Determining Convergence in Gaussian Process Surrogate Model Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Identifying convergence in numerical optimization is an ever-present, difficult, and often subjective task. The statistical framework of Gaussian process surrogate model optimization provides useful 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. 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: Derivative-free Optimization, Computer Simulation, Emulator, Expected Improvement. EWMA

1 Introduction

Black-box derivative-free optimization has a wide variety of applications, especially in the realm of computer simulations (Kolda et al., 2003; Gramacy et al., 2015). 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. So far there have been no reliable solutions for assessing convergence of surrogate model optimization. In this paper, we provide an automated method for determining 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 (Matott et al., 2011), discussed in more detail in Section 4.3, wherein contamination in the groundwater 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 andrequires evaluation of the simulator. Finding the global constrained minimum is a difficult problem where it is easy for optimization routines to temporarily get stuck in a local minimum. (Gaussian process surrogate model optimization should eventually escape local minima if run long enough.) Without knowing the answer in advance, how do we know when to terminate the optimization routine?

The context of this paper is Gaussian process (GP) surrogate model optimization, a statistical modeling approach to derivative-free numerical optimization that constructs a fast approximation to the expensive computer simulation output surface using a statistical surrogate model (Santner et al., 2003). Analysis of the surrogate model allows for efficient exploration the objective solution space. Typically a Gaussian process surrogate model is chosen for its robustness, relative ease of computa-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

tion, and its predictive framework. Arising naturally from the GP predictive distribution (Schonlau et al., 1998), the maximum Expectation of the Improvement distribution (EI) has shown to be a valuable criterion for guiding the exploration of the objective function and shows promise for use as a convergence criterion (Jones et al., 1998; Taddy et al., 2009).

Taddy et al. (2009) 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 a new optimum; that is to say, globally small EI values should be as-sociated with convergence of the algorithm. Thus a simplistic stopping rule might first define somelower EI threshold, then claim convergence upon the first instance of an EI value falling below thisthreshold, as seen in Diwale et al. (2015). This use of EI as a convergence criterion is analogousto other standard convergence identification methods in numerical optimization (e.g., the vanishingstep sizes of a Newton-Raphson algorithm). However, applying this same threshold strategy to theconvergence of surrogate model optimization has not yet been adequately justified. In fact, this useof EI ignores the nature of the EI criterion as a random variable, and oversimplifies the stochasticnature of convergence in this setting. Thus it is no surprise that this treatment of the EI criterion canresult in an inconsistent stopping rule as demonstrated in Figure (1).

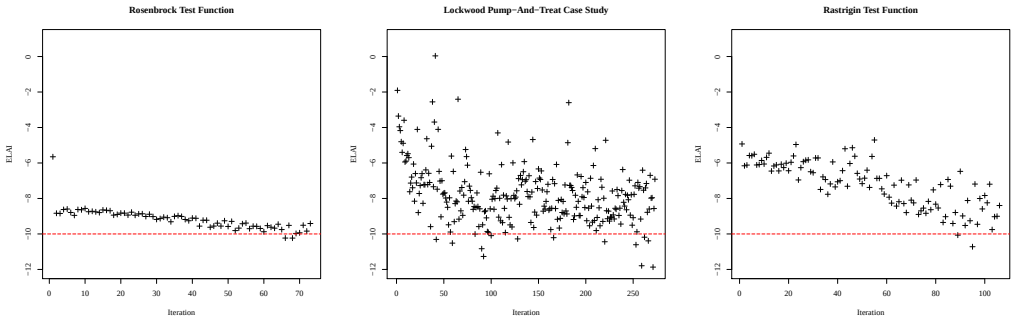


Figure 1: Three Expected Log-normal Approximation to the Improvement series (more details in Section 4) plotted alongside an example convergence threshold value shown as a dashed line at -10.

Because EI is strictly positive but decreasingly small, we find it more productive to work on the log scale, using a log-normal 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 Log-normal Approximation to the Improvement (ELAI) values from three different optimization problems. We will demonstrate later in this paper that convergence is established near the end of each of these series. These three series demonstrate the kind of diversity observed among 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 values 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 invariability is not a necessary condition.

Since the Improvement function is itself random, attempting to set a lower threshold bound on the EI, without consideration of the underlying EI distribution through 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. Throughout this paper we focus on minimization, as maximization can be obtained by minimizing the negative of the function.

