Errors for Linear Systems

When we solve a linear system $Ax = b$ we often do not know A and b exactly, but have only approximations $\hat{A}$ and $\hat{b}$ available. Then the best thing we can do is to solve $\hat{A}\hat{x} = \hat{b}$ exactly which gives a different solution vector $\hat{x}$. We would like to know how the errors of $\hat{A}$ and $\hat{b}$ influence the error in $\hat{x}$.

Example: Consider the linear system $Ax = b$ with

$$\begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

We can easily see that the solution is $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Now let us use the slightly different right hand side vector $\hat{b} = \begin{pmatrix} 2.02 \\ 1.98 \end{pmatrix}$ and solve the linear system $A\hat{x} = \hat{b}$. This gives the solution vector $\hat{x} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$. In this case a small change in the right hand side vector has caused a large change in the solution vector.

Vector norms

In order to measure errors in vectors by a single number we use a so-called *vector norm*.

A **vector norm** $\|x\|$ measures the size of a vector $x \in \mathbb{R}^n$ by a nonnegative number and has the following properties

$$\|x\| = 0 \quad \Rightarrow \quad x = 0$$
$$\|\alpha x\| = |\alpha| \, \|x\|$$
$$\|x + y\| \le \|x\| + \|y\|$$

for any $x, y \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$. There are many possible vector norms. We will use the three norms $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$ defined by

$$\|x\|_1 = |x_1| + \cdots + |x_n|$$
$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{1/2}$$
$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_d|\}$$

If we write $\|x\|$ in an equation without any subscript, then the equation is valid for all three norms (using the same norm everywhere).

If the exact vector is x and the approximation is $\hat{x}$ we can define the **relative error with respect to a vector norm** as $\frac{\|\hat{x}-x\|}{\|x\|}$.

Example: Note that in the above example we have $\frac{\|\hat{b}-b\|_\infty}{\|b\|_\infty} = 0.01$, but $\frac{\|\hat{x}-x\|_\infty}{\|x\|_\infty} = 1$. That means that the relative error of the solution is 100 times as large as the relative error in the given data, i.e., the condition number of the problem is at least 100.

Matrix norms

A *matrix norm* $\|A\|$ measures the size of a matrix $A \in \mathbb{R}^{n \times n}$ by a nonnegative number. We would like to have the property

$$\|Ax\| \le \|A\| \, \|x\| \qquad \text{for all } x \in \mathbb{R}^n \tag{1}$$

where $\|x\|$ is one of the above vector norms $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$. We define $\|A\|$ as the smallest number satisfying (1):

$$\|A\| := \sup_{\substack{x \in \mathbb{R}^n \\ x \ne 0}} \frac{\|Ax\|}{\|x\|} = \max_{\substack{x \in \mathbb{R}^n \\ \|x\| = 1}} \|Ax\|$$

By using the 1, 2, ∞ vector norm in this definition we obtain the matrix norms $\|A\|_1$, $\|A\|_2$, $\|A\|_\infty$ (which are in general different numbers). It turns out that $\|A\|_1$ and $\|A\|_\infty$ are easy to compute:

Theorem:
$\|A\|_\infty = \max\limits_{i=1,\dots,n} \sum\limits_{j=1,\dots,n} |a_{ij}|$ (maximum of *row* sums of absolute values)
$\|A\|_1 = \max\limits_{j=1,\dots,n} \sum\limits_{i=1,\dots,n} |a_{ij}|$ (maximum of *column* sums of absolute values)
Proof: For the infinity norm we have

$$\|Ax\|_\infty = \max_i \left| \sum_j a_{ij} x_j \right| \le \max_i \sum_j |a_{ij}| \, |x_j| \le \left(\max_i \sum_j |a_{ij}| \right) \|x\|_\infty$$

implying $\|A\|_\infty \le \max_i \sum_j |a_{ij}|$. Let i_* be the index where the maximum occurs and define $x_j = \operatorname{sign} a_{i_* j}$, then $\|x\|_\infty = 1$ and $\|Ax\|_\infty = \max_i \sum_j |a_{ij}|$.
For the 1-norm we have

$$\|Ax\|_1 = \sum_i \left| \sum_j a_{ij} x_j \right| \le \sum_j \left(\sum_i |a_{ij}| \right) |x_j| \le \left(\max_j \sum_i |a_{ij}| \right) \|x\|_1$$

implying $\|A\|_1 \le \max_j \sum_i |a_{ij}|$. Let j_* be the index where the maximum occurs and define $x_{j_*} = 1$ and $x_j = 0$ for $j \ne j_*$, then $\|x\|_1 = 1$ and $\|Ax\|_1 = \max_j \sum_i |a_{ij}|$. $\square$

We will not use $\|A\|_2$ since it is more complicated to compute (it involves eigenvalues).

