On the Barzilai and Borwein Choice of Steplength for the Gradient Method

Marcos Raydan

May 1990
(revised October, 1991)

TR90-11

On the Barzilai and Borwein Choice of Steplength for the Gradient Method*

Marcos Raydan[†]

October 1991

Abstract

In a recent paper, Barzilai and Borwein presented a new choice of steplength for the gradient method. Their choice does not guarantee descent in the objective function and greatly speeds up the convergence of the method. We derive a relationship between the gradient method for minimizing a quadratic function and the shifted power method. This relationship allows us to establish the convergence of the Barzilai and Borwein method when applied to the problem of minimizing any strictly convex quadratic function (Barzilai and Borwein considered only 2-dimensional problems). Our point of view also allows us to explain the improvement obtained by using this new choice of steplength.

*During the course of this research the author was a graduate student in the Department of Mathematical Science at Rice University, Houston, Texas. He gratefully acknowledges Universidad Central de Venezuela for support during this period of time.

[†] Department of Mathematics, University of Kentucky, Lexington, Kentucky, 40506

1 Introduction

In order to solve the unconstrained minimization problem, we consider the nonlinear equations problem :

$$\text{find } x_\star \in I\!R^n \text{ such that } \nabla f(x_\star) = 0, \tag{1}$$

where $f : I\!R^n \rightarrow I\!R$. The numerical solution of (1) is usually iterative, moving at each iteration from an estimate x_c of $x_\star$ to a better estimate x_+. In many algorithms, each iteration involves the calculation of a quasi-Newton step, $s_{QN} = -A_c^{-1}\nabla f(x_c)$, where $A_c \in I\!R^{n\text{x}n}$ is an approximation of the Hessian of f at x_c. After each iteration the current A_c is updated to A_+, an approximation of the Hessian of f at x_+. The approximation usually is chosen to satisfy the secant equation,

$$A_+ s_c = y_c, \tag{2}$$

where $s_c = x_+ - x_c$ and $y_c = \nabla f(x_+) - \nabla f(x_c)$.

In the one dimensional case the secant equation completely determines A_+; however if $n > 1$, then many matrices will satisfy the secant equation. So, in addition to obeying (2), the update A_+ must be further restricted to a set of matrices that have desirable properties, See Dennis and Schnabel [3].

Barzilai and Borwein in [2] considered a related but somewhat different approach. They observed that the scalar $\alpha_+ \in I\!R$ that uniquely solves the overdetermined linear system $y_c = \alpha_+ s_c$ in the least squares sense is given by

$$\alpha_+ = \frac{s_c^t y_c}{s_c^t s_c} \tag{3}$$

if $s_c \neq 0$. Hence by restricting the update matrix in the quasi-Newton method to the class of scalar multiples of the identity and then asking that the secant equation be satisfied in the least squares sense they devised the following algorithm

Algorithm 1 (Barzilai and Borwein Method)

```
Given $x_0 \in I\!R^n$,$\alpha_0 \in I\!R$

For k=0,1,...,(until convergence) do

    1. Set $s_k = -\frac{1}{\alpha_k}\nabla f(x_k)$
```

```
  2. Set $x_{k+1} = x_k + s_k$
  3. Set $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$
  4. Set $\alpha_{k+1} = \frac{s_k^t y_k}{s_k^t s_k}$
End do
```

This algorithm requires only O(n) floating point operations per iteration in addition to the gradient evaluation.

If we consider problem (1) when $f(x) = \frac{1}{2}x^tAx - b^tx + c$ is a quadratic function and A is a symmetric positive definite (SPD) matrix, then α_+ in (3) becomes

$$\alpha_+ = \frac{s_c^t A s_c}{s_c^t s_c} \tag{4}$$

and Algorithm 1 becomes

Algorithm 2 (Barzilai and Borwein Method for Quadratics)

```
 Given $x_0 \in I\!R^n$,$\alpha_0 \in I\!R$
For k=0,1,...,(until convergence) do
  1. Set $s_k = -\frac{1}{\alpha_k}\nabla f(x_k)$
  2. Set $x_{k+1} = x_k + s_k$
  3. Set $\alpha_{k+1} = \frac{s_k^t A s_k}{s_k^t s_k}$
End do
```

In the quadratic case, α_{k+1} turns out to be the Rayleigh quotient of A at the vector s_k. Since A is SPD,

$$0 < \lambda_{min} \leq \alpha_k \leq \lambda_{max} \quad for \;\; all \;\; k, \tag{5}$$

where λ_{min} and λ_{max} are respectively the smallest and largest eigenvalues of A. And so, in step 1 there is no danger of dividing by zero.