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 simulation experiments, where each function evaluation may consume considerable computational resources, while the surrogate model can be evaluated quickly. The standard surrogate model in the literature for analysis of computer experiments is a Gaussian process (GP) (Sacks et al., 1989; Santner et al., 2003). A GP is a stochastic process such that when evaluated at any finite collection of points, the collection follows a multivariate Gaussian distribution. A GP is defined by its mean function and its covariance function, with various standard formulations (Abrahamsen, 1997; Stein, 1999). 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 (Gramacy and Lee, 2012).

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 (2008) 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 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` (Gramacy, 2007; Gramacy and Taddy, 2010). 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 a new minimum is to be observed, at new locations of the domain, based upon the predictive distribution of the surrogate model. EI is built upon the improvement function (Jones et al., 1998):

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

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 as 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. The EI criterion is available

in closed form for a stationary GP. For other models the EI criterion can be quickly estimated using Monte Carlo posterior predictive samples at given candidate locations.

2.2 Optimization Procedure

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 function, 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 thebest 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.

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 $\operatorname{argmax}_{\tilde{\boldsymbol{x}}_i} \mathbb{E}\left[\, \mathrm{I}(\tilde{\boldsymbol{x}}_i) \,\right]$ to $\boldsymbol{X}$.
7) Check convergence.
8) If converged exit. Otherwise go to 2).

3 EWMA Convergence Chart

3.1 Statistical Process Control

In Shewhart's seminal book (Shewhart, 1931) 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 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 common to collect an initial set of data from which point estimates of μ and σ^2 may establish an initial standard for control which is further refined as the process proceeds. 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 Log-normal 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, in which case, 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 zero 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 zero 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 log-normal distribution.

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

$$\theta = \log\left(\frac{\omega^2}{\sqrt{\psi + \omega^2}}\right) \qquad \phi = \log\left(1 + \frac{\psi}{\omega^2}\right). \tag{2}$$

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 log-normal, improvement samples, then use relation (2) to directly compute ψ as the Expectation under the Log-normal Approximation to the Improvement (ELAI). The ELAI value is useful for assessing convergence because of the reduced right skew of the log of the posterior predictive improvement distribution. Additionally, the ELAI 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 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 (Lucas and Saccucci, 1990; Scrucca, 2004) 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 can emphasize recent observations and shift 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 $i \in \{1, 2, 3, ...\}$ 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.

Box et al. (1997) describes typical values of λ in the range $0.1 \leq \lambda \leq 0.3$, with large values of λ producing a more myoptic weighting scheme. Furthermore Box et al. (1997) also describes a method for computing optimal choices of λ by minimizing the sum of squared forecasting deviations (S_λ). For each new observation,

$$\hat{\lambda} = \underset{\lambda \in (0,1]}{\operatorname{argmin}} \left(\sum_i \Big(Y_i - Z_i(\lambda) \Big)^2 \right). \tag{4}$$

Through this analysis of S_λ, as seen in Figure (3), it is 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 those typically chosen 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 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. These larger values of λ allow the EWMA to track the movement of ELAI values from pre-convergence into convergence. 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.

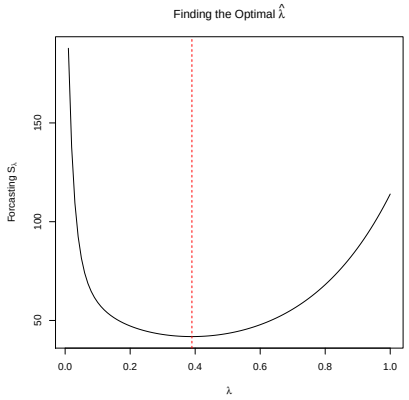


Figure 3: S_λ as calculated for ELAI values derived from optimization of the Rastrigin test function. $\hat{\lambda}$ is shown by the vertical dashed line.

Identifying convergence in this setting, now requires the computation of 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 (1990) show that we can write $\sigma^2_{Z_i}$ in terms of σ^2_Y.

$$\sigma^2_{Z_i} = \sigma^2_Y \left(\frac{\lambda}{2-\lambda} \right) \left[1 - (1-\lambda)^{2i} \right] \tag{5}$$

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. (6).

$$\text{CL}_i = \mu \pm c\sigma_{Z_i} = \mu \pm c \, \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{\lambda}{2-\lambda} \right) \left[1 - (1-\lambda)^{2i} \right]} \tag{6}$$

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, $i \to \infty$, where the control limits are defined by $\mu \pm c\sqrt{\frac{\lambda\sigma^2}{(2-\lambda)n}}$.

