Solution of $f(A)x = b$ with projection methods for the matrix A

Henk A. van der Vorst

Utrecht Universtity, Department of Mathematics, Utrecht, The Netherlands

Abstract. In this paper, we expand on an idea for using Krylov subspace information for the matrix A and the vector b. This subspace can be used for the approximate solution of a linear system $f(A)x = b$, where f is some analytic function. We will make suggestions on how to use this for the case where f is the matrix *sign* function.

1 Introduction

The matrix *sign* function plays an important role in QCD computations, see for instance [12]. In the computational models one has to compute an approximate solution for linear systems of the type

$$(B + \mathrm{sign}(A))x = b, \tag{1}$$

with $A, B \in \mathbb{C}^{n\times n}$, and A and B do not commute. The latter property is an important bottleneck for the efficient computation of subspaces that can be used for the reduction of both A and B.

In [15] an approach was suggested for the usage of a Krylov subspace for the matrix A and a given vector, for instance b, for the computation of approximate solutions of linear systems

$$f(A)x = b,$$

with f an analytic function. The approach in [15] was motivated by the function $f(A) = A^2$, which plays a role in the solution of some biharmonic systems. Furthermore, the proposed methods were outlined for the case that $A = A^H$. However, the approach is easily generalized for non-symmetric complex matrices, as we will see in this paper. We have to pay more attention to the evaluation of f for the reduced system, associated with the Krylov subspace.

In particular, we will discuss some possible approaches for using the Krylov subspace for the computation of $\mathrm{sign}(A)p$ for given vectors p. With the evaluation of the matrix *sign* function one has to be extremely careful. A popular approach, based on a Newton iteration converges fast, but is sensitive for rounding errors, especially when A is ill-conditioned. We will briefly discuss a computational method that was suggested (and analyzed) by Bai and Demmel [3]. This approach can also be combined, in principle, with the subspace reduction technique. Our early experiments, not reported here, indicate that the actual computation of approximate solutions of (1) is complicated because of the occurrence of the matrix B. Since this matrix does not share an eigenvector basis with A, there is little hope that the subspace generated with A can also be used efficiently with B. It seems that we have to experiment with either nested approaches or with mixed subspaces. The current state of the art is that we still have a way to go in our quest for an efficient computational technique for the very large systems that arise in QCD modeling.

2 Krylov subspaces

Krylov subspace methods are well-established methods for the reduction of large linear systems of equations to much smaller size problems. We will explain briefly the idea behind Krylov subspace methods. Given a linear system $Ax = b$, with a large, usually sparse, unsymmetric nonsingular matrix A, then the standard Richardson iteration

$$x_k = (I - A)x_{k-1} + b$$

generates approximate solutions in shifted Krylov subspaces

$$x_0 + K^k(A; r_0) = x_0 + \{r_0, Ar_0, \ldots, A^{k-1}r_0\},$$

with $r_0 = b - Ax_0$, for some given initial vector x_0.
The Krylov subspace projection methods fall in three different classes:

1. The *Ritz-Galerkin approach*: Construct the x_k for which the residual is orthogonal to the current subspace: $b - Ax_k \perp K^k(A; r_0)$.
2. The *minimum residual approach*: Identify the x_k for which the Euclidean norm $||b - Ax_k||_2$ is minimal over $K^k(A; r_0)$.
3. The *Petrov-Galerkin approach*: Find an x_k so that the residual $b - Ax_k$ is orthogonal to some other suitable k-dimensional subspace.

The Ritz-Galerkin approach leads to such popular and well-known methods as Conjugate Gradients, the Lanczos method, FOM, and GENCG. The minimum residual approach leads to methods like GMRES, MINRES, and ORTHODIR. If we select the k-dimensional subspace in the third approach as $K^k(A^H; s_0)$, then we obtain the Bi-CG, and QMR methods. More recently, hybrids of the three approaches have been proposed, like CGS, Bi-CGSTAB, BiCGSTAB(ℓ), TFQMR, FGMRES, and GMRESR.