Barzilai and Borwein [2] also observed, by symmetry, that the scalar $\hat{\alpha}_+$ that uniquely solves the overdetermined linear system $\frac{1}{\hat{\alpha}_+}y_c = s_c$ in the least squares sense is given by

$$\hat{\alpha}_+ = \frac{y_c^t y_c}{s_c^t y_c}$$

if $s_c^t y_c \neq 0$. In the quadratic case, $\hat{\alpha}_+$ becomes

$$\hat{\alpha_+} = \frac{s_c^t A^2 s_c}{s_c^t A s_c},$$

which is the Rayleigh quotient of A at the vector $\sqrt{A} s_c$. Hence, $\hat{\alpha}_+$ also satisfies (5).

In the rest of this paper, we will only consider the Barzilai and Borwein method with the choice of α_+ defined by (3) in the general case and by (4) in the quadratic case. The reason for this is that all results established for Algorithm 2 with the choice α_+ also hold with the choice $\hat{\alpha}_+$ since the method has the same properties in either case.

Notice that, in the Barzilai and Borwein gradient method, the search direction is always the negative gradient of f at x_c as in the gradient method, but the choice of steplength is not the standard choice. In fact, Algorithm 2 would be the steepest descent method for quadratics if we changed (4) to

$$\alpha_+ = \frac{g_+^t A g_+}{g_+^t g_+}, \tag{6}$$

where $g_+ = \nabla f(x_+)$. Since the vector s_+ is a multiple of the vector g_+, then (6) can also be written as $\alpha_+ = \frac{s_+^t A s_+}{s_+^t s_+}$.

Despite the similarities between these two methods, Algorithm 2 is significantly faster than the steepest descent method at the same cost per iteration, see Barzilai and Borwein [2] and also Fletcher [4].

Barzilai and Borwein [2] presented a convergence analysis of their method only in the two-dimensional quadratic case. It is unlikely that their analysis can be extended to higher dimensional problems. In the present work, we establish the convergence of the Barzilai and Borwein gradient method for any strictly convex quadratic function.

This paper is organized as follows. In Section 2 we present a relationship between the gradient method for minimizing a quadratic function and the shifted power method. We believe that this connection is the key to understanding the convergence properties of the Barzilai and Borwein method. In Section 3 we study the convergence of their method applied to a quadratic function with a SPD Hessian.

2 Relationship to the Shifted Power Method

Let us consider the gradient method for problem (1) when f is a differentiable function. For the purpose of comparison we will write the steplength choice in a slightly different way.

Algorithm 3 (Gradient Method)

```
Given $x_0 \epsilon I\!R^n$

For k=0,1,...,(until convergence) do
    1. Choose steplength $\frac{1}{\alpha_k}$
    2. Set $s_k = -\frac{1}{\alpha_k}\nabla f(x_k)$
    3. Set $x_{k+1} = x_k + s_k$
End do
```

Both the steepest descent method and the Barzilai and Borwein method for quadratics are special cases of Algorithm 3 . They differ only in the way the scalars α_k are chosen. Lemma 1 demonstrates a connection between Algorithm 3 and the shifted power method. This relationship will be used to establish our convergence results in Section 3.

Lemma 1 *Let $f(x) = \frac{1}{2}x^t Ax - b^t x + c$ where A is a SPD matrix. Further let $x_\star$ be the unique minimizer of f, $\{x_k\}$ the sequence generated by Algorithm 3 and $e_k = x_\star - x_k$ for all k. Then*

1. $Ae_k = \alpha_k s_k$
2. $e_{k+1} = \frac{1}{\alpha_k}(\alpha_k I - A)e_k$
3. $s_{k+1} = \frac{1}{\alpha_{k+1}}(\alpha_k I - A)s_k$

Proof: Using the fact that $\nabla f(x_k) = Ax_k - b$ and the definition of the step s_k in Algorithm 3, the three claims in Lemma 1 follow directly □

Since A is SPD, the scalars α_k satisfy (5) when they are generated by either (6) or by (4). And so, claim 1 in Lemma 1 allows us to conclude that $\|e_k\|$ tends to zero if and only if $\|s_k\|$ tends to zero. Thus, for the minimization of a quadratic function with a SPD Hessian it suffices to study the behavior of $\{s_k\}$.

