Numerical Linear Algebra
Lecture 3

NLA notions

- Norms
- Sherman-Morrison formula
- Perturbation theory
- Condition number, relative condition number
- Gaussian elimination

Singular values

The singular values of a compact operator $T : X \to Y$ acting between Hilbert spaces X and Y, are the square roots of the eigenvalues of the nonnegative self-adjoint operator $T^*T : X \to X$ (where T^* denotes the adjoint of T). For the case of a matrix A, singular values are computed as: $\sigma = \sqrt{\lambda(A^*A)}$.
The singular values are nonnegative real numbers, usually listed in decreasing order $(\sigma_1(T), \sigma_2(T), ...)$. If T is self-adjoint, then the largest singular value $\sigma_1(T)$ is equal to the operator norm of T.
In the case of a normal matrix A (or $A^*A = AA^*$, when A is real then $A^TA = AA^T$), the spectral theorem can be applied to obtain unitary diagonalization of A as $A = U\Lambda U^*$. Therefore, $\sqrt{A^*A} = U|\Lambda|U^*$ and so the singular values are simply the absolute values of the eigenvalues.

Norm of a vector

We are going to use following such called L_p-norms which are usually denoted by $\|\cdot\|_p$:

$$\|x\|_p = (\sum_{k=1}^{n} |x_k|^p)^{1/p}, \quad p \geq 1.$$

- $p = 1, \|x\|_1 = \sum_{k=1}^{n} |x_k|$, one-norm
- $p = 2, \|x\|_2 = (\sum_{k=1}^{n} |x_k|^2)^{1/2}$, two-norm
- $p = \infty, \|x\|_\infty = \max_{1 \leq k \leq n} |x_k|$, max-norm or infinity-norm.

Norm of a vector

All these norms, as all other vector norms, has following properties:

- $x \neq 0 \rightarrow \|x\| > 0$ (positivity), $\|0\| = 0$.
- $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ (homogeneity)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

Norm of a vector

Norms are defined differently, but they can be compared. Two norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ on a vector space V are called equivalent if there exist positive real numbers C and D such that for all x in V

$$C\,\|x\|_\alpha \leq \|x\|_\beta \leq D\,\|x\|_\alpha\,.$$

For example, let $x = (x_1, ..., x_n)^T$

$$\alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1$$

In this case we have: $\alpha = 1/\sqrt{n}, \beta = 1$.

Norm of a vector

Other examples:

- $$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2$$
- $$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$$
- $$\|x\|_\infty \leq \|x\|_1 \leq n\,\|x\|_\infty\,,$$
- $$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2 \leq n\,\|x\|_\infty$$

If the vector space is a finite-dimensional real or complex one, all norms are equivalent. On the other hand, in the case of infinite-dimensional vector spaces, not all norms are equivalent.

Inner product

The dot product of two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ is defined as:

$x \cdot y = (x, y) = \sum_{k=1}^{n} x_k y_k = x^T y, \quad \|x\|_2 = \sqrt{x^T x}.$
We note that $x^T y$ is a scalar, but xy^T is a matrix.

Example

$$x^T y = [-1, 2, 3] \cdot \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = (-1) \cdot (3) + 2 \cdot 2 + 3 \cdot 1 = 4.$$

$$xy^T = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix} \cdot [3, 2, 1] = \begin{bmatrix} -3 & -2 & -1 \\ 6 & 4 & 2 \\ 9 & 6 & 3 \end{bmatrix}$$

Norm of a vector

Example

$$x = \begin{bmatrix} -1 \\ 2 \\ 3 \\ -5 \end{bmatrix}; \qquad \begin{aligned} ||x||_1 &= |-1| + |2| + |3| + |-5| = 11 \\ ||x||_2 &= \sqrt{(-1)^2 + 2^2 + 3^2 + (-5)^2} = \sqrt{39} \end{aligned}$$

$||x||_\infty = max(|-1|, |2|, |3|, |-5|) = 5.$

Norm of a vector

A vector x is normalized if $||x|| = 1$. If $x \neq 0$ then $\frac{x}{||x||}$ is normalized vector.

