Electronic Transactions on Numerical Analysis.
Volume 47, pp. 127–152, 2017.

ISSN 1068–9613.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

AN OPTIMAL Q-OR KRYLOV SUBSPACE METHOD FOR SOLVING LINEAR SYSTEMS*

GÉRARD MEURANT†

Abstract. Today the most popular iterative methods for solving nonsymmetric linear systems are Krylov methods. In this paper we show how to construct a non-orthogonal basis of the Krylov subspace such that the quasi-orthogonal residual (Q-OR) Krylov method using this basis yields the same residual norms as GMRES up to the final stagnation phase, provided GMRES is not stagnating. In many examples this new Krylov method gives a better maximum attainable accuracy than GMRES with a modified Gram-Schmidt (MGS) implementation. Even though the number of floating point operations per iteration is larger than for GMRES, the optimal Q-OR method offers more potential for parallelism than GMRES with MGS.

Key words. linear systems, Krylov methods, Q-OR algorithm

AMS subject classifications. 65F10

1. Introduction.

We consider the problem of solving linear systems $Ax = b$, where A is a square nonsingular matrix of order n with real entries and b is a real vector of length n. Today the most popular iterative methods for solving such nonsymmetric linear systems are Krylov methods. Assuming the initial guess is zero, they use Krylov subspaces $\mathcal{K}_k(A, b) \equiv \text{span}\{b, Ab, A^2b, \ldots, A^{k-1}b\}$, $k = 1, 2, \ldots$, of growing dimension that are defined by repeated multiplication of the right-hand side b with the matrix A. The approximations x_k to the solution of the linear system are extracted from these subspaces.

Two well-known Krylov methods are FOM (Full Orthogonalization Method) [17] and GMRES (Generalized Minimum RESidual method) [18]. These two methods use an orthonormal basis of the Krylov subspace that is computed by the Arnoldi process [1]. GMRES minimizes the ℓ_2-norm of the residual vector at each iteration and is, in this sense, an optimal method among those using only one matrix-vector product per iteration. In FOM, the residual vectors are orthogonal but the residual norms are larger than or equal to the GMRES residual norms. In [4] it is shown that many Krylov methods can be described as so-called quasi-orthogonal (Q-OR) or quasi-minimum (Q-MR) residual methods. Well-known examples are FOM/GMRES, BiCG/QMR [5, 6], and Hessenberg/CMRH [10, 19]. Here we use a dash in Q-OR and Q-MR for the general methods that use any basis of the Krylov subspace to distinguish them from the QMR method proposed in [6]. All these pairs of methods differ mainly by the basis of the Krylov subspace that is used. FOM/GMRES use an orthogonal basis, BiCG/QMR use a biorthogonal basis, and Hessenberg/CMRH use a basis originating from the LU factorization with pivoting of the Krylov matrix. However, one can use any basis of the Krylov subspace in the Q-OR/Q-MR methods.

The main goal of this paper is to show that one can construct a non-orthogonal basis of the Krylov subspace such that the corresponding Q-OR method yields the same residual norms as GMRES when they are started from the same initial vector (except in case of GMRES stagnation); that is, one can construct what can be considered an optimal Q-OR method. Moreover, we show that this non-orthogonal basis has many interesting mathematical properties.

The content of the paper is as follows. In Section 2 we recall the construction and the properties of general Q-OR methods using any basis of the Krylov subspace. Section 3 is devoted to the construction of a basis for which Q-OR will deliver the same residual norms

*Received January 18, 2017. Accepted August 15, 2017. Published online on October 10, 2017. Recommended by K. Jbilou.
†30 rue du sergent Bauchat, 75012 Paris, France (gerard.meurant@gmail.com).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

as GMRES provided GMRES is not stagnating. In the first subsection it is shown that this goal can be reached, but the corresponding algorithm is not practical since it relies on matrices which are numerically badly conditioned. The second subsection establishes some technical results that will be needed to solve a minimization problem related to the construction of the basis. The third subsection shows how to reliably construct the optimal basis and the upper Hessenberg matrix from which the approximate solution is computed. In Section 4 we study the mathematical properties of the optimal basis. Remarkably, the angles between the basis vectors can be explicitly written as functions of the residual norms. Moreover, a lower bound on the smallest singular value of the matrix whose columns are the basis vectors can be obtained. Some of these properties are also helpful to simplify the implementation of the method. Section 5 provides an implementation of the method. In Section 6 we report some numerical experiments showing that the method delivers the same residual norms as GMRES-MGS, that is, GMRES implemented with the modified Gram-Schmidt algorithm, but, in many cases, with a better maximum attainable accuracy. In Section 7 we propose a simple technique to handle the breakdowns that could occur in the Q-OR optimal method when GMRES is stagnating. Finally, we give some conclusions and perspectives.

2. Q-OR Krylov methods. For simplicity we assume that all the matrices have full rank (except H_k defined below), but the results are also valid when there is an early termination. In particular, the Krylov matrix defined as

$$K = \begin{bmatrix} b & Ab & A^2b & \cdots & A^{n-1}b \end{bmatrix}.$$

is a nonsingular matrix. Without loss of generality we will also assume that $\|b\| = 1$ and that the first iterate is $x_0 = 0$.

We first consider abstract Q-OR methods regardless of the basis that is used. Let us assume that we have an ascending basis v_k, $k = 1, \ldots, n$, of unit norm vectors for $\mathcal{K}_n(A, b)$ with $v_1 = b$. This means that $\{v_1, \ldots, v_k\}$ is a basis of $\mathcal{K}_k(A, b)$ for all $k \leq n$. The unit norm basis vectors are not necessarily orthonormal to each other, but, of course, they are linearly independent. Let V be the matrix whose columns are the basis vectors v_k, $k = 1, \ldots, n$. The matrix V is nonsingular, and there exists a nonsingular upper triangular matrix U (which is the matrix representing the change of basis) such that

$$K = VU. \tag{2.1}$$

Let C be the companion matrix associated with the characteristic polynomial of A denoted by

$$C = \begin{bmatrix} 0 & \cdots & 0 & -\alpha_0 \\ & I_{n-1} & & \vdots \\ & & & -\alpha_{n-1} \end{bmatrix}, \tag{2.2}$$

where I_{n-1} is the identity matrix of order $n-1$. The roots of the monic polynomial with coefficients $\alpha_{n-1}, \ldots, \alpha_0$, where α_0 is the constant coefficient, are the eigenvalues of A. From [3, Theorem 2.1] we know that $H = UCU^{-1}$ is an unreduced upper Hessenberg matrix and that $AV = VH$. This gives an Arnoldi-like relation that is called a Krylov decomposition in [20]; see relation (2.3) below.

Let us now define the Krylov methods we are considering. We proceed as in [4]. Since, without loss of generality, we have chosen a zero starting vector $x_0 = 0$, we define the iterates x_k as

$$x_k = V_k y^{(k)},$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where V_k is the matrix of the k first columns of V. This means that we look for x_k in $\mathcal{K}_k(A,b)$. We have the Arnoldi-like relation

$$AV_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^T = V_{k+1} \underline{H}_k, \tag{2.3}$$

where H_k is the principal submatrix of order k of H, $\underline{H}_k$ is the same matrix appended with the k first entries of the $(k+1)$st row of H, and e_k is the last column of the identity matrix of order k. Thus, the residual vector r_k can be written as

$$r_k = b - Ax_k = V_k e_1 - AV_k y^{(k)} = V_k(e_1 - H_k y^{(k)}) - h_{k+1,k} y_k^{(k)} v_{k+1}, \tag{2.4}$$

where e_1 is the first column of the identity matrix of order k.

The kth iterate $x_k^O = V_k y^{(k)}$ of a Q-OR method is defined (provided that H_k is nonsingular) by computing $y^{(k)}$ as the solution of the linear system

$$H_k y^{(k)} = e_1. \tag{2.5}$$

This annihilates the first term in the rightmost expression of (2.4). Thus, the kth iterate of the Q-OR method is $x_k^O = V_k H_k^{-1} e_1$. Moreover, the residual vector, which we denote as r_k^O, is proportional to v_{k+1}, and $\|r_k^O\| = |h_{k+1,k} y_k^{(k)}|$. In case H_k is singular and x_k^O is not defined, we shall define the residual norm to be infinite, $\|r_k^O\| = \infty$. The vector $y^{(k)}$, i.e., the solution of equation (2.5), is usually computed by transforming the subdiagonal entries of H_k to zero using Givens rotations and then solving an upper triangular system. Note that if b is not of unit norm or if $x_0 \neq 0$, we have to solve $H_k y^{(k)} = \|r_0^O\| e_1$.

In a Q-MR method, the vector $y^{(k)}$ is computed as the solution of the minimization problem

$$\min_y \|e_1 - \underline{H}_k y\|.$$

The Q-MR iterates are always defined contrary to the Q-OR iterates. An optimal Q-MR method is clearly GMRES since it minimizes the residual norm because, in this case, the matrices V_j, $j = 1, \ldots, n$, are orthonormal. However, in this paper we concentrate on Q-OR methods for which we have the following result.

THEOREM 2.1 ([3]). *Let* $\begin{bmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,n} \end{bmatrix}$ *with* $\nu_{1,1} = 1$ *be the first row of* U^{-1}, *the matrix* U *being defined by relation* (2.1). *The entries* $\nu_{1,k+1}$, $k = 0, \ldots, n-1$, *satisfy*

$$|\nu_{1,k+1}| = \frac{1}{\|r_k^O\|},$$

where r_k^O *are the Q-OR residual vectors obtained with the basis* V.

In other words, whatever is the chosen basis, the norms of the residual vectors of the Q-OR method can be read from the first row of the inverse of the upper triangular matrix U that describes the relation between the natural basis and the basis we are using. In the next section we will use this result to construct a basis for which Q-OR will deliver the same residual norms as GMRES. Note that when $\|r_0^O\| \neq 1$, we have $|\nu_{1,k+1}| = \|r_0^O\| / \|r_k^O\|$.

