ON NUMERICAL SOLUTION OF ARBITRARY SYMMETRIC LINEAR SYSTEMS BY APPROXIMATE ORTHOGONALIZATION [1]

CONSTANTIN POPA

Abstract. For many important real world problems, after the application of appropriate discretization techniques we can get symmetric and relatively dense linear systems of equations (e.g. those obtained by collocation or projection-like discretization of first kind integral equations). Usually, these systems are rank-defficient and (very) ill-conditioned, thus classical direct or iterative solvers can not be efficiently applied. Moreover, because they are relatively dense, classical preconditioning techniques (as e.g. Incomplete Decomposition) become useless.
In the present paper we describe an iterative solver for relatively dense symmetric linear systems, in classical or least-squares formulation. The method is based on a modified version of Kovarik's approximate orthogonalization algorithm. We prove that the sequence of approximations so generated converges to the minimal norm solution of the system. Numerical experiments are described for a collocation discretization of two first kind integral equations, one of them appearing in inverse problems related to determination of charge distribution that generates a prescribed electric field.

MSC2000: 65F10, 65F20, 65R20: **AMR:** 106C, 374E

Key words and phrases: symmetric Kovarik algorithm, least-squares problems, minimal norm solution, first kind integral equations.

1 The symmetric Kovarik algorithm

Let A be an $n \times n$ symmetric matrix and $(A)_i, A^t, A^+$ its i-th row, transpose and Moore-Penrose pseudoinverse (see [1]), respectively. By $gk_2(A)$ we shall denote its generalized spectral condition number defined as the square root of the ratio between the biggest and smallest singular values; $\langle \cdot, \cdot \rangle, \| \cdot \|$ will be the Euclidean scalar product and norm on some space $\mathbb{R}^q$. We shall also use the notations $\sigma(B), \rho(B), \| B \|, R(B), N(B)$ for the spectrum, spectral

[1]The paper was supported by the collaboration that the author has with "Gheorghe Mihoc - Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Bucharest.

radius, spectral norm, range and null space of a square matrix B, respectively. All the vectors appearing in the paper will be considered as column vectors. Let $(a_k)_{k\geq 0}$ be the sequence of reals defined $a_j = \frac{1}{2^{2j}}\frac{(2j)!}{(j!)^2}$, $j \geq 0$ and $(q_k)_{k\geq 0}$ a given sequence of positive integers. In [8] we considered the following "symmetric" version of Kovarik's algorithm from [4] (Algorithm A, page 386), extended by the author in [9].

Algorithm KOAS. Let $A_0 = A$; for $k = 0, 1, \ldots,$ do

$$H_k = I - A_k,\ A_{k+1} = f_k(H_k)A_k, \tag{1}$$

where $f_k : I\!R \longrightarrow I\!R$ is the function defined by

$$f_k(x) = 1 + a_1 x + \ldots + a_{q_k} x^{q_k},\ k \geq 0. \tag{2}$$

Theorem 1 *([8]) Let us suppose that the symmetric matrix A satisfies*

$$\| A \| = \rho(A) < 1,\ \langle Ax, x\rangle \ \geq\ 0,\ \forall x \in I\!R^n. \tag{3}$$

Then, the sequence $(A_k)_{k\geq 0}$ generated with the algorithm **KOAS** *converges and*

$$\lim_{k\to\infty} A_k = A^+ A. \tag{4}$$

Remark 1 *For the generalized spectral conditions numbers of the matrices $A_k, k \geq 0$ the following holds*

$$\lim_{k\to\infty} gk_2(A) = gk_2(A^+A) = 1. \tag{5}$$

Thus, the **KOAS** *algorithm can be used as an "iterative preconditioner" for the matrix A (see e.g. [3]).*

The linear convergence of the algorithm **KOAS** is described in the following result (proved in [10]).