A nice overview of Krylov subspace methods, with focus on Lanczos-based methods, is given in [7]. Simple algorithms and unsophisticated software for some of these methods is provided in [4]. Iterative methods with much attention to various forms of preconditioning have been described in [2]. A good

overview on iterative methods was published by Saad [14]; it is very algorithm oriented, with, of course, a focus on GMRES and preconditioning techniques, like threshold ILU, ILU with pivoting, and incomplete LQ factorizations.
An annotated entrance to the vast literature on preconditioned iterative methods is given in [5].

In order to identify optimal approximate solutions in the Krylov subspace we need a suitable basis for this subspace, one that can be extended in a meaningful way for subspaces of increasing dimension. The obvious basis r_0, Ar_0, ..., $A^{i-1}r_0$, for $K^i(A; r_0)$, is not very attractive from a numerical point of view, since the vectors $A^j r_0$ point more and more in the direction of the dominant eigenvector for increasing j and hence the basis vectors will have small mutual angles. This leads to numerically unstable processes.

Instead of the standard basis one usually prefers an orthonormal basis, and Arnoldi [1] suggested to compute this basis as follows. Start with $v_1 \equiv r_0/\|r_0\|_2$. Assume that we have already an orthonormal basis v_1, ..., v_j for $K^j(A; r_0)$, then this basis is expanded by computing $t = Av_j$, and by orthonormalizing this vector t with respect to v_1, ..., v_j. In principle the orthonormalization process can be carried out in different ways, but the most commonly used approach is to do this by a modified Gram-Schmidt procedure [9]. This leads to an algorithm for the creation of an orthonormal basis for $K^m(A; r_0)$, as in Fig 1.

```
v_1 = r_0/||r_0||_2;
for j = 1,..,m-1
    t = A v_j;
    for i = 1,...,j
        h_{i,j} = v_i^H t;
        t = t - h_{i,j} v_i;
    end;
    h_{j+1,j} = ||t||_2;
    v_{j+1} = t/h_{j+1,j};
end
```

Fig. 1. Arnoldi's method with modified Gram–Schmidt orthogonalization

It is easily verified that v_1, ..., v_m form an orthonormal basis for the Krylov subspace $K^m(A; r_0)$ (that is, if the construction does not terminate at a vector $t = 0$). The orthogonalization leads to relations between the v_j, that can be formulated in a compact algebraic form. Let V_j denote the matrix with columns v_1 up to v_j, then it follows that

$$AV_{m-1} = V_m H_{m,m-1}. \tag{2}$$

The m by $m-1$ matrix $H_{m,m-1}$ is upper Hessenberg, and its elements $h_{i,j}$ are defined by the Arnoldi algorithm.

From a computational point of view, this construction is composed from three basic elements: a matrix vector product with A, innerproducts, and updates. We see that this orthogonalization becomes increasingly expensive for increasing dimension of the subspace, since the computation of each $h_{i,j}$ requires an inner product and a vector update.

Note that if A is symmetric, then so is $H_{m-1,m-1} = V_{m-1}^H A V_{m-1}$, so that in this situation $H_{m-1,m-1}$ is tridiagonal. This means that in the orthogonalization process, each new vector has to be orthogonalized with respect to the previous two vectors only, since all other innerproducts vanish. The resulting three term recurrence relation for the basis vectors of $K^m(A; r_0)$ is known as the *Lanczos method* and some very elegant methods are derived from it. In the symmetric case the orthogonalization process involves constant arithmetical costs per iteration step: one matrix vector product, two innerproducts, and two vector updates.

3 Reduced Systems

With equation (2) we can construct approximate solutions for $Ax = b$ in the Krylov subspace $K^m(A; r_0)$. These approximate solutions can be written as $x_m = x_0 + V_m y$, with $y \in \mathbb{R}^n$, since the columns of V_m span a basis for the Krylov subspace. The Ritz-Galerkin orthogonality condition for the residual leads to

$$b - Ax_m \perp \{v_1, \ldots, v_m\},$$

or

$$V_m^H (b - A(x_0 + V_m y)) = 0.$$

Now we use that $b - Ax_0 = r_0 = ||r_0||_2 v_1$, and with (2) we obtain

$$H_{m,m} y = ||r_0|| e_1, \tag{3}$$

with e_1 the first canonical basis vector in $\mathbb{R}^m$. If $H_{m,m}$ is not singular then we can write the approximate solution x_m as

