Other topics in Krylov subspace methods

Professor Tim Moroney

School of Mathematical Sciences, QUT

Outline

- A common framework
- The Hermitian case
- Restarting with deflation
- Matrix functions
- Jacobian-free Newton-Krylov
- Other bits and pieces

A common framework

A matrix problem in n dimensions: e.g. eigenvalue problem, linear system.

A subspace $\mathcal{K}_m$ ("trial space") to project the problem onto. Our focus is on Krylov subspaces. There are m degrees of freedom.

A second subspace $\mathcal{L}_m$ ("test space") with which to impose m suitable constraints: typically that the residual is orthogonal to $\mathcal{L}_m$.

Different choices of $\mathcal{K}_m$ and $\mathcal{L}_m$ give rise to different methods.

Let us put our methods in this framework and see where they fit.

Arnoldi method for eigenvalues (Ritz)

Choose $\mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2b, \ldots, A^{m-1}b\}$ and $\mathcal{L}_m = \mathcal{K}_m$.

We found Ritz values $\theta_i^{(m)}$ and Ritz vectors $u_i^{(m)} = V_m y_i^{(m)} \in \mathcal{K}_m$.

The eigenvalue problem residual is

$$\begin{aligned} r_i^{(m)} &= Au_i^{(m)} - \theta_i^{(m)} u_i^{(m)} \\ &= AV_m y_i^{(m)} - \theta_i^{(m)} V_m y_i^{(m)} \end{aligned}$$

And the constraints imposed are $r^{(m)} \perp \mathcal{K}_m$: i.e. $V_m^* r_i^{(m)} = 0$.

$$\begin{aligned} V_m^*(AV_m y_i^{(m)} - \theta_i^{(m)} V_m y_i^{(m)}) &= 0 \\ V_m^* AV_m y_i^{(m)} - \theta_i^{(m)} V_m^* V_m y_i^{(m)} &= 0 \\ H_m y_i^{(m)} - \theta_i^{(m)} y_i^{(m)} &= 0 \end{aligned}$$

So Ritz values come from solving $\boxed{H_m\, y_i^{(m)} = \theta_i^{(m)} y_i^{(m)}}$

GMRES

Choose $\mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2b, ..., A^{m-1}b\}$ and $\mathcal{L}_m = A\mathcal{K}_m$.

The residual is

$$r^{(m)} = b - Ax^{(m)} = b - AV_m y^{(m)}$$

And the constraints imposed are $r^{(m)} \perp A\mathcal{K}_m$: i.e. $(AV_m)^* r^{(m)} = 0$.

$$\begin{aligned}
(AV_m)^*(b - AV_m y^{(m)}) &= 0 \\
(V_{m+1}\overline{H}_m)^*(\beta V_{m+1} e_1 - V_{m+1}\overline{H}_m y^{(m)}) &= 0 \\
\overline{H}_m^* V_{m+1}^*(\beta V_{m+1} e_1 - V_{m+1}\overline{H}_m y^{(m)}) &= 0 \\
\overline{H}_m^*\overline{H}_m y^{(m)} &= \overline{H}_m^*(\beta e_1)
\end{aligned}$$

which we recognise as the normal equations for the least squares problem

$$\min \|\beta e_1 - \overline{H}_m y\|$$

as used in GMRES.

Methods so far

All methods are built on the Krylov subspace

$$\mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2b, ..., A^{m-1}b\}.$$

The choice of $\mathcal{L}$ and the form of the residual determines the method.

	$\mathcal{L} = \mathcal{K}$	$\mathcal{L} = A\mathcal{K}$
$Au = \lambda u$	Arnoldi (Ritz)	?
$Ax = b$	?	GMRES

Can we fill in the rest of this table?

- Projection method for eigenvalues using $\mathcal{L} = A\mathcal{K}$?
- Projection method for linear systems using $\mathcal{L} = \mathcal{K}$?

Let's tackle the second one first.

FOM

Choose $\mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2b, ..., A^{m-1}b\}$ and $\mathcal{L}_m = \mathcal{K}_m$.

The residual is

$$r^{(m)} = b - Ax^{(m)} = b - AV_m y^{(m)}$$

And the constraints imposed are $r^{(m)} \perp A\mathcal{K}_m$: i.e. $V_m^* r^{(m)} = 0$.

$$V_m^*(b - AV_m y^{(m)}) = 0$$

$$V_m^* b - V_m^* AV_m y^{(m)} = 0$$

$$\beta e_1 - H_m y^{(m)} = 0$$

which is simply the (square) linear system

$$H_m y^{(m)} = \beta e_1$$

This is the Full Orthogonalisation Method (FOM).

Nowhere near as popular as GMRES but we'll find a use for it today.

Arnoldi method for eigenvalues (Harmonic Ritz)

Choose $\mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2b, ..., A^{m-1}b\}$ and $\mathcal{L}_m = A\mathcal{K}_m$.

