UNIVERSITY OF CALIFORNIA
SANTA CRUZ

IDENTIFYING CONVERGENCE IN GAUSSIAN PROCESS SURROGATE MODEL OPTIMIZATION, VIA STATISTICAL PROCESS CONTROL

A document submitted in partial satisfaction of the
requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS AND APPLIED MATHMATICS

by

Nicholas R. Grunloh

July 2014

Approved by:

Professor Herbert Lee

Associate Professor John Musacchio

Table of Contents

Abstract

Identifying convergence in numerical optimization is an ever-present and difficultly 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 is often still subjective. Ideas originally introduced in the field of Statistical Process Control (SPC) are thus used to define convergence in an objective way. The Exponentially Weighted Moving Average (EWMA) chart provides an ideal starting point for adaptation to track convergence via the EWMA convergence chart introduced here.

1 Introduction

Convergence is a bit of a loaded word, that is used in many different quantitative contexts. In many cases the notion of "convergence" can be a frustratingly soft idea, often with an oddly subjective definition and fuzzy interpretation. In each different context of the word, "convergence" may have a slightly different meaning, and with it, "convergence" may carry different implications about the problem at hand. Within the setting of optimization, convergence usually just indicates that we can stop iterating our routines, but even within optimization, convergence can look drastically different from routine to routine. In this paper I aim to give "convergence" a more concrete definition in the context of Gaussian process surrogate model optimization.

By the nature of the stochastic exploration procedures inherent to Gaussian process surrogate model optimization, convergence is also stochastic in nature. Unlike other methods, such as gradient descent or pattern search, convergence in Gaussian process surrogate model optimization is not as straightforward as monitoring a vanishing step-size. In Gaussian process surrogate model optimization the step-size, between locations of function evaluations, is a largely varied random variable and thus it is not a particularly telling feature of the progress of the objective search. Among many practical surrogate modeling applications the claim of convergence may simply depend on the available computation time and the adequacy of the current best solution. However it is obviously preferred to have metrics which expressly indicate convergence. For this purpose, various secondary criteria [**?**], derived from the surrogate model itself, have been monitored. In particular it is common to monitor the

maximum expected-improvement (EI) until it simply falls below a specified threshold [?]. This over simplifies the dynamics of convergence in this setting, as quite small EI values should be expected with some regularity based on the particular topology of the problem and the stochasticity of the criterion itself. For the purpose of formalizing a robust process for tracking this stochastic criterion, I turn to the charting methods of the statistical process control literature. Here I borrow ideas from Shewhart's [?] classic notion of control, to chart the EI, and thus form a more accurate and objectively tangible definition of convergence.

My argument is structured in the following way: Section 1.1 gives a brief overview of context for the optimization used here, Section 1.2 covers the basics of Gaussian process surrogate models, and Section 1.3 explains how these models have been used as efficient derivative-free optimization routines. In Section 1.4 I introduce the use of EI as a convergence criterion, and in Section 1.5 I provide a more in-depth discussion of the statistical process control (SPC) logic I use to consistently identify convergence, via the EI criterion. In Section 2 I tie all of these topics together to outline a charting procedure to identify convergence in a robust way. Finally, in Section 3 I provide some examples of identifying convergence via the methods outlined in Section 2.

1.1 Overview

Derivative-free optimization is an enormously practical and commensurately difficult task. Derivative information has the capacity to efficiently, and rather intuitively, lead the user to an optimal solution. However, derivative information tends to be focused locally, around the starting location of the objective function search, and thus can easily get stuck at local optima. Furthermore, derivative information is often not available in many practical problems. When derivatives are not available we are left to find more creative ways of figuring out which way is up. Examples of this creativity can be seen in the diversity of different techniques employed by some of the more popular methods for derivative-free optimization.

Examples of effective, and greatly varying, derivative-free optimization strategies include: evolutionary algorithms (EA), simulated annealing (SA), pattern search (PS), trust region

methods, as well as surrogate model approaches. Although widely varied, fundamentally these methods share three basic components [?]. Firstly, there is some procedure for exploring the objective function space. Secondly, these methods use derivative-free information from exploratory function evaluations to update the exploratory procedure, and thus, explore more effectively. Thirdly, a well rounded optimization routine is tasked with accurately identifying when it has found an optimal point. By tweaking any one of these components, an optimization routine's behavior, with respect to scope, convergence, and accuracy, may differ dramatically. Thus, an ideal optimization routine would explore a large space quickly, and it would tell you that it has converged to an accurate solution with minimal information required from the objective function. Of course, an optimization routine which embodies *all* of these characteristics does not exist, as of yet, but these are good characteristics to consider when choosing the best optimization routine for a particular problem.

In particular, I consider the convergence properties of Gaussian process surrogate model approaches. The basic idea of surrogate model approaches is to create a statistical approximation (i.e. a model) of the objective function, and use this model to effectively search the objective landscape. Surrogate model-based approaches are primarily of interest due to their ability to deal with functions that are computationally expensive to evaluate because they take special care to minimize the number of objective function evaluations. The typical choice of model in the construction of a surrogate model-based optimization routine is a Gaussian process model [?]. The idea is to work out the next best point to explore by using the surrogate model instead of the objective function. Working with the surrogate model saves objective function computation time, and further more, allows for a careful statistical search of the objective function. I demonstrate how further analysis of Gaussian process surrogate models can be used to identify convergence in this setting.

1.2 Gaussian Process Models

Gaussian Process (GP) models often arise naturally in the context of spatial statistics. In fact, GP models often go by the name kriging due to Danie G. Krige who pioneered the

use of GPs in the context of Geo-spatial data related to mining [**?**]. Thus, it can often be instructive to think about the objective function, f, in the context of a spatial application rather than as an abstract mathematical function, since often times f can resemble, or even actually represent, the classical spatial problems. The choice of a GP as a surrogate model for f hinges on the fundamental idea that f provides a reasonably smooth mapping for relating points in the domain, $\mathbf{x}$, to response values, $z(\mathbf{x})$. That is to say, if we have any hope of finding optima of f, we impose the idea that points close together in the domain should have values in the response that are predictably, and similarly, close. Regardless of the true interpretation of f, by modeling f in this way we may expect f to behave, at least in part, as a spatial quantity; for instance f may just as well represent the elevation of a mountain in space.

