



# On restarted Krylov methods for the approximation of matrix functions

Michael Eiermann, Oliver G. Ernst and Stefan Güttel<sup>1</sup>

<sup>1</sup>Institut für Numerische Mathematik und Optimierung  
Technische Universität Bergakademie Freiberg

Rolling Waves in Leuven  
15. December 2008

## Problem

### Given:

- ▶  $A \in \mathbb{C}^{n \times n}$  large and sparse,
- ▶  $\mathbf{b} \in \mathbb{C}^n$ ,  $\|\mathbf{b}\|_2 = 1$ ,
- ▶ a scalar function  $f$ .

**Compute**  $f(A)\mathbf{b}$ .

## Problem

### Given:

- ▶  $A \in \mathbb{C}^{n \times n}$  large and sparse,
- ▶  $\mathbf{b} \in \mathbb{C}^n$ ,  $\|\mathbf{b}\|_2 = 1$ ,
- ▶ a scalar function  $f$ .

Compute  $f(A)\mathbf{b}$ .

- ▶  $f(z) = z^{-1}$  for solving a linear system of equations,
- ▶  $f(z) = \exp(z)$  for the solution of ordinary differential equations,
- ▶  $f(z) = \cos(z)$  for certain hyperbolic PDEs,
- ▶  $f(z) = \sqrt{z}$  Dirichlet-to-Neumann maps,
- ▶  $f(z) = \text{sign}(z)$  in lattice QCD simulations.
- ▶ ...

# Outline

Arnoldi approximation

Restarted Arnoldi approximation

Optimum gradient method

Conjecture

Convergence

Acceleration strategies

# Arnoldi approximation

Extract approximation  $\mathbf{f}_s$  from Krylov space of order  $s$

$$\mathcal{K}_s(A, \mathbf{b}) = \mathcal{K}_s = \text{span}\{\mathbf{b}, A\mathbf{b}, \dots, A^{s-1}\mathbf{b}\}.$$

## Arnoldi method

1. Compute Arnoldi decomposition

$$AV_s = V_s H_s + h_{s+1,s} \mathbf{v}_{s+1} \mathbf{e}_s^T,$$

with  $V_s = [\mathbf{v}_1, \dots, \mathbf{v}_s]$  is ON basis of  $\mathcal{K}_s$  and  $\mathbf{v}_{s+1} \perp \mathcal{K}_s$ .

2. Define Arnoldi approximation

$$\mathbf{f}_s := V_s f(H_s) V_s^* \mathbf{b}.$$

[Druskin & Knizhnerman, 1989], [Gallooulos & Saad, 1992], ...

## Theorem (Saad, 1992)

*There holds*

$$\mathbf{f}_s = V_s f(H_s) V_s^* \mathbf{b} = p_{s-1}(A) \mathbf{b}$$

*where  $p_{s-1}$  is a polynomial of degree  $s-1$  which interpolates  $f$  at  $\Lambda(H_s)$ .*

The Arnoldi method can therefore be seen as an interpolation method where the nodes are the eigenvalues of  $H_s$  (Ritz values).

## Advantages of Arnoldi method

- ▶ avoids  $A$  (only  $v \rightarrow Av$ ),
- ▶ avoids explicit interpolation,
- ▶ requires only evaluation of  $f(H_s)$  for small matrix  $H_s$ ,
- ▶ stable,
- ▶ usually good extraction quality (provable for normal  $A$ ).

## Drawbacks

As  $s$  gets large,

- ▶ size of  $V_s$  exceeds computer memory,
- ▶  $A$  non-symmetric  $\Rightarrow$  orthogonalization cost for  $V_s$  grows.