Example

$$x^T x = [-1, 2, 3, -5] \cdot \begin{bmatrix} -1 \\ 2 \\ 3 \\ -5 \end{bmatrix} = (-1)\cdot(-1) + 2\cdot 2 + 3\cdot 3 + (-5)^2 = 39$$

$$||x||_2 = \sqrt{39}$$

$$V = \frac{x}{||x||} = \left[\frac{-1}{\sqrt{39}}, \frac{2}{\sqrt{39}}, \frac{3}{\sqrt{39}}, \frac{-5}{\sqrt{39}}\right]^T \Longrightarrow ||V||_2 = 1$$

Matrix norm. Definition

Let K will denote the field of real or complex numbers. Let $K^{m\times n}$ denote the vector space containing all matrices with m rows and n columns with entries in K. Let A^* denotes the conjugate transpose of matrix A.
A matrix norm is a vector norm on $K^{m\times n}$. That is, if $\|A\|$ denotes the norm of the matrix A, then,

- $\|A\| > 0$ if $A \neq 0$ and $\|A\| = 0$ if $A = 0$.
- $\|\alpha A\| = |\alpha|\|A\|$ for all α in K and all matrices A in $K^{m\times n}$.
- $\|A + B\| \leq \|A\| + \|B\|$ for all matrices A and B in $K^{m\times n}$.

Matrix norm. Definition

In the case of square matrices (thus, $m = n$), some (but not all) matrix norms satisfy the following condition, which is related to the fact that matrices are more than just vectors:
$\|AB\| \leq \|A\|\|B\|$ for all matrices A and B in $K^{n\times n}$.
A matrix norm that satisfies this additional property is called a sub-multiplicative norm. The set of all n-by-n matrices, together with such a sub-multiplicative norm, is an example of a Banach algebra.

Matrix norm. Induced norm

If vector norms on K_m and K_n are given (K is field of real or complex numbers), then one defines the corresponding induced norm or operator norm on the space of m-by-n matrices as the following maxima:

$$\|A\| = \max\{\|Ax\| : x \in K^n \text{ with } \|x\| = 1\}$$
$$= \max\left\{\frac{\|Ax\|}{\|x\|} : x \in K^n \text{ with } x \neq 0\right\}.$$

If $m = n$ and one uses the same norm on the domain and the range, then the induced operator norm is a sub-multiplicative matrix norm.

Matrix norm. p-norm

The operator norm corresponding to the p-norm for vectors is:

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

In the case of $p = 1$ and $p = \infty$, the norms can be computed as:

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|,$$

which is simply the maximum absolute column sum of the matrix.

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|,$$

which is simply the maximum absolute row sum of the matrix

Matrix norm. Example

For example, if the matrix A is defined by

$$A = \begin{bmatrix} 3 & 5 & 7 \\ 2 & 6 & 4 \\ 0 & 2 & 8 \end{bmatrix},$$

then we have $||A||_1 = max(5, 13, 19) = 19$ and
$||A||_\infty = max(15, 12, 10) = 15$. Consider another example

$$A = \begin{bmatrix} 2 & 4 & 2 & 1 \\ 3 & 1 & 5 & 2 \\ 1 & 2 & 3 & 3 \\ 0 & 6 & 1 & 2 \end{bmatrix},$$

where we add all the entries in each column and determine the greatest value, which results in $||A||_1 = max(6, 13, 11, 8) = 13$.
We can do the same for the rows and get $||A||_\infty = max(9, 11, 9, 9) = 11$.
Thus 11 is our max.

Matrix norm. Example

In the special case of $p = 2$ (the Euclidean norm) and $m = n$ (square matrices), the induced matrix norm is the spectral norm. The spectral norm of a matrix A is the largest singular value of A i.e. the square root of the largest eigenvalue of the positive-semidefinite matrix A^*A:

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A)$$

where A^* denotes the conjugate transpose of A.

