

IDR: A new generation of Krylov subspace methods?

Jens-Peter M. Zemke
zemke@tu-harburg.de

Institut für Numerische Simulation
Technische Universität Hamburg-Harburg

joint work with:
Martin Gutknecht (ETH Zürich),
Olaf Rendel (TU Hamburg-Harburg),
Anisa Rizvanolli (TU Hamburg-Harburg),
Gerard L.G. Sleijpen (Universiteit Utrecht),
Martin B. van Gijzen (TU Delft)

August 23rd, 2011



Krylov subspace methods

Hessenberg decompositions

Polynomial representations

Perturbations

IDR

IDR and $IDR(s)$

IDREIG

$IDR(s)STAB(\ell)$ and IDRSTABEIG

(Flexible and multi-shift) QMRIDR

Perturbations

Introduction

Krylov subspace methods: approximations

$$\left. \begin{array}{l} \mathbf{x}_k, \underline{\mathbf{x}}_k, \\ \mathbf{y}_k, \underline{\mathbf{y}}_k \end{array} \right\} \in \mathcal{K}_k(\mathbf{A}, \mathbf{q}) := \text{span} \{ \mathbf{q}, \mathbf{A}\mathbf{q}, \dots, \mathbf{A}^{k-1}\mathbf{q} \} = \{ p(\mathbf{A})\mathbf{q} \mid p \in \mathbb{P}_{k-1} \},$$

where

$$\mathbb{P}_{k-1} := \left\{ \sum_{j=0}^{k-1} \alpha_j z^j \mid \alpha_j \in \mathbb{C}, 0 \leq j < k \right\},$$

to solutions of linear systems

$$\mathbf{A}\mathbf{x} = \mathbf{r}_0 (= \mathbf{b} - \mathbf{A}\mathbf{x}_0), \quad \mathbf{A} \in \mathbb{C}^{n \times n}, \quad \mathbf{b}, \mathbf{x}_0 \in \mathbb{C}^n,$$

and (partial) eigenproblems

$$\mathbf{A}\mathbf{v} = \mathbf{v}\lambda, \quad \mathbf{A} \in \mathbb{C}^{n \times n}.$$

Hessenberg decompositions

Construction of basis vectors resembled in structure of arising **Hessenberg decomposition**

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\underline{\mathbf{H}}_k,$$

where

- ▶ $\mathbf{Q}_{k+1} = (\mathbf{Q}_k, \mathbf{q}_{k+1}) \in \mathbb{C}^{n \times (k+1)}$ collects basis vectors,
- ▶ $\underline{\mathbf{H}}_k \in \mathbb{C}^{(k+1) \times k}$ is unreduced extended Hessenberg.

Aspects of **perturbed Krylov subspace methods**: captured with **perturbed Hessenberg decompositions**

$$\mathbf{A}\mathbf{Q}_k + \mathbf{F}_k = \mathbf{Q}_{k+1}\underline{\mathbf{H}}_k,$$

$\mathbf{F}_k \in \mathbb{C}^{n \times k}$ accounts for perturbations (finite precision & inexact methods).

Karl Hessenberg & “his” matrix + decomposition



„Behandlung linearer Eigenwertaufgaben mit Hilfe der Hamilton-Cayleyschen Gleichung“, Karl Hessenberg, 1. Bericht der Reihe „Numerische Verfahren“, [July, 23rd 1940](#), page 23:

Man kann nun die Vektoren $\mathfrak{z}_\nu^{(n-1)}$ ($\nu = 1, 2, \dots, n$) ebenfalls in einer Matrix zusammenfassen, und zwar ist nach Gleichung (55) und (56)

$$(57) \quad (\mathfrak{z}_1, \mathfrak{z}_2, \mathfrak{z}_3, \dots, \mathfrak{z}_n^{(n-1)}) = \alpha \cdot \mathfrak{z}' = \mathfrak{z}' \cdot \mathfrak{P},$$

worin die Matrix \mathfrak{P} zur Abkürzung gesetzt ist für

$$(58) \quad \mathfrak{P} = \begin{pmatrix} \alpha_{10} & \alpha_{11} & \dots & \alpha_{1,n-1} & \alpha_{1n} \\ 1 & \alpha_{21} & \dots & \alpha_{2,n-1} & \alpha_{2n} \\ 0 & 1 & \dots & \alpha_{n-1,n-1} & \alpha_{n-1,n} \\ 0 & 0 & \dots & 1 & \alpha_{nn} \end{pmatrix}$$

- ▶ Hessenberg decomposition, Eqn. (57),
- ▶ Hessenberg matrix, Eqn. (58).

