

Lecture 7: Conditioning and Stability

1.4 Conditioning and Stability.

One can solve linear systems in a finite number of algebraic operations: Given the entries of \mathbf{A} and \mathbf{b} , with sufficient stamina we could compute – by hand, even – the exact answer \mathbf{x} to the system $\mathbf{Ax} = \mathbf{b}$. One does this in elementary linear algebra class with matrices of dimension $n = 3$ or perhaps, if the instructor has a sadistic streak, $n = 4$ or 5 . With considerable effort and sufficient motivation (historical episodes involving architecture and warfare come to mind) one could solve dense systems with, say, $n = 10$ or a bit more.

Computers make quick work of systems of such small size. While one could, in principle, work in exact arithmetic, this approach is rather slow for the large systems that arise from practical applications. Instead, MATLAB (and similar systems) perform *floating point arithmetic*, rounding the entries in \mathbf{A} and \mathbf{b} to nearest values in a finite number system, then performing approximate arithmetic in that system. The implementation in general use today, *IEEE double precision arithmetic*, provides roughly sixteen digits of relative accuracy when it stores and operates on numbers. We will explore the particular features of such arithmetic systems in the next lecture. As present, we are more concerned with how small changes in \mathbf{A} and \mathbf{b} will affect the solution \mathbf{x} .

Since the entries of \mathbf{A} and \mathbf{b} are only stored approximately, we cannot expect the ‘solution’ \mathbf{x} that we subsequently compute to be exact. At best, \mathbf{x} will be the *the exact solution to a ‘nearby’ linear system*. We need to understand how large the discrepancy (or *residual*) $\mathbf{Ax} - \mathbf{b}$ can be. The issue breaks into two subordinate questions. First, how do small changes in \mathbf{A} and \mathbf{b} affect the solution \mathbf{x} ? Next, how does the algorithm that computed \mathbf{x} perform if the arithmetic operations involved are not exact, but only accurate to, say, fifteen digits? The first of these questions concerns the *conditioning* of the mathematical problem; the second concerns the *stability* of the numerical algorithm that we used to compute ‘the solution’. The distinction between conditioning and stability is a fundamental concept in the design and analysis of numerical algorithms.

1.4.1 Abstract condition numbers.

First we address the conditioning of the linear system $\mathbf{Ax} = \mathbf{b}$. The *condition number* of a problem can be abstractly defined as follows.[†] Suppose we wish to compute some (vector-valued) function $\mathbf{f}(\mathbf{y})$. How does the value of \mathbf{f} change when \mathbf{y} is subjected to a perturbation $\delta\mathbf{y}$? Define

$$\delta\mathbf{f} = \mathbf{f}(\mathbf{y} + \delta\mathbf{y}) - \mathbf{f}(\mathbf{y}).$$

(Note that ‘ $\delta\mathbf{y}$ ’ and ‘ $\delta\mathbf{f}$ ’ are both names of vectors; $\delta\mathbf{y}$ *does not* mean ‘the scalar δ times the vector \mathbf{y} ’; thus, e.g., you cannot peel the ‘ δ ’ away from the ‘ \mathbf{y} ’ in $\delta\mathbf{y}$. This usage departs from our notational convention, but is standard for this style of analysis.) We need to compare the size of $\delta\mathbf{f}$ to that of $\delta\mathbf{y}$. Of course, the magnitudes of $\delta\mathbf{y}$ and $\delta\mathbf{f}$ depend on the size of \mathbf{y} and $\mathbf{f}(\mathbf{y})$, and hence we are interested in the *relative* size of these perturbations:

$$\frac{\|\delta\mathbf{f}\|/\|\mathbf{f}(\mathbf{y})\|}{\|\delta\mathbf{y}\|/\|\mathbf{y}\|}.$$

Can the numerator be large when the denominator is small? That is, can small relative changes in \mathbf{y} induce large shifts in the solution? Here $\delta\mathbf{y}$ is a vector, and one expects that some choices for

[†]See Lecture 12 of Trefethen and Bau, which provides the background for the present notes.

that vector in \mathbb{C}^n might stimulate more significant errors than others. To characterize the worst case, maximize over all $\delta\mathbf{y}$ of some fixed size, and continue what happens as the size of that error, $\|\delta\mathbf{y}\|$, goes to zero. The *condition number* of $\mathbf{f}(\mathbf{y})$ is thus defined as

$$\kappa = \lim_{\Delta \rightarrow 0} \max_{\|\delta\mathbf{y}\| \leq \Delta} \frac{\|\delta\mathbf{f}\|/\|\mathbf{f}(\mathbf{y})\|}{\|\delta\mathbf{y}\|/\|\mathbf{y}\|}.$$

