Prof. A. C. Hansen

Mathematical Tripos Part II: Michaelmas Term 2024

Numerical Analysis – Lecture 20

Technique 4.33 (Preconditioning) In $A\boldsymbol{x} = \boldsymbol{b}$, we change variables, $\boldsymbol{x} = P^T\widehat{\boldsymbol{x}}$, where P is a nonsingular $n \times n$ matrix, and multiply both sides with P. Thus, instead of $A\boldsymbol{x} = \boldsymbol{b}$, we are solving the linear system

$$PAP^T\widehat{\boldsymbol{x}} = P\boldsymbol{b} \quad \Leftrightarrow \quad \widehat{A}\widehat{x} = \widehat{\boldsymbol{b}}\,. \tag{4.11}$$

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

The *condition number* of a matrix A is the value $\kappa(A) := \|A\| \cdot \|A^{-1}\|$, so for a symmetric positive definite matrix A it is the ratio between its largest and smallest eigenvalues,

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1\,.$$

The closer is this number to 1, the faster is convergence of CGM. More precisely, for the rate of convergnce of CGM, we have the uppper estimate

$$\|\boldsymbol{e}^{(k)}\|_A \leq 2\rho^k\, \|\boldsymbol{e}^{(0)}\|_A\,, \qquad \rho = \rho_A = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} < 1\,. \tag{4.12}$$

The main idea of preconditioning is to pick P in (4.11) so that $\kappa(\widehat{A})$ is much smaller than $\kappa(A)$, thus accelerating convergence.

To this end, we note that the similarity transform $B \to C^{-1}BC$ preserves spectrum, hence

$$\kappa(\widehat{A}) = \kappa(PAP^T) = \kappa(P^{-1}[PAP^T]P) = \kappa(AP^TP)\,,$$

and if we set

$$S^{-1} := P^TP =: (QQ^T)^{-1},$$

then it is suggestive to choose S as an approximation to A which is easy to Cholesky-factorize, i.e., $S = QQ^T$ (or already in this form), and then take $P = Q^{-1}$. Then $AP^TP = AS^{-1}$ is close to identity, hence

$$\kappa(\widehat{A}) = \kappa(AP^TP) \approx \kappa(I) = 1 \quad \Rightarrow \quad \kappa(\widehat{A}) \ll \kappa(A)\,,$$

and the preconditioned system (4.11) will be solved much faster because of (4.12).

Each step in the CGM for solving $A\boldsymbol{x} = \boldsymbol{b}$ requires one matrix-vector product $A\boldsymbol{y}$, so with $P = Q^{-1}$, additional expense in each step of the CGM for the preconditioned system (4.11) while computing $\widehat{A}\boldsymbol{y} = PAP^T\boldsymbol{y}$ is two additional computations

$$\boldsymbol{u} = P^T\boldsymbol{y} = Q^{-T}\boldsymbol{y}, \qquad \boldsymbol{v} = P\boldsymbol{z} = Q^{-1}\boldsymbol{z},$$

for some $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$, but note that computing $Q^{-1}\boldsymbol{z}$ is the same as solving the linear system $Q\boldsymbol{v} = \boldsymbol{z}$, which is cheap (via forward substitution) as Q is a lower triangular matrix.

Example 4.34 1) The simplest choice of S is $D = \mathrm{diag}\, A$, 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.

3) One can also take $P = L^{-1}$, where L is the lower triangular part of A (maybe imposing some changes). For example, for the Poisson equation, with $m = 20$ hence dealing with 400×400 system, we take P^{-1} as the lower triangular part of A, but change the diagonal elements from 4 to $\frac{5}{2}$. Then we get a computer precision after just 30 iterations.

Example 4.35 For the tridiagonal system $A\boldsymbol{x} = \boldsymbol{b}$ below, we choose the preconditioner as follows.

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \qquad S = QQ^T = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

The matrix S coincides with A except at the $(1,1)$-entry. The matrix $\widehat{A} = Q^{-1}AQ^{-T}$ for the preconditioned CGM has just two distinct eigenvalues, and we recover the exact solution just in two steps. To see the latter, note that $\widehat{A}$ is similar to $Q^{-T}Q^{-1}A = S^{-1}A$, hence it has the same spectrum. Since $A = S + \boldsymbol{e}_1\boldsymbol{e}_1^T$, we have $S^{-1}A = I + \boldsymbol{u}\boldsymbol{e}_1^T$, a rank-1 perturbation of the identity matrix, with all eigenvalues but one equal 1 (the remaining one equal $1 + u_1$).

Remark 4.36 (Rate of convergence of CGM) Here, we prove (4.12). As we have seen, every direction $\boldsymbol{d}^{(i)}$ in CGM is a linear combination of the vectors $(A^s\boldsymbol{r}^{(0)})_{s=0}^i$, therefore, any vector of the form $\widehat{\boldsymbol{x}}^{(k)} = \boldsymbol{x}^{(0)} + \sum_{i=0}^{k-1} a_i\boldsymbol{d}^{(i)}$ can be represented as

$$\widehat{\boldsymbol{x}}^{(k)} = \boldsymbol{x}^{(0)} + \textstyle\sum_{i=0}^{k-1} c_iA^i\boldsymbol{r}^{(0)} . \tag{4.13}$$

Approximation of this kind also arises from various iterative methods of the form

$$\widehat{\boldsymbol{x}}^{(k+1)} = \widehat{\boldsymbol{x}}^{(k)} - \tau_k(A\widehat{\boldsymbol{x}}^{(k)} - \boldsymbol{b}) ,$$

in particular for the steepest descent method.