The eigenvalue problem residual is

$$r_i^{(m)} = Au_i^{(m)} - \theta_i^{(m)} u_i^{(m)} = AV_m y_i^{(m)} - \theta_i^{(m)} V_m y_i^{(m)}$$

And the constraints imposed are $r_i^{(m)} \perp A\mathcal{K}_m$: i.e. $(AV_m)^* r_i^{(m)} = 0$.

$$(AV_m)^*(AV_m y_i^{(m)} - \theta_i^{(m)} V_m y_i^{(m)}) = 0$$

Replace AV_m with $V_m H_m + h_{m+1,m}\, v_{m+1}\, e_m{}^T$ throughout and simplify.

$$(H_m^* H_m + h_{m+1,m}^2\, e_m e_m{}^T) y_i^{(m)} - \theta_i^{(m)} H_m^* y_i^{(m)} = 0$$

Harmonic Ritz values solve the generalised eigenvalue problem

$$\boxed{(H_m^* H_m + h_{m+1,m}^2\, e_m e_m{}^T) y_i^{(m)} = \theta_i^{(m)} H_m^* y_i^{(m)}}$$

The full family

All methods are built on the Krylov subspace

$$\mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2b, ..., A^{m-1}b\}.$$

The choice of $\mathcal{L}$ and the form of the residual determines the method.

	$\mathcal{L} = \mathcal{K}$	$\mathcal{L} = A\mathcal{K}$
$Au = \lambda u$	Arnoldi (Ritz)	Arnoldi (Harmonic Ritz)
$Ax = b$	FOM	GMRES

Terminology:

- Galerkin: $\mathcal{L} = \mathcal{K}$
- Petrov-Galerkin: $\mathcal{L} \neq \mathcal{K}$

Shift-and-invert

Standard Arnoldi tends to converge to exterior eigenvalues.

If you want smallest eigenvalues, one idea is to apply the method to A^{-1} instead of A.

The eigenvalues are reciprocals, so the smallest eigenvalues of A look like exterior eigenvalues of A^{-1} (with eigenvectors unchanged).

Build the Krylov subspace

$$\mathcal{K}_m(A^{-1}, b) = \operatorname{span}\left\{b, A^{-1}b, A^{-2}b, \ldots, A^{-(m-1)}b\right\}$$

and apply Arnoldi as usual from here.

This works well, but requires you can cheaply apply $A^{-1}v$.

(If eigenvalues close to some μ are desired, use instead $(A - \mu I)^{-1}$).

Ritz vs harmonic Ritz values

There is an interpretation of the harmonic Ritz values along these lines.

Ritz: $u \in \mathcal{K}_m$ and $Au - \theta u \perp \mathcal{K}_m$.

Harmonic Ritz: $u \in \mathcal{K}_m$ and $Au - \theta u \perp A\mathcal{K}_m$.

Let $w = Au \in A\mathcal{K}_m$. Then the harmonic Ritz residual in terms of w is

$$w - \theta A^{-1} w \perp A\mathcal{K}_m \implies \theta^{-1} w - A^{-1} w \perp A\mathcal{K}_m$$

So harmonic Ritz looks like an ordinary Ritz method applied to A^{-1}, albeit using the space $A\mathcal{K}_m$. i.e.

Harmonic Ritz: $w \in A\mathcal{K}_m$ and $A^{-1} w - \theta^{-1} w \perp A\mathcal{K}_m$.

In practice the harmonic Ritz values are somewhat better at small eigenvalues but not comparable to true shift-and-invert (which uses $\mathcal{K}_m(A^{-1}, b)$).

The Hermitian case

If $A = A^*$ then there is considerable simplification in the Arnoldi method.

Recall the third Arnolid relation: $V_m^* A V_m = H_m$.

By construction, H_m is upper Hessenberg.

But now also, we see since $A = A^*$ it must be Hermitian.

A matrix that is upper Hessenberg and Hermitian must be tridiagonal.

So H_m is usually relabelled T_m and it is necessarily of the form

$$T_m = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_m \end{bmatrix}$$

The Lanczos process

The implications of H_m becoming the tridiagonal T_m are that most of the inner products we usually compute in the Arnoldi process are already zero!

In fact, at most two previous vectors need subtracting at each iteration (two of the entries in a column of T_m, the third being the resulting norm).

The algorithm is now $O(1)$ work per iteration, $O(m)$ overall, making it readily applicable for very large m.

Loss of orthogonality can appear in the basis, owing to the mathematical enforcement of orthogonality only, rather than explicit numerical enforcement.

If full reorthogonalisation is used, the algorithm cost blows back out to $O(m^2)$, negating much of the advantage.

Selective or partial reorthogonalisation schemes are popular.

The fuller family

We can now add a whole second table to our family of methods – those for Hermitian matrices all get separate names and specialised algorithms.

