Assessing Convergence in Gaussian Process Surrogate Model Optimization

Nicholas R. Grunloh and Herbert K. H. Lee

Abstract

Identifying convergence in numerical optimization is an ever-present, difficult, and often subjective task. The statistical framework provided by Gaussian Process surrogate model optimization provides useful secondary measures for tracking optimization progress; however the identification of convergence via these criteria has often provided only limited success and often requires a more subjective analysis. Here we use ideas originally introduced in the field of Statistical Process Control to define convergence in the context of an robust and objective convergence heuristic. The Exponentially Weighted Moving Average (EWMA) chart provides an ideal starting point for adaptation to track convergence via the EWMA convergence chart introduced here.

Keywords: Convergence, Derivative-free Optimization, Emulator, Expected Improvement.

1 Introduction

Black-box derivative-free optimization has a wide variety of applications, especially in the realm of computer simulations [11, 6]. When dealing with computationally expensive computer models, a key question is that of convergence of the optimization. Because each function evaluation is expensive, one wants to terminate the optimization as early as possible. However for complex simulators, the response surface may be ill-behaved and optimization routines can easily become trapped in a local mode, so one needs to run the optimization sufficiently long to achieve a robust solution. In this paper, we provide an automated method for assessing convergence of Gaussian Process surrogate model optimization by bringing in elements of Statistical Process Control.

Our motivating example is a hydrology application, the Lockwood pump-and-treat problem [13], discussed in more detail in Section 4.3, wherein contamination in the ground-water near the Yellowstone River is remediated via a set of treatment wells. The goal is to minimize the cost of running the wells while ensuring that no contamination enters the river. The contamination constraint results in a complicated boundary that is unknown in advance and requires evaluation of the simulator, and

thus finding the global constrained minimum is a difficult problem where it is easy for optimization routines to, at least temporarily, get stuck in a local minimum. Without knowing the answer in advance, how do we know when to terminate the optimization routine?

The context of this paper is Gaussian Process surrogate model optimization, a statistical modeling approach to derivative-free numerical optimization that constructs a fast approximation to the expensive computer simulation using a statistical surrogate model [9]. Analysis of the surrogate model allows for efficient exploration the objective solution space. Typically a Gaussian Process (GP) surrogate model is chosen for its robustness, relative ease of computation, and its predictive framework [17]. Arising naturally from the GP predictive distribution [18, 4], the maximum Expected Improvement (EI) criterion has shown to be a valuable criterion for guiding the exploration of the objective function and shows promise for use as a convergence criterion [22, 10].

Taddy et al. [22] considers the use of the improvement distribution for identifying global convergence; stating its value for use in applied optimization. The basic idea behind the use of improvement in identifying convergence is that convergence should occur when the surrogate model produces low expectations for discovering new optimum; that is to say, globally small EI values should be associated with convergence of the algorithm. Thus a simplistic stopping rule might first define some lower EI threshold, then claim convergence upon the first instance of an EI value falling below this threshold, as seen in [3]. This use of EI as a convergence criterion is analogous to other standard convergence identification methods in numerical optimization (e.g., the vanishing step sizes of a Newton-Raphson algorithm). However, applying this same threshold strategy to the convergence of surrogate model optimization has not yet been adequately justified. In fact, this use of EI ignores the nature of the EI criterion as a random variable, and oversimplifies the stochastic nature of convergence in this setting. Thus it is no surprise that this treatment of the EI criterion can result in an inconsistent stopping rule as demonstrated in Figure (1).

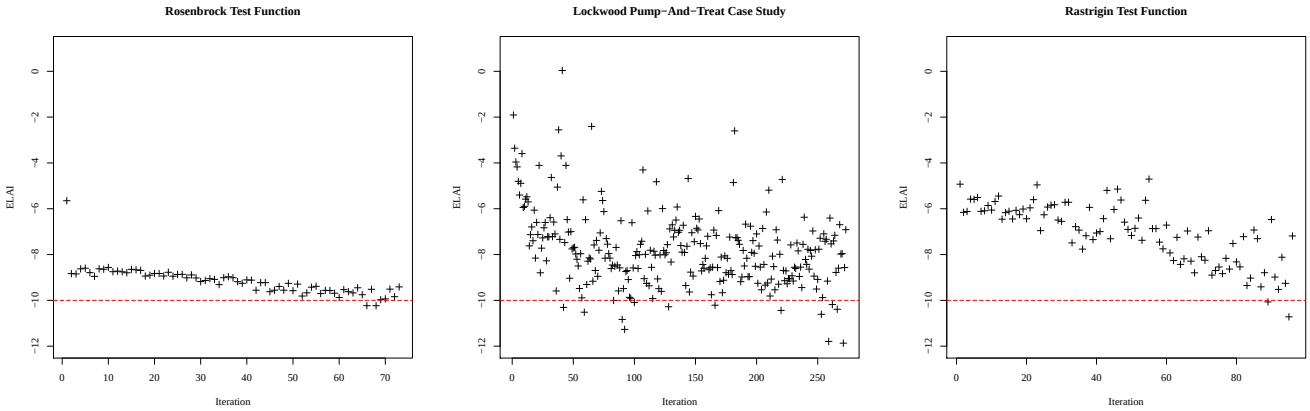


Figure 1: Three ELAI series (more details in Section 4) plotted alongside an example convergence threshold value shown as a dashed line at -10.

Because EI is strictly positive but decreasingly small, we find it more productive to work on the log scale, using a lognormal approximation to the improvement distribution to generate a more appropriate convergence criterion, as described in more detail in Section 3.2. Figure (1) represents three series of the Expected Lognormal Approximation to the Improvement (ELAI) convergence criterion values from three different optimization problems that will be demonstrated later in this paper, where it will be shown that convergence is established near the end of each of these series. These three series demonstrate various ELAI convergence behaviors, and illustrate the difficulty in assessing convergence. In the left-most panel, optimization of the Rosenbrock test function results in a well-behaved series of ELAI values, demonstrating a case in which the simple threshold stopping rule can accurately identify convergence. However the center panel (the Lockwood problem) demonstrates a failure of the threshold stopping rule, as this ELAI series contains much more variance, and thus small ELAI values are observed quite regularly. In the Lockwood example a simple threshold stopping rule could falsely claim convergence within the first 50 iterations of the algorithm. The large variability in ELAI with occasional large values indicates that the optimization routine sometimes briefly settles into a local minimum but is still exploring and is not yet convinced that it has found a global minimum. This optimization run appears to have converged only after the larger ELAI values stop appearing and the variability has decreased. Thus one might ask if a decrease in variability, or small variability, is a necessary condition for convergence. The right-most panel (the Rastrigin test function) shows a case where convergence occurs by meeting the threshold level, but where variability has increased, demonstrating that a decrease in variability is not a necessary condition.

