ITERATIVE ALGORITHMS FOR GRAM-SCHMIDT ORTHOGONALIZATION

by

W. Hoffmann, Amsterdam

Abstract - Zusammenfassung

The algorithms that are treated in this paper are based on the classical and the modified Gram-Schmidt algorithms. It is shown that Gram-Schmidt orthogonalization for constructing a QR factorization should be carried out iteratively to obtain a matrix Q that is orthogonal in almost full working precision. In the formulation of the algorithms, the parts that express manipulations with matrices or vectors are clearly identified to enable an optimal implementation of the algorithms on parallel and/or vector machines. An extensive error analysis is presented. It shows, for instance, that the iterative classical algorithm is not inferior to the iterative modified algorithm when full precision of Q is required. Experiments are reported to support the outcomes of the analysis.

Keywords: Gram-Schmidt orthogonalization, QR factorization, vector algorithms.
AMS subject classification: 65F25, 65G05, 15A23.

In diesem Artikel werden verschiedene Varianten der klassischen und der modifizierten Gram-Schmidt Methode präsentiert. Wir zeigen, dass man für die Konstruktion der QR-Zerlegung die Gram-Schmidt Orthogonalisation iterativ anwenden muss, falls man die Matrix Q ungefähr bis auf Maschinengenauigkeit orthogonal haben will. Die Algorithmen sind so formuliert, dass man alle Operationen mit Matrizen oder Vektoren deutlich identifizieren kann und eine Implementierung auf einem Parallel- oder Vektorcomputer keine Schwierigkeiten bietet. Eine ausführliche Fehleranalyse wird gegeben. Daraus folgt zum Beispiel, dass der iterative klassische Algorithmus nicht schlechter ist als der iterative modifizierte Algorithmus, wenn die Matrix Q so genau wie möglich orthogonal sein muss. Verschiedene Experimente auf einem Vektorcomputer werden beschrieben, welche die Resultate der Fehleranalyse bestätigen.

1. INTRODUCTION

We consider variants of Gram-Schmidt orthogonalization and their suitability for use on super computers. Algorithms for super computers must exploit the parallel and/or vector facilities of the machine to admit for an optimal performance. The numerical stability of the algorithm, however, may require that a formulation which seems to be particularly favourable for use on super computers should be avoided. Some well known variants of the Gram-Schmidt algorithm are good examples of this type of conflict.
The goal of Gram-Schmidt orthogonalization is to construct a QR factorization. This factorization is defined as follows.
Consider an $m \times n$ matrix $A = [a_1,\ldots,a_n]$ with $a_j \in \mathbf{R}^m$ and $m \geq n$. Let $k(j) = \dim(\text{span}(a_1,\ldots,a_j))$ for $j = 1,\ldots,n$ and let $p = k(n)$ {= rank(A)}.
An orthogonal basis $[q_1,\ldots,q_p]$ for $\text{span}(a_1,\ldots,a_n)$ is to be constructed such that $a_i \in \text{span}(q_1,\ldots,q_{k(i)})$, $i = 1,\ldots, n$. In terms of matrix calculation this is equivalent with: construct an orthogonal $m \times p$ matrix Q such that $A = QR$ for a $p \times n$ upper trapezoidal matrix R. If $p = n$, then the problem is called a full-rank problem .
If matrix Q is not used, or is only needed to calculate the product Qv for several vectors v, then the Householder algorithm is to be preferred. If the individual column vectors of matrix Q are wanted (the so called "orthogonal basis" problem), then, in case of a full-rank matrix A, the Gram-Schmidt algorithm is advantageous.
This paper deals with Gram-Schmidt orthogonalization for the case that the matrix has (numerically) full rank, i.e. $p = n$.
For the case $p < n$, the Gram-Schmidt algorithm has been extended with the application of column pivoting; see Businger and Golub [2]. This addition gives quite satisfactory results in most practical cases, but may not detect the right degree of rank deficiency. An adaption of the Gram-Schmidt algorithm which is presented by Chan [3], yields correct results in the general situation with $p \leq n$ and calculates the correct rank of the matrix.

It has become well known that various so called "block QR" algorithms admit efficient performance on super computers. Some of these algorithms, however, appear to be variants of the classical Gram-Schmidt algorithm and it has been acknowledged that the classical algorithm may produce a matrix Q that is far from orthogonal.
The method known as "the modified Gram-Schmidt algorithm" is numerically to be preferred over the classical algorithm; the orthogonality of Q is of the order of machine precision times condition number of the matrix. This may be insufficient for matrices that are ill conditioned. To overcome this shortcoming, the modified algorithm can be applied iteratively, so that almost full machine precision is reached.
We show that the classical algorithm in an iterative fashion can attain that same accuracy in an equal number of iterations. Consequently, constructions that were banned for the sake of accuracy can be accepted in an iterative algorithm.
In section 2 we give definitions of the classical and the modified Gram-Schmidt algorithm through an algorithmic formulation. In section 3 we present the iterative versions of these algorithms. Iterative versions of the Gram-Schmidt algorithm are also presented by Daniel, Gragg, Kaufman and Stewart [4] and by Ruhe[7]. In

section 4 we present the results of numerical experiments which are discussed in section 5. In section 6 we draw our conclusions.

