Comment: Expected Improvement for Efficient Blackbox Constrained Optimization

Hao CHEN and William J. WELCH

Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z4, Canada
(*hao.chen@stat.ubc.ca; will@stat.ubc.ca*)

Gramacy et al. (2016) tackle the important problem of constraints in global optimization of a computer model. Indeed, for realistic applications in engineering and the physical sciences, constraints on the solution are probably the norm. When the constraints are expensive, complex functions of the controllable input variables, statistical modeling of them needs to be flexible. The objective function itself may also require modeling. Any optimization algorithm to guide further computer-model evaluations must somehow combine uncertainty from all components of the formulation—the objective function and the constraints—that require statistical modeling. How to do so in an efficient search for the best feasible solution is not immediately obvious. Thus, the authors are to be congratulated on their general and systematic approach to constrained optimization.

The authors note that the statistical literature is sparse on complicated, simulation-based constraints, but there are exceptions. Aslett et al. (1998) optimized part of an integrated circuit with respect to 20 controllable inputs subject to four simulator-generated constraints. That application had the further complication of 16 uncontrollable noise inputs, leading to a formulation seeking a constrained optimum over the controllable factors such that the constraints all have good mean values and small standard deviations with respect to variation induced by the noise variables. The approach was not as systematic as that proposed by Gramacy et al., however, requiring manual intervention to manage trade-offs between batches of computer-model runs.

A systematic approach to constrained optimization problems of exactly the type described by Gramacy et al. was proposed by Schonlau, Welch, and Jones (1998). Gramacy et al. cite this reference but dismiss it on the grounds that the constraints must be known. That claim is incorrect, however, as Schonlau, Welch, and Jones (1998) explicitly considered complex constraints from blackbox simulation and modeled them via GPs, just as Gramacy et al. do. Thus, there is some overlap in the two approaches. They differ in how the statistical models of the constraint functions and possibly the objective function are used, however.

Our comment concentrates on the similarities and differences of the two methods. We first describe how the Schonlau, Welch, and Jones (1998) algorithm can be viewed as an expected-improvement criterion, adapted in a fairly straightforward way to deal with modeled constraints. We then make empirical comparisons and some concluding remarks.

We first describe the Schonlau, Welch, and Jones (1998) expected-improvement method for constrained optimization. Using similar notation, we have the constrained minimization

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to } c_1(\mathbf{x}) \le 0, \ldots, c_k(\mathbf{x}) \le 0, \mathbf{x} \in \mathbb{B},$$

where $c_1(\mathbf{x}), \ldots, c_k(\mathbf{x})$ are constraints computed by one or more simulators run with inputs $\mathbf{x}$. Initially, we take $f(\mathbf{x})$ as a trivial, known function, as Gramacy et al. do, but later we demonstrate the original Schonlau, Welch, and Jones (1998) algorithm, where $f(\mathbf{x})$ is also a complex function from simulation.

Each constraint function is modeled as a Gaussian process (GP), the approach introduced by Sacks et al. (1989) for an output of a deterministic computer code. All results reported here are based on a constant mean function, a stationary squared-exponential (Gaussian) correlation function, and maximum likelihood (ML) estimation of the mean, variance, and correlation-function parameters. These details could be adapted as necessary. From training data of n runs at any stage of the sequential design, each constraint function $c_j(\mathbf{x})$ is modeled by a fitted GP, $Y_{c_j}(\mathbf{x})$. At any $\mathbf{x}$ considered for the next computer-model run, the GP provides an approximately normal predictive distribution with estimated mean $m_{c_j}(\mathbf{x})$ and standard deviation $s_{c_j}(\mathbf{x})$. Similarly, if the objective is unknown, a GP, $Y_f(\mathbf{x})$, can be fit for $f(\mathbf{x})$. Its approximately normal predictive distribution has mean and standard deviation $m_f(\mathbf{x})$ and $s_f(\mathbf{x})$. All this is now commonplace and is essentially the way that Gramacy et al. model the separate components in their Section 3.2. The following steps depart from their formulation, however.

