IDR versus other Krylov subspace solvers

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

joint work with Olaf Rendel & Anisa Rizvanolli

Institut für Numerische Simulation
Technische Universität Hamburg-Harburg

October 25th, 13:30-14:00
Krylov subspace methods
  Hessenberg decompositions
  Polynomial representations

IDR
  IDR, IDR\((s)\), and IDRE\(_{IG}\)

IDR vs. other Krylov subspace methods
  IDR\(_{TAB}\) and QMRIDR
  Transferring techniques
  Stay close to Arnoldi/Lanczos
Krylov subpace methods: approximations

\[ x_k, x_k', y_k, y_k' \} \in \mathcal{K}_k(A, q) := \text{span} \{ q, Aq, \ldots, A^{k-1}q \} = \{ p(A)q \mid p \in \mathbb{P}_{k-1} \}, \]

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

\[ Ax = r_0 \ (= b - Ax_0), \quad A \in \mathbb{C}^{n \times n}, \quad b, x_0 \in \mathbb{C}^n, \]

and (partial) eigenproblems

\[ Av = v\lambda, \quad A \in \mathbb{C}^{n \times n}. \]
Hessenberg decompositions

Construction of basis vectors resembled in structure of arising Hessenberg decomposition

\[ AQ_k = Q_{k+1} H_k, \]

where

- \( Q_{k+1} = (Q_k, q_{k+1}) \in \mathbb{C}^{n \times (k+1)} \) collects basis vectors,
- \( H_k \in \mathbb{C}^{(k+1) \times k} \) is unreduced extended Hessenberg.

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

\[ AQ_k + F_k = Q_{k+1} H_k, \]

\( F_k \in \mathbb{C}^{n \times k} \) accounts for perturbations (finite precision & inexact methods).
Karl Hessenberg & “his” matrix + decomposition


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

Karl Hessenberg (∗ September 8th, 1904, † February 22nd, 1959)
Important Polynomials

Residuals of OR and MR approximation

\[ x_k := Q_k z_k \quad \text{and} \quad x_k := Q_k z_k \]

with coefficient vectors

\[ z_k := H_k^{-1} e_1 \| r_0 \| \quad \text{and} \quad z_k := H_k^\dagger e_1 \| r_0 \| \]

satisfy

\[ r_k := r_0 - A x_k = R_k(A) r_0 \quad \text{and} \quad r_k := r_0 - A x_k = R_k(A) r_0. \]

Residual polynomials \( R_k, \overline{R}_k \) given by

\[ R_k(z) := \det \left( I_k - z H_k^{-1} \right) \quad \text{and} \quad \overline{R}_k(z) := \det \left( I_k - z H_k^\dagger I_k \right). \]

Convergence of OR and MR depends on (harmonic) Ritz values.
IDR: History repeating

<table>
<thead>
<tr>
<th>Year</th>
<th>IDR</th>
<th>IDR(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>Idea by Sonneveld</td>
<td>2006</td>
</tr>
<tr>
<td>1979</td>
<td>First talk on IDR</td>
<td>2007</td>
</tr>
<tr>
<td>1980</td>
<td>Proceedings</td>
<td>2008</td>
</tr>
<tr>
<td>1989</td>
<td>CGS</td>
<td>2008</td>
</tr>
<tr>
<td>1992</td>
<td>IDR ⟷ BICGSTAB</td>
<td>2010</td>
</tr>
<tr>
<td>1993</td>
<td>BICGSTAB2, BICGSTAB(ℓ)</td>
<td>2011</td>
</tr>
</tbody>
</table>

later “acronym explosion” . . .

