Lecture 16: Iterative solution of a linear system

COMP5930M Scientific Computation

Today

Basic concepts

Standard methods

- Convergence
- Efficiency
- Accuracy

Matrix properties

Iterative methods

- We can design iterative methods to solve a linear system $\mathbf{Ax} = \mathbf{b}$
  - Starting from an initial guess $\mathbf{x}_0$ a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ will be computed by solving linear problems
  - These linear problems should be simpler than the original one (sparser, with special structure etc.)
  - They are an alternative to direct methods, such as Gauss Elimination
- Note: We already use a type of iterative method for the nonlinear problem
  - Newton's method generates a similar sequence of approximations
  - In the nonlinear case no direct solution is possible

Application to the Newton algorithm

- The linear system at each **nonlinear** iteration is specified as $\mathbf{J}\boldsymbol{\delta} = -\mathbf{F}$
- Again we need an initial guess $\boldsymbol{\delta}_0$ for the linear iterative solver
  - We may choose $\boldsymbol{\delta}_0 = \mathbf{0}$
- We must also ensure this linear iterative sequence converges

Iterative methods by splitting the matrix

General iterative method based on splitting: $\mathbf{A}\mathbf{x} = (\mathbf{D} + \mathbf{E})\mathbf{x} = \mathbf{b}$

$$\begin{aligned} \mathbf{D}\mathbf{x}_{k+1} &= \mathbf{b} - \mathbf{E}\mathbf{x}_k \\ \mathbf{x}_{k+1} &= \mathbf{D}^{-1}\mathbf{b} - \mathbf{D}^{-1}\mathbf{E}\mathbf{x}_k \\ \mathbf{x}_{k+1} &= \mathbf{z} + \mathbf{B}\mathbf{x}_k \end{aligned}$$

where $\mathbf{z} = \mathbf{D}^{-1}\mathbf{b}$ and $\mathbf{B} = -\mathbf{D}^{-1}\mathbf{E}$ is called the iteration matrix.

Standard methods: Jacobi iteration

The simplest method is Jacobi iteration, which corresponds to the choice $\mathbf{D} = \mathrm{diag}(\mathbf{A})$ and $\mathbf{E} = \mathbf{A} - \mathbf{D}$ (non-diagonal elements):

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathrm{diag}(\mathbf{A})^{-1}\left(\mathbf{b} - \mathbf{A}\mathbf{x}_k\right)$$

- diag($\mathbf{A}$) is the diagonal of matrix $\mathbf{A}$, easy to invert since all we need to do is take reciprocal d_i^{-1} of each element $i = 1, \ldots, n$

- We usually define the residual $\mathbf{r}_k$ as

$$\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$$

  noting that $\mathbf{r}_k = \mathbf{0}$ implies $\mathbf{x}_k = \mathbf{x}$, and that if the iteration converges, the residuals go to zero, $\lim_{k\to\infty} \mathbf{r}_k \to \mathbf{0}$

Convergence

- The Jacobi iteration converges if **A** is *strictly diagonally dominant*
  - Strictly diagonally dominant requires that for any row i of $\mathbf{A}$, we have

    $$|A_{ii}| \; > \; \Sigma_{j \neq i} |A_{ij}|$$

    which is fairly restrictive

- In practice, it will often converge if **A** is *weakly diagonally dominant*

  $$|A_{ii}| \; \geq \; \Sigma_{j \neq i} |A_{ij}|$$

  but this cannot be proved in general

Efficiency

- Each Jacobi iteration requires a matrix-vector multiplication
  - $\mathcal{O}(n^2)$ expense in general
  - For sparse matrices this reduces to $\mathcal{O}(n)$ expense
- However the overall expense is also determined by the total number of iterations required
  - If we can guarantee a small number of iterations, $K_{\max} \ll n$, we are more efficient than direct algorithms' $\mathcal{O}(n^3)$
- Number of Jacobi iterations is in general LARGE

Standard methods: Gauss-Seidel

A better iterative method is Gauss-Seidel Iteration[1] for $\mathbf{Ax} = \mathbf{b}$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathrm{udiag}(\mathbf{A})^{-1}\left(\mathbf{b} - \mathbf{A}\mathbf{x}_k\right)$$

- udiag($\mathbf{A}$) is the upper-triangular part of the matrix $\mathbf{A}$
- Solution of upper-triangular system can be performed by back-substitution as before
- Computational cost of each Gauss-Seidel iteration is $\mathcal{O}(n^2)$

[1] (Developed by C.F. Gauss around 1820 because he grew tired of performing Gaussian elimination by hand when solving problems of geodetics.)