Matrix norm. Example

Example

$$A = \begin{bmatrix} 1 & -2 & -3 \\ 6 & 4 & 2 \\ 9 & -6 & 3 \end{bmatrix}$$

$$||A||_2 = \max \sqrt{\lambda(A^T A)}$$

$$A^T = \begin{bmatrix} 1 & 6 & 9 \\ -2 & 4 & -6 \\ -3 & 2 & 3 \end{bmatrix}, \; A^T A = \begin{bmatrix} 118 & -32 & 36 \\ -32 & 56 & -4 \\ 36 & -4 & 22 \end{bmatrix}$$

$$\lambda(A^T A) = \begin{bmatrix} 8.9683 \\ 45.3229 \\ 141.7089 \end{bmatrix}; \quad \max \sqrt{\lambda(A^T A)} = \max(2.9947, 6.7322, 11.9042) = 11.9042$$

Matrix norm. Example

Example

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \; A^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \; A^T A - \lambda I = \begin{bmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{bmatrix} = 0;$$

$$\lambda_1 = 1, \; \lambda_2 = 1; \; ||A||_2 = \max \sqrt{\lambda(A^T A)} = \max(1,1) = 1$$

Matrix norm

Any induced norm satisfies the inequality

$$\|A\| \geq \rho(A),$$

where $\rho(A) := \max\{|\lambda_1|, ..., |\lambda_m|\}$ is the spectral radius of A. For a symmetric or hermitian matrix A, we have equality for the 2-norm, since in this case the 2-norm is the spectral radius of A. For an arbitrary matrix, we may not have equality for any norm.

Example

Take

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

the spectral radius $\rho(A)$ of A is 0, but A is not the zero matrix, and so none of the induced norms are equal to the spectral radius of A:

$$\|A\|_1 = 1, \|A\|_\infty = 1, \|A\|_2 = 1.$$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A); A^* A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Matrix norm. "Entrywise" norms

For square matrices we have the spectral radius formula:

$$\lim_{r\to\infty} \|A^r\|^{1/r} = \rho(A).$$

These vector norms treat an $m \times n$ matrix as a vector of size $m \cdot n$, and use one of the familiar vector norms.
For example, using the p-norm for vectors, we get:

$$\|A\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p \right)^{1/p}.$$

The special case $p = 2$ is the Frobenius norm, and $p = \infty$ yields the maximum norm.

Matrix norm. Frobenius norm.

For $p = 2$, this is called the Frobenius norm or the Hilbert - Schmidt norm, though the latter term is often reserved for operators on Hilbert space. This norm can be defined in various ways:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\text{trace}(A^*A)} = \sqrt{\sum_{i=1}^{\min\{m,\,n\}}\sigma_i^2}$$

where A^* denotes the conjugate transpose of A, σ_i are the singular values of A, and the trace function is used. The Frobenius norm is very similar to the Euclidean norm on K_n and comes from an inner product on the space of all matrices. The Frobenius norm is sub-multiplicative and is very useful for numerical linear algebra. This norm is often easier to compute than induced norms and has the useful property of being invariant under rotations.

Matrix norm. Max norm.

The max norm is the elementwise norm with $p = \infty$:

$$\|A\|_{\max} = \max\{|a_{ij}|\}.$$

This norm is not sub-multiplicative.

Matrix norm. Equivalence of norms.

For any two vector norms $||\cdot||_\alpha$ and $||\cdot||_\beta$, we have

$$r\,\|A\|_\alpha \leq \|A\|_\beta \leq s\,\|A\|_\alpha$$

for some positive numbers r and s, for all matrices A in $K^{m\times n}$. In other words, they are equivalent norms; they induce the same topology on $K^{m\times n}$. This is a special case of the equivalence of norms in finite-dimensional Normed vector spaces.

Examples of norm equivalence

For matrix $A \in \mathbb{R}^{m \times n}$ the following inequalities hold:

- $\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\|A\|_2$, where r is the rank of A
- $\|A\|_{\max} \leq \|A\|_2 \leq \sqrt{mn}\|A\|_{\max}$
- $\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty$
- $\frac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1$.