Theorem 2 *The algorithm* **KOAS** *converges linearly, i.e.*

$$\| A_k - A^+ A \| \ \leq\ \left(1 - \frac{\lambda_{min}}{2}\right)^k \| A - A^+ A \|,\ \forall k \geq 0, \tag{6}$$

where λ_{min} is the smallest positive eigenvalue of A.

Remark 2 *Another interesting property of the* **KOAS** *algorithm is related to the relation* $A^+A = P_{R(A^t)}$ *(see e.g. [1]), where by* P_S *we denoted the orthogonal projection onto the vector subspace* $S \subset I\!R^n$. *But, because the matrix* A *is symmetric we get* $A^+A = P_{R(A)}$ *and* $I - A^+A = P_{N(A)}$. *Thus, for a given vector* $b \in I\!R^n$, *from (4) it results that*

$$\lim_{k\to\infty} A_k b = P_{R(A)}(b), \quad \lim_{k\to\infty} (I - A_k)b = P_{N(A)}(b). \tag{7}$$

2 Application to least-squares problems

Let $b \in I\!R^n$ be a given vector. We shall consider the linear least-squares formulation: find $x^* \in I\!R^n$ such that

$$\| Ax^* - b \| = \min\{\| Ax - b \|, x \in I\!R^n\}. \tag{8}$$

In what follows we shall construct an algorithm for the approximation of the minimal norm solution, x_{LS} of (8). The algorithm uses the above **KOAS** together with a recursive modification of the right hand side.

Algorithm KOAS-rhs. Let $A_0 = A, b^0 = b$; for $k = 0, 1, 2, \ldots$ do

$$H_k = I - A_k, \; b^{k+1} = f_k(H_k)b^k, \; A_{k+1} = f_k(H_k)A_k. \tag{9}$$

The following result proves the convergence of **KOAS-rhs** algorithm.

Theorem 3 *In the hypothesis of Theorem 1 the sequence* $(A_k b^k)_{k\geq 0}$ *generated by* (9) *converges and*

$$\lim_{k\to\infty} A_k b^k = A^+ b = x_{LS} \tag{10}$$

Proof. Because A is symmetric we have $N(A^t) = N(A)$, thus

$$b = P_{R(A)}(b) + P_{N(A)}(b), \tag{11}$$

with

$$P_{R(A)}(b) = Ax \tag{12}$$

for some $x \in I\!R^n$. Let $r = rank(A) \leq n$ and Q an $n \times n$ orthogonal matrix such that

$$A = Q\mathrm{diag}(\lambda_1^{(0)}, \ldots, \lambda_r^{(0)}, 0, \ldots, 0)Q^t, \tag{13}$$

where $\lambda_i^{(0)}$ are the nonzero eigenvalues of A. Then, as in [6] we obtain

$$A_k = Q\text{diag}(\lambda_1^{(k)}, \ldots, \lambda_r^{(k)}, 0, \ldots, 0)Q^t.$$

Combining this with (1)-(2) we see that

$$f_k(H_k) = Q\text{diag}(f_k(1-\lambda_1^{(k)}), \ldots, f_k(1-\lambda_r^{(k)}), f_k(1), \ldots, f_k(1))Q^t. \quad (14)$$

On the other hand we know that

$$A^+ = Q\text{diag}(\frac{1}{\lambda_1^{(0)}}, \ldots, \frac{1}{\lambda_r^{(0)}}, 0, \ldots, 0)Q^t, \quad (15)$$

and thus we obtain

$$P_{N(A)} = I - P_{R(A)} = I - A^+A = Q\text{diag}(0, \ldots, 0, 1, \ldots, 1)Q^t. \quad (16)$$

From (16) and (14) we get

$$f_k(H_k)P_{N(A)} = Qdiag(0, \ldots, 0, f_k(1), \ldots, f_k(1))Q^t = f_k(1)P_{N(A)}. \quad (17)$$

