A Novel Termination Criterion for Optimization

Venkatram Padmanabhan and R. Russell Rhinehart*
School of Chemical Engineering, Oklahoma State University

Abstract

A novel method for identification of steady state is demonstrated as the termination criterion for the optimization stage of modeling empirical data. The method is computationally efficient, robust and provides advantages over existing methods. It is described, and its utility is demonstrated on modeling simulated data.

Introduction

Nonlinear, least squares optimization is commonly used to determine model parameter values that best fit the empirical data by minimizing the sum of squared deviations (SSD) of data to model, termed the Objective Function (OF). Such models are commonly used in control and optimization. Common nonlinear optimization methods include Marquardt-Levenberg, Gauss-Newton, Nelder-Mead Simplex, and successive quadratic. Nonlinear optimization proceeds in successive iterations as the search progressively seeks the optimum parameter values, commonly termed decision variables (DV). As the optimum is approached, the optimization procedure needs a criterion to stop the iterations. The criterion should desirably stop the search when subsequent changes in the decision variable values do not improve the objective function value. However, the current stop-optimization criteria of thresholds on values of either objective function, changes in objective function, change in decision variable, or number of iterations require *a priori* knowledge of the appropriate values. They are scale dependent, application dependent, starting point dependent, and optimization algorithm dependent; right choices require human supervision.

This work explains, demonstrates, and evaluates a novel stop-iteration criterion for least squares optimization, which is scale-free and requires no prior knowledge of the optimum. It stops iterations when there is no statistical evidence of improvement relative to the variation in the data.

Corresponding author: R. R. Rhinehart, Chemical Engineering, 423EN, Oklahoma State University, Stillwater, OK 74078-5021, email: rrr@okstate.edu

The technique was initially revealed by Natarajan and Rhinehart [1], and certain aspects of it were quantitatively evaluated by Iyer and Rhinehart [2] for neural network training. This work generically extends the method to a variety of empirical optimization techniques and applications.

Method Description

An observer of an optimization procedure for empirical data will note that the sum of squared deviations (SSD) between the data and the model, the objective function value (OF), drops to an asymptotic minimum with progressive optimization iterations. The novelty of this method of observing progressive improvement on OF, is to calculate the sum of squared deviations (SSD) of a randomly selected subset of data (a different randomly selected subset at each iteration). The random subset SSD will appear as a noisy signal relaxing to its noisy steady state value as iterations progress.

By using a random subset of data to provide a SSD value at each iteration, the "noise" is independently distributed; and, at steady state, when convergence is achieved, the noise reflects the variance in the data. The noise is Chi-Square distributed, with an average equal to the residual variance (model-to-data mismatch). When the noisy signal reaches a statistical steady state, the optimization has progressed to the point where there is no statistically significant improvement in OF with respect to data variance; and optimization should be stopped. Since, the test looks at signal-to-noise ratio, it is scale independent and "right" for any particular application.

There are many ways to determine whether a signal is at steady state, or more properly stated, whether to accept or reject the null hypothesis. We choose the method of Cao and Rhinehart [3, 4] because of its computational simplicity. It presumes no auto-correlation in the noise, a condition which is satisfied by the random selection of data for the objective function value at each

iteration. Steady state is accepted when the ratio statistic in the method is less than unity.

Note that the optimization procedure usually needs the SSD for the full data set. The random subset SSD is only used for the convergence criteria, not on the optimizer logic.

Procedure for Evaluation

The method was examined using three optimization methods on each of three types of data sets. For each of the nine cases, the investigation approach is as follows:

1. The optimization methods were run for excessive iterations, as visually defined.
2. After each optimizer iteration, 20% of the total number of data points were randomly selected to calculate the sum of squared deviations.
3. A plot between the root mean square of the sum of squared deviations of the random subset and the number of iterations is made for visual analysis. The method does not require such a graph.
4. Model parameter values are recorded twice: first when the random subset SSD is determined [3, 4] to be at steady state, and finally after excessive iterations.
5. The models that result from these two parameter sets are visually compared by graphs, and quantitatively compared by analysis of variance. Since the two models are visually indistinguishable, those graphs are not shown here.

