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: Computer Model, 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 cite. 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 [7], discussed in more detail in Section 3.1, 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 and 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 routine that constructs a fast approximation to the expensive computer simulation using a statistical surrogate model [5]. 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 [8]. Arising naturally from the GP predictive distribution [9, 3], the maximum Expected Improvement (EI) criterion has shown to be a valuable criterion for guiding the exploration of the objective function [12]Jones?; furthermore the EI shows promise for use as a convergence criterion cite.

Literature cite recommends considering the EI as a convergence criterion for surrogate model optimization; as of yet, little work has been done to describe what convergence of these algorithms actually looks like in the context of the EI criterion. However, the basic idea behind the use of the EI criterion, as a convergence criterion, is that convergence should occur when the surrogate model produces low expectations for discovering new optima; that is to say, small EI values should be associated with convergence of the algorithm. Thus, a common stopping rule, involving the EI criterion, first defines some lower EI threshold, then claims convergence upon first sight of an EI value falling below this threshold [2], cite. This use of EI as a convergence criteria falls well within the comfort zone of the numerical optimization literature as this is certainly a reasonable approach for monitoring the convergence criteria of other routines (ex. the vanishing step sizes of Newton-Raphson, Pattern Search (PS) methods, etc.). However, applying this same threshold strategy to the convergence of surrogate model optimization, in the context of the EI criterion, 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).

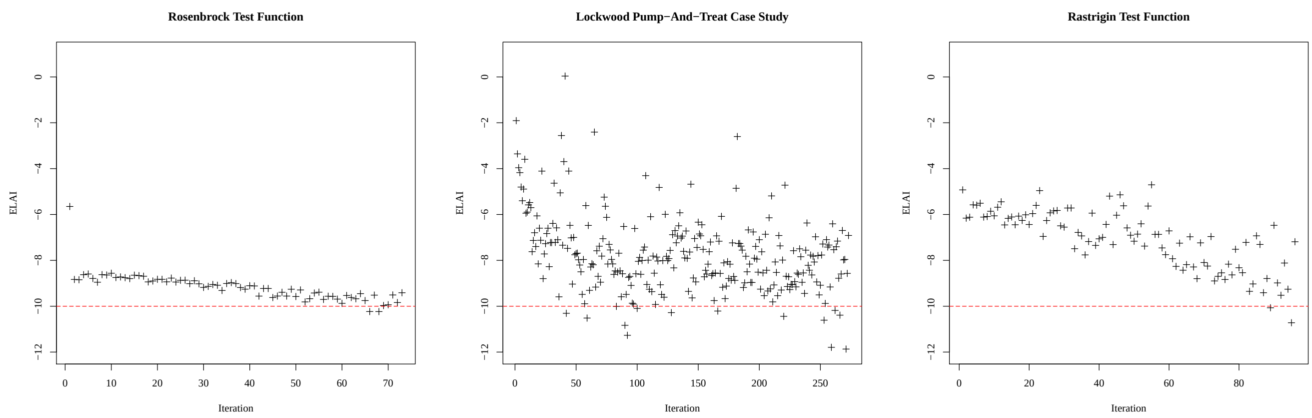

Figure 1: Three ELAI series plotted alongside an example convergence threshold value shown as a dashed line at -10.

Because Expected Improvement 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) 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 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.

2 Gaussian Process Surrogate Model Optimization

*Everybody uses GP, bunch of cites

For more details on the theoretical foundations of Gaussian Processes in computer experiments see [8]. If we have any hope of finding optima of a function, f, we impose the condition that f provides a reasonably smooth mapping for relating points in the domain, $\boldsymbol{x}$, to response values, $z(\boldsymbol{x})$. The $z(\boldsymbol{x})$ are assumed to be particular realizations of an infinitely dimensional generalization of the multivariate normal distribution, $f \sim \text{GP}(m,\ K)$, and thus jointly any set of such realizations, $z(\boldsymbol{x})$, jointly follow a multivariate normal distribution. As part of the GP model specification, we specify the mean function as a linear combination of simple basis functions, $m = \boldsymbol{\beta}^\intercal \mathbf{f}(\boldsymbol{x})$, and we describe the covariance structure, among the dimensions of the $z(\boldsymbol{x})$, through specification of a covariance function, $K(\boldsymbol{x}, \boldsymbol{x}')$. By specifying a homogeneous covariance function, we thus model the relationship between $||\boldsymbol{x} - \boldsymbol{x}'||$ with the correlation structure that we expect to see when jointly considering two such realizations of the GP. The following exponential power family provides an example of a reasonable choice of $K(\boldsymbol{x}, \boldsymbol{x}')$, under the assumption of a reasonably well behaved f,

$$K(\boldsymbol{x},\ \boldsymbol{x}') = \sigma^2 \exp\left\{-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^p}{d}\right\}. \tag{1}$$

Specifying conjugate priors for $\boldsymbol{\beta}$ and σ^2 yields a straightforward Gibbs sampling posterior inferential setting with the exception of the covariance structure parameters requiring Metropolis-Hastings MCMC sampling.