The algorithm proceeds by trying to drive down the minimum feasible value of the objective found so far. Let $f_{\min}$ be the minimum feasible value of f found after n runs. (For simpler notation, we suppress the dependence on n, and $f_{\min}$ can be a large number if no feasible solution has been found so far.) If the constraints are ignored, the improvement $I(\mathbf{x})$ in $f_{\min}$ from a new run at $\mathbf{x}$ has expectation

$$E(I(\mathbf{x})) = \begin{cases} \max(0, f_{\min} - f(\mathbf{x})) & f(\mathbf{x}) \text{ known} \\ s_f(\mathbf{x})\phi(u) + (f_{\min} - m_f(\mathbf{x}))\Phi(u) & f(\mathbf{x}) \text{ unknown,} \end{cases} \quad (1)$$

© 2016 American Statistical Association and the American Society for Quality
TECHNOMETRICS, FEBRUARY 2016, VOL. 58, NO. 1
DOI: 10.1080/00401706.2015.1044119
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/tech*.

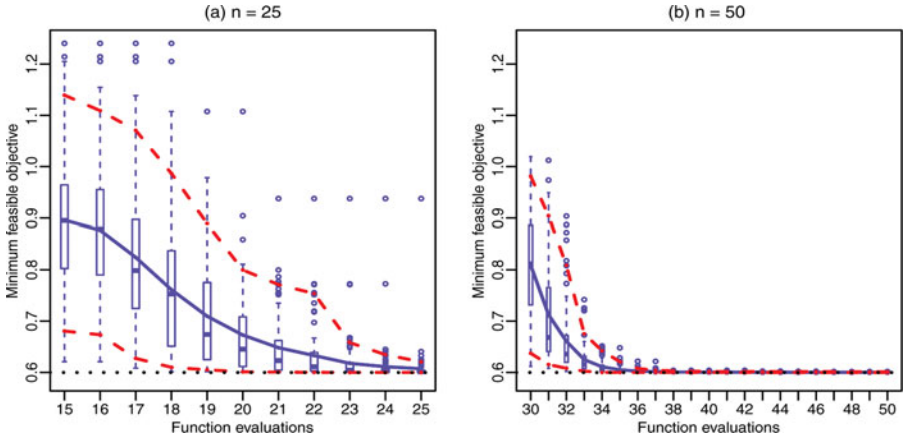

Figure 1. Minimum feasible objective found by the Schonlau, Welch, and Jones (1998) method applied to the illustrative test problem of Gramacy et al. (known $f(\mathbf{x})$). The box-and-whisker plots show distributions over 100 executions of the algorithm. The solid line and the two dashed lines represent mean, 5% (best case), and 95% (worst case) summaries, respectively. The dotted line is the global minimum, 0.5998. The total sample size is (a) $n = 25$ or (b) $n = 50$.

where $u = (f_{\min} - m_f(\mathbf{x}))/s_f(\mathbf{x})$, and ϕ and Φ denote the standard normal probability density function and cumulative distribution function, respectively. The second expression in (1) is the expected improvement (EI) criterion of Jones, Schonlau, and Welch (1998) from the approximately normal predictive distribution. To adapt EI to constrained optimization, Schonlau, Welch, and Jones (1998) also computed the probability that constraint j is met:

$$P(Y_{c_j}(\mathbf{x}) \le 0),$$

which is also straightforward from the normal predictive distribution. Taking account of all k constraint functions, the constrained version of EI in (1) is

$$E(I(\mathbf{x})) \times \prod_{j=1}^{k} P(Y_{c_j}(\mathbf{x}) \le 0), \tag{2}$$

and this is the criterion they used to choose the next run of the computer code. If we assume all GPs are statistically independent, this expression is the expected *feasible* improvement taking account of the constraints. Thus, not only did Schonlau, Welch, and Jones (1998) handle *unknown* constraint functions, their use of the GP models was surprisingly straightforward.