For any s_0, there exist constants $c_1, c_2, ..., c_n$ such that:

$$s_0 = \sum_{i=1}^{n} c_i v_i, \tag{7}$$

where $\{v_1, v_2, ..., v_n\}$ are orthonormal eigenvectors of A associated with the eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_n\}$. Now, using Lemma 1 it follows that for any integer k,

$$s_{k+1} = \frac{1}{\gamma_k} \sum_{i=1}^{n} (\prod_{j=0}^{k} (\alpha_j - \lambda_i)) c_i v_i \tag{8}$$

where

$$\gamma_k = \prod_{j=1}^{k+1} \alpha_j.$$

From (8) we can see that if we use the exact eigenvalues of A as the scalars α_k in Algorithm 3, in any order, then we find the exact solution in p iterations, where p is the number of distinct eigenvalues of A. Unfortunately, we do not know the eigenvalues of A in advance. However, we can choose α_k to be the Rayleigh quotient at s_k (the choice of the steepest descent method) or the Rayleigh quotient at s_{k-1} (the choice of the Barzilai and Borwein method) to approximate the eigenvalues of A. With the choice of the steepest descent method, it can be shown that $s_{k+1}^t s_k = 0$ for all k, and also that the 2-norm of the error decreases at every iteration. Consequently, the scalars α_k tend to take values closer to λ_{max} than to λ_{min}. Therefore, the coefficients associated with the large eigenvalues in (8) will be effectively reduced, while the coefficients associated with the small eigenvalues only decrease very slowly. See Johnson [5] and Akaike [1] for details.

On the other hand, the choice of the Barzilai and Borwein method does not guarantee any orthogonality among the vectors s_k, and the 2-norm of the error might increase at some iterations. In fact, by properties of the Rayleigh quotient, the scalar α_k will approximate the eigenvalue associated with the largest coefficient in the eigenvector expansion of the vector s_{k-1}, regardless of the eigenvalue location in the spectrum. Therefore, the Barzilai and Borwein method is more effective at reducing the coefficients associated with all the eigenvalues in the eigenvector expansion (8).

The behavior of the sequence $\{\alpha_k\}$ for the Barzilai and Borwein method, in particular the approximation of the eigenvalues of the Hessian, is most easily appreciated by considering a particular example.

Example 1

Let $f(x) = \frac{1}{2}x^tAx$ where $A = diag(1,2,12)$. Clearly, f has a unique minimizer at $x_* = (0,0,0)^t$. The first 10 iterations generated by Algorithm 2 starting at $x_0 = (1,1,1)^t$ and $\alpha_0 = 1$ are shown in Table 1. The Table lists the 2-norm of the error, and the 2-norm of the gradient. Also shown are the scalars α_k and the coefficients in the eigenvector expansion (8) associated with the three eigenvalues of A.

Since $\alpha_0 = \lambda_1 = 1$, the column with the coefficients of the eigenvector associated with λ_1 in Table 1 contains zero after the first iteration, and the scalar α_k approximates only the eigenvalues $\lambda_2 = 2$ and $\lambda_3 = 12$ during the rest of the process. Notice that α_k approximates, at each iteration, the eigenvalue with the larger coefficient in the previous iteration. In fact, the bigger the difference between the two coefficients the closer the scalar will be to the eigenvalue.

Notice also that this is not a descent algorithm. Under special circumstances, that will be studied in the next section, the 2-norm of the error e_k, as well as the objective function, increases at some iterations. This is in sharp contrast to the steepest descent method. In fact, if the steepest descent method is used to minimize the function $f(x)$ with the same initial guess x_0, then 165 iterations are required to achieve an error of .3x10^{-29}. Observe that the Barzilai and Borwein method achieved this accuracy in only 10 iterations.

3 Convergence Analysis

In this section we establish the convergence of Algorithm 2 applied to any quadratic function with a SPD Hessian.

