After today's class you should be able to

- define and interpret the condition number of a matrix
- relate nonnegativity preservation and maximum principles of elliptic operators to analogous properties of matrices

Conditioning of Linear Systems Resulting from Finite-Difference Methods

The *conditioning* of a problem $T(u) = 0$ is a more qualitative statement on 'how well-posed' this problem is: Does a small residual $T(v)$ imply a small error $u - v$, with a worst-case amplification factor that is 'not too large'?

Note that this statement is analogous to the stability of the discretisation scheme $T^h(u^h) = 0$, except that the conditioning is a *property of the exact problem* $T(u) = 0$, independent of any numerical method.

If small perturbations in the data of a problem change the result by also just a small amount, then we call such a problem *well-conditioned*. If however small perturbations in the data could change the result dramatically, then we refer to such a problem as *ill-conditioned*.

In this course, we are mostly interested in the conditioning of our linear systems (so $T(u) = Au - b$), which roughly tells us how difficult they are to solve. In practice, the data b (and also the matrix A) will be perturbed as soon as we type them into a computer due to the finite precision available and perhaps additional approximations. It is hence crucial to know how large an effect such perturbations could have on the solution of the linear system in the worst case:

2.2.13 Lemma (Absolute and Relative Condition Number of a Matrix) Let $A \in \mathbb{R}^{n\times n}$ be invertible and $b, \delta b \in \mathbb{R}^n$. Let $x \in \mathbb{R}^n$ be the unique solution of $Ax = b$ and $x + \delta x \in \mathbb{R}^n$ the unique solution of $A(x + \delta x) = b + \delta b$. By $\|\cdot\|$ we denote an arbitrary vector norm as well as the induced matrix norm.

(a) The absolute error $\|\delta x\|$ is bounded as follows:

$$\|\delta x\| \leq \|A^{-1}\|\|\delta b\|. \tag{2.15}$$

The constant $\|A^{-1}\|$ is called *absolute condition number* of A.

(b) The relative error $\frac{\|\delta x\|}{\|x\|}$ is bounded as follows:

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}. \tag{2.16}$$

The constant $\operatorname{cond}(A) = \|A\|\|A^{-1}\|$ is called *(relative) condition number* of A.

Proof. (a) Since $A\delta x = \delta b$, we have $\delta x = A^{-1}\delta b$ and thus, using the definition of an induced matrix norm,

$$\|x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|.$$

(b) Applying the same argument once again,

$$\|b\| = \|Ax\| \leq \|A\|\|x\|,$$

therefore

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}.$$

Now we divide the estimate from (a) by $\|x\|$ and apply this inequality to obtain

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}.$$

□

The condition number not only acts as the worst-case amplification factor of perturbations in the right hand side, but it also appears in the convergence estimates of iterative methods for solving the problem $Ax = b$. For instance, the number of conjugate gradient (CG) iterations required to solve a linear system with symmetric positive definite matrix A up to a given error tolerance is proportional to $\sqrt{\text{cond}_2(A)}$, the root of the condition number with respect to the Euclidean norm $\|\cdot\| = |\cdot|$. Hence, the larger the condition number, the more iterations are required until convergence.

2.2.14 Theorem (Condition Number of the Discrete Laplacian) The finite-difference approximation L^h of the Laplacian with Dirichlet boundary conditions has a condition number

$$\text{cond}_2(L^h) = O\left(\frac{1}{h^2}\right),$$

i.e. it becomes increasingly ill-conditioned as $h \to 0$.

Proof. Since L^h is symmetric positive definite, it has only positive real eigenvalues. With the Euclidean norm,

$$|L^h| = \lambda_{\max}(L^h) \qquad |(L^h)^{-1}| = \lambda_{\max}((L^h)^{-1}) = \frac{1}{\lambda_{\min}(L^h)},$$

hence $\text{cond}_2(L^h) = \frac{\lambda_{\max}(L^h)}{\lambda_{\max}(L^h)}$.

From Lemma 2.2.9 we obtain the eigenvalues in the 1D case with $N = 1/h$ subintervals

$$\lambda_k = \frac{2}{h^2} - \frac{2}{h^2}\cos\frac{k\pi}{N} \qquad k = 1, \dots, N-1$$

with

$$\lambda_{\min} = \lambda_1 = \frac{2}{h^2} - \frac{2}{h^2}\cos\frac{\pi}{N} = \frac{2}{h^2}\left(1 - \cos(\pi h)\right) = \pi^2 + O(h^2)$$
$$\lambda_{\max} = \lambda_{N-1} = \frac{2}{h^2} - \frac{2}{h^2}\cos\frac{(N-1)\pi}{N} = \frac{2}{h^2}\left(1 - \cos(\pi(1-h))\right) = \frac{2}{h^2} + O(1)$$

Consequently,

$$\text{cond}_2(L^h) = \frac{2}{\pi^2 h^2} + O(1) = O\left(\frac{1}{h^2}\right).$$

Using separation of variables, the eigenvalues in 2D are found to be

$$\lambda_{k,l} = \left(\frac{2}{h_1^2} - \frac{2}{h_1^2}\cos\frac{k\pi}{N_1}\right) + \left(\frac{2}{h_2^2} - \frac{2}{h_2^2}\cos\frac{l\pi}{N_2}\right) \qquad k = 1, \dots, N_1 - 1, \quad l = 1, \dots, N_2 - 1$$

with

$$\lambda_{\min} = \lambda_{1,1} = 2\pi^2 + O(h^2)$$
$$\lambda_{\max} = \lambda_{N_1-1,N_2-1} = \frac{2}{h_1^2} + \frac{2}{h_2^2} + O(1)$$

and analogously in higher dimensions. Thus, the $O(1/h^2)$-dependence of the condition number is independent of the number of spatial dimensions[3]. □

Discrete Maximum Principle

[3]The exponent 2 rather refects the fact that we are approximating a 2nd derivative.

