REVUE D'ANALYSE NUMÉRIQUE ET DE THÉORIE DE L'APPROXIMATION
Rev. Anal. Numér. Théor. Approx., vol. 36 (2007) no. 1, pp. 79–87
ictp.acad.ro/jnaat

KOVARIK'S FUNCTION ORTHOGONALIZATION ALGORITHM WITH APPROXIMATE INVERSION[*]

CONSTANTIN POPA[†]

Abstract. Z. Kovarik proposed in 1970 a method for approximate orthogonalization of a finite set of linearly independent vectors from a Hilbert space. This method uses at each iteration a symmetric and positive definite matrix inversion. In this paper we describe an algorithm in which the above matrix inversion step is replaced by an arbitrary odd degree polynomial matrix expression. We prove that this new algorithm converges to the same orthonormal set of vectors as the original Kovarik's method. Some numerical experiments presented in the last section of the paper show us that, even for small degree polynomial expressions the convergence properties of the new algorithm are comparable with those of the original one.

MSC 2000. 65F10, 65F20.

Keywords. Approximate orthogonalization of functions, Kovarik's algorithm, Gram matrix, approximate inverse.

1. A MATRIX APPROACH OF KOVARIK'S ALGORITHM

Let H be a (real) Hilbert space with the scalar product and the associated norm denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. In some space $\mathbb{R}^q$ the vectors will be considered as column vectors and the euclidean scalar product and norm will be denoted by $\langle\cdot,\cdot\rangle_2$ and $\|\cdot\|_2$, respectively. For a $q \times q$ matrix A we shall denote by $A^t, (A)_i, (A)_{ij}, \sigma(A), \rho(A)$ the transpose, i-th row, (i,j)-th element, spectrum and spectral radius, respectively, and by $\| A \|_2$, $\| A \|_\infty$ the matrix norms defined by (see e.g. [1])

$$\| A \|_2 = \sqrt{\rho(A^{\mathrm{t}}A)}, \ \| A \|_\infty = \max_{1\le i\le q} \sum_{j=1}^{q} |(A)_{ij}|. \tag{1}$$

For a linearly independent system of vectors $\Phi = \{\phi_1, \dots, \phi_n\} \subset H$ the Gram matrix $G(\Phi)$, defined by

$$(G(\Phi))_{ij} = \langle \phi_j, \phi_i \rangle, \ i,j = 1,\dots,n \tag{2}$$

[*] This work has been supported by the grant CEEX-05-D11-25/2005.

[†] Faculty of Mathematics and Computer Science, Ovidius University, Blvd. Mamaia 124, Constanţa, Romania, e-mail: cpopa@univ-ovidius.ro.

is symmetric and positive definite (SPD, for short), thus its square root $G(\Phi)^{\frac{1}{2}}$ exists and is also SPD. We define the number trace$(G(\Phi))$ by

$$\text{trace}(G(\Phi)) = \sum_{i=1}^{n} (G(\Phi))_{ii} = \sum_{i=1}^{n} \langle \phi_i, \phi_i \rangle = \sum_{i=1}^{n} \parallel \phi_i \parallel^2 . \tag{3}$$

If $\sigma(G(\Phi)) = \{\sigma_1, \dots, \sigma_n\}$, then trace$(G(\Phi)) = \sigma_1 + \dots + \sigma_n$. For the above system Φ the author considered in [1] the norm $|||\cdot|||$ defined by

$$|||\Phi|||^2 = \frac{1}{n}\text{trace}(G(\Phi)) = \frac{1}{n}\sum_{i=1}^{n} \parallel \phi_i \parallel^2 \tag{4}$$

and proved the inequality

$$|||\Phi|||^2 \leq \rho(G(\Phi)) = \parallel G(\Phi) \parallel_2, \tag{5}$$

where $\parallel G(\Phi) \parallel_2$ is the spectral norm of the matrix $G(\Phi)$. Moreover, for an $n \times n$ matrix $C = (c_{ij})_{i,j}$, he defined the product $\Psi = \Phi \cdot C$ by

$$\Psi = \{\psi_1, \dots, \psi_n\},\ \psi_i = \sum_{j=1}^{n} c_{ij}\phi_j,\ i = 1, \dots, n. \tag{6}$$

