

A Global Optimization Method Using Gaussian Distribution Perturbation

Akaraphunt Vongkunghae

Faculty of Engineering, Naresuan University, Thailand

akaraphuntv@nu.ac.th

Abstract

A novel optimization method and a demonstration on a one-dimensional problem are presented in this paper. A random perturbation having a Gaussian distribution is used to achieve the global minimum. The experiment reveals that the variance of the perturbation is vital for global optimum achievement. However, to reach the global optimum, the large value of the variance is a trade of the weight adjustment accuracy. The result also shows that the system can arbitrarily close the global minimum by selecting the standard deviation parameter controlling the Gaussian distribution shape of the perturbation.

1. Introduction

In the optimization on multi – local optimum surface, the well known problem is that the system cannot reach the true optimum. It is a long term standing problem in the all kinds of adaptive systems. None of the gradient base algorithms [1, 2] can trace the multi local optimum surface. For non-gradient base algorithms, many of them can reach the true optimum of adaptive systems. One of those is the well known genetic algorithm (GA) [3]. In the genetic algorithm and its variants, a large size of data storage

to keep the genetic candidate populations (candidate patterns) are required which are impractical in many real time systems. It also needs expensive operations e.g. cross over and permutation on a large size of population. In practical, the adaptation algorithm should be simply enough for less memory usage while it is able to achieve the true optimum point of its problem space. In [4], the compact genetic algorithm (cGA) was proposed by G. R. Harik *et al.* The cGA uses less amount of memory than simple GA (sGA) does. The cGA is equivalent to sGA only in a certain condition of problem, referred in [4] as the order – one behavior of sGA with uniform crossover. The attempt in [4] is trying to minimize the memory usage while the global optimization is expected as the GA can do in general.

In this paper, a novel adaptation algorithm that uses a less memory usage than the GA does is proposed in the next section. The approach is not based on the genetic algorithm. It is more likely close to the gradient base methods [1, 2] but it does not be stuck in a local optimum as the gradient base methods do.