2. ONE-STEP GRAM-SCHMIDT ALGORITHMS.

For ease of formulation we use the normalizing operator N defined by:
$N(x) = x / \|x\|$, for vectors $x \neq \underline{0}$.
For a full-rank rectangular m × n matrix A, m ≥ n, the orthogonal m × n matrix Q whose columns form an orthogonal basis for the subspace span(a_1,…,a_n) can be defined with the use of projections as follows:

```
1.    Q := [0,…, 0]                          {the m × n zero-matrix}
2.    For j = 1,…, n do
      1.  Q := Q + N ((I - QQ^T) a_j) e_j^T
```

If the elements of the triangular matrix R (=Q^TA) are wanted too, then the description turns into the Classical Gram-Schmidt algorithm, CGS.
Algorithm CGS is given by :

```
1.    Q := [0,…, 0]                          {the m × n zero-matrix}
2.    For j = 1,…, n do
      1.  r_j  :=  Q^T a_j
      2.  t    :=  a_j  -  Q r_j             {t = (I - QQ^T) a_j}
      3.  r_jj :=  || t ||_2
      4.  q_j  :=  t / r_jj                  {q_j = N (t)}
```

The numerical behaviour of this algorithm is very poor in a sense that in many cases the constructed matrix Q is far from orthogonal. This well known result has been shown by Björck [1].
An improved algorithm is the Modified Gram-Schmidt algorithm .
This algorithm exists in two versions; MGSC and MGSR, constructing the matrix R column by column or row by row, respectively. The difference shows only in the way the data is accessed.
Algorithm MGSC has the same structure as algorithm CGS; the difference is that individual elements of vector Q^Ta_j, which is calculated in line 2.1 of CGS, must be calculated sequentially by taking innerproducts with successive columns q_i of Q so that the appropriate multiple of q_i can be subtracted from a_j as soon as its coefficient is available. This repeated modification of column a_j is the crux of the algorithm.

Algorithm MGSC is given by:

```
For j = 1,…, n do
1.    t   :=  a_j
2.    For i = 1,…, j-1 do
      1.  r_ij  :=  q_i^T t
      2.  t   :=  t  -  q_i r_ij
3.    r_jj  :=  ‖ t ‖_2
4.    q_j   :=  t / r_jj                          { q_j = N (t) }
```

The update-rule as given in lines 2.1 and 2.2. of MGSC can as well be applied for each q_i on all columns a_k with $k \geq i$. In that case the i-th row of R is computed as a whole; its elements can be calculated in parallel. This gives rise to algorithm MGSR.
For the description we use the following notation. With $r_{i\bullet}$ we denote the i-th row of R and with $\underline{r}_{i\bullet}$ we denote the part of $r_{i\bullet}$ that is strictly to the right of the diagonal: $(r_{ii+1},\ldots,r_{in})$.

Algorithm MGSR is described by:

```
For i = 1,…, n do
1.    r_ii  :=  ‖ a_i ‖_2
2.    q_i   :=  a_i / r_ii
3.    r_i•  :=  q_i^T [a_i+1,…, a_n]
4.    [a_i+1,…,  a_n] := [a_i+1,…,  a_n] - q_i r_i•  {rank-one  matrix
update}
```

Although the results of both MGS algorithms are an improvement over the results obtained by CGS in the sense that the orthogonality of matrix Q is much better, in many cases the orthogonality is still not good enough. This is reflected in the bounds for $\|Q^TQ - I\|_2$ which is of the order of the product of the machine precision ε and the condition number of the original matrix, as has been shown by Björck [1].

3. Iterative Gram-Schmidt algorithms

Iterative Gram-Schmidt algorithms with improved orthogonality have been presented and analysed by Daniel et al. [4] and Ruhe [7].
We here describe the iterative versions of both the classical and the modified Gram-Schmidt algorithms; the modified algorithm only in the MGSC form. A corresponding iterative version of the modified algorithm in its MGSR form is not possible.