**Cure:** use restarts!

# Restarted Arnoldi approximation

Consider two Arnoldi decompositions

$$\begin{aligned}AV_1 &= V_1 H_1 + h_1 \mathbf{v}_1 \mathbf{e}_s^T, & V_1(:, 1) &= \mathbf{b} \\AV_2 &= V_2 H_2 + h_2 \mathbf{v}_2 \mathbf{e}_s^T, & V_2(:, 1) &= \mathbf{v}_1.\end{aligned}$$

Glued together they satisfy an Arnoldi-like decomposition

$$A[V_1, V_2] = [V_1, V_2] \begin{bmatrix} H_1 & O \\ E_1 & H_2 \end{bmatrix} + h_2 \mathbf{v}_2 \mathbf{e}_{2s}^T$$

Shorter:

$$A \hat{V}_2 = \hat{V}_2 \hat{H}_2 + h_2 \mathbf{v}_2 \mathbf{e}_{2s}^T.$$

The columns of  $\hat{V}_2$  are a (non-orthogonal) basis of  $\mathcal{K}_{2s}(A, \mathbf{b})$ .

After  $k$  restarts

$$A[V_1, \dots, V_k] = [V_1, \dots, V_k] \begin{bmatrix} H_1 & & & & \\ E_1 & H_2 & & & \\ & \ddots & \ddots & & \\ & & & E_{k-1} & H_k \end{bmatrix} + h_k \mathbf{v}_k \mathbf{e}_{ks}^T$$

Shorter:

$$A\hat{V}_k = \hat{V}_k \hat{H}_k + h_k \mathbf{v}_k \mathbf{e}_{ks}^T.$$

The columns of  $\hat{V}_k$  are a (non-orthogonal) basis of  $\mathcal{K}_{ks}(A, \mathbf{b})$ .

Define Arnoldi-like approximation

$$\hat{\mathbf{f}}_k := \hat{V}_k f(\hat{H}_k) \mathbf{e}_1,$$

then

- ▶ there exists an update formula  $\hat{\mathbf{f}}_{k-1} \longrightarrow \hat{\mathbf{f}}_k$  using only  $V_k$  and  $\hat{H}_k$ ,
- ▶ if  $f$  is a rational function in PFE, only  $V_k$  and  $H_k$  are needed [Afanasjew, Eiermann, Ernst, G., 2008],
- ▶ store only  $s$  long Krylov basis vectors at a time,
- ▶ orthogonalization of less than  $s$  vectors per restart.

## Theorem (Eiermann & Ernst, 2006)

*There holds  $\hat{\mathbf{f}}_k = p_{ks-1}(A)\mathbf{b}$  where  $p_{ks-1}$  is a polynomial of degree  $ks - 1$  which interpolates  $f$  at*

$$\Lambda(\hat{H}_k) = \Lambda(H_1) \cup \Lambda(H_2) \cup \dots \cup \Lambda(H_k).$$

## Problem

How do these interpolation points behave?

- ▶ Since  $H_j = V_j^* A V_j$ , better look at the blocks  $V_j$ !

## Optimum gradient method

- ▶ For  $f(z) = 1/z$  and symmetric positive definite  $A$  the restarted Krylov method coincides with the optimum  $s$ -gradient method.
- ▶ For  $s = 1$  this is also known as *steepest descend method*.

SIAM JOURNAL ON MATHEMATICAL ANALYSIS VOL. 4 (1971) pp. 200-201. Commonly  $x_k$  will ultimately behave as though  $F(x) = |Ax - b|^2$ , where  $A$  and  $b$  are an appropriate matrix and vector. For  $n=3$  the authors prove that  $x_k - x^*$  is asymptotically a linear combination of the eigenvectors belonging to the largest and smallest eigenvalue of  $A^T A$  ( $A^T =$  transpose of  $A$ ). Furthermore  $v_{2k} = \text{grad } F(x_{2k}) / |\text{grad } F(x_{2k})|$  has a limit  $v^*$ , and  $|v_{2k} - v^*| \rightarrow 0$  like some  $q^k$ . (Received January 8, 1951.)

**Figure:** Abstract by G.E. Forsythe and T.S. Motzkin for a talk on »Asymptotic properties of the optimum gradient method« at the 446th meeting of the AMS in New York, 1951.