Now, from (9), (11), (12) and (17) we get

$$b^1 = f_0(H_0)b^0 = f_0(H_0)b = f_0(H_0)P_{R(A)}(b) +$$

$$f_0(H_0)P_{N(A)}(b) = f_0(H_0)Ax + f_0(1)P_{N(A)}(b) = A_1x + f_0(1)P_{N(A)}.$$

Using a recursive argument and also (14), (16) and (12), we obtain

$$b^k = A_kx + f_0(1)\cdots f_k(1)P_{N(A)}(b),\ \forall k \geq 0. \quad (18)$$

By mathematical induction we can prove without difficulty that $\forall k \geq 0$, $\langle A_kx, x\rangle \geq 0, \forall x \in I\!R^n$. Moreover, in [8] we have showed that, if $\lambda_i^{(0)} \in (0,1)$ then $f_k(1-\lambda_i^{(k)})\lambda_i^{(k)} \in (0,1), \forall k \geq 0$. From these consideration it results that the matrices H_k are symmetric and positive definite, so will be also $f_k(H_k)$, $\forall k \geq 0$. Using the invertibility of the matrices $f_k(H_k)$ and (9) we can easily prove

$$N(A_k) = N(A),\ \forall k \geq 0,$$

which together with (18) gives us

$$A_kb^k = A_kA_kx = A_k^2x,\ \forall k \geq 0. \quad (19)$$

From (12), (4), (19) and the properties of the Moore-Penrose pseudoinverse A^+ (see e.g. [1]) we then have

$$\lim_{k\to\infty} A_k b^k = \lim_{k\to\infty} A_k A_k x = (A^+A)(A^+A)x =$$

$$(A^+AA^+)(Ax) = A^+Ax = A^+P_{R(A)}(b) = x_{LS}$$

and the proof is complete.

Corollary 1 *In the hypothesis of the above theorem we have*

$$\lim_{k\to\infty} \parallel b^k \parallel = +\infty. \tag{20}$$

Proof. Because $q_k \geq 1, a_k > 0, \forall k \geq 0$ we obtain that

$$f_k(1) = a_0 + a_1 + \ldots + a_{q_k} \geq a_0 + a_1 = 1 + \frac{1}{2} = \frac{3}{2}. \tag{21}$$

From (21), (18) and the orthogonality of the subspaces $N(A)$ and $R(A)$ we then obtain

$$\parallel b^k \parallel^2 = \parallel A_k x \parallel^2 + (f_0(1) \ldots f_k(1))^2 \parallel P_{N(A)}(b) \parallel^2 \geq (\frac{3}{2})^k \parallel P_{N(A)}(b) \parallel^2 \to \infty,$$

because in the inconsistent case $P_{N(A)}(b) \neq 0$. This completes the proof.

3 Numerical experiments

Test problem P1 (see [7])
For a given function $y \in L^2([0,1])$, find $x^* \in L^2([0,1])$ such that

$$\int_0^1 k(s,t)x(t)dt \;=\; y(s),\; s \in [0,1], \tag{22}$$

with

$$k(s,t) = \frac{1}{1+|s-0.5|+t},\; y(s) = \begin{cases} \ln\frac{2.5-s}{1.5-s}, & s \in [0,0.5) \\ \\ \ln\frac{1.5+s}{0.5+s}, & s \in [0.5,1] \end{cases} \tag{23}$$

Remark 3 *The right hand side y was defined as in* (23) *such that the equation* (22) *has the solution $x(t) = 1, \forall t \in [0,1]$.*