We now evaluate the performance of the Schonlau, Welch, and Jones (1998) constrained optimizer on the illustrative test problem in Section 1 of Gramacy et al. We consider two scenarios based on whether $f(\mathbf{x})$ is known or not in (1).

The algorithm starts from an initial Latin hypercube (McKay, Beckman, and Conover 1979) space-filling design. There are either (a) 15 points in the initial design followed by a further 10 guided by criterion (2) for $n = 25$ runs in total, or (b) 30 initial points augmented sequentially by another 20 for a total sample size of $n = 50$. In both cases, the sequential design algorithm is repeated 100 times, from different random starting designs. When maximizing the criterion (2), we use simulated annealing (SA) in the R package `GenSA` beginning with three randomly generated initial values. This optimization is based on inexpensive surrogates. In our reported feasible solutions, there is no tolerance to allow approximate constraint violation.

The results presented in Figure 1(a) show that the best feasible solution found at $n = 25$ is less than about 0.65 for 99 of the 100

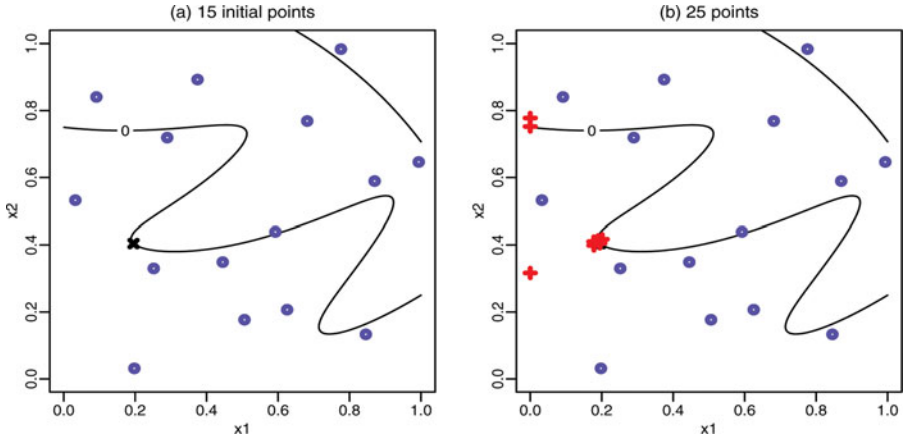

Figure 2. Points chosen by the Schonlau, Welch, and Jones (1998) method when applied to the illustrative test problem of Gramacy et al. (known $f(\mathbf{x})$). The global feasible minimum is at $\times$. Circles show 15 initial points, and the "+" symbols denote 10 additional points, for a total sample size of $n = 25$.

Table 1. Minimum feasible objective found by the Schonlau, Welch, and Jones (1998) method for $n = 25$ and $n = 50$ (known $f(\mathbf{x})$)

	$n = 25$ (15+10)	$n = 50$ (30+20)
95% quantile (worst case)	0.6206	0.6012
Mean	0.6071	0.6001
5% quantile (best case)	0.5999	0.5998

executions of the algorithm. In contrast Figure 2 of Gramacy et al. shows that an *average* of 0.65 is not found by any method until $n > 40$. Our Figure 1(b) shows that by $n = 40$ the global optimum is essentially always found, a better performance than $n = 100$ in Figure 2 of Gramacy et al. The numerical summaries shown in Table 1 for $n = 25$ and $n = 50$ similarly dominate the same metrics presented in the tabular information of Figure 2 of Gramacy et al. For example, for $n = 25$ our 95% quantile of 0.6206 can be compared with a best result of 0.825 across all methods. We do not present any results for $n = 100$ because the Schonlau, Welch, and Jones (1998) method has always already converged by $n = 50$ (actually around $n = 38$).

Figures 2 and 3 illustrate the points chosen for two executions of the algorithm, one for each sample size. It is seen that search mainly concentrates on the global feasible optimum, with occasional search elsewhere. With $n = 50$, as there is little to be gained by continuing to exploit the global optimum, the algorithm makes deeper forays into areas where the objective is even better, just to check whether a feasible solution exists.

