2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Non Zig-Zag Convergence of the Steepest Descent Method in Minimizing Non-quadratic Form

Takashi Ozeki†

†Department of Information Processing, Faculty of Engineering, Fukuyama University
1 Sanzo, Gakuen-cho Fukuyama, Hiroshima, 729-0292 Japan
Email: ozeki@fuip.fukuyama-u.ac.jp

Abstract—In this paper, we discuss the limiting behavior of the search direction of the steepest descent method in minimizing general nonlinear function. It is well known that the search direction of the steepest descent method asymptotically alternates between two fixed directions if the given function is a quadratic form. Moreover, it is conjectured that the property holds for non-quadratic functions. However, we find that some non-quadratic functions do not take zig-zag directions. Also, we give some necessary conditions to cause the special cases.

1. Introduction

In the optimization problems for smooth functions, various kinds of descent methods are frequently used. The steepest descent method is the simplest and the most classical of these descent methods. Since this method involves the use of a gradient vector, it is also a fundamental of all descent methods. Thus it is important to study its properties. For example, it is well known that the search direction of the steepest descent method in minimizing a positive definite quadratic form asymptotically alternates between two fixed directions [1, 2, 3, 4]. Figure 1 shows that the direction of the steepest decent method behaves zig-zag for quadratic form with two variable. This property was conjectured by Forsythe and Motzkin [1] and proved by Akaike [2]. Forsythe further conjectured that the same property holds for general nonlinear function. The reason is that smooth and real-valued functions can be approximated to a certain positive definite quadratic form in the neighborhood of its local minimum by applying Taylor's expansion [5]. Therefore we can expect that the same phenomenon will be observed for non-quadratic form. However, we find that some non-quadratic functions do not have the same behavior and the direction of the steepest descent method do not move zig-zag. Moreover, we give some necessary conditions to cause the special cases.

2. Algorithm of the Steepest Descent Method

Let $f(\boldsymbol{x})$ be real for all $\boldsymbol{x}$ in Euclidean n-space R^n and twice continuous differentiable. In other words, the function belongs to C^2. Also, let $f(\boldsymbol{x})$ take a minimum value.

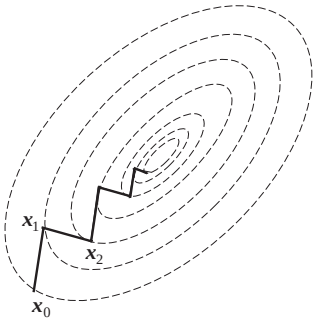

Figure 1: Behavior of the Steepest Descent Method.

The algorithm of the steepest descent method is given as follows.

Algorithm of the steepest descent method

1. Let $\boldsymbol{x}_0$ be an initial vector and put $k = 0$.
2. The search direction is determined by the steepest decent direction $-\nabla f(\boldsymbol{x}_k)$. The step size α_k is given by the smallest positive number which satisfies the next equation:
$$\frac{\partial f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k))}{\partial \alpha} = 0. \tag{1}$$
That is, the step size takes Curry's rule [4]. The next approximation vector is represented by
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k). \tag{2}$$
3. Evaluate a convergent condition. If the condition is not satisfied, increase k and return to the second step.

3. Behavior of the Steepest Descent Method

In this section, we shall discuss the limiting behavior of the search direction of the steepest descent method for general functions. To simplify the discussion, we assume that the function $f(\boldsymbol{x})$ takes a minimum 0 at the origin $\mathbf{0}$ and the Hessian matrix $\nabla^2 f(\mathbf{0})$ is diagonal by using a coordinate transformation. Then, the matrix $\nabla^2 f(\mathbf{0})$ becomes positive. Without loss of generality, we can discuss the behavior of the steepest descent method under these assumptions.

Definition 1 *Let $\boldsymbol{d}_k$ be a normalized vector of the search direction $-\nabla f(\boldsymbol{x}_k)$:*

$$\boldsymbol{d}_k = -\frac{\nabla f(\boldsymbol{x}_k)}{\|\nabla f(\boldsymbol{x}_k)\|}. \tag{3}$$

The following orthogonal theorem is well known as a essential property of the steepest descent method [3, 4].

