PROCEEDINGS OF SCIENCE

Krylov subspace methods and the sign function: multishifts and deflation in the non-Hermitian case[*]

Jacques C.R. Bloch[a], Tobias Breu[a], Andreas Frommer[b], Simon Heybrock[a], Katrin Schäfer[†b], Tilo Wettig[a]

[a] *Institute for Theoretical Physics, University of Regensburg, 93040 Regensburg, Germany*
[b] *Department of Mathematics, University of Wuppertal, 42097 Wuppertal, Germany*

E-mail:
jacques.bloch@physik.uni-regensburg.de,
tobias.breu@physik.uni-regensburg.de,
frommer@math.uni-wuppertal.de,
simon.heybrock@physik.uni-regensburg.de,
schaefer@math.uni-wuppertal.de,
tilo.wettig@physik.uni-regensburg.de

Rational approximations of the matrix sign function lead to multishift methods. For non-Hermitian matrices long recurrences can cause storage problems, which can be circumvented with restarts. Together with deflation we obtain efficient iterative methods, as we show in numerical experiments for the overlap Dirac operator at non-vanishing quark chemical potential for lattices up to size 10^4.

The XXVII International Symposium on Lattice Field Theory
July 26-31, 2009
Peking University, Beijing, China

[*] Supported by the DFG collaborative research center SFB/TR-55 “Hadron Physics from Lattice QCD”.

[†] Speaker.

PoS(LAT2009)043

1. Introduction

In this paper we discuss the approximation of $f(A)b$, where $A \in \mathbb{C}^{n\times n}$ is non-Hermitian and f is a function defined on the spectrum of A such that the extension of f to matrix arguments is defined.[1]

The motivation for this rather general setting comes from quantum chromodynamics (QCD) formulated on a discrete space-time lattice, where $f = \text{sign}$ is of special interest. As the main object relevant for our discussion we are focusing on the overlap Dirac operator [3, 4]. The main numerical effort lies in the inversion of the overlap operator, which is done by iterative methods and requires the repeated application of the sign function of the usual "symmetrized" Wilson operator $H_W = \gamma_5 D_W$ (see [5] for the notation) on a vector.

At zero quark chemical potential μ, the operator H_W is Hermitian. However, one can also study QCD at nonzero μ, which is relevant for many physical systems such as neutron stars, relativistic heavy ion collisions, or the physics of the early universe. The overlap operator has been generalized to this case [5, 6]. The computational challenge is the fact that at non-zero chemical potential H_W becomes non-Hermitian.

This contribution is organized as follows. In Section 2 we review multishift methods which have proven to be successful in the Hermitian ($\mu = 0$) case. We will point out the problems that occur when applying these methods to the non-Hermitian ($\mu \neq 0$) case. In Sections 3 and 4 we present two procedures, restarts and deflation, which — especially when applied in combination — make multishift methods applicable to non-Hermitian matrices. We present our numerical results in Section 5, and conclusions are drawn in Section 6.

2. Multishift methods

First we recall some results for the Hermitian case, i.e., we investigate the computation of $f(A)b$, where $A \in \mathbb{C}^{n\times n}$ is Hermitian. If A is large, $f(A)$ is too costly to compute, while $f(A)b$ can still be obtained in an efficient manner if A is sparse. Krylov subspace methods, i.e., methods that approximate $f(A)b$ in a Krylov subspace $K_k(A,b) = \text{span}\{b, Ab, \ldots, A^{k-1}b\}$, are suitable for this task. We distinguish between two Krylov subspace approaches: direct projection and multishift.

Direct projection methods compute the sign function for the projection of A onto $K_k(A,b)$ and lift the result back to the original space, see [1, 7], or [8, 9] in the context of QCD. These methods are not the topic of this paper, but we will use them for comparison in our numerical results.