Remark 1. If we consider the systems $\Psi = (\phi_1, \dots, \phi_n), \Phi = (\psi_1, \dots, \psi_n)$ as (ordered) vectors in $H^n = H \times H \times \dots \times H$, we can write (6) in the following "matrix by vector multiplication form", very useful in computations (see Section 2)

$$\Psi = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \dots \\ \psi_n \end{bmatrix} = C^{\mathrm{t}} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \dots \\ \phi_n \end{bmatrix}. \tag{7}$$

Let $\Phi^\infty = \{\phi_1^\infty, \dots, \phi_n^\infty\}$ be given by (see (6)–(7))

$$\Phi^\infty = \Phi \cdot (G(\Phi))^{-\frac{1}{2}}. \tag{8}$$

According to [2], Φ^∞ is an orthonormal system, i.e.

$$\langle \phi_j^\infty, \phi_i^\infty \rangle = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{9}$$

and the following result holds.

Lemma 2. *Let $C = (c_{ij})_{i,j}$ be an arbitrary $n \times n$ matrix. Then*

$$|||\Phi^\infty \cdot C||| \ \leq \parallel C \parallel_2 . \tag{10}$$

In [2] Z. Kovarik proposed his approximate orthogonalization algorithm. It will be briefly presented in what follows, together with the corresponding convergence results.

Algorithm 3. Let $\Phi^0 = \Phi$; for $k = 0, 1, \dots$ do

$$G_k = G(\Phi^k); K_k = (I - G_k)(I + G_k)^{-1}; S_k = I + K_k; \Phi^{k+1} = \Phi^k \cdot S_k. \tag{11}$$

Theorem 4. *For any linearly independent system Φ, the sequence $(\Phi^k)_{k\geq 0}$ generated by* (11) *converges to Φ^∞ (in H). Moreover, the following estimate holds*

$$|||\Phi^\infty - \Phi^k||| \ \leq \parallel K_0 \parallel_2^{2^k}, \ \forall k \geq 1. \tag{12}$$

Remark 5. The above convergence of $\Phi^k = \{\phi_1^{(k)}, \dots, \phi_n^{(k)}\}$ to Φ^∞ is equivalent with (see (4))

$$\lim_{k\to\infty} \parallel \phi_i^{(k)} - \phi_i^\infty \parallel = 0, \ i = 1, \dots, n. \tag{13}$$

Now, if we denote by $\lambda_i^{(0)}$ the eigenvalues of G_0, i.e.

$$\sigma(G_0) = \{\lambda_1^{(0)}, \dots, \lambda_n^{(0)}\} \subset (0, \infty), \tag{14}$$

then, from the second equality in (11), we obtain

$$\sigma(K_0) = \left\{\tfrac{1-\lambda_1^{(0)}}{1+\lambda_1^{(0)}}, \dots, \tfrac{1-\lambda_n^{(0)}}{1+\lambda_n^{(0)}}\right\} \subset (-1, 1),$$

thus

$$\parallel K_0 \parallel_2 < \ 1, \tag{15}$$

which tells us that the convergence in (12) is at least quadratic. Moreover, from the equalities (see [2])

$$\Phi^\infty = \Phi^k \cdot G_k^{-\frac{1}{2}}, \ \forall k \geq 0, \tag{16}$$

we obtain

$$\Phi^\infty - \Phi^k = \Phi^\infty \cdot (I - G_k^{\frac{1}{2}}), \ \forall k \geq 0, \tag{17}$$

which together with (9) tell us that the limit in (13) is equivalent with

$$\lim_{k\to\infty} \parallel I - G_k^{\frac{1}{2}} \parallel_2 = 0. \tag{18}$$

Lemma 6. *The Gram matrices $G_k = G(\Phi^k), \ k \geq 0$, can be recursively generated by the following formulas*

$$G_{k+1} = S_k G_k S_k, \ k \geq 0, \tag{19}$$

with K_k and S_k computed as in (11).