1.4.2 Condition number for solving linear systems.

We wish to apply these ideas to analyze the conditioning of the linear system $\mathbf{Ax} = \mathbf{b}$. Here the data ‘ \mathbf{y} ’ comprises \mathbf{A} and \mathbf{b} , and ‘ $\mathbf{f}(\mathbf{y})$ ’ is $\mathbf{A}^{-1}\mathbf{b}$. Rather than attempting to plug these values into the above formula, we perform a more direct analysis.

Suppose $\mathbf{Ax} = \mathbf{b}$ for nonzero \mathbf{x} and \mathbf{b} , and \mathbf{A} is invertible. First, introduce an *infinitesimal* perturbation $\delta\mathbf{A}$ to \mathbf{A} that changes the solution by $\delta\mathbf{x}$:

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}.$$

Multiplying this out gives $\mathbf{Ax} + \mathbf{A}(\delta\mathbf{x}) + (\delta\mathbf{A})\mathbf{x} + (\delta\mathbf{A})(\delta\mathbf{x}) = \mathbf{b}$. Since $\delta\mathbf{A}$ and $\delta\mathbf{x}$ are infinitesimal quantities, we can ignore the quadratic term,

$$\mathbf{Ax} + \mathbf{A}(\delta\mathbf{x}) + (\delta\mathbf{A})\mathbf{x} = \mathbf{b}.$$

Since $\mathbf{Ax} = \mathbf{b}$, we can subtract the first terms on each side and rearrange to obtain

$$\mathbf{A}(\delta\mathbf{x}) = -(\delta\mathbf{A})\mathbf{x}.$$

Inverting \mathbf{A} and taking norms leads to

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}(\delta\mathbf{A})\mathbf{x}\| \leq \|\mathbf{A}^{-1}\| \|\delta\mathbf{A}\| \|\mathbf{x}\|.$$

As we seek to compare the *relative* change in the solution with the change in the \mathbf{A} , we want a $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ term on the right. Multiplying the right hand side by $1 = \|\mathbf{A}\|/\|\mathbf{A}\|$ and dividing by $\|\mathbf{x}\|$ yields

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

It follows that infinitesimal changes in \mathbf{A} can be magnified by no more than

$$\kappa(\mathbf{A}) := \|\mathbf{A}\| \|\mathbf{A}^{-1}\|,$$

which is called the *condition number of \mathbf{A} with respect to solving linear systems*. This is what most people mean when they casually speak of ‘the condition number of \mathbf{A} ’.

It is even simpler to consider how an infinitesimal change $\delta\mathbf{b}$ to \mathbf{b} can affect $\mathbf{Ax} = \mathbf{b}$. Now we have

$$\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b},$$

which, after canceling \mathbf{Ax} and \mathbf{b} from either side, reduces to

$$\delta\mathbf{x} = \mathbf{A}^{-1}(\delta\mathbf{b}).$$

Taking norms and multiplying by $1 = \|\mathbf{Ax}\|/\|\mathbf{b}\|$, we have

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}(\delta\mathbf{b})\| \frac{\|\mathbf{Ax}\|}{\|\mathbf{b}\|} \leq \|\mathbf{A}\| \|\mathbf{A}^{-1}\| \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \|\mathbf{x}\|,$$

which implies

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

Again, the condition number $\kappa(\mathbf{A})$ is the factor that magnifies the perturbation in the data.

It is a useful exercise to piece together the two previous arguments to handle perturbations to both \mathbf{A} and \mathbf{b} simultaneously:

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}.$$

How should we interpret the condition number? It measures the distance of \mathbf{A} from singularity. In particular, one can show (in any induced norm) that there exists a perturbation \mathbf{E} with $\|\mathbf{E}\| = 1/\|\mathbf{A}^{-1}\|$ such that $\mathbf{A} + \mathbf{E}$ is singular (i.e., $\mathbf{A} + \mathbf{E}$ has a zero eigenvalue), and this is the smallest perturbation that makes \mathbf{A} singular. Thus, the *relative* size of the smallest perturbation is $1/\kappa(\mathbf{A})$.

Is there any connection between $\kappa(\mathbf{A})$ and the determinant, $\det(\mathbf{A})$?