$A \neq A^*$	$\mathcal{L} = \mathcal{K}$	$\mathcal{L} = A\mathcal{K}$
$Au = \lambda u$	Arnoldi (Ritz)	Arnoldi (Harmonic Ritz)
$Ax = b$	FOM	GMRES

$A = A^*$	$\mathcal{L} = \mathcal{K}$	$\mathcal{L} = A\mathcal{K}$
$Au = \lambda u$	Lanczos (Ritz)	Lanczos (Harmonic Ritz)
$Ax = b$	Conjugate gradient	MINRES

Notably conjugate gradient is derived as the symmetric version of FOM when A is additionally positive definite, with the optimality property

$$\|\varepsilon^{(m)}\|_A = \min_{p_m(0)=1} \|p_m(A)\varepsilon^{(0)}\|_A$$

where $\varepsilon^{(m)} = x - x^{(m)}$ is the error and the A-norm is $\|v\|_A = \sqrt{v^* A v}$.

Restarted GMRES with deflation

Back to GMRES, we have seen that restarting may be necessary for reasons of memory or runtime.

But this seems wasteful – throwing away the entire Krylov subspace just to start over with a new initial guess.

There are numerous approaches to recycling some of this information, here we present one approach: GMRES with deflated restarting.

The idea is to use the existing Krylov subspace to estimate a few of the smallest eigenvalue/eigenvector pairs.

- Smallest eigenvalues (closest to zero) are those that cause the most trouble for convergence
- Harmonic Ritz projection is the tool for the job

Restarted GMRES with deflation

We can exploit the fact that the GMRES residual vector $r^{(m)} = b - Ax^{(m)}$ and the harmonic Ritz residual vector $r_i'^{(m)} = Au_i^{(m)} - \theta_i u_i^{(m)}$ are parallel (if built from the same Krylov subspace).

Proof: both belong to $\mathcal{K}_{m+1}$ (dim $m+1$) and both are orthogonal to $\mathcal{L} = A\mathcal{K}_m$ (dim m), hence, parallel.

So conceptually starting the next GMRES cycle with $u_i^{(m)}$ generates $Au_i^{(m)}$ as the next vector, and this span already includes the "proper" restart vector $r^{(m)} = r_{\text{new}}^{(0)}$.

If $u_i^{(m)}$ was a true eigenvector u, at the end of the next cycle, the new residual $r_{\text{new}}^{(m)}$ will have no component in that direction.

Proof: GMRES takes $r \perp A\mathcal{K}_m$, but if $u \in \mathcal{K}_m$ then $u \in A\mathcal{K}_m$ also (eigenvector), so $r \perp u$.

Restarted GMRES with deflation

Better yet is to include multiple (k, say) harmonic Ritz vectors, all stuffed at the front of the space.

So the next cycle of GMRES builds an orthonormal basis for

$$\mathcal{S} = \text{span}\left\{ u_1^{(m)}, u_2^{(m)}, \ldots, u_k^{(m)}, r_{\text{new}}^{(0)}, A r_{\text{new}}^{(0)}, \ldots A^{m-k-1} r_{\text{new}}^{(0)} \right\}$$

where as usual $r_{\text{new}}^{(0)}$ is the relabelled residual $r^{(m)}$ from the previous cycle.

If you choose a starting vector s (say) as a linear combination of the $u_i^{(m)}$ then this is truly a Krylov subspace

$$\mathcal{S} = \mathcal{K}_m(A, s)$$

so we still have ourselves a genuine Krylov subspace method.

Restarted GMRES with deflation

$$\mathcal{S} = \text{span}\left\{ u_1^{(m)}, u_2^{(m)}, \ldots, u_k^{(m)}, r_{\text{new}}^{(0)}, A r_{\text{new}}^{(0)}, \ldots A^{m-k-1} r_{\text{new}}^{(0)} \right\}$$

In practice you

- Orthonormalise the harmonic Ritz vectors (Schur decomposition)
- Orthonormalise $r_{\text{new}}^{(0)}$ against each of those
- Build out the rest of the space from there in the usual way

