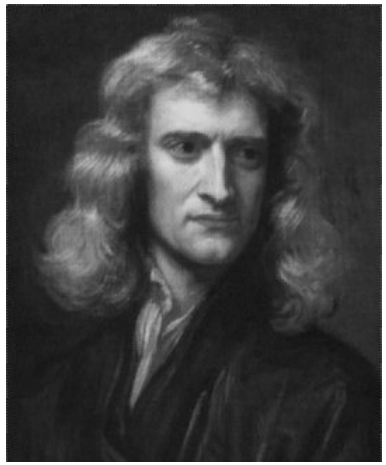
Krylov solvers and definite preconditioning

Jed Brown

VAW, ETH Zürich

2009-01-27

Newton iteration



- Standard form of a nonlinear system

$$F(u) = 0$$

- Iteration

$$\begin{aligned} \text{Solve:} \quad & J(\tilde{u})u = -F(\tilde{u}) \\ \text{Update:} \quad & \tilde{u}_+ \leftarrow \tilde{u} + u \end{aligned}$$

Example (p-Bratu)

Suppose F is a discretization of

$$-\nabla \cdot \big(\eta \nabla u\big) - \lambda e^u - f = 0$$

$$\eta(\gamma) = \Big(\frac{\epsilon^2 + \gamma}{\gamma_0}\Big)^{\frac{p-2}{2}}, \qquad \gamma = \frac{1}{2}\left|\nabla u\right|^2.$$

Then $J(\tilde{u})u$ is a discretization of

$$-\nabla \cdot \big(\eta \nabla u + \eta'(\nabla \tilde{u} \cdot \nabla u)\nabla \tilde{u}\big) - \lambda e^{\tilde{u}} u.$$

Matrices

Definition (Matrix)

A matrix is a linear transformation between finite dimensional vector spaces.

Matrices

Definition (Matrix)

A matrix is a linear transformation between finite dimensional vector spaces.

Definition (Forming a matrix)

Forming or assembling a matrix means defining it's action in terms of entries (usually stored in a sparse format).

Important matrices

1. Sparse (e.g. discretization of a PDE operator)
2. Inverse of *anything* interesting $B = A^{-1}$
3. Jacobian of a nonlinear function $Jy = \lim_{\epsilon \to 0} \frac{F(x+\epsilon y) - F(x)}{\epsilon}$
4. Fourier transform $\mathcal{F}, \mathcal{F}^{-1}$
5. Low rank correction $B = A + uv^T$
6. Schur complement $S = D - CA^{-1}B$
7. Tensor product $A = \sum_e A_x^e \otimes A_y^e \otimes A_z^e$

Important matrices

1. Sparse (e.g. discretization of a PDE operator)
2. Inverse of *anything* interesting $B = A^{-1}$
3. Jacobian of a nonlinear function $Jy = \lim_{\epsilon \to 0} \frac{F(x+\epsilon y) - F(x)}{\epsilon}$
4. Fourier transform $\mathcal{F}, \mathcal{F}^{-1}$
5. Low rank correction $B = A + uv^T$
6. Schur complement $S = D - CA^{-1}B$
7. Tensor product $A = \sum_e A_x^e \otimes A_y^e \otimes A_z^e$

These matrices are dense

Never form them.

Important matrices

1. Sparse (e.g. discretization of a PDE operator)
2. Inverse of *anything* interesting $B = A^{-1}$
3. Jacobian of a nonlinear function $Jy = \lim_{\epsilon \to 0} \frac{F(x+\epsilon y) - F(x)}{\epsilon}$
4. Fourier transform $\mathcal{F}, \mathcal{F}^{-1}$
5. Low rank correction $B = A + uv^T$
6. Schur complement $S = D - CA^{-1}B$
7. Tensor product $A = \sum_e A_x^e \otimes A_y^e \otimes A_z^e$

This matrix is not very sparse

Don't form it.

Important matrices

1. Sparse (e.g. discretization of a PDE operator)
2. Inverse of *anything* interesting $B = A^{-1}$
3. Jacobian of a nonlinear function $Jy = \lim_{\epsilon \to 0} \frac{F(x+\epsilon y) - F(x)}{\epsilon}$
4. Fourier transform $\mathcal{F}, \mathcal{F}^{-1}$
5. Low rank correction $B = A + uv^T$
6. Schur complement $S = D - CA^{-1}B$
7. Tensor product $A = \sum_e A_x^e \otimes A_y^e \otimes A_z^e$

None of these matrices have "entries"

What can we do with a matrix that doesn't have entries?