A formal statistical perspective expresses a GP as an infinitely dimensional generalization of the multivariate normal distribution, such that every realization of a GP is a normal random variable and jointly all such realizations form a multivariate normal distribution. Typically the mean response is modeled using a linear combination of simple basis functions, $\boldsymbol{\beta}^{\intercal}\mathbf{f(x)}$, with a zero mean random process error term, $\epsilon(\mathbf{x})$, such as,

$$z(\boldsymbol{x}) = \boldsymbol{\beta}^{\intercal}\mathbf{f}(\boldsymbol{x}) + \epsilon(\boldsymbol{x}) + \eta(\boldsymbol{x}). \tag{1}$$

Here $\eta(\boldsymbol{x})$ is Gaussian noise, and $\epsilon(\boldsymbol{x})$ is fundamentally governed by a correlation function, $K(\boldsymbol{x}, \boldsymbol{x}')$, such that the covariance is $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 K(\boldsymbol{x}, \boldsymbol{x}')$. By specifying a homogeneous correlation function, we thus model the relationship of $||\boldsymbol{x} - \boldsymbol{x}'||$ with the correlation structure that we expect to see when jointly considering two such realizations of the GP. The following exponential power family provides a common example of such a choice of $K(\boldsymbol{x}, \boldsymbol{x}')$,

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

Considering Eq. (**??**) for every combination of $\mathbf{x}$ and $\mathbf{x}'$ among a particular data set provides a correlation matrix $\mathbf{K}$; thus multiplying by σ^2 creates the likelihood covariance matrix $\mathbf{C}$.

For further discussion of choices of $K(\boldsymbol{x}, \boldsymbol{x}')$ see [**?**]. Equations (**??**) and (**??**) imply the following simple Bayesian model, which forms the basis for many other complex GP models

$$\begin{aligned} \mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{K} \;&\sim\; N_n\Big(\mathbf{F}\boldsymbol{\beta},\; \sigma^2\mathbf{K}\Big) \qquad \sigma^2 \mid a, b \;\sim\; IG(a,\; b) \\ \boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \mathbf{V} \;&\sim\; N_m\big(\boldsymbol{\beta}_0,\; \mathbf{V}\big) \qquad\quad \mathbf{V} \mid \boldsymbol{\Psi}, \nu \;\sim\; IW\big(\boldsymbol{\Psi},\; \nu\big). \end{aligned} \tag{3}$$

Here a, b, $\boldsymbol{\beta}_0$, $\boldsymbol{\Psi}$, and ν are fixed hyper-parameters of the model, and N, IG, and IW represent the Multivariate Normal, Inverse-Gamma, and Inverse-Wishart distributions respectively. Model (**??**) specifies a mostly conjugate, Gibbs sampling, inference setting with the exception of the covariance structure parameters, which require Metropolis-Hastings sampling [**?**].

Bayesian models of this type, not only provide effective inference on f, but they provide a framework for prediction that allows for further efficient exploration of f. In the Bayesian perspective, the parameters of Model (**??**) are random variables. Thus doing inference on these parameters, via MCMC, results not only in point estimates, but entire distributions that completely, and flexibly, describe the present uncertainty. Furthermore Bayesian methods, such as Model (**??**), provide a complete predictive framework for estimating function behavior in unobserved candidate locations, as well as full distributional uncertainty characterization of these unseen locations. By considering the uncertainty and expected behavior of the posterior predictive GP surface across candidate locations, it is thus possible to make informed decisions about where to search for new optima [**?**].

In many cases the assumption of a 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 looks more like the Grand Canyon, as opposed to the Great Plains. Gramacy and Lee [**?**] introduce the idea of allowing this flexibility via a treed partitioning of the domain. This allows separately stationary GP surfaces to fit separately stationary portions of f. For further explanation of partitioned Gaussian process models as well as notes on implementing such models in R, see the R package `tgp` [**?**].

1.3 Optimization

In explaining the construction of optimization routines based on statistical models like model (**??**) as described by [**?**], I view the typical surrogate optimization procedure in terms of the three basic components of optimization that I outline in Section 1.1.1. Firstly, I identify the type of information that has been used in Gaussian process models to effectively explore f. Secondly, I outline the exploration procedure, and how a Gaussian process updates its exploration of f using this information. For the sake of making a concrete argument, I focus on optimization in the context of minimization, but all of these ideas can easily be applied to finding maxima by simply minimizing $-f$.

1.3.1 Expected Improvement

In finding minima via Gaussian process models, the expected-improvement (EI) criterion has been used [**?**], [**?**] to identify candidate points that have the strongest *possibility of encountering new minima*. The EI criterion is fundamentally based on the improvement criterion for each candidate location of the following form,

$$\mathrm{I}(\boldsymbol{x}) \;=\; \max\Big\{\big(f_{min} - f(\boldsymbol{x})\big),\; 0\Big\} \tag{4}$$

In expectation, the $\mathbb{E}\,[\,\mathrm{I}(\boldsymbol{x})\,]$ criterion rewards candidate points not only for a low predictive mean, but also rewards the high uncertainty associated with poorly explored regions. Considering Bayesian models like Model (**??**), we can most efficiently use information in our model by considering $\mathbb{E}\,[\,\mathrm{I}(\boldsymbol{x})\,]$ with respect to the posterior predictive distribution.