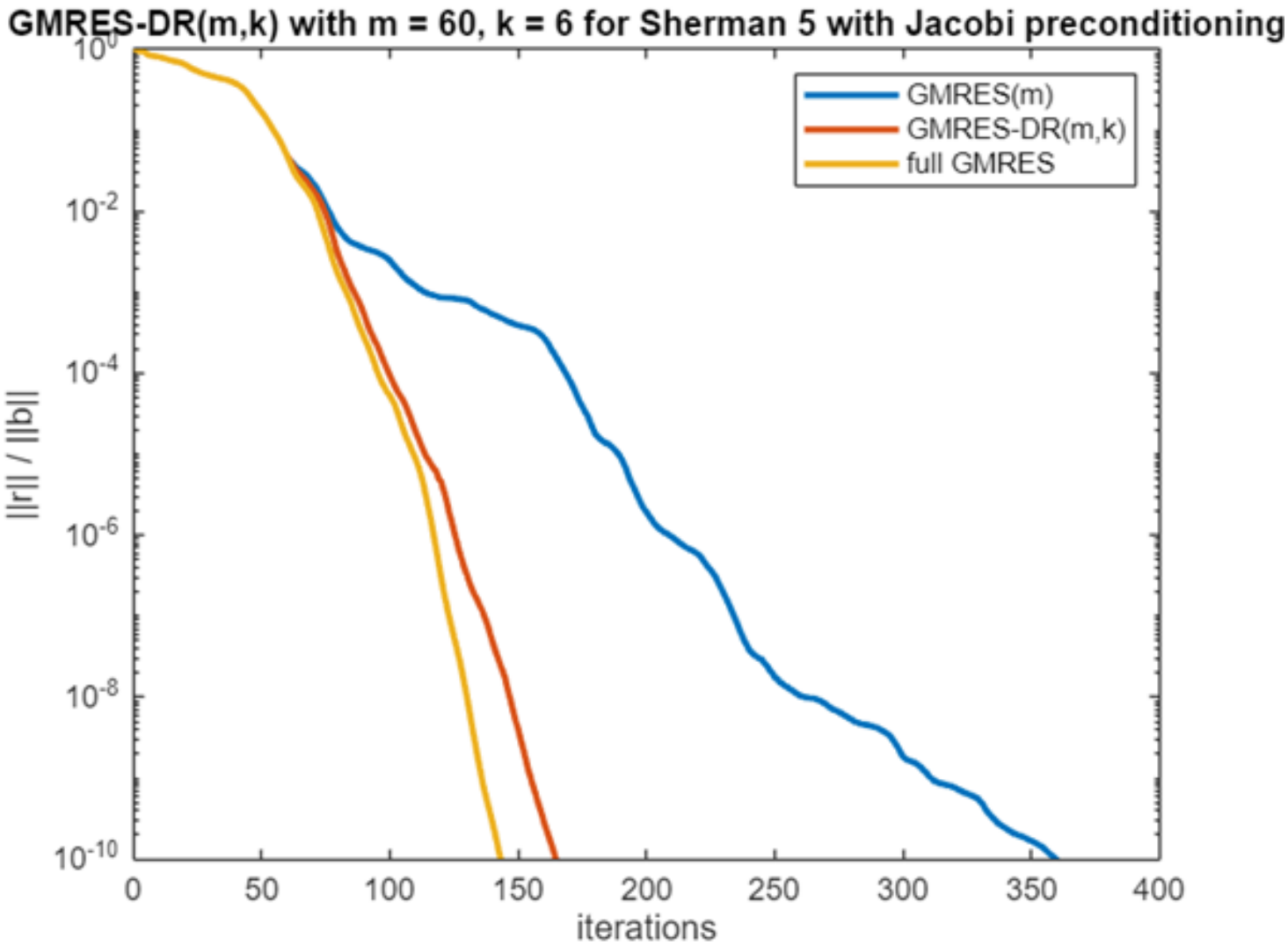
(better still, do the first two steps in m-dimensions and map back up).

You derive a generalised Arnoldi relation $AV_m = V_{m+1}\overline{H}_m$ where $\overline{H}_m$ has the usual upper Hessenberg structure except for a full top-left $(k+1) \times k$ block.

You no longer have $r^{(0)} = V_m(\beta e_1)$, but it's definitely in $\mathcal{S}$, so just calculate $c = V_m^* r^{(0)}$ and solve

$$\min \|c - \overline{H}_m y^{(m)}\| \qquad \text{then} \qquad x^{(m)} = x^{(0)} + V_m y^{(m)}$$

Restarted GMRES with deflation



Shift invariance

Suppose you have multiple shifted linear systems to solve:

$$(A + \sigma_j I)x_j = b$$

You only need the one Krylov subspace, since

$$\mathcal{K}_m(A + \sigma I, b) = \mathcal{K}_m(A, b) = \text{span}\{b, Ab, A^2 b, ..., A^{m-1} b\}$$

The Arnoldi relations carry right over:

$$\begin{aligned}
(A + \sigma I) V_m &= A V_m + \sigma V_m \\
&= V_m H_m + h_{m+1,m} v_{m+1} e_m^T + \sigma V_m \\
&= V_m (H_m + \sigma I) + h_{m+1,m} v_{m+1} e_m^T \\
&= V_{m+1} (\overline{H}_m + \sigma E) \qquad E = [I; 0]
\end{aligned}$$

So you pay a one-off cost building the basis, then you get arbitrarily many cheap shifted solves.

Matrix Functions

Functions of a matrix, or simply "matrix functions" arise in many applications. e.g. exponential integrators, fractional PDEs, Markov chains

As is often the case, we may not really need $f(A)$, but only its action on a vector i.e. $f(A)b$.

Once again, Krylov subspace methods are available to help.

Given the Arnoldi decomposition of $\mathcal{K}_m(A, b)$ with the usual V_m and H_m we know the projection of A onto $\mathcal{K}_m$ is given by $A_{\mathcal{K}} = V_m^* A V_m = H_m$.

Meanwhile b has coordinate vector in the Krylov subspace given by βe_1.