Krylov solvers for $Ax = b$

- Krylov subspace: $\{b, Ab, A^2b, A^3b, \dots\}$
- Convergence rate depends on the spectral properties of the matrix
  - Existance of small polynomials $p_n(A) < \epsilon$ where $p_n(0) = 1$.
  - condition number $\kappa(A) = \|A\| \left\|A^{-1}\right\| = \sigma_{\text{max}}/\sigma_{\text{min}}$
  - spectrum Λ
- For any popular Krylov method $\mathcal{K}$, there is a matrix of size m, such that $\mathcal{K}$ outperforms all other methods by a factor at least $\mathcal{O}(\sqrt{m})$ [Nachtigal et. al., 1992]

Typically...

- The action $y \leftarrow Ax$ can be computed in $\mathcal{O}(m)$
- Aside from matrix multiply, the n^{th} iteration requires at most $\mathcal{O}(mn)$

GMRES

Brute force minimization of residual in $\{b, Ab, A^2b, \dots\}$

1. Use Arnoldi to orthogonalize the nth subspace, producing

$$AQ_n = Q_{n+1}H_n$$

2. Minimize residual in this space by solving the overdetermined system

$$H_n y_n = e_1^{(n+1)}$$

using QR-decomposition, updated cheaply at each iteration.

Properties

- Converges in n steps for all right hand sides if there exists a polynomial of degree n such that $\|p_n(A)\| < \mathit{tol}$ and $p_n(0) = 1$.
- Residual is monotonically decreasing, robust in practice
- In practice, restarted variants are used

GMRES convergence

Theorem (not sharp)

For the GMRES iteration applied to a diagonalizable matrix $A = V\Lambda V^{-1}$

$$\frac{\|r_n\|}{\|r_0\|} \leq \kappa(V) \inf_{\substack{p_n \in P_n \\ p_n(0)=1}} \|p_n\|_\Lambda$$

where P_n *is the space of polynomials of degree at most* n*.*

- For SPD matrices, the convex hull of the spectrum is the interval $[\sigma_{\text{min}}, \sigma_{\text{max}}]$. The polynomial of degree n which is smallest on this set is the shifted Chebyshev polynomial which attains a maximum value

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n$$

  This implies convergence in $\mathcal{O}(\sqrt{\kappa})$ iterations.
- Second order elliptic operators produce $\kappa \in \Theta(h^{-2})$, required number of iterations scales as h^{-1}.

GMRES convergence

Theorem (not sharp)

For the GMRES iteration applied to a diagonalizable matrix $A = V\Lambda V^{-1}$

$$\frac{\|r_n\|}{\|r_0\|} \leq \kappa(V) \inf_{\substack{p_n \in P_n \\ p_n(0)=1}} \|p_n\|_\Lambda$$

where P_n *is the space of polynomials of degree at most* n*.*

- For SPD matrices, the convex hull of the spectrum is the interval $[\sigma_{\text{min}}, \sigma_{\text{max}}]$. The polynomial of degree n which is smallest on this set is the shifted Chebyshev polynomial which attains a maximum value

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n$$

  This implies convergence in $\mathcal{O}(\sqrt{\kappa})$ iterations.
- Second order elliptic operators produce $\kappa \in \Theta(h^{-2})$, required number of iterations scales as h^{-1}.

Preconditioning

Idea: improve the conditioning of the Krylov operator

- Left preconditioning

$$(P^{-1}A)x = P^{-1}b$$

$$\{P^{-1}b, (P^{-1}A)P^{-1}b, (P^{-1}A)^2P^{-1}b, \dots\}$$

- Right preconditioning

$$(AP^{-1})Px = b$$

$$\{b, (P^{-1}A)b, (P^{-1}A)^2b, \dots\}$$

- The product $P^{-1}A$ or AP^{-1} is *not* formed.