Theorem 1 *Two successive search directions are mutually orthogonal. That is, it holds $(\boldsymbol{d}_k, \boldsymbol{d}_{k+1}) = 0$.*

Next, we introduce a symmetric matrix A_k with the size $n \times n$ as follows:

$$A_k \triangleq \int_0^1 \nabla^2 f(t\boldsymbol{x}_{k+1} + (1-t)\boldsymbol{x}_k)dt. \tag{4}$$

If approximation $\boldsymbol{x}_k$ converges to the origin $\mathbf{0}$, the matrix A_k becomes positive for sufficient large k because A_k converges to a positive matrix $\nabla^2 f(\mathbf{0})$. From

$$\begin{aligned} \nabla f(\boldsymbol{x}_{k+1}) &= \nabla f(\boldsymbol{x}_k) + A_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \\ &= \nabla f(\boldsymbol{x}_k) - \alpha_k A_k \nabla f(\boldsymbol{x}_k), \end{aligned} \tag{5}$$

and Theorem 1, by multiplying $\nabla f(\boldsymbol{x}_k)$ to both sides (5), we get

$$\begin{aligned} \alpha_k^{-1} &= \frac{(A_k \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k))}{\|\nabla f(\boldsymbol{x}_k)\|^2} \\ &= (A_k \boldsymbol{d}_k, \boldsymbol{d}_k). \end{aligned} \tag{6}$$

We define a mean by

$$\mu(\boldsymbol{d}_k) \triangleq \alpha_k^{-1} = (A_k \boldsymbol{d}_k, \boldsymbol{d}_k) \tag{7}$$

and a variance V_k by

$$V_k \triangleq \|A_k \boldsymbol{d}_k - \mu(\boldsymbol{d}_k)\boldsymbol{d}_k\|^2. \tag{8}$$

If the given function $f(\boldsymbol{x})$ is a quadratic form, A_k is a constant matrix $A \triangleq \nabla^2 f(\mathbf{0})$. Akaike treated eigenvalues of a positive matrix A as probability values. Then, he considered the set of squares of components of the search vector $\boldsymbol{d}_k$ corresponding to eigenvectors of the matrix A as a probability distribution. Then, $\mu(\boldsymbol{d}_k)$ becomes a mean and V_k becomes a variance. Moreover, he showed the variance increases monotonically and converges to a non-zero number. By using this property, he analyzed the behavior of the search direction [2]. So, we extended this idea to non-quadratic function and examine the variance V_k. However, in the case of non-quadratic function, each eigenvalue of the matrix A_k (sometimes even each eigenvector) changes with k and the variance V_k does not increase monotonically. Nevertheless, we got a next important lemma [7].

Lemma 1 *If $\nabla^2 f(\boldsymbol{x})$ has Lipschitz continuity in the neighborhood of $\mathbf{0}$, the variance V_k converges a nonnegative number.*

From now on, we assume Lipschitz continuity of $\nabla^2 f(\boldsymbol{x})$. Of course, if $f(\boldsymbol{x})$ belongs to C^3, this assumption is automatically satisfied.

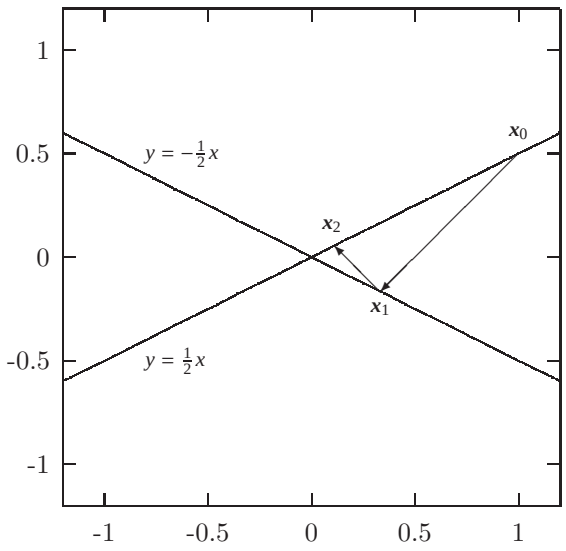

Figure 2: Behavior in the case of quadratic form.