As the Improvement function is itself a random variable, attempting to set a lower threshold bound on the EI, without consideration of the underlying EI distribution over time, over-simplifies the dynamics of convergence in this setting. Instead, we propose taking the perspective of Statistical Process Control (SPC), where a stochastic series is monitored for consistency of the distribution of the most recently observed values. In the next section, we review the statistical surrogate model approach and the use of EI for optimization. In Section 3, we discuss our inspiration from SPC and how we construct our convergence chart. Section 4 provides synthetic and real examples, and then we provide some conclusions in the final section.

2 Gaussian Process Surrogate Model Optimization

The primary motivation for the use of surrogate modeling in optimization is to manage a computationally challenging objective function with the use of a fast and relatively simple functional working model (i.e. the surrogate model) of the problem function. The surrogate model serves as an efficient tool for using function evaluations to infer the expected behavior of the objective function and thus determine where further optima may exist with minimal evaluation of the complex objective function itself. Surrogate modeling is therefore useful for optimizing large computer simulations experiments, where each function evaluation may consume considerable computational resources, while the surrogate model can be evaluted quickly. The standard surrogate model in the literature for analysis of computer experiments is a Gaussian Process (GP) [16, 17]. A GP is a stochastic process such that when evaluated at any finite collection of points, that collection follows a multivariate Gaussian distribution. A GP is defined by its mean function and its covariance function, and various standard formulations exist [1, 21]. Most formulations take advantage of a large degree of smoothness, reflecting a modeling assumption of smoothness in the output of the simulator, in that if the simulator is evaluated at two nearby inputs, then one expects the resulting outputs to be relatively close. A GP can interpolate, which can be useful for a deterministic simulator, or it can smooth, which has a number of practical advantages even for deterministic simulators [8].

In many cases the assumption of a globally smooth f with a homogeneous uncertainty structure can provide an effective and parsimonious model. However for the sake of providing a flexible

surrogate model, it is desirable to have the ability to loosen these restrictions in cases when f may have inherently sharp boundaries, or when numerical simulators have variable stability in portions of the domain. Gramacy and Lee [7] use the idea of allowing subpopulations of flexibility via a treed partitioning of the domain, fitting stationary GP surfaces to separate portions of f. The domain is recursively sub-partitioned and separate hierarchically-linked GP models are fit within each sub-partition. The partitioning scheme is fit via a reversible jump MCMC algorithm, jumping between models with differing partitioning schemes, and averaging over the full parameter space to provide smooth predictions except where the data call for a discontinuous prediction. Partitioning the domain in this way allows parsimonious surrogate models in simple objective function cases and quite flexible surrogate models when the the objective function displays complex behavior. For further explanation of partitioned Gaussian process models as well as notes on implementing such models in R, see the R package `tgp` [5, 9]. Because many of the objective functions of interest are not well modeled by a stationary GP, we use treed GPs as our surrogate models in this paper, but our approach is easily adaptable to a wide variety of surrogate models.

2.1 Expected Improvement

The EI criterion predicts how likely it will be to find a new minimum at a given location based on the predictive surrogate model. EI is built upon the improvement function [10]:

$$\mathrm{I}(\boldsymbol{x}) \;=\; \max\Big\{\big(f_{min} - f(\boldsymbol{x})\big),\; 0\Big\}, \tag{1}$$

where f_{min} is the smallest function value observed so far. EI is the expectation of the improvement function with respect to the posterior predictive distribution of the surrogate model, $\mathbb{E}[\,\mathrm{I}(\boldsymbol{x})\,]$. EI rewards candidates both for having a low predictive mean, as well high uncertainty (where the function has not been sufficiently explored). By definition the improvement function is always non-negative, and the GP posterior predictive $\mathbb{E}[\,\mathrm{I}(\boldsymbol{x})\,]$ is strictly positive. EI is available in closed form for a stationary GP, and for other models can be quickly estimated using Monte Carlo posterior predictive samples at given candidate locations.

2.2 Optimization Procedure

Figure 2: Optimization Procedure

1) Collect an initial set, $\boldsymbol{X}$.
2) Compute $f(\boldsymbol{X})$.
3) Fit surrogate based on evaluations of f.
4) Collect a candidate set, $\tilde{\boldsymbol{X}}$.
5) Compute EI among $\tilde{\boldsymbol{X}}$
6) Add $\text{argmax}_{\tilde{\boldsymbol{x}}_i} \mathbb{E}[\, \text{I}(\tilde{\boldsymbol{x}}_i)\,]$ to $\boldsymbol{X}$.
7) Check convergence.
8) If converged exit. Otherwise go to 2).

Optimization can be viewed as a sequential design process, where locations are selected for evaluation on the basis of how likely they are to decrease the objective funciton, i.e., based on the EI. Optimization begins by initially collecting a set, $\boldsymbol{X}$, of locations to evaluate the true function, f, to gather a basic impression of f. A statistical surrogate model is then fitted with $f(\boldsymbol{X})$ as observations of the true function. Using the surrogate model, a set of candidate points, $\tilde{\boldsymbol{X}}$, are selected from the domain and the EI criterion is calculated among these points. The candidate point that has the highest EI is then chosen as the best candidate for a new minimum and thus, it is added to $\boldsymbol{X}$. The objective function is evaluated at this new location and the surrogate model is refit based on the updated $f(\boldsymbol{X})$. The optimization procedure carries on in this way until convergence. The key contribution of this paper is an automated method for checking convergence, which we develop in the next section.