The idea of multishift methods is to approximate f by a rational function g,

$$f(x) \approx g(x) = \sum_{i=1}^{s} \frac{\omega_i}{x - \sigma_i}\,. \tag{2.1}$$

The systems

$$(A - \sigma_i I)x^{(i)} = b\,, \quad i = 1, \ldots, s \tag{2.2}$$

are treated with standard Krylov subspace methods such as the conjugate gradient method (CG) or the minimal residual method (MINRES), approximating $x^{(i)}$ by ${x_k}^{(i)}$ from a Krylov subspace. Since

[1] The function f can be extended to matrix arguments by, e.g., a spectral definition or a contour integral. For a thorough treatment of matrix functions see [1]; a compact overview is given in [2].

Krylov subspaces are shift invariant, i.e., $K_k(A-\sigma_i I,b) = K_k(A,b)$, the approximations ${x_k}^{(i)}$ can be computed simultaneously using the same subspace for all systems. The desired approximation is then obtained by combining the approximations to the s shifted systems

$$f(A)b \approx x_k = \sum_{i=1}^{s} \omega_i {x_k}^{(i)}\,. \tag{2.3}$$

The core of any such method is the computation of an appropriate basis for the Krylov subspace. For Hermitian matrices an orthonormal basis can be built with short recurrences using the Lanczos process. These short recurrences are essential for the efficiency of the approach.

Turning to non-Hermitian matrices, the computation of an orthogonal basis now requires long recurrences and is usually summarized via the Arnoldi relation

$$AV_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^T\,. \tag{2.4}$$

Here, $V_k = [v_1|\dots|v_k] \in \mathbb{C}^{n\times k}$ is the matrix which contains the computed basis vectors (the Arnoldi vectors), $H_k = {V_k}^{\dagger} A V_k$ is the upper Hessenberg matrix containing the recurrence coefficients $h_{i,j}$, and e_k denotes the k-th unit vector of $\mathbb{C}^k$.

For the rational approximation approach this means that the short-recurrence methods CG and MINRES have to be replaced by multishift versions of the corresponding long-recurrence methods, i.e., the full orthogonalization method (FOM) [10] and the generalized minimal residual method (GMRES) [11], respectively.

Long recurrences slow down computation and increase storage requirements, and thus become inefficient or even infeasible if k, the dimension of the Krylov subspace, becomes large. In this paper we investigate restarts to circumvent this problem for non-Hermitian matrices.

3. Restarts

FOM to solve $Ax = b$ consists of the Arnoldi process to compute the Arnoldi vectors $v_1,\dots,v_k$ as well as the upper Hessenberg matrix $H_k = {V_k}^{\dagger} A V_k$ and of approximating $x \approx x_k = \|b\|_2 V_k {H_k}^{-1} e_1$. The Arnoldi process applied to $A-\sigma_i I$ instead of A produces the same matrices V_k with H_k replaced by the shifted counterpart $H_k - \sigma_i I$. The k-th approximation to $g(A)b$, with $g(x)$ defined in (2.1), is thus given by $\|b\|_2 \sum_{i=1}^{s} V_k (H_k - \sigma_i I)^{-1} e_1$.

To prevent recurrences from becoming too long one can — in this case — use a restart procedure. This means that one stops the Arnoldi process after $k_{\max}$ iterations. At this point we have a, possibly crude, approximation to $g(A)b$, and to allow for a restart one now has to express the error of this approximation anew as the action of a matrix function, $g_1(A)b_1$, say.

A crucial observation concerning multishifts is that for any k the individual residuals ${r_k}^{(i)} = b-(A-\sigma_i I){x_k}^{(i)}$ of the FOM iterates ${x_k}^{(i)}$ are just scalar multiples of the Arnoldi vector v_{k+1}, see, e.g., [10, 12], i.e.,

$${r_k}^{(i)} = {\rho_k}^{(i)} v_{k+1}\,, \quad i = 1,\dots,s \tag{3.1}$$

with collinearity factors ${\rho_k}^{(i)} \in \mathbb{C}$. The error $\Delta_k = g(A)b - x_k$ of the multishift approximation at step k can therefore be expressed as

$$\Delta_k = g_1(A)b_1\,, \quad \text{where } g_1(t) = \sum_{i=1}^{s} \frac{\omega_i {\rho_k}^{(i)}}{t-\sigma_i} \text{ and } b_1 = v_{k+1}\,. \tag{3.2}$$

This allows for a simple restart at step $k_{\max}$ of the Arnoldi process, with the new function g_1 again being rational with the same poles as g. This restart process can also be regarded as performing restarted FOM for each of the individual systems $(A-\sigma_i I)x = b$, $i = 1,\dots,s$ (and combining the individual iterates appropriately), the point being that, even after a restart, we need only a single Krylov subspace for all s systems, see [10].