Algorithm CGSI reads :

```
Q  :=  [0,…, 0]                                  {the m × n zero-matrix}
For j = 1,…, n do
1.    r_j  :=  0
2.    t    :=  a_j
3.    Repeat
      1.  p   :=  t
      2.  s   := Q^T p
      3.  v   := Q s
      4.  t   := p - v
      5.  r_j  := r_j + s
4.    Until  < t perpendicular span(q_1,…, q_j-1) >
5.    r_jj  := ‖ t ‖_2
6.    q_j  :=  t / r_jj                          { q_j = N (t) }
```

Algorithm MGSCI reads :

```
Q  :=  [0,…, 0]                                  {the m × n zero-matrix}
For j = 1,…, n do
1.    r_j  := 0
2.    t    := a_j
3.    Repeat
      1.  p   := t
      2.  For i = 1,…, j-1 do
          1.  s_i  :=  q_i^T t
          2.  t    := t  -  q_i s_i
      3.  r_j  := r_j + (s_1,…, s_j-1,0,…, 0)^T
4.    Until < t perpendicular to span(q_1,…, q_j-1) >
5.    r_jj  := ‖ t ‖_2
6.    q_j  :=  t / r_jj
```

It has been demonstrated by Ruhe [7] that the resulting r_j in the j-th step corresponds with the solution of the equation $Q^TQr_j = Q^Ta_j$ with $Q = [q_1,\ldots,q_{j-1}]$. The CGSI variant corresponds with Gauss-Jacobi iteration for solving that equation and the MGSCI variant with Gauss-Seidel iteration. The resulting accuracy depends on the number of iteration-steps performed.
We would like to emphasize that in Ruhe's analysis the (almost) orthogonality of matrix Q is not used; the goal in the j-th step is to find r_j such that $(a_j - Qr_j)$ is orthogonal to $\text{span}(q_1,\ldots,q_{j-1})$.
The new column q_j is obtained from $q_j := N\,(a_j - Qr_j)$. An implementation of the stopping criterion "t perpendicular to $\text{span}(q_1,\ldots,q_{j-1})$" was not suggested by Ruhe.
For a useful stopping criterion we are inspired by Parlett [6], who analyses Gram-Schmidt orthogonalization for two vectors. He presents an "iterative" orthogonalization algorithm which he attributes to W. Kahan; iterative has been put between quotes because a single reorthogonalization step is sufficient in practice. It provides us with an efficient stopping criterion for algorithm CGSI and it shows to be adequate for MGSCI too.

In the j–th major step of CGSI, vectors $s = fl(Q^T p)$ and $v = fl(Qs)$ are calculated (lines 3.2 and 3.3). A backward error analysis, using one of the customary matrix norms, shows:

$s = (Q + \delta_1 Q)^T p$ with $\|\delta_1 Q\| \le \phi_1(m,j)\, \|Q\| \varepsilon$ and

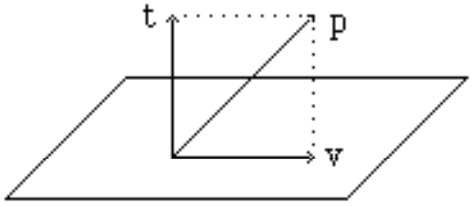
$v = (Q + \delta_2 Q)s$ with $\|\delta_2 Q\| \le \phi_2(m,j)\, \|Q\| \varepsilon$

where ϕ_1 and ϕ_2 stand for low degree polynomials in m and j and ε stands for the effective machine precision (effective means that effects of arithmetic are taken into account).

If v were calculated without error (i.e. $\delta_2 Q = 0$), it would be contained in the column space of Q, regardless the accuracy of s. If s were also calculated without error and Q were exactly orthogonal, then v would be equal to p's projection onto Q's column space.

The vector $t = fl(p - v)$ (line 3.4) is calculated as an approximation to p's component orthogonal to $span(q_1, \ldots, q_{j-1})$. For t we have:

$t = p + \delta p - (v + dv)$ with $\|\delta p\| \le m\, \|p\| \varepsilon$ and $\|\delta v\| \le m\, \|v\| \varepsilon$.



span($q_1, \ldots, q_{j-1}$)

In the following theorem we show that t is close to a vector orthogonal to $span(q_1, \ldots, q_{j-1})$.

Theorem 1.

Let vectors p and $q_1,\ldots,q_{j-1}$ be given in $\mathbf{R}^m$ with $p \notin span(q_1,\ldots,q_{j-1})$.

Suppose $Q = [q_1,\ldots,q_{j-1}]$ and $a \in (0,1)$ are such that $\|Q^T Q - I\|_2 \le \alpha$.

Let scalar μ (> 0), vectors δp and δv and matrices $\delta_1 Q$ and $\delta_2 Q$ be such that

$\|\delta_1 Q\|_2 \le \mu \|Q\|_2$, $\|\delta_2 Q\|_2 \le \mu \|Q\|_2$, $\|\delta p\|_2 \le \frac{1}{\sqrt{2}}\, \mu \|p\|_2$ and $\|\delta v\|_2 \le \frac{1}{\sqrt{2}}\, \mu \|v\|_2$.

Let $s = (Q + \delta_1 Q)^T p$, $v = (Q + \delta_2 Q)s$, $t = p + \delta p - (v + \delta v)$ and let u denote the orthogonal projection of p onto the space perpendicular to $span(q_1,\ldots,q_{j-1})$.

