Numerical Methods

Derek Li

Contents

1 Scientific Computing

1.1 Approximations in Scientific Computation

1.1.1 Absolute Error and Relative Error

Let A be approximate value, T be true value. Absolute error and relative error are defined as follows:

$$\text{Absolute Error} = A - T,$$
$$\text{Relative Error} = \frac{A-T}{T} \text{ assuming } T \neq 0.$$

If numbers written in scientific notation agree to p significant digits, then the magnitude of the relative error is about 10^{-p} (within a factor of 10).

Example 1.1. $A = 5.46729 \times 10^{-12}, T = 5.46417 \times 10^{-12}$. Thus, $A - T = 0.00312 \times 10^{-12}$ and $\frac{A-T}{T} = \frac{3.12\times10^{-3}}{5.46417}$, i.e., the relative error around 10^{-3}.

Example 1.2. $A = 1.00596 \times 10^{-10}, T = 0.99452 \times 10^{-10}$. Thus, $A - T = 0.01144 \times 10^{-10}$ and $\frac{A-T}{T} = \frac{1.144\times10^{-2}}{0.99452}$. A and T agree to 2 significant digits.

1.1.2 Data Error and Computational Error

The difference between exact function values due to error in the input and thus can be viewed as data error.

The difference between the exact and approximate functions for the same input and thus can be considered computational error.

1.1.3 Truncation Error

Truncation error is the difference between the true result (for the actual input) and the result that would be produced by a given algorithm using exact arithmetic. It is due to approximations such as truncating an infinite series, replacing derivatives by finite differences, or terminating an iterative sequence before convergence.

Example 1.3. $f(x) = \sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots, \widehat{f}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}$. The part $-\frac{x^7}{7!} + \cdots$ is the truncation error.

1.1.4 Forward Error and Backward Error

Suppose we want to compute the value of a function, $y = f(x)$, but we obtain instead an approximate value $\widehat{y}$. The discrepancy between the computed and true values, $\Delta y = \widehat{y} - y$, is called the forward error.

The quantity $\Delta x = \widehat{x} - x$, where $f(\widehat{x}) = \widehat{y}$, is called the backward error.

Example 1.4. As an approximation to $y = \sqrt{2}$, the value $\widehat{y} = 1.4$ has a forward error

$$\Delta y = \widehat{y} - y = 1.4 - \sqrt{2} \approx -0.0142$$

or a relative forward error -1.004×10^{-2}. We find that $\sqrt{1.96} = 1.4$, so the backward error is

$$\Delta x = \widehat{x} - x = 1.96 - 2 = -0.04$$

or the relative backward error -2×10^{-2}.

1.1.5 Conditioning

A problem is well-conditioned if a small change to the input produces a small change to the output; ill conditioned if there are some examples for which a small change in the input produces a large change in output.

Consider relative forward error $\frac{\widehat{y}-y}{y} = \frac{f(\widehat{x})-f(x)}{f(x)}$. Since

$$f(\widehat{x}) - f(x) = f'(\widetilde{x})(\widehat{x} - x)$$

for some $\widetilde{x}$ between x and $\widehat{x}$ provided $f'(x)$ exists and is continuous between x and $\widehat{x}$, then

$$\frac{\widehat{y} - y}{y} = \frac{xf'(\widetilde{x})}{f(x)} \frac{\widehat{x} - x}{x} \approx \frac{xf'(x)}{f(x)} \frac{\widehat{x} - x}{x}.$$

$\frac{xf'(x)}{f(x)}$ is called condition number.

Example 1.5. The conditioning of $f(x) = \sqrt{x}, x \geqslant 0$.

Solution. We have $f'(x) = \frac{1}{2\sqrt{x}}$ and thus the condition number is

$$\frac{x}{2\sqrt{x}\sqrt{x}} = \frac{1}{2}.$$

Hence, $f(x)$ is well-conditioned.

Example 1.6. The conditioning of $f(x) = e^x$.

Solution. The condition number is x. Thus, e^x overflows or underflows if $|x|$ is large.

Example 1.7. The conditioning of $f(x) = \sin x$.

Solution. The condition number is $\frac{x \cos x}{\sin x}$.