3 EWMA Convergence Chart

3.1 Statistical Process Control

In Shewhart's seminal book [20] on the topic of control in manufacturing, Shewhart explains that a phenomenon is said to be in control when, "through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future." This notion provides an instructive framework for thinking about convergence because it offers a natural way to consider the distributional characteristics of the EI as a proper random variable. In its most simplified form, SPC considers an approximation of a statistic's sampling distribution as repeated sampling occurs in time. Thus Shewhart can express his idea of control as the expected behavior of random observations from

this sampling distribution. For example, an $\bar{x}$-chart tracks the mean of repeated samples (all of size n) through time so as to expect the arrival of each subsequent mean in accordance with the known or estimated sampling distribution for the mean, $\bar{x}_j \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. By considering confidence intervals on this sampling distribution we can easily draw explicit boundaries (i.e., control limits) to identify when the process is in control and when it is not. Observations violating our expectations (falling outside of the control limits) indicate an out-of-control state. Since neither μ nor σ^2 are typically known, it is of primary importance to use the data carefully to form accurate approximations of these values, thus establishing a standard for control. Furthermore, this logic relies upon the typical asymptotic results of the central limit theorem (CLT), and care should be taken to verify the relevant assumptions required.

It is important to note that we are not performing traditional SPC in this context, the EI criterion will be stochastically decreasing as an optimization routine proceeds. Only when convergence is reached will the EI series look approximately like an in-control process. Thus our perspective is completely reversed from the traditional SPC approach—we start with a process that is out of control, and we determine convergence when the process stabilizes and becomes locally in control. An alternative way to think about our approach is to consider performing SPC backwards in time on our EI series. Starting from the most recent EI observations and looking back, we declare convergence if the process starts in control and then becomes out of control. This pattern generally appears only when the optimization has progressed and reached a local mode. If the optimization were still proceeding, then the EI would still be decreasing and the final section would not appear in control.

3.2 Expected Lognormal Approximation to the Improvement (ELAI)

For the sake of obtaining a robust convergence criterion to track via SPC, it is important to carefully consider properties of the improvement distributions which generate the EI values. The improvement criterion is strictly positive but decreasingly small, thus the improvement distribution is often strongly right skewed, and the EI is far from normal. Additionally, this right skew becomes exaggerated as convergence approaches, due to the decreasing trend in the EI criterion. Together these characteristics of the improvement distribution give the EI criterion inconsistent behavior for tracking convergence via a typical $\bar{x}$-chart.

These issues naturally suggest releasing the bound at 0 by modeling transformations of the improvement, rather than directly considering the improvement distribution on its own. One of the simplest of the many possible helpful transformations in this case would consider the log of the improvement distribution. However due to the MCMC sample-based implementation of the Gaussian Process, and the desire for a large number of samples from the improvement distribution, it is not uncommon to obtain at least one sample that is computationally indistinguishable from 0 in double precision. Thus simply taking the log of the improvement samples can result in numerical failure, particularly as convergence approaches, even though the quantities are theoretically strictly positive. Despite this numerical inconvenience, the distribution of the improvement samples is often very well approximated by the Lognormal distribution.

We avoid the numerical issues by using a model-based approximation. With the desire to model $\mathbb{E}[\ \log \mathrm{I}\ ] \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$, we switch to a Lognormal perspecive. Recall that if a random variable $X \sim \textit{Log-}N(\psi, \nu)$, then another random variable $Y = \log(X)$ is distributed $Y \sim N(\psi, \nu)$. Furthermore, if m and v are, respectively, the mean and variance of a lognormal sample, then the mean, ψ, and variance, ν, of the associated normal distribution are given by the following relation.

$$\psi = \log\left(\frac{m^2}{\sqrt{v+m^2}}\right) \qquad \nu = \log\left(1 + \frac{v}{m^2}\right). \tag{2}$$

Using this relation we do not need to transform any of the improvement samples. We compute the empirical mean and variance of the unaltered, approximately lognormal, improvement samples, then use relation (2) to directly compute ψ as the Expectation under the Lognormal Approximation to the Improvement (ELAI). The ELAI convergence criterion is a useful convergence criterion in this case because of the reduced right skew of the log of the improvement distribution, and the ELAI convergence criterion serves as a computationally robust approximation of the $\mathbb{E}[\ \log \mathrm{I}\ ]$ under reasonable log-normality of the improvements. Furthermore, both the $\mathbb{E}[\ \log \mathrm{I}\ ]$ and ELAI convergence criterion are distributed approximately normally in repeated sampling. This construction allows for more consistent and accurate use of the fundamental theory on which our SPC perspective depends.

3.3 Exponentially Weighted Moving Average

The Exponentially Weighted Moving Average (EWMA) control chart [12, 19] elaborates on Shewhart's original notion of control by viewing the repeated sampling process in the context of a moving average smoothing of series data. Pre-convergence ELAI evaluations tend to be variable and overall decreasing, and so do not necessarily share distributional consistency among all observed values. Thus a weighted series perspective was chosen to follow the moving average of the most recent ELAI observations while still smoothing with some memory of older evaluations. EWMA achieves this robust smoothing behavior, relative to shifting means, by assigning exponentially decreasing weights to successive points in a rolling average among all of the points of the series, thus the EWMA emphasizes recent observations and shifts the focus of the moving average to the most recent information while still providing shrinkage towards the global mean of the series.

If Y_i is the current ELAI value, and Z_i is the EWMA statistic associated with this current value, then the initial value Z_0 is set to Y_0 and for $i > 0$ the EWMA statistic is expressed as,

$$Z_i = \lambda Y_i + (1 - \lambda) Z_{i-1}. \tag{3}$$

Above, λ is a smoothing parameter that defines the weight (i.e. $0 < \lambda \leq 1$) assigned to the most recent observation, Y_i. The recursive expression of the statistic ensures that all subsequent weights geometrically decrease as they move back through the series.