Definition (Preconditioner)

A *preconditioner* $\mathcal{P}$ is a method for constructing a matrix (just a linear function, not assembled!) $P^{-1} = \mathcal{P}(A, A_p)$ using a matrix A and extra information A_p, such that the spectrum of $P^{-1}A$ (or AP^{-1}) is well-behaved.

Preconditioning

Definition (Preconditioner)

A *preconditioner* $\mathcal{P}$ is a method for constructing a matrix $P^{-1} = \mathcal{P}(A, A_p)$ using a matrix A and extra information A_p, such that the spectrum of $P^{-1}A$ (or AP^{-1}) is well-behaved.

- P^{-1} is dense, P is often not available and is not needed
- A is rarely used by $\mathcal{P}$, but $A_p = A$ is common
- A_p is often a sparse matrix, the "preconditioning matrix"
- Matrix-based: Jacobi, Gauss-Seidel, SOR, ILU(k), LU
- Parallel: Block-Jacobi, Schwarz, Multigrid, FETI-DP, BDDC
- Indefinite: Schur-complement

Relaxation

Split into lower, diagonal, upper parts: $A = L + D + U$

Jacobi

$$P^{-1} = D^{-1}$$

- Cheap

Successive over-relaxation (SOR)

$$\left(L + \frac{1}{\omega} D\right) x_{n+1} = \left[\left(\frac{1}{\omega} - 1\right) D - U\right] x_n + \omega b$$

$$P^{-1} = k \text{ iterations starting with } x_0 = 0$$

- Implemented as a sweep
- $\omega = 1$ corresponds to Gauss-Seidel
- Very effective at removing high-frequency components of residual

Factorization

Two phases

- symbolic factorization: determine how much fill occurs, only uses sparsity pattern
- numeric factorization: compute factors

LU decomposition

- Ultimate preconditioner
- Expensive, for $m \times m$ sparse matrix with bandwidth b, traditionally requires $\mathcal{O}(mb^2)$ time and $\mathcal{O}(mb)$ space.
  - Bandwidth scales as $m^{\frac{d-1}{d}}$ in d-dimensions
- Symbolic factorization is problematic in parallel

Incomplete LU

- Allow a limited number of levels of fill: ILU(k)
- Only allow fill for entries that exceed threshold: ILUT
- Very poor scaling in parallel, don't bother beyond 8 PEs.
- No guarantees

Multigrid

Hierarchy: Interpolation and restriction operators

$$\mathcal{I}^{\uparrow} : X_{\text{coarse}} \to X_{\text{fine}} \qquad \mathcal{I}^{\downarrow} : X_{\text{fine}} \to X_{\text{coarse}}$$

- Geometric: define problem on multiple levels, use grid to compute hierarchy
- Algebraic: define problem only on finest level, use matrix structure to build hierarchy

Galerkin approximation

Assemble this matrix: $A_{\text{coarse}} = \mathcal{I}^{\downarrow} A_{\text{fine}} \mathcal{I}^{\uparrow}$

Application of multigrid preconditioner (V-cycle)

- Apply pre-smoother on fine level (any preconditioner)
- Restrict residual to coarse level with $\mathcal{I}^{\downarrow}$
- Solve on coarse level $A_{\text{coarse}} x = r$
- Interpolate result back to fine level with $\mathcal{I}^{\uparrow}$
- Apply post-smoother on fine level (any preconditioner)

Multigrid convergence

Properties

- Textbook: $P^{-1}A$ is spectrally equivalent to identity
- Most theory applies to SPD matrices, often works fine for nonsymmetric matrices such as advection-diffusion
- Good when coefficients in problem are smooth
  - Can fail completely when coefficients have large jumps
- Aggressive coarsening is critical, especially in parallel
- Most theory uses SOR smoothers, ILU can be more robust
- Coarsest component is usually solved semi-redundantly using a direct solver

Domain decomposition

Domain size L, subdomain size H, element size h

Overlapping/Schwarz

- Solve Dirichlet problems on overlapping subdomains
- No overlap: $\mathit{its} \in \mathcal{O}\left(\frac{L}{\sqrt{Hh}}\right)$, Overlap: $\mathit{its} \in \left(\frac{L}{H}\right)$

BDDC and FETI-DP

- Neumann problems on subdomains with coarse grid correction
- $\mathit{its} \in \mathcal{O}\left(1 + \log \frac{H}{h}\right)$