Three common, but distinct optimization methods from MatLAB Version 6.5, were used for this evaluation.

Nelder-Mead Simplex Method: Because it does not use derivative information, the Nelder-Mead method falls in the general class of *direct search methods*. Each iteration of a simplex-based direct search method begins with a simplex, specified by its $n + 1$ vertices (where n is the number of DVs), and the associated function values. The next simplex is obtained by taking the "mirror image" of the point with largest function value.

Marquardt-Levenberg Method: The Marquardt-Levenberg method combines the advantageous functionality of two fundamental methods, namely steepest descent and Newton-Raphson; both use numerical derivative information.

Gauss-Newton method: At each iteration, the Gauss-Newton method recalculates the Jacobian to provide derivative information. It moves along the "best" direction to obtain the new DV values.

Three simple but diverse applications were selected for each optimization approach.

Linear Function: The model equation selected for this linear problem is $y = Ax + B$ and the number of data points is 20. The linear model that was used to generate the data is given by $y = A(x + randn) + B + randn(size(x))$. The '*randn*' function adds Gaussian distributed, zero mean, unity variance, random variation [NID(0,1)] to a particular "*x*" value. Adding uncertainty to the independent variable is a non-conventional practice, but adds realism by simulating uncertainty in experimental control. The "*size(x)*" argument generates a vector of perturbations to the vector of "*y*" values – of the same number of elements as the "*x*" vector.

Nonlinear Function: The model equation selected for this nonlinear problem is $y = A\ln(Bx)$ and the number of data points is 40. The nonlinear model that was used to generate the data is given by $y = A\ln(B(x + randn)) + randn(size(x))$.

Multivariable Function: The model equation selected for this multivariable problem is $z = A\sqrt{x} + B\sqrt{y}$ and the number of data points is 20. The multivariable model that was used to generate the data is given by $z = A\sqrt{(x + randn)} + B\sqrt{(y + randn)} + randn(size(x))$.

Results

Figures 1, 2, and 3 show the plots between the root mean square of the sum of squared deviations of the random subset and the number of iterations using the three optimization methods for the linear model. Each method was run for 40 iterations (judged to be an excessive number by observing that there is no change in the OF.) and the iteration at which the method indicated steady state is shown by a vertical line with the iteration number. The reader may agree that the vertical line marks an iteration, after which, there is no visual evidence of statistical improvement in OF.

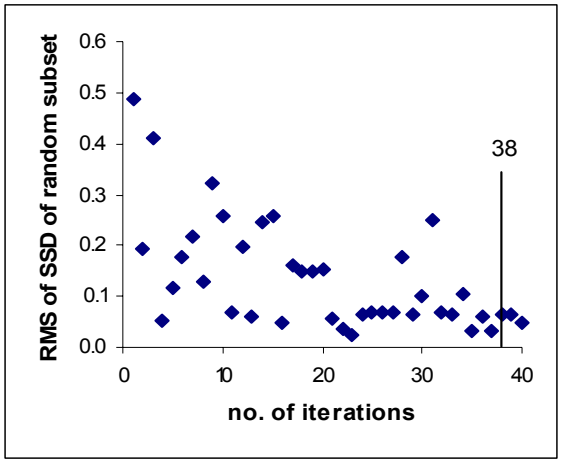

Figure 1 - RMS of SSD of random subset for a linear equation using the Nelder-Mead Simplex