$$x_m = ||r_0||_2 V_m H_{m,m}^{-1} e_1. \tag{4}$$

Note that this expression closely resembles the expression $x = A^{-1}b$ for the exact solution of $Ax = b$. The matrix $H_{m,m}$ can be interpreted as the restriction of A with respect to $v_1, \ldots, v_m$. The vector $||r_0||e_1$ is the expression for the right-hand side with respect to this basis, and V_m is the operator that expresses the solution of the reduced system (in $\mathbb{R}^m$) in terms of the canonical basis for $\mathbb{R}^n$.

Let us from now on assume that $x_0 = 0$. This simplifies the formulas, but is does not pose any further restriction. With $x_0 \neq 0$ we have for $x = w + x_0$ that $A(w + x_0) = b$ or $Aw = b - x_0 = \widetilde{b}$, and we are again in the situation that the initial approximation w_0 for w is $w_0 = 0$.

We can also use the above mechanism for the solution of more complicated systems of equations. Suppose that we want to find approximate solutions for $A^2x = b$, with only the Krylov subspace for A and $r_0 = b$ available. The solution of $A^2x = b$ can be realized in two steps

1. Solve z_m from $Az = b$, using the Ritz-Galerkin condition. With $z = V_m y$ and (2), we have that
$$z = \|b\|_2 V_m H_{m,m}^{-1} e_1.$$
2. Solve x_m from $Ax_m = z_m$, with $x_m = V_m u$. It follows that
$$AV_m u = \|b\|_2 V_m H_{m,m}^{-1} e_1,$$
$$V_{m+1} H_{m+1,m} u_m = \|b\|_2 V_m h_{m,m}^{-1}.$$
The Ritz-Galerkin condition with respect to V_m leads to
$$H_{m,m} u_m = \|b\|_2 H_{m,m}^{-1} e_1.$$

These two steps lead to the approximate solution

$$x_m = \|b\|_2 V_m H_{m,m}^{-2} e_1. \tag{5}$$

If we compare (4) for $Ax = b$ with (5) for $A^2x = b$, then we see that the operation with A^2 translates to an operation with $H_{m,m}^2$ for the reduced system and that is all.

Note that this approximate solution x_m does not satisfy a Ritz-Galerkin condition for the system $A^2x = b$. Indeed, for $x_m = V_m y$, we have that

$$A^2 V_m y = AV_{m+1} H_{m+1,m} y = V_{m+2} H_{m+2,m+1} H_{m+1,m} y.$$

The Ritz-Galerkin condition with respect to V_m, for $b - Ax_m$, leads to

$$V_m^H V_{m+2} H_{m+2,m+1} H_{m+1,m} y = \|b\|_2 e_1.$$

A straight-forward evaluation of $H_{m+2,m+1} H_{m+1,m}$ and the orthogonality of the v_j's, leads to

$$V_m^H V_{m+2} H_{m+2,m+1} H_{m+1,m} = H_{m,m}^2 + h_{m+1,m} h_{m,m+1} e_m e_m^T.$$

That means that the reduced matrix for A^2, expressed with respect to the V_m basis, is given by the matrix $H_{m,m}^2$ in which the bottom right element $h_{m,m}$

is updated with $h_{m+1,m}h_{m,m+1}$. By computing x_m as in (5), we have ignored the factor $h_{m+1,m}h_{m,m+1}$. This is acceptable, since in generic situations the convergence of the Krylov solution process for $Ax = b$ goes hand in hand with small elements $h_{m+1,m}$.