Then the difference between t and u (i.e. the error in t), is bounded as follows:

$$\|u - t\|_2 \le \left\{ \alpha \frac{1+\alpha}{1-\alpha} + \frac{1}{2}\mu\{(2 + \mu)^2 + 1\}(1+\alpha) + 1\} \right\} \|p\|_2 .$$

For instance, for a and $m \le 0.1$, this implies $\|u - t\|_2 \le (1.3\, \alpha + 3.5\, \mu)\, \|p\|_2$.

Proof of Theorem 1.

For the proof of this theorem we need the following results:

Lemma 1. If $\|Q^T Q - I\|_2 \le \alpha$ then $\|Q\|_2^2 \le 1 + \alpha$.

Proof: This result is a direct consequence of the triangle inequality.

Lemma 2. If $\|Q^T Q - I\|_2 \le \alpha < 1$ then $\|I - (Q^T Q)^{-1}\|_2 \le \frac{\alpha}{(1 - \alpha)}$.

Proof: From the identity

$$I - (Q^TQ)^{-1} = \{\{(Q^TQ)^{-1} - I\} + I\}(Q^TQ - I),$$

it follows

$$\|I - (Q^TQ)^{-1}\|_2 \leq \|I - (Q^TQ)^{-1}\|_2 \, \|Q^TQ - I\|_2 + \|Q^TQ - I\|_2.$$

From this, the result follows immediately .

For u we have, using the appropriate projection operator

$$u = (I - Q(Q^TQ)^{-1}Q^T)\,p,$$

so that for the difference u - t we have

$$u - t = (Q + \delta_2 Q)(Q + \delta_1 Q)^T p - Q(Q^TQ)^{-1}Q^T p + (\delta v - \delta p).$$

This yields

$$u - t = \{Q\{I - (Q^TQ)^{-1}\}Q^T + Q\delta_1 Q^T + \delta_2 QQ^T + \delta_2 Q\delta_1 Q^T\}\,p + (\delta v - \delta p).$$

A straightforward calculation yields

$$\|u - t\|_2 \leq \{\{\frac{\alpha}{(1-\alpha)} + 2\mu + \mu^2\}(1 + \alpha) + \frac{1}{\square^2}\,\mu(1 + \frac{\|v\|_2}{\|p\|_2})\}\,\|p\|_2 .$$

From the definition of v we find:

$$\frac{\|v\|_2}{\|p\|_2} \leq (1 + \mu)^2 (1 + \alpha), \qquad (3.1)$$

from which the rest of the proof follows immediately.

□

Consider for certain $j \leq m$ an orthonormal $m \times (j-1)$ matrix $Q = [q_1,\ldots,q_{j-1}]$ and an m-vector p not in the column space of Q. Suppose that for this vector p a single step in algorithm CGSI is carried out to construct an orthonormal basis for $\text{span}(q_1,\ldots,q_{j-1},p)$, then the conditions of the theorem are fulfilled with a small value for a and (possibly) a very small value for m, depending on the sizes of m and j and the effective machine precision. Although the calculated vector t is close to a vector that is perpendicular to the column space of Q, it is not true that consequently t itself is almost orthogonal to that column space. For example, the orthogonality may be (very) bad if t is of the order of the error and is therefore small compared to p.

In the Parlett-Kahan algorithm a reorthogonalization of t against v is prescribed in cases that $\|t\|_2$ is smaller than $\|p\|_2$ divided by a selected accuracy factor κ. This factor must be chosen larger than a constant $\beta > 1$. (They use the value $\beta = (0.83 - \varepsilon)^{-1}$.)

In our situation we conclude that if t is suspected of being not orthogonal to v, it can certainly be suspected of being not orthogonal to columns of Q.

So the decision that a reorthogonalization is required can be made on the same grounds.

The reverse, being the acceptance of t if it is large enough is not evident, but will follow from theorem 2.

The situation that $t = \underline{0}$ does not occur if the given matrix A has (numerically) full rank.

If t is large enough relative to p, then the orthogonality of t with respect to the column space of Q can be estimated as expressed in the following theorem.

Theorem 2.

Let the conditions of theorem 1 be fulfilled and let moreover k be such that $\|t\|_2 \geq \|p\|_2 / \kappa$.

The orthogonality between t and the column space of Q satisfies:

$$\frac{\|Q^T t\|_2}{\|t\|_2} \leq (1 + \frac{1}{\square 2}\alpha)\{\frac{3}{2}\mu + (1+\mu)[\alpha + \mu(1+\alpha) + \frac{1}{\square 2}\mu(1+\mu)(1+\alpha)]\}\kappa .$$

For instance: for α and $\mu < 0.1$ this implies $\frac{\|Q^T t\|_2}{\|t\|_2} \leq (1.2\,\alpha + 3.6\,\mu)\,\kappa$.

Proof of Theorem 2

Next to lemma 1 from theorem 1 we need the following lemma:

Lemma

If $\|Q^T Q - I\|_2 \leq \alpha$ then $\|Q\|_2 \leq 1 + \frac{1}{2}\alpha$.

Proof:

This result follows directly from the inequality $\sqrt{1+x} \leq 1 + x/2$ for $x \geq 0$.