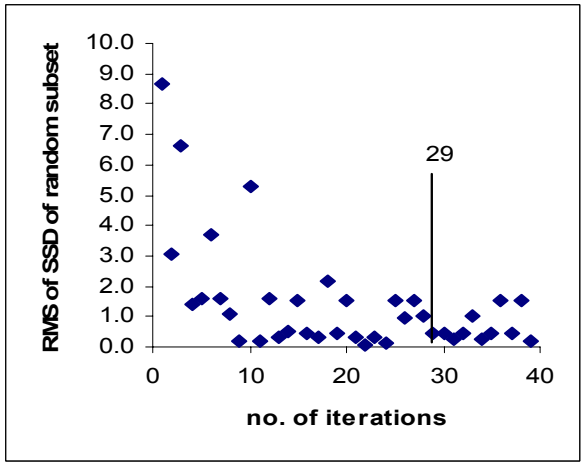

Figure 2 - RMS of SSD of random subset for a linear equation using the Marquardt-Levenberg Method

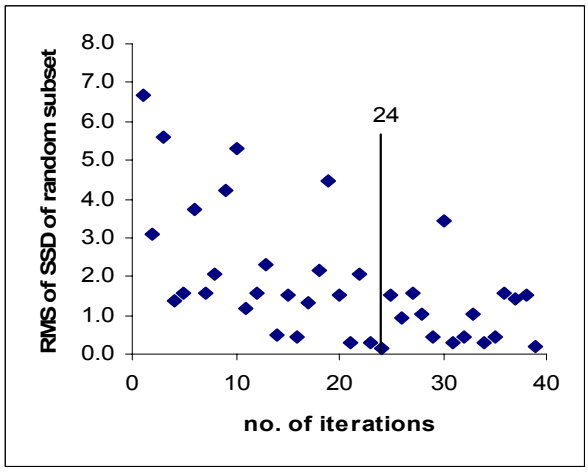

Figure 3 - RMS of SSD of random subset for a linear equation using the Gauss-Newton Method

Table 1: Goodness of fit for the linear model

Method	F-Statistic	p-value
N-M	0.999740	0.499
M-L	0.999800	0.499
G-N	0.999798	0.499

For these three case studies, Table 1 shows the F-statistic and its p-value. The F-statistic is calculated by the ratio of squared residuals, the sum of squared deviations between data and the model based on stopping at steady state to that from excessive iterations.

$$F - statistic = \frac{1}{N-1}\frac{\sum(SSD_1)}{\sum SSD_2}.$$

The optimization result with excessive iterations is accepted as the most perfect model for the particular random realization of the data. It is expected that any model from fewer iterations should not have as good a SSD, and the F-statistic values should be less than 1.0. However, if the new stopping criterion is good, the ratio of SSD measures will be close to unity. The data shows this. The p-value indicates the percentiles of the F distribution. For this case study, p-values of the F-statistics indicate that the curves are statistically indistinguishable.

The actual data and the y-x curves from the two models are not shown because, visually, the curves match the data, and the curves are indistinguishable.

Figures 4, 5, and 6 show the plots between the root mean square of the sum of squared deviations of the random subset and the number of iterations using the Nelder-Mead Simplex, Marquardt-Levenberg and Gauss-Newton methods respectively for the nonlinear model. Each method was run for 75 iterations and the iteration at which the method indicated steady state is shown by a vertical line with the iteration number. The reader may agree that it is an appropriate stage to stop.

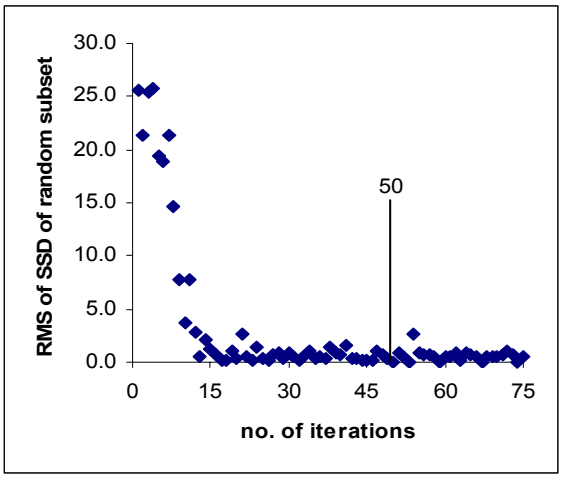

Figure 4 - RMS of SSD of random subset for a nonlinear equation using the Nelder-Mead Simplex