▶ IDR and IDR based methods are old (↝ my generation),
▶ IDR(s) is 5 years “old” (↝ 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 ⟷ all techniques from 90’s applicable!
IDR\((s)\)

IDR spaces:

\[ G_0 := \mathcal{K}(A, q), \quad \text{(full Krylov subspace)} \]
\[ G_j := (\alpha_j A + \beta_j I)(G_{j-1} \cap S), \quad j \geq 1, \quad \alpha_j, \beta_j \in \mathbb{C}, \quad \alpha_j \neq 0, \]

where

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

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

\[ G_j = S_j(P_j, A, \tilde{R}_0) := \left\{ P_j(A)v \mid v \perp \mathcal{K}_j(A^H, \tilde{R}_0) \right\}, \]

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

Image of shrinking space: Induced Dimension Reduction.
IDR(\(s\))

IDR spaces nested:

\[
\{ \mathbf{0} \} = \mathcal{G}_{j_{\text{max}}} \subset \cdots \subset \mathcal{G}_{j+1} \subset \mathcal{G}_j \subset \mathcal{G}_{j-1} \subset \cdots \subset \mathcal{G}_2 \subset \mathcal{G}_1 \subset \mathcal{G}_0.
\]

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 \(\mathbf{G}_{s+1}\) of \(\mathcal{K}_{s+1} \subset \mathcal{G}_0\),

\[
\mathbf{A}\mathbf{V}_s = \mathbf{A}\mathbf{G}_s = \mathbf{G}_{s+1}\mathbf{H}_s, \quad \mathbf{V}_s := \mathbf{G}_s.
\]

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

\[
\mathbf{v}_k = \sum_{i=k-s}^{k} \mathbf{g}_i \gamma_i, \quad \tilde{\mathbf{R}}^H_0 \mathbf{v}_k = \mathbf{0}_s, \quad k \geq s + 1,
\]

\[
\mathbf{g}_{k+1} \nu_{k+1} = (\alpha_j \mathbf{A} + \beta_j \mathbf{I}) \mathbf{v}_k - \sum_{i=k-j(s+1)-1}^{k} \mathbf{g}_i \nu_i, \quad j = \left[ \frac{k-1}{s+1} \right].
\]
Generalized Hessenberg decomposition:

\[ AV_k = AG_k U_k = G_{k+1} H_k, \]

where \( U_k \in \mathbb{C}^{k \times k} \) upper triangular.

Structure of Sonneveld pencils:

\[ H_k = \begin{pmatrix}
\text{\begin{array}{cccccccc}
\times & \times & \times & \times & o & o & o & o \\
+ & \times & \times & \times & o & o & o & o \\
o & + & \times & \times & x & o & o & o \\
o & o & + & \times & x & x & o & o \\
o & o & o & o & + & x & x & x \\
o & o & o & o & o & + & x & x x \\
o & o & o & o & o & o & + & x x x \\
o & o & o & o & o & o & o & + x x x \\
o & o & o & o & o & o & o & o \\
\end{array}\end{pmatrix}}
\]
Eigenvalues of Sonneveld pencil \((H_k, U_k)\) 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 \(A\).

Suppose \(G_{k+1}\) of full rank. Sonneveld pencil \((H_k, U_k)\) as oblique projection:

\[
\hat{G}_k^H (A, I_n) G_k U_k = \hat{G}_k^H (AG_k U_k, G_k U_k) = \hat{G}_k^H (G_{k+1} H_k, G_k U_k) = (I_k^T H_k, U_k) = (H_k, U_k), \tag{1}
\]

here, \(\hat{G}_k^H := I_k^T G_k^\dagger_{k+1}\).