3. An optimal basis for Q-OR. Obviously, from Theorem 2.1, it would be nice to construct the basis in such a way that we have the largest possible values of $|\nu_{1,j}|$, $j = 2, \ldots,$ in the first row of U^{-1}. However, we have the constraint that the basis vectors v_j must be of unit norm. We first construct the basis using the matrix U^{-1} because it is not too difficult to show that, if there is no breakdown, this basis is optimal in the sense that we obtain the same residual norms as GMRES. We do this informally because we will see that, numerically, this is far from being practical and efficient, and we will then look for a more reliable equivalent algorithm.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

3.1. Construction of the basis using U^{-1}. We may directly compute the basis of $\mathcal{K}_n(A,b)$ from the relation $V = KU^{-1}$ obtained from (2.1). In this way we obtain the vectors v_j straightforwardly, but, as we said above, the columns of V have to be of unit norm for the result of Theorem 2.1 to be valid. We denoted by $\nu_{i,j}$ the entries of U^{-1}. Let ν_k be the vector of components $\nu_{i,k}, i = 1, \ldots, k$, the k first components of the kth column of U^{-1}, and let K_k be the matrix of the first k columns of the Krylov matrix K. Using the relation $V = KU^{-1}$, we define

$$\tilde{v}_k = \nu_{1,k} b + \nu_{2,k} A^2 b + \cdots + \nu_{k,k} A^{k-1} b = K_k \nu_k.$$

We would like to have $\|\tilde{v}_k\| = 1$ and $|\nu_{1,k}|$ as large as possible. Then, we can define the basis vector as $v_k = \tilde{v}_k$. This means that we have the constraint $\|K_k \nu_k\| = 1$, which corresponds to

$$\nu_k^T K_k^T K_k \nu_k = \nu_k^T \mathcal{M}_k \nu_k = 1.$$

Since the matrix $\mathcal{M}_k = K_k^T K_k$ is symmetric and positive definite, this is the equation of an (hyper) ellipsoid in $\mathbb{R}^k$ centered at the origin. All the points (that is, vectors) on the surface of the ellipsoid satisfy the constraint. To maximize $|\nu_{1,k}|$ we have to find a point on the surface of this ellipsoid attaining a maximum of the absolute value of the first coordinate.

For simplicity, let us consider the case $k = 3$. A way to solve the problem is to use projective geometry and the fact that an ellipsoid is a transformation of a sphere. A sphere of radius 1 centered at the origin is represented by a matrix S of order 4,

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

A point p is defined as $p = \begin{bmatrix} x & y & z & 1 \end{bmatrix}^T$, and the equation of the unit sphere is $p^T S p = 0$. We apply a transformation T from the sphere to the ellipsoid and obtain

$$(T^{-1}p)^T S (T^{-1}p) = p^T T^{-T} S T^{-1} p = 0.$$

Hence, the equation of the ellipsoid is $p^T W p = 0$ with $W = T^{-T} S T^{-1}$. Now, we are looking for a plane tangent to the ellipsoid and orthogonal to the x-axis. In our coordinate system, a plane is represented by a vector u such that $u^T p = 0$. We recognize that for a point p on the ellipsoid, the plane defined by $u^T = p^T W$ touches the ellipsoid at p since

$$u^T p = p^T W p = 0.$$

It can be shown that this is the only intersection point. The tangent plane is characterized by

$$u^T W^{-1} u = p^T W W^{-1} W p = p^T W p = 0.$$

In this system of coordinates, the equation of our ellipsoid is

$$p^T \begin{bmatrix} \mathcal{M}_3 & 0 \\ 0 & -1 \end{bmatrix} p = 0.$$

Hence,

$$W^{-1} = \begin{bmatrix} \mathcal{M}_3^{-1} & 0 \\ 0 & -1 \end{bmatrix}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

A plane orthogonal to the x-axis is defined by $u^T = \begin{bmatrix} 1 & 0 & 0 & -x \end{bmatrix}$. Therefore, we must have

$$\begin{bmatrix} 1 & 0 & 0 & -x \end{bmatrix} \begin{bmatrix} \mathcal{M}_3^{-1} & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ -x \end{bmatrix} = 0.$$

This yields

$$\begin{bmatrix} 1 & 0 & 0 & -x \end{bmatrix} \begin{bmatrix} (\mathcal{M}_3^{-1})_{1,1} \\ (\mathcal{M}_3^{-1})_{2,1} \\ (\mathcal{M}_3^{-1})_{3,1} \\ x \end{bmatrix} = (\mathcal{M}_3^{-1})_{1,1} - x^2 = 0.$$

Finally, we obtain the first component of the solution by $x = \pm\sqrt{(\mathcal{M}_3^{-1})_{1,1}}$. Note that $\mathcal{M}_3 = K_3^T K_3$ is positive definite, and thus the diagonal entries of its inverse are positive.

However, this does not give the other components of the solution, that is, the other components of a column of U^{-1} which are necessary to compute the columns of H. But, we can compute them using the tangent plane. Let us write the matrix $\mathcal{M}_3$ as

$$\mathcal{M}_3 = \begin{bmatrix} 1 & \hat{\alpha} & \hat{\gamma} \\ \hat{\alpha} & \hat{\beta} & \hat{\delta} \\ \hat{\gamma} & \hat{\delta} & \hat{\omega} \end{bmatrix}.$$

The gradient at (x, y, z) is

$$g = 2 \begin{bmatrix} x + \hat{\alpha}y + \hat{\gamma}z \\ \hat{\alpha}x + \hat{\beta}y + \hat{\delta}z \\ \hat{\gamma}x + \hat{\delta}y + \hat{\omega}z \end{bmatrix}.$$

We obtain the equations

$$\begin{aligned} (x + \hat{\alpha}y + \hat{\gamma}z)x &= 1, \\ \hat{\alpha}x + \hat{\beta}y + \hat{\delta}z &= 0, \\ \hat{\gamma}x + \hat{\delta}y + \hat{\omega}z &= 0. \end{aligned}$$

But, we already know the first component x. The other components are found by solving the linear system

$$\begin{bmatrix} \hat{\beta} & \hat{\delta} \\ \hat{\delta} & \hat{\omega} \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = -x \begin{bmatrix} \hat{\alpha} \\ \hat{\gamma} \end{bmatrix}.$$

Note that the matrix of this system is the trailing principal submatrix of $\mathcal{M}_3$, and the vector on the right-hand side is constructed from the last two entries of the first column. The matrix is nonsingular, but it may happen that the right-hand side is zero, in which case the matrix that we denoted by "U^{-1}" will be singular!

This construction can be straightforwardly generalized to any dimension. Choosing the positive solution, we obtain $x = \sqrt{(\mathcal{M}_k^{-1})_{1,1}}$, and the other components are computed by solving a linear system of order $k-1$ whose matrix and right-hand side are $\mathcal{M}_{2:k,2:k}$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

and $-x\mathcal{M}_{2:k,1}$. In theory, when we have a nonsingular upper triangular matrix U^{-1}, we can compute the basis vectors using $V = KU^{-1}$. The upper Hessenberg matrix H can be computed from U using a result proved in [3]. The submatrices of H can be written as $H_k = U_k C^{(k)} U_k^{-1}$, where U_k is the principal submatrix of U and $C^{(k)}$ is a companion matrix whose last column is $U_k^{-1} U_{1:k,k+1}$. Hence, H_k is nonsingular if U_k^{-1} is nonsingular, that is, all the diagonal entries are nonzero. But, even though we can compute all the columns of U^{-1}, we have no guarantee that the matrix "U^{-1}" is numerically nonsingular; if it is (close to be) singular we have a (near) breakdown of the algorithm.

Unfortunately, this breakdown situation occurs if there is stagnation in GMRES. The real matrices leading to GMRES stagnation have been characterized in [14]. They can be written as $A = ZQRZ^T$, where Z is an orthonormal matrix, R is upper triangular, and Q is an orthonormal matrix such that parts of some columns and of some rows are zero, depending at which iterations the stagnation of the residual norms happens. The right-hand side giving stagnation is $b = Ze_1$. Let us consider, for instance, an initial stagnation of the GMRES residual norms, $\|r_0^G\| = \|r_1^G\| = 1$. Then, the first column of the matrix Q is zero except for the second component, which is equal to 1, that is, $Qe_1 = e_2$; see [14]. Then, we have

$$K_2 = \begin{bmatrix} Ze_1 & AZe_1 \end{bmatrix} = Z \begin{bmatrix} e_1 & QRe_1 \end{bmatrix} = Z \begin{bmatrix} e_1 & \delta Qe_1 \end{bmatrix},$$

with $\delta = (R)_{1,1}$. Therefore, we obtain

$$\mathcal{M}_2 = K_2^T K_2 = \begin{bmatrix} 1 & \delta e_1^T Q e_1 \\ \delta e_1^T Q^T e_1 & \delta^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \delta^2 \end{bmatrix}.$$

The matrix $\mathcal{M}_2$ is diagonal as well as its inverse. Using this, the first two components of the second column of "U^{-1}" that we are looking for are 1 and 0, and the upper triangular matrix we are constructing is singular. This reminds of the relations between FOM and GMRES: when GMRES stagnates, the FOM iterates are not defined. More generally, we obtain zero or small entries on the diagonal of the columns corresponding to the iterations where the GMRES residual norm is (almost) equal to the norm of the previous iteration.

Of course, even when there is no breakdown, this algorithm is not practical since we need to know the moment matrices $\mathcal{M}_k = K_k^T K_k$, and it is well known that, numerically, the column vectors of K_k may loose their linear independence. Moreover, if we are in a case for which we have fast convergence of the Q-OR method, the matrices $\mathcal{M}_k$ tend to become almost singular, and this yields numerical problems in the computation of $(\mathcal{M}_k^{-1})_{1,1}$ and in solving for the other components. Nevertheless, if U^{-1} is nonsingular, H_k is nonsingular for all k, and if we apply the Q-OR method with this basis, we obtain residual vectors whose norms are such that

$$\|r_k^O\|^2 = \frac{1}{(\mathcal{M}_{k+1}^{-1})_{1,1}},$$

with $\mathcal{M}_{k+1} = K_{k+1}^T K_{k+1}$. It is known (see [15] and the references therein) that these values of the residual norms are those that are obtained from GMRES. Hence, maximizing $|\nu_{1,k}|$ yields the GMRES residual norms. This result was also obvious from Theorem 2.1. Therefore, these residual norms are the best ones that we can get with the given Krylov subspace with only one matrix-vector per iteration. In a sense, the construction that we have done before defines an optimal Q-OR method. We have (loosely) proved the following result.

THEOREM 3.1. *Given a matrix A and a starting vector b. If there is no breakdown in the algorithm described in this section, we obtain a basis for the Krylov subspace $\mathcal{K}_n(A, b)$ such that the Q-OR method constructed with this basis yields the same residual norms as GMRES applied to A and b.*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Since constructing U^{-1} is not practical, we would like to try build up the matrix H directly without using the matrix U or its inverse.

3.2. Some technical results. Before constructing the optimal basis using H, we need some technical results.

LEMMA 3.2. *Let A be a square real symmetric positive definite matrix of order n with a spectral factorization $A = Q\Lambda Q^T$ with $\Lambda = diag(\lambda_j)$, λ_j, $j = 1, \ldots, n$, being the eigenvalues of A and $Q^TQ = I$. Let $c \in \mathbb{R}^n$ be not orthogonal to any of the eigenvectors of A and γ be a real positive number. Then, the eigenvalues μ_j, $j = 1, \ldots, n$, of the matrix $A - \gamma cc^T$ are given by the solutions of the secular equation*

$$f(\mu) \equiv 1 - \gamma \sum_{j=1}^{n} \frac{(z_j)^2}{\lambda_j - \mu} = 0, \tag{3.1}$$

with $z = Q^Tc$.