We discretized (22)-(23) by the collocation algorithm from [5], with the collocation points

$$s_i = (i-1)\frac{1}{n-1},\ i = 1, 2, \ldots, n. \tag{24}$$

Thus, we obtained the symmetric system

$$Ax = b, \tag{25}$$

with the $n \times n$ matrix A and $b \in I\!R^n$ given by

$$A_{ij} = \int_0^1 k(s_i, t)k(s_j, t)dt = \begin{cases} \frac{1}{\alpha_i(1+\alpha_i)}, & \text{if } \alpha_i = \alpha_j, \\ \\ \frac{1}{\alpha_i - \alpha_j} \ln \frac{(1+\alpha_j)\alpha_i}{(1+\alpha_i)\alpha_j}, & \text{if } \alpha_i \neq \alpha_j, \end{cases} \qquad b_i = y(s_i), \tag{26}$$

where

$$\alpha_i = 1 + |s_i - \frac{1}{2}|, \quad i = 1, \ldots, n. \tag{27}$$

For $n \geq 3$ we observe that the matrix A from (26) is positive semi-definite with

$$rank(A) = \begin{cases} \frac{n+1}{2}, & \text{if } n \text{ is odd} \\ \\ \frac{n}{2}, & \text{if } n \text{ is even.} \end{cases} \tag{28}$$

We also observe that, because the problem (22)- (23) is consistent (see Remark 3) so will be the system (25). Then, in order to get an inconsistent problem (8) we considered a perturbation of the right hand side b of the form

$$b := b + \delta b, \tag{29}$$

with $\delta b \in I\!R^n$ a randomly generated vector such that $\| \delta b \| = 5\% \| b \|$.

Test problem P2 - Determination of charge distribution generating a given electric field (inverse problem,(see [2], (1.13) - particular case)

For a given function $y \in L^2([0,1])$, find $x^* \in L^2([0,1])$ such that

$$\int_0^1 k(s,t)x(t)dt \ = \ y(s),\ s \in [0,1], \tag{30}$$

with

$$k(s,t) = \frac{1}{\sqrt{(1+(s-t)^2)^3}}, \quad y(s) = s. \tag{31}$$

We discretized (30)-(31) as **P1** before and we obtained the symmetric and positive semidefinite definite system

$$\hat{A}x = b, \tag{32}$$

with the $n \times n$ matrix $\hat{A}$ and $b \in I\!R^n$ given by

$$\hat{A}_{ij} = \int_0^1 k(s_i, t)k(s_j, t)dt, \quad b_i = y(s_i). \tag{33}$$

But, because the exact values $(\hat{A})_{ij}$ from (33) can not be analitically obtained, we approximated them by the rectangles ("midpoint") quadrature formula, with 16 equally spaced points in $[0, 1]$ and we obtained an $n \times n$ matrix A. Moreover, for $n = 16,\ 32,\ 64,\ 128,\ 256$ the matrix A is rank deficient and the system (25) inconsistent (i.e. we must reformulate it also as in (8)). Both the above symmetric least-squares formulations are (very) ill-conditioned ($gk_2(A) \geq 10^8$) and sensitive to round-off errors. Moreover, Kovarik-like algorithms for such kind of problems must be applied with "special care", with respect to (theoretically) zero eigenvalues that can grow and damage the results if the number of iteartions used exceeds a certain value (see for details in this sense the analysis made in [6]). For this, we considered the following numerical "strategy" w.r.t. their numerical solution:
a) for problem **P1** - we considered the absolute ($abser$) and relative ($reler$) errors (with respect to the exact minimal norm solution of (8) x_{LS}), defined by

$$abser = \parallel A_k b^k - x_{LS} \parallel, \quad reler = \frac{abser}{\parallel A_k b^k \parallel}. \tag{34}$$

Then, for $n = 8$ we determined the number of iterations for which

$$reler \leq 0.5, \tag{35}$$

we fixed this number and ran the computer code for other (bigger, more realistic) values of n. In each case we computed $abser$, $reler$, and $stop$, given by the normal equation residual

$$stop = \parallel A^t(AA_k b^k - b) \parallel .$$

The results are presented in Table 1 below for three different choices of the integers q_k.
b) for problem **P2** - the same as for **P1**, but the number of iterations for (35) was determined using the particular value $n = 32$. The results are described in Table 2.