*Expand Treed Partitioning

For additional modeling details, including loosening the assumption of global stationarity, and details about implementing GP models in the context of numerical optimization see the R package `tgp` as well as the associated vignettes [4, 5].

2.1 Expected Improvement

The EI criterion is fundamentally based on the improvement criterion cite which evaluates how possible it may be to encounter new minima at a given location based on the predictive surrogate model. The improvement function takes the following form,

$$\mathrm{I}(\boldsymbol{x}) \;=\; \max\Big\{\big(f_{min} - f(\boldsymbol{x})\big),\; 0\Big\}. \tag{2}$$

By considering the expectation of $\mathrm{I}(\boldsymbol{x})$, candidate locations are not only rewarded for having a low predictive mean, but the $\mathbb{E}[\,\mathrm{I}(\boldsymbol{x})\,]$ also rewards poorly explored locations due to the high uncertainty of $\mathrm{I}(\boldsymbol{x})$ in these places. Notice that by definition the $\mathrm{I}(\boldsymbol{x})$ function is always non-negative, however the GP posterior predictive $\mathbb{E}[\,\mathrm{I}(\boldsymbol{x})\,]$ is a strictly positive criterion. Considering the MCMC inferential setting of our GP surrogate model, the EI criterion can be quickly computed by using posterior predictive $\mathrm{I}(\boldsymbol{x})$ samples at given candidate locations to empirically approximate the $\mathbb{E}[\,\mathrm{I}(\boldsymbol{x})\,]$ calculation.

2.2 Optimization Procedure

The idea for optimization, in this context, is to only evaluate the objective function at locations that have a good chance of providing a new minimum. 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

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

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.

3 EWMA Convergence Chart

3.1 Statistical Process Control

In Shewhart's seminal 1931 book [10] 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. For example, the $\bar{x}$-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 typical sampling distribution for the mean, $\bar{x}_j \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Shewhart expresses his idea of control, in this case, as the expected behavior of random observations from this sampling distribution. By considering confidence intervals on this sampling distribution we can easily draw explicit boundaries (i.e. control limits) to identify which samples are in control, and which are not. Observations violating our expectations (i.e. observations that fall outside of the confidence interval/beyond 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 special care should always be taken to satisfy its requirements.

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 carfully consider properties of the improvement distributions which generate the EI values. Ultimately, the distribution of the EI is asymtotically normal, but the characteristics of the improvement distribution, as convergence approaches, complicate these asymtoptics. The improvement criterion is strictly positive but decreasingly small, thus the improvement distribution is often strongly right skewed; leading to poor asymptotics for the EI distribution. Additionally, this right skew becomes exaggerated as convergence approaches, due to the decreaseing trend in the EI criterion. Together these characteristics of the improvement distribution give the EI criterion inconsistent behaviour for tracking convergence via SPC.

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 could simply 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 in double precision, is computationally indistinguishable from 0. Thus simply taking the log of the improvement samples can be computationally undefined, particularly as convergence approaches.

Rather than simply taking the log of the improvement samples, to determine the more robust statement that $\mathbb{E}[\ \log \mathrm{I}\ ] \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$, it is computationally useful to consider the following approximate model-based perspective. Recall that if a random variable $X \sim \textit{Log-}N(\omega, \nu)$, then another random variable $Y = \log(X)$ is distributed $Y \sim N(\omega, \nu)$. Furthermore, if m and v are, respectively, the mean and variance of a lognormal sample, then the mean, ω, and variance, ν, of the associated normal distribution are given by the following relation cite.

$$\omega = \log\left(\frac{m^2}{\sqrt{v + m^2}}\right) \qquad \nu = \log\left(1 + \frac{v}{m^2}\right). \tag{3}$$

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 (3) to directly compute ω as the Expectation under the Lognormal Approximation to the Improvement (ELAI). The ELAI convergence criterion is a useful convergence criterion in this case because the reduced right skew of the log of the improvement distribution enjoys improved asymptotics, and the ELAI convergence criterion serves as a computationally robust approximation of the $\mathbb{E}[\log \mathrm{I}]$ under reasonable lognormality of the improvements. Furthermore, both the $\mathbb{E}[\log \mathrm{I}]$ and ELAI convergence criterion are distributed normally in repeated sampling. This construction allows for more consistent and accurate use of the fundamental theory on which our SPC prespecitive requires.

