JOURNAL OF APPROXIMATION THEORY 45, 219–231 (1985)

On a Class of Algorithms for Total Approximation

G. A. WATSON

Department of Mathematical Sciences, University of Dundee, Dundee DD1 4HN, Scotland

Communicated by E. W. Cheney

Received May 30, 1984

1. INTRODUCTION

The standard formulation of many linear data fitting problems is as follows: given an $m \times n$ matrix A, with $m > n$, and $\mathbf{b} \in R^m$, find $\mathbf{x} \in R^n$ to minimize $\|\mathbf{r}\|$ where

$$\mathbf{r} = A\mathbf{x} - \mathbf{b}, \tag{1.1}$$

and the norm is a given norm on R^m. It is usually assumed that the expected values of the components of $\mathbf{r}$ are zero, and the appropriate norm to be used depends on the distribution of the errors represented by these components. An underlying assumption, therefore, is that errors are only present in the vector $\mathbf{b}$ (corresponding to dependent variable values). However, it is often the case that the elements of A are also unreliable, for example, if the independent variable values, too, are inexact. One way to take account of this more general errors-in-variables situation is to introduce perturbations into the elements of A and to solve the following *total* approximation problem

$$\text{find } \mathbf{x} \in R^n \text{ to minimize } \|E \vdots \mathbf{r}\| \tag{1.2}$$

where

$$\mathbf{r} = (A + E)\,\mathbf{x} - \mathbf{b},$$

and the norm is now an appropriate matrix norm. If some of the columns of A are known to be error free, then an additional constraint is that the corresponding columns of E are zero. Without loss of generality, it will be assumed that this is true of the first l columns.

If the matrix norm in (1.2) is the l_p norm defined by taking the usual *vector* l_p norm on the elements of the matrix regarded as an extended vec-

0021-9045/85 $3.00

tor in $R^{m \times (n+1)}$, then (1.2) becomes the total l_p approximation problem. An effective way in which this problem may be tackled is to exploit its connection with the following constrained vector norm approximation problem. Let $Z: R^{n+1} \to R^m$ be defined by

$$Z = [A \vdots \mathbf{b}].$$

Then the problem referred to above is

$$\text{find } \mathbf{v} \in R^{n+1} \text{ to minimize } \|Z\mathbf{v}\|_p \text{ subject to } \|\mathbf{v}_2\|_q = 1, \tag{1.3}$$

where the subscripts on the norms indicate the usual l_p and l_q vector norms, where p and q are *dual* in the sense that

$$\frac{1}{p} + \frac{1}{q} = 1, \tag{1.4}$$

and where the vector $\mathbf{v}_2 \in R^{n+1-l}$ is obtained from $\mathbf{v}$ by deleting the first l components. Both (1.2) and (1.3) are non-convex problems (the feasible region in (1.3) is the *outside* of the unit ball), and so it may only be possible to find points satisfying first-order necessary conditions for local solutions (stationary points). The precise relationship between (1.2) and (1.3) is explored in [6]; a basic result is the following.

THEOREM 1 [6]. *Let* $\mathbf{v} \in R^{n+1}$ *be a stationary point of* (1.3) *with* $v_{n+1} \neq 0$. *Then* $\mathbf{x}$ *defined by*

$$\mathbf{v} = \tau \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} \tag{1.5}$$

is a stationary point of the total l_p *approximation problem* (1.2).

The problem (1.3) is in fact meaningful in the errors-in-variables context when p and q are not connected by the relationship (1.4) (or indeed when the norms are replaced by arbitrary norms on R^m and R^{n+1-l}, respectively). For example, the so-called *orthogonal* l_p approximation problem corresponds to the choice $q = 2$ in (1.3) (see [3, 4]). It is shown in [5] that $\mathbf{v}$ solving (1.3) with arbitrary norms $\|\cdot\|_A$ on R^m and $|\cdot\|_B$ on R^{n+1-l} corresponds through (1.5) to a solution of the total approximation problem (1.2) with matrix norm defined on the $m \times (n+1)$ matrix M (with first l columns zero) by

$$\|M\| = \max_{|\mathbf{d}_2|_B = 1} \|M\mathbf{d}\|_A,$$

where the subscript 2 has (and will continue to have) the same connotation as before.

This paper is concerned with the solution of (1.3) for the cases $1 < p$, $q < \infty$ by a class of algorithms essentially proposed by Späth [3] for the orthogonal l_p problem. Numerical experiments reported by Späth suggest that these algorithms possess certain global convergence properties, and this view is reinforced by similar experiments (for the case of the total l_p approximation problem) reported in [6]. It is the intention here to establish a theoretical basis for these properties. In the next section, the class of algorithms is described, and in Section 3 local convergence results are obtained, which generalize those given in [4, 6]. Finally, in Section 4, some global convergence results are proved.