(1) If $x \approx \pm\pi, \pm 2\pi, \pm 3\pi, \cdots, \sin x$ overflows or underflows. Nevertheless, x could be 0 because

$$\left.\frac{\sin(x)}{x}\right|_{x=0} = \left.\frac{x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots}{x}\right|_{x=0} = 1 - \left.\frac{x^2}{3!} + \frac{x^4}{5!} - \cdots\right|_{x=0} = 1.$$

(2) If $|x|$ is big and $\cos x \not\approx 0, \sin x$ overflows or underflows.

1.1.6 Stability of Algorithms

An algorithm is stable if small changes to the input result in small changes to the output. Note that stability of an algorithm does not by itself guarantee that the computed result is accurate.

1.2 Computer Arithmetic

1.2.1 Floating-Point Numbers

In a digital computer, the real number system $\mathbb{R}$ of mathematics is represented by a floating-point number system.

Formally, a floating-point number system $\mathbb{F}$ is characterized by four integers: β (base), p (precision), $[L, U]$ (exponent range). Any floating-point number $x \in \mathbb{F}$ has the form

$$x = \pm d_1.d_2d_3\cdots d_p \times \beta^n,$$

where $0 \leqslant d_i < \beta, L \leqslant n \leqslant U$.

1.2.2 IEEE Floating-Point Standard

System	β	p	L	U
Single-Precision	2	24	-126	127
Double-Precision	2	53	-1022	1023

1.2.3 Normalization

A floating-point system is said to be normalized if the leading digit $d_1 \neq 0$ unless the number represented is zero.

1.2.4 Properties of Floating-Point Systems

There is a smallest positive normalized floating-point number, underflow limit

$$\mathrm{UFL} = \underbrace{1.00\cdots0}_{p \text{ digits}} \times \beta^L.$$

There is a largest floating-point number, overflow limit

$$\mathrm{OFL} = \underbrace{d.dd\cdots d}_{p \text{ digits}} \times \beta^U,$$

where $d = \beta - 1$ and thus

$$\mathrm{OFL} = (\beta - \beta^{1-p})\beta^U = (1 - \beta^{-p})\beta^{U+1}.$$

1.2.5 Rounding

One of the commonly used rounding rules is round-to-nearest which is the default rounding rule in IEEE standard system.

Note 1. Most of the time, there is a unique closest p-digit number.

Note 2. In case of times, we round to the number with an even least significant digit (round to even).

Note 3. In binary, in case of times, we round to the number which has a 0 in its least significant digits.

Example 1.8. Consider a system with $\beta = 10, p = 3, L = -10, U = 10$.

$$1.54 \times 10^1 + 2.56 \times 10^{-1} = 1.5656 \times 10^1 \approx 1.57 \times 10^1.$$

1.2.6 Machine Epsilon

The accuracy of a floating-point system can be characterized by machine epsilon. We define machine epsilon

$$\varepsilon_{\text{mach}} = \beta^{1-p}.$$

There is same number of floating-point numbers in each interval $[\beta^n, \beta^{n+1})$, and numbers are evenly spaced with distance $\beta^n \cdot \varepsilon_{\text{mach}}$.

The machine epsilon is important because it bounds the relative error in representing any nonzero real number x within the normalized range of a floating-point system:

$$|\text{fl}(x) - x| \leqslant \frac{1}{2}\beta^n \varepsilon_{\text{mach}} \Rightarrow \left|\frac{\text{fl}(x) - x}{x}\right| \leqslant \frac{\frac{1}{2}\beta^n \varepsilon_{\text{mach}}}{\beta^n} = \frac{1}{2}\varepsilon_{\text{mach}}.$$

In IEEE, $\text{fl}(a \text{ op } b)$ is closest flouting-point number to a op b (round to nearest and no overflows or underflows), where op means basic operations: $+, -, \times, /, \sqrt{\cdots}$. We also have

$$\left|\frac{\text{fl}(a \text{ op } b) - (a \text{ op } b)}{a \text{ op } b}\right| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}.$$

1.2.7 Subnormals and Gradual Underflow

There is a noticeable gap around zero in floating-point system because of normalization. Subnormal numbers have $d_1 = 0$, which can fill in the gap, but the inequality

$$\left|\frac{\text{fl}(x) - x}{x}\right| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$$

might not hold.