Here, $||\cdot||_p$ refers to the matrix norm induced by the vector p-norm.
Another useful inequality between matrix norms is

$$\|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty}.$$

Sherman - Morrison formula

Suppose A is an invertible square matrix and u, v are vectors. Suppose furthermore that $1 + v^T A^{-1} u \neq 0$. Then the Sherman-Morrison formula states that

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} uv^T A^{-1}}{1 + v^T A^{-1} u}. \quad (1)$$

Here, uv^T is the outer product of two vectors u and v.
If the inverse of A is already known, the formula provides a numerically cheap way to compute the inverse of A corrected by the matrix uv^T.

We verify the properties of the inverse. A matrix Y (in this case the right-hand side of the Sherman - Morrison formula) is the inverse of a matrix X (in this case $A + uv^T$) if and only if $XY = YX = \mathrm{I}$.
We first verify that the right hand side (Y) satisfies $XY = \mathrm{I}$.

$$
\begin{aligned}
XY &= (A + uv^T)\left(A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}\right) \\
&= AA^{-1} + uv^TA^{-1} - \frac{AA^{-1}uv^TA^{-1} + uv^TA^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u} \\
&= I + uv^TA^{-1} - \frac{uv^TA^{-1} + uv^TA^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u} \\
&= I + uv^TA^{-1} - \frac{u(1 + v^TA^{-1}u)v^TA^{-1}}{1 + v^TA^{-1}u}.
\end{aligned}
$$

Note that $v^T A^{-1} u$ is a scalar, so $(1 + v^T A^{-1} u)$ can be factored out, leading to:

$$XY = I + uv^T A^{-1} - uv^T A^{-1} = I.$$

In the same way, it is verified that

$$YX = \left(A^{-1} - \frac{A^{-1} uv^T A^{-1}}{1 + v^T A^{-1} u} \right) (A + uv^T) = I.$$

Perturbation Theory

- Consider linear system $Ax = b$,

Perturbation Theory

- Consider linear system $Ax = b$,
- $\hat{x}$ such that $\hat{x} = \delta x + x$ is its computed solution.

Perturbation Theory

- Consider linear system $Ax = b$,
- $\hat{x}$ such that $\hat{x} = \delta x + x$ is its computed solution.
- Suppose $(A + \delta A)\hat{x} = b + \delta b$.

Perturbation Theory

- Consider linear system $Ax = b$,
- $\hat{x}$ such that $\hat{x} = \delta x + x$ is its computed solution.
- Suppose $(A + \delta A)\hat{x} = b + \delta b$.
- Goal: to bound the norm of $\delta x \equiv \hat{x} - x$.

Perturbation Theory

- Consider linear system $Ax = b$,
- $\hat{x}$ such that $\hat{x} = \delta x + x$ is its computed solution.
- Suppose $(A + \delta A)\hat{x} = b + \delta b$.
- Goal: to bound the norm of $\delta x \equiv \hat{x} - x$.
- Subtract the equalities and solve them for δx

Perturbation Theory

- Consider linear system $Ax = b$,
- $\hat{x}$ such that $\hat{x} = \delta x + x$ is its computed solution.
- Suppose $(A + \delta A)\hat{x} = b + \delta b$.
- Goal: to bound the norm of $\delta x \equiv \hat{x} - x$.
- Subtract the equalities and solve them for δx
- Rearranging terms we get:

$$\delta x = A^{-1}(-\delta A\hat{x} + \delta b)$$

More precisely, let us consider following problems:

$$Ax = b \tag{2}$$

$$(A + \delta A)(x + \delta x) = b + \delta b \tag{3}$$

Subtract (2)-(3) to get:

$$Ax + A\delta x + \delta Ax + \delta A\delta x - Ax = b + \delta b - b. \tag{4}$$

$$\begin{aligned} A\delta x + \delta Ax + \delta A\delta x &= \delta b, \\ (A + \delta A)\delta x + \delta Ax &= \delta b. \end{aligned} \tag{5}$$

$$\begin{aligned} (A + \delta A)\delta x &= \delta b - \delta Ax, \\ (A + \delta A)\delta x &= \delta b - \delta A(\hat{x} - \delta x), \\ A\delta x + \delta A\delta x &= \delta b - \delta A\hat{x} + \delta A\delta x, \\ \delta x &= A^{-1}(\delta b - \delta A\hat{x}). \end{aligned} \tag{6}$$