Proof. From (11) and (6) we get (also using the symmetry of G_k)

$$(G_{k+1})_{ij} = \langle \phi_j^{k+1}, \phi_i^{k+1} \rangle = \sum_{q=1}^n (S_k)_{iq} (\sum_{p=1}^n (S_k)_{jp} \langle \phi_p^{(k)}, \phi_q^{(k)} \rangle) =$$

$$= \sum_{q=1}^n (S_k)_{iq} (\sum_{p=1}^n (S_k)_{jp} (G_k)_{qp}) = \langle (S_k)_i, G_k (S_k)_j \rangle_2 = (S_k G_k S_k)_{ij},$$

which completes the proof. □

2. THE MODIFIED KOVARIK ALGORITHM

As already observed by Z. Kovarik in [2], for applying the algorithm (3) we need to compute at each iteration the inverse $(I+G_k)^{-1}$ of the (SPD) matrix $I+G_k$. We shall avoid this difficulty by replacing the matrix inverse with a polynomial expression with respect to G_k. For this we shall first observe that after the scaling of the system $\Phi = \Phi^0 = \{\phi_1^{(0)}, \dots, \phi_n^{(0)}\}$

$$\psi_i^{(0)} = \frac{1}{\sqrt{\| G_0 \|_\infty + 1}} \phi_i^{(0)}, \ i = 1, \dots, n. \tag{20}$$

If $\Psi^0 = \{\psi_1^{(0)}, \dots, \psi_n^{(0)}\}$ we get

$$\| G(\Psi^0) \|_2 = \| \frac{1}{\| G_0 \|_\infty + 1} G(\Phi^0) \|_2 = \frac{\| G_0 \|_2}{\| G_0 \|_\infty + 1} < 1. \tag{21}$$

Then, for a given sequence of integers $q_k \geq 1$, $k \geq 0$, we shall approximate the inverse $(I+G_k)^{-1}$ in (11) by the truncated Neumann series $S(q_k; G_k)$ defined by

$$S(q_k; G_k) = \sum_{i=0}^{q_k} (-G_k)^i. \tag{22}$$

Then, the modified Kovarik algorithm is the following.

Algorithm 7. Starting with $\Psi^0 = \{\psi_1^{(0)}, \dots, \psi_n^{(0)}\}$ from (20), for $k \geq 0$, do

$$\Gamma_k = G(\Psi^k); K_k = (I - \Gamma_k) S(q_k; \Gamma_k); S_k = I + K_k; \Psi^{k+1} = \Psi^k \cdot S_k. \tag{23}$$

The next (main) result of the paper shows the convergence of the algorithm (7).

Theorem 8. *If $\Psi^0 = \{\psi_1^{(0)}, \dots, \psi_n^{(0)}\}$ is defined as in (20) and all the integers $q_k, k \geq 0$, are odd, then the sequence $(\Psi^k)_{k \geq 0}$ generated in (23) converges to Φ^∞ from (8) in the sense (13), i.e.*

$$\lim_{k \to \infty} \| \psi_i^{(k)} - \phi_i^\infty \| = 0, \ \forall i = 1, \dots, n. \tag{24}$$

In order to prove it we need some auxiliary results which will be presented below.

Lemma 9. *Let $\Gamma_0 = G(\Psi^0)$ and Q be an $n \times n$ orthonormal matrix such that*

$$\Gamma_0 = Q^{\mathrm{t}} \ \mathrm{diag}(\gamma_1^{(0)}, \dots, \gamma_n^{(0)}) \ Q \tag{25}$$

with (see (21))

$$\sigma(\Gamma_0) = \{\gamma_1^{(0)}, \dots, \gamma_n^{(0)}\} \subset (0,1). \tag{26}$$

Then, if q_k are all odd numbers we have

$$\Gamma_k = Q^{\mathrm{t}} \ \mathrm{diag}(\gamma_1^{(k)}, \dots, \gamma_n^{(k)}) \ Q, \tag{27}$$

with $\gamma_i^{(k)}, i = 1, \dots, n$, *recursively defined by*

$$\gamma_i^{(k+1)} = \left(\frac{2 + (\gamma_i^{(k)})^{q_k+1}(\gamma_i^{(k)} - 1)}{1 + \gamma_i^{(k)}} \right)^2 \cdot \gamma_i^{(k)} \in (0, 1), \forall k \geq 1. \tag{28}$$