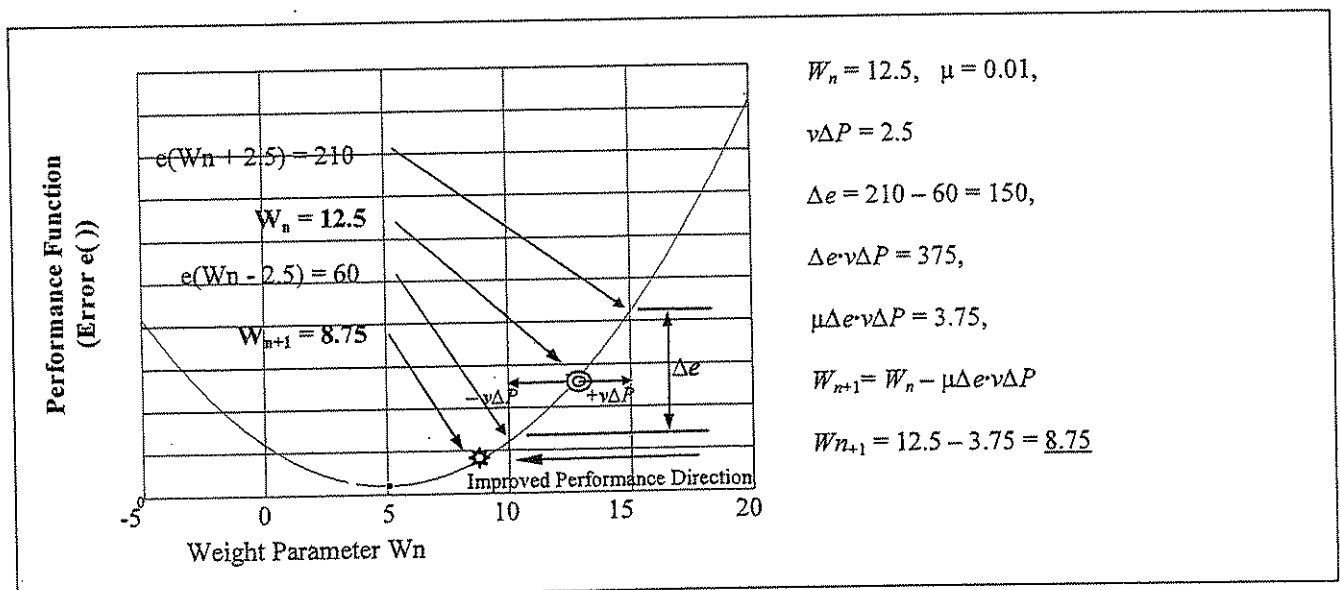


Figure 1: An Example of Calculations

2. Algorithm

The expression of this proposed algorithm for the case of minimization is

$$\Delta e = e(W_n + v\Delta P) - e(W_n - v\Delta P) \quad (1)$$

$$W_{n+1} = W_n - \mu \cdot \Delta e \cdot v\Delta P \quad (2)$$

where

W_{n+1} is the adapted weight or coefficient vector of system,

W_n is the old weight or old coefficient vector of system,

ΔP is a perturbation vector having a normal Gaussian probability distribution, i.e. its standard deviation is 1.0,

v is a distribution factor to assign the distribution shape of the Gaussian perturbation $v\Delta P$. It is always positive i.e., the v is the standard deviation of the $v\Delta P$,

$e(W_n + v\Delta P)$ is the evaluated performance at a point with a positive perturbation ($W_n + v\Delta P$),

$e(W_n - v\Delta P)$ is the evaluated performance at a point with a negative perturbation ($W_n - v\Delta P$),

Δe is the difference of evaluated performances between the points of positive perturbation and negative perturbation,

μ is a learning rate of adaptation.

The principle behind the equation (2) is the correlation between Δe and $v\Delta P$, shown in Figure 1. The weight W_n should be adapted in a direction that the performance error is minimized as the gradient base algorithms do by inverting the gradient sign in its weight adaptation calculation. However to save computation power and memory use, an alternative way to calculate the direction for weight adaptation to achieve the least error can be done by looking at the instantaneous correlation $\Delta e \cdot v\Delta P$. That is the correlation between the $e(W_n + v\Delta P) - e(W_n - v\Delta P)$ and the $v\Delta P$. Once W_n is perturbed by a positive $v\Delta P$ and the error difference result is positive, the instantaneous correlation $\Delta e \cdot v\Delta P$ is then a positive value. For the case of minimization to reduce the error, the W_n must be adapted in opposite direction comparing to the sign of the correlation $\Delta e \cdot v\Delta P$. Therefore, there is a minus sign in front of the $\mu \Delta e \cdot v\Delta P$ as show in the equation (2). In Figure 1, the current value of W_n is 12.5 and the size of the perturbation $v\Delta P$ is 2.5. After the calculation to adapt the current weight W_n is done, the value of the adapted weight (W_{n+1}) is 8.75. From the

performance curve of Figure 1, the error value $e(8.75)$ of the adapted weight is less than the error value $e(12.5)$. Therefore, minimizing error is as the result.

To do maximization, the equation (2) can be modified using the positive sign instead of the negative sign as expressed in the following equation (3).

$$W_{n+1} = W_n + \mu \cdot \Delta e \cdot v\Delta P \quad (3)$$

3. Least Correlation Principle

Once, this algorithm had been demonstrated to my colleagues, there were many responses. Some said that the algorithm is similar to the gradient base algorithms. However, we found that we are not able to use the Taylor series [5] to have a linearization at the point W_n as we all are doing in the analysis of the gradient base methods. To understand how the equation (1) and the equation (2) work, we first look at an example, the minimization performance function in the equation (4).

$$e(w) = m \cdot (w + 2)^2 + a \cdot \sin(6.2832 \cdot w) + b \cdot \sin(5.0265 \cdot (w - 10)) \quad (4)$$

The performance curve $e(w)$, the parameters m , a , and b are equal to 50, 700, and 1000 respectively, is shown in Figure 2. Note that the function has many local minimum points. The simulated expected values of the correlation $E[\Delta e \cdot v\Delta P]$, when v is 0.70, is also shown along with the performance curve. To see how the weight W_n in the equation (2) converges to a solution, take the expected value of the equation (2). The expectation [6] of the equation (2) is the equation (5)

$$E[W_{n+1}] = E[W_n] - \mu \cdot E[\Delta e \cdot v\Delta P] \quad (5)$$

Clearly the optimal weight W_n is reached when the term $E[\Delta e \cdot v\Delta P]$ is equal to zero. In Figure 2, the $E[\Delta e \cdot v\Delta P]$ is zero when W is equal to -2.00. With the equations (1) and (2) and demonstration in Figure 2, the coefficient W_n will converge to the optimum -2.00 which is the point that the correlation of $\Delta e \cdot v\Delta P$ is the least or equal to zero.