There also exists a restarted version of multishift GMRES, see [11] for a detailed derivation.

4. Deflation

In [8] two deflation approaches were proposed which use eigensystem information, namely Schur vectors (Schur deflation) or left and right eigenvectors (LR deflation) corresponding to some “critical” eigenvalues. Critical eigenvalues are those which are close to a singularity of f. If they are not reflected very precisely in the Krylov subspace, we get a poor approximation. In case of the sign function the critical eigenvalues are those close to the imaginary axis. Here, we describe LR deflation (see [13] for the reason why this is the method of choice) and show how it can be combined with multishifts and restarts.

Let $R_m = [r_1|\dots|r_m]$ be the matrix containing the right eigenvectors and ${L_m}^\dagger = [l_1|\dots|l_m]^\dagger$ the matrix containing the left eigenvectors corresponding to m critical eigenvalues of the matrix A. This means that we have

$$AR_m = R_m\Lambda_m \quad \text{and} \quad {L_m}^\dagger A = \Lambda_m {L_m}^\dagger\,, \tag{4.1}$$

where Λ_m is a diagonal matrix containing the m critical eigenvalues. Since left and right eigenvectors are biorthogonal, we can normalize them such that ${L_m}^\dagger R_m = I_m$. The matrix $P = R_m {L_m}^\dagger$ represents an oblique projector onto the subspace $\Omega_R = \mathrm{span}\{r_1,\dots,r_m\}$.

We now split $f(A)b$ into the two parts

$$f(A)b = f(A)(Pb) + f(A)(I-P)b\,. \tag{4.2}$$

Since we know the left and right eigenvectors which make up P, we directly obtain

$$x_P \equiv f(A)(Pb) = f(A)R_m L_m^\dagger b = R_m f(\Lambda_m)(L_m^\dagger b)\,, \tag{4.3}$$

which can be computed exactly. The remaining part $f(A)(I-P)b$ can then be approximated iteratively by using a multishift method. Thus $f(A)b$ is now approximated in augmented Krylov subspaces $\Omega_R + K_k(A,(I-P)b)$,

$$x_k = \underbrace{\;x_P\;}_{\in\Omega_R} + \underbrace{\sum_{i=1}^{s}\omega_i x_k^{(i)}}_{\in K_k(A,(I-P)b)}\,. \tag{4.4}$$

Theoretically, we have

$$K_k(A,(I-P)b) = (I-P)K_k(A,(I-P)b) \subseteq \mathrm{range}(I-P)\,, \tag{4.5}$$

see [13]. In computational practice, however, components outside of range$(I-P)$ will show up gradually when building $K_k(A,(I-P)b)$ due to rounding effects in floating-point arithmetic. It is thus necessary to reapply $I-P$ from time to time in order to eliminate these components.

Since the only effect of LR deflation is the replacement of b by $(I-P)b$, no modifications of the restart algorithm are necessary.

5. Algorithms

We combine multishift methods with restarts and deflation. We assume that the original function f is replaced by a rational function (given by the shifts σ_i and weights ω_i) which approximates the original function sufficiently well after deflation.

Depending on the underlying multishift method (FOM or GMRES), we get LR-deflated multishift FOM (FOM-LR) or LR-deflated multishift GMRES (GMRES-LR). Algorithm 1 gives an algorithmic description of FOM-LR. (For an algorithmic description of GMRES-LR we refer to [13].) The notation FOM-LR(m,k) indicates that we LR-deflate a subspace of dimension m and that we restart FOM after a cycle of k iterations. The vector x is the approximation to $f(A)b$. After the completion of each cycle we perform a projection step to eliminate numerical contamination by components outside of range$(I-P)$.