Proof. According to Lemma 6 and (23) we obtain

$$\Gamma_{k+1} = S_k \Gamma_k S_k, \tag{29}$$

with

$$S_k = I + (I - \Gamma_k) S(q_k; \Gamma_k). \tag{30}$$

Now we shall use an induction argument. For this, let Γ_k be as in (27). Using the orthonormality of Q, (22) and (29)–(30) we obtain

$$S_k = Q^{\mathrm{t}} \ \mathrm{diag}(d_1^{(k)}, \dots, d_n^{(k)}) \ Q \tag{31}$$

with

$$d_i^{(k)} = 1 + (1 - \gamma_i^{(k)}) \sum_{j=0}^{q_k} (-1)^j (\gamma_i^{(k)})^j > 0, \ i = 1, \dots, n. \tag{32}$$

Then, because $\gamma_i^{(k)} \in (0, 1)$ we obtain (by also taking into account that q_k is odd)

$$d_i^{(k)} = 1 + \frac{(1 - \gamma_i^{(k)})(1 - (\gamma_i^{(k)})^{q_k+1})}{1 + \gamma_i^{(k)}}, \tag{33}$$

which gives us, together with (27) and (29),

$$\Gamma_{k+1} = Q^{\mathrm{t}} \ \mathrm{diag}(\gamma_1^{k+1}, \dots, \gamma_n^{(k+1)}) \ Q,$$

with $\gamma_i^{(k+1)}$ as in (28). So, it rests us to prove that

$$\gamma_i^{(k+1)} \in (0, 1), \ \forall i = 1, \dots, n. \tag{34}$$

For this, we consider the function $f : [0, \infty) \longrightarrow \mathbb{R}$

$$f(x) = \left(\frac{2 + x^{q_k+1}(x - 1)}{1 + x} \right)^2 \cdot x \tag{35}$$

and observe that, $\forall x \in (0, 1)$,

$$f(x) - 1 = \frac{-(1 - x)^2 - x(1 - x)x^{q_k+1}[4 - (1 - x)x^{q_k+1}]}{(1 + x)^2} < 0. \tag{36}$$

Then, from (28) and (36), it results (34) and the proof is complete. □

Lemma 10. *For any* $i \in \{1, \dots, n\}$, *if* q_k *are all odd numbers then the sequence* $(\gamma_i^{(k)})_{k \geq 0}$, *recursively defined in* (28), *is strictly increasing and*

$$\lim_{k \to \infty} \gamma_i^{(k)} = 1. \tag{37}$$

Proof. Using again $f(x)$ from (35) we get, for $x \in (0,1)$,

$$f(x) - x = x \cdot \frac{3 + x - (1-x)x^{q_k+1}}{1+x} \cdot \frac{(1-x)(1-x^{q_k+1})}{1+x} \; > \; 0. \tag{38}$$

Then, by a recursive argument we obtain that the sequence $(\gamma_i^{(k)})_{k \geq 0}$ is strictly increasing and bounded,

$$\gamma_i^{(k)} \in (\gamma_i^{(0)}, 1), \; \forall k \geq 0. \tag{39}$$

If $\gamma_i \in [0,1]$ will denote its limit, i.e.

$$\gamma_i = \lim_{k \to \infty} \gamma_i^{(k)}, \; i = 1, \dots, n, \tag{40}$$

and we shall prove that $\gamma_i = 1$. If this is not true true, i.e.

$$0 < \gamma_i < 1 \tag{41}$$

($\gamma_i \neq 0$ because $\gamma_i^{(0)} > 0$ and $(\gamma_i^{(k)})_{k \geq 0}$ is strictly increasing), then from (28) and (38) we obtain

$$\gamma_i^{(k+1)} - \gamma_i^{(k)} > \frac{\gamma_i^{(k)}(1-\gamma_i^{(k)})^2}{1+\gamma_i^{(k)}}. \tag{42}$$