Proof. See [7, 9]. We are looking for an eigenpair (μ, y) such that

$$(A - \gamma cc^T)y = \mu y.$$

This yields $c^Ty = \gamma c^T(A - \mu I)^{-1}cc^Ty$. But $c^Ty \neq 0$ since otherwise y would be an eigenvector of A and c would be orthogonal to such an eigenvector. Therefore, μ must satisfy the equation

$$1 - \gamma c^T(A - \mu I)^{-1}c = 0.$$

Using the spectral factorization of A and $z = Q^Tc$, we obtain the secular equation (3.1). ∎

COROLLARY 3.3. *Using the hypotheses and notation of Lemma 3.2, let*

$$\gamma_{opt} = \frac{1}{\sum_{j=1}^{n} \frac{(z_j)^2}{\lambda_j}} = \frac{1}{c^TA^{-1}c}.$$

Then $A - \gamma cc^T$ is positive definite if $0 \leq \gamma < \gamma_{opt}$, positive semi-definite if $\gamma = \gamma_{opt}$. and indefinite if $\gamma > \gamma_{opt}$.

Proof. Obviously the matrix is positive definite if $\gamma = 0$. So, let us assume $\gamma > 0$. The function f defined in (3.1) has poles at the eigenvalues λ_j of A, which are positive. The function f is decreasing in between the poles. Hence, since $\gamma > 0$, there is a strict interlacing between the eigenvalues of A and the eigenvalues of $A - \gamma cc^T$. Only the smallest eigenvalue μ_1 of $A - \gamma cc^T$ can eventually be negative. The function f tends to 1 from below when $\mu \to -\infty$. From (3.1) we have

$$f(0) = 1 - \gamma \sum_{j=1}^{n} \frac{(z_j)^2}{\lambda_j}.$$

This value is positive when $\gamma < \gamma_{opt}$, and $A - \gamma cc^T$ is singular when $\gamma = \gamma_{opt}$. ∎

PROPOSITION 3.4. *Let B be an $n \times k$ real matrix with $n > k$ and $d \in \mathbb{R}^n$, $d \neq 0$, such that the matrix $B_d = (B, -d)$ is of full rank $k + 1$. We also assume that $B^Td \neq 0$. Let $\nu \neq 0$ be a given vector in $\mathbb{R}^k$ such that the vector ν appended with a zero at the bottom is not orthogonal to any eigenvector of $B_d^TB_d$. Let*

$$\gamma_{opt} = \min_{y \in \mathbb{R}^k, \nu^Ty \neq 0} \frac{\|d - By\|^2}{(\nu^Ty)^2}. \tag{3.2}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Then,

$$\gamma_{opt} = \frac{\alpha}{\alpha \nu^T (B^T B)^{-1} \nu + \omega^2}, \tag{3.3}$$

with $\alpha = d^T d - d^T B (B^T B)^{-1} B^T d$ *and* $\omega = d^T B (B^T B)^{-1} \nu$. *Moreover, if* $\omega \neq 0$, *a solution* y_{opt} *of the minimization problem* (3.2) *is given by*

$$y_{opt} = (B^T B)^{-1} B^T d + \frac{\alpha}{\omega} (B^T B)^{-1} \nu. \tag{3.4}$$

Proof. Note that $\gamma_{opt} \geq 0$, but, by our hypothesis, d is not in the range of B, and there is no vector y such that $\|d - By\| = 0$. Hence, $\gamma_{opt} > 0$.

We use the same technique as in [16] and [11]. We consider the problem

$$\sup_{\gamma} \gamma, \quad \text{such that } \|d - By\|^2 \geq \gamma (\nu^T y)^2 \quad \text{for all } y \in \mathbb{R}^k, \nu^T y \neq 0. \tag{3.5}$$

From [11, Theorem 2] we know that this yields the solution of the minimization problem (3.2). We have

$$\|d - By\|^2 = d^T d - y^T B^T d - d^T B y + y^T B^T B y, \qquad (\nu^T y)^2 = y^T \nu \nu^T y.$$

Therefore, the constraint in (3.5) can be written in matrix form as

$$\begin{bmatrix} y^T & 1 \end{bmatrix} C(\gamma) \begin{bmatrix} y \\ 1 \end{bmatrix} \geq 0, \qquad C(\gamma) = \begin{bmatrix} B^T B - \gamma \nu \nu^T & -B^T d \\ -d^T B & d^T d \end{bmatrix}. \tag{3.6}$$

We note that the matrix $C(\gamma)$ can be expressed as

$$C(\gamma) = B_d^T B_d - \gamma \begin{bmatrix} \nu \\ 0 \end{bmatrix} \begin{bmatrix} \nu^T & 0 \end{bmatrix}, \qquad B_d \equiv \begin{bmatrix} B & -d \end{bmatrix}.$$

Hence, by our hypothesis, $C(\gamma)$ is a rank-one perturbation of a symmetric positive definite matrix. We are in a position to apply the results of Corollary 3.3. There is a unique value $\gamma = \gamma_{opt}$ for which the matrix $C(\gamma)$ is singular. If $0 \leq \gamma < \gamma_{opt}$, then the matrix $C(\gamma)$ is positive definite. Therefore, γ_{opt} is the largest value for which the relation (3.6) is true for all vectors y. Since the last component of the vector in the rank-one modification is zero, from Corollary 3.3, to compute the denominator of γ_{opt}, we need the value of

$$\begin{bmatrix} \nu^T & 0 \end{bmatrix} (B_d^T B_d)^{-1} \begin{bmatrix} \nu \\ 0 \end{bmatrix},$$

and therefore we have to find the first principal block of the inverse of $C(0)$. It is given by the inverse of the Schur complement (which is positive definite) and

$$\gamma_{opt} = \frac{1}{\nu^T \left(B^T B - \frac{1}{d^T d} B^T d d^T B \right)^{-1} \nu}.$$

We use the Sherman-Morrison formula (see [8]) to obtain

$$\gamma_{opt} = \frac{1}{\nu^T \left[(B^T B)^{-1} + (B^T B)^{-1} \frac{B^T d d^T B}{d^T d - d^T B (B^T B)^{-1} B^T d} (B^T B)^{-1} \right] \nu}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

With $\alpha = d^T d - d^T B(B^T B)^{-1} B^T d$ and $\omega = d^T B(B^T B)^{-1}\nu$, we obtain the relation (3.3). The value γ_{opt} is the minimum of (3.2) that we were looking for.

From our hypotheses there exists a vector $y_{opt} \neq 0$ such that

$$\gamma_{opt} = \frac{\|d - By_{opt}\|^2}{(\nu^T y_{opt})^2}.$$

It can be computed in the following way. When $\gamma = \gamma_{opt}$, the matrix $C(\gamma)$ has a zero eigenvalue. Therefore, we solve

$$(B^T B - \gamma_{opt}\nu\nu^T)y_{opt} = B^T d. \tag{3.7}$$

It turns out that we also have $d^T d - d^T B y_{opt} = 0$ since

$$d^T d - d^T B(B^T B - \gamma_{opt}\nu\nu^T)^{-1} B^T d = 0.$$

The vector y_{opt}, the solution of (3.7), is obtained from the Sherman-Morrison formula.

$$\begin{aligned} y_{opt} &= (B^T B - \gamma_{opt}\nu\nu^T)^{-1} B^T d, \\ &= (B^T B)^{-1} B^T d + (B^T B)^{-1} \frac{\gamma_{opt}\nu\nu^T}{1 - \gamma_{opt}\nu^T (B^T B)^{-1}\nu} (B^T B)^{-1} B^T d, \\ &= (B^T B)^{-1} B^T d + \frac{\gamma_{opt}[\nu^T (B^T B)^{-1} B^T d]}{1 - \gamma_{opt}\nu^T (B^T B)^{-1}\nu} (B^T B)^{-1}\nu. \end{aligned}$$

Now, $\nu^T (B^T B)^{-1} B^T d = d^T B(B^T B)^{-1}\nu = \omega$. Let $t = (B^T B)^{-1}\nu$. We have to consider the factor

$$\frac{\gamma_{opt}\omega}{1 - \gamma_{opt}\nu^T t}, \qquad \text{with} \quad \gamma_{opt} = \frac{\alpha}{\alpha\nu^T t + \omega^2}.$$

Some algebra yields,

$$\frac{\gamma_{opt}\omega}{1 - \gamma_{opt}\nu^T t} = \frac{\alpha}{\omega},$$

and this proves the result (3.4) for y_{opt}. □

Let us remark that from Proposition 3.4, we see that in order to compute γ_{opt} and y_{opt}, we need to solve two linear systems

$$(B^T B)t = \nu, \qquad (B^T B)s = B^T d.$$

The vector s is the solution of the least-squares problem $\min_y \|d - By\|$. The solution of the problem in Proposition 3.4 is the solution of this least-squares problem plus a correction depending on ν. We also remark that the coefficient α is a Schur complement in the matrix $C(0)$ which is symmetric positive definite. Therefore, α is positive.

Let us now consider the case $B^T d = 0$. Then, $\gamma_{opt} = 1/\nu^T (B^T B)^{-1}\nu$, and $\|d - By\|^2 = d^T d + y^T B^T B y$. Therefore we have

$$y^T (B^T B - \gamma_{opt}\nu^T\nu)y = -d^T d < 0.$$

The matrix within the parenthesis is positive semi-definite, and there is no vector y that could satisfy this identity. Hence, there is no finite solution to the minimization problem (3.2).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

3.3. The construction of the basis using H. To construct the Q-OR optimal basis using the relation $AV = VH$, that is, without using the inverse of U, we have to relate H to the first row of U^{-1}.

LEMMA 3.5. *The entries $\nu_{1,j}$ of the first row of the inverse of U are related to the entries $h_{i,j}$ of H by*

$$\nu_{1,k+1} = -\frac{1}{h_{k+1,k}} \sum_{j=1}^{k} \nu_{1,j} h_{j,k}, \qquad k = 1, \ldots, n-1. \tag{3.8}$$

Proof. We have seen that $H = UCU^{-1}$, where C is the companion matrix defined in (2.2). Multiplying on the left by U^{-1}, we have $U^{-1}H = CU^{-1}$. From the structure of C and the fact that U^{-1} is upper triangular, the entries of the first row of CU^{-1} are zero except for the last one in position $(1, n)$. Writing the entry $(1, k)$ of $U^{-1}H$ for $k < n$, we obtain

$$\sum_{j=1}^{k+1} \nu_{1,j} h_{j,k} = 0 \;\Rightarrow\; \nu_{1,k+1} = -\frac{1}{h_{k+1,k}} \sum_{j=1}^{k} \nu_{1,j} h_{j,k}. \qquad \square$$

Hence, we have a relation between the first row of U^{-1} and the kth column of H. Let us assume that at step k of the algorithm we have already computed $\nu_{1,j}$, $j = 1, \ldots, k$, and we would like to find $h_{j,k}$, $j = 1, \ldots, k+1$, to maximize the absolute value of $\nu_{1,k+1}$, knowing that $h_{k+1,k}$ has to be chosen to obtain a vector v_{k+1} of unit norm. From the Arnoldi-like relation (2.3), the subdiagonal entry $h_{k+1,k}$ is given by the norm of the vector

$$\tilde{v}_k = Av_k - \sum_{j=1}^{k} h_{j,k} v_j, \tag{3.9}$$

and the next basis vector is $v_{k+1} = \tilde{v}_k / h_{k+1,k}$ with $h_{k+1,k} = \|\tilde{v}_k\|$. Then, from Lemma 3.5,

$$|\nu_{1,k+1}| = \frac{|\nu^T y|}{\|d - By\|},$$

with

$$d = Av_k, \quad B = V_k = \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix}, \quad y = \begin{bmatrix} h_{1,k} & \cdots & h_{k,k} \end{bmatrix}^T, \quad \nu = \begin{bmatrix} \nu_{1,1} & \cdots & \nu_{1,k} \end{bmatrix}.$$

We would like to maximize $|\nu_{1,k+1}|^2$. Hence, we wish to solve

$$\frac{1}{|\nu_{1,k+1}|^2} = \min_{y \in \mathbb{R}^k, \nu^T y \neq 0} \frac{\|d - By\|^2}{(\nu^T y)^2}.$$

The minimizer of this problem was given in Proposition 3.4. When it exists, the solution y_{opt} is obtained by solving two linear systems with the matrix $V_k^T V_k$. We outline in a moment how to simplify the algorithm and to compute the solution efficiently. Let us first consider a straightforward implementation of the algorithm (with $x_0 = 0$):

ALGORITHM 0.