The situation when $\| Z\mathbf{v} \|_p = 0$ at a solution (corresponding to $\mathbf{r} = \mathbf{0}$ in (1.1)) is of little interest. Therefore it is assumed in what follows that Z has full rank.

2. The Class of Algorithms

For all values of p, q in the range $1 < p, q < \infty$, (1.3) may be written as (find $\mathbf{v} \in R^{n+1}$ to)

$$\text{minimize } \| Z\mathbf{v} \|_p^p \text{ subject to } \| \mathbf{v}_2 \|_q^q = 1, \tag{2.1}$$

which has differentiable objective and constraint functions. From an algorithmic point of view, it is important to be able to define the diagonal matrices

$$D(\mathbf{v}) = \text{diag}\{ |(Z\mathbf{v})_i|^{p-2}, i = 1, 2, \ldots, m \},$$
$$C(\mathbf{v}) = \text{diag}\{ 0, 0, \ldots, 0, |v_{l+1}|^{q-2}, \ldots, |v_{n+1}|^{q-2} \},$$

which will be assumed to exist for all $\mathbf{v}$ of interest. When $p = 1$, the solution to (2.1) is characterized by certain zero components of $Z\mathbf{v}$ (see [2]) so some elements of $D(\mathbf{v})$ will become increasingly large as p tends to 1; however, provided that p is not too close to 1, it is normally possible to work with $D(\mathbf{v})$ except for pathological cases. The problem (2.1) may then be written

$$\text{minimize } \mathbf{v}^{\mathrm{T}} J(\mathbf{v})\, \mathbf{v} \text{ subject to } \mathbf{v}^{\mathrm{T}} C(\mathbf{v})\, \mathbf{v} = 1 \tag{2.2}$$

where

$$J(\mathbf{v}) = Z^{\mathrm{T}} D(\mathbf{v})\, Z.$$

The Kuhn Tucker first-order necessary conditions for $\mathbf{v}^*$ to solve (2.2) correspond to the existence of a scalar μ^* such that

$$pJ(\mathbf{v}^*)\,\mathbf{v}^* - \mu^* q C(\mathbf{v}^*)\,\mathbf{v}^* = \mathbf{0}. \tag{2.3}$$

Thus $\mathbf{v}^*$ is an eigenvector of the generalized eigenvalue problem

$$J(\mathbf{v}^*)\,\mathbf{v} = \gamma C(\mathbf{v}^*)\,\mathbf{v}, \tag{2.4}$$

with eigenvalue $\gamma = \mathbf{v}^{*\mathrm{T}} J(\mathbf{v}^*)\,\mathbf{v}^*$, which is positive by assumption. In particular, if $p = q = 2$ (corresponding to the total least-squares problem) $\mathbf{v}^*$ is an eigenvector corresponding to the smallest eigenvalue of $Z^{\mathrm{T}}Z$. An analysis of this problem, and a method of solution based on the singular value decomposition of Z, is given in [1].

If $q \neq 2$, and $C(\mathbf{v}^*)$ contains some zero elements (which it must do if $l \neq 0$), then (2.4) is deficient in the sense that not all the eigenvalues are finite. However, if $D(\mathbf{v}^*)$ exists and is positive definite, then the eigenvalue problem

$$C(\mathbf{v}^*)\,\mathbf{v} = \lambda J(\mathbf{v}^*)\,\mathbf{v} \tag{2.5}$$

has a full set of real non-negative eigenvalues (at least l of which are zero) and the eigenvector $\mathbf{v}^*$ now corresponds to the eigenvalue λ^*, say, where

$$\lambda^* = 1/\mathbf{v}^{*\mathrm{T}} J(\mathbf{v}^*)\,\mathbf{v}^* = 1/\|Z\mathbf{v}^*\|_p^p.$$

The natural generalization of the basic method suggested by Späth [3] for the case $q = 2$, $l = 0$ has at the i^{th} iteration an approximation $\mathbf{v}^{(i)}$ to the solution of (2.1), with $\|\mathbf{v}_2^{(i)}\|_q = 1$, and defines $\mathbf{v}^{(i+1)}$ as the eigenvector (correctly normalized and assumed to be unique) corresponding to the *largest* eigenvalue of the generalized eigenproblem

$$C(\mathbf{v}^{(i)})\,\mathbf{v} = \lambda J(\mathbf{v}^{(i)})\,\mathbf{v}. \tag{2.6}$$