Algorithm 1. Restarted FOM-LR(m,k)

```
{Input $m$, $k = k_{\max}$, $A$, $\{\sigma_1,\ldots,\sigma_s\}$, $\{\omega_1,\ldots,\omega_s\}$, $b$, $L = L_m$, $R = R_m$, $\Lambda = \Lambda_m$}
$x = x_P = Rf(\Lambda)L^{\dagger}b$
$r = (I-P)b$
$\rho^{(i)} = 1$, $i = 1,\ldots,s$
while not all systems are converged do {loop over restart cycles}
    $\beta = \|r\|_2$
    $v_1 = r/\beta$
    compute $V_k$, $H_k$ by running $k$ steps of Arnoldi with $A$
    $y_k^{(i)} = \beta\rho^{(i)}(H_k - \sigma_i I_k)^{-1}e_1$, $i = 1,\ldots,s$
    $x = x + V_k \sum_{i=1}^{s} \omega_i y_k^{(i)}$
    $r = v_{k+1}$
    $\rho^{(i)} = -h_{k+1,k} e_k^T y_k^{(i)}$, $i = 1,\ldots,s$
    $r = (I-P)r$ {projection step}
end while
```

Note that a combination of deflation and a multishift method based on the two-sided Lanczos algorithm is also possible, see [13]. Of course, since two-sided Lanczos already gives short recurrences, there is no need to restart here.

6. Numerical results

For our numerical experiments we turn to $f = \text{sign}$. In the Hermitian case, the sign function of A can be approximated using the Zolotarev best rational approximation, see [14] and, e.g., [15, 16].

Using the Zolotarev approximation on non-Hermitian matrices gives rather poor results, unless all eigenvalues are close to the real axis. A better choice for generic non-Hermitian matrices is the rational approximation originally suggested by Kenney and Laub [17] and used by Neuberger [18, 19] for vanishing chemical potential,

$$\operatorname{sign}(t) \approx g_s(t)\,, \quad \text{where } g_s(t) = \frac{(t+1)^{2s} - (t-1)^{2s}}{(t+1)^{2s} + (t-1)^{2s}}\,. \tag{6.1}$$

The partial fraction expansion of g_s is known to be

$$g_s(t) = t\sum_{i=1}^{s} \frac{\omega_i}{t^2 - \sigma_i} \quad \text{with } \omega_i = \frac{1}{s}\cos^{-2}\left(\frac{\pi}{2s}\left(i-\frac{1}{2}\right)\right), \quad \sigma_i = -\tan^2\left(\frac{\pi}{2s}\left(i-\frac{1}{2}\right)\right), \tag{6.2}$$

see [17, 18]. Note that actually one uses $g(ct)$, where the parameter $c > 0$ is chosen to minimize the number of poles s needed to achieve a given accuracy. If the spectrum of A is known to be contained in the union of two circles $C(m,r) \cup C(-m,r)$, where $C(m,r)$ is the circle $\{|z-m| \le r\}$ and m and r are real with $0 < r < m$, then $c = ((m+r)(m-r))^{-1/2}$ is optimal, see [13, 16].

Figure 1 shows the performance of FOM-LR in comparison to the direct projection method. The k-th approximation in the latter is given as

$$x_P + \|(I-P)b\|_2 V_k \operatorname{sign}(H_k) e_1\,, \tag{6.3}$$

where $\operatorname{sign}(H_k)$ is computed via Roberts' method, see [1]. The relative performance of the two approaches depends on the parameters of the problem, such as the lattice size, the deflation gap, and the size of the Krylov subspace. For more details, see [13]. We add that in the meantime an improved method to compute $\operatorname{sign}(H_k)$ in the direct approach has been developed, see [20] in these proceedings.