So it makes sense to propose the Krylov approximation

$$f(A)b \approx \beta V_m f(H_m) e_1$$

Matrix Functions

This formula can be further motivated by the following argument. Consider the Cauchy integral formula for $f(A)$

$$f(A)b := \frac{1}{2\pi i}\int_\Gamma f(z)(zI - A)^{-1}\, b\, \mathrm{d}z$$

We will use Krylov methods to approximate $x(z) = (zI - A)^{-1}b$. Substitute the FOM approximation (shifted system!)

$$x(z) \approx x_m(z) = \beta V_m (zI - H_m)^{-1} e_1$$

So

$$\begin{aligned} f(A)b &\approx \frac{1}{2\pi i}\int_\Gamma f(z)\beta V_m (zI - H_m)^{-1} e_1\, \mathrm{d}z \\ &= \beta V_m \left[\frac{1}{2\pi i}\int_\Gamma f(z)(zI - H_m)^{-1}\, e_1\, \mathrm{d}z\right] \\ &= \beta V_m f(H_m) e_1 \end{aligned}$$

(If $f(z) = 1/z$ this recovers FOM itself.)

Matrix Function residual

How do we find a suitable residual for $f(A)b$? For $f = \exp$ at least, we can define an ODE-based residual, since $\exp(tA)b$ is the solution to

$$u'(t) = Au(t), \quad u(0) = b$$

So define the ODE residual

$$\rho_m(t) := u'_m(t) - Au_m(t)$$

and substitute

$$u_m(t) = \beta V_m \exp(tH_m)e_1$$
$$u'_m(t) = \beta V_m H_m \exp(tH_m)e_1$$

and

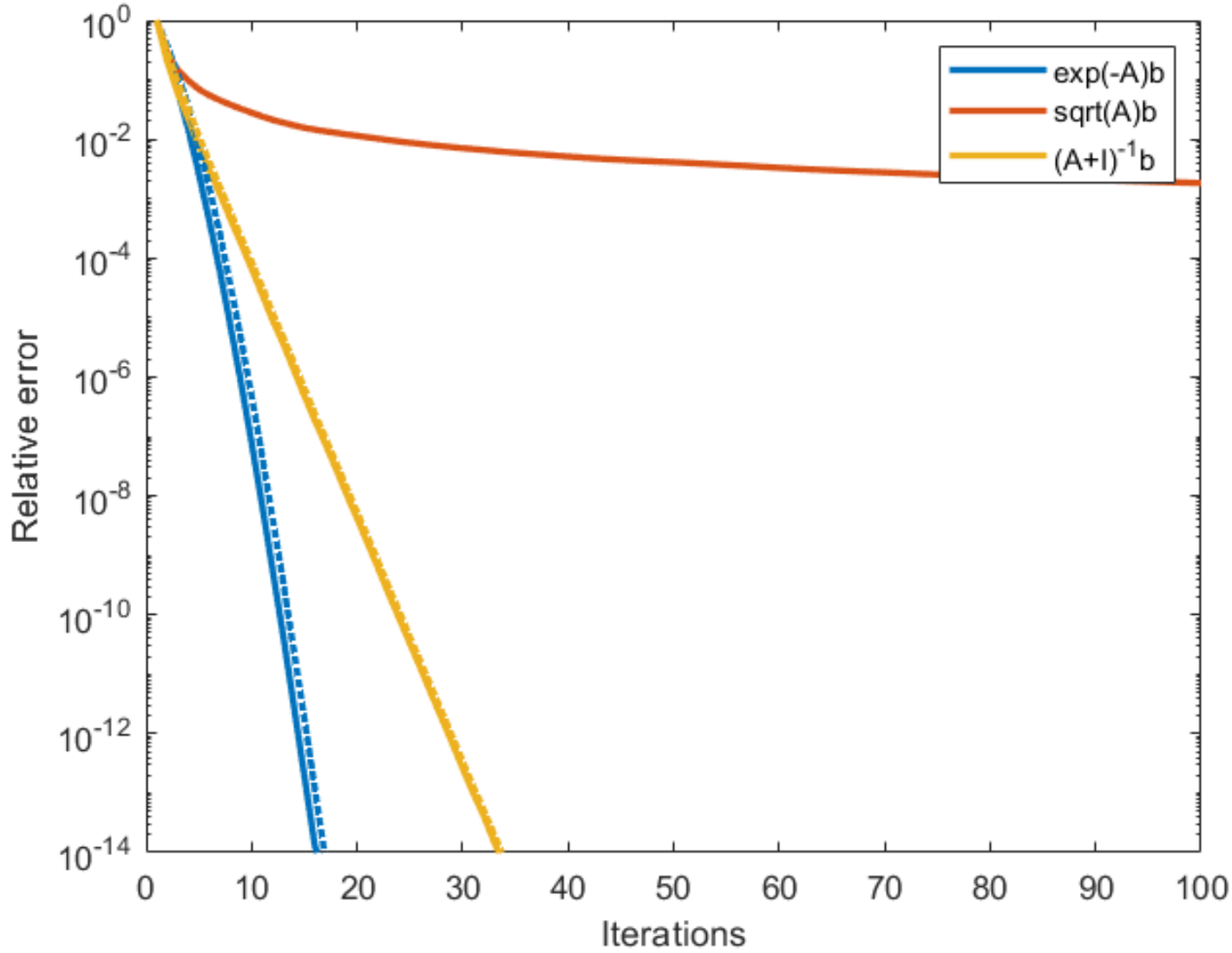
$$\begin{aligned} Au_m(t) &= \beta AV_m \exp(tH_m)e_1 \\ &= \beta(V_m H_m + h_{m+1,m} v_{m+1}\, e_m^T) \exp(tH_m)e_1 \end{aligned}$$

to derive

$$\|\rho_m(t)\| = \|u'_m(t) - Au_m(t)\| = \beta h_{m+1,m} |e_m^T exp(tH_m)e_1|$$

Matrix Function convergence

```
A = gallery('tridiag', 1000); b = rand(size(A,1),1);
```



Extended Krylov Subspace Methods

Matrix functions computed using Krylov subspace methods are naturally polynomial-based approximations.