For any initial error e_0, there exist constants $d_1^0, d_2^0, ..., d_n^0$ such that:

$$e_0 = \sum_{i=1}^{n} d_i^0 v_i,$$

iteration	$\|\epsilon_k\|_2$	$\|\nabla f(x_k)\|_2$	α_k	coefficients of eigenvectors associated with		
				$\lambda_1 = 1$	$\lambda_2 = 2$	$\lambda_3 = 12$
0	0.17d+01	0.12d+02	0.1000d+01	-.10d+01	-.20d+01	-.12d+02
1	0.11d+02	0.13d+03	0.1165d+02	0.00d+00	0.17d+00	0.11d+02
2	0.88d+00	0.42d+01	0.1199d+02	0.00d+00	0.14d+00	-.32d+00
3	0.69d+00	0.13d+01	0.1045d+02	0.00d+00	0.13d+00	0.71d-04
4	0.55d+00	0.11d+01	0.2000d+01	0.00d+00	0.56d+00	-.55d-04
5	0.45d-04	0.54d-03	0.2000d+01	0.00d+00	0.80d-06	0.27d-03
6	0.22d-03	0.27d-02	0.1199d+02	0.00d+00	0.65d-14	-.23d-03
7	0.16d-08	0.19d-07	0.1200d+02	0.00d+00	0.54d-14	0.16d-08
8	0.26d-13	0.53d-13	0.1200d+02	0.00d+00	0.45d-14	0.00d+00
9	0.22d-13	0.44d-13	0.2000d+01	0.00d+00	0.22d-13	0.00d+00
10	0.31d-29	0.63d-29	0.2000d+01	0.00d+00	-.32d-29	0.00d+00

Table 1: Barzilai and Borwein method for Example 1

where $\{v_1, v_2, ..., v_n\}$ are orthonormal eigenvectors of A associated with the eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_n\}$.

Using Lemma 1 we obtain for any integer k,

$$\epsilon_{k+1} = \sum_{i=1}^{n} d_i^{k+1} v_i, \tag{9}$$

where

$$d_i^{k+1} = \prod_{j=0}^{k} (\frac{\alpha_j - \lambda_i}{\alpha_j}) d_i^0.$$

We observe that the convergence properties of the sequence $\{e_k\}$ will depend on the behavior of each one of the sequences $\{d_i^k\}$, $1 \leq i \leq n$. Later in this section, we will prove that each of these sequences converges to zero. First let us establish the Q-linear convergence of Algorithm 2 applied to a quadratic function with a SPD Hessian that satisfies the admittedly restrictive condition

$$\lambda_{max} < 2 * \lambda_{min}. \tag{10}$$

Lemma 2 *Let $f(x) = \frac{1}{2}x^tAx - b^tx + c$ where A is SPD and satisfies (10). Let $\{x_k\}$ be the sequence generated by Algorithm 2 and $x_\star$ the unique minimizer of f. Then the sequence $\{x_k\}$ converges Q-linearly to $x_\star$ in the Euclidean norm with convergence factor $\hat{c} = (\lambda_{max} - \lambda_{min})/\lambda_{min}$.*

Proof: Using Lemma 1 and (9) we obtain for any k,

$$e_{k+1} = \frac{1}{\alpha_k}(\alpha_k I - A)\sum_{i=1}^{n} d_i^k v_i = \sum_{i=1}^{n}(\frac{\alpha_k - \lambda_i}{\alpha_k})d_i^k v_i.$$

By the orthonormality of the eigenvectors we have

$$\|e_{k+1}\|_2^2 = \sum_{i=1}^{n}(d_i^k)^2(\frac{\alpha_k - \lambda_i}{\alpha_k})^2 \leq \max_i(\frac{\alpha_k - \lambda_i}{\alpha_k})^2\|e_k\|_2^2. \qquad (11)$$

From (10), recalling that α_k obeys (5),

$$\max_i |\frac{\alpha_k - \lambda_i}{\alpha_k}| \leq \frac{\lambda_{max} - \lambda_{min}}{\lambda_{min}} < 1. \qquad (12)$$

Combining (11) and (12) gives

$$\|e_{k+1}\|_2 \leq \hat{c}\|e_k\|_2 \quad \textit{where} \quad \hat{c} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{min}} < 1 \ \Box$$

Now we can explain why the norm of the error might increase at some iterations when the spectrum of A does not satisfy (10). Let us first divide the spectrum of A into two subintervals

$$Left = [\lambda_{min}, \frac{\lambda_{max}}{2}] \quad \textit{and} \quad Right = (\frac{\lambda_{max}}{2}, \lambda_{max}].$$

Clearly, if the spectrum of A obeys (10) then the Left interval is empty and Lemma 2 says that the error decreases at each iteration. If we force the scalars α_k to be in the Right interval, by a similar argument, $\{x_k\}$ converges Q-linearly to $x_\star$ for any $x_0 \in I\!R^n$. But Algorithm 2 moves the scalars α_k dynamically within the spectrum of A. If at the j^{th} iteration $\alpha_j \in Left$, then the coefficient associated with the eigenvalues λ_i that satisfy $\lambda_i > 2\alpha_j$, will be amplified by the factor $|\frac{\alpha_j - \lambda_i}{\alpha_j}| > 1$ (i.e., $|d_i^{j+1}| > |d_i^j|$), and this could lead to an increase of $\|e_{j+1}\|$ with respect to $\|e_j\|$. In general, the sequences $\{d_i^k\}$ defined in (9) will increase at some iterations. However, the sequence $\{d_1^k\}$ associated with the eigenvalue λ_{min} will decrease at every iteration.

