Fast Iterative Solvers for Sequences of Linear Systems

Eric de Sturler, Department of Mathematics
Virginia Tech
sturler@vt.edu
http://www.math.vt.edu/people/sturler

Institute for Computer Science, Academy of Sciences of the Czech Republic,
Prague, Czech Republic, October 19-26, 2008
Collaborators and Applications

- Shun Wang, Mike Parks, Chris Siefert, Greg Mackey, UIUC/Sandia
- Misha Kilmer, Tufts University
- Glaucio Paulino, Bob Dodds, Philippe Geubelle, UIUC
- Ernst (TU Freiberg), Cai (Boulder), Zabaras (Cornell)
- Emilio Silva, Luis Motta, USP
- Phani Nukala, ORNL
- Hector Klié, University of Texas at Austin
- Ceperley, Jeongnim Kim, UIUC

Applications: crack propagation, fatigue & fracture, tomography, topology optimization, nonlinear equations, materials science and computational physics, large-scale fracture in disordered materials, electromagnetic wave propagation, stochastic FE/uncertainty, QMC

Funded by DOE, NSF
Overview

- Sequences of linear systems
- Examples
  - Crack propagation
  - Topology optimization
  - Tomography
- Krylov methods crash course
- Solvers for sequences of linear systems
- Why recycle a search space?
- Experimental results and analysis of recycling
- Conclusions and future work
- Good Reading
Sequences of Linear Systems

- Computational problems often involve a sequence of systems with small or localized changes
  - Evolutionary problems, nonlinear problems and optimization, parameter estimation
  - Accurate simulation requires solution of hundreds to thousands of large, sparse, linear systems
  - Crack propagation, topology optimization, tomography

- Fast solution by exploiting the slowly changing nature of problem or special structural changes (reuse old solutions – obvious idea)

- Recycle (adapt & reuse) search spaces from previous problems

- Adapt preconditioners (especially AMR)
Example: Crack Propagation
(with Guebelle, Maiti, and Parks)
Example: Topology Optimization

Optimize material distribution, $\rho$, in design domain

Minimize compliance $u^T K(\rho) u$, where $K(\rho) u = f$

(with Paulino and Wang)
Topology Optimization - example
Diffusive Tomography

- Developed for medical imaging
  - Near IR not significantly absorbed by tissue, but it is highly scattered.
  - Photons diffuse through tissue in density waves
- Microstructure (electrical permittivity, magnetic permeability) is not resolvable.
- Macro-structure (optical absorption and diffusivity) is recoverable from light emerging from tissue
- Applications: Breast and brain imaging where local rise in absorption is tied to presence of oxygenated hemoglobin

(with Misha Kilmer)
Diffusion Forward Model

- Photon fluence due to source input $s$ and frequency $\omega$ given by solution to the following PDE

$$-\nabla \cdot (D(r) \nabla f^s_{\omega}(r)) + \left[ \mu(r) + i \frac{\omega}{c} \right] f^s_{\omega}(r) = g^s(r)$$

plus boundary conditions

- Desired are parameters, $p$, for diffusivity, $D$, and absorption, $\mu$

- In matrix form

$$A_{\omega}(D(p), \mu(p)) f^s_{\omega} = g$$

with $A$ sparse, symmetric positive definite real part

- Solution at some positions known from measurements. Find parameters that match solution to data.
Consider $Ax = b$ and no (or explicit) preconditioning. Given $x_0$ and $r_0 = b - Ax_0$, compute optimal update $z$ from $K^m(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}$:

$$\min_{z \in K^m(A, r_0)} \|b - A(x_0 + z)\|_2 \iff \min_{z \in K^m(A, r_0)} \|r_0 - Az\|_2$$

Let $K_m = \begin{bmatrix} r_0 & Ar_0 & A^2r_0 & \cdots & A^{m-1}r_0 \end{bmatrix}$, then $z = K_m \zeta$,

and we must solve the following least squares problem

$$AK_m \zeta \approx r_0 \iff \begin{bmatrix} Ar_0 & A^2 r_0 & \cdots & A^m r_0 \end{bmatrix} \zeta \approx r_0$$