3.1. In the case of positive variance

If the variance V_k converges to a positive number, we got a next result [7]. This theorem is an extension of Akaike theorem [2] to non-quadratic function .

Theorem 2 *Let assume the variance V_k converges to a positive number.*

1. *The search direction $\boldsymbol{d}_k$ can be asymptotically approximated by a linear combination of two eigenvectors of the matrix $\nabla^2 f(\mathbf{0})$.*
2. *The direction $\boldsymbol{d}_k$ asymptotically alternates between two vectors that are mutually orthogonal.*

Figure 2 shows the behavior of the steepest descent method in minimizing quadratic form $f(x,y) = \frac{1}{2}x^2 + y^2$ started from $\boldsymbol{x}_0 = (1, 0.5)$. In this case, approximation $\boldsymbol{x}_k$ moves between two lines. On the other hand, Figure 3 shows the behavior of the steepest descent method in minimizing non-quadratic form $f(x,y) = \frac{1}{2}x^2 + \frac{1}{3}x^3 + y^2$ started from $\boldsymbol{x}_0 = (\frac{-1+\sqrt{5}}{2}, 0.5)$. In this case, approximation $\boldsymbol{x}_k$ moves between two curved lines. Whether quadratic form or not, the search direction moves zig-zag at right angles.

Next, we give a sufficient condition in which the variance converges to a positive number.

Definition 2 *We call a function $f(\boldsymbol{x})$ is separable if the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is diagonal.*

Example 1 Let the function $f(x,y)$ be as follows:

$$f(x,y) = \frac{1}{2}x^2 + y^2 + x^3. \tag{9}$$

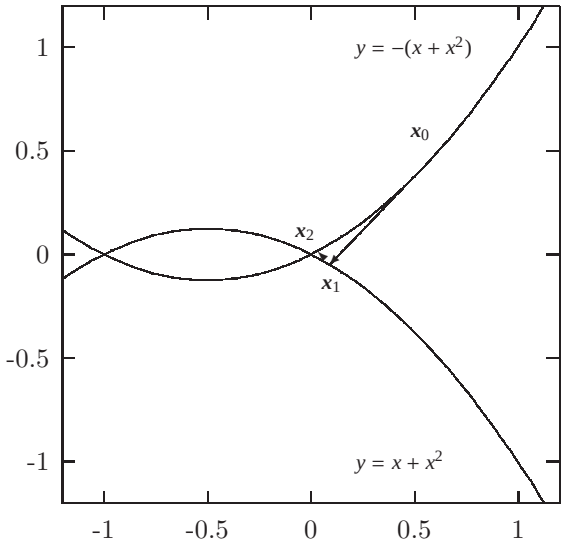

Figure 3: Behavior in the case of non-quadratic form.

Then, we get the Hessian matrix

$$\nabla^2 f(x,y) = \begin{pmatrix} 1+6x & 0 \\ 0 & 2 \end{pmatrix}. \tag{10}$$

Therefore, this function is separable. On the other hand, the function

$$f(x,y) = \frac{1}{2}x^2 + y^2 + x^2 y \tag{11}$$

is not separable. Because the Hessian matrix

$$\nabla^2 f(x,y) = \begin{pmatrix} 1+2y & 2x \\ 2x & 2 \end{pmatrix} \tag{12}$$

is not diagonal. Of course, quadratic form is obviously separable.

The following theorem give a sufficient condition in which the variance V_k converges to a positive number.

Theorem 3 *Let suppose that the steepest descent method does not converge by finite iterations. If the function* $f(\boldsymbol{x})$ *is separable, the variance does not converge to zero. Therefore, it holds theorem* 2.