By the nature of the Bayesian construction of models like Model (**??**), criteria such as the improvement criterion, $\mathrm{I}(\boldsymbol{x})$, are random variables, and as such, we can learn their distributions via Markov Chain Monte Carlo (MCMC) methods. The distribution of $\mathrm{I}(\boldsymbol{x})$, a posteriori, can be obtained by considering samples from the posterior predictive distribution at each candidate location and computing the necessary statistics to form $\max\Big\{\big(f_{min} - z(\tilde{\boldsymbol{x}})\big),\; 0\Big\}$ as an approximation of Eq. (**??**). Furthermore, the mean of these posterior

predictive $\mathrm{I}(\boldsymbol{x})$ samples provide an empirical solution for finding an EI for each candidate location [**?**]. Thus, truncating these $\mathrm{I}(\boldsymbol{x})$ samples at 0 and finding the candidate location with the maximum $\mathbb{E}[\ \mathrm{I}(\boldsymbol{x})\ ]$, identifies the EI criterion described.

1.3.2 Exploration 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. An optimization scheme based on models like Model (**??**) starts 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.

Figure 1: 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).

1.4 A Convergence Criterion

Iterating the above mentioned optimization procedure and tracking the value of EI at each iteration gives the user a sense for what the algorithm thinks is left to learn about finding a minimum of f. For instance, recall that EI is giving us information about the *possibility of encountering a new optimum.* For determining convergence, you can imagine exploring f with respect to EI, in the following way. Initially, as we enter the objective landscape of f, we do not really know what to expect. However we set-out to explore this space, and as we

explore, we develop expectations about the topography of f. Each exploratory sample of f provides some potential for a new optimum, although as we learn more about this function, the expectation that we will see something new begins to diminish. If we continue to explore this landscape, we will come to understand the environment so well that we will virtually never expect to see something new. In fact, continued exploration of the space may become boring. Thus the key to efficiently identifying convergence is to quickly realize that the expectation of finding a new, and substantial, optimum is sufficiently low. That is to say, convergence occurs, in this context, when the expectation of finding new minima is low enough, so that continued search of the objective function is expected to be unproductive.

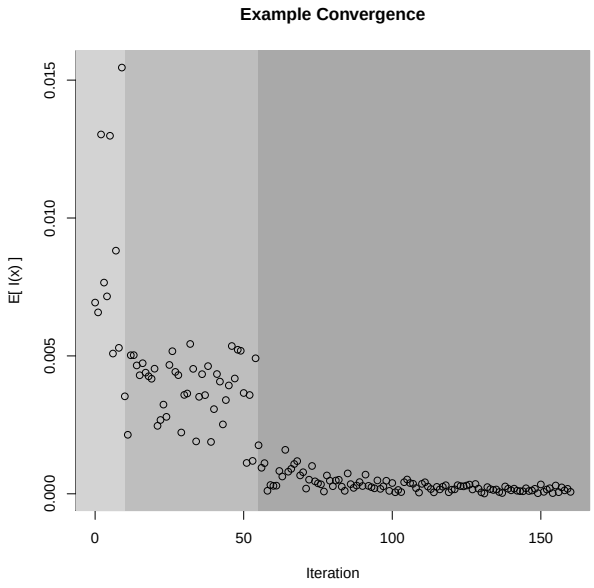


Figure 2: A fabricated EI progression, made to clearly demonstrate the typical three stage convergence pattern.

Following from this intuitive story about the behavior of EI; quantitatively EI follows a stochastically non-stationary decreasing function as iterations of the optimization routine pass. Initially EI tends to start in at an optimistically high value, depending on the initial size of $\boldsymbol{X}$. As several iterations of the optimization procedure continue to sweep through, EI tends to enter a fairly stable region of intermediate values where the algorithm is figuring out the major features of f. Eventually the value of EI will converge in probability to 0, but by construction it can not decrease below 0.

1.5 Statistical Process Control

In identifying convergence, I find the notion of "control", from the SPC literature, to be in the same spirit as the notion of "convergence" in optimization. In Shewhart's seminal 1931 book [?] 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 is not only an instructive framework for thinking about convergence, but it offers this framework with a comforting sense of finitude. The phrase "within limits" gives us a hope of drawing some line in the sand; turning the previously subjective burden of identifying convergence, into a simple objective task that even a computer can accomplish.

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, say m, repeated samples, of size n, 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 our 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.

2 Identifying Convergence

Figure (**??**) can bee seen to resemble an $\bar{x}$-chart, and with some modifications it is not hard see how tracking EI values could naturally fall into the framework of the SPC logic. Recall that each point in this figure is the mean of $\mathrm{I}(\boldsymbol{x})$ MCMC samples at the most promising candidate location, in the current iteration. Thus as iterations of the optimization procedure pass, we form a repeated sampling situation for the EI values to consider via SPC. The idea behind identifying convergence in this setting, is to establish a state of pre-convergence; we then claim that we have achieved convergence when we observe EI values that indicate a move from this pre-convergence state, into a state of control about some converged EI distribution.

Considering the EI values in this way requires tactful consideration of how to evaluate the arrival of EI observations. In order to achieve the above described perspective of convergence, the goal is to establish control among the most recently observed EI values as they move from the initial pre-convergence values into control. Thus, it is necessary to consider the progression of EI values in the reverse order for the sake of SPC. That is to say, I consider the most recently observed EI value as the first value to be tracked in the SPC repeated sampling. This construction allows moving average methods, described in section 2.2, to establish a standard of control that is based on the most up-to-date EI information. In many ways the EI criterion falls naturally into the typical $\bar{x}$-chart setting, but as we have already seen, careful consideration of the properties of the EI convergence behavior illustrate some practical and theoretical concerns for identifying convergence using the SPC logic.

Firstly, recall that EI is a stochastically decreasing function of the iteration number. The nonstationary decreasing nature of the EI values can easily lead to premature identification of convergence. The initially very high EI values combined with the overall decreasing trending of the series often sets up a trap which quickly cause initial values to exceed the control limits of an $\bar{x}$-chart. I introduce the Exponentially Weighted Moving Average (EWMA) chart as a way of diffusing this effect.

