SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

AN ALGORITHM TO IMPROVE NEARLY ORTHONORMAL SETS OF VECTORS ON A VECTOR PROCESSOR*

BERNARD PHILIPPE†

Abstract. The symmetric orthogonalization, which is obtained from the polar decomposition of a matrix, is optimal. We propose an iterative algorithm to compute this orthogonalization on vector computers. It is especially efficient when the original matrix is near an orthonormal matrix.

Key words. polar decomposition, iterative method, square root, vector computer

AMS(MOS) subject classification. 65F25

Introduction. In the computation of the eigenvectors of a Hermitian matrix, it is necessary to check the orthonormality of the computed vectors, since for close eigenvalues there is an accompanying loss of orthogonality. Usually, especially when the vectors have been computed by inverse iteration, the Gram–Schmidt orthonormalization is performed on the groups of eigenvectors corresponding to close eigenvalues. If the residual is checked before and after this orthogonalization, a loss of accuracy appears. This should not be surprising since Gram–Schmidt orthogonalization corresponds to a QR factorization which depends on the ordering of the vectors. So, instead of a QR factorization, a polar decomposition seems to be preferred because it leads to an orthonormalization which is the best in some sense. This process has been called "Symmetric Orthogonalization" by Lowdin in [LO70].

In this paper, the optimal properties of symmetric orthogonalization are described in § 1. In this section it is also shown that, to orthonormalize a matrix A, it is sufficient to compute $A\,(A^*A)^{-1/2}$.

In § 2, an iterative scheme, which computes $S^{-1/2}$, where S is a Hermitian positive definite matrix, is analyzed and shown to be efficient on vector processors.

In § 3, the complete algorithm for the symmetric orthogonalization is given and experiments are presented.

1. Polar decomposition. In this section, the polar decomposition of a matrix and its application are described. This decomposition is a well-known factorization and a satisfactory presentation is given by Higham in [HA84].

THEOREM 1.1. *Let $A \in \mathbf{C}^{n\times p}$, $n \geqq p$. Then there exists a matrix $U \in \mathbf{C}^{n\times p}$ and a unique Hermitian positive semidefinite matrix $H \in \mathbf{C}^{p\times p}$ such that*

$$A = UH, \qquad U^*U = I_p.$$

If rank $(A) = p$ *then H is positive definite and U is uniquely determined.*

Proof. See [G59]. □

This factorization can be obtained directly from the singular value decomposition of the initial matrix. The SVD insures the existence of unitary matrices $P \in \mathbf{C}^{n\times n}$ and $Q \in \mathbf{C}^{p\times p}$ such that

$$P^*AQ = \underline{D} \tag{1.1}^1$$

---

* Received by the editors March 18, 1985; accepted for publication October 16, 1986.

† Institut de Recherche en Informatique et Systèmes Aléatoires, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France.

[1] Notation: $\underline{M}$ is the $n \times p$ matrix, obtained by the suppression of the $n - p$ last columns of the matrix $M \in \mathbf{C}^{n\times n}$.

with $D = \text{diag}(\sigma_1, \cdots, \sigma_p, 0, \cdots, 0) \in \mathbf{C}^{n\times n}$ where $0 \leqq \sigma_1 \leqq \cdots \leqq \sigma_p$. Because $A = P\underline{D}Q^*$, then

$$U = \underline{P}Q^* \quad \text{and} \quad H = QD'Q^* = (A^*A)^{1/2},$$

with $D' \in \mathbf{C}^{p\times p}$ and $D' = \text{diag}(\sigma_1, \cdots, \sigma_p)$. When the matrix A is of full rank, the factorization can be performed by using the following theorem.