Note that for $A, B \in \mathbb{R}^{n \times n}$ we have $\|AB\| \le \|A\| \, \|B\|$ since

$$\|ABx\| \le \|A\| \, \|Bx\| = \|A\| \, \|B\| \, \|x\| \, .$$

The following results about matrix norms will be useful later:

Lemma 1: $\|A - B\| < \frac{1}{\|A^{-1}\|}$ implies that B is nonsingular.
Proof: For $b := Ax$ we have $\|x\| = \left\|A^{-1}b\right\| \le \left\|A^{-1}\right\| \|b\|$ and therefore

$$\|Bx\| = \|Ax + (A-B)x\| \ge \|b\| - \|(A-B)x\| \ge \frac{1}{\|A^{-1}\|} \|x\| - \|A - B\| \, \|x\| \, .$$

Hence $Bx = 0$ and $\frac{1}{\|A^{-1}\|} - \|A - B\| > 0$ imply $x = 0$, i.e., B is nonsingular.

Lemma 2: For given vectors $x, y \in \mathbb{R}^n$ with $x \ne 0$ there exists a matrix $E \in \mathbb{R}^{n \times n}$ with $Ex = y$ and $\|E\| = \frac{\|y\|}{\|x\|}$.

Proof: For the infinity-norm we have $\|x\|_\infty = |x_j|$ for some j. Let $a \in \mathbb{R}^n$ be the vector with $a_j = 1$, $a_k = 0$ for $j \ne k$ and let

$$E = \frac{1}{\|x\|} y a^\top ,$$

then (i) $a^\top x = \|x\|_\infty$ implies $Ex = y$ and (ii) $\left\|ya^\top v\right\|_\infty = \left|a^\top v\right| \|y\|_\infty$ with $\left|a^\top v\right| \le \|v\|_\infty$ implies $\|E\| \le \frac{\|y\|}{\|x\|}$.
For the 1-norm we use $a \in \mathbb{R}^n$ with $a_j = \operatorname{sign}(x_j)$ since $a^\top x = \|x\|_1$ and $\left|a^\top v\right| \le \|v\|_1$.
For the 2-norm we use $a = x / \|x\|_2$ since $a^\top x = \|x\|_2$ and $\left|a^\top v\right| \le \|a\|_2 \|v\|_2 = \|v\|_2$.

Condition numbers

Let x denote the solution vector of the linear sytem $Ax = b$. If we choose a slightly different right hand side vector $\hat{b}$ then we obtain a different solution vector $\hat{x}$ satisfying $A\hat{x} = \hat{b}$. We want to know how the relative error $\left\|\hat{b} - b\right\| / \|b\|$ influences the relative error $\|\hat{x} - x\| / \|x\|$ ("error propagation"). We have $A(\hat{x} - x) = \hat{b} - b$ and therefore

$$\|\hat{x} - x\| = \left\|A^{-1}(\hat{b} - b)\right\| \le \left\|A^{-1}\right\| \left\|\hat{b} - b\right\| .$$

On the other hand we have $\|b\| = \|Ax\| \leq \|A\| \, \|x\|$. Combining this we obtain

$$\frac{\|\hat{x}-x\|}{\|x\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\left\|\hat{b}-b\right\|}{\|b\|}.$$

The number $\operatorname{cond}(A) := \|A\| \left\|A^{-1}\right\|$ is called **condition number** of the matrix A. It determines how much the relative error of the right hand side vector can be amplified. The condition number depends on the choice of the matrix norm: In general $\operatorname{cond}_1(A) := \|A\|_1 \left\|A^{-1}\right\|_1$ and $\operatorname{cond}_\infty(A) := \|A\|_\infty \left\|A^{-1}\right\|_\infty$ are different numbers.