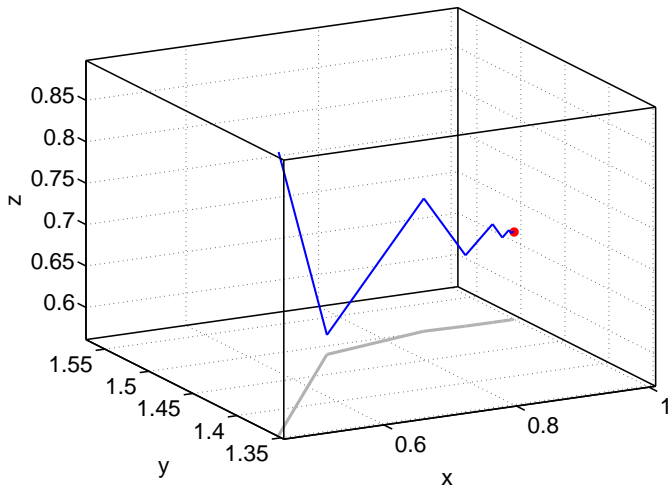


Figure: Iterates of steepest descent for  $A = \text{diag}(1, 2, 3)$  and  $\mathbf{b} = (1, 3, 2)^T$

In 1954, Hirotogu Akaike proved in his paper »*On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method*« for arbitrary  $n$  and  $s = 1$ :

**THEOREM 4;** *In the optimum gradient method with respect to the metric  $\| \cdot \|_P$*

- i)  $\varepsilon_k (=x_k - x)$  *tends to be approximated by a linear combination of two fixed eigenvectors of  $A'PA$  with the eigenvalues equal to  $\text{Max}(\lambda_i; (\varepsilon_0, \xi_i) \neq 0)$  and  $\text{Min}(\lambda_i; (\varepsilon_0, \xi_i) \neq 0)$ , respectively, and*
- ii)  $\varepsilon_k$  *alternates asymptotically in two fixed directions.*

In 1968, Forsythe considered the transformation  $T$  with

$$V_{k+1}(:, 1) = TV_k(:, 1) = p_s(A) V_k(:, 1)$$

and proved

(3.7) **Definition.** By a *continuum* we mean a closed connected set in  $E_n$ , with the understanding that a single point is a continuum.

(3.8) **Theorem.** Fix  $s$  with  $1 \leq s \leq n-1$ . Let  $y_0 = (\eta_1^{(0)}, \dots, \eta_n^{(0)})^T$  be any vector in  $\Sigma^*$  with  $\eta_i^{(0)} \neq 0$  ( $i=1, \dots, n$ ). For  $k=0, 1, \dots$ , define  $y_{k+1} = T y_k$ , where  $T$  was defined in (3.6). Then the set of limit points of the sequence  $\{y_{2k} : k=0, 1, 2, \dots\}$  of normalized gradients is a continuum  $R \subset \Sigma^*$ . Moreover, for any point  $r$  in  $R$ , we have  $r = T^2 r = T(T r)$ .

Forsythe wrote:

The author has programmed a number of test cases with  $s=2$ , to investigate the nature of the set  $R$ . In every case,  $R$  appeared to be a single point. *The author conjectures that  $R$  is always a single point in theorem (3.8).* So far, this has been proved only for  $s=1$ , and we give the proof in (4.12).

If the conjecture is true, then as  $k \rightarrow \infty$

$$V_k, V_{k+1}, V_{k+2}, V_{k+3}, \dots = V_a, V_b, V_a, V_b, \dots$$

Remember  $H_k = V_k^* A V_k$  and therefore

$$H_k, H_{k+1}, H_{k+2}, H_{k+3}, \dots = H_a, H_b, H_a, H_b, \dots$$

In a few slides: The eigenvalues of  $H_k$  are interpolation nodes of our restarted Arnoldi method.

## An extended conjecture

Forsythe only considered angles between the first vectors in the block matrices

$$V_k, V_{k+1}, V_{k+2}, V_{k+3}, \dots$$

(maybe due to memory limitations?)



Nowadays we observe a beautiful symmetry by visualizing the mutual angles between all vectors...