Lemma 3 *The sequence $\{d_1^k\}$ converges to zero Q-linearly with convergence factor $\hat{c} = 1 - (\lambda_{min}/\lambda_{max})$.*

Proof: For any positive integer k,

$$d_1^{k+1} = (\frac{\alpha_k - \lambda_{min}}{\alpha_k})d_1^k.$$

Since α_k satisfies (5), we have

$$0 < \frac{\lambda_{min}}{\lambda_{max}} \leq \frac{\lambda_{min}}{\alpha_k} \leq 1.$$

And so,

$$|d_1^{k+1}| = (1 - \frac{\lambda_{min}}{\alpha_k})|d_1^k| \leq \hat{c}|d_1^k|,$$

where

$$\hat{c} = 1 - \frac{\lambda_{min}}{\lambda_{max}} < 1 \ \square$$

In the proof of our convergence theorem, we will use the following result.

Lemma 4 *If the sequences $\{d_1^k\}, \{d_2^k\}, ..., \{d_l^k\}$ all converge to zero for a fixed integer l, $1 \leq l < n$. Then,*

$$\liminf_{k\to\infty} |d_{l+1}^k| = 0 \ .$$

Proof: Suppose, by way of contradiction, that there exists a constant $\epsilon > 0$ such that

$$(d_{l+1}^k)^2\lambda_{l+1}^2 > \epsilon \text{ for all } k \ .$$

By (9), Lemma 1 and the orthonormality of the eigenvectors $\{v_1, v_2, ..., v_n\}$, we can see that the Rayleigh quotient α_{k+1} can be written as

$$\alpha_{k+1} = \frac{\sum_{i=1}^n (d_i^k)^2\lambda_i^3}{\sum_{i=1}^n (d_i^k)^2\lambda_i^2} \ . \tag{13}$$

Since the sequences $\{d_1^k\}, ..., \{d_l^k\}$ all converge to zero, there exists $\hat{k}$ sufficiently large such that

$$\sum_{i=1}^{l}(d_i^k)^2\lambda_i^2 < \frac{\epsilon}{2} \text{ for all } k \geq \hat{k} \ . \tag{14}$$

By (13) and (14), we obtain

$$\frac{(\sum_{i=l+1}^{n}(d_i^k)^2\lambda_i^2)\lambda_{l+1}}{\frac{\epsilon}{2}+(\sum_{i=l+1}^{n}(d_i^k)^2\lambda_i^2)} \leq \alpha_{k+1} \leq \lambda_{max} \ . \tag{15}$$

Since

$$\sum_{i=l+1}^{n}(d_i^k)^2\lambda_i^2 \geq (d_{l+1}^k)^2\lambda_{l+1}^2 > \epsilon \ ,$$

then, by using (15), it follows that

$$\frac{2}{3}\lambda_{l+1} \leq \alpha_{k+1} \leq \lambda_{max} \text{ for all } k \geq \hat{k} \ .$$

Finally, using the fact that $d_{l+1}^{k+1} = ((\alpha_k - \lambda_{l+1})/\alpha_k)d_{l+1}^k$, we obtain for all $k \geq \hat{k}+1$,

$$|d_{l+1}^{k+1}| = |1 - \frac{\lambda_{l+1}}{\alpha_k}||d_{l+1}^k| \leq \hat{c}|d_{l+1}^k| \ ,$$

where

$$\hat{c} = \max(\frac{1}{2}, 1 - \frac{\lambda_{l+1}}{\lambda_{max}}) < 1 \ ,$$

which is a contradiction. Therefore, $\liminf_{k\to\infty} |d_{l+1}^k| = 0$ □

Theorem 1 establishes the convergence of the Barzilai and Borwein method when applied to a quadratic function with a SPD Hessian.