Example: In the above example we have

$$\operatorname{cond}_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty = \left\|\begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix}\right\|_\infty \left\|\begin{pmatrix} 25.25 & -24.75 \\ -24.75 & 25.25 \end{pmatrix}\right\|_\infty = 2 \cdot 50 = 100$$

and therefore

$$\frac{\|\hat{x}-x\|}{\|x\|} \leq 100 \frac{\left\|\hat{b}-b\right\|}{\|b\|}$$

which is consistent with our results above (b and $\hat{b}$ were chosen so that the worst possible error magnification occurs).

The fact that the matrix A in our example has a large condition number is related to the fact that A is close to the singular matrix $B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

The following result shows that $\frac{1}{\operatorname{cond}(A)}$ indicates how close A is to a singular matrix:

Theorem: $$\min_{B \in \mathbb{R}^{n\times n},\, B \text{ singular}} \frac{\|A-B\|}{\|A\|} = \frac{1}{\operatorname{cond}(A)}$$

Proof: (1) Lemma 1 shows: B singular implies $\|A-B\| \geq \frac{1}{\|A^{-1}\|}$.

(2) By the definition of $\left\|A^{-1}\right\|$ there exist $x, y \in \mathbb{R}^n$ such that $x = A^{-1}y$ and $\left\|A^{-1}\right\| = \frac{\|x\|}{\|y\|}$. By Lemma 2 there exists a matrix $E \in \mathbb{R}^{n\times n}$ such that $Ex = y$ and $\|E\| = \frac{\|y\|}{\|x\|}$. Then $B := A - E$ satisfies $Bx = Ax - Ex = y - y = 0$, hence B is singular and $\|A-B\| = \|E\| = \frac{1}{\|A^{-1}\|}$. $\square$

Example: The matrix $A = \begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix}$ is close to the singular matrix $B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ so that $\frac{\|A-B\|_\infty}{\|A\|_\infty} = \frac{.02}{2} = .01$. By the theorem we have that $0.01 \leq \frac{1}{\operatorname{cond}_\infty(A)}$ or $\operatorname{cond}_\infty(A) \geq 100$. As we say above we have $\operatorname{cond}_\infty(A) = 100$, i.e., the matrix B is really the closest singular matrix to the matrix A.

When we solve a linear system $Ax = b$ we have to store the entries of A and b in the computer, yielding a matrix $\hat{A}$ with rounded entries $\hat{a}_{ij} = fl(a_{ij})$ and a rounded right hand side vector $\hat{b}$. If the original matrix A is singular then the linear system has no solution or infinitely many solutions, so that any computed solution is meaningless. How can we recognize this on a computer? Note that the matrix $\hat{A}$ which the computer uses may no longer be singular.

Answer: We should compute (or at least estimate) $\operatorname{cond}(\hat{A})$. If $\operatorname{cond}(\hat{A}) < \frac{1}{\varepsilon_M}$ then we can guarantee that *any* matrix A which is rounded to $\hat{A}$ must be nonsingular: $|\hat{a}_{ij} - a_{ij}| \leq \varepsilon_M |a_{ij}|$ implies $\left\|\hat{A} - A\right\| \leq \varepsilon_M \|A\|$ for the infinity or 1-norm. Therefore $\frac{\|\hat{A}-A\|}{\|\hat{A}\|} \leq \frac{\varepsilon_M}{1-\varepsilon_M} \approx \varepsilon_M$ and $\operatorname{cond}(\hat{A}) < \frac{1-\varepsilon_M}{\varepsilon_M} \approx \frac{1}{\varepsilon_M}$ imply $\frac{\|\hat{A}-A\|}{\|\hat{A}\|} < \frac{1}{\operatorname{cond}(\hat{A})}$. Hence the matrix A must be nonsingular by the theorem.