We will now take a closer look at the structure of the discretised Poisson-Dirichlet (and other, not even necessarily elliptic) problems. The better we understand the properties of the 'big linear system' $L^h u^h = f^h$, the better we will understand what characteristic features of the continuous problem $Lu = f$ are preserved under a 'suitable' discretisation scheme. Furthermore, we will later use all the information that we can possibly gather on the matrix L^h to select a numerical method for solving the discrete problem that is guaranteed to converge, and that additionally exploits all the structure of L^h to compute u^h as efficiently as somehow possible.

Let's begin with a property that is slightly weaker[4] than the elliptic maximum principle from Theorem 2.1.4. Operators L that satisfy the elliptic maximum principle preserve nonnegativity of the source term and the boundary values: if $Lu \geq 0$ in the domain and if $u \geq 0$ on the boundary, then the solution u will also be ≥ 0 in the entire domain. This property of nonnegativity preservation is usually called *monotonicity* in the context of matrices.

2.2.15 Definition (Monotone Matrix) A square matrix $A \in \mathbb{R}^{n\times n}$ is said to be *monotone* if for all vectors $x \in \mathbb{R}^n$ with the property $Ax \geq 0$ it already follows that $x \geq 0$.

Do you still remember that we already came across this idea in 2.1.6? Back then, we noticed that (invertible) matrices with componentwise nonnegative inverse[5] $A^{-1} \geq 0$ preserve the sign of the right hand side data. Here is the complete statement of this observation:

2.2.16 Lemma (Monotonicity) A matrix $A \in \mathbb{R}^{n\times n}$ is monotone if and only if it is nonsingular and inverse-nonnegative.

Proof. Let $A \in \mathbb{R}^{n\times n}$ be monotone.

- A is nonsingular:

- $A^{-1} \geq 0$:

Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and inverse-nonnegative.

□

[4] but also satisfied by elliptic operators with zeroth-order coefficient $c \geq 0$

[5] A shorter expression for "having a componentwise nonnegative inverse matrix" is "inverse-nonnegative".

In reality, we are never able to explicitly compute the inverse of our very large matrices. The inverse of a sparse matrix is usually no longer sparse, and thus an enormous amount of computation would be required to calculate all n^2 entries of the inverse matrix and check their sign. Therefore, we need another criterion that implies that our big matrix is monotone, one that can be easily verified. After two more definitions, we'll be ready to present such a practicable criterion for monotonicity!

2.2.17 Definition (Diagonal Dominance) If in the i^{th} row of a matrix $A \in \mathbb{R}^{n\times n}$ the absolute value of the diagonal entry is

- greater than or equal to the sum of the absolute values of the off-diagonal terms

  then we say that A is *weakly diagonally dominant* in this row;
- greater than the sum of the absolute values of the off-diagonal terms

  then we say that A is *strictly diagonally dominant* in this row.

A matrix A with the properties that

(i) A is weakly diagonally dominant in all rows

(ii) A is strictly diagonally dominant in at least one row

(iii) for all rows i_0 there exists a chain of indices $i_0 \to i_1 \to \cdots \to i_s$ to a strictly diagonally dominant row i_s such that all $a_{i_{l-1}i_l} \neq 0$ $(l = 1, \ldots, s)$

is called *weakly chained diagonally dominant*. If, instead of (iii), there even holds that

(iv) for any two rows i_0, i_s there exists a chain of indices $i_0 \to i_1 \to \cdots \to i_s$ such that all $a_{i_{l-1}i_l} \neq 0$ and all $a_{i_l i_{l-1}} \neq 0$ $(l = 1, \ldots, s)$

then A is called *irreducibly diagonally dominant*.

The two chain properties describe how data from the right hand side and information from each component of the solution propagate through the linear system. As can be seen from the conditions (iii) and (iv), they describe the structure of the sparsity pattern of A.

With the weaker chain property (iii), information from row i_s is referred to in row i_{s-1}. Then the equation in row i_{s-2} refers to the component i_{s-1} of the solution, and hence indirectly also to i_s. Finally, row i_0 directly or indirectly depends on the components $i_1, i_2, \ldots, i_s$ of the solution and the right hand side. It is not required for (iii) that conversely row i_s also depends on row i_0.

This is the difference to the stronger property (iv). Here, information is shared globally and all rows directly or indirectly refer to themselves and all other rows.

2.2.18 Definition (Z-Matrices, L-Matrices and M-Matrices) A matrix $A \in \mathbb{R}^{n\times n}$ is called

- *Z-matrix* if all off-diagonal entries are negative or zero:
- *L-matrix* if all off-diagonal entries are negative or zero and all diagonal entries are positive:
- *M-matrix*, if it is a monotone Z-matrix.