Such an augmented floating-point system is sometimes said to exhibit gradual underflow, since it extends the lower range of magnitudes representable rather than underflowing to zero as soon as the minimum exponent value would otherwise be exceeded.

Example 1.9 (Underflow to a Subnormal)**.** Consider a system with $\beta = 10, p = 3, L = -10, U = 10$.

$$1.01 \times 10^{-5} \times 2.02 \times 10^{-6} = 2.0402 \times 10^{-11} = 0.20402 \times 10^{-10} \approx 0.20 \times 10^{-10}.$$

Example 1.10 (Underflow to Zero)**.**

$$1.01 \times 10^{-6} \times 2.02 \times 10^{-7} = 2.0402 \times 10^{-13} = 0.0020402 \times 10^{-10} \approx 0.00 \times 10^{-10}.$$

1.2.8 Exceptional Values

The IEEE floating-point standard provides two additional special values to indicate exceptional situations:

- Inf, which stands for infinity, results from dividing a finite number by zero or Inf + Inf.
- NaN, which stands for not a number, results from an undefined or indeterminate operation such as $\frac{0}{0}, 0 \times \mathrm{Inf}, \frac{\mathrm{Inf}}{\mathrm{Inf}}$, or Inf − Inf.

Example 1.11 (Overflow)**.**

$$3.56 \times 10^5 \times 5.41 \times 10^5 = 19.2596 \times 10^{10} \to \mathrm{Inf}.$$

1.2.9 Floating-Point Arithmetic

IEEE Standard guarantees that if a and b are IEEE floating-point numbers, then $\mathrm{fl}(a \text{ op } b)$ is the correctly rounded value of a op b (could be $\pm$inf, NaN, underflow).

If a and b are normalized floating-point numbers and if no overflows or underflows occur in computing a op b, then

$$\mathrm{fl}(a \text{ op } b) = (a \text{ op } b)(1+\delta),$$

where $|\delta| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}, \delta = \frac{\mathrm{fl}(a \text{ op } b) - (a \text{ op } b)}{a \text{ op } b}$ is the relative error.

If there is an underflow, we may get $|\delta| > \frac{1}{2}\varepsilon_{\text{mach}}$.

Example 1.12. Suppose $L = -10, p = 3$ approximate $\mathrm{fl}(2.02 \times 10^{-6} \times 1.11 \times 10^{-6})$.

The exact value is $2.2422 \times 10^{-12} = 0.022422 \times 10^{-10} \approx 0.02 \times 10^{-10}$. We know $\delta \approx -0.108$, but $\frac{1}{2}\varepsilon_{\text{mach}} = 0.005$.

1.2.10 Cancellation

Here is an example of catastrophic cancellation.

Example 1.13. Suppose $p = 3, \mathrm{fl}(1.00 \times 10^3 + 1.00 \times 10^7 - 1.00 \times 10^7)$.

First step: $\mathrm{fl}(1.0000 \times 10^7 + 0.0001 \times 10^7) = \mathrm{fl}(1.0001 \times 10^7) \to 1.00 \times 10^7$. (Note: it is not a underflow.)

Second step: $\mathrm{fl}(1.00 \times 10^7 - 1.00 \times 10^7) = 0$.

If you compute a sum and some of the intermediate values are much larger in magnitude than final result, then the relative error in the computed sum may be very inaccurate.

One exception in sum: suppose $a \geqslant 0, b \geqslant 0, c \geqslant 0$ (or $a \leqslant 0, b \leqslant 0, c \leqslant 0$) and assume no overflows or underflows.

$$\begin{aligned}\mathrm{fl}((a+b)+c) &= [(a+b)(1+\delta_1)+c](1+\delta_2) \\ &:= (a+b+c)(1+\widehat{\delta}_1)(1+\delta_2) := (a+b+c)(1+\widetilde{\delta}),\end{aligned}$$

where $|\delta_1|, |\delta_2| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$. We can show that $|\widehat{\delta}_1| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$ and $|\widetilde{\delta}| \leqslant 1.01\varepsilon_{\text{mach}}$.