Typical values of λ can range from $0.1 \leq \lambda \leq 0.3$, with a default value of λ around 0.2, as described by Box et al. [2]. Large values of λ assign more weight to recent observations in the series, allowing for a more flexible fit for unstable series. However, the choice of a large λ may over-fit the Z_i to noise in the Y_i. It is thus desirable to choose to the smallest λ which still provides good forecasts of future observations in the series. Box et al. [2, p. 87] explains how to choose an optimal value for λ by

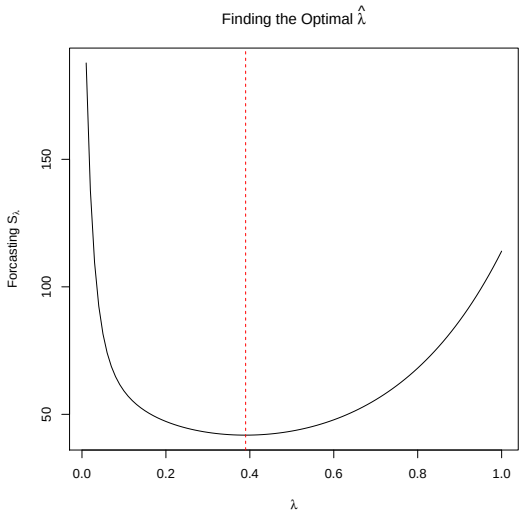


Figure 3: S_λ as calculated over a fine grid of possible λ values for ELAI values derived from optimization of the Rastrigin test function. The optimal forecasting $\hat{\lambda}$ is shown by the vertical dashed line.

choosing the $\hat{\lambda}$ which minimizes the sum of squared fore-

casting deviations (S_λ) for each new observation. Through this analysis of S_λ, as seen in Figure (3), is it evident that EWMA charts can be very robust to reasonable choices of λ, due to the small first and second derivatives of S_λ for a large range of sub-optimal choices of λ around $\hat{\lambda}$. In fact, Figure (3) shows that for $\lambda \in [0.2, 0.6]$, S_λ stays within 10% of its the minimum possible value.

It is interesting to note that for the example series used in Figure (3), the optimal $\hat{\lambda} \approx 0.4$ exceeds the recommended upper limit of 0.3 for λ. Discrepancies between the optimal values of λ chosen here, and the typically chosen values of λ can be naturally attributed to the differing context in which we apply EWMA as compared to the typical SPC application. The typical use of EWMA in SPC begins with the premise of a relatively stable (in-control) series and attempts to identify new out-of-control observations which would indicate some change in the data generating process. However our use of EWMA, to identify convergence, begins with an out-of-control series and we wish to identify when the series falls into control (i.e. convergence). As a result, ELAI values for tracking convergence are inherently less stable than typical SPC applications of EWMA. Due to the decreased stability of the series, in this context, the optimal forecasting $\hat{\lambda}$ may often fall above the traditionally recommended upper limit for λ, in-order to better follow the more active moving averages inherent to the unstable pre-convergence series. In this context we want to reiterate that while it is useful to borrow the EWMA machinery often used in SPC, we are approaching the whole process backwards, in that we are starting with an "out of control" process and waiting to see when it settles down into control, and thus our approach should be viewed as SPC-inspired rather than a formal application of SPC.

Identifying convergence relies upon carefully defining the control limits on the EWMA statistic. As in the simplified $\bar{x}$-chart, defining the control limits for the EWMA setting amounts to considering an interval on the sampling distribution of interest. In the EWMA case we are interested in the sampling distribution of the Z_i. Assuming that the Y_i are *i.i.d.* then Lucas and Saccucci [12] show that we can write $\sigma^2_{Z_i}$ in terms of σ^2_Y.

$$\sigma^2_{Z_i} = \sigma^2_Y \left(\frac{\lambda}{2-\lambda} \right) \left[1 - (1-\lambda)^{2i} \right] \tag{4}$$

Thus if $Y_i \overset{i.i.d.}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$, then the sampling distribution for Z_i is $Z_i \sim N\left(\mu, \sigma^2_{Z_i}\right)$. Furthermore by

choosing a confidence level through choice of a constant c, the control limits based on this sampling distribution are seen in Eq. (5).

$$\mathrm{CL}_i = \mu \pm c\sigma_{Z_i} = \mu \pm c\,\frac{\sigma}{\sqrt{n}}\,\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]} \tag{5}$$

Notice that since $\sigma^2_{Z_i}$ has a dependence on i, the control limits do as well. Looking back through the series brings us away from the focus of the moving average, at i, and thus the control limits widen until the limiting case as, $i \to \infty$, the control limits are defined by $\mu \pm c\,\sqrt{\frac{\lambda\sigma^2}{(2-\lambda)n}}$.

At first glance it is not clear that the Y_i are in fact *i.i.d.* Indeed the early iterations of the convergence processes seen in Figure (1) certainly do not display *i.i.d.* Y_i. However as the series approaches convergence, the Y_i eventually do enter a state of control see Figure (**??**). For these controlled Y_i an *i.i.d.* approximation is quite reasonable. The realization of such a controlled region of the series defines the notion of consistency which is what allows for the identification of convergence.

3.4 The Control Window

The final structural feature of the EWMA convergence chart for identifying convergence is the so called *control window*. The control window contains a fixed number, w, of the most recently observed Y_i. Only information from the w points currently residing inside the control window is used to calculate the control limits, but the EWMA statistic is still computed for all Y_i values. Initially, the convergence algorithm is allowed to fill the control window, by collecting an initial set of w observations of the Y_i. As new observations arrive, the oldest Y_i value is removed from the control window, thus allowing for the inclusion of a new Y_i.

The purpose of the control window is two-fold. Firstly it serves to dichotomizes the series for evaluating subsets of the Y_i for distributional consistency. Secondly it offers a structural way for basing the standard for consistency (i.e. the control limits) only on the most recent and relevant information in the series.

The size of the control window, w, may vary from problem to problem based on the difficulty of optimization in each case. A reasonable way of choosing w is to consider the number of observations neccissary to establish a standard of control. The choice of w will naturally increase as the difficulty

of the optimization problem increases. Just as in other sample size calculations, the the choice of an optimal w must consider the cost of premature identification of convergence (i.e. poor inference) associated with underestimating w, against the cost of continuing to sample after convergence has occurred (i.e. the cost of over sampling) associated with overestimating w. Providing a default choice of w is somewhat arbitrary without careful analysis of the particulars of the objective function behavior and the costs of each successive objective function evalutation, however as a recommendation for starting this analysis, one may consider $w \geq 15p$ as a rule of thumb. This recommendation considers the dimensionality, p, of the objective function as well as represents the prior assertion that premature identification of convergence is a worse error than computing extraneous objective function evaluations.

For example, if the objective function must be searched over a large domain, particularly in many dimensions, optimization will naturally take many function evaluations to gather adequate information to reflect confident identification of convergence. Thus the EI criterion, and by extension the ELAI criterion, may display high variance, associated with high uncertainty, as well as be slow to decrease in mean value from the initial state of preconvergence into convergence. Jointly the high ELAI variance and the slow decreasing mean value of the ELAI criterion may make it hard to identify convergence; in these cases large values of w are required to discern this relatively slight signal in the context of increased noise. Similar effects may be observed for highly multimodal objective functions, as the regular discovery of new modes will increase the variance of the ELAI criterion, and disguise any decreasing mean value, among the noise inherent to the search of such functions.