Now we assume that we perturb both the right hand side vector b and the matrix A:

Theorem: Assume $Ax = b$ and $\hat{A}\hat{x} = \hat{b}$. If A is nonsingular and $\left\|\hat{A} - A\right\| \leq 1/\left\|A^{-1}\right\|$ there holds

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)\frac{\|\hat{A}-A\|}{\|A\|}} \left(\frac{\left\|\hat{b} - b\right\|}{\|b\|} + \frac{\left\|\hat{A} - A\right\|}{\|A\|} \right)$$

Proof: Let $E = \hat{A} - A$, hence $A\hat{x} = \hat{b} - E\hat{x}$. Subtracting $Ax = b$ gives $A(\hat{x} - x) = (\hat{b} - b) - E\hat{x}$ and therefore

$$\|\hat{x} - x\| \leq \left\|A^{-1}\right\| \left(\left\|\hat{b} - b\right\| + \|E\| \|\hat{x}\| \right).$$

Dividing by $\|x\|$ and using $\|b\| \leq \|A\| \|x\| \iff \|x\| \geq \|b\| / \|A\|$ gives

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \left\|A^{-1}\right\| \|A\| \left(\frac{\left\|\hat{b} - b\right\|}{\|b\|} + \frac{\|E\|}{\|A\|} \frac{\|\hat{x}\|}{\|x\|} \right)$$

Now we have $\frac{\|\hat{x}\|}{\|x\|} \leq \frac{\|x\|+\|\hat{x}-x\|}{\|x\|} = 1 + \frac{\|\hat{x}-x\|}{\|x\|}$. By putting $\frac{\|\hat{x}-x\|}{\|x\|}$ on the left hand side and solving for it we obtain the assertion. $\square$

If $\operatorname{cond}(A)\frac{\|\hat{A}-A\|}{\|A\|} \ll 1$ we have that both the relative error in the right hand side vector and in the matrix are magnified by $\operatorname{cond}(A)$.

If $\operatorname{cond}(A)\frac{\|\hat{A}-A\|}{\|A\|} = \left\|A^{-1}\right\| \left\|\hat{A} - A\right\| \geq 1$ then by the theorem for the condition number the matrix $\hat{A}$ may actually be singular, so that the solution $\hat{x}$ is no longer well defined.

Computing the condition number

We have seen that the condition number is very useful: It tells us what accuracy we can expect for the solution, and how close our matrix is to a singular matrix.

In order to compute the condition number we have to find A^{-1}. This takes $n^3 + O(n^2)$ operations, compared with $\frac{n^3}{3} + O(n^2)$ operations for the LU-decomposition. Therefore the computation of the condition number would make the solution of a linear system 3 times as expensive. For large problems this is not reasonable.

However, we do not need to compute the condition number with full machine accuracy. Just knowing the order of magnitude is sufficient. Assume that we pick a vector c and solve the linear sytem $Az = c$. Then $z = A^{-1}c$ and $\|z\| \leq \left\|A^{-1}\right\| \|c\|$ or

$$\left\|A^{-1}\right\| \geq \frac{\|z\|}{\|c\|}.$$

This gives us a lower bound for $\left\|A^{-1}\right\|$, and the cost of this operation is only $n^2 + O(n)$. The trick is to pick c such that $\frac{\|z\|}{\|c\|}$ becomes as large as possible, so that the lower bound is close $\left\|A^{-1}\right\|$. There are a number of heuristic methods available which achieve fairly good lower bounds: (i) Pick $c = (\pm 1, \ldots, \pm 1)$ and pick the signs so that the forward susbstitution gives a large vector, (ii) picking $\tilde{c} := z$ and solve $A\tilde{z} = \tilde{c}$ often improves the lower bound. The Matlab functions `condest(A)` and `1/rcond(A)` use similar ideas to give lower bounds for $\operatorname{cond}_1(A)$. Typically they give an estimated condition number c with $c \leq \operatorname{cond}_1(A) \leq 3c$ and require the solution of 2 or 3 linear systems which costs $O(n^2)$ operations if the LU decomposition is known. (However, the Matlab commands `condest` and `rcond` only use the matrix A as an input value, so they have to compute the LU decomposition of A first and need $\frac{n^3}{3} + O(n^2)$ operations.)

Computation in machine arithmetic and residuals

When we run Gaussian elimination on a computer each single operation causes some roundoff error, and instead of the exact solution x of a linear system we only get an approximation $\hat{x}$. As explained above we should select the pivot candidate with the largest absolute value to avoid unnecessary subtractive cancellation, and this usually is a numerically stable algorithm. However, there is no theorem which guarantees this for partial pivoting (row interchanges). (For “full pivoting” with row *and* column interchanges some theoretical results exist. However, this algorithm is more expensive, and for all practical examples partial pivoting seems to work fine.)

