{1\%}$ limits with criteria that search more globally as compared to the more local criteria. Intuition may lead us to believe that the more global the search, the quicker the criteria will sample points around the three global optima dispersed through the design space. Rather, quite the opposite occurs. Although the model sees there are potentially better points in the three valleys, it defers sampling additional points there until more of the uncertainty has been reduced.

Another result worth noting is the observation that the WB3 criterion now lives up to expectations by having the lowest RMS error. It is possible that the equally good fitting models enable the criterion to search better for the minimax subproblem solution. Of course, this does not help explain why the EI(10) criterion performed so poorly. If it was more globally searching by design, then the RMS would have been presumable lower than for the EI criteria with lower g values. This matter remains unresolved and warrants further investigation.

6. Cooling Schedule for EI

As described in Section 3, the effect of increasing the parameter g in the EI criterion is to attempt to search more globally - although in practice, results thus far do not definitively support that claim. However, a good method for choosing the value of g is not immediately obvious. Too high of a value could prevent EGO from converging on a good solution in a reasonable amount of time. Too low of a value could allow EGO to overlook areas of high uncertainty as it searches too locally. This points towards an opportunity for improvement presented here.

Another test was performed on the Branin function using the same modeling improvements as in the last section. With the kriging parameters fixed, the impact of changing g as a function of iteration number was explored. The effect was similar to a Simulated Annealing algorithm whereby the goal is to start searching globally, then refine the search to more local concerns as time goes on. The heuristic cooling schedule used in this work is shown in Table 6. Table 7 shows the results.

Table 6: Cooling schedule

Iteration	g value
1 - 4	20
5 - 9	10
10 - 19	5
20 - 24	2
25 - 34	1
35 -	0

Table 7: Branin cooling schedule example

	cool	EI (1)	EI (2)	EI (5)	EI (10)
$x_{1\%}$	26	24	24	24	24
$3x_{1\%}$	42	28	28	30	+
$f_{1\%}$	53	24	24	30	+
x_*	1.6	59.2	35.5	17.9	100
RMS	35.9	33.9	41.6	31.6	100

As compared to the EI criteria with fixed g values, the cooling schedule does not appear to fair very well for the first three metrics. Both the EI(1) and the EI(5) beat the cooling method on all but the x_* metric. However, when looking at the accuracy of the optimum sampled point, the cooling scheduled EI criterion outperforms the others. Thus, the cooling schedule may not find solutions as fast as its competitors, but it finds them more accurately, once the g parameter has been lowered. It also competes fairly well at the RMS metric. In short, the cooling schedule approach shows promise.

7. Constraint Handling

One concern with the EGO algorithm is the handling of constraints. Currently, the literature tentatively suggests multiplying the value of the expected improvement by the probability that the point is in fact feasible [10]. However, the value of the infill criterion may be impacted too strongly. Consider the case where the infill criterion has small values near the border of the constraint with only slightly worse values a small distance from the constraint boundary. By multiplying the infill criterion by the probability of feasibility, the already small values near the constraint boundary may become such that EGO now prefers points further from the constraint where the infill criterion is not as strongly reduced. In essence, this keeps the algorithm from exploring points directly along the constraint boundary where the true optimum lies.

A remedy proposed here is to switch from the probability method to a penalty method after a set number of iterations. The idea is that, with the probability method, the model of the constraint function presumably becomes more accurate around constraint boundary once several points have been sampled in that vicinity. With more confidence in where the exact location of the constraint boundary lies, the algorithm can now directly approach this region with a penalty method. With a penalty method, a large constant (i.e., a penalty) is added to the criterion in order to restrict it from choosing infill samples in the infeasible design region.

This idea was tested on Ex3. Below we see a side-by-side comparison of the function calls made by the EI criterion with the probability method (left) and the switch to penalty method after ten iterations using the same infill criterion (right). The contours are of the true functions, and the initial samples are shown as circles and the infill points as x's.

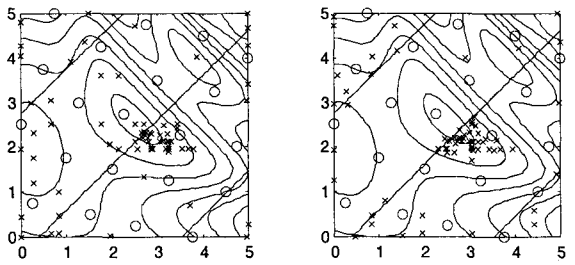

Figure 11: Ex3 constraint handling for EI criterion

While both methods cluster around the global optimum, the probability method tends to sample more frequently in the infeasible space (see Figure 8) than the penalty method. The penalty method clusters more sample points around the optimum on the feasible side of the constraint. For optimization problems where strict feasibility is important, using a penalty method after several iterations may prove useful.