Do this accurately and efficiently every iteration for increasing $m$. Arnoldi recurrence: $AV_m = V_{m+1}H_m$, where $v_1 = r_0 / \|r_0\|_2$, $V_m = V_{m+1}H_m$, and range$(V_{m+1}) = \text{range}(K_{m+1})$

$$\|r_0 - AV_m y_m\|_2 = \|V_{m+1}e_1\|_2 \|r_0\|_2 - V_{m+1}H_m y_m\|_2 = \|e_1\|_2 \|r_0\|_2 - H_m y_m\|_2$$
Krylov Spaces

Krylov space is a *space of polynomials in a matrix times a vector*.

Krylov space inherits the approximation properties that polynomials on the real line or in the complex plane possess.

Let $B$ be diagonalizable, $B = VA V^{-1}$ (for simplicity). Then $B^2 = VA V^{-1}VA V^{-1} = V A^2 V^{-1}$ and generally $B^k = VA^k V^{-1}$.

So, the polynomial $p_m(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_m t^m$ applied to $B$ gives

$$p_m(B) = V \left( \alpha_0 I + \alpha_1 A + \alpha_2 A^2 + \cdots + \alpha_m A^m \right) V^{-1}$$

and hence

$$p_m(B) = V p_m(A) V^{-1} = V \text{diag}(p_m(\lambda_1), \ldots, p_m(\lambda_n)) V^{-1}$$

The polynomial is applied to the eigenvalues individually.

Approximate solutions to linear systems, eigenvalue problems, and more general problems using polynomial approximation.
Approximation by Matrix Polynomials

Let $B = V \Lambda V^{-1}$, let $\Lambda(B) \subset \Omega \subset \mathbb{C}$.

If $p_m(t) \approx \frac{1}{t}$ for all $t \in \Omega$, then $p_m(B) \approx B^{-1}$.

Let $y = V\zeta$. Then $p_m(B)y = \sum_i v_i p_m(\lambda_i) \zeta_i \approx \sum_i v_i \frac{\zeta_i}{\lambda_i} = B^{-1}y$

Furthermore, let $\varepsilon \approx 0$ and $|\lambda_i - \lambda_j| > \delta$ (for some eigenvalue $\lambda_i$).

If $p_m(t) = \begin{cases} \varepsilon, & t \in \Omega \text{ and } |t - \lambda_i| > \delta, \\ 1, & t = \lambda_i, \end{cases}$

then $p_m(B)y \approx v_i \zeta_i$.

If we can construct such polynomials for modest $m$ we have an efficient linear solver or eigensolver.
Solvers for Sequences of Linear Systems

- Iterative (Krylov) methods build search space and compute solution by projection
- Building search space often dominates cost
- Initial convergence often poor, reasonable size search space needed, then superlinear convergence
- Get fast convergence rate and good initial guess immediately by recycling selected search spaces from previous systems
- How to select the right subspace to recycle?
How to Select the Right Space to Recycle?

- Typically, a recurring subspace exists such that almost any Krylov space (from any starting vector) has large components in that space (reason why restarting is bad).

- Optimality derives from orthogonal projection: new search directions should be far from this recurring subspace (after resolving it) for fast convergence.

- If such a recurring subspace persists (approx) from one system to the next, it can be recycled.

- Typically true when changes to problem are small and/or highly localized.