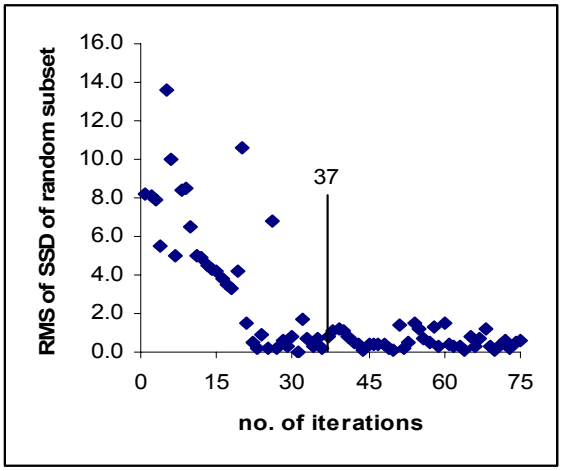

Figure 5 - RMS of SSD of random subset for anonlinear equation using the Marquardt-Levenberg Method

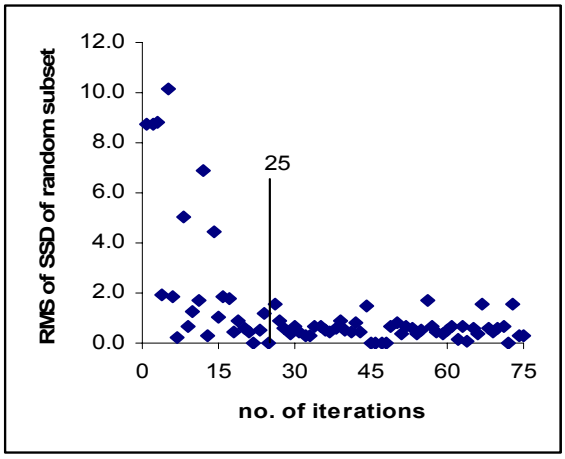

Figure 6 - RMS of SSD of random subset for a nonlinear equation using the Gauss-Newton Method

Table 2: Goodness of fit for the nonlinear model

Method	F-Statistic	p-value
N-M	0.99070	0.488
M-L	0.99998	0.498
G-N	0.99537	0.493

For the above three case studies, Table 2 shows the F-statistic and its p-value. Again, for this case study, p-values of the F-statistics indicate that the curves are statistically indistinguishable.

Again, the actual data and the y-x curves from the two models are not shown because, visually, the curves match the data, and the curves are indistinguishable.

Figures 7, 8, and 9 show the plots between the root mean square of the sum of squared deviations of the random subset and the number of iterations using the Nelder-Mead Simplex, Marquardt-Levenberg and Gauss-Newton methods respectively for the multivariable model. Each method was run for 35 iterations and the iteration at which the method indicated steady state is shown by a vertical line with the iteration number. The reader may agree that the vertical line marks an iteration, after which, there is no visual evidence of statistical improvement in OF.

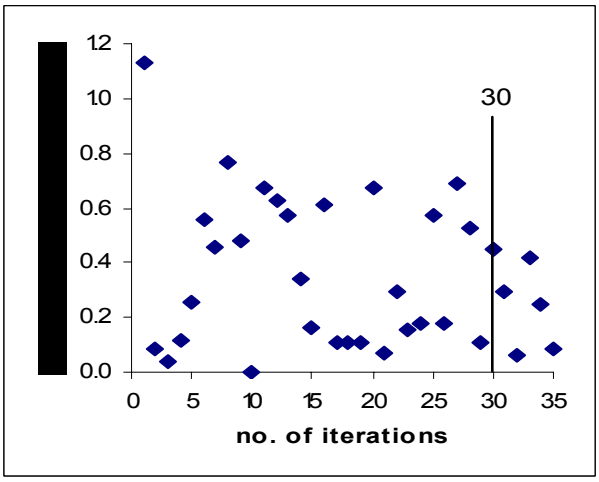

Figure 7 - RMS of SSD of random subset for a multi-variable equation using the Nelder-Mead