Initialization phase

$v_1 = b/\|b\|$, $v_1^A = Av_1$,
$\omega = v_1^T v_1^A$, $\alpha = (v_1^A)^T v_1^A - \omega^2$,
$h_{1,1} = \omega + \frac{\alpha}{\omega}$,
$\tilde{v} = v_1^A - h_{1,1} v_1$, $h_{2,1} = \|\tilde{v}\|$,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$v_2 = \frac{1}{h_{2,1}}\tilde{v},\ v_2^A = Av_2,$
$\nu_{1,1} = 1, \nu_{1,2} = -\frac{h_{1,1}}{h_{2,1}},$
$\nu = \begin{bmatrix}\nu_{1,1} & \nu_{1,2}\end{bmatrix}^T.$
End of initialization
For $k = 2, \ldots$
1. solve of $V_k^T V_k\, t = \nu$,
2. $v_k^{tA} = V_k^T v_k^A$,
3. solve of $V_k^T V_k\, s = v_k^{tA}$,
4. $\omega = (v_k^{tA})^T t,\ \alpha = (v_k^A)^T v_k^A - (v_k^{tA})^T s$,
5. $$h_{1:k,k} = \begin{bmatrix} h_{1,k} \\ \vdots \\ h_{k,k} \end{bmatrix} = s + \frac{\alpha}{\omega} t,$$
6. $$\tilde{v} = v_k^A - V_k\, h_{1:k,k},\ h_{k+1,k} = \|\tilde{v}\|,\ \nu_{1,k+1} = -\frac{1}{h_{k+1,k}} \nu^T h_{1:k,k},$$
$$\nu = \begin{bmatrix}\nu_{1,1} & \cdots & \nu_{1,k+1}\end{bmatrix}^T,$$
7. $v_{k+1} = \frac{1}{h_{k+1,k}}\tilde{v}$ and $v_{k+1}^A = Av_{k+1}$,
8. if needed, solve $H_k y_k = \|b\| e_1,\ x_k = V_k y_k$.

End For k.

We will later discuss how to efficiently solve the linear systems in steps 1 and 3. This will provide an improved implementation of the method.

4. Mathematical properties of the optimal basis. In this section we prove some properties of the basis that will help us simplify the implementations of the algorithm. We use the same notation as in the previous section,

$$t = (V_k^T V_k)^{-1}\nu, \qquad s = (V_k^T V_k)^{-1} V_k^T A v_k, \qquad \nu = \begin{bmatrix}\nu_{1,1} & \cdots & \nu_{1,k}\end{bmatrix}^T,$$

$$\omega = (V_k^T A v_k)^T t, \qquad \alpha = \|Av_k\|^2 - (V_k^T A v_k)^T s.$$

We assume that $\nu_{1,j} \neq 0,\ j = 1, \ldots, k$, and that the absolute values of the $\nu_{1,j}$'s are strictly increasing. That is, we assume that GMRES is not stagnating. First, we note that we could expect the basis to be closer and closer to orthogonality as the method converges. More precisely, the vector $\tilde{v}_k$ in (3.9) is

$$\tilde{v}_k = (I - V_k(V_k^T V_k)^{-1} V_k^T) A v_k - \frac{\alpha}{\omega} V_k (V_k^T V_k)^{-1} \nu. \tag{4.1}$$

The first term on the right-hand side is in the orthogonal complement of the subspace $\mathcal{K}_k(A, b)$, and the second one is in $\mathcal{K}_k(A, b)$. We remark that if we took $\alpha = 0$ for every iteration, we would construct an orthogonal basis. In a moment, we highlight the significance of the term α/ω. Looking at the angles between the basis vectors, from (4.1), we have

$$V_k^T v_{k+1} = \frac{1}{h_{k+1,k}} V_k^T \tilde{v}_k = -\frac{\alpha}{\omega h_{k+1,k}} \begin{bmatrix} \nu_{1,1} \\ \vdots \\ \nu_{1,k} \end{bmatrix}. \tag{4.2}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The next lemma shows how this expression can be simplified.

LEMMA 4.1. *The basis vectors satisfy*

$$V_{k+1}^T v_{k+1} = \frac{1}{\nu_{1,k+1}} \begin{bmatrix} \nu_{1,1} \\ \vdots \\ \nu_{1,k} \\ \nu_{1,k+1} \end{bmatrix}. \tag{4.3}$$

Proof. From relation (4.2) we have

$$V_{k+1}^T v_{k+1} = \frac{1}{h_{k+1,k}} \begin{bmatrix} V_k^T \\ v_{k+1}^T \end{bmatrix} \tilde{v}_k = \begin{bmatrix} -\frac{\alpha}{\omega h_{k+1,k}} \nu \\ 1 \end{bmatrix}.$$

We need to consider $\alpha/(\omega h_{k+1,k})$. We first remark that $\omega = \nu^T s$ because $\nu^T s = t^T V_k^T A v_k$. Then, we compute $h_{k+1,k}^2 = \|\tilde{v}_k\|^2$ using (3.9),

$$h_{k+1,k}^2 = \|Av_k\|^2 - 2(Av_k, V_k h_{1:k,k}) + (V_k^T V_k h_{1:k,k}, h_{1:k,k}),$$

and

$$V_k h_{1:k,k} = V_k \left(s + \frac{\alpha}{\omega} t \right), \qquad V_k^T V_k h_{1:k,k} = V_k^T A v_k + \frac{\alpha}{\omega} \nu.$$

This yields

$$h_{k+1,k}^2 = \|Av_k\|^2 - (Av_k, V_k s) - \frac{\alpha}{\omega}[(Av_k, V_k t) - (\nu, s)] + \left(\frac{\alpha}{\omega}\right)^2 (\nu, t).$$

But,

$$\|Av_k\|^2 - (Av_k, V_k s) = \alpha, \qquad (Av_k, V_k t) - (\nu, s) = 0.$$

Therefore,

$$h_{k+1,k}^2 = \alpha + \left(\frac{\alpha}{\omega}\right)^2 (\nu, t).$$

Inserting ω into the first term of the right-hand side of this relation, we obtain

$$h_{k+1,k}^2 = \frac{\alpha}{\omega} \left(\nu^T s + \frac{\alpha}{\omega} \nu^T t \right).$$

Therefore,

$$\frac{\alpha}{h_{k+1,k}\omega} = \frac{h_{k+1,k}}{\nu^T s + \frac{\alpha}{\omega} \nu^T t} = \frac{h_{k+1,k}}{\nu^T h_{1:k,k}} = -\frac{1}{\nu_{1,k+1}}.$$

This proves the relation (4.3). ∎

From Lemma 4.1 we observe that the cosines of the angles between the basis vectors are given by ratios of the residual norms since

$$\frac{|\nu_{1,j}|}{|\nu_{1,k+1}|} = \frac{\|r_k^O\|}{\|r_{j-1}^O\|}, \qquad j = 1, \ldots, k+1.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The right-hand side of this identity becomes small when the method converges fast enough. The matrix $V_k^T V_k$ is

$$V_k^T V_k = \begin{bmatrix} 1 & \frac{1}{\nu_{1,2}} & \frac{1}{\nu_{1,3}} & \cdots & \frac{1}{\nu_{1,k}} \\ \frac{1}{\nu_{1,2}} & 1 & \frac{\nu_{1,2}}{\nu_{1,3}} & \cdots & \frac{\nu_{1,2}}{\nu_{1,k}} \\ \frac{1}{\nu_{1,3}} & \frac{\nu_{1,2}}{\nu_{1,3}} & 1 & \cdots & \frac{\nu_{1,3}}{\nu_{1,k}} \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{1}{\nu_{1,k}} & \frac{\nu_{1,2}}{\nu_{1,k}} & \cdots & & 1 \end{bmatrix}. \tag{4.4}$$

Below, we explore the inverse of this matrix in detail. The next lemma shows that, mathematically, the vector t is zero except for the last component.

LEMMA 4.2. *The solution of $V_k^T V_k \nu = t$ is*

$$t = (V_k^T V_k)^{-1} \nu = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \nu_{1,k} \end{bmatrix}. \tag{4.5}$$

Proof. Using Lemma 4.1 we have

$$\nu_{1,k} V_k^T v_k = \nu.$$

Therefore,

$$\nu = \nu_{1,k} V_k^T V_k e_k = V_k^T V_k \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \nu_{1,k} \end{bmatrix}. \qquad \square$$

Next, let us prove that the basis is semi A-orthogonal (or conjugate), that is, $v_i^T A v_k = 0$, for $i > k$. Hence, our Q-OR optimal method is a right conjugate direction method (RCD). For a left conjugate direction method (LCD), see [22].

LEMMA 4.3. *The matrix $V_k^T A V_k$ is upper triangular.*

Proof. Let us consider $v_i^T A v_k$, $i > k$. From the Arnoldi-like relation (2.3), we have

$$A v_k = h_{k+1,k} v_{k+1} + V_k h_{1:k,k}.$$

This yields

$$v_i^T A v_k = \sum_{j=1}^{k+1} h_{j,k} v_i^T v_j.$$

From relation (4.4), the i-th row of $V^T V$ is

$$\begin{bmatrix} \frac{\nu_{1,1}}{\nu_{1,i}} & \cdots & \frac{\nu_{1,i-1}}{\nu_{1,i}} & 1 & \frac{\nu_{1,i}}{\nu_{1,i+1}} & \cdots \frac{\nu_{1,i}}{\nu_{1,n}} \end{bmatrix}.$$

Then, with $i > k$ and using the relation (3.8),

$$\begin{aligned} \sum_{j=1}^{k+1} h_{j,k} v_i^T v_j = \sum_{j=1}^{k+1} h_{j,k} \frac{\nu_{1,j}}{\nu_{1,i}} &= h_{k+1,k} \frac{\nu_{1,k+1}}{\nu_{1,i}} + \sum_{j=1}^{k} h_{j,k} \frac{\nu_{1,j}}{\nu_{1,i}}, \\ &= \frac{1}{\nu_{1,i}} [h_{k+1,k} \nu_{1,k+1} - h_{k+1,k} \nu_{1,k+1}] = 0. \qquad \square \end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

It is interesting to remark that the inverse of the matrix $V_k^T V_k$ has a particular structure.