Karl Hessenberg (* September 8th, 1904, † February 22nd, 1959)

Important Polynomials

Residuals of **OR** and **MR** approximation

$$\mathbf{x}_k := \mathbf{Q}_k \mathbf{z}_k \quad \text{and} \quad \underline{\mathbf{x}}_k := \mathbf{Q}_k \underline{\mathbf{z}}_k$$

with coefficient vectors

$$\mathbf{z}_k := \mathbf{H}_k^{-1} \mathbf{e}_1 \|\mathbf{r}_0\| \quad \text{and} \quad \underline{\mathbf{z}}_k := \underline{\mathbf{H}}_k^\dagger \mathbf{e}_1 \|\mathbf{r}_0\|$$

satisfy

$$\mathbf{r}_k := \mathbf{r}_0 - \mathbf{A} \mathbf{x}_k = \mathcal{R}_k(\mathbf{A}) \mathbf{r}_0 \quad \text{and} \quad \underline{\mathbf{r}}_k := \mathbf{r}_0 - \mathbf{A} \underline{\mathbf{x}}_k = \underline{\mathcal{R}}_k(\mathbf{A}) \mathbf{r}_0.$$

Residual polynomials $\mathcal{R}_k, \underline{\mathcal{R}}_k$ given by

$$\mathcal{R}_k(z) := \det(\mathbf{I}_k - z \mathbf{H}_k^{-1}) \quad \text{and} \quad \underline{\mathcal{R}}_k(z) := \det(\mathbf{I}_k - z \underline{\mathbf{H}}_k^\dagger \mathbf{I}_k).$$

Convergence of **OR** and **MR** depends on (harmonic) **Ritz values**.

Perturbed OR methods

Setting changes when perturbations enter the stage, here, OR method.

In perturbed case

$$\mathbf{A}\mathbf{Q}_k + \mathbf{F}_k = \mathbf{Q}_{k+1}\mathbf{H}_k$$

polynomial representation

$$\mathbf{r}_k = \mathcal{R}_k(\mathbf{A})\mathbf{r}_0 - \sum_{\ell=1}^k z_{\ell k} \mathcal{R}_{\ell+1:k}(\mathbf{A})\mathbf{f}_{\ell} + \mathbf{F}_k \mathbf{z}_k$$

(all trailing square Hessenberg matrices are assumed to be regular).

Here,

$$\mathcal{R}_{\ell+1:k}(z) := \det(\mathbf{I}_{k-\ell} - z\mathbf{H}_{\ell+1:k}^{-1}).$$

Convergence: $\mathbf{F}_k \mathbf{z}_k$ bounded (inexact methods) & $\mathcal{R}_{\ell+1:k}(\mathbf{A})$ “small”.

IDR: History repeating

IDR

1976 Idea by Sonneveld
 1979 First talk on IDR
 1980 Proceedings
 1989 CGS
 1992 IDR \rightsquigarrow BICGSTAB
 1993 BICGSTAB2, BICGSTAB(ℓ)
 later “acronym explosion” ...

IDR(s)

2006 Sonneveld & van Gijzen
 2007 First presentation & report
 2008 SIAM paper (SISC)
 2008 IDR(s)BIO
 2010 IDR(s)STAB(ℓ), IDREIG
 2011 flexible & multi-shift QMRIDR
 later “acronym explosion”?

- ▶ IDR and IDR based methods are old (\rightsquigarrow my generation),
- ▶ IDR(s) is 5 years “old” (\rightsquigarrow my son’s generation).

IDR is based on Lanczos’s method; IDR(s) is based on Lanczos($s, 1$).

IDR(s) is a Krylov subspace method \rightsquigarrow all techniques from 90’s applicable!