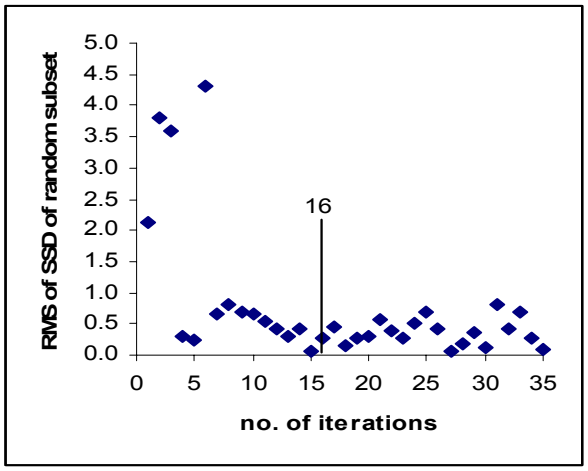

Figure 8 - RMS of SSD of random subset for a multivariable equation using the Marquardt-Levenberg Method

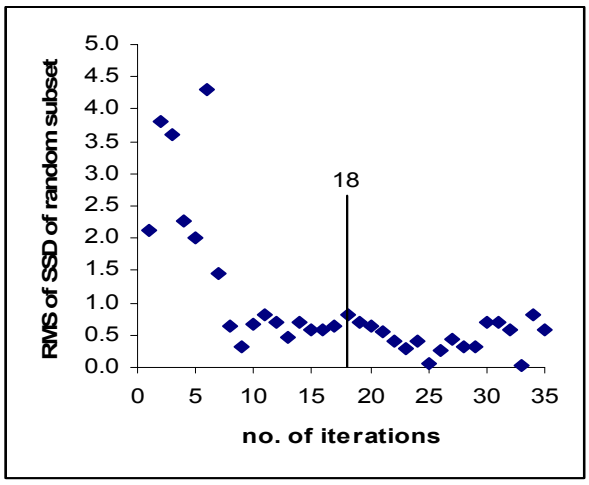

Figure 9 - RMS of SSD of random subset for a multivariable equation using the Gauss-Newton Method

Table 3: Goodness of fit for the multivariable model

Method	F-Statistic	p-value
N-M	0.91463	0.419
M-L	0.99832	0.498
G-N	0.96611	0.468

For the above three case studies, Table 3 shows the F-statistic and its p-value. Again, for this case study, p-values of the F-statistics indicate that the curves are statistically indistinguishable.

The actual data and the y-x curves for the two models are not shown because, visually, the curves match the data, and the curves are indistinguishable.

From all the above graphs, it can be observed that the RMS value of the sum of squared deviations of a random subset starts at a high value and gradually decreases until it reaches a noisy steady state.

Discussion

In an earlier investigation this technique was used as the stopping criterion for both the Levenberg-Marquardt and error backpropagation methods for neural network training [2]. While the number of decision variables (15 to 30 weights) was larger than the number in this work (2 model coefficients), the application was of one type. While all problems in this work were low dimensional, this work extends the applications and optimizations to demonstrate the practicality of this steady state stopping criterion on a wider variety of problems.

Since these optimization applications were of low dimension, the optimization approaches immediately started "down hill" to minimize the Objective Function value. By contrast, in the prior work with many decision variables, the improvement in the OF value in the initial iterations was often slight, and the plot of random subset SSD with respect to iteration number would appear to be at steady state initially. This would stop the optimization prior to making progress. Consequently, the broader, two-condition rule, "Stop optimization when steady state is identified subsequent to a transient period." was not initially described in the introduction to this work. That additional logic would not affect these results, but would be part of a general routine, as illustrated in Figure 10.