- Taking norms and triangle inequality leads us to

$$\|\delta x\| \leq \|A^{-1}\|(\|\delta A\| \cdot \|\hat{x}\| + \|\delta b\|)$$

- Taking norms and triangle inequality leads us to

$$\|\delta x\| \leq \|A^{-1}\|(\|\delta A\| \cdot \|\hat{x}\| + \|\delta b\|)$$

- Rearranging inequality gives us

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \|A^{-1}\| \cdot \|A\| \cdot (\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|\hat{x}\|})$$

where $k(A) = \|A^{-1}\| \cdot \|A\|$ is the condition number of the matrix A

Lemma

Let $\|\cdot\|$ satisfy $\|AB\| \leq \|A\| \cdot \|B\|$. Then $\|X\| < 1$ implies that $I - X$ is invertible. $(I - X)^{-1} = \sum_{i=0}^{\infty} X^i$, and

$$\|(I - X)^{-1}\| \leq \frac{1}{1 - \|X\|}. \tag{7}$$

- Recall

$$\delta x = A^{-1}(-\delta A\hat{x} + \delta b)$$

We can rewrite the above equation by multiplying by A as

$$\begin{aligned} A\delta x &= -\delta A\hat{x} + \delta b, \\ A\delta x + \delta A\hat{x} &= \delta b, \\ A\delta x + \delta A(x + \delta x) &= \delta b, \\ \delta Ax + (A + \delta A)\delta x &= \delta b. \end{aligned}$$

Solving this equation in the form $\delta Ax + (A + \delta A)\delta x = \delta b$ for δx gives:

- Recall

$$\delta x = A^{-1}(-\delta A\hat{x} + \delta b)$$

We can rewrite the above equation by multiplying by A as

$$\begin{aligned} A\delta x &= -\delta A\hat{x} + \delta b, \\ A\delta x + \delta A\hat{x} &= \delta b, \\ A\delta x + \delta A(x + \delta x) &= \delta b, \\ \delta Ax + (A + \delta A)\delta x &= \delta b. \end{aligned}$$

Solving this equation in the form $\delta Ax + (A + \delta A)\delta x = \delta b$ for δx gives:

-

$$\begin{aligned} \delta x &= \big((A + \delta A)^{-1}(-\delta Ax + \delta b) \\ &= [A(I + A^{-1}\delta A)]^{-1}(-\delta Ax + \delta b) \\ &= (I + A^{-1}\delta A)^{-1}A^{-1}(-\delta Ax + \delta b)\big) \end{aligned} \tag{8}$$

- Takes norms from both sides and then divide them by $\|x\|$:

$$\begin{aligned}
\frac{\|\delta x\|}{\|x\|} &\leq \|(I + A^{-1}\delta A)^{-1}\| \cdot \|A^{-1}\| \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \\
&\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|} \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \ (\textit{Lemma}\ ) \\
&= \frac{\|A^{-1}\| \cdot \|A\|}{1 - \|A^{-1}\| \cdot \|A\| \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|x\|} \right) \qquad (9) \\
&\leq \frac{k(A)}{1 - k(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)
\end{aligned}$$

Perturbation Theory

- Takes norms from both sides and then divide them by $\|x\|$:

$$\begin{aligned}
\frac{\|\delta x\|}{\|x\|} &\leq \|(I + A^{-1}\delta A)^{-1}\| \cdot \|A^{-1}\| \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \\
&\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|} \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \ (\textit{Lemma}\ ) \\
&= \frac{\|A^{-1}\| \cdot \|A\|}{1 - \|A^{-1}\| \cdot \|A\| \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|x\|} \right) \\
&\leq \frac{k(A)}{1 - k(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)
\end{aligned} \tag{9}$$

- expresses the relative error $\frac{\|\delta x\|}{\|x\|}$ in the solution as multiple of the relative errors $\frac{\|\delta A\|}{\|A\|}$ and $\frac{\|\delta b\|}{\|b\|}$ in the input.

