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 [5], 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 [4]. 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 [6]. Arising naturally from the GP predictive distribution [7, 2], the maximum Expected Improvement (EI) criterion has shown to be a valuable criterion for guiding the exploration of the objective function [9]Jones?; furthermore the EI shows promise for use as a convergence criterion cite.

Literature cite recommends considering the EI as a convergence criterion for GP 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 GP 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 [1], 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 GP 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).

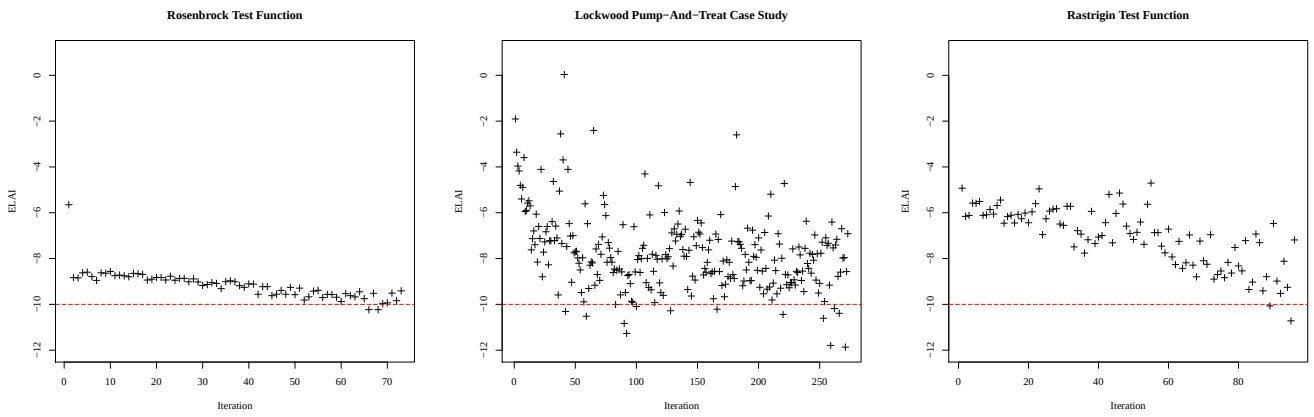

Figure 1: Three ELAI series plotted alongside an example convergence threshold value shown as a dashed line at -10.

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, as described in more detail in Section XX. Figure (1) represents three series of the Expected Lognormal Approximation to the Improvement (ELAI) 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

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, $f \sim \mathrm{GP}(m,\ K)$, 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, $m = \boldsymbol{\beta}^\top \mathbf{f}(\boldsymbol{x})$, 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 cite posterior inferential setting with the exception of the covariance structure parameters requiring Metropolis-Hastings MCMC sampling cite.

For more details on the theoretical foundations of Gaussian Processes in computer experiments see [6]. 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 [3, 4].

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 GP 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

Figure 2: Optimization Procedure

1) Collect an initial set, $\boldsymbol{X}$.
2) Compute $f(\boldsymbol{X})$.
3) Fit GP model 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).

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 GP model is then fitted with $f(\boldsymbol{X})$ as observations of the true function. Using this model, a set of candidate points, $\tilde{\boldsymbol{X}}$, are randomly 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 GP model is refit based on the updated $f(\boldsymbol{X})$. The optimization procedure carries on in this

way until convergence.

3 Statistical Process Control

In Shewhart's seminal 1931 book [8] 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.

4 EWMA Convergence Chart

4.1 Expected Lognormal Approximation to the Improvement (ELAI)

In the spirit of obtaining a robust solution through the use of the CLT, as applied to the distribution of the EI, it is important to carefully consider properties of the improvement distribution. Firstly, the improvement values are strictly positive but decreasingly small, thus the improvement distribution is often strongly right skewed. This right skew becomes exaggerated as convergence approaches (i.e.

when the improvement values become small). Thus if large sample assumptions (from the improvement distribution) are challenged, then the asymptotic EI distribution may become less accurate at the time of convergence (i.e. when we desire accuracy most). Secondly, since the improvement distribution is strictly positive, the EI should also be bounded at 0. Thus an unfettered normal distribution will always struggle to capture the EI distribution.

These issues naturally suggest modeling the log improvement, rather than the improvement distribution on its own. 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 indistinguishable from 0. Thus simply taking the log of the improvement samples can be computationally undefined, particularly as convergence approaches.

Thus 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.

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

Using this relation we do not need to transform any of the improvement samples. We compute the empirical mean of the unaltered, approximately lognormal, improvement samples, then use relation (3) to directly compute the Expectation of the Lognormal Approximation to the Improvement (ELAI). The reduced right skew of the log improvement improves the asymptotics of the $\mathbb{E}[\ \log \mathrm{I}\ ]$ distribution; the ELAI serves as a computationally robust approximation of the $\mathbb{E}[\ \log \mathrm{I}\ ]$, and thus both the $\mathbb{E}[\ \log \mathrm{I}\ ]$ and ELAI are distributed approximately normal in repeated sampling.

4.1.1 Exponentially Weighted Moving Average

4.1.2 The Control Window

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.