Proof. We have

$$(a+b)(1+\delta_1)+c \leqslant (a+b)(1+\frac{1}{2}\varepsilon_{\text{mach}})+c \leqslant (a+b+c)(1+\frac{1}{2}\varepsilon_{\text{mach}}).$$

Similarly,

$$(a+b+c)(1-\frac{1}{2}\varepsilon_{\text{mach}}) \leqslant (a+b)(1+\delta_1)+c := (a+b+c)(1+\widehat{\delta}_1).$$

Therefore, $|\widehat{\delta}_1| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$.

Hence,

$$\begin{aligned}|\widetilde{\delta}| &\leqslant |\widehat{\delta}_1|+|\delta_2|+|\widehat{\delta}_1\delta_2| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}+\frac{1}{2}\varepsilon_{\text{mach}}+\frac{1}{4}\varepsilon_{\text{mach}}^2\\ &= (1+\frac{1}{4}\varepsilon_{\text{mach}})\varepsilon_{\text{mach}} \leqslant 1.01\varepsilon_{\text{mach}},\end{aligned}$$

when p is large. □

In multiplication, the situation could be better. Assume no overflows or underflows.

$$\begin{aligned}\mathrm{fl}((a*b)*c) &= [(a*b)(1+\delta_1)]*c(1+\delta_2) = (a*b*c)(1+\delta_1)(1+\delta_2)\\ &= (a*b*c)(1+\widehat{\delta}),\end{aligned}$$

where $|\delta_1|, |\delta_2| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}, |\widehat{\delta}| \leqslant 1.01\varepsilon_{\text{mach}}$.

Example 1.14 (Catastrophic Cancellation)**.** $f(x)=\sqrt{1+x^2}-1$.

(1) $f(x)$ is well conditioned because

$$\frac{xf'(x)}{f(x)} = \frac{x^2}{\sqrt{1+x^2}(\sqrt{1+x^2}-1)} = 1+\frac{1}{\sqrt{1+x^2}} \in [1,2].$$

(2) $\mathrm{fl}(\sqrt{1+x^2}-1)$ does not always give an accurate result in the relative error sense. If x is small enough, $\mathrm{fl}(\sqrt{1+x^2}-1) \to 0$. Recall that if $A=0, T\neq 0, \frac{A-T}{T}=-1$, which is a bad relative error because there is no same digit.

(3) We could change the function to a mathematically equivalent value that has a much smaller floating-point error: $\frac{x^2}{\sqrt{1+x^2}+1}$.

Proof. Want to show $\mathrm{fl}\left(\frac{x^2}{\sqrt{1+x^2}+1}\right) = \frac{x^2}{\sqrt{1+x^2}+1}\left(1+\widetilde{\delta}\right)$. We have

$$
\begin{aligned}
\mathrm{fl}\left(\frac{x^2}{\sqrt{1+x^2}+1}\right) &= \frac{x^2(1+\delta_1)(1+\delta_5)}{\left[\sqrt{[1+x^2(1+\delta_1)](1+\delta_2)}(1+\delta_3)+1\right](1+\delta_4)} \\
&:= \frac{x^2(1+\delta_1)(1+\delta_5)}{\left[\sqrt{(1+x^2)(1+\widehat{\delta}_1)(1+\delta_2)}(1+\delta_3)+1\right](1+\delta_4)} \\
&:= \frac{x^2(1+\delta_1)(1+\delta_5)}{\left(\sqrt{1+x^2}(1+\widetilde{\delta}_1)(1+\widetilde{\delta}_2)(1+\delta_3)+1\right)(1+\delta_4)} \\
&:= \frac{x^2(1+\delta_1)(1+\delta_5)}{(\sqrt{1+x^2}+1)(1+\widetilde{\delta}_1)(1+\widetilde{\delta}_2)(1+\widetilde{\delta}_3)(1+\delta_4)} \\
&:= \frac{x^2}{\sqrt{1+x^2}+1}(1+\delta_1)(1+\delta_5)(1+\widehat{\widehat{\delta}}_1)(1+\widehat{\widehat{\delta}}_2)(1+\widehat{\widehat{\delta}}_3)(1+\widehat{\widehat{\delta}}_4) \\
&:= \frac{x^2}{\sqrt{1+x^2}+1}(1+\widetilde{\delta}),
\end{aligned}
$$

where $|\delta_i| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$.

Note that (1) $\left|\widehat{\delta}_1\right| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$ since $\widehat{\delta}_1 = \frac{x^2}{1+x^2}\delta_1 \leqslant \frac{x^2}{2(1+x^2)}\varepsilon_{\text{mach}} \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$.