Use deflated pencil for Lanczos Ritz values (Gutknecht, Z. (2010): IDRE\text{IG})

First: IDR(\(s\))ORES, Olaf Rendel: IDR(\(s\))BIO, Anisa Rizvanolli: IDR(\(s\))STAB(\(\ell\)).
IDRStab (Sleijpen’s implementation) recursively computes “(extended) Hessenberg matrices of basis matrices and residuals” \((k \geq 1)\):

\[
\begin{align*}
G^{(k)}_{11}, r^{(k)}_{11} & \quad G^{(k)}_{12}, r^{(k)}_{12} & \quad \ldots & \quad G^{(k)}_{1,\ell+1}, r^{(k)}_{1,\ell+1} \\
G^{(k)}_{21}, r^{(k)}_{21} & \quad G^{(k)}_{22}, r^{(k)}_{22} & \quad \ldots & \quad G^{(k)}_{2,\ell+1}, r^{(k)}_{2,\ell+1} \\
G^{(k)}_{32}, r^{(k)}_{32} & \quad \ldots & \quad \vdots & \quad \vdots \quad \vdots \\
& \quad G^{(k)}_{\ell+1,\ell+1}, r^{(k)}_{\ell+1,\ell+1} \\
G^{(k)}_{\ell+2,\ell+1} & \quad & \quad & \\
& \quad & & \\
\end{align*}
\]

Initialization using Arnoldi’s method:

\[
\begin{align*}
G^{(1)}_{21} & = AG^{(1)}_{11} = (G^{(1)}_{11}, g_{\text{tmp}})H^{(0)}_s, \\
r^{(1)}_{11} & = r_0 - G^{(1)}_{21} \alpha^{(1)} = (I - G^{(1)}_{21} (\tilde{R}^H_0 G^{(1)}_{21})^{-1} \tilde{R}^H_0) r_0, \quad r^{(1)}_{21} = Ar^{(1)}_{11}.
\end{align*}
\]
Columnwise update (IDR part) such that diagonal blocks

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

In particular, with $\tilde{v}_i \in \mathcal{G}_{j-1} \cap S$,

$$\beta^{(j)}_i = (\tilde{R}_0^H \mathcal{G}_{j-1})^{-1} \tilde{R}_0^H (A \tilde{v}_i)$$

$$\Rightarrow (A \tilde{v}_i) - G_{j,j-1} \beta^{(j)}_i = A(\tilde{v}_i - G_{j-1,j-1} \beta^{(j)}_i) \in \mathcal{G}_j \cap S$$

Every new vector in $\mathcal{G}_j \cap S$ is orthonormalized with respect to the others.

Thus, for the IDR-IDRStab pencil relating (STAB-purified) diagonal blocks,

- $\beta^{(j)} \in \mathbb{C}^{s \times s}$ couples $G_{jj}$ and $G_{j,j-1} = AG_{j-1,j-1} \sim U_k,$
- $H^{(j)}_{s-1} \in \mathbb{C}^{s \times (s-1)}$ couples result with others in same block $\sim H_k.$

All other blocks in column treated in same manner.
Residual updates en détail (i ≤ j, r_{j+1, j}^{(k)} = A r_{j, j}^{(k)}):

\[ r_{i, j}^{(k)} = r_{i, j-1}^{(k)} - G_{i+1, j}^{(k)} \alpha^{(j)}, \quad r_{j, j}^{(k)} = (I - G_{j+1, j}^{(k)} (\tilde{R}_0^H G_{j+1, j}^{(k)})^{-1} \tilde{R}_0^H) r_{j, j-1}^{(k)}. \]

Here,

\[ \alpha^{(j)} := (\tilde{R}_0^H G_{j+1, j}^{(k)})^{-1} \tilde{R}_0^H r_{j, j-1}^{(k)}, \]

\( \alpha^{(j)} \) relating \( r_{j, j-1}^{(k)} = A r_{j-1, j-1}^{(k)} \) (old) and \( r_{j, j}^{(k)} \) (new) via \( G_{j+1, j}^{(k)} = A G_{j, j}^{(k)} \rightleftharpoons U_k. \)