This may be obtained by the application of the power method with initial approximation $\mathbf{v}^{(i)}$. Clearly, the correct normalization of $\mathbf{v}^{(i+1)}$ is always possible if $J(\mathbf{v}^{(i)})$ is non-singular. If k steps of the power method are applied, then the inner iteration has the form

$$J(\mathbf{v}^{(i)})\,\mathbf{d}^{(i_j)} = C(\mathbf{v}^{(i)})\,\mathbf{d}^{(i_{j-1})}, \qquad j = 1, 2, \ldots, k, \tag{2.7}$$

where

$$\mathbf{d}^{(i_0)} = \mathbf{v}^{(i)},$$

and

$$\mathbf{v}^{(i+1)} = \mathbf{d}^{(i,k)} / \| \mathbf{d}_2^{(i,k)} \|_q.$$

In practice, (2.7) is solved for $\mathbf{d}^{(i,k)}$ by forming the QR factors of $D^{1/2}(\mathbf{v}^{(i)})\, Z$ and using forward and backward substitution with the matrix R. If k steps of the power method are applied at every value of i, then the outer iteration process for the algorithm may be written

$$\mathbf{d}^{(i)} = Q(\mathbf{v}^{(i)})^k\, \mathbf{v}^{(i)}, \tag{2.8}$$

$$\mathbf{v}^{(i+1)} = \mathbf{d}^{(i)} / \| \mathbf{d}_2^{(i)} \|_q, \qquad i = 0, 1, 2, \ldots,$$

where

$$Q(\mathbf{v}) = J(\mathbf{v})^{-1} C(\mathbf{v})$$

and is assumed to exist for all i. The initial approximation is arbitrary except that $\| \mathbf{v}_2^{(0)} \|_q = 1$. In his numerical experiments, Späth [3] observed (when $l = 0$, $q = 2$) that this algorithm converged for all values of p in the range $1 < p < p'$, where $p' \approx 2.7$ (depending on Z). Further, the convergence was independent of the value of k used (for example, taking $k = 1$ was satisfactory). The rest of this paper is concerned with an analysis of (2.8), which in particular goes some way towards explaining this phenomenon.

3. Local Convergence Analysis

Let $\mathbf{v}^*$ be a fixed point of the iteration (2.8), so that $\mathbf{v}^*$ is an eigenvector of the generalized eigenproblem

$$Q(\mathbf{v}^*)\, \mathbf{v} = \lambda \mathbf{v} \tag{3.1}$$

corresponding to the eigenvalue $\lambda^* = 1/\mathbf{v}^{*\mathrm{T}} J(\mathbf{v}^*)\, \mathbf{v}^*$, and therefore also a stationary point of (2.1). Let the remaining n eigenvalues of (3.1) be $\lambda_1, \lambda_2, \ldots, \lambda_n$ with corresponding eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ (clearly (3.1) is non-defective) normalized so that

$$\mathbf{z}_i^{\mathrm{T}} J(\mathbf{v}^*)\, \mathbf{z}_j = \delta_{ij}, \qquad i, j = 1, 2, \ldots, n \tag{3.2}$$

$$\mathbf{z}_i^{\mathrm{T}} J(\mathbf{v}^*)\, \mathbf{v}^* = 0, \qquad i = 1, 2, \ldots, n. \tag{3.3}$$

Let

$$w_i = \lambda_i / \lambda^*, \qquad i = 1, 2, \ldots, n. \tag{3.4}$$

Then second-order conditions for $\mathbf{v}^*$ to solve (2.1) may be used to obtain the following results.

THEOREM 2 [6]. (i) *If* $\mathbf{v}^*$ *solves* (2.1), *then*

$$w_i \leqslant \frac{p-1}{q-1}, \qquad i = 1, 2, \ldots, n. \tag{3.5}$$

(ii) *If* (3.5) *holds with strict inequality for each* i, $i = 1, 2, \ldots, n$, *then* $\mathbf{v}^*$ *is a local minimum of* (2.1).

Before proving the main result of this section, a piece of notation and a preliminary lemma are required. For any $\mathbf{w} \in R^{n+1}$, any $(n+1) \times (n+1)$ matrix $T(\mathbf{v})$ whose elements are continuously differentiable functions of $\mathbf{v} \in R^{n+1}$, $\nabla(T(\mathbf{v}))\,\mathbf{w}$ will denote the $(n+1) \times (n+1)$ matrix which satisfies

$$T(\mathbf{v}+\mathbf{s})\,\mathbf{w} = T(\mathbf{v})\,\mathbf{w} + [\nabla(T(\mathbf{v}))\,\mathbf{w}]\,\mathbf{s} + O(\|\mathbf{s}\|^2)$$

for any $\mathbf{s} \in R^{n+1}$.