Substitution of the expressions for t and v yields:

$$Q^T t = Q^T (p - v + \delta p - \delta v)$$
$$= (Q + \delta_1 Q)^T p - Q^T v + Q^T(\delta p - \delta v) - \delta_1 Q^T p .$$

This implies

$$Q^T t = (I - Q^T Q)s - Q^T \delta_2 Q s + Q^T(\delta p - \delta v) - \delta_1 Q^T p,$$

which yields

$$\| Q^T t \|_2 \leq (\alpha + \mu \|Q\|_2^2) \| s \|_2 + \frac{1}{2}\mu \|Q\|_2 (1 + \frac{\|v\|_2}{\|p\|_2}) \|p\|_2 + \mu \|Q\|_2 \|p\|_2 .$$

With the use of formula 3.1, the observation $\|p\|_2 \leq \|t\|_2 \kappa$ and the inequality $\| s \|_2 \leq (1 + \mu) \|Q\|_2 \|p\|_2$, the result follows directly.

□

According to this theorem we use the following stopping criterion:

In line 4 of CGSI the condition on orthogonality can be implemented as:

$\|t\|_2 > \|p\|_2 / \kappa$ {for some positive $\kappa \geq \beta$ } .

The reliability on the orthogonality of Q diminishes with larger values of κ .

Conclusions from extensive experimenting on the choice of κ are reported in the next section; a provisional statement is that algorithm CGSI gives good results with the choice $\kappa = 2$.

From Ruhe's observation that the calculation of r_j is equivalent with Gauss-Jacobi iteration (c.q. Gauss-Seidel iteration) on the linear system $Q^T Q r_j = Q^T a_j$ and from the fact that the iteration matrix $Q^T Q$ is close to a diagonal matrix, we may conclude that in general the calculated solution is very accurate. In practice we can assume that the calculated r_j gives a small residue which says that for a constant γ that is not too big the following bound holds :

$$\|Q^T a_j - Q^T Q r_j\|_2 \leq \gamma.\varepsilon \| a_j \|_2 .$$

Let us focus on algorithm CGSI. For vector t we have theoretically:

$$t = a_j - Q r_j .$$

so that for the quantity $\|Q^T t\|_2$ we find

$$\|Q^T t\|_2 \leq \gamma.\varepsilon \| a_j \|_2 .$$

If a single iteration step is sufficient for the current value of j, we have $p = a_j$ so that for the quotient $\frac{\|Q^T t\|_2}{\|t\|_2}$ we have

$$\frac{\|Q^T t\|_2}{\|t\|_2} \leq \gamma.\varepsilon\,\kappa\,\frac{\|a_j\|_2}{\|p\|_2} = \gamma.\varepsilon\,\kappa\,. \qquad (3.2)$$

If an extra iteration step is needed, we are in the situation that $a_j \approx Qr_j$ i.e. $\|t\| << \|a_j\|$ (in all our experiments we never observed the need for more than one extra iteration step). The extra iteration can be interpreted as a calculation of t and of the residual vector $Q^T t$ in extended precision. Also in this situation we will find a good relative accuracy. In either situation, the bound in theorem 2, which depends on $\|Q^TQ - I\|_2$, may be therefore (much) too pessimistic.

Suppose that $\tilde{Q}$ denotes the matrix that follows from adding the calculated column q_j (i.e. $N(t)$) to columns $(q_1,\ldots,q_{j-1})$. In successive steps of the algorithm, matrix $\tilde{Q}$ replaces Q.

So we are concerned with a bound for $\|\tilde{Q}^T\tilde{Q} - I\|_2$ in relation to the bound for $\|Q^TQ - I\|_2$. This is settled in the following theorems.
In theorems 3 and 4 we treat the case that the added column has a spectral norm that is exactly equal to one; in theorem 5 we cover the effect of rounding errors.

Theorem 3.
Let $Q = [q_1,\ldots,q_{j-1}]$ and $\alpha \in (0,1)$ be such that $\|Q^TQ - I\|_2 \leq \alpha$.

Let q_j with $\|q_j\|_2 = 1$ be such that $\|Q^Tq_j\|_2 \leq \omega$. Define $\tilde{Q} = [Q:q_j]$ then

$$\|\tilde{Q}^T\tilde{Q} - I\|_2 \leq \frac{1}{2}\left(\alpha + \sqrt{\alpha^2 + 4\omega^2}\right).$$

Proof of Theorem 3.
If τ is an eigenvalue with maximal modulus of the $(j \times j)$ symmetric matrix $([Q:q_j]^T[Q:q_j] - I)$ then $\|\tilde{Q}^T\tilde{Q} - I\|_2 = |\tau|$.
Suppose that $(x:\delta)^T$ defines a partitioning of an eigenvector corresponding with τ and define $V = Q^TQ - I$ and $w = Q^Tq_j$ then the following equations hold:

$$Vx + \delta w = \tau x \quad \text{and} \quad w^Tx = \tau\delta\,. \qquad (3.3)$$

If $\delta = 0$ then $Vx = \tau x$ from which we find $|\tau| \leq \alpha$ which ends the proof for this case.
For $\delta \neq 0$ we may assume $x \neq \underline{0}$; the case $x = \underline{0}$ can only occur for $\tau = 0$ which satisfies the inequality to be proven in a trivial way.
Eliminating δ from equations (3.3) yields

$$Vx + \tau^{-1}ww^Tx = \tau x.$$

This leads to a quadratic equation in τ by taking innerproducts with x and multiplication with τ.