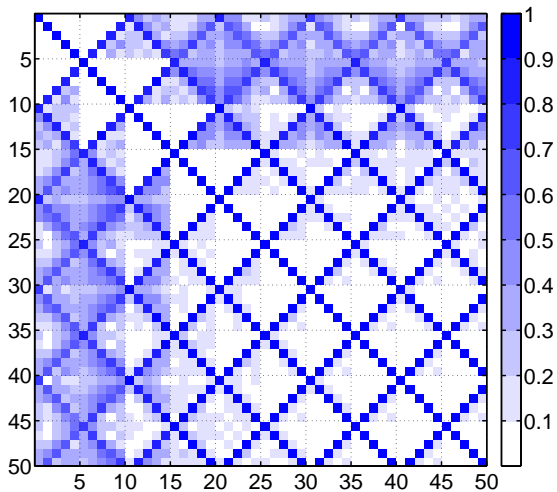


Figure:  $s = 5$ , each block is entrywise  $\sqrt{|\cdot|}$  of  $V_j^* V_k$ .

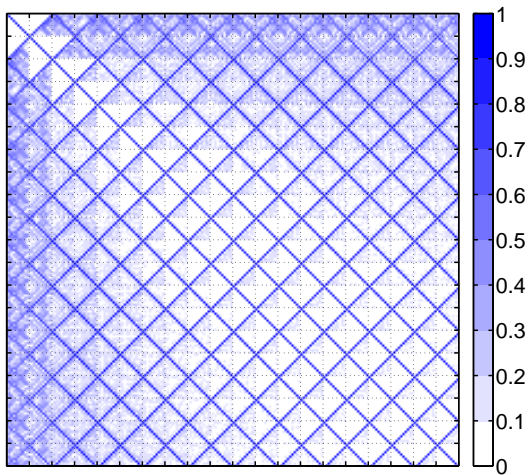


Figure:  $s = 10$ , each block is entrywise  $\sqrt{|\cdot|}$  of  $V_j^* V_k$ .

## Conjecture

Asymptotically, the normalized directions of the optimum  $s$ -gradient method are

$$\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{s-1}, \mathbf{y}_s,$$

$$\mathbf{y}_{s+1}, \mathbf{y}_s, \dots, \mathbf{y}_3, \mathbf{y}_2$$

repeated cyclically.

These vectors span an  $A$ -invariant space of dimension  $s + 1$ .

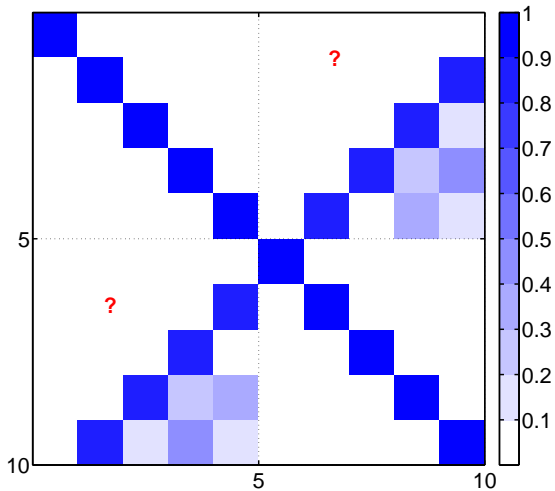


Figure:  $s = 5$ , zoom in previous picture.

## A strong local version of the conjecture

### Lemma

Let

$$\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_s,$$

$$\mathbf{y}_{s+1}, \mathbf{y}_{s+2}, \dots, \mathbf{y}_{2s}$$

be the normalized directions produced by 2 consecutive restarts of the optimum  $s$ -gradient method. Then

$$\mathbf{y}_{s+1} \perp \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_s \quad (\text{by construction})$$

$$\mathbf{y}_{s+2} \perp \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{s-1}$$

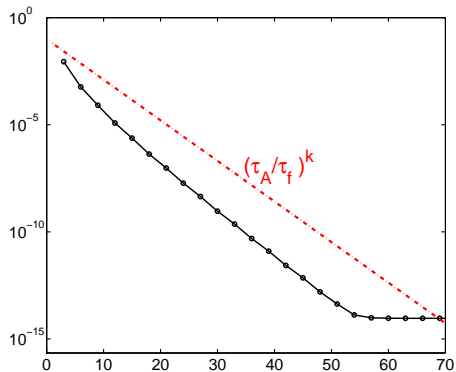
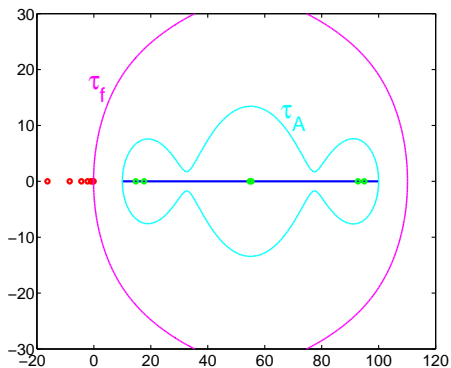
$$\vdots$$