LEMMA 1. *Let* $G(\mathbf{v}) = \nabla(Q(\mathbf{v})^k)\,\mathbf{v}$. *Then*

$$G(\mathbf{v}^*) = (q-2) \sum_{j=1}^{k} (\lambda^*)^{j-1} Q(\mathbf{v}^*)^{k-j+1} - (p-2) \sum_{j=1}^{k} (\lambda^*)^{j} Q(\mathbf{v}^*)^{k-j}.$$

Proof.

$$\begin{aligned} G(\mathbf{v}) &= \nabla(Q(\mathbf{v})^k)\,\mathbf{v} \\ &= \sum_{j=1}^{k} Q(\mathbf{v})^{k-j} \nabla(Q(\mathbf{v}))\, Q(\mathbf{v})^{j-1}\,\mathbf{v}, \end{aligned}$$

so that

$$G(\mathbf{v}^*) = \sum_{j=1}^{k} (\lambda^*)^{j-1} Q(\mathbf{v}^*)^{k-j} \nabla(Q(\mathbf{v}^*))\,\mathbf{v}^*. \tag{3.6}$$

Also

$$\nabla(C(\mathbf{v})\,\mathbf{v}) = (q-1)\,C(\mathbf{v}),$$

and so

$$\begin{aligned} (q-1)\,C(\mathbf{v}) &= \nabla(J(\mathbf{v})\,J(\mathbf{v})^{-1} C(\mathbf{v})\,\mathbf{v}) \\ &= \nabla(J(\mathbf{v})\,Q(\mathbf{v})\,\mathbf{v}) \\ &= \nabla(J(\mathbf{v}))\,Q(\mathbf{v})\,\mathbf{v} + J(\mathbf{v})\,\nabla(Q(\mathbf{v})\,\mathbf{v}). \end{aligned} \tag{3.7}$$

Now if $\mathbf{u}$ is a constant vector in R^{n+1},

$$\begin{aligned}\nabla(J(\mathbf{v}))\,\mathbf{u} &= \nabla(J(\mathbf{v})\,\mathbf{u}) \\ &= (p-2)\,Z^{\mathrm{T}}\,WZ,\end{aligned}$$

where

$$W = \operatorname{diag}\{(Z\mathbf{u})_i\,(Z\mathbf{v})_i\,|(Z\mathbf{v})_i|^{p-4},\qquad i=1,2,\ldots,m\},$$

so that

$$\nabla(J(\mathbf{v}^*))\,Q(\mathbf{v}^*)\,\mathbf{v}^* = (p-2)\lambda^*\,J(\mathbf{v}^*).$$

Thus, from (3.7),

$$(q-1)\,C(\mathbf{v}^*) = (p-2)\lambda^*\,J(\mathbf{v}^*) + J(\mathbf{v}^*)\,(Q(\mathbf{v}^*) + \nabla(Q(\mathbf{v}^*))\,\mathbf{v}^*),$$

and so

$$\begin{aligned}\nabla(Q(\mathbf{v}^*))\,\mathbf{v}^* &= J(\mathbf{v}^*)^{-1}\,((q-1)\,C(\mathbf{v}^*) - (p-2)\lambda^*\,J(\mathbf{v}^*)) - Q(\mathbf{v}^*) \\ &= (q-2)\,Q(\mathbf{v}^*) - (p-2)\lambda^* I.\end{aligned}$$

The result now follows from (3.6). ∎

THEOREM 3. *Sufficient conditions for* (2.8) *to converge locally to* $\mathbf{v}^*$ *are that*

$$1 + (q-1)\,w_i < p < 2 + (q-2)\,w_i + (1-w_i)\,(1+w_i^k)/(1-w_i^k),\quad i=1,2,\ldots,n. \tag{3.8}$$