Through the observation that the Rayleigh quotient $\frac{x^TVx}{x^Tx}$ is bounded by α we find from this quadratic:

$$|\tau| \leq \frac{1}{2}\left(\alpha + \sqrt{\alpha^2 + 4\omega^2}\right),$$

which ends the proof.

□

Using a fixed upperbound for $\|Q^Tq_j\|$, independently of j, the departure from orthogonality of matrix Q can be expressed as in the following theorem:

Theorem 4.

Let $Q_j = [q_1,\ldots,q_j]$., for $j = 1,\ldots,n$. If $\|q_k\|_2 = 1$, for $k = 1,\ldots,j$ and ω is such that $\| Q_{k-1}^T q_k \|_2 \le \omega$ for $k = 2,\ldots,j$ then $\|Q_j^T Q_j - I \|_2 \le \omega\sqrt{2j}$.

Proof of theorem 4:

For the proof we use the following lemma

Lemma:

The elements of the row $(a_1, a_2, \ldots, a_k, \ldots)$, defined by the recurrence relation

$a_1 = 1;\ a_{k+1} = \frac{1}{2}(a_k + \sqrt{a_k^2 + 4})$, $k = 1, 2, \ldots$

satisfy $a_k < \sqrt{2k}$.

Proof:

The proof is by mathematical induction.

For $k = 1$ the result holds.

Assume that $a_k = f\sqrt{2k}$ for some value of $f < 1$. Using the definition of a_{k+1} we find

$$\left(\frac{a_{k+1}}{\sqrt{2k+2}}\right)^2 = \frac{f^2k + 1 + \sqrt{f^4k^2 + 2f^2k}}{2k + 2} < \frac{f^2k + 1}{k + 1}$$

From this we conclude $\frac{a_{k+1}}{\sqrt{2k+2}} < 1$ which ends the proof of this lemma.

We like to comment that the given bound is rather sharp for relatively small values of k already, as can be concluded from simple calculations which yield for example $\frac{a_{50}}{\sqrt{100}} \approx 0.99$.

The final proof of theorem 4 follows directly from the application of theorem 3 and the above lemma.

□

In practice, the norms of the columns of Q are not exactly equal to one. The consequences of this are considered in the next theorem.

Theorem 5.

Assume that $Q_j = [q_1,\ldots,q_j]$ is calculated in floating-point arithmetic.

Let $\delta_k = \|q_k\|_2$, for $k = 1,\ldots,j$ and $D_j = \mathrm{diag}(\delta_1,\ldots,\delta_j)$.

Let $\omega > 0$ be such that $\| Q_{k-1}^T q_k \|_2 \le \omega\, \delta_k$ for $k = 2,\ldots,j$

and let $\sigma \in (0,1)$ be such that $|\, \delta_k^2 - 1\,| \le \sigma$ for $k = 1,\ldots, j$.

Then $\|Q_j^T Q_j - I\|_2 \le \sigma + \omega\,(1 + \sigma)\sqrt{2j}$.

Proof of theorem 5:

Observe that theorem 4 is applicable for matrix $Q_j D_j^{-1}$.

If W_j is defined by $W_j = D_j^{-1} Q_j^T Q_j D_j^{-1}$ we find accordingly

$$\|W_j - I\|_2 \le \omega\sqrt{2j}.$$

For the spectrum of W_j, $\lambda(W_j)$, we have consequently

$$\lambda(W_j) \in [1 - \omega\sqrt{2j}, 1 + \omega\sqrt{2j}] .$$

Combining this with $\lambda(D_j) \in [\sqrt{1 - \sigma}, \sqrt{1 + \sigma}]$ yields

$$(1 - \omega\sqrt{2j})(1 - \sigma) \le \|D_j W_j D_j\|_2 \le (1 + \omega\sqrt{2j})(1 + \sigma),$$

so that for the spectrum of $(Q_j^T Q_j - I)$ we have

$$\lambda(Q_j^T Q_j - I) \in [-\sigma - \omega\,(1 - \sigma)\sqrt{2j}, \sigma + \omega\,(1 + \sigma)\sqrt{2j}]$$

from which the desired result directly follows.