THEOREM 1.2. *Let $A \in \mathbf{C}^{n\times p}$ with* rank $(A) = p \leqq n$. *Then the polar decomposition $A = UH$ is given by*

$$U = A(A^*A)^{-1/2} \quad \textit{and} \quad H = (A^*A)^{1/2}.$$

Proof. From the SVD (1.1), we have

$$(A^*A)^{-1/2} = QD'^{-1}Q^*$$

and

$$A(A^*A)^{-1/2} = P\underline{D}Q^*(QD'^{-1}Q^*)$$
$$= \underline{P}Q^*. \qquad \square$$

There is an algorithm [HB84] that computes $(A^*A)^{-1/2}$ and $(A^*A)^{1/2}$ simultaneously. Here, because we are only looking for the matrix U in the polar decomposition, we use the formulation of Theorem 1.2. Transforming a matrix A into the matrix U is an orthonormalization procedure which we call symmetric orthogonalization. This transformation is different from the usual one which corresponds to the QR factorization. In the following theorem the optimal properties of this symmetric orthogonalization are described.

THEOREM 1.3. *Let $A \in \mathbf{C}^{n\times p}$ with $p \leqq n$ and let $A = UH$ be a polar decomposition. Then*

$$\|A - U\| = \min_{Q\in\mathbf{U}} \|A - Q\|$$

where U *is the subset of all orthonormal matrices of $\mathbf{C}^{n\times n}$. This result is true for both the Euclidean norm and the Frobenius norm.*

Proof. For $p = n$, this result was proved by Fan and Hoffman in [FH55]. Its extension for $p \leqq n$ is straightforward.

2. Computation of the inverse of a square-root.

2.1. Scalar schemes. When Newton's method is used to find the positive root of the polynomial $f(x) = sx^2 - 1$, where $s > 0$, the iterative scheme obtained is

(I) given x_0, $\quad x_{m+1} = (1/2)(x_m + 1/(sx_m))$,

whereas if Newton's method is applied to the function $f(x) = 1/x^2 - s$ the scheme becomes

(II) given x_0, $\quad x_{m+1} = x_m + x_m(1 - sx_m^2)/2$.

When they are convergent, these schemes are quadratically convergent; let $e_m = x_m - s^{-1/2}$ be the error at step m. For (I) and (II) this quantity satisfies the following:

$$e_{m+1} = K_i e_m^2, \qquad i = \text{I, II}$$

with $K_{\text{I}} = 1/(2x_m)$ and $K_{\text{II}} = -s^{1/2}(s^{1/2}x_m + 2)/2$. Hence, close to the solution, the ratio of the convergence rates of the two schemes is equal to

$$K_{\text{II}}/K_{\text{I}} \simeq -3.$$

The domains of convergence for the two schemes are exhibited in the next result.

PROPOSITION 2.1. *For any positive numbers x_0 and s, the sequence $\{x_m\}$ defined by scheme* (I) *converges quadratically to $s^{-1/2}$.*

For any positive number s, the condition $0 < x_0 < \sqrt{3}s^{-1/2}$ insures that the sequence $\{x_m\}$ defined by scheme (II) *converges quadratically to $s^{-1/2}$.*

Proof. Let us consider the quantity $u_m = s^{1/2}x_m$; the convergence of the sequence $\{x_m\}$ to $s^{-1/2}$ is then equivalent to the convergence of the sequence $\{u_m\}$ to 1. So, the schemes become

$$(\mathrm{I}') \qquad \text{given } u_0 = x_0 s^{1/2}, \qquad u_{m+1} = (u_m + 1/(u_m))/2$$

and

$$(\mathrm{II}') \qquad \text{given } u_0 = x_0 s^{1/2}, \qquad u_{m+1} = u_m + u_m(1 - u_m^2)/2.$$

The function $u \to g(u) = (u + 1/u)/2$ transforms the interval $(0, +\infty)$ into the interval $[1, +\infty)$ and satisfies the following:

$$u > 1 \text{ implies } 0 < g(u) - 1 = (u-1)^2/(2u) < (u-1)/2.$$

The last inequality proves that scheme (I') is always convergent.

The function $u \to g(u) = u + u(1 - u^2)/2$ transforms the interval $(0, \sqrt{3})$ into the interval $(0, 1]$. If we consider u such that $0 < u < 1$ then

$$0 < 1 - g(u) = (u+2)(1-u)^2/2 < 1 - u.$$

So if $0 < u_0 < \sqrt{3}$ scheme (II') converges. □

In this situation it appears that scheme (I) must be preferred to scheme (II). The generalization of scheme (II) to the matrix situation is much more interesting, since its computation is expressed with matrix multiplications. Moreover, the differences in convergence between (I) and (II) are not as great as in the scalar case.