3.2. In the case of zero variance

When the variance V_k converges to zero, the direction does not always behave zig-zag. In this subsection, we shall examine the special case. If the variance V_k converges to zero, from (8), all accumulation points $\boldsymbol{r}_\alpha$ of the sequence $\{\boldsymbol{d}_k\}$ must satisfy the equation

$$\|A\boldsymbol{r}_\alpha - \mu(\boldsymbol{r}_\alpha)\boldsymbol{r}_\alpha\| = 0. \tag{13}$$

Therefore, the search vector $\boldsymbol{d}_k$ must accumulate to eigenvectors of the matrix $A = \nabla^2 f(\boldsymbol{0})$. This is a first necessary condition in which the direction does not behave zig-zag. This depends on how we take an initial vector $\boldsymbol{x}_0$. Next, from Theorem 3, we also need that the function is not separable. Now, we give a simple example to satisfy these necessary conditions.

Example 2 Let the function $f(x,y)$ be as follows:

$$f(x,y) = \frac{1}{2}x^2 + y^2 + x^3 y - xy^3. \tag{14}$$

Then, we get the Hessian matrix

$$\nabla^2 f(x,y) = \begin{pmatrix} 1+6xy & 3(x^2-y^2) \\ 3(x^2-y^2) & 2-6xy \end{pmatrix}. \tag{15}$$

Therefore, this function is not separable. Since the vector $\boldsymbol{d}_k$ is in two dimensional space R^2, if $\boldsymbol{d}_k$ is an eigenvector of the matrix $A = \nabla^2 f(0,0)$, the next direction $\boldsymbol{d}_{k+1}$ also become an eigenvector. Therefore, every search direction $\boldsymbol{d}_k$ is always an eigenvector. From

$$\nabla f(x,y) = \begin{pmatrix} x + 3x^2 y - y^3 \\ 2y + x^3 - 3xy^2 \end{pmatrix}, \tag{16}$$

if we put an initial approximation $\boldsymbol{x}_0 = (1, -\frac{1}{3})$, this point satisfies $2y - 3xy^2 + x^3 = 0$. Therefore, every search direction $\boldsymbol{d}_k$ is an eigenvector of the matrix

$$\nabla^2 f(0,0) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{17}$$

Hence, it holds

$$\begin{aligned} \lim_{k\to\infty} V_k &= \lim_{k\to\infty} -(A_k\boldsymbol{d}_k, \boldsymbol{d}_{k+1}) \\ &= \lim_{k\to\infty} -(A\boldsymbol{d}_k, \boldsymbol{d}_{k+1}) = 0. \end{aligned} \tag{18}$$

We confirmed that the variance V_k converges to zero. Moreover, it holds $\boldsymbol{d}_{k+2} = -\boldsymbol{d}_k$. Consequently, the direction $\boldsymbol{d}_k$ takes one of four vectors

$$(-1,0),\ (0,1),\ (1,0),\ (0,-1)$$

cyclically and the sequence $\{\boldsymbol{x}_k\}$ converges to the origin spirally (See Figure 4). This case is never seen in quadratic form. However, the search direction does not always behave non zig-zag even if the variance converges to zero and the search direction accumulates eigenvectors. For example, put non-separable function

$$f(x,y) = \frac{1}{2}x^2 + \frac{3}{2}y^2 + x^2 y + xy^2 \tag{19}$$

and an initial vector $\boldsymbol{x}_0 = (-\frac{1}{2}, -\frac{1}{8})$. Although every search direction is an eigenvector of the matrix $\nabla^2 f(0,0)$, it alternates only two directions and the approximation $\boldsymbol{x}_k$ moves zig-zag (See Figure 5). In both examples, the approximation $\boldsymbol{x}_k$ moves between two curved lines. How two curved lines curve each other is important.

Finally, we give a theorem about the order of convergence.