By contrast, strongly unimodal functions will enjoy a relatively fast decrease in the ELAI criterion in the presence of relatively small variability. This higher signal-to-noise ratio makes for easier identification of convergence, and thus allows for a smaller choice of w. However if w is chosen to be too small, the algorithm may be over eager to claim convergence and the reccomendation of $w \geq 15p$ is particularly apt here to guard against false identification of convergence.

3.5 Identifying Convergence

In identifying convergence, we not only desire that the ELAI convergence criterion reaches a state of control, but we desire that the ELAI series demonstrates a move from a state of pre-convergence to a

consistent state of convergence. To recognize the move into convergence we combine the notion of the control window with the EWMA framework to construct the so called, *EWMA Convergence Chart*. Since we expect EI values to decrease upon convergence, the primary recognition of convergence is that new ELAI values demonstrate values that are consistently lower than initial pre-converged values.

Firstly, we require that all exponentially weighted Z_i values inside the control window fall within the control limits. This ensures that the most recent ELAI values demonstrate distributional consistency within the bounds of the control window. Secondly, since we wish to indicate a move from the initial pre-converged state of the system, we require at least one point, from beyond the initial control window, to fall outside the defined EWMA control limits. This second rule suggests that the new ELAI observations have established a state of control which is significantly different from the previous pre-converged ELAI observations. Jointly enforcing these two rules implies convergence based on the notion that convergence enjoys a state of consistently decreased expectation of finding new minima in future function evaluations.

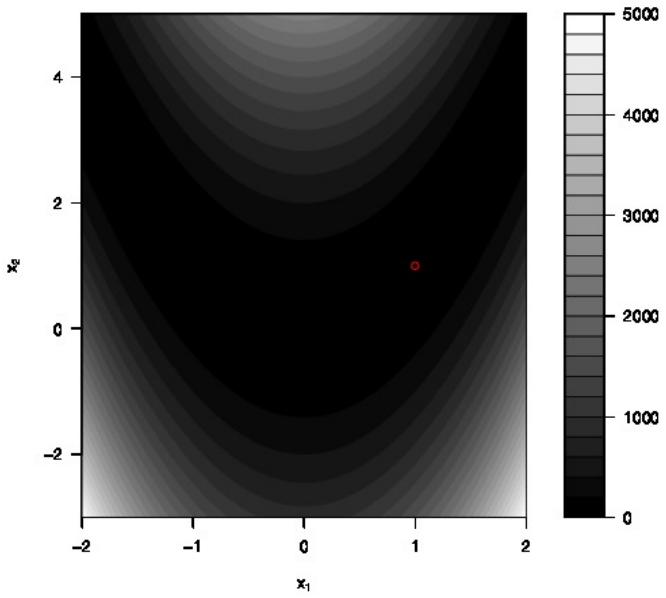
Considering the optimization procedure outlined in Figure (2), the check for convergence indicated in step 7) amounts to computing new EWMA Z_i values, and control limits, from the inclusion of the most recent observation of the improvement distribution, and checking if the subsequent set of Z_i satisfy both of the above rules of the EWMA convergence chart. Satisfying one, or none, of the convergence rules indicates insufficient exploration and further iterations of optimization are required to gather more information about the objective function.

4 Examples

We first look at two synthetic examples from the optimization literature, where the true optimum is known, so we can be sure we have converged to the true global minimum. We then provide a real world example from hydrology.

4.1 Rosenbrock

The Rosenbrock function [15] was an early test problem in the optimization literature. It combines a narrow, flat parabolic valley with steep walls, and thus it can be difficult for gradient-based methods. It is generalizable to higher dimensions, but we use the two-dimensional version here.



$$\begin{aligned} f(x_1, x_2) &= 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2 \\ \text{Minimum} &: f(1, 1) = 0 \end{aligned} \tag{6}$$

Convergence is non-trivial to assess, because optimization routines can take a while to explore the relatively flat, but non-convex, valley floor for the global minimum. Here we focus on the region $-2 \leq x_1 \leq 2$, $-3 \leq x_2 \leq 5$. While the region around the mode presents some minor challenges, this problem is unimodal, and thus represents a relatively easier optimization problem. This example illustrates a well-behaved convergence process.

We estimate λ via the minimum S_λ estimator, $\hat{\lambda} \approx 0.5$, and use the minimum default value $w = 30$. Figure 4 shows the result of surrogate model optimization at convergence, as assessed by our method. The right panel shows the best function value (y-axis) found so far at each iteration (x-axis), and verifies that we have found the global minimum. The left panel shows our convergence chart, with the window indicated by the vertical line. The dashed horizontal lines show the control

