

Mathematical Tripos Part II: Michaelmas Term 2021

Numerical Analysis – Lecture 20

Convergence of CG The following theorem gives an important characterization of the CG method.

Theorem 4.33 *Let A be symmetric positive definite. After k iterations of the conjugate gradient method, the error $e^{(k)} = x^* - x^{(k)}$ satisfies*

$$\|e^{(k)}\|_A = \min_{P_k} \|P_k(A)e^{(0)}\|_A$$

where the minimization is over all polynomials P_k of degree $\leq k$ that satisfy $P_k(0) = 1$.

Proof. We know from Lecture 18 (Equation (4.7) applied recursively) that $e^{(k)}$ is obtained from $e^{(0)}$ by projecting out the components $d^{(0)}, \dots, d^{(k-1)}$ (in the inner product $\langle \cdot, \cdot \rangle_A$). This means that

$$\|e^{(k)}\|_A = \min_v \|e^{(0)} - v\|_A$$

where the minimization is over all $v \in \text{span}(d^{(0)}, \dots, d^{(k-1)})$. For the conjugate gradient method, this subspace is the same as $\text{span}(r^{(0)}, \dots, A^{k-1}r^{(0)})$, and since $r^{(0)} = Ae^{(0)}$, this means that any such v can be written as $v = \sum_{i=1}^k c_i A^i e^{(0)}$, i.e., $e^{(0)} - v = P_k(A)e^{(0)}$ with $P_k(t) = 1 - \sum_{i=1}^k c_i t^i$ is a degree k polynomial with $P_k(0) = 1$. $\square$

Remark 4.34 *If A has s distinct eigenvalues $\lambda_1, \dots, \lambda_s > 0$, then with $P_s(t) = \prod_{i=1}^s (1 - t/\lambda_i)$ we have $\deg P_s = s$, $P_s(0) = 1$, and $P_s(A) = 0$. Thus this shows that the CG method terminates after s iterations, recovering the result of Theorem 4.29.*

Corollary 4.35 *Let A be symmetric positive definite, and assume that all its eigenvalues lie in $[l, L]$ where $0 < l < L$. Then after k iterations of the conjugate gradient method, the error $e^{(k)} = x^* - x^{(k)}$ satisfies*

$$\|e^{(k)}\|_A \leq 2\rho^k \|e^{(0)}\|_A \leq 2(1 - \sqrt{l/L})^k \|e^{(0)}\|_A, \quad \rho = \frac{\sqrt{L} - \sqrt{l}}{\sqrt{L} + \sqrt{l}} < 1.$$

Proof. First note that for any polynomial P_k we have

$$\|P_k(A)e^{(0)}\|_A \leq \left(\max_{\lambda \in \text{spec}(A)} |P_k(\lambda)| \right) \|e^{(0)}\|_A$$

where $\text{spec}(A)$ is the set of eigenvalues of A (its spectrum). To see why, let $w_1, \dots, w_n$ be an orthogonal basis of eigenvectors of A such that $e^{(0)} = \sum_i w_i$. Since the w_i are eigenvectors of A , they are also pairwise orthogonal wrt A -inner product, and so $\|e^{(0)}\|_A^2 = \sum_i \|w_i\|_A^2$. In addition $P_k(A)e^{(0)} = \sum_i P_k(\lambda_i)w_i$ and so

$$\begin{aligned} \|P_k(A)e^{(0)}\|_A^2 &= \left\| \sum_i P_k(\lambda_i)w_i \right\|_A^2 = \sum_i |P_k(\lambda_i)|^2 \|w_i\|_A^2 \\ &\leq \left(\max_{\lambda \in \text{spec}(A)} |P_k(\lambda)|^2 \right) \|e^{(0)}\|_A^2 \end{aligned}$$

as desired.

We know that the eigenvalues of A are all in $[l, L]$, so we consider the problem of finding the polynomial P_k of degree k , such that $P_k(0) = 1$, and that minimizes the value

$$\max_{x \in [l, L]} |P_k(x)|.$$

This problem has a classical solution $P_k = T_k^*$, where T_k^* is the Chebyshev polynomial on the interval $[l, L]$, which is obtained by dilation and translation of the standard Chebyshev polynomial T_k given on the interval $[-1, 1]$, namely

$$P_k(x) = T_k \left(2 \frac{L-x}{L-l} - 1 \right) / T_k \left(\frac{L+l}{L-l} \right).$$