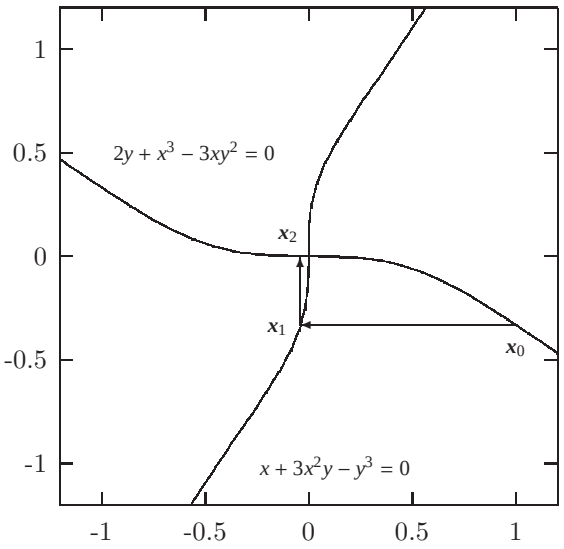

Figure 4: Search directions move four directions cyclically.

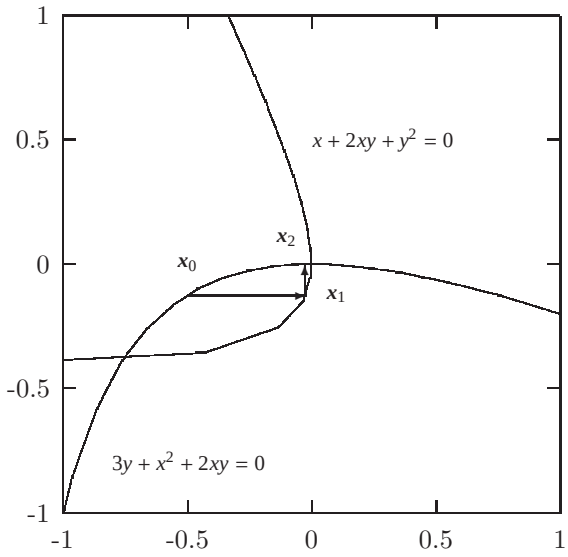

Figure 5: Search directions alternate two directions.

Theorem 4 *If the variable V_k converges to zero, the order of convergence is super linear.*

Usually, the steepest descent method converges lineally because the search direction take an eigenvector infrequently.

4. Conclusions

We discussed the limiting behavior of the search direction of the steepest descent method in minimizing nonlinear function. Unlike in the case of quadratic form, the Hessian matrix changes at each iteration. By the same way that Akaike treated quadratic form, we also analyzed the variance of the search direction corresponding to the Hessian matrix. If the variance converges to a positive number, the search direction asymptotically alternates between two vectors. This is the same behavior with quadratic form. However, we found the existence of a special case that the variance converges to zero and the direction takes one of four directions cyclically. This case never seen in quadratic form. However, this is a simple counter-example with two variable. Are there any counter-examples with more than three variable? This problem is not solved.

References

[1] G. E. Forsythe and T. S. Motzkin, "Asymptotic Properties of the Optimum Gradient Method," *Bull. AMS*, vol.57, p.183, 1951(Abstract).

[2] H. Akaike, "On a Successive Transformation of Probability Distribution and Its Application to the Analysis of the Optimum Gradient Method," *Ann. Inst. Stat. Math.*, vol.11, pp.1–16, 1959.

[3] J. Kowalik and M. R. Osborne, *Methods for Unconstrained Optimization Problems*, American Elsevier Publishing Company, New York, 1968.

[4] H. Konno and H. Yamashita, *Nonlinear Programming* (Japanese), Nikagiren, Tokyo, 1978.

[5] G. E. Forsythe, "On the Asymptotic Directions of the s-Dimensional Optimum Gradient Method," *Numer. Math.*, vol.11, pp.57–76, 1968.

[6] T. Ozeki and T. Iijima, "Behavior of the Steepest Descent Method in Minimizing Rayleigh Quotient," *IEICE Trans. Fundamentals*, vol. E80-A, no. 1, pp. 176-182, Jan. 1997.

[7] T. Ozeki, "Behavior of the Steepest Descent Method in Minimizing Nonlinear Function," *The Memoirs of the Faculty of Engineering Fukuyama University*, vol. 26, pp. 113-123, Dec. 2002.

[8] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, W. H. Freeman and Company, 1963.