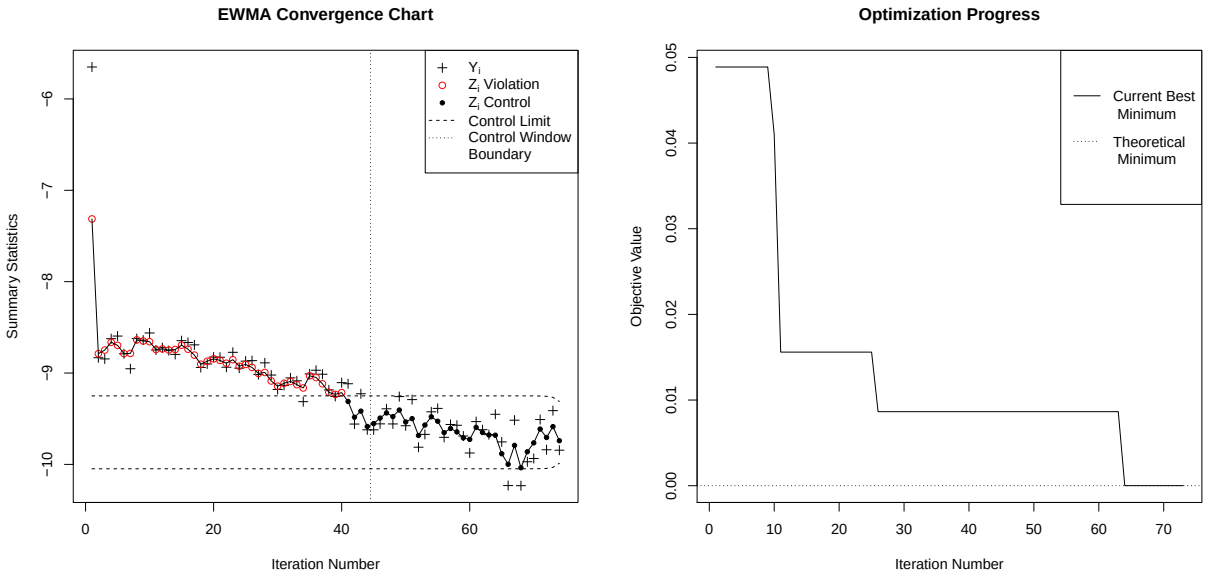
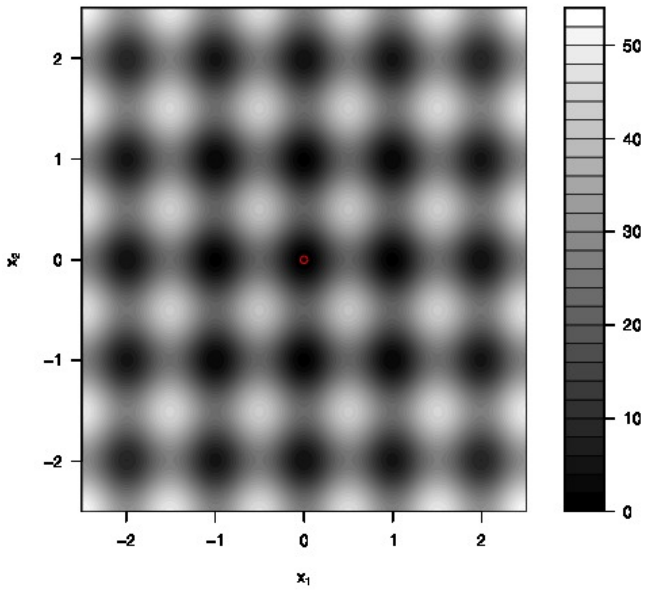


Figure 4: Rosenbrock function: Convergence chart on the left, optimization progress on the right.

limits. At iteration 74 is the first time that all EWMA points in the control window are within the control limits, and thus we declare convergence. Note that the EWMA points generally trend downward until the global minimum has been found.

4.2 Rastrigin

The Rastrigin function is a commonly used test function for evaluating the performance of global optimization schemes such as gentetic algorithms [23]. The global behavior of Rastrigin is dominated by the spherical function, $\sum_i x_i^2$, however Rastrigin has been oscillated by the cos(.) function, and shifted, so that it achieves a global minimum value of 0 at the lowest point, of its lowest trough at (0, 0).



$$\begin{aligned} f(x_1, x_2) &= \sum_{i=1}^{2} \left[x_i^2 - 10\cos(2\pi x_i)\right] + 2(10) \\ \text{Minimum} &: f(0,0) = 0 \end{aligned} \tag{7}$$

Rastrigin is generalizable to an arbitrarly many dimensions, but to develop intution, this example considers Rastrigin over the 2 dimensional square $-2.5 \leq x_i \leq 2.5$. Rastrigin is a highly multimodal function, and as such, the many similar modes present a challenge for identifying convergence. The multimodality of this problem increases the variablility of the EI criterion, and thus represents a moderately difficult optimization problem in this context. It should be noted that by increasing the size of the search domain, either by increasing the bounds of the search square and/or increasing the dimension of the domain would make this example considerably more difficult and a less intuitive example for developing the choice of w.

$\hat{\lambda}$ in this example is calculated to be about 0.4 . The decreased value of $\hat{\lambda}$, relative to Rosebrock, increases the smoothing capabilities of EWMA proceedure, as a response to the increased noise in th ELAI series.The added noise of the ELAI criterion, in this case, comes from the regular discovery of dramatic new modes as optimization proceeds. Due to the increased complexity of Rastrigin relative to Rosenbrock, a larger w is needed to recognize convergence in the presence of increased noise in the ELAI criterion. In this application $w = 40$, was chosen by experimentation. Although larger

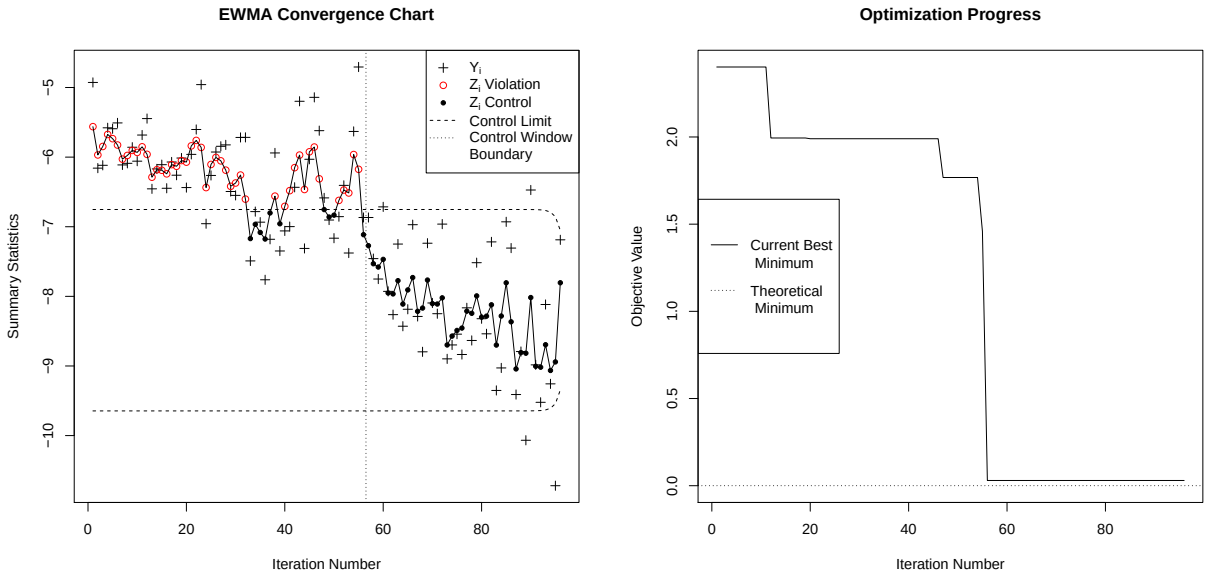


Figure 5: Rastrigin function: Convergence chart on the left, optimization progress on the right.

choices of w produce equally consistent identification of convergence, they do so with more function evaluations.