The variation of the real function $g(x) = \frac{x(1-x)^2}{1+x}$, $x \in (0,1)$ gives us the relation

$$g(x) \geq \min\{g(\gamma_i^{(0)}), g(\gamma_i)\} > 0, \; \forall x \in [\gamma_i^{(0)}, \gamma_i]. \tag{43}$$

Thus, if we define $\epsilon_0 > 0$ by

$$\epsilon_0 = \min\{g(\gamma_i^{(0)}), g(\gamma_i)\}, \tag{44}$$

from (42) we obtain

$$\gamma_i^{(k+1)} > \gamma_i^{(k)} + \epsilon_0, \; \forall k \geq 0,$$

i.e.

$$\gamma_i^{(k+1)} > \gamma_i^{(0)} + k \cdot \epsilon_0, \; \forall k \geq 0. \tag{45}$$

The above inequality tells us that it exists an integer $k_0 \geq 1$ such that

$$\gamma_i^{(k+1)} > \gamma_i^{(k_0+1)} > \gamma_i,$$

which contradicts (39)–(41). It results that (37) is true and the proof is complete. □

Proof. (of Theorem 8.) From (27) and (37) we obtain that $\lim_{k \to \infty} \Gamma_k = I$, thus

$$\lim_{k \to \infty} \parallel I - \Gamma_k^{\frac{1}{2}} \parallel_2 = 0. \tag{46}$$

Then, according to (16)–(18) we shall obtain (24) if we can prove for the systems Ψ^k equalities similar with (16), i.e.

$$\Phi^\infty = \Psi^k \cdot \Gamma_k^{-\frac{1}{2}},\ \forall k \geq 0. \tag{47}$$

For this we shall use the mathematical induction. First, for $k = 0$ from (20), (21), (8) we get

$$\Psi^0 \cdot \Gamma_0^{-\frac{1}{2}} = \left(\frac{1}{\sqrt{\| G_0 \|_\infty + 1}}\, \Phi^0\right) \cdot (\sqrt{\| G_0 \|_\infty + 1}\, G_0^{-\frac{1}{2}}) = \Psi^0 \cdot G_0^{-\frac{1}{2}} = \Phi^\infty. \tag{48}$$

Let now $k \geq 0$ be an integer such that (47) holds for it. Then, using (23), (29) and (7) we successively obtain

$$\Psi^{k+1} \cdot \Gamma_{k+1}^{-\frac{1}{2}} = (\Psi^k \cdot S_k)(S_k \Gamma_k S_k)^{-\frac{1}{2}} = S_k^{\mathrm{t}} \Psi^k (S_k \Gamma_k S_k)^{-\frac{1}{2}} =$$

$$= [S_k^{\mathrm{t}} (S_k \Gamma_k S_k)^{-\frac{1}{2}}] \Psi^k = ((S_k \Gamma_k S_k)^{-\frac{1}{2}} S_k)^{\mathrm{t}} \Psi^k = \Psi^k \cdot [(S_k \Gamma_k S_k)^{-\frac{1}{2}} S_k], \tag{49}$$

which together with (27), (31) and (32) gives us the equality

$$(S_k \Gamma_k S_k)^{-\frac{1}{2}} S_k = (Q^{\mathrm{t}} D_k T_k D_k Q)^{-\frac{1}{2}} Q^{\mathrm{t}} D_k Q, \tag{50}$$

where we used the notations

$$D_k = \mathrm{diag}(d_1^{(k)}, \ldots, d_n^{(k)}),\ T_k = \mathrm{diag}(\gamma_1^{(k)}, \ldots, \gamma_n^{(k)}). \tag{51}$$