Our aim in applying the EWMA framework in this context is to recognize the fundamental notion of control that EWMA enforces in the newly arriving EI values, as optimization proceeds. Convergence often arrises as a subtle shift of the EI distribution into place. In this context a more traditional $\bar{x}$ chart will often overlook convergence as a subtle random fluctuation, when in fact it is often this subtle signal that we aim to pick-up. EWMA is amoung the better techniques for recognizing such subtlely shifting means Aerne et al. (1991); Zou et al. (2009), while maintaining the capability to detect abrupt shifts in mean. As convergence approaches the newly arriving Y_i begin to fit into the $i.i.d.$ EWMA framework and the Z_i increasingly begin to fall within the EWMA control limits. EWMA's recognition of such a controlled region in the newly arriving ELAI values, indicates the notion of distributional consistency that is neccisary for defining convergence for stochastic measures of convergence, such as EI.

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, however to assess convergence the EWMA statistic is 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.First, it serves to dichotomize the series for evaluating subsets of the Y_i for distributional consistency.Second, 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 necessary to establish a standard of control. In this setting w is a kind of sample size, and as such the choice of w will naturally increase as the variability in the ELAI series increases. Just as in other sample size calculations, the choice of an optimal w must consider the cost of poor inference (premature identification of convergence) associated with underestimating w, against the cost of over sampling (continuing to sample after convergence has occurred) 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 evaluation.

For the purpose of exploring the behavior of w in examples presented here, we use the following proceedure for educating the choice of w. We hand tune w for two informative known example functions (i.e. Rosebrock and Rastrigin). From exploration of w in known examples, it is clear that w increases directly with ELAI variance. Furthermore, if one considers the form of sample size calculations based on classical power analysis, sample size increases directly proportional with the sample variance. Thus we linearly extrapolate the choice of w for the Lockwood case study based on a defalt starting value of 30 (based on sampling conventions) with a slope term structured to make use of the proportionality of w with the observed ELAI variance, $\hat{w} = \frac{\Delta w}{\Delta \mathrm{V(ELAI)}}\hat{v} + 30$. Further exploration of the exact form of an estimator of w is left to be discovered, although the connection of w with sample size calculations is a promising line of research in itself.

3.5 Identifying Convergence

In identifying convergence, we not only desire that the ELAI series 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.

First, we require that all exponentially weighted Z_i values inside the control window to fall within the control limits. This ensures that the most recent ELAI values demonstrate distributional consistency within the bounds of the control window. Second, since we wish to indicate a move from the initial pre-converged state of the system, we require at least one point 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.

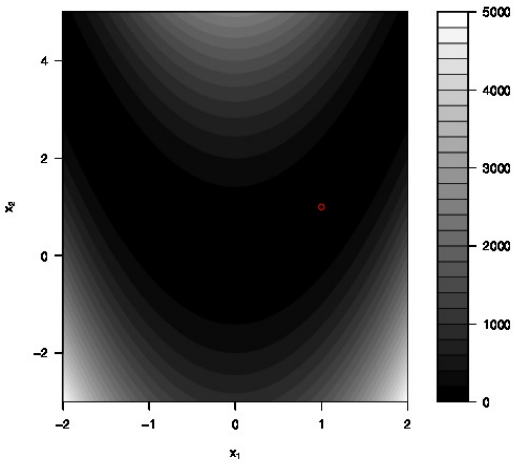
Considering the optimization procedure outlined in Figure (2), the check for convergence indicated in step 7) amounts to computing new EWMA Z_i values, and control limits, from the inclusion of the most recent observation of the improvement distribution, and checking if the subsequent set of Z_i satisfy 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 tune the EWMA Convergence Charts for each of these synthetic examples, then extrapolate the choice of w to provide a real world example from hydrology.

4.1 Rosenbrock

The Rosenbrock function (Rosenbrock, 1960) 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. 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, in the context of GP surrogate model optimization. This example illustrates a well-behaved convergence process.



$$f(x_1, x_2) = 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2$$
Minimum : $f(1, 1) = 0$

We estimate λ via the minimum S_λ estimator, $\hat{\lambda} \approx 0.5$. Due to the relative simplicity of this problem we find that $w = 30$ results in a well behaved convergence pattern with a final ELAI variance of 0.35 . 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 the convergence chart, with the control window to the right of the vertical line, and the control limits indicated by the dashed lines.Iteration 74 is the first time that all EWMA points in the control window are observed within the control limits, and thus we declare convergence. This declaration of convergence comes after the global minimum has been found, but not too many iterations later, just enough to establish convergence. Note that the EWMA points generally trend downward until the global minimum is found at iteration 63.