2.2. Matrix schemes. Let S be a Hermitian positive definite matrix of order p and let $0 < s_1 \leqq \cdots \leqq s_p$ be its eigenvalues. First of all we remark that the only schemes to be considered are those which correspond to the application of the scalar schemes in every eigendirection when the initial guess commutes with S. Because we are only interested in polynomial schemes, we consider the following schemes that are based on (II) of the scalar case:

$$(\Sigma_a) \qquad \text{given } T_0, \qquad T_{m+1} = T_m + \alpha T_m(I - T_m S T_m) + \beta(I - T_m S T_m)T_m$$

where α and β are two nonnegative parameters satisfying $\beta = \frac{1}{2} - \alpha$. The quantity $Z_m = I - T_m S T_m$ is called the residual at step m.

THEOREM 2.2. *Let $K(S) > 1$ be the condition number of S (ratio of the extremal eigenvalues). If $K(S) < 17 + 6\sqrt{8}$ then $S^{-1/2}$ is a point of attraction of the iteration $(\Sigma_{1/4})$; this condition becomes $K(S) < 9$ for the iterations (Σ_0) or $(\Sigma_{1/2})$.*

Proof. $V = S^{-1/2}$ is a fixed point of the polynomial

$$F_a : T \to T + \alpha T(I - TST) + \beta(I - TST)T.$$

Let us compute its differential application at V. If $T = V + W$ then

$$\begin{aligned} I - TST &= -VSW - WSV + O(W^2) \\ &= -V^{-1}W - WV^{-1} + O(W^2). \end{aligned}$$

Hence

$$T(I - TST) = -W - VWV^{-1} + O(W^2), \qquad (I - TST)T = -V^{-1}WV - W + O(W^2).$$

So, for every matrix W

$$F_\alpha(V+W)-F_\alpha(V)=(1/2)W-\alpha VWV^{-1}-\beta V^{-1}WV+O(W^2).$$

Hence the differential application is given at V by

$$F'_\alpha(V)W=(1/2)W-\alpha VWV^{-1}-\beta V^{-1}WV.$$

To use Ostrowsky's Theorem [OR70], it is necessary to prove that the spectral radius of $F'_\alpha(V)$ is smaller than 1. By using a similarity transformation, we may assume that the matrix V is diagonal:

$$V=\text{diag}\,(1/\sqrt{s_1},\cdots,1/\sqrt{s_p}).$$

Then it can be proved that the spectrum of $F'_\alpha(V)$ is the set

$$\sigma(F'_\alpha(V))=\{\mu_{ij}|\mu_{ij}=1/2-\alpha\sqrt{s_i/s_j}-\beta\sqrt{s_j/s_i},\quad i,j=1,n\}.$$

Let λ be any $\sqrt{s_i/s_j}$. We look for the largest interval $\mathbf{I}$ such that if $\lambda\in\mathbf{I}$ and $1/\lambda\in\mathbf{I}$ then $|1/2-\alpha\lambda-\beta/\lambda|<1$. It is easy to see that this is equivalent to solving the system

$$\alpha\lambda^2-3\lambda/2+\beta<0,\qquad \beta\lambda^2-3\lambda/2+\alpha<0. \tag{2.1}$$

If $\alpha=\beta=1/4$ then (2.1) is equivalent to

$$\lambda^2-6\lambda+1<0$$

and then $\mathbf{I}=(1/\lambda_0,\lambda_0)$ with $\lambda_0=3+\sqrt{8}$. Hence $\sigma(F'_{1/4}(V))\subset\mathbf{I}$ is equivalent to $K(S)<(3+\sqrt{8})^2=17+6\sqrt{8}$.

If $\alpha=1/2$ and $\beta=0$ then (2.1) is equivalent to

$$\lambda^2-3\lambda<0,\qquad 1-3\lambda<0$$

and then $\mathbf{I}=(1/3,3)$. Hence $\sigma(F'_{1/2}(V))\subset\mathbf{I}$ is equivalent to $K(S)<9$.