IDR(s)

IDR spaces:

$$\mathcal{G}_0 := \mathcal{K}(\mathbf{A}, \mathbf{q}), \quad (\text{full Krylov subspace})$$

$$\mathcal{G}_j := (\alpha_j \mathbf{A} + \beta_j \mathbf{I})(\mathcal{G}_{j-1} \cap \mathcal{S}), \quad j \geq 1, \quad \alpha_j, \beta_j \in \mathbb{C}, \quad \alpha_j \neq 0,$$

where

$$\text{codim}(\mathcal{S}) = s, \quad \text{e.g.,} \quad \mathcal{S} = \text{span} \{ \tilde{\mathbf{R}}_0 \}^\perp, \quad \tilde{\mathbf{R}}_0 \in \mathbb{C}^{n \times s}.$$

Interpreted as **Sonneveld spaces** (Sleijpen, Sonneveld, van Gijzen 2010):

$$\mathcal{G}_j = \mathcal{S}_j(P_j, \mathbf{A}, \tilde{\mathbf{R}}_0) := \left\{ P_j(\mathbf{A})v \mid v \perp \mathcal{K}_j(\mathbf{A}^H, \tilde{\mathbf{R}}_0) \right\},$$

$$P_j(z) := \prod_{i=1}^j (\alpha_i z + \beta_i).$$

Image of shrinking space: **Induced Dimension Reduction.**

IDR(s)

IDR spaces nested:

$$\{\mathbf{0}\} = \mathcal{G}_{j_{\max}} \subsetneq \cdots \subsetneq \mathcal{G}_{j+1} \subsetneq \mathcal{G}_j \subsetneq \mathcal{G}_{j-1} \subsetneq \cdots \subsetneq \mathcal{G}_2 \subsetneq \mathcal{G}_1 \subsetneq \mathcal{G}_0.$$

How many vectors in $\mathcal{G}_j \setminus \mathcal{G}_{j+1}$? In generic case, $s + 1$.

Stable basis: Partially orthonormalize basis vectors \mathbf{g}_k , $1 \leq k \leq n$:

Arnoldi: compute orthonormal basis of $\mathcal{K}_{s+1} \subset \mathcal{G}_0$,

$$\mathbf{A}\mathbf{G}_s = \mathbf{G}_{s+1}\mathbf{H}_s.$$

“Lanczos”: perform intersection $\mathcal{G}_j \cap \mathcal{S}$, map, and orthonormalize,

$$\mathbf{v}_k = \sum_{i=k-s}^k \mathbf{g}_i \gamma_i, \quad \tilde{\mathbf{R}}_0^H \mathbf{v}_k = \mathbf{0}_s, \quad k \geq s + 1,$$

$$\mathbf{g}_{k+1} \nu_{k+1} = (\alpha_j \mathbf{A} + \beta_j \mathbf{I}) \mathbf{v}_k - \sum_{i=k-j(s+1)-1}^k \mathbf{g}_i \nu_i, \quad j = \left\lfloor \frac{k-1}{s+1} \right\rfloor.$$

Eigenvalues of **Sonneveld pencil** $(\mathbf{H}_k, \mathbf{U}_k)$ are roots of residual polynomials. Those distinct from roots of

$$P_j(z) = \prod_{i=1}^j (\alpha_i z + \beta_i), \quad \text{i.e.,} \quad z_i = -\frac{\beta_i}{\alpha_i}, \quad 1 \leq i \leq j$$

converge to eigenvalues of \mathbf{A} .

Suppose \mathbf{G}_{k+1} of full rank. Sonneveld pencil $(\mathbf{H}_k, \mathbf{U}_k)$ as **oblique projection**:

$$\begin{aligned} \widehat{\mathbf{G}}_k^H(\mathbf{A}, \mathbf{I}_n) \mathbf{G}_k \mathbf{U}_k &= \widehat{\mathbf{G}}_k^H(\mathbf{A} \mathbf{G}_k \mathbf{U}_k, \mathbf{G}_k \mathbf{U}_k) \\ &= \widehat{\mathbf{G}}_k^H(\mathbf{G}_{k+1} \underline{\mathbf{H}}_k, \mathbf{G}_k \mathbf{U}_k) = (\underline{\mathbf{I}}_k^T \underline{\mathbf{H}}_k, \mathbf{U}_k) = (\mathbf{H}_k, \mathbf{U}_k), \end{aligned} \quad (1)$$

here, $\widehat{\mathbf{G}}_k^H := \underline{\mathbf{I}}_k^T \mathbf{G}_{k+1}^\dagger$.