LEMMA 4.4. *The inverse of the matrix $V_k^T V_k$ in (4.4) is tridiagonal.*

Proof. This is readily seen from the structure of the matrix in relation (4.4); see [13]. But, this can also be proved directly as follows. From relation (2.3) we have

$$V_k^T A V_k = V_k^T V_k H_k + h_{k+1,k} V_k^T v_{k+1} e_k^T.$$

Let $R_k = V_k^T A V_k - h_{k+1,k} V_k^T v_{k+1} e_k^T$. By Lemma 4.3 the matrix R_k is upper triangular. We have $(V_k^T V_k)^{-1} = H_k R_k^{-1}$. This shows that $(V_k^T V_k)^{-1}$ is upper Hessenberg. However, it is a symmetric matrix, and therefore it is tridiagonal. ∎

It turns out that one can find exact expressions for the diagonal and subdiagonal entries of the inverse of the matrix $V_k^T V_k$ in (4.4). Let α_j be the diagonal entries and β_j the subdiagonal entries. We have

$$\begin{aligned}
\alpha_1 &= \frac{\nu_{1,2}^2}{\nu_{1,2}^2 - 1}, & \beta_1 &= -\frac{\nu_{1,2}}{\nu_{1,2}^2 - 1}, \\
\alpha_i &= \frac{\nu_{1,i-1}^2}{\nu_{1,i}^2 - \nu_{1,i-1}^2} + \frac{\nu_{1,i+1}^2}{\nu_{1,i+1}^2 - \nu_{1,i}^2}, & \beta_i &= -\frac{\nu_{1,i}\nu_{1,i+1}}{\nu_{1,i+1}^2 - \nu_{1,i}^2}, \qquad i = 2, \ldots, k-1, \\
\alpha_k &= \frac{\nu_{1,k}^2}{\nu_{1,k}^2 - \nu_{1,k-1}^2}.
\end{aligned}$$

This result helps us to obtain a lower bound for the singular values of V_k.

THEOREM 4.5. *Assume $|\nu_{1,j+1}| \neq |\nu_{1,j}|$, $j = 1, \ldots, n-1$. Let μ_i, $i = 1, \ldots, k$, be defined as*

$$\mu_1 = \frac{|\nu_{1,2}|}{\nu_{1,2}^2 - 1}, \quad \mu_i = \frac{|\nu_{1,i-1}|\,|\nu_{1,i}|}{\nu_{1,i}^2 - \nu_{1,i-1}^2} + \frac{|\nu_{1,i}|\,|\nu_{1,i+1}|}{\nu_{1,i+1}^2 - \nu_{1,i}^2}, \quad 1 < i < k, \quad \mu_k = \frac{|\nu_{1,k-1}|\,|\nu_{1,k}|}{\nu_{1,k}^2 - \nu_{1,k-1}^2}.$$

The smallest singular value σ_k of V_k is bounded from below by

$$\sigma_k \geq \frac{1}{(\max_i(\alpha_i + \mu_i))^{\frac{1}{2}}}.$$

Proof. Using Gerschgorin circles, the eigenvalues of the tridiagonal matrix $(V_k^T V_k)^{-1}$ are located in the union of the intervals $[\alpha_i - \mu_i, \alpha_i + \mu_i]$, $i = 1, \ldots, k$, where $\mu_1 = |\beta_1|$, $\mu_i = |\beta_{i-1}| + |\beta_i|$, $i = 2, \ldots, k-1$, and $\mu_k = |\beta_{k-1}|$. Hence, the smallest eigenvalue of $V_k^T V_k$ is bounded from below by $1/\max_i(\alpha_i + \mu_i)$. ∎

From Theorem 4.5 we have

$$\alpha_1 + \mu_1 = \frac{1}{1 - \frac{1}{\nu_{1,2}^2}} \left(1 + \frac{1}{|\nu_{1,2}|}\right), \qquad \alpha_k + \mu_k = \frac{1}{1 - \frac{\nu_{1,k-1}^2}{\nu_{1,k}^2}} \left(1 + \frac{|\nu_{1,k-1}|}{|\nu_{1,k}|}\right),$$

and

$$\alpha_i + \mu_i = \frac{1}{\frac{\nu_{1,i}^2}{\nu_{1,i-1}^2} - 1} \left(1 + \frac{|\nu_{1,i}|}{|\nu_{1,i-1}|}\right) + \frac{1}{1 - \frac{\nu_{1,i}^2}{\nu_{1,i+1}^2}} \left(1 + \frac{|\nu_{1,i}|}{|\nu_{1,i+1}|}\right), \quad i = 2, \ldots, k-1.$$

Without stagnation, the values $|\nu_{1,i}|$ must be strictly increasing since they are the inverses of the GMRES residual norms. Hence, all the terms are positive. But, we see that if some

ETNA
Kent State University and
Johann Radon Institute (RICAM)

consecutive values of $|\nu_{1,i}|$ are close, then the maximum of $\alpha_i + \mu_i$ will be large, and the lower bound of the smallest singular value will be small. On the contrary, if the sequence $|\nu_{1,i}|$ is increasing rapidly, then the lower bound is large, and the basis is well behaved.

Considering the other ends of the intervals we have

$$\alpha_1 - \mu_1 = \frac{1}{1 - \frac{1}{\nu_{1,2}^2}} \left(1 - \frac{1}{|\nu_{1,2}|}\right), \qquad \alpha_k - \mu_k = \frac{1}{1 - \frac{\nu_{1,k-1}^2}{\nu_{1,k}^2}} \left(1 - \frac{|\nu_{1,k-1}|}{|\nu_{1,k}|}\right).$$

From $|\nu_{1,2}| \geq 1$, $|\nu_{1,k}| \geq |\nu_{1,k-1}|$, we obtain $\alpha_1 - \mu_1 \geq 0$ and $\alpha_k - \mu_k \geq 0$, respectively. For the other intervals from the Gerschgorin circles, we have

$$\alpha_i - \mu_i = \frac{1}{\frac{\nu_{1,i}^2}{\nu_{1,i-1}^2} - 1} \left(1 - \frac{|\nu_{1,i}|}{|\nu_{1,i-1}|}\right) + \frac{1}{1 - \frac{\nu_{1,i}^2}{\nu_{1,i+1}^2}} \left(1 - \frac{|\nu_{1,i}|}{|\nu_{1,i+1}|}\right), \quad i = 2, \ldots, k-1.$$

The first term on the right-hand side is negative, and the second one is positive. Hence, we do not know if $\alpha_i - \mu_i$ is positive for $i = 2, \ldots, k-1$. A necessary and sufficient condition for $\alpha_i - \mu_i$ to be positive is

$$|\nu_{1,i}|\,|\nu_{1,i+1}^2 - \nu_{1,i-1}^2| \geq |\nu_{1,i-1}|\,|\nu_{1,i+1}^2 - \nu_{1,i}^2| + |\nu_{1,i+1}|\,|\nu_{1,i}^2 - \nu_{1,i-1}^2|.$$

When $\min_i(\alpha_i - \mu_i) > 0$, we obtain an upper bound for σ_1, the largest singular value of V_k.

Unfortunately, some of the mathematical properties of the basis vectors that we proved above are not verified up to machine precision in floating point computations. We can only use a few of them. For instance, the computation of the new unnormalized vector $\tilde{v}_k$ in (3.9) can be simplified since from relation (4.5),

$$V_k h_{1:k,k} = V_k s + \frac{\alpha}{\omega} \nu_{1,k} v_k.$$

But,

$$\omega = t^T V_k^T A v_k = \nu_{1,k} v_k^T A v_k.$$

Assuming $\nu_{1,k} \neq 0$, yields

$$\tilde{v}_k = A v_k - V_k s - \beta v_k, \qquad \beta = \frac{\alpha}{v_k^T A v_k}.$$

Moreover, the k first entries of column k of H are given by

$$h_{1:k,k} = s + \beta e_k.$$

We observe that we have a breakdown of the algorithm only if $v_k^T A v_k = 0$. This means that this method cannot be used for skew-symmetric matrices.

The relations giving the inverse of $V_k^T V_k$ have to be used carefully to solve $V_k^T V_k s = v_k^{tA}$ at each iteration since this will gradually lead to a discrepancy with the values that can be computed from the vectors v_j. This is unfortunate since it would have given a big saving in the number of dot products. However, these relations can be used if we restart the algorithm and if the restart parameter m is not too large. This will be considered in a forthcoming paper.

Let us prove some more relations between the quantities involved in the algorithm. We have

$$h_{k+1,k}^2 = \alpha + \left(\frac{\alpha}{\omega}\right)^2 \nu_{1,k}^2 = \alpha + \frac{\alpha^2}{(v_k^T A v_k)^2} = \alpha + \beta^2.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We also have

$$h_{k+1,k}^2 = \alpha + h_{k+1,k}^2 \frac{\nu_{1,k}^2}{\nu_{1,k+1}^2},$$

which yields

$$h_{k+1,k}^2 = \alpha \frac{\nu_{1,k+1}^2}{\nu_{1,k+1}^2 - \nu_{1,k}^2}.$$

From the Arnoldi-like relation (2.3), we have

$$V_k^T \tilde{v}_k = V_k^T A v_k - V_k^T V_k s - \beta V_k^T v_k,$$

which yields $\beta = -v_k^T \tilde{v}_k$. Moreover, the normalization of $\tilde{v}_k$ gives us

$$V_k^T v_{k+1} = -\frac{\beta}{h_{k+1,k}} V_k^T v_k.$$

Using the expression for $V_{k+1}^T v_{k+1}$, we obtain

$$\frac{\nu_{1,k}}{\nu_{1,k+1}} = -\frac{\beta}{h_{k+1,k}} = -\frac{\beta}{\sqrt{\alpha + \beta^2}}.$$

Hence,

$$\frac{|\nu_{1,k}|}{|\nu_{1,k+1}|} \leq 1.$$

We have seen that the residual vectors are proportional to the basis vectors since $r_k = -h_{k+1,k} y_k^{(k)} v_{k+1}$. Moreover, from the Arnoldi-like relation (2.3), we have

$$h_{k+1,k}\, v_{k+1} = A v_k - V_k h_{1:k,k}.$$

This implies that the basis vectors are obtained by the application of a polynomial in A to the initial residual $v_{k+1} = q_k(A)b$, where q_k is a polynomial of degree k. Hence, the residual vectors are also given in polynomial form as $r_k^O = p_k(A)b$, where the so-called residual polynomial p_k of order k is such that $p_k(0) = 1$. The polynomials q_k satisfy

$$q_0(\lambda) \equiv 1, \quad q_1(\lambda) = \frac{1}{h_{2,1}}(\lambda - h_{1,1}),$$

$$q_k(\lambda) = \frac{1}{h_{k+1,k}} \left[\lambda q_{k-1}(\lambda) - \sum_{i=1}^{k} h_{i,k} q_{i-1}(\lambda) \right], \qquad k = 2, \ldots$$

On the other hand, since H_k is an upper Hessenberg matrix, it is well known that $\det(\lambda I - H_k)$ is a polynomial $s_k(\lambda)$ that satisfies $s_1(\lambda) = \lambda - h_{1,1}$ and the recurrence relation

$$s_k(\lambda) = (\lambda - h_{k,k}) s_{k-1}(\lambda) - \sum_{i=1}^{k-1} h_{i,k} \prod_{j=i+1}^{k} h_{j,j-1} s_{i-1}(\lambda), \qquad k = 2, \ldots$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

It can be shown that the two polynomials q_k and s_k are equal up to a scaling factor, and therefore they have the same roots. The roots of the residual polynomials are the eigenvalues of H_k. These eigenvalues can be considered, when k increases, as approximations to the eigenvalues of A. They are distinct from the Ritz values that are obtained from the Arnoldi algorithm.