Proof. Let

$$\begin{aligned}\mathbf{v} &= \mathbf{v}^* + \sum_{i=1}^{n} \theta_i \mathbf{z}_i \\ &= \mathbf{v}^* + \boldsymbol{\varepsilon}, \qquad \text{say,}\end{aligned}$$

where it is assumed that the numbers θ_i, $i=1,2,\ldots,n$, are small. Then

$$\begin{aligned}\|\mathbf{v}_2\|_q^q &= \|\mathbf{v}_2^* + \boldsymbol{\varepsilon}_2\|_q^q \\ &= \|\mathbf{v}_2^*\|_q^q + q\boldsymbol{\varepsilon}^{\mathrm{T}} C(\mathbf{v}^*)\,\mathbf{v}^* + O(\|\boldsymbol{\varepsilon}\|^2) \\ &= 1 + q\sum_{i=1}^{n} \theta_i \mathbf{z}_i^{\mathrm{T}}\,C(\mathbf{v}^*)\,\mathbf{v}^* + O(\|\boldsymbol{\varepsilon}\|^2) \\ &= 1 + O(\|\boldsymbol{\varepsilon}\|^2)\end{aligned}$$

using (3.3). Thus $\mathbf{v}$ is correctly normalized to first order in $\boldsymbol{\varepsilon}$. Now define $\boldsymbol{\delta}$ by

$$\alpha(\mathbf{v}^* + \boldsymbol{\delta}) = Q(\mathbf{v})^k(\mathbf{v}^* + \boldsymbol{\varepsilon}), \tag{3.9}$$

where α is the normalization constant such that

$$\boldsymbol{\delta} = \sum_{i=1}^{n} \phi_i \mathbf{z}_i.$$

As before, $\mathbf{v}^* + \boldsymbol{\delta}$ is correctly normalized to first order in $\boldsymbol{\delta}$. In addition $\|\boldsymbol{\delta}\| \sim \|\boldsymbol{\varepsilon}\|$ since $J(\mathbf{v}^*)$ is non-singular. Then Taylor expansion of the right-hand side of (3.9) about $\mathbf{v}^*$ gives

$$\alpha(\mathbf{v}^* + \boldsymbol{\delta}) = Q(\mathbf{v}^*)^k(\mathbf{v}^* + \boldsymbol{\varepsilon}) + G(\mathbf{v}^*)\,\boldsymbol{\varepsilon} + O(\|\boldsymbol{\varepsilon}\|^2),$$

where

$$G(\mathbf{v}) = \nabla(Q(\mathbf{v})^k)\,\mathbf{v}.$$

Equating zero-order terms gives

$$\alpha = (\lambda^*)^k,$$

and equating first-order terms gives

$$(\lambda^*)^k \phi_i \mathbf{z}_i = Q(\mathbf{v}^*)^k \theta_i \mathbf{z}_i + G(\mathbf{v}^*)\,\theta_i \mathbf{z}_i, \qquad i = 1, 2, \ldots, n.$$

Therefore, using Lemma 1,

$$(\lambda^*)^k \phi_i = \lambda_i^k \theta_i + (q-2) \sum_{j=1}^{k} (\lambda^*)^{j-1} (\lambda_i)^{k-j+1} \theta_i$$

$$- (p-2) \sum_{j=1}^{k} (\lambda^*)^j \lambda_i^{k-j} \theta_i, \qquad i = 1, 2, \ldots, n,$$

and so

$$\phi_i/\theta_i = w_i^k + (q-2)\, w_i(1 - w_i^k)/(1 - w_i)$$

$$- (p-2)(1 - w_i^k)/(1 - w_i), \; i = 1, 2, \ldots, n.$$

Since local convergence is implied by $|\phi_i/\theta_i| < 1$, $i = 1, 2, \ldots, n$, the result follows. ∎

COROLLARY. *Let second-order sufficient conditions (i.e., strict inequality in* (3.5)) *hold at* $\mathbf{v}^*$. *Then local convergence is guaranteed when*

(i) $p < 2 + (q-2)\, w_i + (1 - w_i)(1 + w_i^k)/(1 - w_i^k)$, $i = 1, 2, \ldots, n$,

(ii) $p < 3 + (q-1)\, w_i$, $i = 1, 2, \ldots, n$, *if* $k = 1$,

(iii) $p < 3 + (q-3)\, w_i$, $i = 1, 2, \ldots, n$, *if* $k \to \infty$, *and* $|w_i| \leqslant 1$, $i = 1, 2, \ldots, n$,

(iv) $p < 2 + \sqrt{1 + w_i}$, $i = 1, 2, \ldots, n$ *if* $k = 1$ *and* $1/p + 1/q = 1$.

These results suggest that there are advantages in an algorithm based on the simple choice $k = 1$ in (2.8). In particular, it is clear that in this case local convergence is normally guaranteed for all p, $1 < p < 3$, irrespective of the value of q.

4. Global Convergence Analysis

It is convenient to denote the objective function of (2.1) by $F(\mathbf{v})$. If $J(\mathbf{v})$ is defined, then

$$F(\mathbf{v}) = \mathbf{v}^{\mathrm{T}} J(\mathbf{v})\, \mathbf{v}.$$

A key result in the analysis of this section is the following lemma, the proof of which is straightforward.