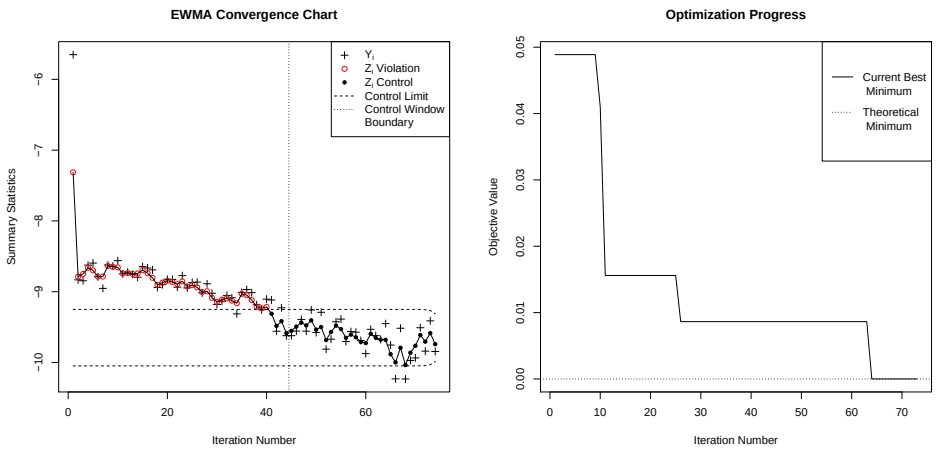
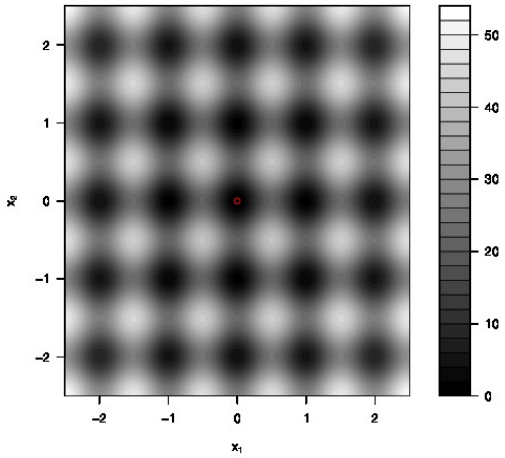


Figure 4: Rosenbrock function: Convergence chart on the left, optimization progress on the right.

4.2 Rastrigin

The Rastrigin function is a commonly used test function for evaluating the performance of global optimization schemes such as genetic algorithms (Whitley et al., 1996). The global behavior of Rastrigin is dominated by the spherical function, $\sum_i x_i^2$, however Rastrigin has been oscillated by the cosine function and vertically shifted so that it achieves a global minimum value of 0 at the lowest point of its lowest trough at (0, 0).



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

Rastrigin is generalizable to arbitrarily many dimensions, but to develop intuition, 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 variability 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.

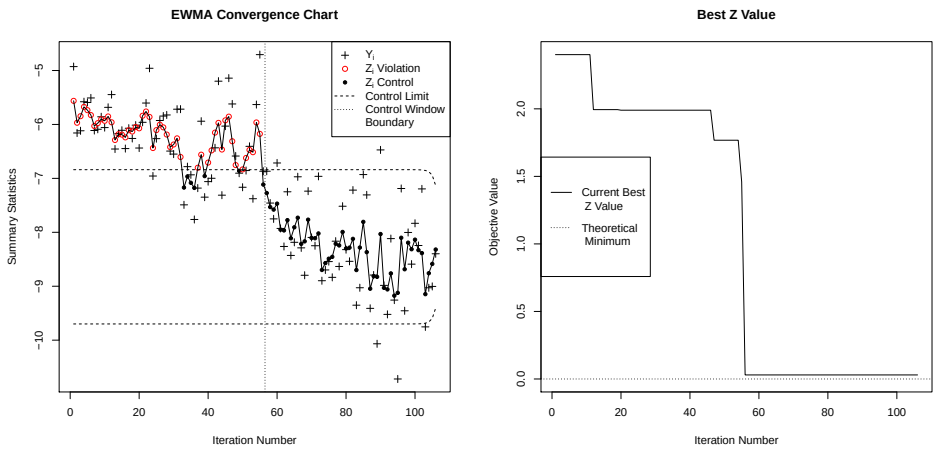


Figure 5: Rastrigin function: Convergence chart on the left, optimization progress on the right.