We can go one step further, and try to solve

$$(A^2 + \alpha A + \beta I)x = b,$$

with Krylov subspace information obtained for $Ax = b$ (with $x_0 = 0$). The Krylov subspace $K^m(A, r_0)$ is shift invariant, that is

$$K^m(A, r_0) = K^m(A - \sigma I, r_0),$$

for any scalar $\sigma \in \mathbb{C}$. The matrix polynomial $A^2 + \alpha A + \beta I$ can be factored into

$$A^2 + \alpha A + \beta I = (A - \omega_1 I)(A - \omega_2 I).$$

Proceeding as for A^2, that is solving the given system in two steps, and imposing a Ritz-Galerkin condition for each step, leads to the approximate solution

$$x_m = ||b||_2 V_m (H_{m,m}^2 + \alpha H_{m,m} + \beta I_m)^{-1} e_1.$$

The generalization to higher degree polynomial systems

$$p_n(A)x \equiv \left(A^n + \alpha_{n-1}A^{n-1} + \ldots + \alpha_0 I\right) x = b$$

is straight forward and leads to an approximate solution of the form

$$x_m = ||b||_2 V_m p_n(H_{m,m})^{-1} e_1.$$

If f is an analytic function, then we can compute the following approximate solution x_m for the solution of $f(A)x = b$:

$$x_m = ||b||_2 V_m f(H_{m,m})^{-1} e_1. \tag{6}$$

All these approximations are equal to the exact solution if $h_{m+1,m} = 0$. Because $h_{n+1,n} = 0$, we have the exact solution after at most n iterations. The hope is, of course, that the approximate solutions are sufficiently good after $m \ll n$ iterations. There is little to control the residual for the approximate solutions, since in general $f(A)$ may be an expensive function. We use the Krylov subspace reduction in order to avoid expensive evaluation of $f(A)p$ for $p \in \mathbb{R}^n$. A possibility is to compare successive approximations x_m and to base a stopping criterion on this comparison.

4 Computation of the inverse of $f(H_{m,m})$

The obvious way of computing $f(H_{m,m})^{-1}$ is to reduce $H_{m,m}$ first to some convenient canonical form, for instance to diagonal form. If $H_{m,m}$ is symmetric (in that case $H_{m,m}$ is tridiagonal) then it can be orthogonally transformed to diagonal form:

$$Q_m^H H_{m,m} Q_m = D_m,$$

with Q_m an m by m orthogonal matrix and D_m a diagonal matrix. We then have that

$$Q^H f(H_{m,m})^{-1} Q = f(D_m)^{-1},$$

and this can be used for an efficient and stable computation of x_m. If $H_{m,m}$ is neither symmetric nor (close to) normal (that is $H_{m,m}^H H_{m,m} = H_{m,m} H_{m,m}^H$, then the transformation to diagonal form cannot be done by an orthogonal operator. If $H_{m,m}$ has no Jordan blocks, the transformation can be done by

$$X_m^{-1} H_{m,m} X_m = D_m.$$

This decomposition is not advisable if the condition number of X_m is much larger than 1. In that case it is much better to reduce the matrix $H_{m,m}$ to Schur form:

$$Q_m^H H_{m,m} Q_m = U_m,$$

with U_m an upper triangular matrix. The eigenvalues of $H_{m,m}$ appear along the diagonal of U_m. If A is real, then the computations can be kept in real arithmetic if we use the property that $H_{m,m}$ can be orthogonally transformed to generalized Schur form. In a generalized Schur form, the matrix U_m may have two by two non-zero blocks along the diagonal (but its strict lower triangular part is otherwise zero). These two by two blocks represent complex conjugate eigenpairs of $H_{m,m}$. For further details on Schur forms, generalized Schur forms, and their computation, see [9].

5 Numerical examples

Our numerical examples have been taken from [15]. These experiments have been carried out for diagonal real matrices A, which does not mean a loss of generality. In exact arithmetic the Krylov subspace generated with A, and v, coincides with the Krylov subspace generated with $Q^H AQ$ and $Q^H v$, in the sense that

$$K^m(A; v) = QK^m(Q^H AQ, Q^H v),$$

if Q is an orthogonal matrix ($Q^H Q = I$). In other words, transformation with an orthogonal Q leads to another orientation of the orthogonal basis, but the Krylov subspace method leads to the same approximate solutions. Of course, round-off patterns may be different, but round-off does not have dramatic effects on Krylov subspace methods other than an occasional small delay in the number of iterations. Therefore, we may expect similar behavior for more general matrices with the same eigenvalue distribution.

