

## 5. EFFICIENT SOLUTION OF THE LINEAR SYSTEMS ARISING FROM FINITE ELEMENT DISCRETIZATION

**5.1. Optimization methods.** We have shown that the finite element discretization of Poisson's equation leads to the solution of a linear system  $Ax = b$ , in which  $A$  is a symmetric matrix. It is also easy to check that  $A$  is a positive definite matrix, i.e.,  $x^T Ax > 0$  for  $x \neq 0$ . For such a problem, the solution of this system is also the minimizer of the functional  $\phi(x) = \frac{1}{2}x^T Ax - x^T b$ . Note the minimum will occur where  $\nabla\phi(x) = 0$ . But  $\nabla\phi(x) = Ax - b$ , so the solution of the minimization problem is the solution of the linear system of equations.

A typical minimization algorithm is to let  $\{p^k\}_{k \geq 0}$  be a set of search directions and  $\{\alpha_k\}_{k \geq 0}$  a set of scalars and define an iteration

$$x^{k+1} = x^k + \alpha_k p^k.$$

The simplest example is the method of steepest descent, in which we choose

$$p^k = -\nabla\phi(x^k) = -[Ax^k - b].$$

To determine the best choice of  $\alpha_k$ , we then minimize  $\phi(x^k + \alpha_k p^k)$  with respect to  $\alpha_k$ , considering  $x^k$  and  $p^k$  now fixed. Since

$$\phi(x^k + \alpha_k p^k) = \frac{1}{2} [(x^k)^T Ax^k + 2\alpha_k (p^k)^T Ax^k + \alpha_k^2 (p^k)^T Ap^k] - x^T b - \alpha_k p^T b,$$

minimizing with respect to  $\alpha_k$  gives:

$$(p^k)^T Ax^k + \alpha_k (p^k)^T Ap^k - (p^k)^T b = 0,$$

i.e.,

$$\alpha_k = \frac{(p^k)^T (b - Ax^k)}{(p^k)^T Ap^k} = \frac{(p^k)^T p^k}{(p^k)^T Ap^k}.$$

If we make the simpler choice,  $\alpha_k = \alpha$  for all  $k$ , then we get the iteration

$$x^{k+1} = x^k - \alpha[Ax^k - b] = [I - \alpha A]x^k + \alpha b.$$

If we let  $x$  denote the exact solution of  $Ax = b$ , then we get the error equation

$$x - x^{k+1} = x - [I - \alpha A]x^k - \alpha b = [I - \alpha A](x - x^k) + \alpha Ax - \alpha b = [I - \alpha A](x - x^k).$$

Iterating this equation, we find that

$$x - x^k = [I - \alpha A]^k (x - x^0).$$

A well known result from linear algebra says that this iteration will converge for all  $x^0 \in \mathbb{R}^n$  if and only if  $\rho(I - \alpha A) < 1$ , where if  $M$  is an  $n \times n$  matrix with eigenvalues  $\mu_i$ , then  $\rho(M) = \max_i |\mu_i|$ . Now if  $\lambda$  is an eigenvalue of  $A$ , then  $1 - \alpha\lambda$  is an eigenvalue of  $I - \alpha A$  (with the same eigenvector). Hence, for convergence, we need  $-1 < 1 - \alpha\lambda < 1$  for all eigenvalues  $\lambda$  of the matrix  $A$ . Since  $A$  is positive definite, all its eigenvalues are positive, so we require  $0 < \alpha < 2/\lambda$ , i.e.,  $0 < \alpha < 2/\rho(A)$ .

To determine the optimal choice of the parameter  $\alpha$ , we minimize the norm of the iteration matrix  $I - \alpha A$ . If we consider  $\|I - \alpha A\|_2$ , then since  $A$  is assumed symmetric, so is  $I - \alpha A$ .