3.3 Exponentially Weighted Moving Average

The EWMA control chart is based upon Shewhart's original notion of control, however the EWMA control chart veiws the repeated sampling process in the context of moving average smoothing of series data. Since the ELAI convergence criterion stochastically slides into convergence, a series perspective is appropriate here; in particular the EWMA perspective has shown to be well suited for tracking the progression of means that are subject to subtle drifting processes [11, ?], just as displayed by the ELAI criterion upon convergence. 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 smoothing with some memory of the past.

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,

$$Z_i = \lambda Y_i + (1 - \lambda) Z_{i-1}. \tag{4}$$

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. [1]. Additionally Box et al. explains that EWMA charts are relatively robust to reasonable choices of λ, as evident by relatively small errors in the sum of squared forcasting deviations (S_λ) associated with reasonable choices of λ. Box et al.'s analysis of S_λ also suggests an optimal choice for λ by choosing $\hat{\lambda}$ to minimize S_λ [1, p. 87] for each new observation. Figure (3) shows the robustness of EWMA to suboptimal choices of λ since for the large region $\lambda \in [0.2, 0.6]$, S_λ stays within 10% of its the minimum possible value at $\hat{\lambda}$ for the optimization of the Rastrigin test function.

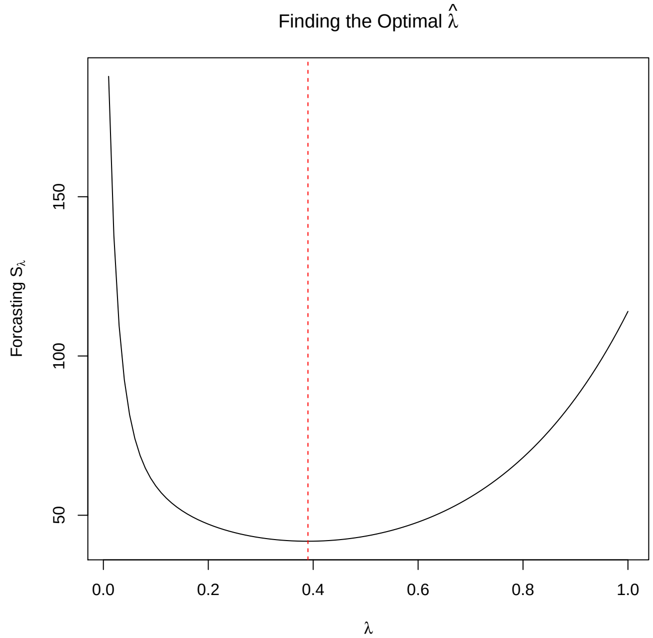

Figure 3: Prior to the observation of Y_i, all available values are used to define a forcasted value for Z_i using Eq. (4). S_λ is subsisquently calculated upon observation of Y_i and Z_i. Here $\hat{\lambda}$ is shown by the vertical the dashed line.

In general large values of λ assign more weight to recently observed values, and thus past observations effect the moving average less. Conversely, small values of λ assign less weight to recent observations, and thus small values of λ provide more smoothing across the effects of past observations. Hence larger values of λ tend to be better suited for dealing with consistently shifting series, and small values of λ are more well suited for series with fairly consistent expected values.

For identifying convergence it is of primary importantance to define the control limits on the EWMA statistic in this setting. As in the simplified $\bar{x}$-chart, defineing the control limits in 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 [6] 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 the $Y_i \overset{i.i.d.}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ the sampling distribution for Z_i is $Z_i \sim N\left(\mu, \sigma^2_{Z_i}\right)$. Furthermore by choose a confidence level through choice of a constant c, the control limits based on this sampling distribution are seen in Eq. (6).

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

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, $i \to \infty$, the contol 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 (4). For these controlled Y_i an *i.i.d.* assumption is very reasonable. The realization of such a controlled region of the series defines the notion of consitency which in-turn is part of what allows for the identification of convergence here.

3.4 The Control Window

The final structural feature of the EWMA convergence chart that is necissary for identifying the distributional consistency we use to identify convergence is the so called *control window*. The control window is a window containing a fixed number, w, of the most recently observed Y_i. The idea being, 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 algorithm is allowed to fill the control window, by collecting and initial set of w observations of the Y_i. As new observations arrive, the oldest 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 subset of the Y_I for distibutional consistency. Secondly it offers a structural way for basing the standard for consistency only on the most recent and relevant information in the series. This is important due to the subtle way in which ELAI values sliding to convergence.

The size, w, of the control window is an important parameter for correctly identifying convergence. Because w may vary from problem to problem it is ultimately left as a tuning parameter of the system. Choosing the correct value of w presents an interesting decision problem since underestimating the size of the control window may lead to premature convergence, but if w is too large, we compute unnecessary objective function evaluations. As a general trend, harder optimization problems require larger values of w since the EI criterion follows a less structured decreasing pattern as new modes are discovered at irregular patterns.