In the same way, the reader can prove that $\sigma(F'_0(V))\subset\mathbf{I}$ is equivalent to $K(S)<9$. □

Remark 2.3. (i) If T_0 is Hermitian, then scheme $(\Sigma_{1/4})$ can be expressed in a better way by

$$\begin{aligned}&\text{given } T_0,\\ &T'_{m+1}=T_m+(1/2)T_m(I-T_mST_m),\\ &T_{m+1}=(1/2)(T'^*_{m+1}+T'_{m+1}).\end{aligned}$$

This expression proves that $(\Sigma_{1/4})$ is actually equivalent to using $(\Sigma_{1/2})$ and adding a symmetrization at every step. This formulation is cheaper in terms of operation count than the original one.

(ii) Considering the scheme

$$\text{given } T_0,\qquad T_{m+1}=(1/2)(T_m+(ST_m)^{-1})$$

which is based on the scalar scheme (I), the associated function G defined by

$$G\colon T\to(1/2)(T+(ST)^{-1})$$

has the same differential application as the $(F_0)'$. So, the local convergence of this scheme is only insured if $K(S)<9$. This scheme has been studied by Laasonen in [LA58].

THEOREM 2.4. *Let $\rho(S)$ be the spectral radius of S. If $\mu<(3/\rho(S))^{1/2}$ then the scheme*

$$T_0=\mu I,\qquad T_{m+1}=T_m+(1/2)T_m(I-T_mST_m) \tag{$\bar{\Sigma}$}$$

is quadratically convergent.

Moreover, if $K(S) < 9$ *then this scheme is locally stable. This condition can be weakened into* $K(S) < 17 + 6\sqrt{8}$ *if a symmetrization is performed at every step on* T_m.

Proof. By induction, it is clear that every iterate T_m is Hermitian and commutes with S. Because the subspace of the matrices commuting with S is included in the kernel of the differential application which is defined in Theorem 2.2 then scheme $(\bar{\Sigma})$ has a quadratic convergence as soon as it is convergent. In this situation, the scheme is equivalent to using the scalar scheme (II) in every eigendirection. Using the initial guess μ to compute $s_i^{-1/2}$, $i = 1, p$ with the scalar scheme (II) the conditions

$$\mu < \sqrt{3} s_i^{-1/2}, \qquad i = 1, p$$

must be true to insure convergence. These are also sufficient conditions (see Proposition 2.1). So, the first result of the theorem is proved.

If we assume now that this scheme is perturbed by rounding errors, we can no longer insure that T_m commutes with S. The condition $K(S) < 9$ (or $K(S) < 17 + 6\sqrt{8}$ if a symmetrization of T_m occurs at every step) is sufficient to insure that a perturbation due to rounding errors will decrease in the succeeding steps at least in a neighborhood of the solution, since $S^{-1/2}$ is a point of attraction of the iteration (Theorem 2.2). □

PROPOSITION 2.5. *The residual of scheme* $(\bar{\Sigma})$, *not considering rounding errors, satisfies*

$$Z_{m+1} = (3/4)Z_m^2 + (1/4)Z_m^3. \tag{2.2}$$

Proof.

$$T_{m+1}^2 = (T_m + (1/2)T_m Z_m)^2$$

$$= T_m^2 + (1/4)T_m^2 Z_m^2 + T_m^2 Z_m.$$

Hence

$$Z_{m+1} = I - ST_m^2 - (1/4)ST_m^2 Z_m^2 - ST_m^2 Z_m.$$

When we use $-ST_m^2 = Z_m - I$, the result of the proposition follows. □

Remark 2.6. Formula 2.2 appears to be of interest because it can split the computation of T_{m+1} into two tasks since Z_{m+1} can be evaluated from Z_m only. However, this formula cannot be used repeatedly without updating the residual from its definition $Z_m = I - T_m S T_m$.

3. Computation of the symmetric orthogonalization.

3.1. Application of scheme $(\bar{\Sigma})$. Let us come back to the matrix $A \in \mathbf{C}^{n \times p}$, assuming rank $(A) = p \leqq n$. To orthogonalize this matrix with a symmetric orthogonalization, it is necessary to compute $S^{-1/2}$ where $S = A^*A$ (§ 1). To insure the stability of $(\bar{\Sigma})$ the condition number of S is assumed to be smaller than $(17 + 16\sqrt{8})$.