- Currently, we use two methods that differ in how they capture the recurring subspace.
Near Optimal Methods (single system)

- GMRES optimal in number of iterations, but very expensive in time and memory (unless convergence very rapid)
- Limit resources: restart after $m$ steps with best solution as initial guess or orthogonalize only against latest $m$ vectors
- Unfortunately, these strategies can slow down convergence drastically and even prevent convergence
- Solution: keep some carefully selected subspace
- Various possible choices (and ways to use them)
  - Optimality requires orthogonal projection. After correction subspace important only for orthogonality relations. You cannot maintain orthogonality against discarded subspace
  - Approximate invariant subspace. Based on assumptions on normality and spectrum. Not for general problems, but often works
Convergence GMRES (motivational)

Consider $Ax = b$, and relate convergence to polynomials.

$$x_m = x_0 + z_m \text{ where } z_m \in \text{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{m-1}r_0\}$$

$$r_m = r_0 - Az_m \in \text{span}\{r_0, Ar_0, \ldots, A^mr_0\}$$

Assume $A = U\Lambda U^{-1}$ (diagonalizable), then residual at step $m$

$$\min_{z \in K^m(A, r_0)} \|r_0 - Az\| = \min_{p_m(0) = 1} \|p_m(A)r_0\| \leq \|r_0\||U||U^{-1}| \min_{p_m(0) = 1, \lambda \in \Lambda(A)} \max |p_m(\lambda)|$$

For normal matrix this bound is sharp. For highly nonnormal matrix this bound may not be useful.

$\kappa(U)$ small: convergence determined by minimal polynomial

Clustered eigenvalues yield fast convergence: preconditioning

Eigenvalues surrounding origin yields very poor convergence.
GMRES-DR / GCRO-DR

Consider again the bound on the residual norm

$$\min_{z \in K^m(A, r_0)} \|r_0 - Az\| = \min_{p_m(0)=1} \|p_m(A)r_0\| \leq \|r_0\|\|U\||U^{-1}\| \min_{p_m(0)=1} \max_{\lambda \in \Lambda(A)} |p_m(\lambda)|$$

If $\kappa(U)$ is not large we can improve this bound by removing those eigenvalues that make $\min_{p_m(0)=1} \max_{\lambda \in \Lambda(A)} |p_m(\lambda)|$ large.

For PDEs these are often (but not always) the small eigenvalues. One way of doing this is to augment the Krylov space with the corresponding approximate eigenvectors.

In general, we do not have the exact eigenvalues and eigenvectors but approximations.

For strongly nonsymmetric problems this approach is dubious

(GMRES-DR for single system - Morgan’02, GCRO-DR for sequence of systems - Parks, de Sturler et al. ’06)
How do we get approximate eigenpairs? From the (augmented) Arnoldi recurrence. Let $AV_m = V_{m+1} H_m$. The eigenpair approximation over the space $\text{Range}(V_m)$ is given by

$$AV_m y - \theta V_m y \perp V_m \Leftrightarrow V_m^H AV_m y = \theta y$$

(note $H_m = V_m^* AV_m$)

The approximate eigenpair $(\theta, V_m y)$ is called a Ritz pair. Since the Ritz pairs for small eigenvalues are often inaccurate, we use the harmonic Ritz pairs, associated with $A^{-1}$ over same space

$$A^{-1} (AV_m) y - \theta AV_m y \perp AV_m \Leftrightarrow H_m^* y - \theta H_m^* H_m y = 0$$

After solving the generalized eigenvalue problem we have the harmonic Ritz pair $(\frac{1}{\theta}, V_m y)$. New approximate eigenpairs computed after every cycle
GCROT  (de Sturler’99)

- Measure angles between subsequent search spaces
- Small angles indicate that orthogonalization is needed wrt to that space
- Otherwise explore search space over which already optimized; moreover, we reintroduce components indirections that had been removed – recurring subspace
- We can show that it is optimal to maintain orthogonality wrt space associated with smallest angles
- Resulting method (single system): GCROT
- Use space generated also for recycling: rGCROT (Parks, de Sturler, et al. ’06)
GCROT: Selective orthogonality

Compare iterations of restarted GMRES with GCROT, which maintains orthogonality against sequence of selected subspaces. Time-wise GCROT has additional advantage of working with a smaller subspaces (cheaper iterations).