(2) Suppose $|\delta| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$, we have

$$
1-\frac{1}{2}\varepsilon_{\text{mach}} \leqslant \sqrt{1-\frac{1}{2}\varepsilon_{\text{mach}}} \leqslant \sqrt{1+\delta} \leqslant \sqrt{1+\frac{1}{2}\varepsilon_{\text{mach}}} \leqslant 1+\frac{1}{2}\varepsilon_{\text{mach}},
$$

provided $\frac{1}{2}\varepsilon_{\text{mach}} \leqslant 2$. Therefore, $\exists\widehat{\delta}$ s.t. $\sqrt{1+\delta} = 1+\widehat{\delta}$ for $\left|\widehat{\delta}\right| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$.

(3) Suppose $|\delta| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$, we have

$$
\frac{1}{1+\delta} \leqslant \frac{1}{1-\frac{1}{2}\varepsilon_{\text{mach}}} = \frac{2}{2-\varepsilon_{\text{mach}}} = 1+\frac{1}{2-\varepsilon_{\text{mach}}}\varepsilon_{\text{mach}} \leqslant 1+\frac{1.01}{2}\varepsilon_{\text{mach}},
$$

for $\varepsilon_{\text{mach}} \leqslant 0.02$. Similarly,

$$
1-\frac{1.01}{2}\varepsilon_{\text{mach}} \leqslant \frac{1}{1+\delta}
$$

and thus $\left|\widehat{\delta}\right| \leqslant \frac{1.01}{2}\varepsilon_{\text{mach}}$.

Therefore, we have $|\delta_1|, |\delta_5| \leqslant \frac{1}{2}\varepsilon_{\text{mach}}$, $\left|\widehat{\widehat{\delta}}_i\right| \leqslant \frac{1.01}{2}\varepsilon_{\text{mach}}$, and $\left|\widetilde{\delta}\right|$ is kind of small. $\square$

2 Systems of Linear Equations

2.1 Review: Existence and Uniqueness

An $n \times n$ matrix A is said to be nonsingular if it satisfies any one of the following equivalent conditions:

- A has an inverse, i.e., $\exists A^{-1}$, an $n \times n$ matrix, s.t. $AA^{-1} = A^{-1}A = I$.
- $\det(A) \neq 0$.
- $\operatorname{rank}(A) = n$ (recall that the rank of a matrix is the maximum number of linearly independent rows or columns it contains).
- $\forall \mathbf{z} \neq \mathbf{0}, A\mathbf{z} \neq \mathbf{0}$.

Otherwise, the matrix is singular.

2.2 Conditioning

The form is

$$A\mathbf{x} = \mathbf{b},$$

where A and $\mathbf{b}$ are input and $\mathbf{x}$ is output.

2.2.1 Vector Norm

Definition 2.1 (p-Norm)**.** $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, p \in \mathbb{N}$.

Definition 2.2 (Manhattan Norm)**.** $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$.

Definition 2.3 (Euclidean Norm)**.** $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$.

Definition 2.4 (∞-Norm)**.** $\|\mathbf{x}\| = \max_{1 \leqslant i \leqslant n} |x_i|$.

Example 2.1. Let $\mathbf{x} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$, then we have $\|\mathbf{x}\|_1 = 1 + 2 = 3, \|\mathbf{x}\|_2 = \sqrt{1+4} = \sqrt{5}, \|\mathbf{x}\|_\infty = \max(1, 2) = 2$.

Theorem 2.1. $\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^n$.

Proof. Suppose $\max_{1 \leqslant i \leqslant n} |x_i| = |x_k|$, then

$$\|\mathbf{x}\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| = |x_k| = \sqrt{x_k^2} \leqslant \sqrt{x_1^2 + \cdots + x_k^2 + \cdots + x_n^2} = \|\mathbf{x}\|_2.$$

Besides,

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \leqslant (\sqrt{x_1})^2 + \cdots + (\sqrt{n})^2 = \|\mathbf{x}_1\|.$$

$\square$

Theorem 2.2. $\|\mathbf{x}\|_2 \leqslant \sqrt{n}\|\mathbf{x}\|_\infty$.