4. Infinite Range of Gaussian Distribution

In Figure 2, the true global minimum point is at -0.28. Because the performance characteristic in Figure 2 is not a symmetrical function, the point that the expected value is equal to zero is not on the true global optimum [7]. This is a reason that the random perturbation with Gaussian distribution is used in this algorithm for reaching to the true global optimum. The random number with Gaussian distribution [6] can be from infinite minus to infinite positive. Once the coefficient W_n converges to the optimum of the least correlation, seeking the true minimum near the least correlation optimum can be done by looking at the minimum value of sampling $e(W_n + v\Delta P)$, $e(W_n - v\Delta P)$. The value W_{min} giving the minimum of the performance function $e(w)$ will be

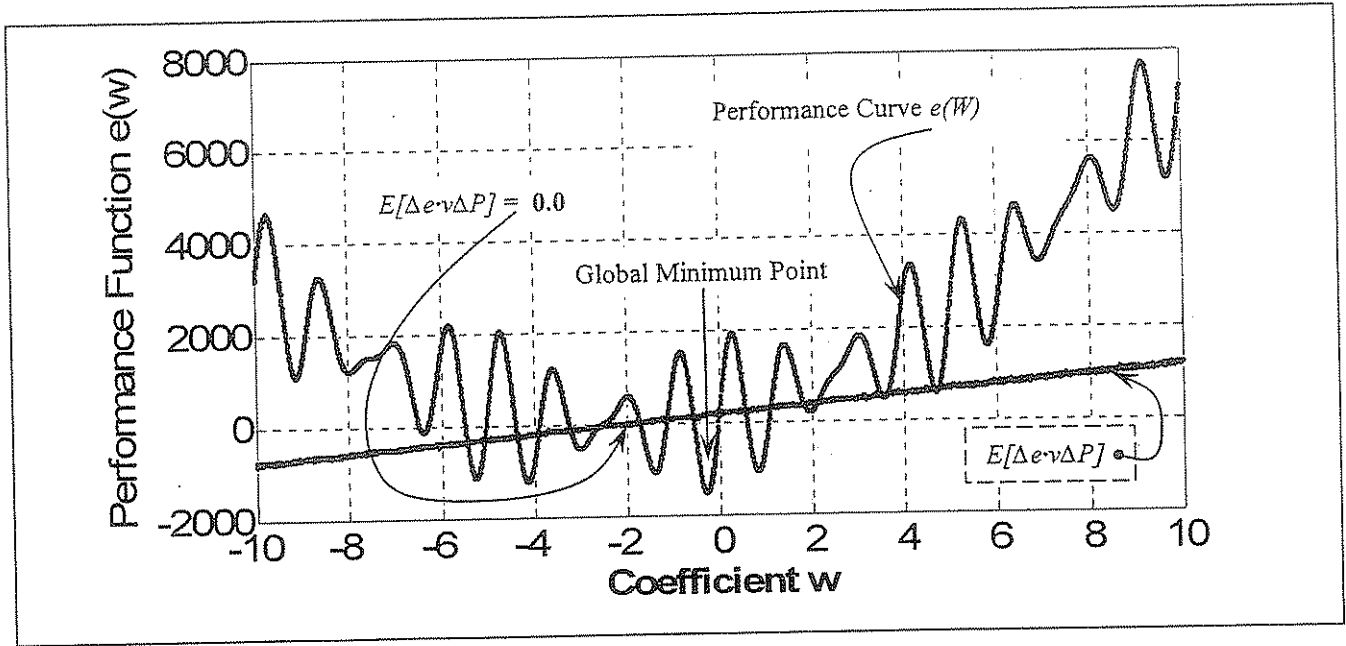


Figure 2: The curves of the performance function $e(w)$ and the expected value $E[\Delta e \cdot v \Delta P]$, $m = 50$, $a = 700$, $b = 1000$ and $v = 0.70$

updated and closer to the true minimum value every time of adaptation cycles. The expression for updating the value of W_{min} is

$$W_{min} = \begin{cases} W_n + v\Delta P & \begin{cases} e(W_n + v\Delta P) < e(W_{minold}) \\ e(W_n + v\Delta P) < e(W_n - v\Delta P) \\ e(W_n + v\Delta P) < e(W_n) \end{cases} \\ W_n - v\Delta P & \begin{cases} e(W_n - v\Delta P) < e(W_{minold}) \\ e(W_n - v\Delta P) < e(W_n + v\Delta P) \\ e(W_n - v\Delta P) < e(W_n) \end{cases} \\ W_n & \begin{cases} e(W_n) < e(W_{minold}) \\ e(W_n) < e(W_n + v\Delta P) \\ e(W_n) < e(W_n - v\Delta P) \end{cases} \\ W_{minold} & \text{otherwise} \end{cases} \quad (6)$$