Theorem

Let A be non-singular. Then
$min\left\{\frac{\|\delta A\|_2}{\|A\|_2} : A+\delta A \quad singular\right\} = \frac{1}{\|A^{-1}\|_2 \cdot \|A\|_2} = \frac{1}{k(A)}$. *Therefore the distance to the nearest singular matrix (ill-posed problem)*$=\frac{1}{condition\ number}$

Pertubation theory. Second approach.

Let us consider $Ax = b$. Let the computed solution $\hat{x} = \delta x + x$. Thus, $\delta x = \hat{x} - x = \hat{x} - A^{-1}b$.
Let us define residual as $r = A\hat{x} - b$. Then $\hat{x} = A^{-1}(r + b)$ and $\delta x = \hat{x} - x = \hat{x} - A^{-1}b = A^{-1}(r + b) - A^{-1}b$.
Thus, $\delta x = A^{-1}r$ and we can write

$$||\delta x|| \leq ||A^{-1}r|| \leq ||A^{-1}|| \cdot ||r||.$$

This is the simplest way to estimate δx.

Theorem

Let $r = A\hat{x} - b$ Then there exists a δA such that $\|\delta A\| = \frac{\|r\|}{\|\hat{x}\|}$ and $(A + \delta A)\hat{x} = b$. No δA of smaller norm and satisfying $(A + \delta A)\hat{x} = b$ exists. Thus, δA is the smallest possible backward error (measured in norm). This is true for any vector norm and its induced norm (or $\|\cdot\|_2$ for vectors and $\|\cdot\|_F$ for matrices).

Proof.

$(A + \delta A)\hat{x} = b$ if $\delta A \cdot \hat{x} = -r$, so $\|r\| = \|\delta A \cdot \hat{x}\| \leq \|\delta A\| \cdot \|\hat{x}\|$, implying $\|\delta A\| \geq \frac{\|r\|}{\|\hat{x}\|}$. We complete the proof only for the two-norm and its induced matrix norm. Choose $\delta A = \frac{-r \cdot \hat{x}^T}{\|\hat{x}\|_2^2}$. We can easily verify that $\delta A \cdot \hat{x} = -r$ and $\|\delta A\|_2 = \frac{\|r\|_2}{\|\hat{x}\|_2}$ □

Relative condition number

Let δA be a small componentwise relative perturbation in A and

$$|\delta A| \leq \varepsilon |A|, \quad |\delta b| \leq \varepsilon |b|. \tag{10}$$

Here, $|A|, |\delta A|$ are matrices of absolute values of $A, \delta A$, correspondingly. We use our perturbation equation

$$\delta x = A^{-1}(-\delta A \hat{x} + \delta b). \tag{11}$$

Now we use triangle inequality to get

$$\begin{aligned} |\delta x| &\leq |A^{-1}|(|\delta A| \cdot |\hat{x}| + |\delta b|) \\ &\leq |A^{-1}|(\varepsilon |A| \cdot |\hat{x}| + \varepsilon |b|) \\ &\leq \varepsilon(|A^{-1}|(|A| \cdot |\hat{x}| + |b|)). \end{aligned} \tag{12}$$

Using any vector norm (infinity, one-, Frobenious with $|| \, |z| \, || = ||z||$) we have

$$||\delta x|| \leq \varepsilon || \, |A^{-1}| \, \cdot (\, |A| \, \cdot |\hat{x}| + |b|)||. \tag{13}$$

If we assume that $\delta b = 0$ then the estimate above can be weakened to the bound

$$||\delta x|| \leq \varepsilon \| \ |A^{-1}| \cdot |A| \ \| \cdot \|\hat{x}\|. \tag{14}$$

or

$$\frac{||\delta x||}{||\hat{x}||} \leq \varepsilon \| \ \ |A^{-1}| \ \cdot \ |A| \ \|. \tag{15}$$

Then we define as $k_{CR}(A) \equiv || \ |A^{-1}| \cdot \ |A| \ ||$ the componentwise relative condition number of A also called as Bauer condition number or Skeel condition number.
Theorem about the distance from A to the nearest singular matrix is also valid for $k_{CR}(A)$.