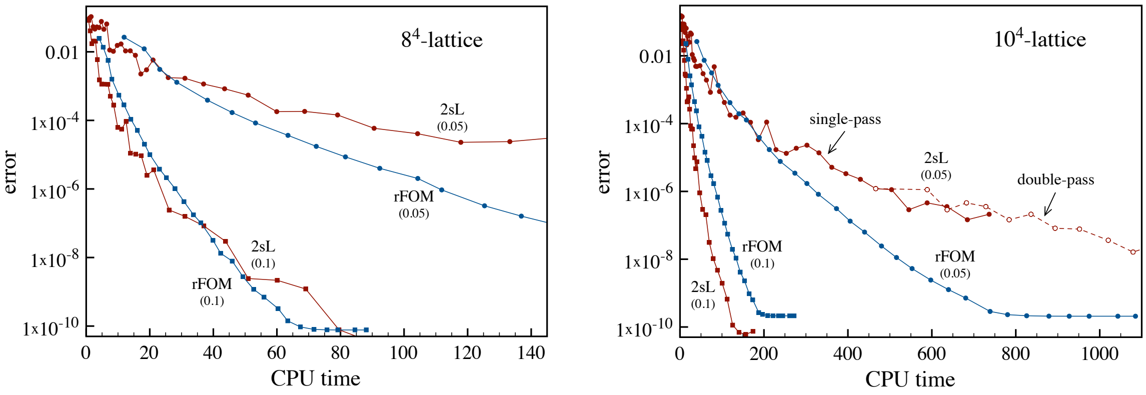

Figure 1: Comparison of the accuracy of the restarted FOM-LR algorithm (rFOM) and the direct two-sided Lanczos-LR method (2sL) as a function of the CPU time in seconds for an 8^4 (left) and a 10^4 (right) lattice configuration, using $\mu = 0.3$ in both cases. Each plot shows data for two different deflation gaps, given in parentheses. The restart size used in the restarted FOM-LR algorithm is $k_{\max} = 30$ for the 8^4 lattice and $k_{\max} = 40$ for the 10^4 lattice.

Figure 2 is meant to convey a warning. It shows that the projection step after each restart, as formulated in Algorithm 1, may be crucial to ensure convergence. In both plots we give results for Algorithm 1 and a variant thereof in which the projection step is omitted. The right plot shows that this may destroy convergence, the left plot shows that this is not necessarily so. Since the CPU time is increased only marginally by the projection step, the latter should always be included.

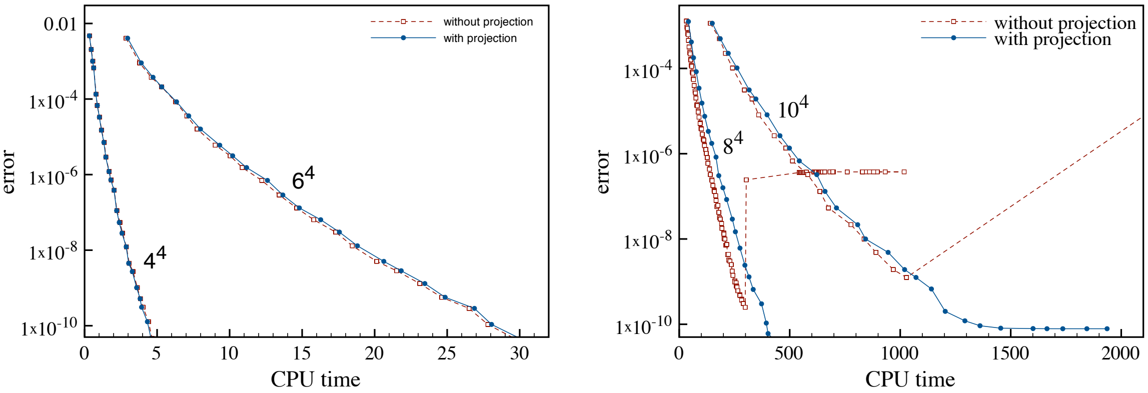

Figure 2: Error vs CPU time for the FOM-LR algorithm with and without re-orthogonalization for 4^4 and 6^4 (left) as well as 8^4 and 10^4 (right) lattices. We again used $\mu = 0.3$ in all cases.

7. Conclusion

We have presented an algorithm, FOM-LR, to approximate the action of the sign function of a non-Hermitian matrix on a vector. This algorithm combines LR deflation and a rational approximation to the sign function, which is computed by a restarted multishift method. The latter has fixed storage requirements determined by the restart parameter (maximum size of the Krylov subspace) and the degree of the rational approximation. Occasionally, additional projections of the Krylov vectors are necessary for numerical stability.