Convection-diffusion problem with strong convection
Results for Crack Propagation

Problem from Guebelle & Maiti, UIUC

Convergence results with and without recycling using IC(0) preconditioner
Convergence Crack Propagation

Convergence for five consecutive problems – total number of matrix-vector products for full GMRES

Convergence for GCRODR (recycling)
Topology Opt. Convergence results

Results: 84x28x14 mesh (107K dofs) and 180x60x30 (1M dofs) (on PC)

Currently more complicated models up to 2M dof (on PC)
Optimization Result

Optimization using 180 x 60 x 30 mesh, more than 1M dofs.
Approximately 130 optimization steps required
What Makes a Good Recycling Method?

- Methods need to satisfy three main requirements
  1. Identify and converge to effective recycle space in modest number of iterations (recurring subspace, invariant subspaces) and do so over solution of multiple (changing) systems
  2. Yield significant convergence improvement for recycle space of modest dimension
  3. Converge quickly to perturbed recycle space for updated matrix, effective mechanism for cheaply updating recycle space to reflect changes. How quickly does the method learn/adapt?

- Use knowledge of application to tune the recycling (recurring subspace, invariant subspace, solution space)

- Type of matrix update, problem and algorithm dependent
  - timesteps, rank-k update in quasi-Newton method, small change or structured, localized nonlinear behavior (crack propagation), line search, other structured updates
Fast Solution of Sequences: Issues

- Analysis of convergence of a method with recycling when keeping approximate (invariant, solution) subspace
- How do relevant subspaces change under changes in the matrix?
  - Perturbation of invariant subspaces
  - Perturbation of solution subspaces
- How quickly does recycled space adapt to new matrix?
- Use knowledge of application to tune the recycling (subspace)
  - If rhs similar recycle solution spaces as well
- Type of matrix update, problem and algorithm dependent
  - timesteps, rank-k update in quasi-Newton method, localized nonlinear behavior (crack propagation), line search
- Nature of PDE and changes in parameters
- Varying behavior over time: certain modes stationary while others still change (does the method learn?)
- Multiple parameterized matrices and right hand sides
Convergence

Consider $Ax = b$, solved by GCRO-DR or symmetric version

$Q_l$ is invariant subspace of dim $l$. $C_k$ approximates subspace $Q_l$ (where $k \geq l$), and $\delta = \| (I - \Pi_C) \Pi Q \|_2 < 1$. Let $r_0 = (I - \Pi_C) b$, $v_1 = r_0 / \| r_0 \|$ and $(I - \Pi_C) A V_m = V_{m+1} H_m$ generate new search space.

Relate convergence of GCRO-DR to ‘deflated problem’, where all components in $Q_l$ have been removed from residual/rhs.

non-Hermitian:

$$\min_{d_1 \in V_m + C_k} \| b - d_1 \|_2 \leq \min_{d_2 \in (I-P_Q)V_m} \| (I - P_Q) r_1 - d_2 \|_2 + \frac{\delta}{1 - \delta} \| P_Q \|_2^2 \| (I - \Pi_V) r_1 \|_2$$

Hermitian:

$$\min_{d_1 \in V_m + C_k} \| b - d_1 \|_2 \leq \frac{1 - \delta}{\delta} \min_{d_2 \in (I - \Pi_Q)V_m} \| (I - \Pi_Q) r_0 - d_2 \|_2$$

(joint work with Mike Parks following ideas from Simoncini&Szyld’05)
Convergence

Hermitian:

\[
\min_{d_1 \in V_m + C_k} \| b - d_1 \|_2 \leq \frac{1 - \delta}{\delta} \min_{d_2 \in (I - \Pi Q) V_m} \| (I - \Pi Q) r_0 - d_2 \|_2
\]

However, \( \min_{d_2 \in (I - P_Q) V_m} \| (I - P_Q) r_0 - d_2 \|_2 \) not quite ‘deflated problem’.

\[
\text{Range}(V_m) = K_m \left( (I - \Pi C) A (I - \Pi C), (I - \Pi C) b \right)
\]