$\hat{\lambda}$ in this example is calculated to be about 0.4 . The decreased value of $\hat{\lambda}$, relative to Rosenbrock, increases the smoothing capabilities of the EWMA procedure, as a response to the increased noise in the ELAI series. The added noise of the ELAI series here 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 $w = 50$ was found to work well, with a final ELAI variance of 1.71 .

Figure (5) shows the convergence chart (left) and the optimization progress of the algorithm (right) after 105 iterations of optimization. Although the variability of the ELAI criterion 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 seen 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 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). Specifically we consider the Lockwood pump and treat problem, originally presented by Matott et al. (2011).

The Lockwood pump and treat case study considers an industrial site, along the Yellowstone River in Montana, with groundwater contaminated by chlorinated solvents. If left untreated, this contaminated groundwater may contaminate the Yellowstone river, as dictated by the hydrology of the system. In order 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 treatment facility. Due to the cost of running these pumps, it is desirable to determine how to best allocate the pumping effort among 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 quantity 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 contamination boundary. Each x_i is bounded on the 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.

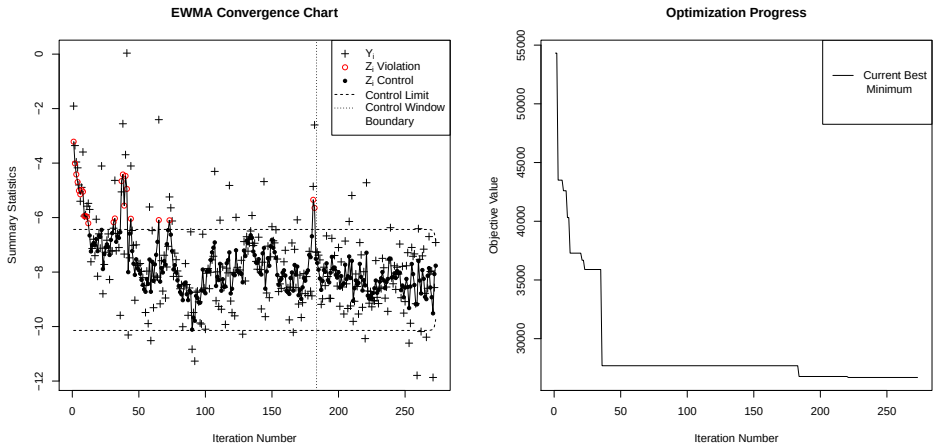


Figure 6: Lockwood Case-study: Convergence chart on the left, optimization progress on the right.

By using the fitted values of w and the observed ELAI variance in each of the two previous examples we extrapolate an apporpriate value of w for this case study based on an observed ELAI variance of 2.86 , resulting in an estimated w of $93 \approx \left(\frac{60-30}{1.71-0.35}\right) 2.86 + 30$, as discussed in section 3.4.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 the ELAI criterion due to the large search domain, as well as the complicated contamination boundary among the six wells.Furthermore these features of the objective function complicate fit of the surrogate model and thus more function evaluations are required to produce an accurate model of f.

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 $w = 93$. Convergence in this case does not occur with a dramatic shift in the mean level of the ELAI criterion, but rather convergence occurs as the series stabilizes after large ELAI values move beyond the control limit. Interestingly the last major spike in the ELAI series is observed alongside the discovery 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 eventually 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 diminishing returns for this problem, by the analysis of Gramacy et al. (2015) on the same problem, as seen in 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 an objective standard for identifying convergence in the presence of the inherent stochasticity of the improvement criterion in this setting. The examples provided here demonstrate how the EWMA convergence chart may accurately and efficiently identify convergence in the context of GP surrogate model optimization. We note that our approach could be applied with any optimization algorithm that allows computation of an expected improvement at each iteration.