Whether FOM-LR or a direct method (i.e., the two-sided Lanczos-LR method) performs better depends on many details of the problem. Some of them have been mentioned in Section 6. Others include implementation issues such as optimized linear algebra libraries, and ultimately parallelization.

References

[1] N. J. Higham, *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, 2008.

[2] A. Frommer and V. Simoncini, *Matrix Functions*, vol. 13 of *Mathematics in Industry*, ch. 3, pp. 275–303. Springer, Heidelberg, 2008.

[3] R. Narayanan and H. Neuberger, *A construction of lattice chiral gauge theories*, *Nucl. Phys.* **B443** (1995) 305–385, [hep-th/9411108].

[4] H. Neuberger, *Exactly massless quarks on the lattice*, *Phys. Lett.* **B417** (1998) 141–144, [hep-lat/9707022].

[5] J. C. R. Bloch and T. Wettig, *Overlap Dirac operator at nonzero chemical potential and random matrix theory*, *Phys. Rev. Lett.* **97** (2006) 012003, [hep-lat/0604020].

[6] J. C. R. Bloch and T. Wettig, *Domain-wall and overlap fermions at nonzero quark chemical potential*, *Phys. Rev.* **D76** (2007) 114511, [arXiv:0709.4630].

[7] H. van der Vorst, *An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A*, *J. Comput. Appl. Math.* **18** (1987) 249–263.

[8] J. C. R. Bloch, A. Frommer, B. Lang, and T. Wettig, *An iterative method to compute the sign function of a non- Hermitian matrix and its application to the overlap Dirac operator at nonzero chemical potential*, *Comput. Phys. Commun.* **177** (2007) 933–943, [arXiv:0704.3486].

[9] J. C. R. Bloch, T. Breu, and T. Wettig, *Comparing iterative methods to compute the overlap Dirac operator at nonzero chemical potential*, *PoS* **LATTICE2008** (2008) 027, [arXiv:0810.4228].

[10] V. Simoncini, *Restarted full orthogonalization method for shifted linear systems*, *BIT Numerical Mathematics* **43** (2003) 459–466.

[11] A. Frommer and U. Glässner, *Restarted GMRES for shifted linear systems*, *SIAM J. Sci. Comput.* **19** (1998) 15–26.

[12] A. Frommer, *BiCGStab(l) for families of shifted linear systems*, *Computing* **70** (2003) 87–109.

[13] J. C. R. Bloch, T. Breu, A. Frommer, S. Heybrock, K. Schäfer, and T. Wettig, *Short-recurrence Krylov subspace methods for the overlap Dirac operator at nonzero chemical potential*, arXiv:0910.1048.

[14] E. I. Zolotarev, *Application of elliptic functions to the question of functions deviating least and most from zero*, *Zap. Imp. Akad. Nauk. St. Petersburg* **30** (1877) 5.

[15] D. Ingerman, V. Druskin, and L. Knizhnerman, *Optimal finite difference grids and rational approximations of the square root. I. Elliptic problems*, *Comm. Pure Appl. Math.* **53** (2000) 1039–1066.

[16] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. A. van der Vorst, *Numerical methods for the QCD overlap operator. I: Sign-function and error bounds*, *Comput. Phys. Commun.* **146** (2002) 203–224, [hep-lat/0202025].

[17] C. Kenney and A. Laub, *A hyperbolic tangent identity and the geometry of Padé sign function iterations*, *Numer. Algorithms* **7** (1994) 111–128.

[18] H. Neuberger, *A practical implementation of the overlap Dirac operator*, *Phys. Rev. Lett.* **81** (1998) 4060–4062, [hep-lat/9806025].

[19] H. Neuberger, *The overlap Dirac operator*, in *Numerical challenges in Lattice Quantum Chromodynamics* (A. Frommer, T. Lippert, B. Medeke, and K. Schilling, eds.), pp. 1–17, Springer Berlin, 2000. hep-lat/9910040.

[20] J. C. R. Bloch and S. Heybrock, *A nested Krylov subspace method for the overlap operator*, *PoS* **LAT2009** (2009) 025, [arXiv:0910.2918].