Determining Convergence in Gaussian Process Surrogate Model Optimization

Nicholas Grunloh and Herbie Lee (Applied Mathematics and Statistics, Baskin School of Engineering, University of California, Santa Cruz)

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. Here we develop a novel approach that adapts ideas originally introduced in the field of statistical process control to define a robust convergence criterion based upon the improvement function.

Gaussian Process Surrogates

The typical approach to modeling a black-box function is with a Gaussian Process (GP) surrogate model (Santner et al., 2003):

$$Y(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{h}(\mathbf{x}) + Z(\mathbf{x}) + \epsilon$$

where Y is a scalar output, $\boldsymbol{\beta}$ is a vector of regression coefficients, h is typically the identity function plus inclusion of an intercept, and Z a zero-mean GP with spatial covariance kernel $C(\cdot,\cdot)$ and possible error term (nugget) $\epsilon \sim N(0, \sigma_\epsilon^2)$. Following the literature, we use a Gaussian correlation structure $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\sum_{k=1}^{m}(x_{ik} - x_{jk})^2/d_k\right\}$. For additional flexibility, we used Treed Gaussian Processes (Gramacy and Lee, 2012), which allows for non-stationarity and possible discontinuities.

Optimization with Expected Improvement

Global minimization can be attempted using a surrogate model by sequentially choosing a new function evaluation at the input that maximizes the expected improvement (EI), where the improvement function is

$$I(\mathbf{x}) = \max\left\{y_{min} - Y(\mathbf{x})\ ,\ 0\right\}$$

where $y_{min} = \min\{y_1, \ldots, y_n\}$ and the expectation is taken with respect to the posterior predictive distribution of the surrogate (Jones et al., 1998). We can do optimization by sequentially evaluating the point with the largest EI at each iteration, then updating the model and repeating. But when do you stop?

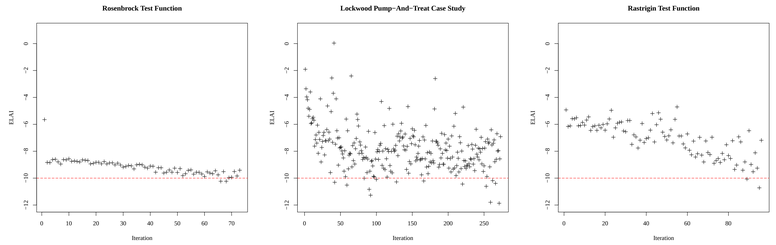
When to Declare Convergence?

Some have suggested using EI for convergence (Diwale et al., 2015), declaring convergence when the largest EI value falls below a predetermined threshold, much like many stopping criteria in numerical algorithms. The idea is that our surrogate model predicts that there is very low probability of finding any other points that are more optimal.

However, this approach ignores the fact that EI is a random variable and oversimplies the stochastic nature of convergence in this setting. Thus it results in inconsistent behavior.

Examples of EI Monitoring

Three examples of Expected Log-normal Approximation to the Improvement (ELAI), where the log scale is used because EI is strictly positive but decreasingly small:



The first example (Rosenbrock) is an ideal well-behaved case, where convergence does occur when the threshold is met. The second example (Lockwood) shows a failure of the threshold, as it is met too early, and while the variability is still too large. But small variability is also not sufficient or necessary, as the third example shows, where variability is increasing but meeting the threshold does indicate convergence.

Working on the log sacle is important because the EI distribution is right skewed, and become more skewed as convergence approaches. Numerically, some EI values get evaluated to be 0 in double precision, even though they are theoretically positive. Thus we use a model-based approximation of a log-normal, transforming the empirical mean and variance via the log-normal.

Statistical Process Control

Consider a control chart, used to monitor a process in equilibrium to watch for it to deviate (go out of control). We take inspiration from this approach, but use it in reverse. As optimization proceeds, EI decreases as we find lower values and learn more about the function. Once we find the minimum, we achieve convergence, and at this point, EI should settle into an approximately stable distribution. Thus we want to see when our out of control process becomes an in-control process.