As for any optimization algorithm, a converged solution may only be considered as good as the algorithm's exploration 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 embedded, 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, or that it has converged to a local mode; for well-tuned surrogate modeling strategies, this point should correspond with the realization of an optimal solution. In either case, the EWMA convergence chart identifies the moment at which it is beneficial to stop iterating the routine and reflect upon the results.

The EWMA convergence chart presented here is intended as a starting point for establishing an appropriate analysis of convergence for sequential surrogate modeling optimization algorithms. Details of the particular implementation of these methods may improve through further analysis of model usage and parameter estimation. The strategy presented here for transforming the improvement distribution via the Log-normal approximation to the improvement distribution has shown to be an empirically effective and computationally simple solution to better meet the assumptions of the EWMA control charting methodology. However, some applications may find it worthwhile to explore other transformations which could result in higher overall transformed signal to noise ratios, across a more broad set of improvement distributions. For example, rather than adopting the ELAI transformed estimate from the improvement distribution, it may be computationally feasible to apply the two-parameter Box-Cox transformation (Box and Cox, 1964) to the improvement samples,

$$y_i^{(\boldsymbol{\lambda})} = \begin{cases} \frac{(y_i+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(y_i+\lambda_2) & \lambda_2 = 0 \end{cases} \tag{9}$$

thus alleviating any difficulties due to numerical truncation of the improvement samples at 0, while finding a flexible transformation to reduce skew. It should be noted that this approach adds additional computational expense, while our ELAI transformation requires minimal computation. Additionally the EWMA convergence chart could benefit from a more precise method for choosing the control window size parameter, w, although developing such a method would be a major research project itself; our empirical solution has worked well in practice.Although improvements to the details of these methods may exist, the fundamental consideration of the stochastic nature of convergence in this setting would remain, and SPC offers a nice framework for its inclusion.

References

Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions. Technical Report 917, Norsk Regnesentral/Norwegian Computing Center.

Aerne, L. A., Champ, C. W., and Rigdon, S. E. (1991). Evaluation of control charts under linear trend. *Communications in Statistics-Theory and Methods*, 20(10):3341–3349.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211–252.

Box, G. E. P., Luceño, A., and Paniagua-Quiñones, M. D. C. (1997). *Statistical Control by Monitoring and Adjustment*. Wiley, New York.

Diwale, S. S., Lymperopoulos, I., and Jones, C. (2015). Optimization of an airborne wind energy system using constrained gaussian processes with transient measurements. In *First Indian Control Conference*, number EPFL-CONF-199719.

Gramacy, R. B. (2007). tgp: an r package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software*, 19(9):1–46.

Gramacy, R. B., Gray, G. A., Le Digabel, S., Lee, H. K. H., Ranjan, P., Wells, G., and Wild, S. M. (2015). Modeling an augmented lagrangian for blackbox constrained optimization. *Technometrics*. to appear, preprint arXiv:1403.4890.

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

Gramacy, R. B. and Lee, H. K. H. (2012). Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722.

Gramacy, R. B. and Taddy, M. (2010). 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(6):1–48.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

Kolda, T. G., Lewis, R. M., and Torczon, V. (2003). Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45:385–482.

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

Matott, S. L., Leung, K., and Sim, J. (2011). Application of matlab and python optimizers to two case studies involving groundwater flow and contaminant transport modeling. *Computers & Geosciences*, 37(11):1894–1899.

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

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

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, 4:409–423.

Santner, T. J., Williams, B. J., and Notz, W. (2003). *The design and analysis of computer experiments*. Springer, New York.

Schonlau, M., Jones, D., and Welch, W. (1998). Global versus local search in constrained optimization of computer models. In *New Developments and applications in experimental design*, number 34 in IMS Lecture Notes - Monograph Series, pages 11–25. JSTOR.

Scrucca, L. (2004). qcc: an r package for quality control charting and statistical process control. *R News*, 4/1:11–17.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. D. Van Nostrand Company, New York.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer, New York.

Taddy, M. A., Lee, H. K. H., Gray, G. A., and Griffin, J. D. (2009). Bayesian guided pattern search for robust local optimization. *Technometrics*, 51(4):389–401.

Whitley, D., Rana, S., Dzubera, J., and Mathias, K. E. (1996). Evaluating evolutionary algorithms. *Artificial Intelligence*, 85(1-2):245–276.

Zou, C., Liu, Y., and Wang, Z. (2009). Comparisons of control schemes for monitoring the means of processes subject to drifts. *Metrika*, 70(2):141–163.