In practical, the shapes of the performance surface are unknown. If it is a symmetrical shape, the converged weight W_n is at the same place of the true optimum. Occasionally when it is not, the weight W_n converges far from the true optimum. Using the Gaussian perturbation allows the weight adjustment is able to cross the hills of performance surface. The weight W_n is adjusted by the equations (1) and (2) and moved to its optimum depending on its performance surface shape. At the same time, with its infinite random range property, the true global minimum can iteratively be detected.

5. Experiments

To demonstrate the ability of the algorithm in general, the equation (4) is used as the performance

function. We aim to find what the value of W at the minimum of the function $e(W)$. The parameters for the equation (4) are the same as shown at the caption of Figure 2. The learning rate μ and the distribution factor v are 0.00005 and 0.70 respectively. Figure 3 shows the adaptation curves of minimization. The top graph is the adaptation curve of performance $e(W_n)$. The solid line is the error $e(W_n)$ and the dash line is the minimum value $e(W_{min})$ from the equation (6). With the initial conditions $W_n = 10$ at the iteration 0 and $W_n = -10$ at the iteration 1000, both of the $e(W_n)$ and $e(W_{min})$ lines are generally decreasing along the iteration number.

At the tails of adaptations with both initial conditions, the $e(W_n)$ converge to the least correlation optimum at the point that $E[\Delta e \cdot v \Delta P]$ is equal to zero as demonstrated in Figure 2. The $e(W_{min})$ of the both cases of initial conditions converges to the true minimum about -1526 while the $e(W_n)$ converges to the least correlation optimum about 300.

The bottom graph of Figure 3 is the adaptation curve of weight W . The solid line for W_n converges to the least correlation optimum -2.00. The dash line for W_{min} converges to the value -0.28 which is the desired value corresponding to the global minimum point in Figure 2.

The learning curve which is the solid line in the top graph of Figure 3 is not a typical one as we see in general adaptation processes. There appear some oscillations which is the result of climbing out the local minima of the performance surface. The solid line is always on the top of dash line. At the point that the solid line $e(W_n)$ is touching the dash line $e(W_{min})$, the adaptation is going around a local minimum. Moving out from local minima, the $e(W_n)$ must be increased and be decreased to local minimum repeatedly while the W_n is moving closer to the optimum W .