Theorem 1 *Let $f(x)$ be a strictly convex quadratic function. Let $\{x_k\}$ be the sequence generated by Algorithm 2 and $x_\star$ the unique minimizer of f. Then, either $x_j = x_\star$ for some finite j, or the sequence $\{x_k\}$ converges to $x_\star$.*

Proof: We need only consider the case in which there is no finite integer j such that $x_j = x_\star$. Hence, it suffices to prove that the sequence $\{e_k\}$ converges to zero.

From (9) and the orthonormality of the eigenvectors we have

$$\|e_k\|_2^2 = \sum_{i=1}^{n}(d_i^k)^2.$$

And so, the sequence of errors $\{e_k\}$ converges to zero if and only if each one of the sequences $\{d_i^k\}$ for $i = 1, 2, ..., n$ converges to zero.

Suppose, by way of contradiction, that some of the sequences $\{d_i^k\}$ are not converging to zero. In particular, let us suppose that p is the smallest integer between 1 and n for which the sequence $\{d_p^k\}$ does not converge to zero. By Lemma 3, we can see that $p \geq 2$.

Since $\{d_1^k\}, ...\{d_{p-1}^k\}$ all converge to zero, then for a given $\epsilon > 0$ there exists $\hat{k}$ sufficiently large such that

$$\sum_{i=1}^{p-1} (d_i^k)^2 \lambda_i^2 < \frac{\epsilon}{2} \text{ for all } k \geq \hat{k} \ . \tag{16}$$

By Lemma 4, it follows that $\liminf_{k \to \infty} |d_p^k| = 0$. Hence, there exists $k_p \geq \hat{k}$ such that

$$(d_p^{k_p})^2 \lambda_p^2 < \epsilon \ .$$

Now, to study the behavior of the sequence $\{d_p^k\}$ for $k \geq k_p$, we define

$$M_\epsilon = \sup_{i \geq k_p} \{(d_p^i)^2\} \ .$$

Let us say that $k_0 > k_p$ is any positive integer for which $(d_p^{k_0 - 1})^2 \lambda_p^2 < \epsilon$ and $(d_p^{k_0})^2 \lambda_p^2 > \epsilon$. From (13) and (16), it follows that

$$\frac{2}{3} \lambda_p \leq \alpha_k \leq \lambda_{max} \ , \tag{17}$$

for all integer k that satisfies $k_0 + 1 \leq k \leq j$, where j is the first positive integer greater than k_0 for which $(d_p^j)^2 \lambda_p^2 < \epsilon$.

Finally, using the bound

$$|d_p^{k_0+1}| \leq (\frac{\lambda_{max} - \lambda_{min}}{\lambda_{min}})^2 |d_p^{k_0-1}| \ ,$$

and the fact that

$$|d_p^{k+1}| \leq \max(\frac{1}{2}, 1 - \frac{\lambda_p}{\lambda_{max}}) |d_p^k| \ ,$$

whenever α_k satisfies (17), we conclude that

$$M_\epsilon \leq (\frac{\lambda_{max} - \lambda_{min}}{\lambda_{min}})^4 \frac{\epsilon}{\lambda_p^2} \ .$$

Since $\epsilon > 0$ can be chosen arbitrary small, then $\limsup_{k\to\infty} (d_p^k)^2 = 0$. Hence, $\lim_{k\to\infty} |d_p^k| = 0$, which is a contradiction. Therefore, the sequence $\{e_k\}$ converges to zero $\square$

Notice that with the choice of $\hat{\alpha}_{k+1} = \frac{s_k^t A^2 s_k}{s_k^t A s_k}$ instead of $\alpha_{k+1} = \frac{s_k^t A s_k}{s_k^t s_k}$, equality (13) can be written as

$$\hat{\alpha}_{k+1} = \frac{\sum_{i=1}^n (d_i^k)^2 \lambda_i^4}{\sum_{i=1}^n (d_i^k)^2 \lambda_i^3}.$$

Then, by a similar argument, we conclude that the convergence result established in Theorem 1 for Algorithm 2 with the choice of α_{k+1} also holds with the choice of $\hat{\alpha}_{k+1}$.

Acknowledgements

The author wishes to thank Richard Byrd, John E. Dennis and Richard A. Tapia for helpful comments and suggestions.

References

[1] H. AKAIKE. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Statist. Math. Tokyo*, 11:1–17, 1959.

[2] J. BARZILAI and J.M. BORWEIN. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

[3] J.E. DENNIS Jr. and R.B. SCHNABEL. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[4] R. FLETCHER. Low storage methods for unconstrained optimization. *Lectures in Applied Mathematics (AMS)*, 26:165–179, 1990.

[5] CLAES JOHNSON. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, 1990.