In order to define an initial guess, the spectral radius $\rho(S)$ of the matrix S has to be estimated. In fact, the ∞-norm is used instead of this spectral radius. From Theorem 2.4 a number μ is then computed:

$$\mu = \sqrt{3/\|S\|_\infty} \leqq \sqrt{3/\rho(S)}.$$

By choosing $T_0 = \mu I$, we ensure the convergence of scheme $(\bar{\Sigma})$. The first iteration can be skipped since it is easy to compute the following:

$$T_1 = (3/2)\mu I - (1/2)\mu^3 S.$$

However, if the matrix $\Delta = S - I$ is small (i.e., $\rho(\Delta) < 1$), the initial guess can be much closer to the solution by choosing the Taylor approximation of order k of $(I + \Delta)^{-1/2}$

$$T_0 = I + \sum_{i=1}^{k} (-1)^i \binom{-1/2}{i} \Delta^i.$$

After m iterations the magnitude of the error is given by

$$T_m - S^{-1/2} = O(\Delta^{(k+1)2^m}).$$

The computation of T_m involves $(k - 1) + 3m$ matrix multiplications. Then, the best order to choose is always smaller than 5. For a required precision ε, an estimation of the best order of the Taylor approximation is given by the author in [PH85] and depends on the ratio $(\log \varepsilon / \log \|\Delta\|_\infty)$.

This algorithm is related to the algorithm which is described in [BB71]: here the matrix $T = (A^*A)^{-1/2}$ is computed before performing the multiplication $A \times T$, hence the iterative part of the algorithm is in $O(p^3)$ flops while it was in $O(np^2)$ in [BB71]. Moreover, the introduction of a symmetrization on the iterate at every stage improves the stability when needed.

3.2. Algorithm. Summarizing the previous considerations, we have the following algorithm.

begin
  $S := A^* \times A$;
  $\Delta := I - S$;
  $\delta := \|\Delta\|_\infty$;
  if $(\delta < \varepsilon)$ **then**
      nothing to do ;
    elseif $(\delta < 1)$ **then**
      $k :=$ Taylor approximation order ;
      $T :=$ Taylor approximation of order k ;
      sym := false ;
    else
      $\mu = \sqrt{3/\delta}$;
      $T := (3/2)\mu I - (1/2)\mu^3 S$;
      sym := true ;
    endif ;
    iter := 0 ;
      loop :
        $\delta 0 := \delta$;
        $Z := I - T \times S \times T$;
        $\delta := \|Z\|_\infty$;
        if $(\delta < \varepsilon)$ **then** exit of the loop **endif** ;
        if $(\delta > \delta 0)$ **then** divergence **endif** ;
        iter := iter + 1 ;
        $T := (1/2)T \times (2I + Z)$;
        if (sym) **then** $T := (1/2)(T^* + T)$ **endif** ;
      endloop ;
  $A := A \times T$;
end.

TABLE 1
Symmetric orthogonalization on CRAY 1.

$\delta = \|\bar{Q}^T\bar{Q} - I\|_\infty$ before orthog.	With symmetriz.	Taylor order	# of iteration(s)	Elapsed time (unity: 10^{-1} s)
$0.24\,E-3$	No	3	0	0.35
$0.41\,E-3$	No	3	1	0.45
$0.54\,E-2$	No	2	1	0.42
$0.22\,E-1$	No	3	1	0.45
$0.81\,E-1$	No	4	2	0.58
0.39	No	4	3	0.69
$0.27\,E+1$	Yes		14	1.72
$0.34\,E+1$	Yes		7	1.01

To measure the cost of the computation, it is assumed that $A \in \mathbf{R}^{n\times p}$. Following [GVL83] a flop is defined as the amount of computation involved in a triad: $a := a + b \times c$. Then the cost of:

$S := A^T \times A$ is p^2n flops (or $1/2p^2n$ flops if symmetry of S is taken into account),

one iteration of $(\bar{\Sigma})$ is $\approx 3p^3$ flops,

$A := A \times T$ is p^2n flops.