But, because D_k and T_k are diagonal matrices they commute, thus (Q being orthonormal)

$$Q^{\mathrm{t}} D_k T_k D_k Q = (Q^{\mathrm{t}} (D_k T_k D_k)^{\frac{1}{2}} Q)^2$$

and from (50), (27) and (51) we obtain

$$(S_k \Gamma_k S_k)^{-\frac{1}{2}} S_k = Q^{\mathrm{t}} D_k^{-\frac{1}{2}} T_k^{-\frac{1}{2}} D_k^{-\frac{1}{2}} Q Q^{\mathrm{t}} D_k Q = Q^{\mathrm{t}} T_k^{-\frac{1}{2}} Q = \Gamma_k^{-\frac{1}{2}}. \tag{52}$$

Then, the relations (51), (49) and the induction hypothesis give us $\Phi^\infty = \Psi^{k+1} \cdot \Gamma_{k+1}^{-\frac{1}{2}}$ and the proof is complete. □

Remark 11. In the other cases for the integers q_k, i.e. $q_k =$ constant (even) or $q_k =$ arbitrary, the modified algorithm (22)–(23) does not always converge.

3. NUMERICAL EXPERIMENTS

We considered in our experiments a regular discretization of (0, 1): $N \geq 2$, $h = 1/N,\ n = N - 1,\ x_i = ih,\ i = 1, \ldots, n$ and the piecewise linear (one dimensional) finite element basis $\Phi = \Phi^0 = \{\phi_1^{(0)}, \ldots, \phi_n^{(0)}\}$, $\phi_i^{(0)} : [0,1] \longrightarrow \mathbb{R}$ defined by (see e.g. [3])

$$\phi_i^{(0)}(x) = \begin{cases} \frac{x - x_{i-1}}{h}, & x \in (x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{h}, & x \in (x_i, x_{i+1}) \\ 0, & \text{else.} \end{cases}$$

As the Hilbert space H we considered $H_0^1((0,1))$, with the scalar product

$$\langle \phi, \psi \rangle = \int_0^1 \frac{\mathrm{d}\phi}{\mathrm{dx}} \frac{\mathrm{d}\psi}{\mathrm{dx}}.$$

In this context, the Gram matrix $G_0 = (\langle \phi_j^{(0)}, \phi_i^{(0)} \rangle)_{i,j=1,\dots,n}$ is given by

$$G_0 = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

We first applied the original Kovarik algorithm (11) (using (19)), for different values of $N \geq 2$ and the stopping test

$$\| G_{k+1} - G_k \|_\infty \leq 10^{-3}. \tag{53}$$

The numbers of iterations for obtaining (53) are described in Table 1.

Then, for $N = 128$ we applied the modified Kovarik algorithm (22)–(23) (with (29)) for q_k = constant (odd), $\forall k \geq 0$ and the same stopping rule (53). The numbers of iterations for these tests are presented in Table 2.

The last tests were made as those from above, but only for the first three (odd) values of q_k and different values of $N \geq 2$. The numbers of iterations are described in Table 3.

FINAL REMARKS AND COMMENTS

(1) We observe that for very small values of q_k (which means less computational effort per iteration in (22)–(23)), we got good enough results (Tables 2 and 3) by comparing them with those for the algorithm (11) (Table 1).
(2) All the tests indicate a "mesh-independent" behaviour for both Kovarik, original and modified algorithms.
(3) All the numerical experiments were performed with the numerical linear algebra software "OCTAVE", freely available on the Internet.

Table 1. Algorithm 3

$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
7	8	9	10	11

Table 2. Algorithm 7 with $N = 128$

$q_k = 1$	$q_k = 3$	$q_k = 5$	$q_k = 7$	$q_k = 9$	$q_k = 11$
28	22	19	17	16	15

Table 3. Algorithm 7

	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$
$q_k=1$	16	26	27	28	29
$q_k=3$	19	20	21	22	23
$q_k=5$	16	17	18	19	20

REFERENCES

[1] Björck, A., *Numerical methods for least squares problems*, SIAM, Philadelphia, 1996.

[2] Kovarik, Z., *Some iterative methods for improving orthogonality*, SIAM J. Numer. Anal **7(3)**, pp. 386–389, 1970.

[3] Trottenberg, U., Oosterlee, C. and Schüller, A., *Multigrid*, Academic Press, New York, 2001.

Received by the editors: September 11, 2006.