8. Simulation-Based Example

In this section, a Hybrid Electric Vehicle (HEV) simulation called ADVISOR [9] is used to explore the capabilities of EGO to work with computer simulations. ADVISOR predicts the fuel economy, acceleration performance and emissions of a wide range of vehicles. On an Ultra 10 SunSparc workstation, each function evaluation requires approximately two minutes to compute, expensive enough to warrant judicious use of function evaluations through surrogate modeling techniques. More details on the simulation can be found in the following works [1], [4].

The design problem illustrated here is to maximize the fuel economy in miles per gallon (m.p.g.) of a mid-sized hybrid passenger car subject to a set of performance constraints established by the Partnership for Next Generation Vehicles (PNGV), a US government-led organization. Additional constraints are imposed to ensure that the vehicle is able to sustain the state of charge (SOC) in the batteries. The design variables are the size of the engine (in kW), the size of the electric motor (in kW), and the size of the battery pack (in number of modules). The problem formulation is summarized below.

maximize $f(\boldsymbol{x})$ = m.p.g.

$\boldsymbol{x}$ = {engine size, motor size, battery size}

15 kW $\leq$ engine size $\leq$ 150 kW

5 kW $\leq$ motor size $\leq$ 50 kW

5 modules $\leq$ battery size $\leq$ 70 modules

subject to:

g_1: maximum speed $\geq$ 85 mph

g_2: maximum acceleration $\geq$ 0.5 g's

g_3: 5 second distance $\geq$ 140 feet

g_4: 0-60 time $\leq$ 12 seconds

g_5: 0-85 time $\leq$ 23.4 seconds

g_6: 40-60 (passing) time $\leq$ 5.3 seconds

g_7: 55 (cruising) gradability $\geq$ 6.5%

g_8: maximum launch grade $\geq$ 30%

g_9: change in $SOC_{urban} \leq 0.5\%$

g_{10}: change in $SOC_{highway} \leq 0.5\%$

8.1 Results

Four tests were run on the simulation-based example for 200 function evaluations each. Because of the poor performance of the WB3 criterion in locating minima, it was decided that it would not be tested with the others. The last test (shown as "cool") used the two alterations discussed in Sections 6 and 7 with the EI criterion. Specifically, it used the cooling schedule shown in Table 6,

and the constraint handling switched over from the probability method to the penalty method at iteration 25.

In Table 8, we show the values of each of the three design variables at their best sample value, the value of the fuel economy prediction at that point, and the largest value of any constraint function (not including the SOC constraints) at the best sample point (the number of that constraint given in parentheses).

Table 8: Simulation-based example

	Engine (kW)	Motor (kW)	Battery (modules)	m.p.g.	max(g_i)
EI	45.83	34.80	18.27	49.18	-.03 (g_6)
WB1	46.40	39.17	23.06	48.58	-.10 (g_7)
WB2	48.75	46.63	19.63	47.47	-.40 (g_6)
cool	45.83	40.28	20.65	48.73	-.06 (g_7)

It can be seen that either the passing time or the gradability constraint are closest to being active in all four cases. This agrees with prior experience with the HEV simulation. The WB2 criterion had the poorest results. Its best sample point still had quite a bit of slack left in the constraint where improvements on the objective function could have been made. This is most likely due to the fact that there were ten constraints to be modeled in this problem. The compounding effect of the probability multiplications may have been detrimental.

When the penalty switching was used with the last test, the constraint came much closer to activity. Still, the original EI function discovered a slightly better region by reducing the size of the components significantly.

9. Discussion

The focus of this paper was to compare the effectiveness of a variety of infill sampling criteria in a global optimization framework. The results shown here do not indicate that any one of the criteria are superior in all respects. The local searching criteria such as EI(1) and WB1 performed well at accurately locating optima in these tests, but they did not necessarily do so the most efficiently. Also, EI(10) and WB3, intended to search more globally throughout the design space, did not always do so as well as other criteria. The results shown here have lead to more questions, opening the discussion up to new directions.

Some effort was made to alleviate the analytical difficulties in order to better compare the alternative infill sampling criteria. Accounting for the modeling inaccuracy has helped clarify the results somewhat, but we still cannot be sure how much of the differences in the results are contributable to the actual differences among the criteria, and how much is simply an artifact of the imperfect searching process during the maximization of the infill sampling criteria. Regardless, results have shown that the impact of having an accurate fitting to the data sample is crucial to the success of EGO.