The diagonal matrix A is of order 900. Its eigenvalues are 0.034, 0.082, 0.127, 0.155, 0.190. The remaining 895 eigenvalues are uniformly distributed over the interval $[0.2, 1.2]$. This type of eigenvalue distribution is more or less what one might get with preconditioned Poisson operators. Now suppose that we want to solve $A^2 x = b$, with b a vector with all ones. We list the results for two different approaches:

m	$\|\|b - A^2 x_m^{new}\|\|_2$	$\|\|b - A^2 x_m^{old}\|\|_2$
0	$0.21E2$	$0.21E2$
10	0.18	0.15
20	$0.27E-2$	$0.16E-1$
30	$0.53E-5$	$0.63E-2$
40	$0.16E-8$	$0.28E-2$
50		$0.36E-2$
60		$0.10E-2$
70		$0.49E-4$
80		$0.18e-5$
100		$0.21e-8$

Table 1. Residual norms for approaches A and B

A We generate the Krylov subspace with A and b and determine the approximate solution x_m^{new} as in (5).

B We generate the Krylov subspace for the operator A^2 and the vector b (the 'classical' approach). This leads to approximations denoted as x_m^{old}.

In Table 1 we have listed the norms of the residuals for the two approaches for some values of m. The analysis in [15] shows that the much faster convergence for the new approach could have been expected. Note that the new approach also has the advantage that there are only m matrix vector products with A for the new approach. For the classical approach we need $2m$ matrix vector products with A, assuming that vectors like A^2p are computed by applying A twice. Usually, the matrix vector product is the CPU-dominating factor in the computations, since they operate in $\mathbb{R}^n$. The oprations with $H_{m,m}$ are carried out in $\mathbb{R}^m$, and in typical applications $m \ll n$.

In [15] also an example is given for a more complicated function of A, namely the solution of

$$e^A x = b,$$

with A the same diagonal matrix as in the previous example, and b again the vector with all ones. This is a type of problem that one encounters in the solution of linear systems of ODEs. With the Krylov subspace for A and r_0 of dimension 20, a residual

$$||r_m||_2 \equiv ||b - e^A x_m||_2 \approx 8.7E-12$$

was observed, working in 48 bits floating point precision.

Others have also suggested to work with the reduced system for the computation of, for instance, the exp function of a matrix, as part of solution schemes for (parabolic) systems of equations. See, e.g. [10,8,11].

6 Matrix sign function

The matrix sign function $sign(A)$ for a nonsingular matrix A, with no eigenvalues on the imaginary axis, is defined as follows [3,13]. Let

$$A = X \operatorname{diag}(J_+, J_-)X^{-1}$$

denote the decomposition of $A \in \mathbb{C}^{n\times n}$. The eigenvalues of J_+ lie in the right half plane, and those of J_- are in the left half plane. Let I_+ denote the identity matrix with the same dimensions as J_+, and I_- the identity matrix corresponding to J_-. Then

$$\operatorname{sign}(A) \equiv X \operatorname{diag}(I_+, -I_-)X^{-1}.$$

The sign function can be used, amongst others, to compute invariant subspaces, for instance those corresponding to the eigenvalues of A with positive real parts. It plays also an important role in QCD (cf [12]). The Jordan decomposition of A is not a useful vehicle for the computation of this function. It can be shown that $\operatorname{sign}(A)$ is the limit of the Newton iteration

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-1}), \quad \text{for} \quad k = 0, 1, \ldots, \quad \text{with} \quad A_0 = A.$$

see [3]. Unfortunately, the Newton iteration is also not suitable for accurate computation if A is ill-conditioned. Bai and Demmel consider more accurate ways of computation, which rely on the (block) Schur form of A:

$$B = Q^H AQ = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}.$$

The matrix Q is orthonormal, and it can easily be shown that

$$\operatorname{sign}(A) = Q \operatorname{sign}(B)Q^H.$$

Let this decomposition be such that B_{11} contains the eigenvalues of A with positive real part. Then let R be the solution of the Sylvester equation

$$B_{11}R - RB_{22} = -B_{12}.$$

This Sylvester equation can also be solved by a Newton iteration process. Then Bai and Demmel proved that

$$\operatorname{sign}(A) = Q \begin{bmatrix} I & -2R \\ 0 & -I \end{bmatrix} Q^H.$$

See [3] for further details, stability analysis, and examples of actual computations.