Standard methods: Gauss-Seidel

A better iterative method is Gauss-Seidel Iteration[1] for $\mathbf{Ax} = \mathbf{b}$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathrm{udiag}(\mathbf{A})^{-1}\left(\mathbf{b} - \mathbf{A}\mathbf{x}_k\right)$$

- udiag($\mathbf{A}$) is the upper-triangular part of the matrix $\mathbf{A}$
- Solution of upper-triangular system can be performed by back-substitution as before
- Computational cost of each Gauss-Seidel iteration is $\mathcal{O}(n^2)$
- Converges for all symmetric, positive definite $\mathbf{A}$, i.e. all eigenvalues $\lambda > 0$, or if $\mathbf{A}$ is strictly diagonally dominant

[1] (Developed by C.F. Gauss around 1820 because he grew tired of performing Gaussian elimination by hand when solving problems of geodetics.)

Standard methods: Gauss-Seidel

A better iterative method is Gauss-Seidel Iteration[1] for $\mathbf{A}\mathbf{x} = \mathbf{b}$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathrm{udiag}(\mathbf{A})^{-1}\left(\mathbf{b} - \mathbf{A}\mathbf{x}_k\right)$$

- udiag($\mathbf{A}$) is the upper-triangular part of the matrix $\mathbf{A}$
- Solution of upper-triangular system can be performed by back-substitution as before
- Computational cost of each Gauss-Seidel iteration is $\mathcal{O}(n^2)$
- Converges for all symmetric, positive definite $\mathbf{A}$, i.e. all eigenvalues $\lambda > 0$, or if $\mathbf{A}$ is strictly diagonally dominant

[1] (Developed by C.F. Gauss around 1820 because he grew tired of performing Gaussian elimination by hand when solving problems of geodetics.)

Residual vs. error

- We can easily compute the residual $\mathbf{r}_k$, but this is not the same as the error $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$
  - Even though $\mathbf{r}_k = \mathbf{0}$ implies $\mathbf{e}_k = \mathbf{0}$
  - (Compare to Lecture 3 convergence discussion on $F(x_k)$ and $|x^* - x_k|$)
- We can show that they are related through

$$\mathbf{A}\mathbf{e}_k \;=\; \mathbf{A}\left(\mathbf{x} - \mathbf{x}_k\right) \;=\; \mathbf{b} - \mathbf{A}\mathbf{x}_k \;=\; \mathbf{r}_k$$

- The precise relationship is determined by properties of $\mathbf{A}$

Eigenvalues and Eigenvectors

Assume our matrix $\mathbf{A}$ is symmetric, $\mathbf{A} = \mathbf{A}^T$, and that we can construct matrix $\mathbf{Q}$, such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} \ = \ \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix

- The diagonal entries of $\mathbf{D}$ are the eigenvalues, λ_i, of $\mathbf{A}$
- The columns of $\mathbf{Q}$ are the eigenvectors, $\mathbf{q}_i$, of $\mathbf{A}$

Non-symmetric $\mathbf{A}$ can be treated similarly but there is no guarantee that a full set of eigenvalues and eigenvectors exists.

Matrix norm

- There are several ways to define the norm of a matrix $\|\mathbf{A}\|$
- For symmetric $\mathbf{A}$, we define the norm as the spectral radius

$$\|\mathbf{A}\| = \max_i |\lambda_i|$$

- Consequently

$$\|\mathbf{A}^{-1}\| = \frac{1}{\min_i |\lambda_i|}$$

- The spectral norm has the expected properties of a norm $\|\mathbf{A}\| = 0$ implies $\mathbf{A} = \mathbf{0}$ and $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$
- It is consistent with the vector norm, $\|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\| \|\mathbf{x}\|_2$

Convergence of iterative methods

Recall the iterative method based on splitting: $\mathbf{Ax} = (\mathbf{D} + \mathbf{E})\mathbf{x} = \mathbf{b}$

$$\mathbf{x}_{k+1} \quad = \quad \mathbf{z} + \mathbf{B}\mathbf{x}_k$$

where $\mathbf{z} = \mathbf{D}^{-1}\mathbf{b}$ and $\mathbf{B} = -\mathbf{D}^{-1}\mathbf{E}$ is called the iteration matrix.