LEMMA 2. *For any* $a \in R$, $b \in R$, *with* $b \neq 0$ *if* $p < 2$,

$$|a|^p - |b|^p - \tfrac{1}{2} p\, |b|^{p-2}(a^2 - b^2) \leqslant 0, \qquad 1 < p \leqslant 2$$
$$\geqslant 0, \qquad 2 \leqslant p < \infty.$$

COROLLARY 1. *Let* $1 < p \leqslant 2$, *let* $\mathbf{v} \in R^{n+1}$ *be such that* $J(\mathbf{v})$ *is defined, and let* $\mathbf{d} \in R^{n+1}$ *be arbitrary. Then*

$$F(\mathbf{d}) \leqslant F(\mathbf{v}) + \tfrac{1}{2} p(\mathbf{d}^{\mathrm{T}} J(\mathbf{v})\, \mathbf{d} - \mathbf{v}^{\mathrm{T}} J(\mathbf{v})\, \mathbf{v}).$$

Proof. Set $a = (Z\mathbf{d})_i$, $b = (Z\mathbf{v})_i$ in Lemma 2 and sum over i. ∎

COROLLARY 2. *Let* $2 \leqslant q < \infty$, *and let* $\mathbf{v} \in R^{n+1}$, $\mathbf{d} \in R^{n+1}$ *be arbitrary. Then*

$$\|\mathbf{d}_2\|_q^q \geqslant \|\mathbf{v}_2\|_q^q + \tfrac{1}{2} q(\mathbf{d}^{\mathrm{T}} C(\mathbf{v})\, \mathbf{d} - \mathbf{v}^{\mathrm{T}} C(\mathbf{v})\, \mathbf{v}).$$

Proof. Set $a = (\mathbf{d}_2)_i$, $b = (\mathbf{v}_2)_i$ in Lemma 2 and sum over i. ∎

A global convergence result is now given for the iteration (2.8) performed with $k = 1$. Notice that there is no restriction on the value of q.

THEOREM 4. *Let* $1 < p \leqslant 2$, *and let the sequence* $\{\mathbf{v}^{(i)}\}$ *be defined by* (2.8) *with* $k = 1$, *with* $\mathbf{v}^{(0)}$ *arbitrary except that* $\| \mathbf{v}_2^{(0)} \|_q = 1$. *Then*

(i) $F(\mathbf{v}^{(i+1)}) < F(\mathbf{v}^{(i)})$ *unless* $\mathbf{v}^{(i)}$ *is a stationary point of* (2.1),

(ii) *the limit points of* $\{\mathbf{v}^{(i)}\}$ *at which* Q *is defined are stationary points of* (2.1).

Proof. Let $\mathbf{v} \in R^{n+1}$, $\| \mathbf{v}_2 \|_q = 1$ be such that $J(\mathbf{v})$ is defined (and therefore positive definite) and let $\mathbf{d}$ satisfy

$$J(\mathbf{v})\,\mathbf{d} = C(\mathbf{v})\,\mathbf{v}. \tag{4.1}$$

Further, let γ be such that $\| (1/\gamma)\,\mathbf{d}_2 \|_q = 1$. Then by Corollary 1 of Lemma 2,

$$F\left(\frac{1}{\gamma}\mathbf{d}\right) \leqslant F(\mathbf{v}) + \frac{1}{2}p\left(\frac{1}{\gamma^2}\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{d} - \mathbf{v}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{v}\right). \tag{4.2}$$

Now, by convexity

$$\left\| \frac{1}{\gamma}\mathbf{d}_2 \right\|_q^q \geqslant \| \mathbf{v}_2 \|_q^q + q\left(\frac{1}{\gamma}\mathbf{d} - \mathbf{v}\right)^{\mathrm{T}} C(\mathbf{v})\,\mathbf{v},$$

so that

$$\mathbf{d}^{\mathrm{T}}C(\mathbf{v})\,\mathbf{v} \leqslant \gamma,$$

or

$$\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{d} \leqslant \gamma \qquad \text{using (4.1).}$$

Thus

$$\frac{1}{\gamma^2}\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{d} \leqslant \frac{1}{\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{d}}. \tag{4.3}$$

By the Cauchy–Schwartz inequality

$$(\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{v})^2 \leqslant (\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{d})(\mathbf{v}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{v})$$

so that

$$1 \leqslant (\mathbf{d}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{d})(\mathbf{v}^{\mathrm{T}}J(\mathbf{v})\,\mathbf{v}) \qquad \text{from (4.1).}$$

Using (4.3), it then follows from (4.2) that

$$F\left(\frac{1}{\gamma}\mathbf{d}\right) \leqslant F(\mathbf{v})$$

with equality only if $\mathbf{d}$ and $\mathbf{v}$ are parallel, that is, if $\mathbf{v}$ is a stationary point of (2.1). Therefore (i) is proved.