Let us assume that the data are real. Since $K = VU$, we have

$$V = KU^{-1} \Longrightarrow v_{k+1} = KU^{-1}e_{k+1} = \pm\frac{r_k^O}{\|r_k^O\|} = \pm\frac{1}{\|r_k^O\|}p_k(A)b.$$

Let $p_k(A)b = \xi_k^{(k)}A^kb + \cdots + \xi_1^{(k)}Ab + b$. By identification of the coefficients, we obtain

$$U^{-1}e_{k+1} = \pm\frac{1}{\|r_k^O\|}\begin{bmatrix} 1 \\ \xi_1^{(k)} \\ \vdots \\ \xi_k^{(k)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We can also express $U_j^{-1} = \hat{U}_j^{-1}D_j^{-1}$, for $j = 1, \ldots, n-1$, by means of

$$D_j = \begin{bmatrix} 1 & & & \\ & \|r_1^O\| & & \\ & & \ddots & \\ & & & \|r_{j-1}^O\| \end{bmatrix} S_j, \quad \hat{U}_j^{-1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ & \xi_1^{(1)} & \cdots & \xi_1^{(j-1)} \\ & & \ddots & \vdots \\ & & & \xi_{j-1}^{(j-1)} \end{bmatrix},$$

where S_j is a diagonal matrix with ± 1 on the diagonal. The columns of the matrix $\hat{U}_j^{-1}$ contain the coefficients of the residual polynomials.

5. Implementation of the algorithm. Let us investigate how the computation in the algorithm at step k can be organized. As said, we have to solve a linear system with the matrix $V_k^TV_k$. We would like to compute the Cholesky factorization of this matrix incrementally. Let us assume that we know the lower triangular matrix L_{k-1} such that $L_{k-1}L_{k-1}^T = V_{k-1}^TV_{k-1}$. Then, denote

$$L_kL_k^T = \begin{bmatrix} V_{k-1}^TV_{k-1} & z_k \\ z_k^T & 1 \end{bmatrix},$$

with $z_k = V_{k-1}^Tv_k$ and

$$L_k = \begin{bmatrix} L_{k-1} & 0 \\ \ell_k^T & \ell_{k,k} \end{bmatrix}.$$

It is well known that by identification, we find that ℓ_k is obtained by solving the equation $L_{k-1}\ell_k = z_k = V_{k-1}^Tv_k$ and $\ell_{k,k} = \sqrt{1 - \ell_k^T\ell_k}$.

If we use this Cholesky factorization, then there will occur triangular solves that can slow down the computations. Therefore it seems more promising to construct the inverses of the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

matrices L_k incrementally. Doing so, we can replace the triangular solves by matrix-vector products. We have,

$$\tilde{L}_k = L_k^{-1} = \begin{bmatrix} L_{k-1}^{-1} & 0 \\ -\frac{1}{\ell_{k,k}} \ell_k^T L_{k-1}^{-1} & \frac{1}{\ell_{k,k}} \end{bmatrix}.$$

We note that

$$\ell_k^T \ell_k = v_k^T V_{k-1} L_{k-1}^{-T} L_{k-1}^{-1} V_{k-1}^T v_k = v_k^T V_{k-1} (V_{k-1}^T V_{k-1})^{-1} V_{k-1}^T v_k = v_k^T P_{k-1} v_k,$$

where P_{k-1} is the orthogonal projector onto the Krylov subspace of dimension $k-1$. This yields

$$\ell_{k,k} = \|(I - P_{k-1}) v_k\|.$$

Note that $\ell_{k,k} = 0$ if and only if $v_k = P_{k-1} v_k$, which means that $v_k \in \mathcal{K}_{k-1}(A, b)$ and V_k is not of full rank. We remark that

$$P_{k-1} v_k = V_{k-1} L_{k-1}^{-T} L_{k-1}^{-1} V_{k-1}^T v_k = V_{k-1} L_{k-1}^{-T} \ell_k.$$

We can easily find $(P_{k-1} v_k)^T = \ell_k^T L_{k-1}^{-1} V_{k-1}^T$ since $\ell_k^T L_{k-1}^{-1}$ has to be computed to obtain the last row of L_k^{-1}. The interest of computing $\ell_{k,k}$ in this way is that we are sure that $\ell_{k,k} \geq 0$. The algorithm reads as follows(with $x_0 = 0$):

ALGORITHM 1 (Q-OR-optinv).

Initialization phase

$v_1 = b/\|b\|$, $v_1^A = A v_1$, $\tilde{L}_1 = 1$,
$\omega = v_1^T v_1^A$, $\alpha = (v_1^A)^T v_1^A - \omega^2$,
$h_{1,1} = \omega + \frac{\alpha}{\omega}$,
$\tilde{v} = v_1^A - h_{1,1} v_1$, $h_{2,1} = \|\tilde{v}\|$,
$v_2 = \frac{1}{h_{2,1}} \tilde{v}$, $v_2^A = A v_2$,
$\nu_{1,1} = 1$, $\nu_{1,2} = -\frac{h_{1,1}}{h_{2,1}}$,
$\nu = \begin{bmatrix} \nu_{1,1} & \nu_{1,2} \end{bmatrix}^T$.

End of initialization

For $k = 2, \ldots$

1. $v_k^V = V_{k-1}^T v_k$, $v_k^{tA} = V_k^T v_k^A$,
2. $\ell_k = \tilde{L}_{k-1} v_k^V$,
3. $y_k^T = \ell_k^T \tilde{L}_{k-1}$,
4. if $\ell_k^T \ell_k < 1$, $\ell_{k,k} = \sqrt{1 - \ell_k^T \ell_k}$, else $(p_k^v)^T = y_k^T V_{k-1}^T$, $\ell_{k,k} = \|v_k - p_k^v\|$,
5. $$\tilde{L}_k = \begin{bmatrix} \tilde{L}_{k-1} & 0 \\ -\frac{1}{\ell_{k,k}} y_k^T & \frac{1}{\ell_{k,k}} \end{bmatrix},$$
6. $\ell_A = \tilde{L}_k v_k^{tA}$, $s = \tilde{L}_k^T \ell_A$,
7. $\alpha = (v_k^A)^T v_k^A - \ell_A^T \ell_A$, $\beta = \frac{\alpha}{(v_k^{tA})_k}$,
8. $$h_{1:k,k} = \begin{bmatrix} h_{1,k} \\ \vdots \\ h_{k,k} \end{bmatrix} = s + \beta e_k,$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

9.

$$\tilde{v} = v_k^A - V_k\, h_{1:k,k},\ h_{k+1,k} = \|\tilde{v}\|, \qquad \nu_{1,k+1} = -\frac{1}{h_{k+1,k}}\nu^T h_{1:k,k},$$
$$\nu = \begin{bmatrix}\nu_{1,1} & \cdots & \nu_{1,k+1}\end{bmatrix}^T,$$

10. $v_{k+1} = \frac{1}{h_{k+1,k}}\tilde{v}$ and $v_{k+1}^A = Av_{k+1}$,
11. if needed, solve $H_k y^{(k)} = \|b\|e_1$ using Givens rotations, $x_k = V_k y^{(k)}$.

End For k.

Note that the modulus of $\nu_{1,k+1}$ gives the inverse of the (relative) norm of the Q-OR residual at iteration k. Hence, we can compute the basis vectors, stop the iterations using $\nu_{1,k+1}$, and then reduce the upper Hessenberg matrix to upper triangular form to compute the final approximate solution.

Unfortunately, in step 1, we have to compute $V_{k-1}^T v_k$ and $V_k^T v_k^A$, that is, $2k-1$ dot products while there are only k dot products in the Arnoldi process (but, of course, when the basis is orthonormal, $V_k^T V_k = I$). This is the price to pay for having a non-orthogonal basis. The reader may wonder why we have derived an algorithm which delivers the same residual norms as GMRES but with more floating point operations. However, the dot products in Q-OR-optinv are all independent, and they can be computed in parallel contrary to the dot products in the MGS implementation of GMRES.

As in the algorithm using the matrix U, there may also be breakdowns in the algorithm using H. A problem occurs if there is an index k for which $v_k^T A v_k = 0$. Hence, the smallness of $v_k^T A v_k$ must be tested in step 7.

Of course, the previous algorithm can also be restarted every m iterations as it is done for GMRES. Moreover, we can also easily use a preconditioner.

In one iteration of Q-OR-optinv, we have $2k$ dot products of length n. However, the dot products $V_{k-1}^T v_k$ and $V_k^T v_k^A$ in step 1 can be computed as one dense matrix-matrix product $V_k^T[v_k, Av_k]$, where the first matrix is $k \times n$ and the second one is $n \times 2$. There are two triangular matrix-vector products of order k and two of order $k-1$ as well as two dot products of length k. The two other operations are the dense matrix-vector product $V_k h_{1:k,k}$ and a (sparse) matrix-vector product Av_{k+1}.

As we said above, even though we have more floating point operations in Q-OR-optinv than in GMRES, this algorithm could be interesting for parallel computers since GMRES with the modified Gram-Schmidt implementation is not fully parallel. It is not the topic of this paper, but on a parallel computer, most of the operations can be replicated on every processing unit. The global operations are the matrix-vector products $V_{k-1}^T v_k$ and $V_k^T v_k^A$ in step 1, the dot product $(v_k^A)^T v_k^A$, and an addition of vectors $v_k^A - V_k\, h_{1:k,k}$. The first two of these operations can be done in parallel using optimized codes for dense matrix-matrix products. Depending on the way V_k is stored, the last operation can be done without communications.