We can show that the error, $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$, satisfies the relation:

$$\mathbf{e}_{k+1} = \mathbf{B}\mathbf{e}_k,$$

so the iteration converges iff the spectral norm $\|\mathbf{B}\| < 1$.

In fact, since $||\mathbf{e}_{k+1}|| \leq ||\mathbf{B}|| \, ||\mathbf{e}_k||$ then $||\mathbf{e}_k|| \leq ||\mathbf{B}||^k \, ||\mathbf{e}_0|| \to 0$.

Condition number κ

- We define the matrix condition number κ as

$$\kappa = \|\mathbf{A}\| \, \|\mathbf{A}^{-1}\|.$$

  If $\mathbf{A}$ is symmetric, it follows that $\kappa = \frac{\max|\lambda_i|}{\min|\lambda_i|}$

- Small condition number implies the error and residual are directly proportional
  - termed *well-conditioned* linear problem

  Large condition number implies there is no direct proportionality
  - termed *ill-conditioned* linear problem

- Condition number for FDM matrix is typically $\mathcal{O}(h^{-2})$

Relationship between error and residual

Since $\mathbf{e}_k = \mathbf{A}^{-1}\mathbf{r}_k$, we can estimate

$$\|\mathbf{e}_k\| = \|\mathbf{A}^{-1}\mathbf{r}_k\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{r}_k\|.$$

We divide both sides by $\|\mathbf{x}_k\|$ and note that

$$\|\mathbf{A}\|^{-1} = \frac{\kappa(\mathbf{A})}{\|\mathbf{A}\|}$$

so that

$$\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}_k\|} \leq \kappa(\mathbf{A})\frac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\|}.$$

Relationship between error and residual

Since $\mathbf{e}_k = \mathbf{A}^{-1}\mathbf{r}_k$, we can estimate

$$\|\mathbf{e}_k\| = \|\mathbf{A}^{-1}\mathbf{r}_k\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{r}_k\|.$$

We divide both sides by $\|\mathbf{x}_k\|$ and note that

$$\|\mathbf{A}\|^{-1} = \frac{\kappa(\mathbf{A})}{\|\mathbf{A}\|}$$

so that

$$\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}_k\|} \leq \kappa(\mathbf{A})\frac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\|}.$$

Relative error $\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}_k\|}$ is bounded from above by the relative residual $\frac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\|}$ and a constant factor equal to the condition number $\kappa(\mathbf{A})$

Relationship between error and residual

Since $\mathbf{e}_k = \mathbf{A}^{-1}\mathbf{r}_k$, we can estimate

$$\|\mathbf{e}_k\| = \|\mathbf{A}^{-1}\mathbf{r}_k\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{r}_k\|.$$

We divide both sides by $\|\mathbf{x}_k\|$ and note that

$$\|\mathbf{A}\|^{-1} = \frac{\kappa(\mathbf{A})}{\|\mathbf{A}\|}$$

so that

$$\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}_k\|} \leq \kappa(\mathbf{A})\frac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\|}.$$

Relative error $\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}_k\|}$ is bounded from above by the relative residual $\frac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\|}$ and a constant factor equal to the condition number $\kappa(\mathbf{A})$

Summary

- Iterative linear solver generates a sequence $\{\mathbf{x}_k\}_{k=0}^{K_{\max}}$ by solving "simpler" linear problems with $\mathcal{O}(n^2)$ cost per iteration

- Iteration drives residual $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$ close to zero

- When matrix **A** is well-conditioned, $\kappa(\mathbf{A}) = \mathcal{O}(1)$, the residual $\mathbf{r}_k$ and the true error $\mathbf{e}_k = \mathbf{x}^* - \mathbf{x}_k$ are proportional to each other and the iteration drives $\mathbf{e}_k$ close to zero

- For the FDM matrix, decreasing grid size h makes the Jacobian increasingly ill-conditioned (PROBLEM)