So, \( d_2 \neq p(A)(I - P_Q) b \)

However, in practice we see (numerically) very similar behavior

Why?
Convergence

Let $A$ Hermitian, $\|A\|_2 = 1$, $Q$ associated with small eigenvalues

Let $\begin{bmatrix} C & W \end{bmatrix}$ and $\begin{bmatrix} Q & Y \end{bmatrix}$ be unitary and $C^*Q = \Phi \begin{bmatrix} \Omega & 0 \end{bmatrix} \Psi^*$ (SVD) then

\[
Q\Psi = C\Phi \begin{bmatrix} \Omega \\ 0 \end{bmatrix} + W \begin{bmatrix} \Sigma \\ 0 \end{bmatrix},
\]

\[
Y\Theta = C\Phi \begin{bmatrix} -\Sigma & 0 & 0 \\ 0 & 0 & I \end{bmatrix} + W \begin{bmatrix} \Omega & 0 & 0 \\ 0 & I & 0 \end{bmatrix}, \quad \text{etc.}
\]

($\sigma_{\text{max}} = \delta$)

Furthermore, let $Y \begin{bmatrix} \varphi_1 & \ldots & \varphi_i \end{bmatrix}$ be (almost) invariant subspace with largest eigenvalue $\mu < 1$ then

$Q$ component in $\left( WW^* A WW^* \right)^k r_0$ behaves as $\mu^k \Omega^k \Sigma$
Convergence

Hermitian:

$$\min_{d_1 \in V_j + C_k} \| b - d_1 \|_2 \leq \min_{d_2 \in (I - P_Q)V_j} \left\| (I - P_Q)r_1 - d_2 \right\|_2 + \frac{\delta}{1 - \delta} \left\| (I - \Pi_V)r_1 \right\|_2$$

Good speed-up for modest $\delta$, say $O(0.1) - O(0.01)$; we mainly see convergence of deflated problem.

If exact deflation effective, we see immediate improvement.

For modest $\delta$, sensitive to perturbation (rapid deterioration as $\delta$ increases).

So, we need to update regularly for good convergence. Deterioration from couple of updates to the matrix will make the old deflation space ineffective.
Convergence - alternatives

Hermitian case: consider $\lambda(WW^* A WW^*)$ directly.

Let $A = \begin{bmatrix} Q Y_1 Y_2 \end{bmatrix} \text{diag}(A_Q, A_{Y_1}, A_{Y_2}) \begin{bmatrix} Q Y_1 Y_2 \end{bmatrix}^*$ and
$$R(C_k) \subseteq R\left(\begin{bmatrix} Q Y_1 \end{bmatrix}\right),$$
$$\varepsilon_1 = \max_\lambda \text{diag}(A_Q, A_{Y_1}),$$
$$\eta = \max \lambda_{Y_1} - \min \lambda_{Y_1}.$$

Then
$$\lambda(WW^* A WW^*) = \{0\} \cup \left[\lambda_{Y_1,\min} - \eta, \lambda_{Y_1,\max} + \eta + \delta^2 \varepsilon_1\right] + \lambda(A_{Y_2})$$

Nonhermitian case: interested in $(I - P_Q)V_j$ where $A_W V_j = V_{j+1}H_j$

Consider a backward error for Krylov decomp.
$$(A + E)\hat{V}_j = \hat{V}_{j+1}\hat{H}_j$$

and analyze convergence for perturbed $A$ (following Stewart’02)
Behavior of $\delta$ over Multiple Systems

<table>
<thead>
<tr>
<th>System</th>
<th>401</th>
<th>402</th>
<th>403</th>
<th>404</th>
<th>405</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Cycle 1</td>
<td>0.9994</td>
<td>0.2380</td>
<td>0.0464</td>
<td>0.0607</td>
<td>0.0794</td>
</tr>
<tr>
<td>After Cycle 1</td>
<td>0.9180</td>
<td>0.1448</td>
<td>0.0463</td>
<td>0.0606</td>
<td>0.0787</td>
</tr>
<tr>
<td>After Cycle 2</td>
<td>0.2446</td>
<td>0.0302</td>
<td>0.0416</td>
<td>0.0568</td>
<td>0.0690</td>
</tr>
<tr>
<td>After Cycle 3</td>
<td>0.2331</td>
<td>0.0302</td>
<td>0.0415</td>
<td>0.0567</td>
<td>0.0684</td>
</tr>
</tbody>
</table>