Remark 4 *We have also to observe that for all dimensions n and for both problems the value* $\| x_{LS} \|$ *was very big (of order* $10^7 - 10^9$*), thus the absolute error abser appears so big. On the other hand, we verified the accuracy of the approximations in each case by plotting the corresponding solution; we then observed that, up to a unit difference concerning the order of magnitude, the solutions have the same "shape" (variation) as the exact one* x_{LS} *which for such big numerical vales is an enough good comparison criterion.*

Note. All the computations were made with the Numerical Linear Algebra software package OCTAVE, freely available under the terms of the GNU General Public License, see `www.octave.org`.

Table 1. Number of iterations: 70/46/37									
n	$q_k = 1, \forall k$			$q_k = 2, \forall k$			$q_k = 3, \forall k$		
	stop	abser	reler	stop	abser	reler	stop	abser	reler
8	10^{-6}	10^8	0.36	10^{-7}	10^8	0.28	10^{-6}	10^8	0.38
16	10^{-5}	10^6	0.11	10^{-5}	10^7	0.1	10^{-6}	10^7	0.15
32	10^{-5}	10^6	0.09	10^{-5}	10^7	0.16	10^{-7}	10^7	0.38
64	10^{-5}	10^7	0.07	10^{-5}	10^7	0.2	10^{-5}	10^7	0.37
128	10^{-5}	10^6	0.09	10^{-4}	10^7	0.11	10^{-4}	10^7	0.59
256	10^{-4}	10^6	0.27	10^{-4}	10^7	0.35	10^{-4}	10^7	0.37

Table 1: Results for the problem P1

Table 2. Number of iterations: 73/47/39									
n	$q_k = 1, \forall k$			$q_k = 2, \forall k$			$q_k = 3, \forall k$		
	stop	abser	reler	stop	abser	reler	stop	abser	reler
32	10^{-6}	10^7	0.27	10^{-6}	10^7	0.47	10^{-6}	10^7	0.24
64	10^{-6}	10^7	0.31	10^{-6}	10^7	0.52	10^{-6}	10^7	0.26
128	10^{-5}	10^7	0.34	10^{-5}	10^7	0.56	10^{-6}	10^7	0.27
256	10^{-6}	10^7	0.36	10^{-6}	10^7	0.60	10^{-6}	10^7	0.3

Table 2: Results for the problem P2

References

[1] Boullion, L. T. and Odell, P. L., *Generalized inverse matrices*, Wiley-Interscience, New York, 1971.

[2] Engl, H.W., Hanke, M., Neubauer A., *Regularization of inverse problems*, Kluwer Academic Publ., Dordrecht, 2000.

[3] Evans, D. J. and Popa, C. , *Projections and preconditioning for inconsistent least-squares problems*, Intern. J. Computer Math., **78(4)**(2001), 599-616.

[4] Kovarik, S., *Some iterative methods for improving orthogonality*, SIAM J. Num. Anal., **7(3)**(1970), 386-389.

[5] Kress, R., *Linear integral equations*, Springer Verlag, Berlin, 1989.

[6] Mohr, M., Popa, C., Rüde, U., *Regularization by a Kovarik type algorithm for symmetric matrices*, to appear as technical report, Lehrstuhl für Informatik 10 (Systemsimulation), FAU Erlangen-Nürnberg.

[7] Pelican, E., Popa C., *Some remarks on a collocation method for first kind integral equations*, Report **03-1**, Lehrstuhl für Informatik 10 (Systemsimulation), FAU Erlangen-Nürnberg.

[8] Popa, C., *On a modified Kovarik algorithm for symmetric matrices*, Annals of "Ovidius" Univ. Constanta, Series Mathematics, vol.**XI(1)**(2003), 147-156.

[9] Popa, C., *Extension of an approximate orthogonalization algorithm to arbitrary rectangular matrices*, Linear Alg. Appl., **331**(2001), 181-192.

[10] Popa, C., *Some properties and applications of a modified Kovarik algorithm*, to appear in Bulletin of "Politehnica" University of Timisoara.