6. Numerical experiments with Q-OR-optinv. In this section we consider a few linear systems with nonsymmetric matrices, and we compare Q-OR-optinv (Algorithm 1) with GMRES. The acronym GMRES will refer to GMRES with modified Gram-Schmidt orthogonalization. The matrices come from the Matrix Market[1] or Tim Davis' SuiteSparse Matrix Collection[2] at Texas A&M University (formerly known as University of Florida matrix collection).

[1] http://math.nist.gov/MatrixMarket/
[2] https://sparse.tamu.edu/

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Example 1. For the first problem, the matrix A is *fs 680 1* of order 680 scaled by the inverse of its diagonal. It has 2184 nonzero entries. The norm of A is 3.8168, and its condition number is $8.6944\ 10^3$, the smallest singular value being $4.3900\ 10^{-4}$. This matrix is non-normal. Let e be the vector with all its components equal to 1. Then, the right-hand side is $b = Ae$, and $x_0 = 0$.

Figure 6.1 displays the difference of the true residual norms of GMRES and Q-OR-optinv as a function of the iteration number. It is smaller than 10^{-13} except after the final stagnation. Figure 6.2 shows the true residual norms of GMRES (plain curve) and Q-OR-optinv (dashed curve). We see that they almost coincide except for the final stagnation. Q-OR-optinv yields a better maximum attainable accuracy than GMRES by a factor of 9.

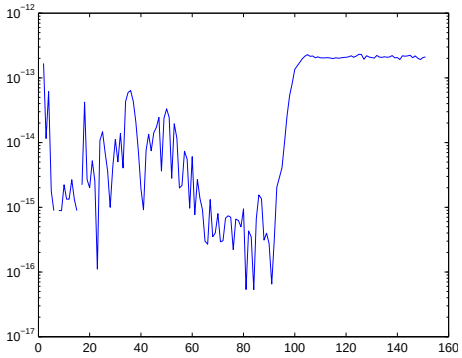

FIG. 6.1. *fs 680 1c: difference between the residual norms of GMRES and Q-OR-optinv.*

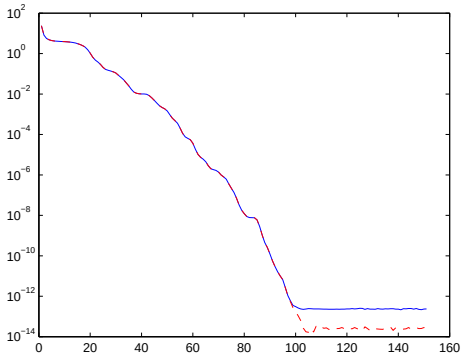

FIG. 6.2. *fs 680 1c: residual norms of GMRES (plain) and Q-OR-optinv (dashed).*

Table 6.1 shows the maximum attainable accuracies with variants of GMRES and Q-OR-optinv. CGS refers to the classical Gram-Schmidt algorithm. We use it also with reorthogonalization and double reorthogonalization. MGS denotes the modified Gram-Schmidt algorithm, and GMRES-Householder is an implementation using Householder reflections to generate the basis; see [21]. GMRES-CGS has a much larger maximum attainable accuracy than the other methods. Both GMRES-CGS and GMRES-MGS need a reorthogonalization to have a maximum attainable accuracy comparable to what is obtained with Q-OR-optinv. Note that doing a double reorthogonalization improves the results and that GMRES-Householder is worse than GMRES-MGS with reorthogonalization.

TABLE 6.1
fs 680 1c: true residual norms after 150 iterations.

Method	$\|b - Ax_{150}\|$
GMRES-CGS	$6.8377\ \ 10^{-11}$
GMRES-CGS with reorth.	$2.79327\ 10^{-14}$
GMRES-CGS with double reorth.	$1.75040\ 10^{-14}$
GMRES-MGS	$2.36046\ 10^{-13}$
GMRES-MGS with reorth.	$2.51184\ 10^{-14}$
GMRES-MGS with double reorth.	$1.59114\ 10^{-14}$
GMRES-Householder	$1.51153\ 10^{-13}$
Q-OR-optinv	$2.59770\ 10^{-14}$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Example 2. The second example is the matrix *raefsky1 3242*. This matrix is of order 3242 with 293409 nonzero entries. Its norm is 3.7095, and its condition number is $1.2885\ 10^4$. The smallest singular value is $2.8789\ 10^{-4}$. This matrix is non-normal. In Figure 6.3 we observe that the difference of the residual norms of GMRES and Q-OR-optinv is almost less than 10^{-16} for most of the iterations. We have the same conclusions as for the first example; see Figure 6.4. There is a factor of 1.8 between the maximum attainable accuracies.

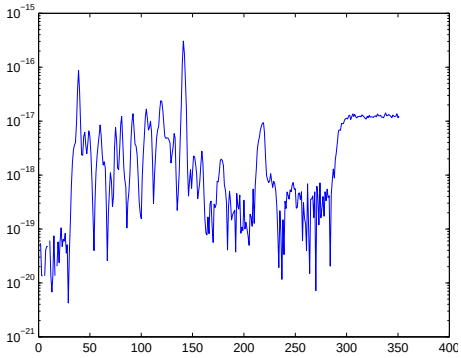

FIG. 6.3. *raefsky1 3242: difference between the residual norms of GMRES and Q-OR-optinv.*

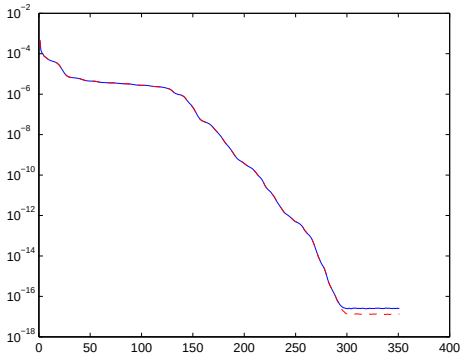

FIG. 6.4. *raefsky1 3242: residual norms of GMRES (plain) and Q-OR-optinv (dashed).*

Table 6.2 shows the maximum attainable accuracies with variants of GMRES and Q-OR-optinv. The best maximum attainable accuracy is obtained with Q-OR-optinv which is slightly better than GMRES-CGS with reorthogonalization.

TABLE 6.2
raefsky1 3242: true residual norms after 350 iterations.

Method	$\|b - Ax_{350}\|$
GMRES-CGS	$5.00382\ 10^{-11}$
GMRES-CGS with reorth.	$1.72941\ 10^{-17}$
GMRES-CGS with double reorth.	$1.83714\ 10^{-17}$
GMRES-MGS	$2.55978\ 10^{-17}$
GMRES-MGS with reorth.	$1.85462\ 10^{-17}$
GMRES-MGS with double reorth.	$1.90070\ 10^{-17}$
GMRES-Householder	$2.42497\ 10^{-17}$
Q-OR-optinv	$1.32132\ 10^{-17}$

Example 3. The third example is the symmetric matrix *Trefethen 500*. This matrix is of order 500 with 8478 nonzero entries. Its norm is $3.5712\ 10^3$, and its condition number is $3.1856\ 10^3$. The smallest singular value is 1.1210. Figure 6.5 displays the difference of the residual norms of GMRES and Q-OR-optinv as a function of the iteration number for 150 iterations. It is smaller than 10^{-14} except after the final stagnation. Figure 6.2 shows the residual norms of GMRES (plain curve) and Q-OR-optinv (dashed curve). Q-OR-optinv yields a better maximum attainable accuracy than GMRES by a factor of 11.77.

Table 6.3 displays the maximum attainable accuracies with variants of GMRES and Q-OR-optinv. The best result is obtained with Q-OR-optinv which is an order of magnitude

ETNA
Kent State University and
Johann Radon Institute (RICAM)

better than the other methods.

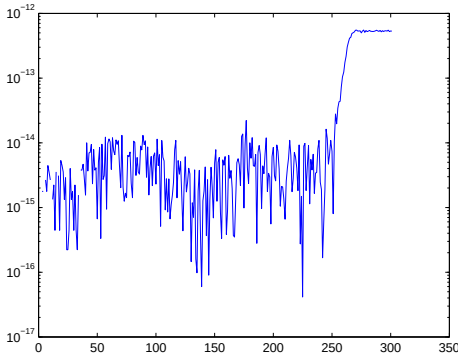

FIG. 6.5. *Trefethen 500: difference between the residual norms of GMRES and Q-OR-optinv.*

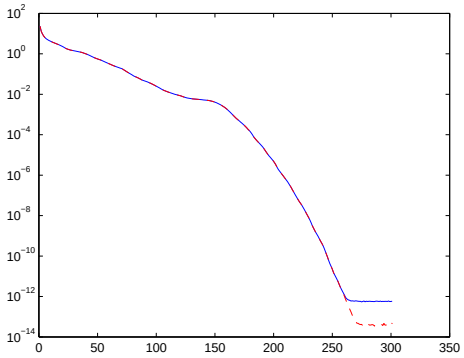

FIG. 6.6. *Trefethen 500: residual norms of GMRES (plain) and Q-OR-optinv (dashed).*

TABLE 6.3
Trefethen 500: true residual norms after 300 iterations.

Method	$\|b - Ax_{300}\|$
GMRES-CGS	$9.27328\ 10^{-12}$
GMRES-CGS with reorth.	$3.39014\ 10^{-13}$
GMRES-CGS with double reorth.	$2.93607\ 10^{-13}$
GMRES-MGS	$5.80063\ 10^{-13}$
GMRES-MGS with reorth.	$2.95675\ 10^{-13}$
GMRES-MGS with double reorth.	$3.28948\ 10^{-13}$
GMRES-Householder	$5.46732\ 10^{-13}$
Q-OR-optinv	$4.92909\ 10^{-14}$

Example 4. The fourth example arises from using the SUPG scheme (Streamwise upwind Galerkin) to discretize a convection-diffusion equation on a square with a mesh size of $1/41$; see [12]. The diffusion coefficient is 0.01. The stabilization coefficient is computed automatically. This matrix is of order 1600 and has 13924 nonzero entries. Its norm is $4.8716\ 10^{-2}$, and the condition number is 40.518. The smallest singular value is $9.8379\ 10^{-4}$. This matrix is non-normal. Once again, the maximum attainable accuracy is slightly better with Q-OR-optinv than with GMRES by a factor of 2.43; see Figures 6.7 and 6.8.

Table 6.4 displays the maximum attainable accuracies with variants of GMRES and Q-OR-optinv. Again, Q-OR-optinv gives the best maximum attainable accuracy even though GMRES-MGS with reorthogonalization is not too far away.