Hence,

$$\|I - \alpha A\|_2 = \rho(I - \alpha A) = \max_i |1 - \alpha \lambda_i|,$$

where  $\lambda_i$  are the eigenvalues of  $A$ . Since  $A$  is positive definite, we have that  $0 < \lambda_1 \leq \dots \leq \lambda_n$ . Then  $\max_i |1 - \alpha \lambda_i| = \max\{|1 - \alpha \lambda_1|, |1 - \alpha \lambda_n|\}$  and this maximum will occur where these two quantities are equal, i.e.,  $1 - \alpha \lambda_1 = \alpha \lambda_n - 1$ . Hence, the optimal value is  $\alpha = 2/(\lambda_1 + \lambda_n)$ . In this case,

$$\rho(I - \alpha A) = 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_n} = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{(\lambda_n/\lambda_1) - 1}{(\lambda_n/\lambda_1) + 1}.$$

Let  $\kappa = \|A\|_2 \|A^{-1}\|_2$  be the condition number measured in the  $\|\cdot\|_2$  norm. Since  $A$  is symmetric and positive definite,  $\|A\|_2 = \rho(A) = \lambda_n$ . Since the eigenvalues of  $A^{-1}$  are the reciprocals of the eigenvalues of  $A$ ,  $\|A^{-1}\|_2 = \rho(A^{-1}) = 1/\lambda_1$ . Hence,  $\kappa = \lambda_n/\lambda_1$ . Thus,  $\rho(I - \alpha A) = (\kappa - 1)/(\kappa + 1)$ , and we have proved the following result.

**Theorem 8.** *If  $A$  is symmetric and positive definite, then the iteration scheme defined by  $x^{k+1} = [I - \alpha A]x^k + \alpha b$ , with  $\alpha = 2/(\lambda_1 + \lambda_n)$  satisfies:*

$$\|x - x^k\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x - x^0\|_2, \quad \kappa = \lambda_{\max}(A)/\lambda_{\min}(A).$$

For the solution of Poisson's problem by standard finite elements, we can show that there is a constant independent of  $h$  such that  $\kappa(A) \leq c^2 h^{-2}$ . Thus, implementing this iteration in its present form leads to a small reduction in error ( $1 - O(h^2)$ ) and slow convergence.

To get a more precise understanding of what the method is doing, we consider an eigenfunction expansion of the error, i.e., we suppose that  $A\phi_i = \lambda_i\phi_i$ , where  $\{\phi_i\}_{i=1}^N$  are a set of orthonormal eigenvectors of  $A$ . We then set  $e^k = x - x^k$  and write

$$e^0 = \sum_{i=1}^N [(e^0)^T \phi_i] \phi_i.$$

Suppose we choose  $\alpha = \lambda_N$ , the largest eigenvalue of  $A$ . Then

$$e^k = [I - \alpha A]^k e^0 = \sum_{i=1}^N [(e^0)^T \phi_i] (1 - \lambda_i/\lambda_N)^k \phi_i.$$

Now for large eigenvalues  $1 - \lambda_i/\lambda_N$  is small, so the high frequency components of the error are damped out quickly, while for small eigenvalues  $1 - \lambda_i/\lambda_N \approx 1$ , and there is not much decay in the error and so the low frequency components are not changed much. Thus, a few iterations of this method has the effect of "smoothing" the error. We shall come back to this idea in a later lecture.

**5.2. Conjugate-Gradient method (CG).** A better choice of search directions  $\{p^k\}$  is to choose them to be  $A$ -orthogonal, i.e., to satisfy  $(p^j)^T A p^i = 0$  for  $i \neq j$ . In this case, the best choice of the  $\alpha_k$  are given by

$$\alpha_k = \frac{(p^k)^T [b - A x^0]}{(p^k)^T A p^k}.$$

The CG method generates the directions  $p^k$  recursively using the Gram-Schmidt orthogonalization process (see references for precise description). For the CG method, we get the following error estimate:

$$\|x - x^k\|_A \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x - x^0\|_A,$$

where  $\|x\|_A^2 = x^T A x$ . Since now  $\sqrt{\kappa}$  enters, the reduction is like  $1 - O(h)$ , better than before, but still slow.

In practice, one uses the idea of preconditioning. Instead of solving the system  $Ax = b$ , we solve the system  $BAx = Bb$ , where  $B$  is an approximation to  $A^{-1}$  that is easily computable. Then the rate of convergence depends on the condition number of  $BA$  instead of  $A$ . If  $B$  is a good approximation to  $A^{-1}$ , then  $BA \approx I$ , and so  $\kappa(BA)$  will be close to 1, and we will get a substantial error reduction at each iteration.