Note that $\delta$ improves over multiple systems
Convergence with Periodic Update

GCRO-DR(40,20)
Perturbation of Invariant Subspaces

\[ A = \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \text{diag}(\Lambda_1, \Lambda_2, \Lambda_3) \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix}^*, \quad A + E = ??? \]

Perturbation \( E \) is small, but often too large to simply assume that invariant subspace for small eigenvalues, \( \text{Range}(V_1) \), survives.

Especially since we cannot assume \( \lambda_1^{(2)} - \lambda_{k1}^{(1)} \) is (relatively) large

However, from underlying physics/model we may expect that \( E \) should not affect \( V_1 \) much.

For example, small change in crack length (and other small nonlinear deformations) should not change smoothest modes of the (breaking) plate much.

Of course, many small changes make a large change ...
Perturbation of Invariant Subspaces

\[ A = \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \text{diag}(\Lambda_1, \Lambda_2, \Lambda_3) \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix}^* , \quad A + E = ??? \]

Perturbation \( E \) of \( A \) is small, but too large to simply assume that invariant subspace for small eigenvalues, \( \text{Range}(V_1) \), survives.

However, \( E \) concentrated in high frequency modes

\[ \lambda^{(2)}_1 - \lambda^{(1)}_{k_1} \text{ small but } \left\| \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T E \right\|_F = \varepsilon < \gamma_1 \left( \lambda^{(2)}_1 - \lambda^{(1)}_{k_1} \right) \]

\[ \|E\|_F \text{ not small, but } \|V_3^T E\|_F \approx \|E\|_F \text{ and } \|V_3^T E\|_F < \gamma_2 \left( \lambda^{(3)}_1 - \lambda^{(1)}_{k_1} \right) \]

Then \( \tan \vartheta_1 \left( \text{Range}(V_1), \text{Range}(\hat{V}_1) \right) = O(\varepsilon) \) (small)

(Kilmer and de Sturler’06)
Test Case: Tomography (with Misha Kilmer)

- Reconstruct medium by measuring how signals propagate
- Parameterize medium and optimize parameters by matching measured signal with computed signal (at receivers)
- Forward problem \[-\nabla \cdot \left( a(x, \omega; p) \nabla u_j \right) + m(x, \omega; p) u = f_j\]
- Have to solve (forward) problem many times (optimization)
- Problem is Hermitian for zero frequency and nonzero frequency gives imaginary shift
- Multiple sources give multiple right hand sides
- Nonlinear least squares/Gauss-Newton with line search
- First few steps fix background parameters, later steps mainly change shape of tumor: ‘diffusion’ jump in small region
- Change in matrix concentrated in high frequency modes
- Lot of opportunity to exploit structure
Iteration Counts for Recycle Version

![Graph showing iteration counts for recycle version.]
Conclusions and Future Work

- Recycling search spaces is very effective
- Techniques for recycling are fairly cheap
- Accurate recycle space not needed for fast convergence but typically need regular updating to track changes in problem
- Localized changes in problem yield small changes in ‘small eigenvalue’ invariant subspace (application dependent) – useful for computational mechanics/tomography/…
- Tune recycling for particular applications and/or nonlinear iteration. Yields further improvement (tomography)
  - For Broyden type methods very cheap recycling possible
- Best recycle space is open question; nontrivial issues even in the symmetric case (come back next year)
- Software available soon (Trilinos, Sandia, some in PETSc)
Stop! You're going too far
Good Reading

- All available at http://www.math.vt.edu/people/sturler/index.html