We basically perform SPC backwards in time. Starting with the most recent iteration, we look backwards in time and see if we have an in-control process that becomes out of control. That is the signature of convergence.

We use an Exponentially Weighted Moving Average (EWMA) control chart (Lucas and Saccucci, 1990), which provides smoothing to account for highly stochastic nature of EI. The exponential weighting also helps when the series is decreasing, as is typical pre-convergence. Denote the ELAI values by Y_i. EWMA tracks $Z_i = \lambda Y_i + (1-\lambda)Z_{i-1}$.

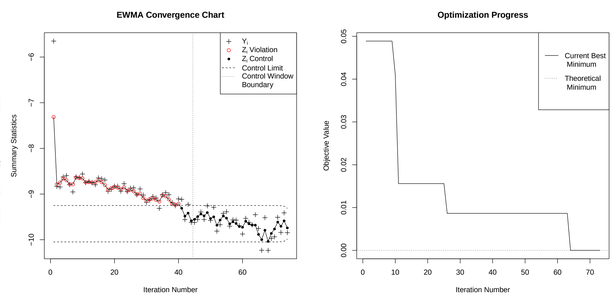
Control Window

We define a control window of the w most recent observations from which we compte the control limits of the EWMA chart. Starting from the most recent point in time, we look backwards, using the first w observations to get the control limits, then looking to see if any points further back in time go outside those limits. If so, we declare convergence. If we are pre-convergence, then the w points won't be in equilibrium and the control limits will be quite large, typically including all runs from the beginning.

The choice of w depends on the difficulty of the problem, it needing to be sufficiently large to establish control, but not too large as that means extra iterations beyond those necessary. Our default is $w = 15p$ where p is the dimension of the problem.

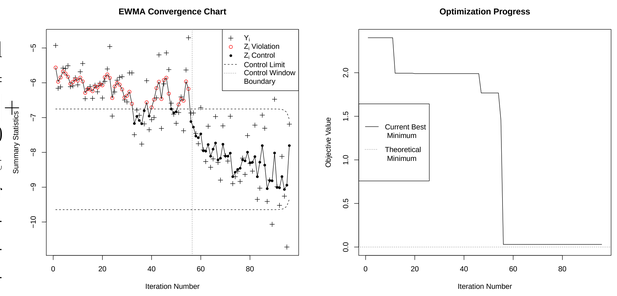
Rosenbrock Test Function

$f(x_1, x_2) = 100\left(x_2 - x_1^2\right)^2 + (1-x_1)^2$. We use $w = 30$ and estimate $\hat{\lambda} \approx 0.5$. The EWMA convergence chart goes into control as we find the true global minimum.



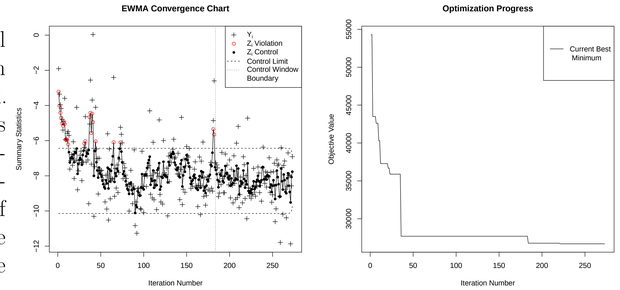
Rastrigin Test Function

This highly multi-modal function is $f(x_1, x_2) = \sum_{i=1}^{2}\left[x_i^2 - 10\cos(2\pi x_i)\right] + 2(10)$. We use $w = 40$ and estimate $\hat{\lambda} \approx 0.4$. It takes a little time after finding the global minimum for the EWMA convergence chart to establish convergence.



Lockwood Case Study

For this six-dimensional groundwater remediation problem from Mayer et al. (2002), the objective is to find the lowest operation cost configuration (pumping rates) of six extraction wells while avoiding spread of the contamination. We use $w = 90$ and estimate $\hat{\lambda} \approx 0.4$.



References

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*, no. EPFL-CONF-199719.

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

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.

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

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.

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