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

\[ r_{11}^{(k+1)} = r_{1, \ell+1}^{(k)} - \sum_{i=1}^{\ell} r_{i+1, \ell+1}^{(k)} \gamma_i^{(\ell)}, \quad \left\{ \begin{array}{l} G_{11}^{(k+1)} = G_{1, \ell+1}^{(k)} - \sum_{i=1}^{\ell} G_{i+1, \ell+1}^{(k)} \gamma_i^{(\ell)} , \\ G_{21}^{(k+1)} = G_{2, \ell+1}^{(k)} - \sum_{i=1}^{\ell} G_{i+2, \ell+1}^{(k)} \gamma_i^{(\ell)}. \end{array} \right. \]

Anisa Rizvanolli: \( \rightleftharpoons \) Lanczos-IDRSTAB pencil for eigenvalues, IDRSTABEig.
Structure of (STAB-purified) IDR-IDRSTAB pencil

IDR-IDRStab pencil: uH

IDR-IDRStab pencil: U

nz = 84

nz = 181
Structure of (undeflated) Lanczos-IDR\textsuperscript{stab} pencil
MR methods: use **extended** Hessenberg matrix

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

IDR based: **generalized** Hessenberg decomposition,

\[ \mathbf{AV}_k = \mathbf{AG}_k \mathbf{U}_k = \mathbf{G}_{k+1} \mathbf{H}_k. \]

Thus,

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

Simplified residual bound (block-wise orthonormalization):

\[ \| \mathbf{r}_k \| = \| \mathbf{r}_0 - \mathbf{A} \mathbf{x}_k \| \leq \| \mathbf{G}_{k+1} \| \cdot \| \mathbf{e}_1 \| \| \mathbf{r}_0 \| - \mathbf{H}_k \mathbf{z}_k \| \]

\[ \leq \sqrt{\frac{k + 1}{s + 1}} \cdot \| \mathbf{e}_1 \| \| \mathbf{r}_0 \| - \mathbf{H}_k \mathbf{z}_k \|. \]

Implementation based on short recurrences possible.
QMRIDR

Other Krylov-paradigms possible, e.g., flexible QMRIDR:

\[
P_j(A)v_k = (\alpha_jA + \beta_jI)v_k \leadsto (\alpha_jAP_k^{-1} + \beta_jI)v_k = A\tilde{v}_k + \beta_jv_k,
\]

\[
\tilde{v}_k := P_j^{-1}v_k\alpha_j, \quad A\tilde{V}_k = G_{k+1}H_k.
\]

Generalized Hessenberg relation, generically no longer generalized Hessenberg decomposition, as generically

\[
A\tilde{V}_k \neq AG_k\tilde{U}_k
\]

for every (upper triangular) \(\tilde{U}_k\).

Computation of flexible MR iterate and flexible MR approximation:

\[
z_k := H_k^+e_1\|r_0\|, \quad x_k := \tilde{V}_kz_k.
\]

Flexible IDR variants algorithmically very easy to implement.
Multi-shift is a technique developed for shifted systems

\[(A - \sigma I)x^{(\sigma)} = r_0, \quad \sigma \in \mathbb{C}.\]

We look for quasi-optimal approximations of the form

\[x^{(\sigma)} \approx x_k^{(\sigma)} := V_k z_k^{(\sigma)}.\]

Since \(AV_k = AG_k U_k = G_{k+1} H_k\), and since we use \(G_{k+1} e_1 \|r_0\| = r_0\),

\[r_k^{(\sigma)} = r_0 - (A - \sigma I)x_k^{(\sigma)} = G_{k+1} \left( e_1 \|r_0\| - (H_k - \sigma U_k)z_k^{(\sigma)} \right).\]

Thus, \(z_k^{(\sigma)}\) quasi-optimal:

\[z_k^{(\sigma)} := (H_k - \sigma U_k)^{\dagger} e_1 \|r_0\|.\]

Various extensions for IDR\textsuperscript{stab}: Olaf Rendel, Z. ~→ QMRIDR\textsuperscript{STAB}. 
In (Sonneveld, 2010) a relation between IDR and GMRES for the case of random shadow vectors was pointed out.