Subtracting both parts of (4.13) from the exact solution $\boldsymbol{x}^*$ we obtain $\widehat{\boldsymbol{e}}^{(k)} = \boldsymbol{e}^{(0)} - \sum_{i=0}^{k-1} c_iA^i\boldsymbol{r}^{(0)}$, and since $\boldsymbol{r}^{(0)} = A\boldsymbol{e}^{(0)}$, we can express the error $\widehat{\boldsymbol{e}}^{(k)} = \boldsymbol{x}^* - \widehat{\boldsymbol{x}}^{(k)}$ as

$$\widehat{\boldsymbol{e}}^{(k)} = (I - \textstyle\sum_{i=1}^{k} c_iA^i)\,\boldsymbol{e}^{(0)} = P_k(A)\,\boldsymbol{e}^{(0)}, \tag{4.14}$$

where P_k is a polynomial of degree $\le k$, which satisfies $P_k(0) = 1$.

Now we make use of the following.

Theorem 4.37 (Non-examinable) *Given $A \in \mathbb{R}^{n\times n}$, $A > 0$, let $\{\boldsymbol{d}^{(k)}\}_{k=0}^{m-1}$ be a set of the conjugate directions, i.e., $(A\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(i)}) = 0$ for $i < k$, and consider*

$$F(\boldsymbol{x}^{(k)}) := \|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|_A^2 = \|\boldsymbol{e}^{(k)}\|_A^2 .$$

Then the value of $F(\boldsymbol{x}^{(m+1)})$ obtained through the CGM coincides with the minimum of $F(\boldsymbol{y})$ taken over all $\boldsymbol{y} = \boldsymbol{x}^{(0)} + \sum_{k=0}^{m} c_k\boldsymbol{d}^{(k)}$ simultaneously, namely

$$\arg\min_{c_0,\ldots,c_m} F(\boldsymbol{y}) = \boldsymbol{x}^{(m+1)} = \boldsymbol{x}^{(0)} + \sum_{k=0}^{m} \alpha_k\boldsymbol{d}^{(k)} .$$

Hence, at the k-th stage, the CGM produces the vector $\boldsymbol{x}^{(k)}$ that minimizes the functional

$$F(\widehat{\boldsymbol{x}}^{(k)}) = \|\widehat{\boldsymbol{e}}^{(k)}\|_A^2 = (A\widehat{\boldsymbol{e}}^{(k)}, \widehat{\boldsymbol{e}}^{(k)})$$

over all vectors $\widehat{\boldsymbol{x}}^{(k)}$ of the form $\widehat{\boldsymbol{x}}^{(k)} = \boldsymbol{x}^{(0)} + \sum_{i=0}^{k-1} a_i\boldsymbol{d}^{(i)}$, hence over all $\widehat{\boldsymbol{e}}^{(k)}$ of the form (4.14). Expressing $\boldsymbol{e}^{(0)}$ as $\boldsymbol{e}^{(0)} = \sum \gamma_i\boldsymbol{w}_i$, where $(\boldsymbol{w}_i)$ are orthonormal eigenvectors of A, we find from (4.14) that $\widehat{\boldsymbol{e}}^{(k)} = \sum_i \gamma_iP_k(\lambda_i)\boldsymbol{w}_i$, and $A\widehat{\boldsymbol{e}}^{(k)} = \sum_i \gamma_iP_k(\lambda_i)\lambda_i\boldsymbol{w}_i$, and respectively

$$\|\widehat{\boldsymbol{e}}^{(k)}\|_A^2 = \textstyle\sum_i [P_k(\lambda_i)]^2\lambda_i\gamma_i^2 \le \displaystyle\max_{\lambda\in\sigma(A)} [P_k(\lambda)]^2\, \|\boldsymbol{e}^{(0)}\|_A^2 .$$

Hence, because of the minimization property of CGM,

$$\|\boldsymbol{e}^{(k)}\|_A = \min_{P_k} \|\widehat{\boldsymbol{e}}^{(k)}\|_A \le \min_{P_k} \max_{\lambda\in\sigma(A)} |P_k(\lambda)|\, \|\boldsymbol{e}^{(0)}\|_A .$$

Now, assume that, for the spectrum $\sigma(A)$, we know the largest and the smallest eigenvalues, or some lower and upper bounds, say, $0 < m \le \lambda \le M$. Then the following minimization problem, on the class of polynomials of degree k, arises:

$$P_k(0) = 1, \qquad \max_{x \in [m,M]} |P_k(x)| \to \min \,.$$

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

$$T_k(x) = \cos k\theta, \qquad x = \cos\theta, \qquad \theta \in [0, \pi] \,.$$

One can show that $|T_k^*(x)| \le 2\rho^k$ on the interval $[m, M]$, hence the rate of convergence of CGM admits the following estimate:

$$\|\boldsymbol{e}^{(k)}\|_A \le 2\rho^k \|\boldsymbol{e}^{(0)}\|_A, \qquad \rho = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} < 1, \qquad \sigma(A) \in [m, M] \,.$$