Figure (5) shows the convergence chart (left) and the optimization progress of the algorithm (right) after 95 iterations of optimization. Although the variablity of the ELAI crterion increases as optimization proceeds, large ELAI values stop arriving after iteration 55, coincidently with the surrogate model's discovery of the Rastrigin's main mode, as see in the right panel of Figure (5). Furthermore notice that optimization progress in Figure (5, right) demonstrates that convergence in this case does indeed represent approximate identification of the theoretical minimum of the the function, as indicated by the dashed horizontal line at the theoretical minimum.

4.3 Lockwood Case Study

The previous examples have focused on analytical functions with known minima. This is done for the sake of developing an intuition for tuning the EWMA convergence chart parameters and to ensure that our methods correspond to the identification of real optima. Here we apply the EWMA convergence chart in the practical optimization setting of pump and treat optimization problems as formulated by Mayer et al. (2002) [14]. Specifically we consider the Lockwood pump and treat problem, originally presented by Matott et al. [13].

The Lockwood pump and treat case study considers an industrial site, along the Yellowstone River, in Montana, with groundwater contaminated by chorinated solvents. If left untreated, this contaminated groundwater may contaminate the Yellowstone river, as dictated by the hydrology of the system. Inorder to control this contaminated groundwater, a total of six pumps situated over two plumes, of the contaminated groundwater, are used to redirect groundwater away from the river to a treament facility. Due to the cost of running these pumps, over a long period of time, it is desirable to determine how to best allocate the pumping effort amoung these pumps so as to determine the lowest cost pumping strategy to protect the river. Pumping each of these six wells at different rates can drastically change the groundwater behavior, and thus a numerical simulation of the system is required to predict the behavior of the system at a given set of pumping rates.

The objective function, $f(\boldsymbol{x})$, to be minimized in this case, can be expressed as the sum of the pumping rates for each pump (a quanity proportional to the expense of running the pumps in USD), with additional large penalties associated with any contamination of the river.

$$f(\boldsymbol{x}) = \sum_{i=1}^{6} x_i + 2\big[c_a(\boldsymbol{x}) + c_b(\boldsymbol{x})\big] + 20000\big[\mathbb{1}_{c_a(\boldsymbol{x})>0} + \mathbb{1}_{c_b(\boldsymbol{x})>0}\big] \tag{8}$$

Here $c_a(\boldsymbol{x})$ and $c_b(\boldsymbol{x})$ are outputs of a simulation, indicating the amount of contamination, if any, of the river as a function of the pumping rates, $\boldsymbol{x}$, for each of the six wells. Any amount of contamination of the river results in a large stepwise penalty which introduces a discontinuity into the objective function, at the contaimination boundary. Each x_i is bounded on the large interval, $0 \leq x_i \leq 20,000$, representing a large range of possible management schemes. The full problem defines a six-dimensional optimization problem, to determine the optimal rate at which to pump each well, so as to minimize the loss function defined in Eq (8). Since the loss function is defined over a large and continuous domain, and running the numerical simulation of the system is computationally expensive, this example presents an ideal situation for use with surrogate model based optimization.

Again λ was chosen via the minimum S_λ estimator to be $\hat{\lambda} \approx 0.4$ in this case.This level of smoothing is required here to reduce the noise in ELAI criteion due to the large search domain, as well as the complicated contamination boundary amoung the six wells. Furthermore these featres of the objective function complicate fit of the surrogate model and thus more function evaluations are

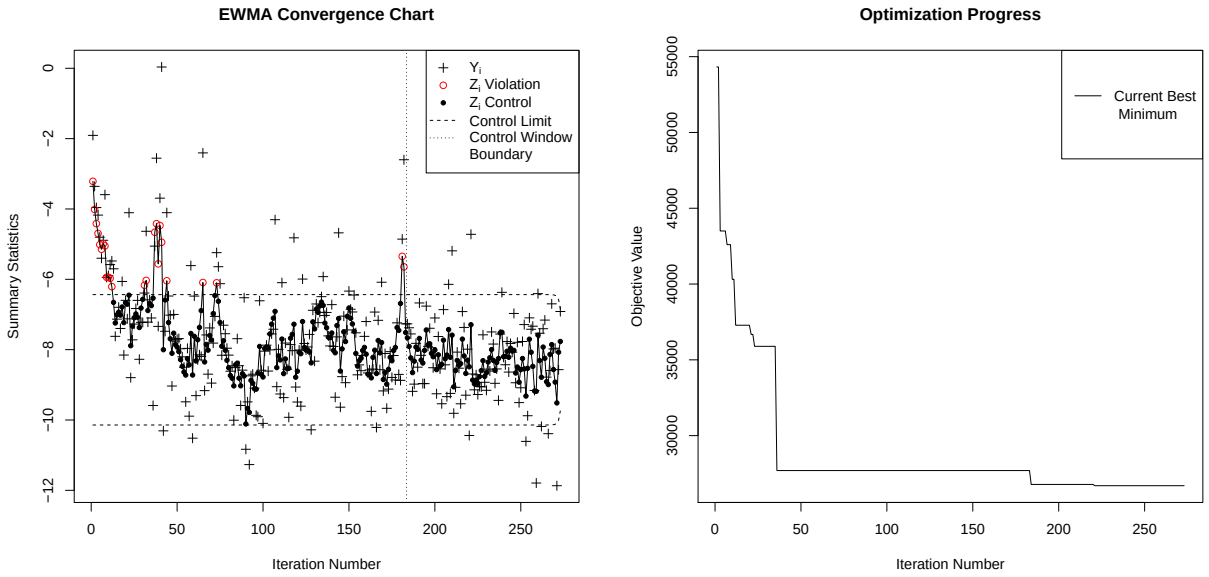


Figure 6: Lockwood Case-study: Convergence chart on the left, optimization progress on the right.

required to produce an accurate model of f. As a result, the control window size, w, must increase to provide the initial surrogate model enough information to yeild reasonable accuracy. Here w was chosen to be 90 iterations, as determined by the adequate initial surrogate model behaviour as well as consistent identification of convergence.

