Convergence of the CG Method

William McLean

Computational Maths Seminar, 23 September 2008

Previously in this seminar ...

- Symmetric positive-definite matrix A.
- Linear system $A\boldsymbol{x} = \boldsymbol{b}$.
- CG iterates $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ...
- residual $\boldsymbol{r}_k = \boldsymbol{b} - A\boldsymbol{x}_k$.
- Krylov subspace $V_k = \operatorname{span}\{\boldsymbol{r}_0, A\boldsymbol{r}_0, A^2\boldsymbol{r}_0, \ldots, A^{k-1}\boldsymbol{r}_0\}$.
- Galerkin equations: $\boldsymbol{x}_k - \boldsymbol{x}_0 \in V_k$ satisfies

$$\langle A\boldsymbol{x}_k, \boldsymbol{w}\rangle = \langle \boldsymbol{b}, \boldsymbol{w}\rangle \quad \text{for all } \boldsymbol{w} \in V_k.$$

- Optimality property $\|\boldsymbol{x}_k - \boldsymbol{x}\|_A = \min_{\boldsymbol{y}\in\boldsymbol{x}_0+V_k} \|\boldsymbol{y} - \boldsymbol{x}\|_A$.
- Finite termination property: $\boldsymbol{x}_k \neq \boldsymbol{x}$ for $0 \le k \le k^*$ and $\boldsymbol{x}_{k^*+1} = \boldsymbol{x}$, for some $k^* < n$.
- Alternative bases:

$$\begin{aligned} V_k &= \operatorname{span}\{\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{k-1}\}, & \langle \boldsymbol{r}_j, \boldsymbol{r}_k\rangle &= 0 \quad \text{if } j \neq k, \\ &= \operatorname{span}\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{k-1}\}, & \langle \boldsymbol{p}_j, \boldsymbol{p}_k\rangle_A &= 0 \quad \text{if } j \neq k. \end{aligned}$$

We want to estimate the *error vector*

$$\boldsymbol{e}_k = \boldsymbol{x}_k - \boldsymbol{x}.$$

Lemma 1. *For each $k \le k^* + 1$ we have*

$$\boldsymbol{e}_k = \boldsymbol{e}_0 + \sum_{j=1}^{k} c_{kj} A^j \boldsymbol{e}_0$$

for some $c_{kj} \in \mathbb{R}$

Proof. Since $A\boldsymbol{e}_0 = A(\boldsymbol{x}_0 - \boldsymbol{x}) = A\boldsymbol{x}_0 - \boldsymbol{b} = -r_0$ we have

$$V_k = \operatorname{span}\{\, A\boldsymbol{e}_0, A^2\boldsymbol{e}_0, \ldots, A^k\boldsymbol{e}_0 \,\},$$

so the result follows from

$$\boldsymbol{e}_k - \boldsymbol{e}_0 = (\boldsymbol{x}_k - \boldsymbol{x}) - (\boldsymbol{x}_0 - \boldsymbol{x}) = \boldsymbol{x}_k - \boldsymbol{x}_0 \in V_k.$$

□

Thus, we may write

$$\boldsymbol{e}_k = P_k(A)\boldsymbol{e}_0 \quad \text{where} \quad P_k(\tau) = 1 + \sum_{j=1}^{k} c_{kj}\tau^j.$$

Lemma 2. *The error in the kth CG iterate satisfies*

$$\|\boldsymbol{e}_k\|_A = \min \|Q(A)\boldsymbol{e}_0\|_A,$$

where the minimum is over all polynomials Q with degree $\le k$ satisfying $Q(0) = 1$.

Proof. Let $Q(\tau) = 1 + \sum_{j=1}^{k} d_j\tau^j$ then

$$\boldsymbol{x} + Q(A)\boldsymbol{e}_0 = \underbrace{\boldsymbol{x} + \boldsymbol{e}_0}_{=\boldsymbol{x}_0} + \sum_{j=1}^{k} d_j A^j \boldsymbol{e}_0 \in \boldsymbol{x}_0 + V_k$$

so the optimality property gives

$$\|\boldsymbol{e}_k\|_A = \|\boldsymbol{x}_k - \boldsymbol{x}\|_A \le \big\|\big(\boldsymbol{x} + Q(A)\boldsymbol{e}_0\big) - \boldsymbol{x}\big\|_A = \|Q(A)\boldsymbol{e}_0\|_A.$$

□

Eigensystem:

$$A\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j, \qquad \langle \boldsymbol{v}_j, \boldsymbol{v}_k\rangle = \delta_{jk}, \qquad 0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n.$$

Lemma 3. *For any $\boldsymbol{w} \in \mathbb{R}^n$ and any real polynomial Q,*

$$\|Q(A)\boldsymbol{w}\|_A \le \max_{1\le j\le n} |Q(\lambda_j)|\|z\|_A.$$

Proof. The eigenvectors form an orthonormal basis for $\mathbb{R}^n$, so

$$\boldsymbol{w} = \sum_{j=1}^{n} \langle \boldsymbol{w}, \boldsymbol{v}_j \rangle \boldsymbol{v}_j$$

and

$$\|\boldsymbol{w}\|_A^2 = \langle \boldsymbol{w}, A\boldsymbol{w} \rangle = \sum_{j=1}^{n} \langle \boldsymbol{w}, \boldsymbol{v}_j \rangle \langle \boldsymbol{v}_j, A\boldsymbol{w} \rangle = \sum_{j=1}^{n} \lambda_j \langle \boldsymbol{w}, \boldsymbol{v}_j \rangle^2.$$