Question 1: How much error do we have to accept for $\frac{\|\hat{x}-x\|}{\|x\|}$? This is the *unavoidable error* which occurs even for an ideal algorithm where we only round the input values and the output value to machine accuracy, and use infinite accuracy for all computations.

When we want to solve $Ax = b$ we have to store the entries of A, b in the computer, yielding a matrix $\hat{A}$ and a right hand side vector $\hat{b}$ of machine numbers so that $\frac{\|\hat{A}-A\|}{\|A\|} \le \varepsilon_M$ and $\frac{\|\hat{b}-b\|}{\|b\|} \le \varepsilon_M$. An ideal algorithm would then try to solve this linear system exactly, i.e., compute a vector $\hat{x}$ such that $\hat{A}\hat{x} = \hat{b}$. Then we have

$$\frac{\|\hat{x}-x\|}{\|x\|} \le \frac{\operatorname{cond}(A)}{1-\operatorname{cond}(A)\varepsilon_M}(\varepsilon_M + \varepsilon_M) \approx 2\operatorname{cond}(A)\varepsilon_M$$

if $\operatorname{cond}(A) \ll 1/\varepsilon_M$. Therefore the unavoidable error is $2\operatorname{cond}(A)\varepsilon_M$.

Question 2: After we computed $\hat{x}$ how can we check how good our computation was? The obvious thing to check is $\hat{b} := A\hat{x}$ and to compare it with b. The difference $r = \hat{b} - b$ is called the residual. As $Ax = b$ and $A\hat{x} = \hat{b}$ we have

$$\frac{\|\hat{x}-x\|}{\|x\|} \le \operatorname{cond}(A)\frac{\left\|\hat{b}-b\right\|}{\|b\|}$$

where $\left\|\hat{b}-b\right\| / \|b\|$ is called the *relative residual*. We can compute (or at least estimate) $\operatorname{cond}(A)$, and therefore can obtain an upper bound for the error $\|\hat{x}-x\| / \|x\|$.

If $\left\|\hat{b}-b\right\| / \|b\|$ is not much larger than ε_M then the computation was numerically stable: Just perturbing the input slightly from b to $\hat{b}$ and then doing everything else exactly would give the same result $\hat{x}$.

But it can happen that the relative residual is much larger than ε_M, and yet the computation is numerically stable. We obtain a better way to measure numerical stability by considering perturbations of the matrix A:

Assume we have a computed solution $\hat{x}$. If we can find a slightly perturbed matrix $\tilde{A}$ such that

$$\frac{\left\|\tilde{A}-A\right\|}{\|A\|} \le \varepsilon, \qquad \tilde{A}\hat{x} = b \tag{2}$$

where ε not much larger than ε_M, then the computation is numerically stable: Just perturbing the matrix within the roundoff error and then doing everything exactly gives the same result as our computation.

How can we check whether such a matrix $\tilde{A}$ exists? Compute the "weighted residual"

$$\rho := \frac{\left\|\hat{b}-b\right\|}{\|A\|\,\|\hat{x}\|}.$$

Then:

1. If $\hat{x}$ is the solution of a slightly perturbed problem (2) we have $\rho \le \varepsilon$.
2. If $\rho \le \varepsilon$ then $\hat{x}$ is the solution of a slightly perturbed problem (2).

Proof:

1. Let $E = \tilde{A} - A$. Then $(A+E)\hat{x} = b$ or $\hat{b} - b = -E\hat{x}$ yielding

$$\left\|\hat{b}-b\right\| \le \|E\|\,\|\hat{x}\|, \qquad \frac{\left\|\hat{b}-b\right\|}{\|A\|\,\|\hat{x}\|} \le \frac{\|E\|}{\|A\|} \le \varepsilon.$$

2. Let $y := b - \hat{b}$. Using Lemma 2 we get a matrix E with $E\hat{x} = y$ and $\|E\| = \frac{\|y\|}{\|\hat{x}\|}$. Then $\tilde{A} := A + E$ satisfies $\tilde{A}\hat{x} = (A+E)\hat{x} = \hat{b} + (b - \hat{b}) = b$ and $\frac{\|E\|}{\|A\|} = \frac{\|b-\hat{b}\|}{\|\hat{x}\|\|A\|} \le \varepsilon$.