Proof. Since

$$\|\mathbf{x}\|_2^2 = x_1^2 + \cdots + x_n^2 \leqslant n \max_{1\leqslant i\leqslant n} |x_i|^2 = n\left(\max_{1\leqslant i\leqslant n} |x_i|\right)^2 = n\|\mathbf{x}\|_\infty^2,$$

then $\|\mathbf{x}\|_2 \leqslant \sqrt{n}\|\mathbf{x}\|_\infty$. □

Property 2.1 (Properties of Norms)**.**

- $\|\mathbf{x}\| > 0$ for $\mathbf{x} \neq \mathbf{0}$.
- $\|\gamma\mathbf{x}\| = |\gamma| \cdot \|\mathbf{x}\|$ for $\gamma \in \mathbb{R}$.
- $\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.

Corollary 2.1. $\|\mathbf{0}\| = 0$ and $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leqslant \|\mathbf{x}-\mathbf{y}\|$.

Proof. Suppose $\mathbf{x} \neq \mathbf{0}$, then $\|\mathbf{0}\| = \|0\mathbf{x}\| = |0| \cdot \|\mathbf{x}\| = 0$.

Besides, $\|(\mathbf{x}-\mathbf{y}) + \mathbf{y}\| \leqslant \|\mathbf{x}-\mathbf{y}\| + \mathbf{y} \Rightarrow \|\mathbf{x}\| - \|\mathbf{y}\| \leqslant \|\mathbf{x}-\mathbf{y}\|$. Similarly, we can prove $\|\mathbf{y}\| - \|\mathbf{x}\| \leqslant \|\mathbf{x}-\mathbf{y}\|$. Wherefore, $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leqslant \|\mathbf{x}-\mathbf{y}\|$. □

2.2.2 Matrix Norm

Definition 2.5 (Matrix Norm)**.** $\|A\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$ and such a matrix norm is said to be induced by or subordinate to the vector form.

Definition 2.6. $\|A\|_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$, which is the maximum column sum.

Definition 2.7. $\|A\|_\infty = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|$, which is the maximum row sum.

Definition 2.8. $\|A\|_2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^TA\}$.

Definition 2.9 (Frobenius Norm)**.** $\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}$.

Property 2.2 (Properties of Matrix)**.**

- $\|A\| > 0$ if $A \neq O$.
- $\|\gamma A\| = |\gamma| \cdot \|A\|$ for $\gamma \in \mathbb{R}$.
- $\|A+B\| \leqslant \|A\| + \|B\|$.
- $\|AB\| \leqslant \|A\| \cdot \|B\|$.
- $\|A\mathbf{x}\| \leqslant \|A\| \cdot \|\mathbf{x}\|$.

Property 2.3. $\|A\| = \max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$.

Proof. We have

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{x}\neq\mathbf{0}} \left\|\frac{1}{\mathbf{x}}A\mathbf{x}\right\| = \max_{\mathbf{x}\neq\mathbf{0}} \left\|A\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\right\| = \max_{\mathbf{y}=\frac{\mathbf{x}}{\|\mathbf{x}\|},\mathbf{x}\neq\mathbf{0}} \|A\mathbf{y}\| = \max_{\|\mathbf{y}\|} \|A\mathbf{y}\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

□

Definition 2.10 (Norm Equivalence)**.** Two norms are said to be equivalent if $\exists c_1, c_2 > 0$ s.t. $c_1\|\mathbf{x}\|_a \leqslant \|\mathbf{x}\|_b \leqslant c_2\|\mathbf{x}\|_a, \forall \mathbf{x} \in \mathbb{R}^n$.

Note that c_1, c_2 can depend on n but not $\mathbf{x}$.

2.2.3 Matrix Condition Number

Assume A is nonsingular $n \times n$ matrix and $A\mathbf{x} = \mathbf{b}$. Suppose we change $\mathbf{b}$ but not A, then

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}},$$

with $\Delta\mathbf{b} = \mathbf{b} - \hat{\mathbf{b}}, \Delta\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}}$.