□

4. NUMERICAL EXPERIMENTS

All experiments were carried out on the CYBER 205 computer of SARA, the Academic Computer Centre in Amsterdam. For this machine the value of e is about 5_{10}-14. We carried out experiments with algorithms CGSI and MGSCI on a large number of matrices having various numbers of rows and columns and different sorts of distributions for their singular values. The smallest matrices consisted of 50 rows and 25 columns; the largest matrices of 210 rows and 200 columns.
The matrices are constructed by multiplying a given diagonal matrix (singular values) from both sides by random orthogonal matrices. The maximal singular value is always equal to 1 and the smallest varies between 0.1 and $_{10}$-12 so that the condition number of the matrices is between 10 and $_{10}12$.
We have observed that the distribution of the singular values within the interval $[\sigma_{min},\ldots,\sigma_{max}]$ is of little importance for the resulting orthogonality of Q.

The number of iterations performed depends on parameter κ; for all matrices used, κ has been given the values 2, 10, $_{10}2$, $_{10}3$, … , $_{10}10$, in successive experiments. The effect of taking a smaller value for κ is that in some cases a second iteration is necessary to calculate the next column of Q; a third iteration never occurred.

In table 1 we show a representative selection of our test results; it shows the typical behaviour of algorithms CGSI and MGSCI for various values of parameter κ.
The average number of iterations per column is denoted by ν; the departure from orthogonality is measured in the l_1-norm, and given by $\| Q^TQ - I \|_1$.
All matrices used in the selection described in table 1 have m = 210 and n= 100; the singular values are distributed equally over the interval [(conditionnumber) $^{-1}$, 1].

We also carried out a number of experiments with matrices that are close to a matrix of rank one.
A representative result is described in table 2; the matrix that is used has m = 50 and n = 25 ,the largest singular value is equal to 1 and the remaining 24 singular values are distributed equally in $[1.0_{10}$-11 , 1.0_{10}-10].

cond. nr.	k	ν (= avg. nr. iter. per col)		error in $(Q^TQ - I)$	
		CGSI	MGSCI	CGSI	MGSCI
10	2	1.1	1.1	2.6_{10}-13	1.3_{10}-13
	10	1	1	3.0_{10}-13	1.6_{10}-13
$_{10}$+4	2	1.78	1.78	1.8_{10}-13	8.9_{10}-14
	10	1.57	1.57	3.3_{10}-12	3.1_{10}-13
	$_{10}$+2	1.26	1.26	3.1_{10}-10	3.4_{10}-12
	$_{10}$+3	1.02	1.02	7.8_{10}- 9	1.3_{10}-11
	$_{10}$+4	1	1	9.8_{10}- 9	1.7_{10}-11
$_{10}$+7	2	1.86	1.86	2.1_{10}-13	7.7_{10}-14
	10	1.76	1.76	1.1_{10}-12	2.7_{10}-13
	$_{10}$+2	1.58	1.58	5.9_{10}-10	4.6_{10}-12
	$_{10}$+3	1.42	1.42	6.5_{10}- 8	1.3_{10}-10
	$_{10}$+4	1.27	1.27	5.3_{10}- 6	7.0_{10}-10
	$_{10}$+5	1.13	1.13	1.8_{10}- 4	3.0_{10}- 9
	$_{10}$+6	1.01	1.01	6.5_{10}- 3	1.0_{10}- 8
	$_{10}$+7	1	1	6.5_{10}- 3	1.0_{10}- 8
$_{10}$+10	2	1.89	1.89	2.1_{10}-13	7.8_{10}-14
	10	1.81	1.81	7.6_{10}-12	2.1_{10}-13
	$_{10}$+2	1.71	1.71	3.6_{10}-10	4.9_{10}-12
	$_{10}$+3	1.6	1.6	8.4_{10}- 8	3.8_{10}-11
	$_{10}$+4	1.51	1.51	3.1_{10}- 6	4.0_{10}-10
	$_{10}$+5	1.29	1.39	1.0_{10}+0	1.2_{10}- 8
	$_{10}$+6	1.06	1.28	"	8.5_{10}- 8
	$_{10}$+7	1	1.20	"	2.1_{10}- 7
	$_{10}$+8	1	1.09	"	2.8_{10}- 6
	$_{10}$+9	1	1.01	"	1.8_{10}- 5
	$_{10}$+10	1	1	"	1.8_{10}- 5

table 1.

cond. nr.	k	ν (= avg. nr. iter. per col)		error in (QTQ - I)	
		CGSI	MGSCI	CGSI	MGSCI
$_{10}$+11	$_{10}$+8	1.96	1.96	6.1_{10}-14	3.0_{10}-14
	$_{10}$+9	1	1.48	1.0_{10}+0	1.7_{10}- 5
	$_{10}$+10	1	1.08	"	2.2_{10}- 4
	$_{10}$+11	1	1	"	2.3_{10}- 4

table 2.

5. DISCUSSION

We observed that for all matrices in all experiments the decomposition is accurate, which means that we always find matrices Q and R such that the norm of the residue, $\| A - QR \|_2$, is of the order of magnitude of the machine precision relatively to $\| A \|_2$, even in cases where Q is far from orthogonal.