Neglecting the influence of the STAB-part, i.e., focusing on Lanczos\((s, 1)\), the deviation of IDR from GMRES is described using stochastic arguments.

As a rule of thumb:

\[
\text{As } s \text{ tends to infinity, the convergence curves of } \text{Lanczos}(s, 1) \text{ tend to the convergence curve of full GMRES.}
\]

In practice, the first steps of IDR/QMRIDR and Arnoldi/GMRES coincide, as we ideally start IDR with these methods.

We present some examples that depict the relations in (Sonneveld, 2010), show additionally the effects of finite precision, and relate GMRES to QMR\((s, 1)\) and to QMRIDR\((s)\).

We remark that the prototype IDR algorithm suffered from instability for large values of \(s\). We only consider new, stable implementations.
“Exact” Lanczos \((s, 1)\) versus full GMRes

- Stay close to Arnoldi/Lanczos
- "Exact" Lanczos \((s, 1)\) versus full GMRes
- Matrix-vector-multiplications: residuals in log-scale
- Full GMRes vs. Lanczos \((s, 1)\), \(s=1, \ldots, 40\), full reorthogonalization
“Finite precision” Lanczos\((s, 1)\) versus full GMRes

Lanczos(s, 1) vs. full GMRes, s=1,...,40, finite precision

- full GMRes
- Lanczos(s, 1), s=1..40

residuals in log-scale

matrix–vector–multiplications
"Exact" QMR\(^{(s, 1)}\) versus full GMRes

QMR\((s, 1)\) vs. full GMRes, \(s=1,\ldots,40\), full reorthogonalization

residuals in log-scale

matrix-vector-multiplications

full GMRes
QMR\((s, 1)\), \(s=1..40\)

TUHH
Jens-Peter M. Zemke
IDR @ Kyoto 2011
2011-10-25
26 / 32
"Finite precision" QMR$(s, 1)$ versus full GMRES

QMR$(s, 1)$ vs. full GMRes, $s=1,...,40$, finite precision

residuals in log-scale

matrix-vector-multiplications

full GMRes
QMR$(s, 1)$, $s=1..40$
Finite precision $\text{QMRIDR}(s)$ versus full GMRes

![Graph showing residuals in log-scale for QMRIDR(s) and full GMRes](image)

- Full GMRes
- QMRIDR(s), $s=1..40$

Matrix-vector-multiplications vs. residuals in log-scale

Jens-Peter M. Zemke

IDR @ Kyoto 2011
A comparison: IDR based eigenvalue solvers

comparison of different Krylov eigenvalue approximations

- Lanczos
- IDR(1)OResEig
- IDR(5)OResEig
- IDR(5)Stab(1)Eig
- IDR(5)Stab(3)Eig
- Arnoldi

distance to eigenvalue in log-scale

matrix-vector multiplications
Flexible QMRIDR$_s$
The new implementations of IDR, i.e., IDR$^{\text{STAB}}$, QMRIDR, its combinations, and the eigensolver counterparts, are very promising.

The new IDR implementations provide a “smooth” transition between Arnoldi/GMRES ($s \to \infty$) and Lanczos/QMR ($s \to 1$).

The matrix generalization of Hessenberg decompositions to generalized Hessenberg decompositions and generalized Hessenberg relations allows for a simple application of standard Krylov subspace techniques.

The dependence on the parameters ($s$, $\tilde{R}_0$, the STAB-part, ...) has to be analyzed carefully.

An error analysis and a description of the finite precision behavior is desperately needed.
どうもありがとうございました。

Thank you very much for inviting me to 京都大学.

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.

An extended abstract can be found in the proceedings:

IDR versus other Krylov subspace solvers, Z., 2011.
On the convergence behaviour of IDR(s).
Technical Report 10-08, Department of Applied Mathematical Analysis, Delft University of Technology, Delft.