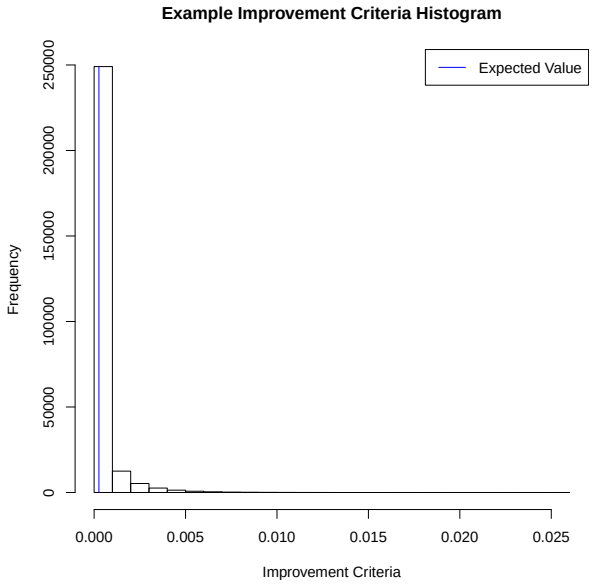


Figure 3: An example $I(\boldsymbol{x})$ sample histogram, demonstrating the extreme right skew. Additionally, $\mathbb{E}[\, I(\boldsymbol{x})\,]$ is shown in blue.

The second major concern brought about by the application of SPC on EI is the distribution of $I(\boldsymbol{x})$. Upon investigation of this distribution it is quickly clear that $I(\boldsymbol{x})$ typically follows a strongly right skewed distribution, due to the non-negative construction of the $I(\boldsymbol{x})$ criterion. This is not a fatal property of the distribution of $I(\boldsymbol{x})$ in terms of the central limit theorem (CLT), although it does yield nearly worst case asymptotics in terms of the convergence of the sampling distribution to normality. A simple solution for getting more asymptotic results could include increasing the sample size of the draws of $I(\boldsymbol{x})$. However it is important to consider that by the nature of MCMC implementation of models like Model **??**, increasing the number of samples of $I(\boldsymbol{x})$ entails taking more samples of every single parameter in the model. Increasing the sample size of $I(\boldsymbol{x})$ is a perfectly valid solution to this problem, if computation memory is of no concern, but it is not a particularly robust and satisfying solution. Additionally, since the $I(\boldsymbol{x})$ criterion is naturally bounded at 0, an unfettered normal distribution will always struggle to model the EI criterion since the normal distribution will never respect its boundary conditions. Thus, modeling these data so as to find appropriate transformations to improve their asymptotics, and better model the boundary conditions of the problem, is a worthwhile consideration.

In the following sections I address each of these concerns in turn. As I address each issue, I modify my method for tracking the information contained in the EI criterion so as to robustly identify convergence as inspired by SPC.

2.1 Improved Normality

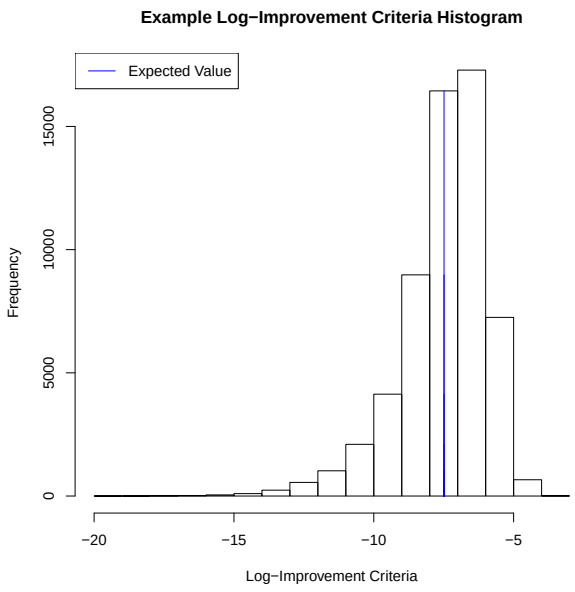


Figure 4: An example $\log \mathrm{I}(\boldsymbol{x})$ sample histogram, demonstrating improved skew. Additionally, $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ is shown in blue.

For the sake of improving the asymptotic convergence of the EI sampling distribution to normality it is of much desire to transform the distribution of $\mathrm{I}(\boldsymbol{x})$ so as to be less skewed. A simple and often effective transformation for getting skewed distributions to more resemble normality is a simple log transformation of the data. In this case simply log transforming the MCMC samples of the $\mathrm{I}(\boldsymbol{x})$ criterion is not always possible. By the definition of the $\mathrm{I}(\boldsymbol{x})$ criterion, an $\mathrm{I}(\boldsymbol{x})$ sample can at its lowest take a value of 0; see Eq. (**??**). MCMC samples that explore $\mathrm{I}(\boldsymbol{x})$ values as low as 0 will render the log function useless, as log is undefined for values ≤ 0. Thus rather than actually transforming the data in this way, I propose modeling these samples with a log-normal distribution.

Recall that if a random variable $X \sim \textit{Log-}N(\mu, \sigma^2)$, then another random variable $Y = \log(X)$ is distributed $Y \sim N(\mu, \sigma^2)$. Furthermore, if m and v are, respectively, the mean and variance of a log-normal sample, then the mean, μ, and variance, σ^2, of the associated normal distribution are given by the following relation,

$$\mu = \ln\left(\frac{m^2}{\sqrt{v+m^2}}\right) \qquad \sigma^2 = \ln\left(1+\frac{v}{m^2}\right). \tag{5}$$

Using this relation we do not need to transform any of the $\mathrm{I}(\boldsymbol{x})$ samples; we can instead jump straight to the mean of the samples that would have resulted from transforming these $\mathrm{I}(\boldsymbol{x})$ samples. That is to say, by using relation **??** we immediately get a value for $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ without actually taking the log of $\mathrm{I}(\boldsymbol{x})$ samples. Considering the distribution of $\log \mathrm{I}(\boldsymbol{x})$,

seen in figure **??**, the asymptotics ensuring the normality of the distribution of $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ as compared with that of $\mathbb{E}[\ \mathrm{I}(\boldsymbol{x})\ ]$ are far favorable.