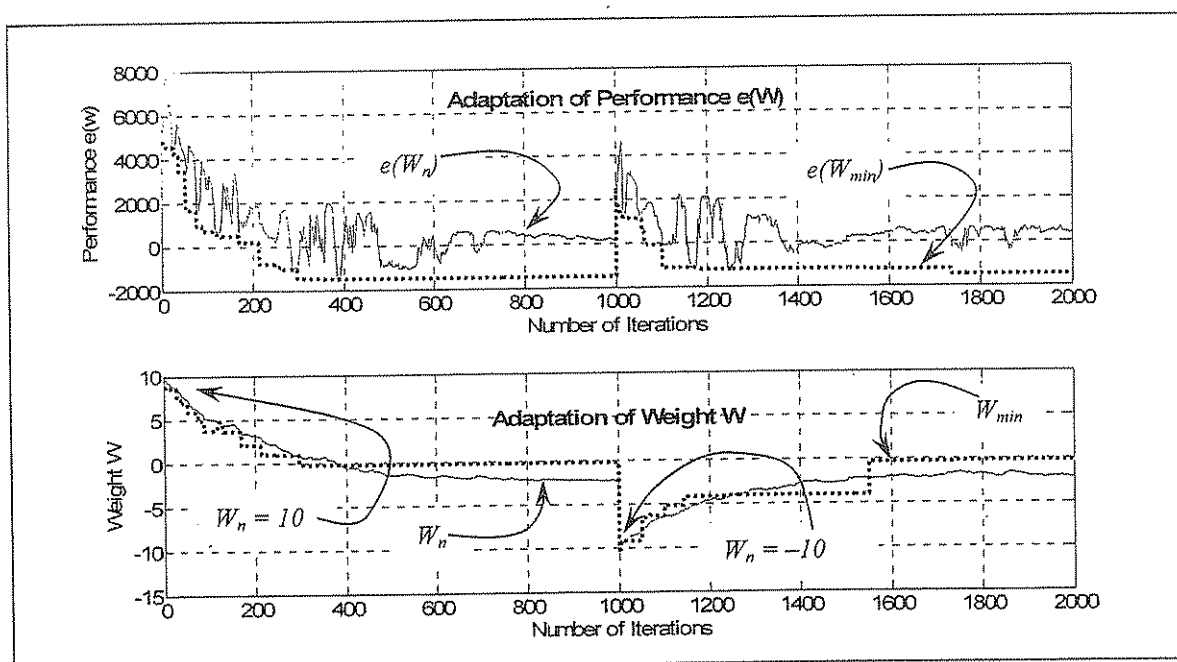


Figure 3: Adaptation Curves

- The top graph is the adaptation curve of performance $e(W)$. The solid line is the adaptation performance $e(W_n)$. The dash line is the detected minimum value of $e(W_{min})$ from the equation (6) using Gaussian random property.
- The bottom graph is the adaptation of weight W . The solid line is the adapted weight W_n . The dash line is the detected weight W_{min} giving minimum value of $e(W_{min})$.

6. Conclusions

This paper demonstrated the ability of a novel adaptation method which has a high potential for using in smart systems such as artificial neural network systems [2, 8], adaptive filters [1]. The case of unsymmetrical error surface is given as an example. The author intends to demonstrate for discussions and to point out the real practical problems. Usually we cannot know what the shape of performance function.

To make the adaptation can climb over the hills; the distribution factor v must be large enough and wide cover the largest distant of two adjacent local minima. If the performance shape is symmetry, the point of the least correlation optimum and the global optimum are at the same place. Minimum detection using the equation (6) may not be needed.

However, if the performance shape is an unsymmetrical shape, the global minimum still be detected by the equation (6). We are taking a benefit from the infinite random range property of Gaussian perturbation. The global optimum will be discovered eventually because of the infinite random range of Gaussian perturbation.

7. Future Works

More researches are required for understanding the parameters μ and v to the adaptation process. Proof of its convergence and its implemented application are not provided here. These are opened research areas to many researchers. For the interested researchers in the

proof of convergence, R.C. Tumlade in [7] has laid a good foundation that everyone may follow his work.

The other well know minimization methods, the gradient descent technique, the least mean square algorithm, and back propagation algorithm in [1, 2] never use the infinite size perturbation technique. Only MadalineII network in [2] uses a small magnitude of perturbation with the minimal disturbance principle to train to multilayer networks. In [2] said that the training rule for MadalineII network is equivalent to the back propagation algorithm when the perturbation size is small. The proposed algorithm here does not agree to the minimal disturbance principle which is the basic principle of the other weight minimization methods. Therefore to make fair performance comparisons of the proposed algorithm to the other gradient based methods, carefully setting up experiments is necessary to have several view points for consideration.

References

- [1] Bernard Widrow and S. D. Stearns, *Adaptive Signal Processing*, Englewood Cliffs, NJ, Prentice-Hall, 1985.
- [2] Bernard Widrow and Michael A. Lehr, "30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation", *Proc. IEEE*, Sept. 1990, Vol. 79, No. 9, pp. 1415 – 1442.
- [3] Goldberg, David E (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer Academic Publishers, Boston, MA.

- [4] G. R. Harik, F. G. Lobo, and D. E. Goldberg, "The Compact Genetic Algorithm", *IEEE Trans. on Evolutionary Computation*, Vol. 3, NO. 4, November 1999.
- [5] E. Kreyszig, "Advanced Engineering Mathematics", Edition 7th, John Wiley & Sons, Inc. New York, 1993, ISBN 0-471-55380-8.
- [6] Athanasios Papoulis, "Probability, Random Variables, and Stochastic Processes", Edition 3th, 1991, McGraw-Hill Inc., New York, US, ISBN 0-07-048477-5.
- [7] R. C. Turnblade, "Random optimization and multipole adaptive control," Proceeding of a Symposium : Adaptive Control System, Garden City, Long island, New York, October 1960, page 143 – 177.
- [8] E. B. Baum, "On the capabilities of multilayer perceptrons," *J. Complexity*, Sept, 1988, Vol. 4 pp. 193 – 215.