Because Z has full rank, $\{\mathbf{v}^{(i)}\}$ is a bounded sequence (using part (i)), and so has limit points. Let the subsequence $\{\mathbf{v}^{(j_i)}\} \to \mathbf{v}^*$ as $i \to \infty$, with $Q(\mathbf{v}^*)$ defined. Further, going to another subsequence if necessary (which is not renamed)

$$\{\mathbf{v}^{(j_i+1)}\} \to \mathbf{w}^* \qquad \text{as } i \to \infty.$$

Now $F(\mathbf{v}^{(i)})$ is a decreasing sequence, bounded below, and so is convergent, to F^*, say. By continuity of F, it follows that

$$F(\mathbf{v}^*) = F(\mathbf{w}^*) = F^*. \tag{4.4}$$

From (2.8), for each i,

$$\begin{aligned} J(\mathbf{v}^{(j_i)})\,\mathbf{d}^{(j_i)} &= C(\mathbf{v}^{(j_i)})\,\mathbf{v}^{(j_i)}, \\ \mathbf{v}^{(j_i+1)} &= \mathbf{d}^{(j_i)}/\|\,\mathbf{d}_2^{(j_i)}\,\|_q. \end{aligned} \tag{4.5}$$

Letting $i \to \infty$ and using continuity,

$$\begin{aligned} J(\mathbf{v}^*)\,\mathbf{d}^* &= C(\mathbf{v}^*)\,\mathbf{v}^*, \\ \mathbf{w}^* &= \mathbf{d}^*/\|\,\mathbf{d}_2^*\,\|_q. \end{aligned}$$

If $\mathbf{v}^*$ is not a stationary point of (2.1), then by part (i) $F(\mathbf{w}^*) < F(\mathbf{v}^*)$, a contradiction of (4.4) which completes the proof. ∎

The final theorem applies to (2.8) when $k > 1$. In fact it requires that k be sufficiently large that a "close enough" approximation is obtained to the maximum eigenvalue of the generalized eigenproblem (2.6) at each step. Let $\mathbf{v} \in R^{n+1}$, $\|\,\mathbf{v}_2\,\|_q = 1$ be such that $Q(\mathbf{v})$ is defined and let $\bar{\lambda}$ denote the largest eigenvalue of the generalized eigenproblem

$$Q(\mathbf{v})\,\mathbf{d} = \lambda\,\mathbf{d}. \tag{4.6}$$

Let $\mathbf{d}_j$ satisfy

$$\mathbf{d}_j = Q(\mathbf{v})^j\mathbf{v}, \qquad j = 1, 2, \ldots,$$

with $\|\,(\mathbf{d}_j)_2\,\|_q^q = \gamma_j^q$. Then it follows that

$$\frac{1}{\gamma_j}\,\mathbf{d}_j \to \bar{\mathbf{d}} \qquad \text{as } j \to \infty, \tag{4.7}$$

where $\bar{\mathbf{d}}$ is the eigenvector, suitable normalized, corresponding to $\bar{\lambda}$ (or some linear combination of the eigenvectors if $\bar{\lambda}$ is a multiple eigenvalue). Now by definition

$$\begin{aligned}\bar{\lambda} &= \frac{\bar{\mathbf{d}}^{\mathrm{T}} C(\mathbf{v})\,\bar{\mathbf{d}}}{\bar{\mathbf{d}}^{\mathrm{T}} J(\mathbf{v})\,\bar{\mathbf{d}}} \\ &= \max_{\mathbf{y}\neq 0} \frac{\mathbf{y}^{\mathrm{T}} C(\mathbf{v})\,\mathbf{y}}{\mathbf{y}^{\mathrm{T}} J(\mathbf{v})\,\mathbf{y}} \\ &\geqslant \frac{\mathbf{v}^{\mathrm{T}} C(\mathbf{v})\,\mathbf{v}}{\mathbf{v}^{\mathrm{T}} J(\mathbf{v})\,\mathbf{v}} \\ &= \frac{1}{\mathbf{v}^{\mathrm{T}} J(\mathbf{v})\,\mathbf{v}}\end{aligned}$$

with equality holding only if $\mathbf{v}$ is an eigenvector of (4.7), in other words a stationary point of (2.1). Thus if $\mathbf{v}$ is not a stationary point, then for k sufficiently large (4.7) implies that