For a rational approximation (pole at the origin) we can instead build the extended Krylov subspace

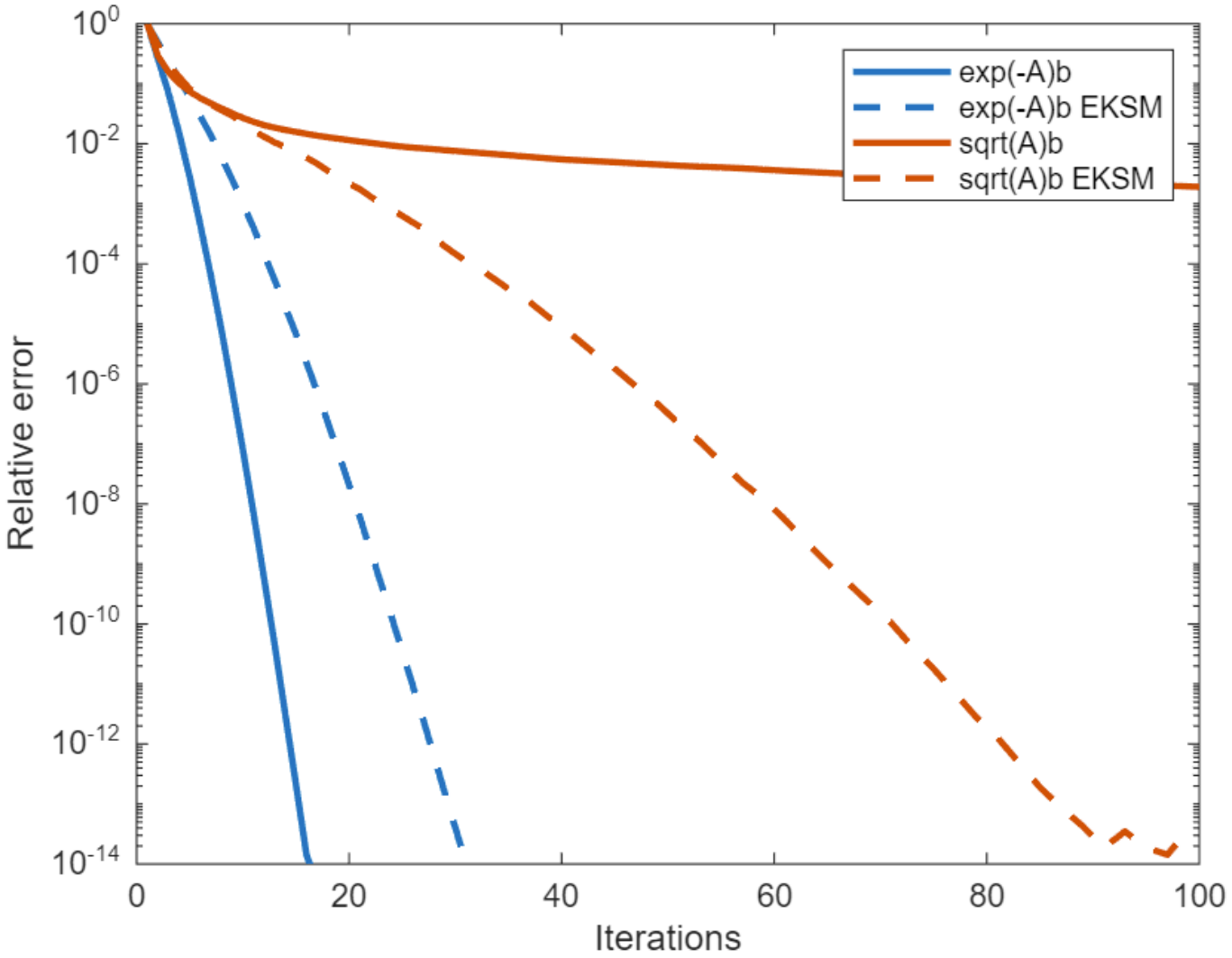
$$\mathcal{K}_{\ell,m}(A, v) = \operatorname{span}\left\{A^{-\ell+1}v, \ldots, A^{-1}v, v, Av, \ldots, A^{m-1}v\right\}$$

This can be attractive when

- Your matrix is cheap to solve linear systems (e.g. tridiagonal)
- Your function is not well approximated by a polynomial on the spectrum of A (e.g. $\sqrt{A}$)

Matrix function convergence (EKSM)

```
A = gallery('tridiag', 1000); b = rand(size(A,1),1);
```



Flexible GMRES

Standard GMRES with (right) preconditioning uses $\mathcal{K}_m(AM^{-1}, b)$.