2.2 Exponentially Weighted Moving Average

In general moving average methods use the idea of a rolling average to smooth data that arrive as a series. EWMA methods achieve this smoothing by assigning exponentially decreasing weights to successive points in a rolling average among all of the points of a series. This disproportionately focuses the attention of the moving average on recent information (i.e. the most relevant information in this case), while still smoothing the overall series progression with at least some memory of past values. These properties of the EWMA have shown to provide a robust solution for tracking the progression of means that are subject to subtle drifting processes [**?**], such as one might expect to see in convergence.

Since early $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ values generally behave differently than later values, I use the EWMA procedure to track the progression of $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ values in reverse order. This has the effect of forgiving the wild fluctuations of early inexperienced explorations and highlighting the most recent experiences with f. Furthermore, the exponentially decreasing weights are well suited for monitoring convergence in this case because they have the ability to smooth out the initial EI fluctuations, while still having the resolution to pick out the subtle shifts inherent to the convergence process. Further dynamics of EWMA are well explained by Box et al. [**?**]; additionally EWMA charts, among other common control charts, can easily be implemented in R by using the R package `qcc` [**?**].

If Y_i is the current value of $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$, 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,

$$Z_i = \lambda Y_i + (1 - \lambda) Z_{i-1}. \tag{6}$$

Above, λ is a 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 moving back through the series.

Typically values of λ range from $0.1 \leq \lambda \leq 0.3$, with a default value of $\lambda = 0.2$, as described by Box et al. [**?**]. Additionally Box et al. explains how to estimate a $\hat{\lambda}$ so as to minimize sum of squared deviation (S_λ) of the resulting EWMA series. 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 large shifts, and small values of λ are more sensitive to small shifts. Often it is the case with EWMA that the best choice of λ is based on "expert opinions" related to the underlying data generating process. For identifying convergence, "expert opinions" come in the form of an in-depth understanding of how EI values behave relative to the specific search phenomena related to the expected behavior of f.

For identifying convergence it is also important to define the control limits on the statistic seen in Eq. (**??**). Again, this amounts to considering an interval on the sampling distribution of interest. In this case we are interested in the sampling distribution of the Z_i, if the Y_i are *i.i.d.* then Lucas and Saccucci [**?**] show that we can write $\sigma^2_{Z_i}$ in terms of σ^2_Y.

$$\sigma^2_{Z_i} = \sigma^2_Y \left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda^{2i})\right] \tag{7}$$

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 if we choose a confidence level through a choice of the constant c, the control limits based on this sampling distribution are seen in Eq. (**??**) follow on the next page.

$$\begin{aligned} \mathrm{CL}_i &= \mu \pm c\sigma_{Z_i} \\ &= \mu \pm c\,\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda^{2i})\right]} \end{aligned} \tag{8}$$

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 when traversing backwards through the series resulting directly from the geometrically decreasing weights.

At this point it is necessary to take a step back from the modeling details and motivate the use of such models on the real $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ behavior. At first glance it is not clear that the Y_i are in fact *i.i.d.* Indeed the early iterations of the process certainly do not display *i.i.d.* looking Y_i values. However once the process proceeds far enough, the Y_i enter a state of control; thus for the values that are in control, the *i.i.d.* assumption is very reasonable. The fact that the pre-convergence Y_i do not necessarily demonstrate control may in large part have to do with their lack of *i.i.d.*ness, and thus identifying control in this way validates the appropriateness of this model by identifying control. That is to say, in this application the concept of *i.i.d* samples serves as part of the check for control; realizing control validates the assumption of *i.i.d.* samples and vice versa.

Another detail worth further recognition, is the asymptotics on which all of the this logic is built. Even after log transformation of the $\mathbb{E}[\ \mathrm{I}(\boldsymbol{x})\ ]$ criterion, $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ is still only asymptotically normal. Even though the distribution of $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ is likely to be quite normal, via the CLT, it is still worth mentioning that any non-normality left in the model is left for EWMA's own robustness to cope with. Furthermore, constructing a robust model for identifying convergence is of primary concern, especially considering the many varied conditions this model must perform under. In considerations of this type, Box et al. [**?**] argues that EWMA is a robust and parsimonious solution in many cases, especially with respect to disputes in the distribution and stationarity of the Y_i. Thus the EWMA logic should provide the sturdy and robust framework necessary for dealing with data of this type.

2.3 The Control Window

Typically in SPC some initial number of samples are gathered and deeply investigated to establish an initial standard for control that is not wildly off-base. Investigation in this setting amounts to careful analysis of the conditions related to points that fall beyond the control limits; often with the intention of attributing some deterministic reason for the lack of control. Observations investigated in this way are often discarded from the calculation of initial control limits because they are found not to represent the desired state of control, and thus through their exclusion the standard for control is elevated.

In the setting of optimization it is not desirable to study each pass of our routine so carefully, and often such analysis would not be meaningful. However, for the sake of keeping in the spirit of SPC, as well as other practicalities, I offer a compromise that frames the setting for which convergence is to be identified. I propose the idea of a sliding window of fixed size, w, that I call the *control window*. 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 values as before. Initially, I allow the algorithm to fill the control window, by collecting w observations of f. As new observations of f arrive, the oldest value is removed from the control window, thus allowing for the inclusion of a new value. This offers an automated way of basing the standard for convergence only on the most updated and relevant information.