$$\frac{\mathbf{d}^{\mathrm{T}} C(\mathbf{v})\,\mathbf{d}}{\mathbf{d}^{\mathrm{T}} J(\mathbf{v})\,\mathbf{d}} > \frac{1}{\mathbf{v}^{\mathrm{T}} J(\mathbf{v})\,\mathbf{v}}, \tag{4.8}$$

where $\mathbf{d} = \mathbf{d}_k$. The proof of the following theorem requires that, at each iteration of (2.8), k be large enough that the corresponding inequality

$$\frac{\mathbf{d}^{(i)\mathrm{T}} C(\mathbf{v}^{(i)})\,\mathbf{d}^{(i)}}{\mathbf{d}^{(i)\mathrm{T}} J(\mathbf{v}^{(i)})\,\mathbf{d}^{(i)}} > \frac{1}{\mathbf{v}^{(i)\mathrm{T}} J(\mathbf{v}^{(i)})\,\mathbf{v}^{(i)}} \tag{4.9}$$

is satisfied at each step. The above argument shows that this is always possible away from a stationary point.

THEOREM 5. *Let* $1 < p \leqslant 2$, $2 \leqslant q < \infty$, *and let the sequence* $\{\mathbf{v}^{(i)}\}$ *be defined by* (2.8) *with* $\mathbf{v}^{(0)}$ *arbitrary except that* $\|\mathbf{v}_2^{(0)}\|_q = 1$. *Then if* (4.9) *holds at each step*

(i) $F(\mathbf{v}^{(i+1)}) < F(\mathbf{v}^{(i)})$ *unless* $\mathbf{v}^{(i)}$ *is a stationary point of* (2.1),

(ii) *the limit points of* $\{\mathbf{v}^{(i)}\}$ *at which* J *is defined are stationary points of* (2.1).

Proof. Let $\mathbf{v} \in R^{n+1}$, $\|\mathbf{v}_2\|_q = 1$ be such that $J(\mathbf{v})$ is defined and let $\mathbf{d}$ satisfy

$$\mathbf{d} = Q(\mathbf{v})^k \mathbf{v}, \tag{4.10}$$

and the inequality (4.8). Let γ be such that $\|(1/\gamma)\,\mathbf{d}_2\|_q = 1$. Then by Corollary 1 of Lemma 2, (4.2) is satisfied. Also by Corollary 2 of Lemma 2,

$$\left\| \frac{1}{\gamma}\,\mathbf{d}_2 \right\|_q^q \geqslant \|\mathbf{v}_2\|_q^q + \frac{1}{2}q\left(\frac{1}{\gamma^2}\,\mathbf{d}^{\mathrm{T}}C(\mathbf{v})\,\mathbf{d} - \mathbf{v}^{\mathrm{T}}C(\mathbf{v})\,\mathbf{v}\right),$$

so that

$$\frac{1}{\gamma^2} \leqslant \frac{1}{\mathbf{d}^{\mathrm{T}}C(\mathbf{v})\,\mathbf{d}}. \tag{4.11}$$

It follows from (4.2), using (4.8) and (4.11), that $F((1/\gamma)\,\mathbf{d}) < F(\mathbf{v})$ unless $\mathbf{v}$ is a stationary point of (2.1), and therefore (i) is proved.

Part (ii) follows as in Theorem 4. ∎

Unfortunately these theorems do not give a complete analysis of the convergence of (2.8). In particular they leave open the question of global convergence when $k > 1$ in (2.8) but (4.9) is not satisfied at every step. The situation in this case remains unresolved.

References

1. G. H. Golub and C. F. van Loan, An analysis of the total least squares problem, *SIAM J. Numer. Anal.* **17** (1980), 883–893.
2. M. R. Osborne and G. A. Watson, An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem, *SIAM J. Sci. Statist. Comput.* **6** (1985), 410–424.
3. H. Späth, On discrete linear orthogonal L_p-approximation, *Z. Angew. Math. Mech.* **62** (1982), 354–355.
4. G. A. Watson, Numerical methods for linear orthogonal L_p approximation, *IMA J. Numer. Anal.* **2** (1982), 275–287.
5. G. A. Watson, The total approximation problem, *in* "Approximation Theory IV," (C. K. Chui, L. L. Schumaker, and J. D. Ward, Eds.), Academic Press, New York, 1983.
6. G. A. Watson, The numerical solution of total l_p approximation problems, *in* "Numerical Analysis, Dundee 1983," (D. F. Griffiths, Ed.), Springer-Verlag, Berlin, 1984.