Since $A\mathbf{x} = \mathbf{b}$, then $\|\mathbf{b}\| = \|A\mathbf{x}\| \leqslant \|A\| \cdot \|\mathbf{x}\|$, if $\|A\| = \max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$, which is because

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\| \Rightarrow \|A\mathbf{x}\| \leqslant \|A\| \cdot \|\mathbf{x}\|.$$

Also,

$$A\Delta\mathbf{x} = A(\mathbf{x} - \hat{\mathbf{x}}) = A\mathbf{x} - A\hat{\mathbf{x}} = \mathbf{b} - \hat{\mathbf{b}} = \Delta\mathbf{b} \Rightarrow \Delta\mathbf{x} = A^{-1}\Delta\mathbf{b},$$

and wherefore

$$\|\Delta\mathbf{x}\| = \|A^{-1}\Delta\mathbf{b}\| \leqslant \|A^{-1}\| \cdot \|\Delta\mathbf{b}\|.$$

As a consequence,

$$\frac{\|\Delta\mathbf{x}\|}{\|A\| \cdot \|\mathbf{x}\|} \leqslant \frac{\|A^{-1}\| \cdot \|\Delta\mathbf{b}\|}{\|A\| \cdot \|\mathbf{x}\|} \leqslant \frac{\|A^{-1}\| \cdot \|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \Leftrightarrow \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \|A\| \cdot \|A^{-1}\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Let $\mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|$, which is the matrix condition number.

Property 2.4. $\mathrm{cond}(A) \geqslant 1$.

Proof. We have $1 = \|I\| = \|A^{-1}A\| \leqslant \|A^{-1}\| \cdot \|A\| = \mathrm{cond}(A)$. □

Property 2.5. $\mathrm{cond}(I) = 1$.

Proof. $\mathrm{cond}(I) = \|I\| \cdot \|I^{-1}\| = \|I\| \cdot \|I\| = 1$. □

Property 2.6. $\mathrm{cond}(\gamma A) = A, \gamma \in \mathbb{R}\backslash\{0\}$.

Proof. $\mathrm{cond}(\gamma A) = \|\gamma A\| \cdot \|(\gamma A)^{-1}\| = \|\gamma A\| \cdot \|\gamma^{-1}A^{-1}\| = |\gamma\gamma^{-1}|\|A\| \cdot \|A^{-1}\| = \|A\| \cdot \|A^{-1}\|$. □

Property 2.7. $\operatorname{cond}(A) = \dfrac{\max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}}{\min\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}} = \dfrac{\max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}{\min\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}$.

Proof. We have $\operatorname{cond}(A) = \|A\| \cdot \|A^{-1}\| = \max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A^{-1}\mathbf{x}\|}{\|\mathbf{x}\|}$.

Besides, $\|A^{-1}\| = \max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A^{-1}\mathbf{x}\|}{\|\mathbf{x}\|} := \max\limits_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{y}\|}{\|A\mathbf{y}\|} = \max\limits_{\mathbf{y}\neq\mathbf{0}} \frac{1}{\|A\mathbf{y}\|/\|\mathbf{y}\|} = \frac{1}{\min\limits_{\mathbf{y}\neq\mathbf{0}} \frac{\|A\mathbf{y}\|}{\|\mathbf{y}\|}} = \frac{1}{\min\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}}$.

Therefore,

$$\operatorname{cond}(A) = \frac{\max\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}}{\min\limits_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}} = \frac{\max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}{\min\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|},$$

which means $\frac{\text{How much you can expand}}{\text{How much you can contract}}$. ☐

Property 2.8. For any p-norm, if D is a diagonal matrix and all $d_i \neq 0$, then

$$\operatorname{cond}_\infty(D) = \frac{\max\limits_{1\leqslant i\leqslant n} |d_i|}{\min\limits_{1\leqslant i\leqslant n} |d_i|}.$$

Proof. Since $D = \begin{pmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \cdots & d_n \end{pmatrix}$ is a diagonal matrix, then

$$D^{-1} = \begin{pmatrix} d_1^{-1} & 0 & 0 & \cdots & 0 \\ 0 & d_2^{-1} & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \cdots & d_n^{-1} \end{pmatrix}.$$

Therefore,

$$\operatorname{cond}(D)_\infty = \|D\|_\infty \cdot \|D^{-1}\|_\infty = \frac{\max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|_\infty}{\min\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|_\infty} = \frac{\max\limits_{1\leqslant i\leqslant n} |d_i|}{\min\limits_{1\leqslant i\leqslant n} |d_i|}.$$

☐

If we change both A and $\mathbf{b}$, then

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \operatorname{cond}(A) \left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

Notice that, if $\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \approx \varepsilon_{\text{mach}}$, $\frac{\|\Delta A\|}{\|A\|} \approx \varepsilon_{\text{mach}}$, and thus $\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \operatorname{cond}(A)\varepsilon_{\text{mach}}$.