The size, w, of the control window is important 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. Thus it may be important to consider these two opposing forces when choosing an appropriate value for w. I recommend conservatively large values for w because I regard premature convergence to be a greater problem than extraneous function evaluations. As a default value $w = 30$ has provided me a reasonable starting point for further analysis. I have found that choosing

w based on the value of λ seems to be an efficient way of tuning w. As a general trend, the larger the value of λ, the more fluctuation present in the EWMA statistic. Thus for good results, large λ, naturally imply large values of w for an accurate representation of the increased fluctuations of the EWMA statistic in the repeated sampling average. Conversely for small λ, smaller values of w are acceptable.

2.4 EWMA Convergence Chart

Recall that this work is meant to apply at step 7) of figure **??**. Thus, at each iteration we can obtain a new $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ value to track via the EWMA procedure outlined in Section 2.2. If we then apply the control window to this EWMA tracking procedure, interesting rules for identifying convergence naturally arise.

The typical EWMA identifies out-of-control observations by simply identifying which Z_i values exceed the established control limits. This may be adequate for a check of simple control, but is not a stringent enough standard for determining convergence. In identifying convergence we not only desire that the converged state has reached a state of control, but we also desire a state of control that indicates significantly lower expectations for finding new minima in the future. This suggest a dual use of the the control window in suggesting convergence. Firstly, I require that all Z_i values inside the control window fall within the control limits. Secondly, since we wish to indicate a move from the initial pre-converged state of the system, I require at least one point from beyond the control window to fall outside the defined control limits. A set of Z_i satisfying both of these conditions indicate the desired state of convergence. Satis-

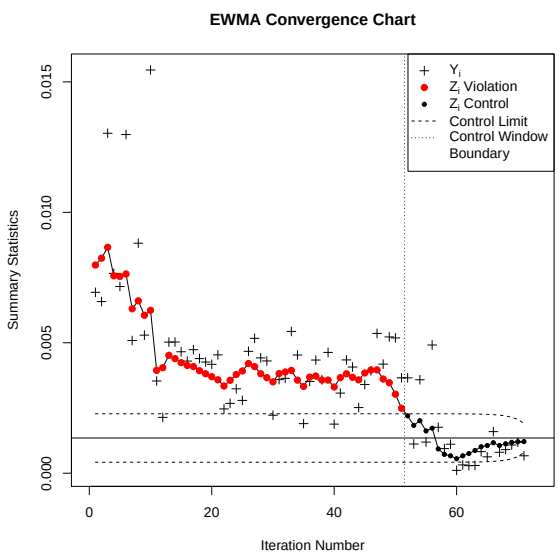


Figure 5: An example EWMA convergence chart based on the fabricated EI progression in figure **??**. This chart is converged.

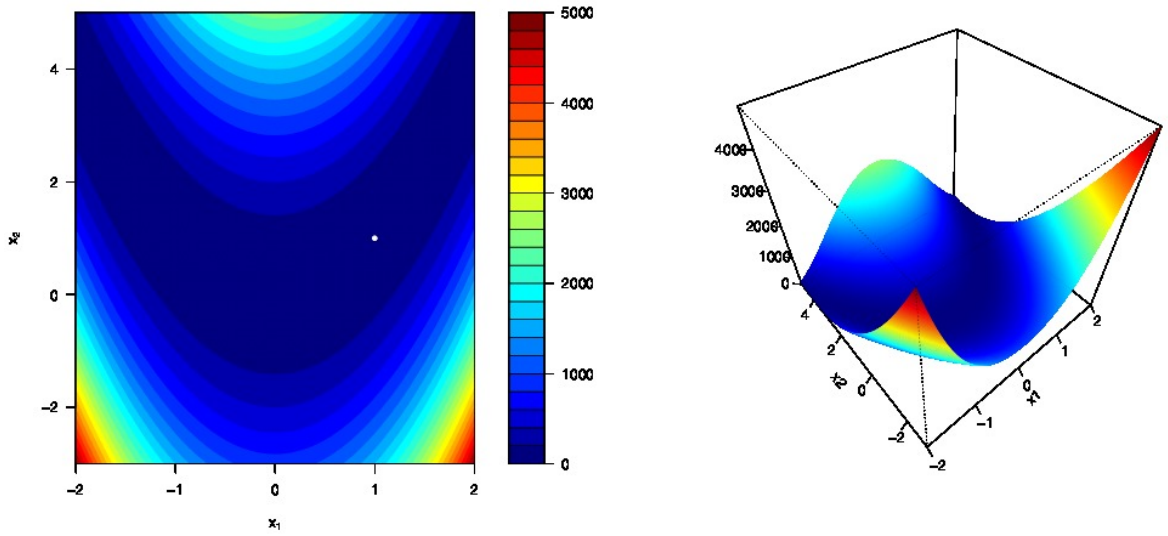
fying only one of these conditions indicates insufficient exploration of the objective landscape and thus further iterations of procedure **??** are required.

Together these rules capture the notion that the process of convergence is a slide from a pre-convergence state, into a converged state of control. In the converged state, Z_i values in the control window (i.e. recent values) indicate control in the classical sense. Since the control window partitions the Z_i into new and old values, the control window provides a mechanism for identifying when control has shifted into a state of control that has moved sufficiently from the initial pre-convergence state. These added considerations for the behavior of current Z_i values, relative to old observations, differentiate the classical notion of control charts from what I call a convergence chart.

3 Examples

In this section I provide a series of test function examples, as well as a case study demonstrating the use of the EWMA convergence chart for monitoring convergence. The test function examples are meant to highlight strategies for tuning λ and w relative to the observed behavior of f, as well as demonstrate various states of convergence with respect to various objective function behaviors. Additionally, I consider the Lockwood pump-and-treat case study as a practical example of how the EWMA convergence chart may be used in less tidy optimization problems.