The idea of using a cross-validation scheme to achieve better model accuracy was successfully tested. However, the cost of cross-validation is quite high, and the EGO algorithm already has substantial overhead. Perhaps a better alternative is to use jack-knifing, whereby a supplementary set of data is left aside for checking the accuracy of the model. The cost of fitting a model this way is lower, and in the case of many real world simulation problems, the designer already has quite a few prior computer runs at their disposal from earlier studies. Alternatively, cross-validation could be performed ahead of time when the designer has a reasonable database of simulation runs and the model parameters held fixed throughout the optimization as was done in Section 5. This one time cost could reduce later computations enough to warrant cross-validation.

The effect of the constraint handling was also mentioned. It was shown that the metamodeling strategy has an aversion to approaching the constraint boundary. This effect is amplified by the number of constraints increases. The penalty method attempted here shows promise, but better methods need to be developed for dealing with the multiple constraints inherent to most difficult computer simulations.

Another modeling issue that has been ignored is the validity of the Gaussian distribution assumption. The type of kriging models used here rely heavily on this assumption. Work must be done to ensure that optimization of generic computer simulations can be undertaken without violating this assumption. Data transformations have been suggested to aid in this endeavor, but are beyond the scope of this paper.

Convergence properties are a major weakness to the EGO algorithm. Even though some convergence criteria have been proposed for the EI function, none exist for these alternative infill sampling criteria. And while tests here indicate a clustering of the samples near the optima, there is no guarantee that a dense cluster will form in the limit. Of course, practical optimization in engineering fields may not require such mathematical rigor if the sole purpose is to locate a better design. In addition, EGO has the capability of restart. This is an advantageous property that allows the user to terminate the search at any time, then continue where they left off. There are no problem-specific parameters such as Hessian updates within the algorithm that are lost. This advantage has proven useful during the testing on the HEV simulation.

Finally, it is clear that the paper shown here is only a starting point. While the alternative infill sampling criteria discussed here did not do much to illicit great enthusiasm, there is hope. Several other infill sampling criteria exist in the geostatistics literature that may prove quite useful for solving these types of computer simulation problems. Further work will be done in evaluating potential candidates.

Acknowledgments

This research was partially supported by the General Motors Satellite Research Laboratory and the U.S. Army Automotive Research Center for Modeling and Simulation of Ground Vehicles. This support is gratefully acknowledged.

References

[1] Assanis, D., G. Delagrammatikas, R. Fellini, Z. Filipi, J. Liedtke, N. Michelena, P. Papalambros, D. Reyes, D. Rosenbaum, A. Sales, M. Sasena, 1999. An Optimization Approach to Hybrid Electric Propulsion System Design, *Journal of Mechanics of Structures and Machines*, Automotive Research Center Special Edition Issue, E. J. Haug (ed.), **27**, no. 4: 393-421.

[2] Dixon, L. C. W. and G. P. Szegö. The Global Optimisation Problem: An Introduction, *Towards Global Optimisation 2*, North-Holland Publishing Company, 1978.

[3] Fellini, R., P. Papalambros, T. Weber, 2000. Application of a Product Platform Design Process to Automotive Powertrains, *Proceedings of the 8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Long Beach, CA, Sept. 2000, AIAA-2000-4849.

[4] Fellini, R., N. Michelena, P. Papalambros, M. Sasena, 1999. *Optimal Design of Automotive Hybrid Powertrain Systems*, IEEE EcoDesign '99 Conference, Tokyo, Japan.

[5] Gomez, S. and A. Levy, 1982. The tunneling method for solving the constrained global optimization problem with several non-connected feasible regions, A. Dold and B. Eckmann (ed.): *Lecture Notes in Mathematics 909*, Springer-Verlag, pp. 34-47.

6] Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, NY.

'] Jones, Donald R., Matthias Schonlau, and William J. Welch, 1998. "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, **13**, no. 4: 455-92.

[8] Jones, D. R., C. D. Perttunen, and B. E. Stuckman, 1993. "Lipschitzian Optimization Without the Lipschitz Constant", *Journal of Optimization Theory and Application*, **79**: 157-181.

[9] National Renewable Energy Laboratory, Department of Energy, web site, http://www.ctts.nrel.gov/analysis/ .

[10] Schonlau, Matthias. *Computer Experiments and Global Optimization.*, 1997. Doctoral Dissertation, Department of Statistics, University of Waterloo.

[11] Watson, Alan G, and Randal J. Barnes, 1995. "Infill Sampling Criteria to Locate Extremes," *Mathematical Geology*, **27**, no. 5: 589-608.