Thomas Lippert [1] has experimented with alternative schemes for the computation of sign(A), in particular the Schultz iteration

$$A_{k+1} = \frac{1}{2} A_k (3I - A_k^2).$$

We do not know of any stability analysis for this approach; our own preliminary experiments suggest a rather slow convergence.

Suppose that we want to solve sign$(A)x = b$. Then, in view of the previous section, we suggest to start with constructing the Krylov subspace $K^m(A; b)$, and to compute the sign function for the reduced matrix $H_{m,m}$. This leads to the following approximation for the solution of sign$(A) = b$:

$$x_m = ||b||_2 V_m \text{sign}(H_{m,m})^{-1} e_1. \tag{7}$$

Our preliminary experiments with this approach are encouraging. The actual problem in QCD, however, is often to solve an essentially different equation

$$(C + \text{sign}(A))x = b.$$

See, for instance, [12]. We have not yet succeeded in identifying efficient computational schemes for this shifted equation. An obvious way to solve this equation is to apply a nested Krylov solution process. In the outer process one forms the Krylov subspace for $C + \text{sign}(A)$ and b, and for each matrix vector product involved in this process one computes approximations for sign$(A)p$, for given vector v, using the process that we have outlined above. Our experiments, not reported here, show that this is a very expensive process. We are currently experimenting with nested Newton iterations for the inverse of $C + \text{sign}(A)$ and the sign function. The idea is to carry out these Newton processes in an inexact way and to accelerate the process, in a way that has been described in [6].

References

1. Arnoldi, W. E.: The principle of minimized iteration in the solution of the matrix eigenproblem. *Quart. Appl. Math.* **9** (1951) 17–29
2. Axelsson, O.: *Iterative Solution Methods.* Cambridge University Press, Cambridge (1994)
3. Bai, Z., Demmel, J.: Using the matrix sign function to compute invariant subspaces. *SIAM J. Matrix Anal. Applic.* **19** (1998) 205–225
4. Barrett, R., Berry, M., Chan, T., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., van der Vorst, H.: *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods.* SIAM, Philadelphia, PA (1994)
5. Bruaset, A.M.: *A Survey of Preconditioned Iterative Methods.* Longman Scientific & and Technical, Harlow, UK (1995)

[1] Private communication

6. Fokkema, D.R., Sleijpen, G.L.G., van der Vorst, H.A.: Accelerated inexact Newton schemes for large systems of nonlinear equations. *SIAM J. Sci. Comput.* **19** (1998) 657–674
7. Freund, R.W., Golub, G.H., Nachtigal, N.M.: Iterative solution of linear systems. In *Acta Numerica 1992*. Cambridge University Press, Cambridge (1992)
8. Gallopoulos, E., Saad, Y.: Efficient solution of parabolic equations by Krylov approximation methods. *SIAM J. Sci. Statist. Comput.* **13** (1992) 1236–1264
9. Golub, G.H., Van Loan, C.F.: *Matrix Computations*. The Johns Hopkins University Press, Baltimore (1996)
10. Hochbruck, M., Lubich, C.: On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.* **34** (1997) 1911–1925
11. Meerbergen, K., Sadkane, M.: Using Krylov approximations to the matrix exponential operator in Davidson's method. *Appl. Numer. Math.* **31** (1999) 331–351
12. Neuberger, H.: The Overlap Dirac Operator. in: Frommer, A., Lippert, Th., Medeke, B., Schilling, K. (edts.). Numerical Challenges in Lattice Quantum Chromodynamics. Proceedings of the Interdisciplinary Workshop on Numerical Challenges in Lattice QCD, Wuppertal, August 22-24, 1999. Series Lecture Notes in Computational Science and Engineering (LNCSE). Springer Verlag, Heidelberg (2000)
13. Roberts, J.: Linear model reduction and solution of the algebraic Riccati equation. *Inter. J. Control* **32** (1980) 677–687
14. Saad, Y.: *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston (1996)
15. van der Vorst, H.A.: An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A. *J. Comp. and Appl. Math.* **18** (1987) 249–263