3.1 Rosenbrock



$$\begin{aligned} f(x_1, x_2) &= 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2 \\ \text{Minimum} &: f(1, 1) = 0 \end{aligned} \tag{9}$$

The Rosenbrock function [**?**] is a classic optimization benchmark, and thus it is only fitting to begin with an analysis of its long and narrow, flat parabolic valley. Exploring the banana-shaped valley is an exercise in self control, as the flat valley floor seems to endlessly meander around the curve of the banana with relatively little descent as compared with the steep 4^{th} order polynomial valley walls. This stark difference in scale tempts optimization routines to prematurely claim convergence all along the lengthy valley floor. As indicated above, the true global minimum value is attained jarringly off centered at (1, 1), where the objective value eventually falls to 0.

For the sake of this analysis I focus on the pictured intervals, in this example I am focusing on Rosenbrock with $-2 \leq x_1 \leq 2$; $-3 \leq x_2 \leq 5$. This interval subjects the surrogate model to the flat valley floor, but since we start optimization with a naively chosen $\boldsymbol{X}$, initially the model is forced to make some consideration of the steep valley walls.

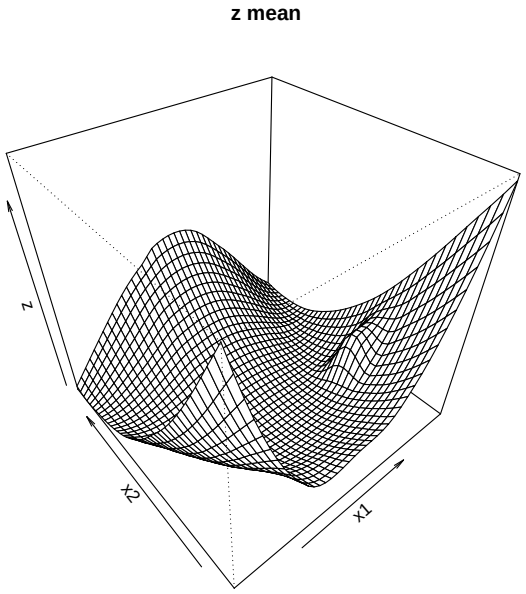


Figure 6: The GP mean predictive surface of the Rosenbrock function after 70 iterations.

As optimization, proceeds only the maximum EI candidate location is added to $\boldsymbol{X}$, and thus for the sake of optimization no time is wasted searching for a minimum in the steep walls. Furthermore, since the maximum EI candidate location typically falls somewhere in the valley floor, more and more points from inside the valley are gathered. This forms a very good model for what the true shape of the Rosenbrock function looks like inside this valley, but gives only the necessary impression of the surrounding walls needed to identify that they do not contain minima, as seen in Figure (**??**).

Since the Rosenbrock function is not highly multimodal and does not contain any hard to discover features, it provides an example of the kind of function that will provide the default EI behavior. The basic strategy for tuning λ and w starts by first choosing a value for λ, to remove a degree of freedom from the system. Since Rosenbrock provides an example of default convergence behavior, the default λ value is appropriate, $\lambda = 0.2$. After choosing λ, the choice of w can be made based on the chosen value of λ. Again since λ was chosen in accordance with the default value, the default value of $w = 30$ is also appropriate here.

Recall that for constructing an EWMA convergence chart it is necessary to first fill the control window so as to establish some initial opinion of the searching process. Figure (**??**, *left*) shows this initial, pre-convergence state, for the Rosenbrock function as described thus far.

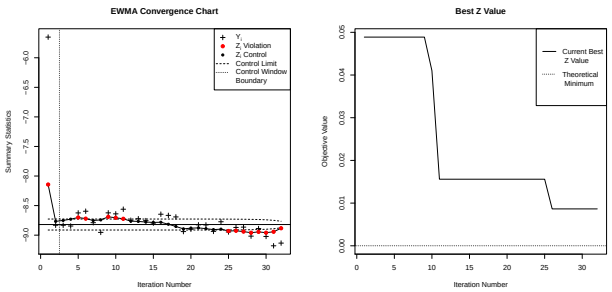


Figure 7: (*left*) The initial pre-converged state of the EWMA convergence chart. (*right*) The cumulative smallest objective function value at each iteration.

The EWMA convergence chart shows an out-of-control state in the control window, with violations of the upper control limit for older values and violations of the lower control limit for more recently observed values. This pattern indicates that the $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ values are still in a steady decreasing state, and more iterations of the optimization procedure are required. As the optimization procedure is allowed to proceed these violations move out of the control window and eventually a set of $\mathbb{E}[\ \log \mathrm{I}(\boldsymbol{x})\ ]$ values fill the control window which do not display any violations inside the control window, as seen in Figure (**??**, *left*). Furthermore Figure (**??**, *left*) demonstrates convergence because virtually all of the values beyond the control window violate the upper control limit.

As a check of the accuracy of the EWMA convergence chart, the right panels of both Figures (**??**) and (**??**) track the smallest objective function evaluation observed. Notice that in Figure (**??**, *right*) the surrogate model appears to have found a suboptimal location within Rosenbrock's valley, but the EWMA convergence chart does not indicate convergence. In Figure (**??**, *left*) the EWMA convergence chart does indicate convergence shortly after the surrogate model finds a value that is indistinguishable from Rosenbrock's theoretical minimum value.