This polynomial satisfies $P_k(0) = 1$, and since $|T_k(t)| \leq 1$ for all $t \in [-1, 1]$, we have

$$|P_k(x)| \leq \left| T_k \left(\frac{L+l}{L-l} \right) \right|^{-1} = |T_k(1+2\eta)|^{-1}, \quad \eta = l/(L-l),$$

for all $x \in [l, L]$. Using the inequality

$$T_k(t) \geq \frac{1}{2} \left(t + \sqrt{t^2 - 1} \right)^k$$

valid for all $|t| \geq 1$, one can show that $|T_k(1+2\eta)| \leq \left(\frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}} \right)^k$, which gives the desired result. $\square$

For a symmetric positive definite matrix A , let $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} > 1$ be its *condition number*. We saw that the convergence rate of the steepest descent method is $\approx (1 - \frac{1}{\kappa(A)})^k$, whereas the CG method achieves the better rate of $\left(1 - \frac{1}{\sqrt{\kappa(A)}} \right)^k$.

Remark 4.36 The condition number defined above can be written as $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$ where $\|\cdot\|_2$ is the operator norm of A . This quantity measures the sensitivity of the matrix inverse operation, in a relative error sense. Let $\phi(A) = A^{-1}$ be the matrix inverse operation, and consider a perturbation $\tilde{A} = A + H$. The relative sensitivity is defined as:

$$\frac{\|\phi(\tilde{A}) - \phi(A)\|_2 / \|\phi(A)\|_2}{\|\tilde{A} - A\|_2 / \|A\|_2} = \frac{\text{output relative error}}{\text{input relative error}}.$$

One can show that for H small, this quantity is bounded above by $\kappa(A)$.

Preconditioning In $Ax = b$, we change variables, $x = P^T \hat{x}$, where P is a nonsingular $n \times n$ matrix, and multiply both sides with P . Thus, instead of $Ax = b$, we are solving the linear system

$$PAP^T \hat{x} = Pb \Leftrightarrow \hat{A} \hat{x} = \hat{b}. \quad (4.11)$$

Note that symmetry and positive definiteness of A imply that $\hat{A} = PAP^T$ is also symmetric and positive definite since $\langle \hat{A}y, y \rangle = \langle PAP^T y, y \rangle = \langle AP^T y, P^T y \rangle > 0$. Therefore, we can apply conjugate gradients to the new system. This results in the solution $\hat{x}$, hence $x = P^T \hat{x}$. This procedure is called the *preconditioned conjugate gradient method* and the matrix P is called the *preconditioner*.

The main idea of preconditioning is to pick P in (4.11) so that $\kappa(\hat{A})$ is much smaller than $\kappa(A)$, thus accelerating convergence. Ideally, one would like to choose P so that $PAP^T = I$, however this amounts to inverting A ! Instead, we look for an approximation S of A that is easy to invert, or Cholesky-factorize. If we let $S = LL^T$ this Cholesky factorization, and take $P = L^{-1}$, then $PAP^T = L^{-1}AL^{-T}$ which is similar to $S^{-1}A \approx I$. Possible choices of S include:

Example 4.37 1) The simplest choice of S is $D = \text{diag } A$, then $P = D^{-1/2}$ in (4.11).

2) Another possibility is to choose S as a band matrix with small bandwidth. For example, solving the Poisson equation with the five-point formula, we may take S to be the tridiagonal part of A .

Example 4.38 Consider the tridiagonal system $Ax = b$, and let S be defined by:

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \quad S = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} = LL^T, \quad \text{with } L = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}.$$

The matrix S coincides with A except at the $(1, 1)$ -entry. This matrix S happens to have a simple Cholesky factorization $S = LL^T$. Using $P = L^{-1}$, we note that PAP^T has only two distinct eigenvalues, and so the CG method converges in two iterations. To see why, note that $A = S + e_1 e_1^T$, so that

$$S^{-1}A = I + u e_1^T,$$

a rank-1 perturbation of the identity matrix, with all eigenvalues but one equal 1 (the remaining one equal $1 + u_1$).

Matlab demo: Download the Matlab GUI for *Preconditioning of Conjugate Gradient* from <https://www.damtp.cam.ac.uk/user/hf323/M21-II-NA/demos/precond/precond.html>. Run the GUI to solve different systems of linear equations, trying different preconditioners P . You can select from some preset preconditioners but can propose your own customised preconditioners as well. What does preconditioning do to the spectrum of the system matrix?