The cost of the Gram–Schmidt orthogonalization is $p^2n + np$ flops. So if $n \gg p$ the symmetric orthogonalization is about twice as expensive as the Gram–Schmidt process, but it is based only on matrix multiplications. If $n = p$ the symmetric orthogonalization becomes more expensive for the computation of $T^{-1/2}$.

An alternative way to compute the symmetric orthogonalization would be to perform the SVD of A or to diagonalize S. In both cases, the number of flops is larger (see [PH85]). Moreover these algorithms are much more difficult to vectorize.

3.3. Experiments. In this section, the results of experiments on a CRAY 1[2] are discussed. An orthogonal matrix $Q \in \mathbf{R}^{201\times 61}$ was constructed from a unitary vector u by $Q = I - 2uu^T$. This matrix Q was randomly perturbed into a matrix $\bar{Q}$ whose column vectors were still normalized (to be in a situation similar to when finding eigenvectors). Both orthogonalizations (symmetric and Gram–Schmidt) were performed on $\bar{Q}$. For symmetric orthogonalization, the results for different magnitudes of perturbation are exhibited in Table 1. For each run, the algorithm is defined by the value of the quantity δ. If δ is smaller than 1 then the initial guess is obtained by a Taylor expansion whose order is given in Table 1. If δ is larger than 1 then the initial guess is μI, where μ is computed from δ (§ 3.2). In this last case, a symmetrization on the iterate occurs at every stage.

After orthogonalization, the residual $\|\bar{Q}^T\bar{Q} - I\|_\infty$ was always in the range $[10^{-13}, 10^{-12}]$. The elapsed time for the Gram–Schmidt process was 0.59×10^{-1} s.

For each run, the distance between the perturbed matrix and its orthogonalized matrix was very close to the residual given in the first column of Table 1 for Gram–Schmidt. For the symmetric orthogonalization, the distance was only half of this residual. Some cases of divergence were obtained with perturbation of larger magnitude. These cases correspond to matrices $\bar{Q}^T\bar{Q}$ with a small eigenvalue which implies a large condition number. In these situations, the solution was almost reached before the rounding errors became important because of increasing magnitude at every iteration.

[2] This CRAY 1 is managed by the Conseil Scientifique du Centre de Calcul Vectoriel pour la Recherche, Palaiseau, France.

Conclusion. Even when the result of a computation should be an orthonormal set of vectors (e.g., for the eigenvectors of a Hermitian matrix), there is often a loss of orthogonality which occurs due to rounding errors. In this situation the orthogonalization process should preserve the quality of the original set. As has been proved, the symmetric orthogonalization is optimal. The iterative scheme which is proposed in this paper is efficient on vector processors since it uses only matrix multiplications. This scheme is numerically stable when the ratio of the extremal singular values is smaller than $3 + \sqrt{8}$.

Acknowledgment. The author would like to thank the referees for their valuable criticism of a first version of this paper. He is also grateful to B. Parlett and A. Sameh for their helpful remarks.

REFERENCES

[BB71] A. BJORCK AND C. BOWIE, *An iterative algorithm for computing the best estimate of an orthogonal matrix*, SIAM J. Numer. Anal., 8 (1971), pp. 358–364.

[FH55] K. FAN AND A. HOFFMAN, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[G59] F. R. GANTMACHER, *The Theory of Matrices*, Volume One, Chelsea, New York, 1959.

[GVL83] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[HA84] N. J. HIGHAM, *Computing the polar decomposition—with applications*, Numerical Analysis Report No. 94, Univ. of Manchester, Manchester, England, 1984.

[HB84] ———, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–550.

[LA58] P. LAASONEN, *On the iterative solution of the matrix equation* $AX^2 - I = 0$, Math. Tables Aids Comput., 12 (1958), pp. 109–116.

[LO70] P. LOWDIN, *Advances in Quantum Chemistry*, Vol. 5, Academic Press, New York, 1970.

[OR70] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[PA80] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ 1980.

[PH85] B. PHILIPPE, *Approximating the square root of the inverse of a matrix*, Cedar document No. 108, CSRD, Univ. of Illinois, Urbana, IL 1985.