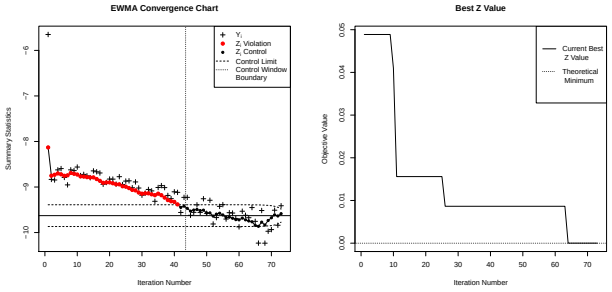
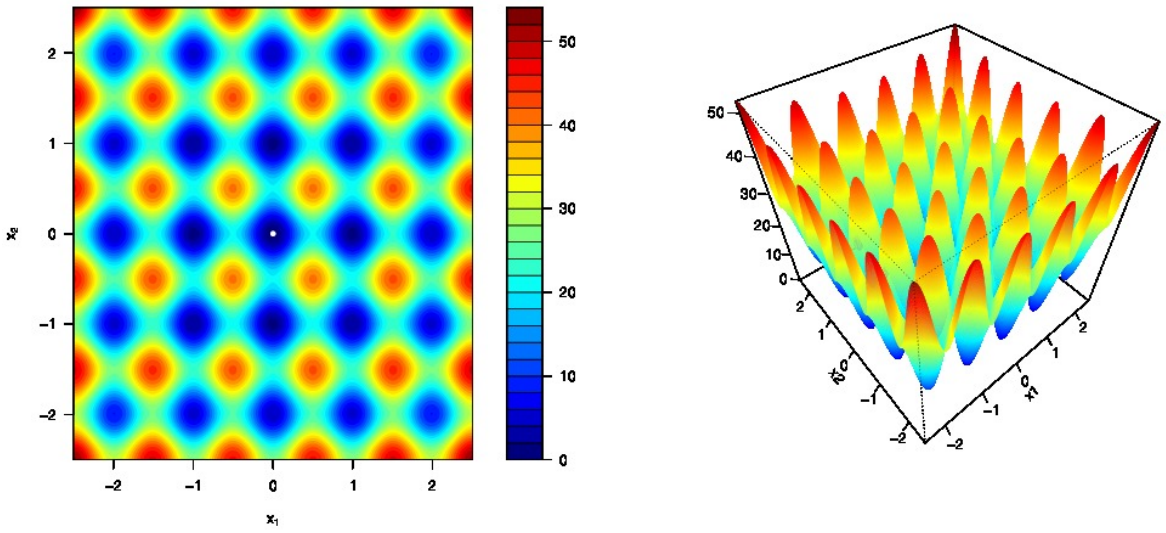


Figure 8: (*left*) The converged state of the EWMA convergence chart. (*right*) The cumulative smallest objective function value at each iteration.

3.2 Rastrigin



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

The Rastrigin function [?] is a commonly used test function on genetic algorithms; in this setting it offers an interesting example for considering convergence for highly multimodal objective functions. The global behavior of Rastrigin is dominated by the spherical function, $\sum_i x_i^2$, however Rastrigin has been oscillated by the cos(.) function, and shifted, so that it achieves a global minimum value of 0 at the lowest point, of its lowest trough at (0, 0).

Tuning λ and w in this case requires careful consideration of how Rastrigin's many modes effect the exploration process. Again the basic parameter tuning strategy is to first determine an appropriate value for λ, then tune w conditionally on the expected behavior of the EWMA statistic for the chosen value of λ. Rastrigin's many modes mean that as the objective landscape is explored, the algorithm will regularly find features indicative of possible new minima. Initially this drives up the value of the EI criterion, but as each of these modes are adequately explored, and disregarded, the EI value experiences large downward shifts. To accurately track these large shifts it is required to choose a large value for λ. Recall that typical values for λ lie in the range $0.1 \leq \lambda \leq 0.3$; in this case it is sufficient to

simply choose the upper threshold of this typical range, thus λ is set to $\lambda = 0.3$. Since the increased value for λ results in a more actively fluctuating EWMA statistic, the choice of w must reflect this knowledge. In the default case $w = 30$, but to account for the increased fluctuations brought about by a larger λ value, w is increased to $w = 40$.

As before, the initial EWMA convergence chart, seen in Figure (**??**, *left*), indicates a lack of control, as demonstrated by the violations of the control limits inside the control window. This out-of-control chart indicates that the algorithm has not yet converged, which is confirmed Figure (**??**, *right*), showing that initially the smallest objective value is far from the theoretical minimum. Considering Figure (**??**, *right*), after about 55 iterations the algorithm finds nearly the theoretical minimum, but it takes until about iteration 90 for enough of the large EI values to move out of the control window so that the EWMA control chart can claim convergence. Notice that since the value of λ was increased here, the fluctuations in the EWMA statistic have also increased as compared with Figure (**??**, *left*). The larger value of w is needed to alleviate the possibility of introducing small sample biases, introduced by the increased EWMA fluctuations, in long-run average estimates in computing the control limits. Thus decreasing w here would increase the prevalence of false claims of convergence.

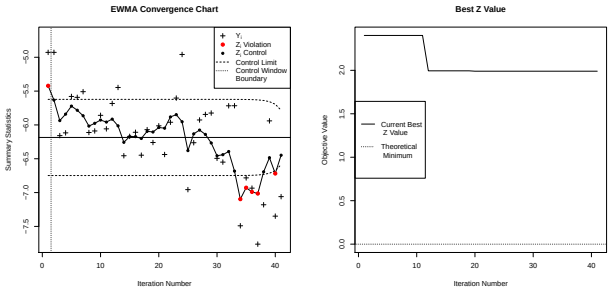


Figure 9: Initial EWMA convergence chart and smallest objective function value.

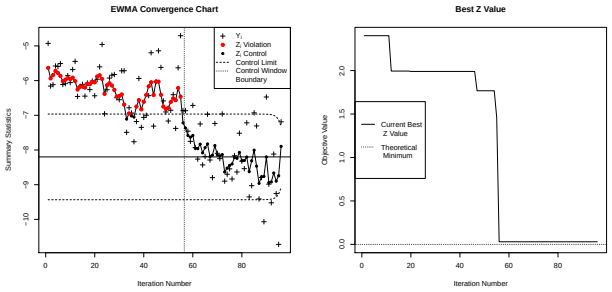


Figure 10: Final EWMA convergence chart and smallest objective function value.