$$\mathbf{y}_{2s} \perp \mathbf{y}_1.$$

# Convergence analysis

Interpolation in cyclically repeated nodes  $\vartheta_1, \dots, \vartheta_{2s}$  is well understood:

- ▶ look at the level lines  $|w(z)| = |(z - \vartheta_1) \cdots (z - \vartheta_{2s})| = \tau^2$ .



# Acceleration strategies

Having computed an Arnoldi decomposition

$$AV_1 = V_1 H_1 + h_1 \mathbf{v}_1 \mathbf{e}_m^T,$$

we can modify it

$$\begin{aligned} AV_1 &= V_1 (H_1 - h_1 \mathbf{x} \mathbf{e}_m^T) + h_1 (\mathbf{v}_1 + V_1 \mathbf{x}) \mathbf{e}_m^T \\ &= V_1 \tilde{H}_1 + h_1 \tilde{\mathbf{v}}_1 \mathbf{e}_m^T. \end{aligned}$$

- ▶ The Arnoldi approximation  $V_1 f(\tilde{H}_1) V_1^* \mathbf{b} = p_{s-1}(A) \mathbf{b}$  where  $p_{s-1}$  interpolates  $f$  at  $\Lambda(\tilde{H}_1)$ .
- ▶  $\Lambda(\tilde{H}_1)$  can be arbitrarily altered with  $\mathbf{x}$ .
- ▶ This can be done after each restart.

Thick restarts ([Sorensen, 1992], [Morgan, 2000]) can be applied:

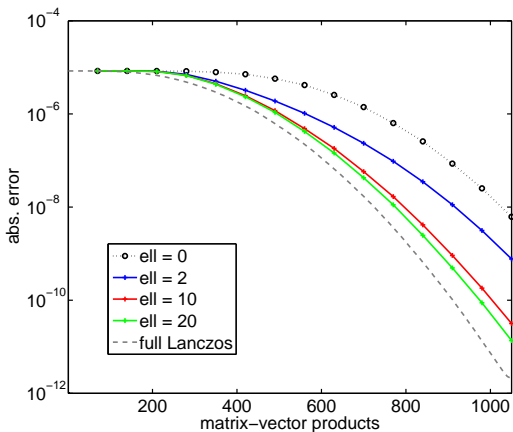


Figure: Solving Maxwell's equation by  $\exp(A)\mathbf{b}$ ,  $A \in \mathbb{C}^{n \times n}$  symmetric,  $n = 565,326$ ,  $s = 70$  [Eiermann, Ernst & G., in prep.].

# Summary

- ▶ Restarted Krylov methods are promising if storage is limited.
- ▶ Thick-restarts can accelerate the restarted methods.
- ▶ (Thick-) restarted Krylov methods are analyzed using Walsh's theory on interpolation polynomials (at least for Hermitian  $A$ ).
- ▶ We have extended the Forsythe conjecture.
- ▶ For  $s = 1$ , the Forsythe-Motzkin conjecture can be proven purely algebraic (see [Afanasjew, Eiermann, Ernst & G., 2008]).

# Summary

- ▶ Restarted Krylov methods are promising if storage is limited.
- ▶ Thick-restarts can accelerate the restarted methods.
- ▶ (Thick-) restarted Krylov methods are analyzed using Walsh's theory on interpolation polynomials (at least for Hermitian  $A$ ).
- ▶ We have extended the Forsythe conjecture.
- ▶ For  $s = 1$ , the Forsythe-Motzkin conjecture can be proven purely algebraic (see [Afanasjew, Eiermann, Ernst & G., 2008]).
  
- ▶ HAPPY BIRTHDAY to Adhemar!