But what if the preconditioner varies each time it is applied? e.g. the preconditioner is itself an iterative method, solved only to some tolerance.

Arnoldi step needs modifying: $z_j = M_j^{-1} v_j$, $w = Az_j$ then orthogonalise w as usual to get v_{j+1}. So we need to record both z_j and v_j:

$$AZ_m = V_{m+1}\overline{H}_m$$

$$\min \|\beta e_1 - \overline{H}_m y^{(m)}\|$$

$$x^{(m)} = Z_m y^{(m)}$$

Double the memory cost, storing both matrices V_{m+1} and Z_m.

But flexibility to vary the preconditioner with each step.

Jacobian-Free Newton-Krylov

Consider Newton's method for the nonlinear system $F(x) = 0$. Each step requires solving

$$J(x^k)\delta x^{(k)} = -F(x^{(k)})$$

where $J = \partial F/\partial x$ is the Jacobian matrix. A Krylov method for this system requires Jacobian-vector products $J(x^{(k)})v$ to build the Krylov subspace.

But a Jacobian-vector product is nothing but a directional derivative of F:

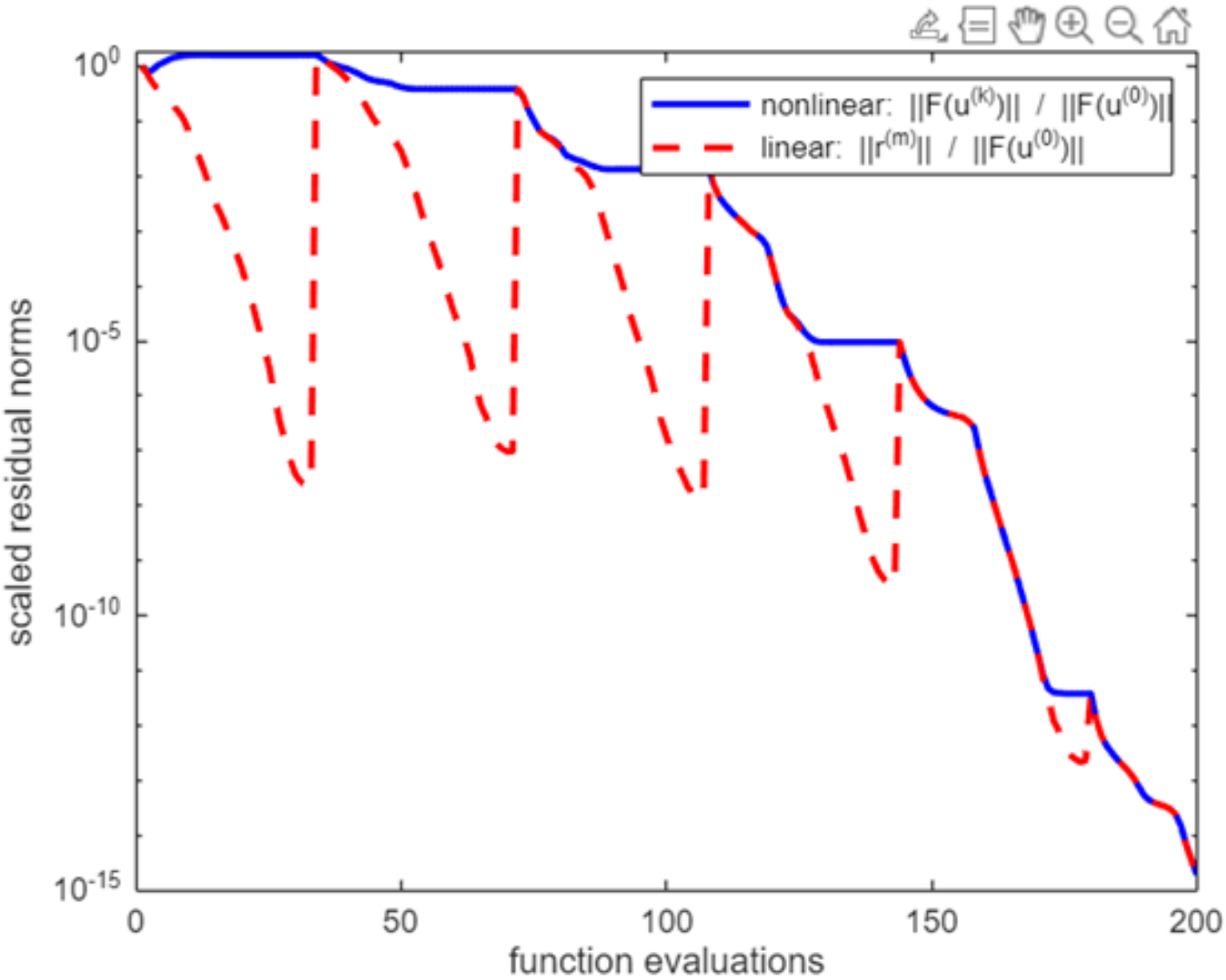
$$J(x^k)v = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} F(x^{(k)} + tv)$$

So using automatic differentiation (or finite differences), Jacobian-vector products can be calculated (or approximated) entirely matrix-free.