Also,

$$\langle Q(A)\boldsymbol{w}, \boldsymbol{v}_j \rangle = \langle \boldsymbol{w}, Q(A)\boldsymbol{v}_j \rangle = \langle \boldsymbol{w}, Q(\lambda_j)\boldsymbol{v}_j \rangle = Q(\lambda_j)\langle \boldsymbol{w}, \boldsymbol{v}_j \rangle$$

and thus

$$\begin{aligned} \|Q(A)\boldsymbol{w}\|_A^2 &= \sum_{j=1}^{n} \lambda_j \langle Q(A)\boldsymbol{w}, \boldsymbol{v}_j \rangle^2 = \sum_{j=1}^{n} \lambda_j Q(\lambda_j)^2 \langle \boldsymbol{w}, \boldsymbol{v}_j \rangle^2 \\ &\le \Bigl(\max_{1\le j\le n} Q(\lambda_j) \Bigr)^2 \sum_{j=1}^{n} \lambda_j \langle \boldsymbol{w}, \boldsymbol{v}_j \rangle^2 = \Bigl(\max_{1\le j\le n} Q(\lambda_j) \|\boldsymbol{w}\|_A \Bigr)^2. \end{aligned}$$

□

Problem: find Q of degree k such that $Q(0) = 1$ and $|Q(\tau)|$ is small for $\lambda_1 \le \tau \le \lambda_n$.

The *Chebyshev polynomial* T_k is defined by

$$T_k(\cos\theta) = \cos k\theta,$$

or equivalently, since $\cos(i\theta) = \cosh\theta$,

$$T_k(\cosh\theta) = \cosh k\theta.$$

It follows that

$$T_k(s) = \tfrac{1}{2}\bigl[(s + \sqrt{s^2-1})^k + (s - \sqrt{s^2-1})^k\bigr] \quad \text{for } |s| > 1.$$

Put

$$s = \frac{2\tau - (\lambda_1 + \lambda_n)}{\lambda_n - \lambda_1} \quad \text{and} \quad s_0 = -\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} < -1,$$

so that $-1 \le s \le 1$ for $\lambda_1 \le \tau \le \lambda_n$, and $s = s_0$ when $\tau = 0$. Choose

$$Q(\tau) = \frac{T_k(s)}{T_k(s_0)}$$

then Q is a polynomial of degree k satisfying $Q(0) = 1$ and

$$\max_{1\le j\le n} |Q(\lambda_j)| \le \max_{\lambda_1\le\tau\le\lambda_n} |Q(\tau)| = \max_{-1\le s\le 1} \left| \frac{T_k(s)}{T_k(s_0)} \right| = \frac{1}{|T_k(s_0)|}.$$

Lemma 4.

$$T_k(s_0) = \frac{(-1)^k}{2}\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right] \quad \textit{where} \quad \kappa = \frac{\lambda_n}{\lambda_1} \geq 1.$$

Proof. Since $s_0 = -(\kappa+1)/(\kappa-1)$,

$$s_0^2 - 1 = \frac{(\kappa+1)^2 - (\kappa-1)^2}{(\kappa-1)^2} = \frac{(2\kappa)(2)}{(\kappa-1)^2} = \left(\frac{2\sqrt{\kappa}}{\kappa-1}\right)^2$$

and so

$$s_0 + \sqrt{s_0^2-1} = -\frac{\kappa+1-2\sqrt{\kappa}}{\kappa-1} = -\frac{(\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)(\sqrt{\kappa}-1)} = -\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1},$$

whereas

$$s_0 - \sqrt{s_0^2-1} = \frac{1}{s_0+\sqrt{s_0^2-1}} = -\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}.$$

□

We call κ the *spectral condition number* of A.

Theorem 5. *The error in the kth CG iterate satisfies*

$$\|\boldsymbol{e}_k\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|\boldsymbol{e}_0\|_A.$$

Proof. Follows from the preceding lemmas, because

$$|T_k(s_0)| \geq \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k.$$

□

If κ is large then

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = \frac{1-\kappa^{-1/2}}{1+\kappa^{-1/2}} = 1 - 2\kappa^{-1/2} + O(\kappa^{-1}),$$

which means

$$\log\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \approx -2\kappa^{-1/2}.$$

Thus, we can achieve

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k < \epsilon$$

by taking

$$k \geq \tfrac{1}{2}\sqrt{\kappa}\log\frac{1}{\epsilon}.$$

Unfortunately, if A is a finite element stiffness matrix arising from a nodal basis using a quasi-uniform partition with element size h, then $\kappa = O(h^{-2}) \to \infty$ if we refine the mesh. Thus, putting $\epsilon = h^2$,

$$\text{expected number of CG iterations} = O\bigl(h^{-1}\log h\bigr) = O\bigl(n^{1/d}\log n\bigr),$$

which grows with n.