We now repeat the above study assuming the objective function is unknown in (1). Because the sample size is fairly large for a simple-to-model, linear objective function, a small nugget (10^{-5}) is added to the correlation matrix to attain numerical stability. Results are given in Figure 4 and Table 2. There is modest deterioration of performance for $n = 25$ relative to the known-objective results but no practical difference for $n = 50$. Thus, even when modeling $f(\mathbf{x})$ too, the worst-case and average results still dominate those in Figure 2 of Gramacy et al.

We now make some summarizing remarks. The method we have illustrated here has a number of conceptual advantages. First, it is simple and follows directly from the properties of GPs. It can be implemented as a straightforward wrapper calling

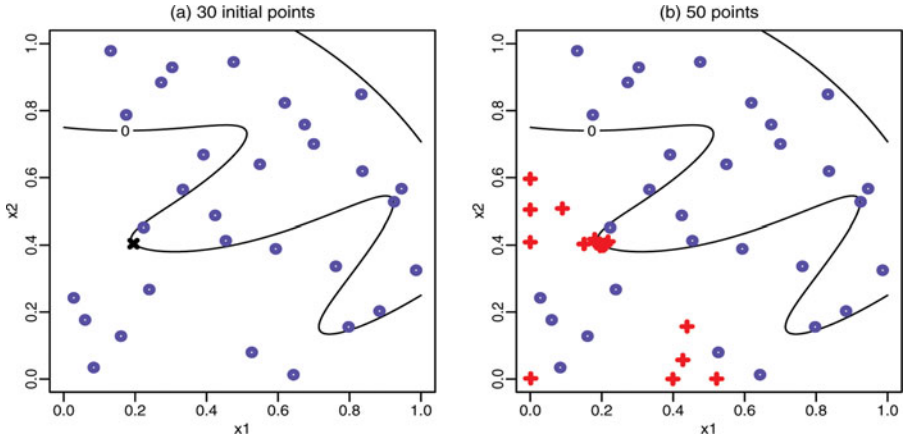

Figure 3. Points chosen by the Schonlau, Welch, and Jones (1998) method when applied to the illustrative test problem of Gramacy et al. (known $f(\mathbf{x})$). The global feasible minimum is at $\times$. Circles show 30 initial points, and the "+" symbols denote 20 additional points, for a total sample size of $n = 50$.

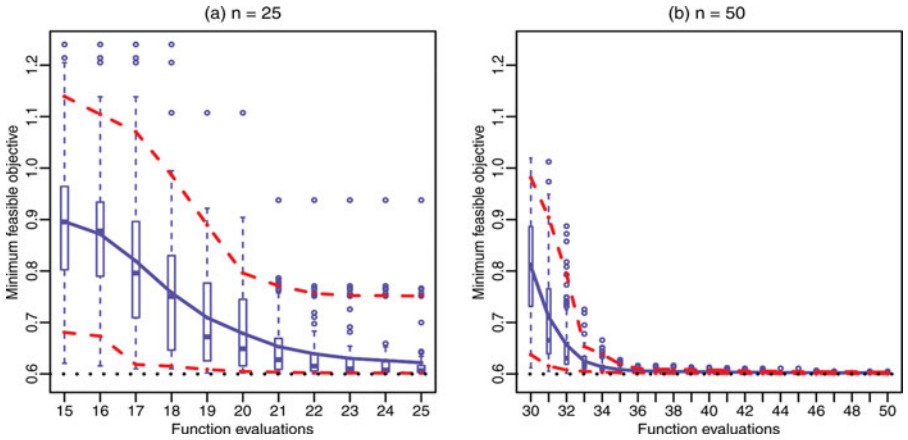

Figure 4. Minimum feasible objective found by the Schonlau, Welch, and Jones (1998) method applied to the illustrative test problem of Gramacy et al. (unknown $f(\mathbf{x})$). The box-and-whisker plots show distributions over 100 executions of the algorithm. The solid line and the two dashed lines represent mean, 5% (best case), and 95% (worst case) summaries, respectively. The dotted line is the global minimum, 0.5998. The total sample size is (a) $n = 25$ or (b) $n = 50$.