Example 2.2. $A = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$.

Solution. Recall that $A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, so $AA^{-1} = \frac{1}{\det(A)} \begin{pmatrix} ad - bc & -ab + ab \\ cd - cd & -bc + ad \end{pmatrix}$.

Hence, $A^{-1} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$, and $\operatorname{cond}_\infty(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty = 4 \times 5 = 20$.

Definition 2.11. A matrix A is close to singular means if you change some elements of A by a little bit, you will get a singular matrix, i.e., $\exists E$ s.t. $\frac{\|E\|}{\|A\|} \propto \frac{1}{\text{cond}(A)}$ and $A + E$ is singular.

Property 2.9. Suppose A is close to singular, γA for $\gamma \in \mathbb{R}\backslash\{0\}$ is close to singular.

Proof. Suppose $A + E$ is singular and $\frac{\|E\|}{\|A\|}$ is small. We have $\gamma A + \gamma E = \gamma(A + E)$, and

$$\frac{\|\gamma E\|}{\|\gamma A\|} = \frac{|\gamma|\|E\|}{|\gamma|\|A\|} = \frac{\|E\|}{\|A\|}.$$

$\square$

Note that the determinant does not tell you if a matrix is close to singular, and the counter example is I and γI.

Suppose A is $n \times n$ matrix, and the standard way of computing A^{-1} requires about n^3 arithmetic operators so that we want to estimate $\text{cond}(A)$ without computing A^{-1}. We have

$$\mathbf{y} = A\mathbf{z} \Rightarrow \mathbf{z} = A^{-1}\mathbf{y} \Rightarrow \frac{\|\mathbf{z}\|}{\|\mathbf{y}\|} \leqslant \|A^{-1}\|.$$

We iterate choosing sequence of $\mathbf{z}$'s to try to make $\frac{\|\mathbf{z}\|}{\|\mathbf{y}\|}$ as large as possible.

2.2.4 Residual

Suppose $A\mathbf{x} = \mathbf{b}$, and we get computed solution $\widehat{\mathbf{x}}$. The residual is

$$\mathbf{r} = \mathbf{b} - A\widehat{\mathbf{x}} \Rightarrow A\widehat{\mathbf{x}} = \mathbf{b} - \mathbf{r} = \widehat{\mathbf{b}}.$$

Therefore, $\Delta\mathbf{b} = \mathbf{b} - \widehat{\mathbf{b}} = \mathbf{r}$. Let $\Delta\mathbf{x} = \mathbf{x} - \widehat{\mathbf{x}}$.

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \text{cond}(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} = \text{cond}(A)\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

Hence, nearly all good linear equation solvers give a small $\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$.

2.3 Solving Linear Systems

2.3.1 Back Solve

It is easy to solve $U\mathbf{x} = \widehat{\mathbf{b}}$, where U is upper triangular matrix.

Start at bottom equation:

$$u_{nn}x_n = \widehat{b}_n \Rightarrow x_n = \frac{\widehat{b}_n}{u_{nn}}.$$

Next, consider second to last equation:

$$u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n = \widehat{b}_{n-1} \Rightarrow x_{n-1} = \frac{\widehat{b}_{n-1} - u_{n-1,n}}{x}{}_n u_{n-1,n-1}.$$

Keep the procedure and we will solve every x_i.

Total work for back solve is n dim $+ \sum_{k=1}^{n}(k-1)(A+M) = n$ dim $+ \frac{n(n-1)}{2}(A+M)$.

2.3.2 Gaussian Elimination and LU Factorization

Theorem 2.3. $\mathbf{x}$ is a solution of $A\mathbf{x} = \mathbf{b} \Leftrightarrow \mathbf{x}$ is a solution of $M_{n-1} \cdots M_2 M_1 \mathbf{x} = M_{n-1} \cdots M_2 M_1 \mathbf{b}$, where M_i's are nonsingular.