Suppose that D is any nonsingular diagonal matrix and B is any nonsingular matrix such that $A = DB$. Then

$$\begin{aligned} k_{CR}(A) = k_{CR}(DB) &= \| \, |(DB)^{-1}| \cdot |DB| \, \| \\ &= \| \, |B^{-1}D^{-1}| \cdot |DB| \, \| \\ &= \| \, |B^{-1}| \cdot |B| \, \| = k_{CR}(B). \end{aligned} \tag{16}$$

The equation abobe means that if DB is badly scaled, i.e. B is well-conditioned and DB is badly conditioned then we still hope to get accurate solution for $Ax = (DB)x = b$.

Theorem
The smallest $\varepsilon > 0$ such that there exist $|\delta A| \leq \varepsilon |A|$ and $|\delta b| \leq \varepsilon |b|$ satisfying $(A + \delta A)\hat{x} = b + \delta b$ is called the componentwise relative backward error which can be expressed as

$$\varepsilon = \max_i \frac{|r_i|}{(|A| \cdot |\hat{x}| + |b|)_i} \tag{17}$$

where $r_i = (A\hat{x} - b)_i$.

Pivot element

- The pivot or pivot element is the element of a matrix, an array, or some other kind of finite set, which is selected first by an algorithm (e.g. Gaussian elimination, Quicksort, Simplex algorithm, etc.), to do certain calculations. In the case of matrix algorithms, a pivot entry is usually required to be at least distinct from zero, and often distant from it; in this case finding this element is called pivoting.
- Pivoting may be followed by an interchange of rows or columns to bring the pivot to a fixed position and allow the algorithm to proceed successfully, and possibly to reduce round-off error.

Examples of systems that require pivoting

In the case of Gaussian elimination, the algorithm requires that pivot elements not be zero. Interchanging rows or columns in the case of a zero pivot element is necessary. The system below requires the interchange of rows 2 and 3 to perform elimination.

$$\left[\begin{array}{ccc|c} 1 & -1 & 2 & 8 \\ 0 & 0 & -1 & -11 \\ 0 & 2 & -1 & -3 \end{array}\right]$$

The system that results from pivoting is as follows and will allow the elimination algorithm and backwards substitution to output the solution to the system.

$$\left[\begin{array}{ccc|c} 1 & -1 & 2 & 8 \\ 0 & 2 & -1 & -3 \\ 0 & 0 & -1 & -11 \end{array}\right]$$

Examples of systems that require pivoting

Furthermore, in Gaussian elimination it is generally desirable to choose a pivot element with large absolute value. This improves the numerical stability. The following system is dramatically affected by round-off error when Gaussian elimination and backwards substitution are performed.

$$\left[\begin{array}{cc|c} 0.00300 & 59.14 & 59.17 \\ 5.291 & -6.130 & 46.78 \end{array}\right]$$

This system has the exact solution of $x_1 = 10.00$ and $x_2 = 1.000$, but when the elimination algorithm and backwards substitution are performed using four-digit arithmetic, the small value of a_{11} causes small round-off errors to be propagated. The algorithm without pivoting yields the approximation of $x_1 \approx 9873.3$ and $x_2 \approx 4$.

Examples of systems that require pivoting

In this case it is desirable that we interchange the two rows so that a_{21} is in the pivot position

$$\left[\begin{array}{cc|c} 5.291 & -6.130 & 46.78 \\ 0.00300 & 59.14 & 59.17 \end{array}\right].$$

Considering this system, the elimination algorithm and backwards substitution using four-digit arithmetic yield the correct values $x1 = 10.00$ and $x2 = 1.000$.