Use **deflated pencil** for Lanczos Ritz values (Gutknecht, Z. (2010): IDREIG).
First: IDR(s)ORES, **Olaf Rendel**: IDR(s)BIO, **Anisa Rizvanolli**: IDR(s)STAB(ℓ).

IDRSTAB

IDR(s)STAB(l) (Tanio & Sugihara; Sleijpen & van Gijzen): combine ideas of IDR(s) and BICGSTAB(l).

IDRSTAB (Sleijpen's implementation) recursively computes "(extended) Hessenberg matrices of basis matrices and residuals" ($k \geq 1$):

$$\begin{array}{cccc}
 \mathbf{G}_{11}^{(k)}, \mathbf{r}_{11}^{(k)} & \mathbf{G}_{12}^{(k)}, \mathbf{r}_{12}^{(k)} & \cdots & \mathbf{G}_{1,\ell+1}^{(k)}, \mathbf{r}_{1,\ell+1}^{(k)} \\
 \mathbf{G}_{21}^{(k)}, \mathbf{r}_{21}^{(k)} & \mathbf{G}_{22}^{(k)}, \mathbf{r}_{22}^{(k)} & \cdots & \mathbf{G}_{2,\ell+1}^{(k)}, \mathbf{r}_{2,\ell+1}^{(k)} \\
 & \mathbf{G}_{32}^{(k)}, \mathbf{r}_{32}^{(k)} & \ddots & \vdots \\
 & & \ddots & \mathbf{G}_{\ell+1,\ell+1}^{(k)}, \mathbf{r}_{\ell+1,\ell+1}^{(k)} \\
 & & & \mathbf{G}_{\ell+2,\ell+1}^{(k)}
 \end{array}
 \quad
 \begin{array}{l}
 \mathbf{G}_{i,j}^{(k)} \in \mathbb{C}^{n \times s}, \quad \mathbf{r}_{i,j}^{(k)} \in \mathbb{C}^n, \\
 \mathbf{G}_{i+1,j}^{(k)} = \mathbf{A}\mathbf{G}_{i,j}^{(k)}, \quad \mathbf{r}_{i+1,j}^{(k)} = \mathbf{A}\mathbf{r}_{i,j}^{(k)}, \\
 \tilde{\mathbf{R}}_0^H \mathbf{G}_{ii}^{(k)} = \mathbf{O}_s, \quad \tilde{\mathbf{R}}_0^H \mathbf{r}_{ii}^{(k)} = \mathbf{o}_s, \\
 (\mathbf{G}_{ii}^{(k)})^H \mathbf{G}_{ii}^{(k)} = \mathbf{I}_s.
 \end{array}$$

Initialization using Arnoldi's method:

$$\begin{aligned}
 \mathbf{G}_{21}^{(1)} &= \mathbf{A}\mathbf{G}_{11}^{(1)} = (\mathbf{G}_{11}^{(1)}, \mathbf{g}_{\text{tmp}}) \underline{\mathbf{H}}_s^{(0)}, \\
 \mathbf{r}_{11}^{(1)} &= \mathbf{r}_0 - \mathbf{G}_{21}^{(1)} \boldsymbol{\alpha}^{(1)} = (\mathbf{I} - \mathbf{G}_{21}^{(1)} (\tilde{\mathbf{R}}_0^H \mathbf{G}_{21}^{(1)})^{-1} \tilde{\mathbf{R}}_0^H) \mathbf{r}_0, \quad \mathbf{r}_{21}^{(1)} = \mathbf{A}\mathbf{r}_{11}^{(1)}.
 \end{aligned}$$

IDRSTAB

Columnwise update (IDR part) such that diagonal blocks

- ▶ form basis of $\mathcal{G}_j \setminus \mathcal{G}_{j+1}$ with expansion $\mathcal{G}_j = \mathbf{A}(\mathcal{G}_{j-1} \cap \mathcal{S}) \rightsquigarrow \boldsymbol{\beta}^{(j)} \in \mathbb{C}^{s \times s}$,
- ▶ are orthonormalized $\rightsquigarrow \underline{\mathbf{H}}_{s-1}^{(j)} \in \mathbb{C}^{s \times s-1}$

All other blocks in column treated in same manner.