The comparison of this steady state stop optimization criterion to the conventional operator-decision based on cross validation in training neural networks concluded that the automated method gave equivalent RMS values and chose to stop with fewer iterations [2]. The automation advantage of this method was subsequently used in evaluating the probability of finding a global minimum in training thousands of neural networks [4]. This

work, supports that finding on a variety of conventional applications.

Conclusion

The application of a novel stopping criterion for optimization, based on identifying steady state of a random subset sum of squared deviations with respect to iteration number, formerly explored for neural network training, has been extended to demonstrate advantages on a variety of empirical modeling optimization applications. The novel stop optimization criterion gives equivalent results (as measured by model residuals) to the best possible results, with a sufficient (not excessive) number of iterations and without *a priori* knowledge of the optimization problem (scale, end-point values, and other classic stopping criteria).

Acknowledgement

The authors appreciate partial financial support from the Edward E. and Helen Turner Bartlett Foundation.

References

1. Natarajan, S., and R. R. Rhinehart, "Automated Stopping Criteria for Neural Network Training," Proceedings of the 1997 American Control Conference, Proceedings, June 4-6, 1997, Albuquerque, NM, paper TP09-4, pp. 2409-2413.
2. Iyer, M. S., and R. R. Rhinehart, "A Novel Method to Stop Neural Network Training," Proceedings of the 2000 American Control Conference, June 28-30, 2000, Chicago, IL, Paper WM17-3, pp 929-933
3. Cao, S., and R. Russell Rhinehart, "An Efficient Method for Online Identification of Steady State" J. Process Control, 5 (1995): 363.
4. Cao, S., and R. Russell Rhinehart, "Critical Values for Steady State Identifier" J. Process Control, 7 (1997): 149.
5. Iyer, M. S., and R. R. Rhinehart, "A Method to Determine the Required Number of Neural Network Training Repetitions," IEEE Transactions on Neural Networks, Vol. 10, No. 2, March, 1999, pp. 427-432.

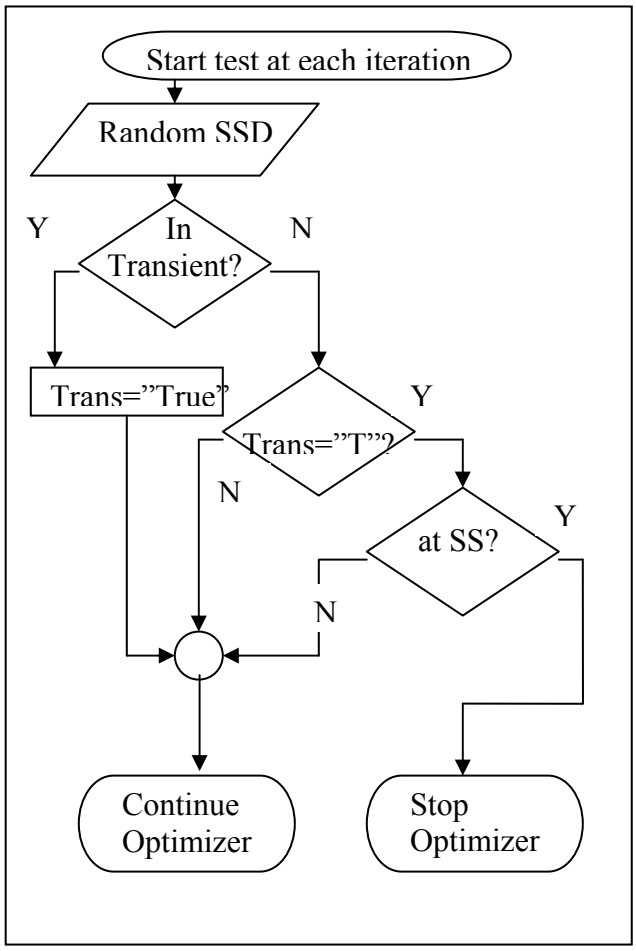

Figure 10 – Flowchart for Stopping Criterion