Example 5. The matrix *bcsstk14* of order 1806 has 63454 nonzero entries. Its norm is $1.1923\ 10^{10}$ as well as its condition number. It is a symmetric matrix. The eigenvalues are real with the smallest one being equal to 1. We use a left Gauss-Seidel preconditioner since, otherwise, the methods converge too slowly. It makes the matrix nonsymmetric. This example is quite interesting since if we continue iterating with GMRES, after reaching a minimum, the norm of the true residual increases (which is not in accordance with the theoretical properties

ETNA
Kent State University and
Johann Radon Institute (RICAM)

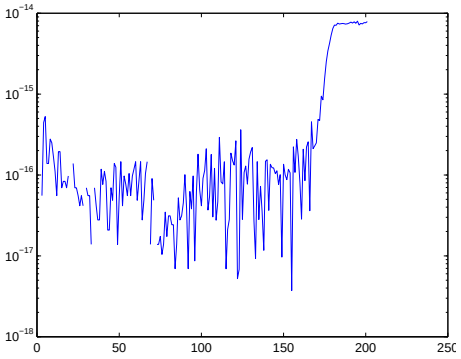

FIG. 6.7. *Supg001 1600: difference between the residual norms of GMRES and Q-OR-optinv.*

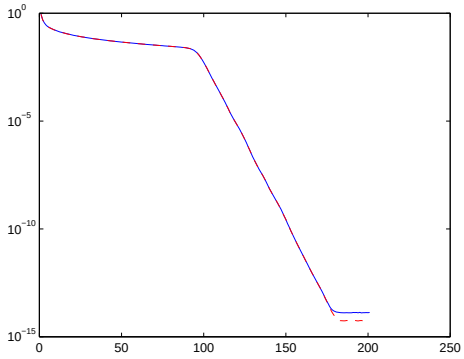

FIG. 6.8. *Supg001 1600: residual norms of GMRES (plain) and Q-OR-optinv (dashed).*

TABLE 6.4
Supg001 1600: true residual norms after 200 iterations.

Method	$\|b - Ax_{200}\|$
GMRES-CGS	$1.54043\ 10^{-13}$
GMRES-CGS with reorth.	$7.05585\ 10^{-15}$
GMRES-CGS with double reorth.	$7.23790\ 10^{-15}$
GMRES-MGS	$1.33776\ 10^{-14}$
GMRES-MGS with reorth.	$6.70649\ 10^{-15}$
GMRES-MGS with double reorth.	$6.70339\ 10^{-15}$
GMRES-Householder	$1.03229\ 10^{-14}$
Q-OR-optinv	$5.50626\ 10^{-15}$

of GMRES), whereas this is not the case with reorthogonalization or with Q-OR-optinv; see Figures 6.9 and 6.10. It happens because the matrices whose columns are the Arnoldi basis vectors are no longer orthonormal. Their smallest singular values become small after iteration 400.

Table 6.5 shows the maximum attainable accuracies with variants of GMRES and Q-OR-optinv. The best maximum attainable accuracy is given by Q-OR-optinv, but the methods with reorthogonalization give almost the same result.

TABLE 6.5
bcsstk14 1806: true residual norms after 600 iterations.

Method	$\|b - Ax_{600}\|$
GMRES-CGS	$2.77022\ 10^{-2}$
GMRES-CGS with reorth.	$1.14254\ 10^{-11}$
GMRES-CGS with double reorth.	$9.41352\ 10^{-12}$
GMRES-MGS	$1.17811\ 10^{-6}$
GMRES-MGS with reorth.	$1.15320\ 10^{-11}$
GMRES-MGS with double reorth.	$1.20222\ 10^{-11}$
GMRES-Householder	$9.34302\ 10^{-9}$
Q-OR-optinv	$1.04931\ 10^{-11}$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

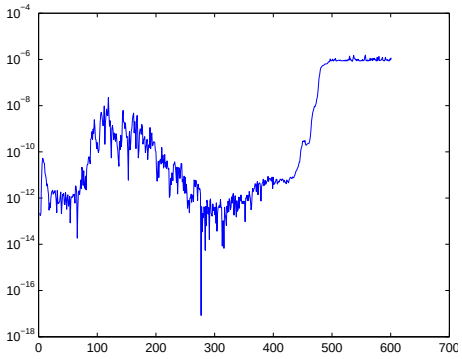

FIG. 6.9. *Bcsstk14 1806: difference between the residual norms of GMRES and Q-OR-optinv, Gauss-Seidel preconditioning.*

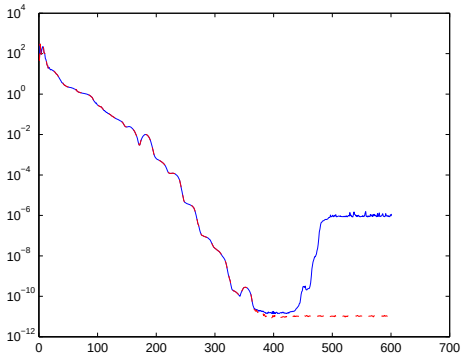

FIG. 6.10. *Bcsstk14 1806: residual norms of GMRES (plain) and Q-OR-optinv (dashed), Gauss-Seidel preconditioning.*

Other examples. Even though for most examples, Q-OR-optinv gives a maximum attainable accuracy better than or equal to GMRES-MGS, there are some problems for which it is the opposite case. One such problem is *e05r0500* of order 236. With this matrix, GMRES has a phase of quasi-stagnation for more than 150 iterations, and then the true residual norm decreases rapidly. The residual norm after 250 iterations is $2.1615\ 10^{-11}$ for GMRES and $1.0196\ 10^{-8}$ for Q-OR-optinv. There are also examples with matrices in block Jordan form for which the accuracy of Q-OR-optinv is worse than the accuracy of GMRES-MGS.

7. Handling of breakdowns. A stagnation of the residual norms in GMRES corresponds to a small (or zero) value of $v_k^T A v_k$ in step 7 of the Q-OR-optinv algorithm. One possible way to cure this breakdown is to allow the algorithm to be sub-optimal for the given iteration. This can be achieved by testing $v_k^T A v_k$ and, if it is too small, modifying the previous components of ν as $(1 - t_j)\nu_j$, where t_j is a random number in $(0, 1)$ with a uniform distribution.

Let us consider an example of a linear system of order 10. The matrix and the right-hand side are constructed using the results of [2] to obtain the following GMRES residual norms for $k = 0, 1, \ldots, 9$,

$$1, 0.9, 0.5, 0.1, 0.1, 0.1, 0.05, 0.01, 0.001, 0.0001.$$

In Figure 7.1 the plain curve shows the GMRES residual norms, the dashed curve displays the Q-OR-optinv residual norms if we do not apply any remedy to the breakdowns; we see that the breakdowns hamper convergence. The curve with stars is what we obtain when applying the remedy described above. We have an increase of the residual norm for a while but then we recover the GMRES convergence curve. However, this way of handling the possible breakdowns requires further theoretical and numerical studies.

8. Conclusion. In this paper we have shown that it is possible to construct a non-orthogonal basis for the Krylov subspace such that the Q-OR corresponding method yields the same residual norms as GMRES. Even though there are more floating point operations than in GMRES, this Q-OR optimal method gives in many cases a better maximum attainable accuracy than GMRES-MGS. It also offers more opportunity for parallelization. It remains to study the stability of the new method and to implement it on a parallel computer. Moreover, we plan to investigate truncated versions of the Q-OR method as well as the possibility of using the tridiagonal inverse of $V_k^T V_k$ when the method is used with restarts.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

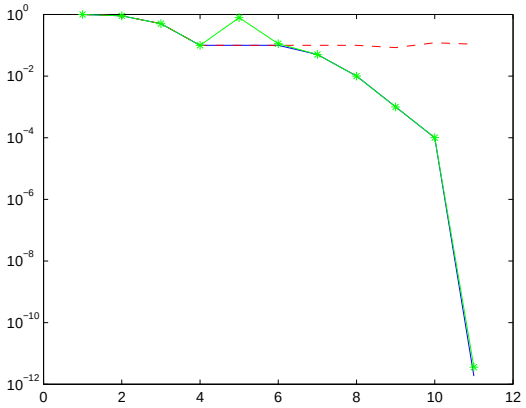

FIG. 7.1. *True residual norms, GMRES (plain), Q-OR-optinv (dashed), Q-OR-optinv with cure (stars).*

Acknowledgments. The author thanks an anonymous referee for his/her comments which helped improving the writing of this paper.

REFERENCES

[1] W. E. ARNOLDI, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[2] J. DUINTJER TEBBENS AND G. MEURANT, *Any Ritz value behavior is possible for Arnoldi and for GMRES*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 958–978.
[3] ———, *On the convergence of Q-OR and Q-MR Krylov methods for solving nonsymmetric linear systems*, BIT, 56 (2016), pp. 77-97.
[4] M. EIERMANN AND O. G. ERNST, *Geometric aspects of the theory of Krylov subspace methods*, Acta Numer., 10 (2001), pp. 251–312.
[5] R. FLETCHER, *Conjugate gradient methods for indefinite systems*, in *Numerical Analysis (Proc 6th Biennial Dundee Conf., Univ. Dundee, Dundee, 1975)*, G. A. Watson, ed., Lecture Notes in Math. vol 506, Springer, Berlin, 1976, pp. 73–89.
[6] R. W. FREUND AND N. M. NACHTIGAL, *QMR: a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[7] G. H. GOLUB, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd, ed., Johns Hopkins University Press, Baltimore, 1996.
[9] G. H. GOLUB AND G. MEURANT, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton, 2010.
[10] M. HEYOUNI AND H. SADOK, *On a variable smoothing procedure for Krylov subspace methods*, Linear Algebra Appl., 268 (1998), pp. 131–149.
[11] D. JIBETEAN AND E. DE KLERK, *Global optimization of rational functions: a semidefinite programming approach*, Math. Program., 106 (2006), pp. 93–109.
[12] J. LIESEN AND Z. STRAKOŠ, *GMRES convergence analysis for a convection-diffusion model problem*, SIAM J. Sci. Comput., 26 (2005), pp. 1989–2009.
[13] G. MEURANT, *A review on the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.
[14] ———, *Necessary and sufficient conditions for GMRES complete and partial stagnation*, Appl. Numer. Math., 75 (2014), pp. 100–107.
[15] G. MEURANT AND J. DUINTJER TEBBENS, *The role eigenvalues play in forming GMRES residual norms with non-normal matrices*, Numer. Algorithms, 68 (2015), pp. 143–165.
[16] J. NIE, J. DEMMEL, AND M. GU, *Global minimization of rational functions and the nearest GCDs*, J. Global Optim., 40 (2008), pp. 697–718.
[17] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[18] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[19] H. SADOK, *CMRH: a new method for solving nonsymmetric linear systems based on the Hessenberg reduction algorithm*, Numer. Algorithms, 20 (1999), pp. 303–321.
[20] G. STEWART, *Matrix Algorithms, Volume II: Eigensystems*, SIAM, Philadelphia, 2001.
[21] H. F. WALKER, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 152–163.
[22] J. Y. YUAN, G. H. GOLUB, R. J. PLEMMONS, AND W. A. G. CECÍLIO, *Semi-conjugate direction methods for real positive definite systems*, BIT, 44 (2004), pp. 189–207.