Residual updates en détail ($i \leq j$, $\mathbf{r}_{j+1,j}^{(k)} = \mathbf{A}\mathbf{r}_{j,j}^{(k)}$):

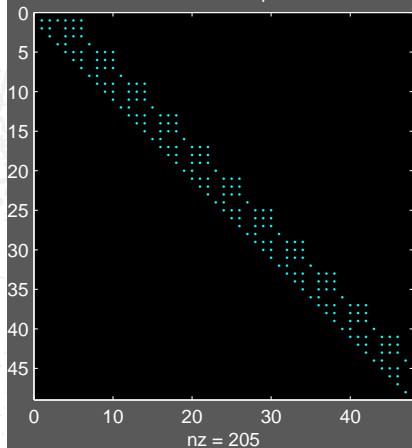
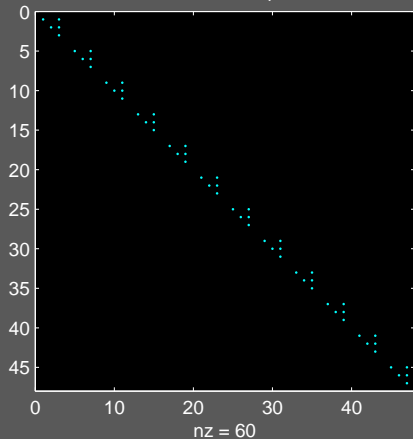
$$\mathbf{r}_{i,j}^{(k)} = \mathbf{r}_{i,j-1}^{(k)} - \mathbf{G}_{i+1,j}^{(k)}\boldsymbol{\alpha}^{(j)}, \quad \mathbf{r}_{j,j}^{(k)} = (\mathbf{I} - \mathbf{G}_{j+1,j}^{(k)}(\tilde{\mathbf{R}}_0^H \mathbf{G}_{j+1,j}^{(k)})^{-1} \tilde{\mathbf{R}}_0^H) \mathbf{r}_{j,j-1}^{(k)}.$$

New cycle (STAB part, $\mathbf{r}_{21}^{(k+1)} = \mathbf{A}\mathbf{r}_{11}^{(k+1)}$, $\gamma^{(\ell)} \in \mathbb{C}^s$ such that $\|\mathbf{r}_{11}^{(k+1)}\| = \min$):

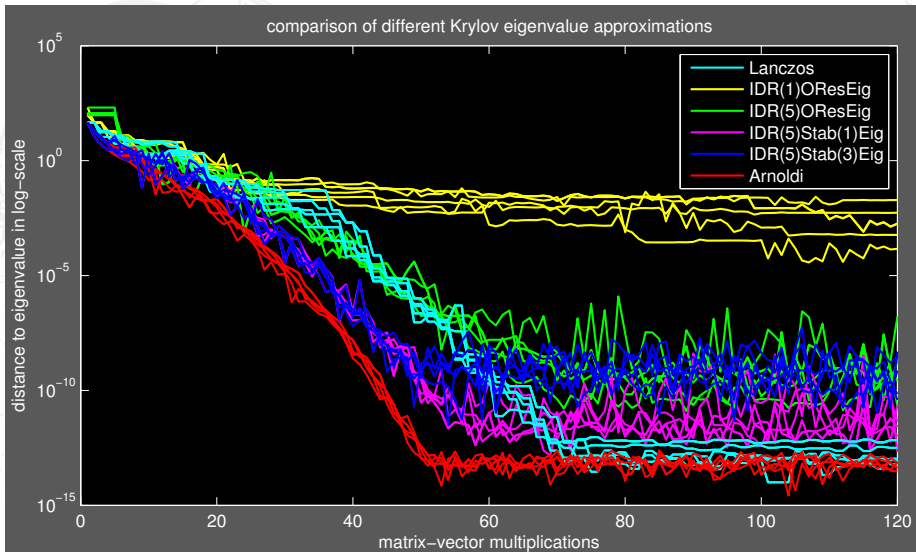
$$\mathbf{r}_{11}^{(k+1)} = \mathbf{r}_{1,\ell+1}^{(k)} - \sum_{i=1}^{\ell} \mathbf{r}_{i+1,\ell+1}^{(k)} \gamma_i^{(\ell)}, \quad \begin{cases} \mathbf{G}_{11}^{(k+1)} = \mathbf{G}_{1,\ell+1}^{(k)} - \sum_{i=1}^{\ell} \mathbf{G}_{i+1,\ell+1}^{(k)} \gamma_i^{(\ell)}, \\ \mathbf{G}_{21}^{(k+1)} = \mathbf{G}_{2,\ell+1}^{(k)} - \sum_{i=1}^{\ell} \mathbf{G}_{i+2,\ell+1}^{(k)} \gamma_i^{(\ell)}. \end{cases}$$