Table 2. Minimum feasible objective found by the Schonlau, Welch, and Jones (1998) method for $n = 25$ and $n = 50$ (unknown $f(\mathbf{x})$)

	$n = 25$ (15+10)	$n = 50$ (30+20)
95% quantile (worst case)	0.7512	0.6035
Mean	0.6217	0.6019
5% quantile (best case)	0.6016	0.6002

off-the-shelf GP software. Second, it avoids the need to model highly nonlinear functions generated by maximum operators in the augmented Lagrangian or in the sequence of quadratic problems. We speculate that this enhances predictive accuracy, particularly in critical regions where the constraints are binding. Gramacy et al., too, turn in their Section 3.2 to modeling the constraint (and possibly objective) functions separately. But substituting the surrogates into the augmented Lagrangian causes mathematical or numerical implementation difficulties. Complexities abound.

The method we illustrated is interpreted as based on a constrained expected improvement under an assumption of independent GP models for the various components. Modeling the dependencies, that is, trade-offs, might have advantages for constrained expected improvement as well as augmented Lagrangian methods.

The simple approach we have illustrated works well for the Gramacy et al. test problem. Of course, that is just one example, and the authors' methods are likely superior for some types of problem. We were unable to access the hydrology application, for instance, to make comparisons. If the authors can carry out the computations we have outlined for the application, it might shed light on when one or more of their methods stand out.

In conclusion, the authors have brought the critical problem of constrained blackbox optimization to the attention of statisticians in this work. They are also to be congratulated for bringing together statistical modeling and augmented-Lagrangian approaches. Their methods considerably enhance the range of tools available to practitioners.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council, Canada.

REFERENCES

Aslett, R., Buck, R. J., Duvall, S. G., Sacks, J., and Welch, W. J. (1998), "Circuit Optimization via Sequential Computer Experiments: Design of an Output Buffer," *Applied Statistics*, 47, 31–48. [12]

Gramacy, R. B., Gray, G. A., Le Digabel, S., Lee, H. K. H., Ranjan, P., Wells, G., and Wild, S. M. (2016), "Modeling an Augmented Lagrangian for Blackbox Constrained Optimization," *Technometrics*, 58, 1–11. [12]

Jones, D. R., Schonlau, M., and Welch, W. J. (1998), "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, 13, 455–492. [12]

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. [13]

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–423. [12]

Schonlau, M., Welch, W. J., and Jones, D. R.. (1998), "Global Versus Local Search in Constrained Optimization of Computer Models," in *New Developments and Applications in Experimental Design*, eds. N. Flournoy, W. F. Rosenberger, and W. K. Wong, Hayward, CA: Institute of Mathematical Statistics, pp. 11–25. [12,13]

Comment: "Modeling an Augmented Lagrangian for Blackbox Constrained Optimization" by Gramacy et al.

Yichen CHENG

Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
Seattle, WA 98109
(*ycheng@fredhutch.org*)

Faming LIANG

Department of Biostatistics
University of Florida
Gainesville, FL 32611
(*faliang@ufl.edu*)

We congratulate the authors for such a wonderful contribution to the literature of optimization problems.

Classical optimization methods can be summarized into two categories: mathematical programming approaches and statistical approaches. Each of the approaches has its own merits and shortcomings. The mathematical programming approaches aim at exploitation: given the current information collected, they produce a direction to search for finer solutions at the next iteration. So, this convergence can be fast. However, they often overlook points that are not local, and thus often lead to a local solution. Their statistical counterparts usually have good properties in that global solutions are guaranteed asymptotically. However, such a global search is slower than local search, and in the case

© 2016 American Statistical Association and the American Society for Quality
TECHNOMETRICS, FEBRUARY 2016, VOL. 58, NO. 1
DOI: 10.1080/00401706.2015.1040927