3.5 Identifying Convergence

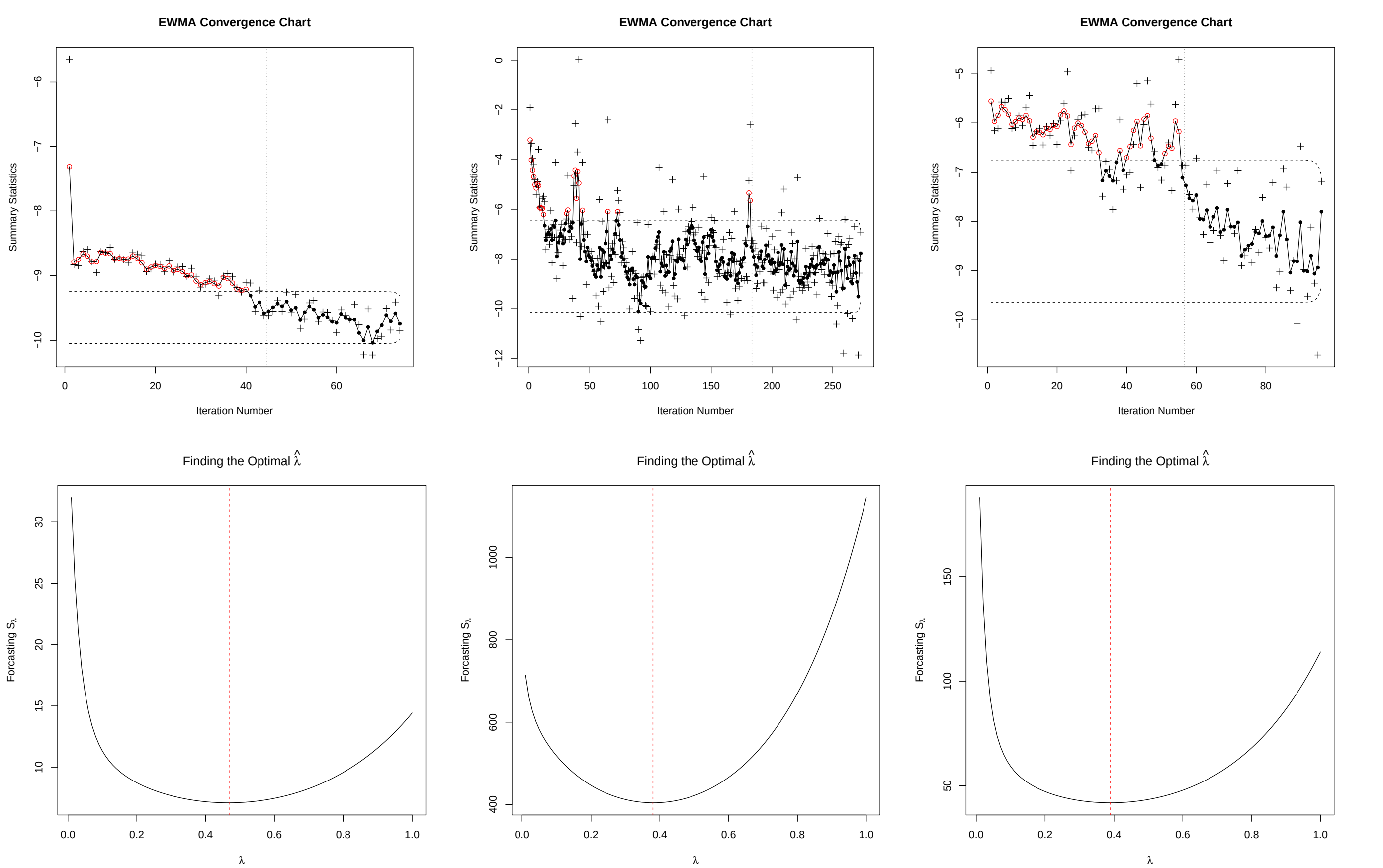

Figure 4: *left* : Rosebrock | *center* : Lockwood | *right* : Rastrigin

4 Examples

5 Conclusion

References

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

[2] 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*, number EPFL-CONF-199717, 2014.

[3] David Ginsbourger, Rodolphe Le Riche, Laurent Carraro, et al. A multi-points criterion for deterministic parallel global optimization based on gaussian processes. 2008.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] Yan Su, Lianjie Shu, and Kwok-Leung Tsui. Adaptive ewma procedures for monitoring processes subject to linear drifts. *Computational statistics & data analysis*, 55(10):2819–2829, 2011.

[12] 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.