Anisa Rizvanolli: \rightsquigarrow Lanczos-IDRSTAB pencil for eigenvalues, IDRSTABEIG.

Structure of (undeflated) Lanczos-IDRSTAB pencil

Lanczos-IDRStab pencil: uH Lanczos-IDRStab pencil: U 

A comparison: IDR based eigenvalue solvers



QMRIDR

MR methods: use **extended Hessenberg matrix**

$$\underline{\mathbf{x}}_k := \mathbf{Q}_k \underline{\mathbf{z}}_k, \quad \underline{\mathbf{z}}_k := \underline{\mathbf{H}}_k^\dagger \mathbf{e}_1 \|\mathbf{r}_0\|.$$

IDR based: **generalized** Hessenberg decomposition,

$$\mathbf{A} \mathbf{V}_k = \mathbf{A} \mathbf{G}_k \mathbf{U}_k = \mathbf{G}_{k+1} \underline{\mathbf{H}}_k.$$

Thus,

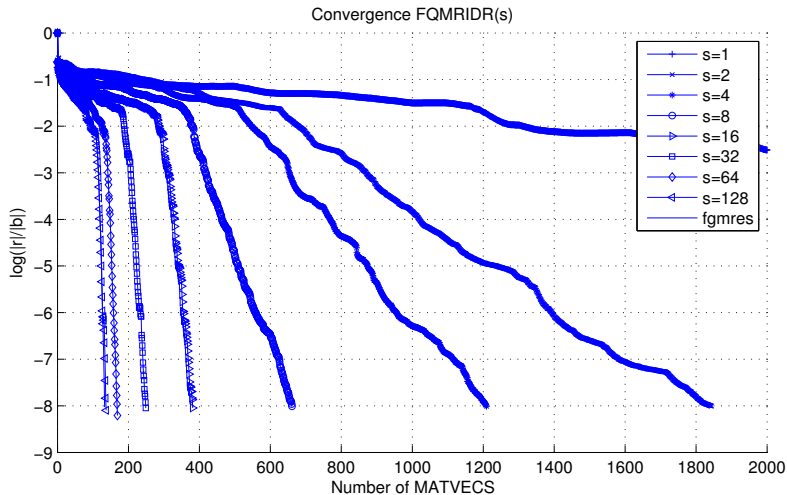
$$\underline{\mathbf{x}}_k := \mathbf{V}_k \underline{\mathbf{z}}_k = \mathbf{G}_k \mathbf{U}_k \underline{\mathbf{z}}_k, \quad \underline{\mathbf{z}}_k := \underline{\mathbf{H}}_k^\dagger \mathbf{e}_1 \|\mathbf{r}_0\|.$$

Other Krylov-paradigms possible, e.g., flexible (& multi-shift) QMRIDR:

$$P_j(\mathbf{A}) \mathbf{v}_k = (\alpha_j \mathbf{A} + \beta_j \mathbf{I}) \mathbf{v}_k \rightsquigarrow (\alpha_j \mathbf{A} \mathbf{P}_j^{-1} + \beta_j \mathbf{I}) \mathbf{v}_k = \mathbf{A} \tilde{\mathbf{v}}_k + \beta_j \mathbf{v}_k,$$

$$\tilde{\mathbf{v}}_k := \mathbf{P}_j^{-1} \mathbf{v}_k \alpha_j, \quad \mathbf{A} \tilde{\mathbf{v}}_k = \mathbf{G}_{k+1} \underline{\mathbf{H}}_k \quad (\text{gen. Hessenberg relation}).$$

Olaf Rendel, Gerard Sleijpen, Martin van Gijzen: \rightsquigarrow **QMRIDRStab**.