The convergence chart for monitoring the optimization of the lockwood case study, is shown in the left panel of Figure (6), as computed with $\hat{\lambda} \approx 0.4$ and $w = 90$.Convergence in this case does not occure with a dramatic shift in the mean level of the ELAI criterion, but rather convergence occures as optimization the series stabalizes after large ELAI value move beyonfd the control limit.Interestingly the last major spike in the ELAI series is observed alongside the disovery of the final major jump in the current best minimum value as seen at about iteration 180 in the right panel of Figure (6). The EWMA convergence chart identifies convergence as the EWMA statistic associated with this final ELAI spike eventally exits the control window at iteration 270. The solution shown here corresponds to $f(\boldsymbol{x}) \approx 26696$ at $\boldsymbol{x} \approx [0, 6195, 12988, 3160, 1190, 3163]$. This solution is well corroborated as a point of dimishing returns for this problem, by the analysis of Gramacy et al. [6] on the same problem, as seen their average EI surrogate modeling behavior.

5 Conclusion

Adapting the notion of control from the SPC literature, the EWMA convergence chart outlined here aims to provide a objective standard for identifying convergence.The examples provided demonstrate how the EWMA convergence chart may accurately, and efficiently, identify convergence in the context of statistical surrogate model optimization.

As for any optimization algorithm, a converged solution may only be considered as good as the algorithms consideration of f. Thus poorly tuned surrogate modeling strategies may never optimize f to their fullest extent, but the EWMA convergence chart presented here may still claim convergence in these cases. The EWMA convergence chart may only consider convergence in the context of the algorithm in which it is imbeded, and thus should be interpreted as a means of identifying when an algorithm has converged rather than when the lowest minimum has been found. For poorly tuned surrogate modeling strategies the EWMA convergence chart may only identify that the algorithm has reached a point of diminishing returns; for correctly tuned surrogate modeling strategies this point should correspond with the realization of an optimal solution. In either poor or correct surrogate tuning, the EWMA convergence chart identifies the moment at which it is beneficial to stop iterating the routine and reflect upon the results.

Admmitadly the use of the EWMA convergence chart comes with the addition of it's own parameters which themselves require estimation. The addition of these parameters can be easily justified under a divide and conquer mentality; thus replacing the original large subjective task of appropriatly identifying convergence with relatively simple parameter estimation problems. The choice of λ has been shown to be relatively robust to suboptimal choices, and furthermore estimation of the minimum sum of the squared forcasting errors $\hat{\lambda}$ is a simple in practice. The estimation of w is more subtle, but follows from reasonable intuition of the problem. The choice of w would ideally consider an objective measure of the complexity of f as well as the dimensionality of the domain, p. For simplicity the recommendation $w \geq 15p$ has shown to work quite well, although it contains no explicate consideration of the observed complexity of f.

- tuning parameters added in the spirit of reducing hard problems into a series of esier ones
  - convergence is hard and massively subjective

  - interpreting convergence charts is easier
  - tuning λ is objective and robust
  - tuning w can be subjective (requires large simulation study to choose.)

- choose w
  - complexity of f (??entropy??)
  - Dimension of f
  - Emplerical results here: $w \approx 15p$; p is the dimension

- 2-parameter box-cox EI transformation instead of ELAI

References

[1] Petter Abrahamsen. *A review of Gaussian random fields and correlation functions*. Norsk Regnesentral/Norwegian Computing Center, 1997.

[2] George E. P. Box, Alberto Luceño, and María Del Carmen Paniagua-Quiñones. *Statistical Control by Monitoring and Adjustment*. Wiley, New York, NY, 1997.

[3] Sanket Sanjay Diwale, Ioannis Lymperopoulos, and Colin Jones. Optimization of an airborne wind energy system using constrained gaussian processes. In *IEEE Multi-Conference on Systems and Control*, 2014.

[4] David Ginsbourger, Rodolphe Le Riche, Laurent Carraro, et al. A multi-points criterion for deterministic parallel global optimization based on gaussian processes. In *Journal of Global Optimization, in revision*. Citeseer, 2009.

[5] Robert B. Gramacy. tgp: an r package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software*, 19(9):6, 2007.

[6] Robert B Gramacy, Genetha A Gray, Sebastien Le Digabel, Herbert KH Lee, Pritam Ranjan, Garth Wells, and Stefan M Wild. Modeling an augmented lagrangian for blackbox constrained optimization. *arXiv preprint arXiv:1403.4890*, 2014.

[7] Robert B. Gramacy and Herbert H. K. Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483), 2008.

[8] Robert B Gramacy and Herbert KH Lee. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012.

[9] Robert B. Gramacy and Matthew Taddy. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed gaussian process models. *Journal of Statistical Software*, 33(i06), 2012.

[10] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[11] Tamara G. Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45:385–482, 2003.

[12] James M. Lucas and Michael S. Saccucci. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12, 1990.

[13] Shawn L. Matott, Kenny Leung, and Junyoung Sim. Application of matlab and python optimizers to two case studies involving groundwater flow and contaminant transport modeling. *Computers & Geosciences*, 37(11):1894–1899, 2011.

[14] Alex S Mayer, CT Kelley, and Cass T Miller. Optimal design for problems involving flow and transport phenomena in saturated subsurface systems. *Advances in Water Resources*, 25(8):1233–1256, 2002.

[15] Howard H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184, 1960.

[16] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[17] Thomas J. Santner, Brian J. Williams, and William Notz. *The design and analysis of computer experiments*. Springer, 2003.

[18] Matthias Schonlau, William J. Welch, and Donald R. Jones. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series*, pages 11–25, 1998.

[19] Luca Scrucca. qcc: an r package for quality control charting and statistical process control. *R News*, 4/1:11–17, 2004.

[20] Walter A. Shewhart. *Economic control of quality of manufactured product*, volume 509. ASQ Quality Press, 1931.

[21] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer, 1999.

[22] Matthew A. Taddy, Herbert H. K. Lee, Genetha A. Gray, and Joshua D. Griffin. Bayesian guided pattern search for robust local optimization. *Technometrics*, 51(4):389–401, 2009.

[23] Darrell Whitley, Soraya Rana, John Dzubera, and Keith E. Mathias. Evaluating evolutionary algorithms. *Artificial Intelligence*, 85(1-2):245–276, 1996.