This leads to a Jacobian-free Newton-Krylov method (JFNK).

(Depending on your preconditioner it could even be fully matrix-free.)

JFNK with over-solving



Jacobian-Free Newton-Krylov

Further efficiencies are possible by matching the accuracy of the inner Krylov solves to the outer Newton level accuracy (avoid over-solving).

We're solving

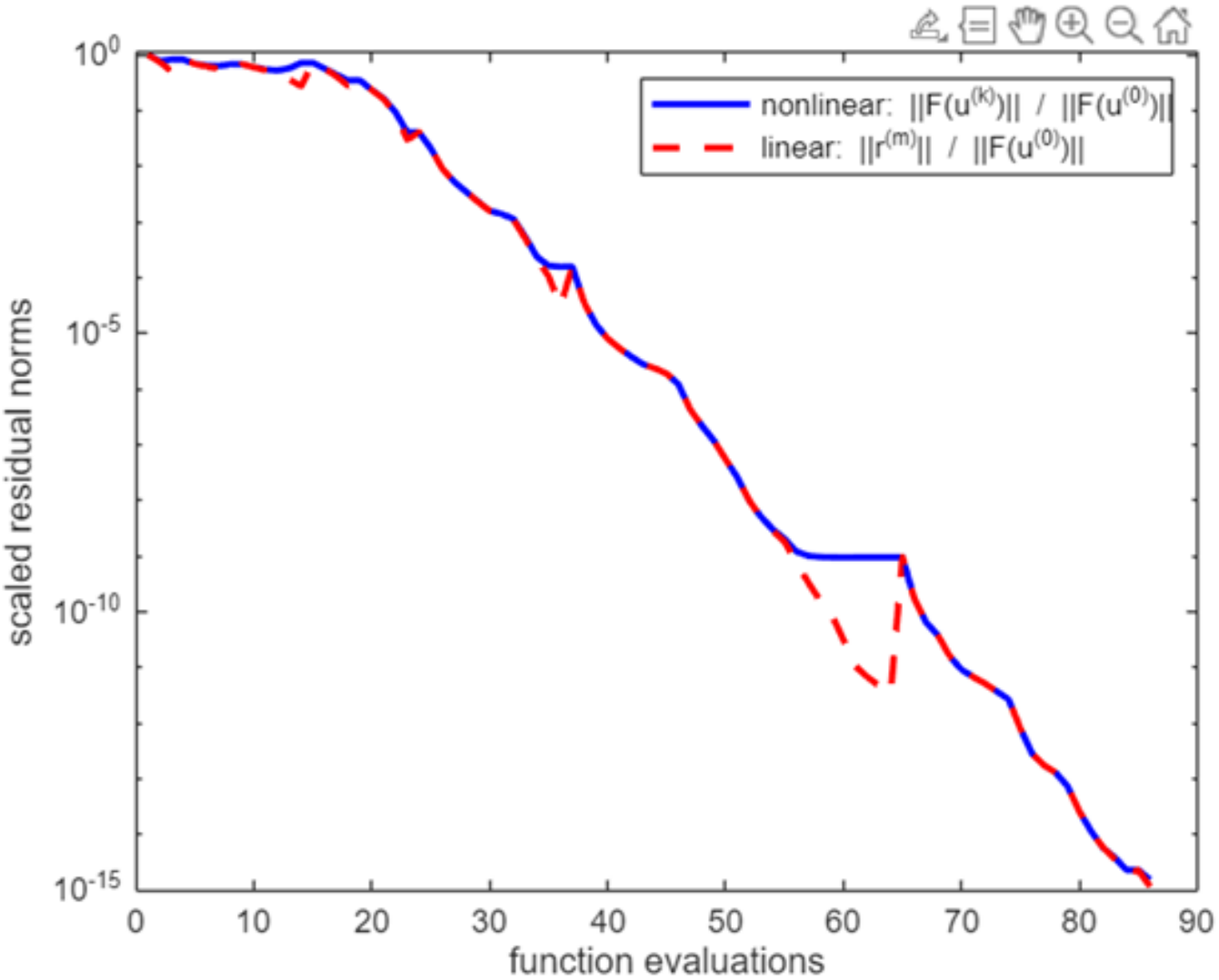
$$J(x^k)\delta x^{(k)} = -F(x^{(k)})$$

Use tolerance for inner (linear) solve

$$\|J(x^k)\delta x^{(k)} + F(x^{(k)})\| \leq \gamma_k \|F(x^{(k)})\|$$

where the "forcing term" $\gamma_k < 1$ and is updated as the iterations proceed.

JFNK reducing over-solving



Conclusion

In this series we have covered:

- Arnoldi process
- Arnoldi method for eigenvalues
- GMRES
- Preconditioning
- Other related Krylov methods
- Deflation
- Matrix functions
- Various other tips and tricks