Flexible QMRIDR(s)

Perturbations

IDR based on **short recurrences**, i.e., Lanczos based.

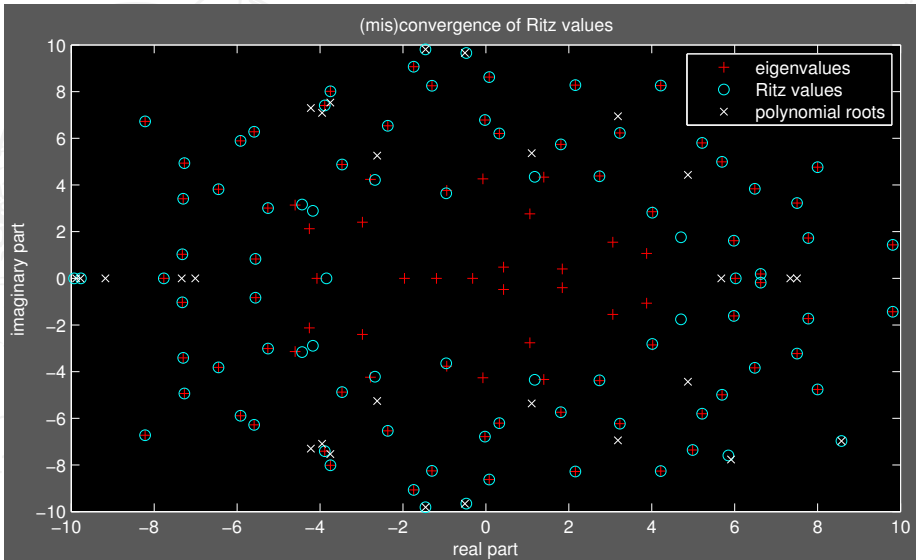
↪ Behaviour in **finite precision**? Inexact methods? General perturbations?

Lanczos	IDR
deviation multiple Ritz values delay of convergence attainable accuracy: condition analysis by Chris Paige	deviation ghost polynomial roots delay of convergence attainable accuracy: worse than Lanczos thus far no error analysis available

But:

- ▶ IDR transpose-free,
- ▶ easy to implement,
- ▶ more stable (for large values of s),
- ▶ often close to “optimal” methods (for large values of s).

IDR(3)STAB(3): “Ghost polynomial roots”



Conclusion and Outlook

- ▶ IDR is both **old** (original IDR, CGS, BICGSTAB, BICGSTAB2, BICGSTAB(ℓ), ...) and **new** (IDR(s), IDRSTAB, QMRIDR, ...).
- ▶ IDR can be included in the framework of Krylov subspace methods using **generalized Hessenberg decompositions**.
- ▶ New developments double old developments at **increased speed**.
- ▶ IDR based methods **bridge the gap** between short- and long-term recurrences.
- ▶ IDR based methods offer **more freedom** in parameters (e.g., the choice of the additional polynomials).

ILAS related:

- ▶ The analysis & development of IDR based methods is a **new branch of Krylov subspace methods**.
- ▶ The pencils of IDR based methods are **specially structured pencils** (adapted backward stable algorithms; perturbation theory, ...).

Thank you for your attention!

In case of questions feel free to ask Anisa, Olaf & myself at any time.

This talk is partially based on the following technical reports:

Eigenvalue computations based on IDR, Martin H. Gutknecht and Z., Bericht 145, Institut für Numerische Simulation, TUHH, 2010,

Flexible and multi-shift induced dimension reduction algorithms for solving large sparse linear systems, Martin B. van Gijzen, Gerard L.G. Sleijpen, and Z., Bericht 156, Institut für Numerische Simulation, TUHH, 2011.