For all matrices, with both the modified and the classical iterative Gram-Schmidt algorithm, the choice $\kappa = 2$ results in a matrix Q that is orthogonal to almost full precision; the condition of the matrix and the distribution of the singular values is only reflected in the number of columns that needs a second iteration to yield this good orthogonality.

For matrices that are not well conditioned, the orthogonality becomes worse with larger values of k for both CGSI and MGSCI.
If κ is given so large a value that no column of Q needs a second iteration, the results of the one-step classical and the one-step modified Gram-Schmidt respectively are produced. In this one-step situation we observe for the modified Gram-Schmidt algorithm that the orthogonality of Q (i.e. $\| Q^TQ - I \|_2$) is roughly equal to ε times the conditionnumber of the matrix, which confirms the bounds for the modified algorithm as given by Björck (cf. remark at the end of section 2). With the classical algorithm in that situation the results are bad; for every matrix with a conditionnumber larger than $1/\sqrt{\varepsilon}$ the result was $\| Q^TQ - I \|_2 \approx 1$.
Using the modified Gram-Schmidt algorithm we observe that for all values of κ the orthogonality is roughly bounded by $\kappa\,\varepsilon$.
This is according to theorem 5 applied with $\omega \approx \kappa\,\varepsilon$ (discarding the factor $\sqrt{n}$). In view of formula 3.2, we conjecture that the iterated modified Gram-Schmidt algorithm produces columns q_j that satisfy $\| Q_{j-1}{}^T q_j \|_2 \approx< \kappa\,\varepsilon$, also for large values of κ. This inspires us to the following conjecture:

CONJECTURE: If the iterated modified Gram-Schmidt algorithm is used with a value of $\kappa \geq 2$, then the resulting matrix Q satisfies: $\| Q^TQ - I \|_2 \approx< \kappa\,\varepsilon\sqrt{n}$.

For all matrices we tested, no matter its condition, this relation was fulfilled; we did not observe a relation of a similar sort for the classical Gram-Schmidt algorithm. This conjecture has the following application. For larger values of κ, the number of columns that need a second iteration may diminish. So, if the wanted accuracy is denoted by η (η should not be chosen smaller than $\approx 2 \times \varepsilon$), the factor κ may be chosen according to $\kappa = \text{maximum}\left(\frac{\eta}{\varepsilon\sqrt{n}}, 2\right)$.
Both CGSI and MGSCI can be used to solve the orthogonal basis problem. The operation count for these algorithms is νmn^2 flops.
The solution of the orthogonal basis problem with Householders method requires $2 \times (mn^2 - n^3/3)$ flops (see for instance Golub & Van Loan [5] p.152), so that for $\nu < 2 - (2n)/(3m)$ the iterative Gram-Schmidt algorithms require less operations.

6. CONCLUSIONS

Regarding the use of algorithms CGSI and MGSCI we come to the following conclusions.
For any small value of parameter κ ($\kappa = 2$, for example, will do) the orthogonality of the resulting matrix Q is of the order ε for both the modified and the classical iterative Gram-Schmidt algorithm.
Consequently, from a numerical point of view, there is no reason to prefer the iterative modified algorithm over the iterative classical algorithm **when full precision is wanted**.
Hence, a choice between these two algorithms can be made on considerations regarding efficient execution of the resulting code. For instance, we can use a fast matrix - vector multiplication routine in CGSI but not in MGSCI. This consideration clearly favours CGSI.

On the other hand, when **less than full precision is sufficient,** then the use of algorithm MGSCI may be advantageous; according to the conjecture stated in the previous section, the parameter κ can be given an optimal value to cut down on iterations.
In cases that the wanted accuracy η is much larger than ε, for example $\eta = \sqrt{\varepsilon}$, this strategy yields a rather large value for κ, which may result in considerable savings on the number of 'reorthogonalizations'. This is especially valuable for matrices with a bad condition.

ACKNOWLEDGEMENTS
The author wants to thank professor T.J. Dekker for his contribution to the proof of theorem 5.

LITERATURE
1. A. Björck, Solving linear least squares problems by Gram-Schmidt orthogonalization, *BIT* 7, 1-21 (1967).
2. P. Businger and G.H. Golub, Linear least squares solutions by Householder transformations, *Numer. Math.* 1, 269-276 (1965).
3. T.F. Chan, Rank revealing QR factorizations, *Linear Algebra Appl.* 88/89, 67-82 (1987).
4. J.W. Daniel, W.B. Gragg, L. Kaufman and G.W. Stewart, Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization, *Math. Comp.* 30,772-795 (1976).
5. G.H. Golub and C.F. van Loan, *Matrix Computations,* North Oxford Academic, Oxford 1983.
6. B.N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs N.J. 1980.
7. A. Ruhe, Numerical aspects of Gram-Schmidt orthogonalization of vectors, *Linear Algebra Appl.* 52/53, 591–601 (1983).