Let $B \in \mathbb{R}^{n\times n}$ be a symmetric positive-definite matrix, and put

$$\tilde{A} = B^{-1}A, \qquad \tilde{\boldsymbol{b}} = B^{-1}\boldsymbol{b}, \qquad \langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T B\boldsymbol{y},$$

so that $A\boldsymbol{x} = \boldsymbol{b}$ iff

$$\tilde{A}\boldsymbol{x} = \tilde{\boldsymbol{b}}.$$

Notice that $\tilde{A}$ is symmetric and positive-definite with respect to the new inner product,

$$\begin{aligned}\langle \tilde{A}\boldsymbol{x}, \boldsymbol{y}\rangle &= (B^{-1}A\boldsymbol{x})^T B\boldsymbol{y} = (A\boldsymbol{x})^T B^{-1}B\boldsymbol{y} = \boldsymbol{x}^T A\boldsymbol{y}\\ &= \boldsymbol{x}^T BB^{-1}A\boldsymbol{y} = \boldsymbol{x}^T B\tilde{A}\boldsymbol{y} = \langle \boldsymbol{x}, \tilde{A}\boldsymbol{y}\rangle,\end{aligned}$$

with

$$\langle \tilde{A}\boldsymbol{x}, \boldsymbol{x}\rangle = \boldsymbol{x}^T A\boldsymbol{x} > 0 \quad \text{for } \boldsymbol{x} \neq \boldsymbol{0}.$$

We call B a *preconditioner*, and use $\tilde{A}$ and $\langle\cdot,\cdot\rangle$ to construct the the *preconditioned* CG iterates $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ... These iterate satisfy the same error bound as before but with κ replaced by

$$\tilde{\kappa} = \frac{\lambda_{\max}(\tilde{A})}{\lambda_{\min}(\tilde{A})}.$$

Thus, to speed up the convergence we seek a (symmetric positive-definite) B such that

1. $\tilde{\kappa} \ll \kappa$
2. we can cheaply compute the action of B^{-1} on a vector.

Notice $\tilde{\kappa} = 1$ if $B = A$, whereas $\tilde{\kappa} = \kappa$ if $B = I$. In practice, we seek a B that is *spectrally equivalent* to A:

$$c\boldsymbol{x}^T B\boldsymbol{x} \leq \boldsymbol{x}^T A\boldsymbol{x} \leq C\boldsymbol{x}^T B\boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n,$$

$\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$
Solve $B\boldsymbol{s}_k = \boldsymbol{r}_k$ to find $\boldsymbol{s}_0$
$\boldsymbol{p}_0 = \boldsymbol{s}_0$
$\alpha_0 = \|\boldsymbol{s}_0\|^2/\|\boldsymbol{p}_0\|^2_{\tilde{A}} = \boldsymbol{s}_0^T\boldsymbol{r}_0/(\boldsymbol{p}_0^T A\boldsymbol{p}_0)$
$\boldsymbol{x}_1 = \boldsymbol{x}_0 + \alpha_0\boldsymbol{p}_0$
do $k = 1, n-1$
    $\boldsymbol{r}_k = \boldsymbol{b} - A\boldsymbol{x}_k = \boldsymbol{r}_{k-1} - \alpha_{k-1}A\boldsymbol{p}_{k-1}$
    if ($\|\boldsymbol{r}_k\| <$ tol) then
        exit
    end if
    Solve $B\boldsymbol{s}_k = \boldsymbol{r}_k$ to find $\boldsymbol{s}_k$
    $\beta_k = \|\boldsymbol{r}_k\|^2/\|\boldsymbol{r}_{k-1}\|^2 = \boldsymbol{s}_k^T\boldsymbol{r}_k/\boldsymbol{s}_{k-1}^T\boldsymbol{r}_{k-1}$
    $\boldsymbol{p}_k = \boldsymbol{r}_k + \beta_k\boldsymbol{p}_{k-1}$
    $\alpha_k = \|\boldsymbol{r}_k\|^2/\|\boldsymbol{p}_k\|^2_{\tilde{A}} = \boldsymbol{s}_k^T\boldsymbol{r}_k/(\boldsymbol{p}_k^T A\boldsymbol{p}_k)$
    $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k$
end do

Figure 1: Preconditioned conjugate gradient method

with $C > c > 0$ independent of h. Then $\tilde{\kappa}$ is bounded independently of h.

Let

$$\boldsymbol{r}_k = \boldsymbol{b} - A\boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{s}_k = \tilde{\boldsymbol{b}} - \tilde{A}\boldsymbol{x}_k = B^{-1}\boldsymbol{r}_k.$$

Since

$$\|\boldsymbol{s}_k\|^2 = \boldsymbol{s}_k^T B\boldsymbol{s}_k = \boldsymbol{s}_k^T\boldsymbol{r}_k,$$

we see that the *preconditioned conjugate gradient method* takes the form shown in Figure 1.