Machine Epsilon

Consider the discrete subset **F** of real numbers **R** to be our floating point number system depending on the precision we choose then for all $x \in \mathbf{R}$, there exists $x' \in \mathbf{F}$ such that

$$\frac{|x - x'|}{|x|} \leq \epsilon_{machine} \tag{1}$$

$f : \mathbb{R} \to \mathbb{F}$ is a round off function then

$$f'(x) = x(1 + \epsilon) \tag{2}$$

i.e., there exists an ϵ with $|\epsilon| \leq \epsilon_{machine}$ such that equation 2 is true.

$$\begin{aligned} \text{In single precision: } \epsilon_{machine} &\approx \frac{2^{-23}}{2} \approx 5.96 \times 10^{-8} \\ \text{In double precision: } \epsilon_{machine} &\approx \frac{2^{-52}}{2} \approx 1.11 \times 10^{-16} \end{aligned} \tag{3}$$

Fundamental operations in floating point arithmetic

Classical arithmetic operations are $+, -, \times$ and $/$
One a computer, we have analogous operations on $\mathbb{F}$, denote these operations ——————————Missing————-

Let us consider $x, y \in \mathbb{F}$ and $*$ denotes the arithmetic operations, there exists ϵ with $|\epsilon| \leq \epsilon_{machine}$ such that

$$x * y = f'(x * y) = (x * y)(1 + \epsilon) \tag{4}$$

i.e., every operation of floating point arithmetic is exact upto a relative error ϵ of size atleast ϵ

Conditioning and Stability

Conditioning: sensitivity of a mathematical problem to perturbations in input

y = f(x),

- x $\to$ input to the problem (data)
- f $\to$ represents the problem
- y $\to$ represents a solution

- What happens to **y** when the given input **x** is perturbed slightly?

1 Absolute contiion number

If the small perturbation in **x** is denoted by δx then let the resulting perturbation in the solution be represented as $\delta\mathbf{f}$ i.e., δf = f(x + δx) - f(x). Then the absolute condition number $k' = k(x)$ of the problem **f** at **x** is given by

$$\mathbf{K}(x) = \max_{\delta x} \left(\frac{||\delta \mathbf{f}||}{||\delta \mathbf{x}||} \right) \tag{5}$$

for infinitesimally small δf and δx

If **f** has a derivative, we can evaluate Jacobian matrix J(x) as $J_{ij} = \frac{\delta \mathbf{f}_i}{\delta x_j}$

We have $\delta\mathbf{f} \approx J(x)\delta\mathbf{x}$, equality $||\delta\mathbf{x}|| \to 0$

$$\begin{aligned} K(x) &= \max_{\delta x} \frac{||J(x)\delta \mathbf{x}||}{||\delta \mathbf{x}||} \\ K(x) &= ||J(x)|| \end{aligned} \tag{6}$$

2 Relative Condition Number

Assume δx is infinitesimal

$$K^2 = \max_{\delta x} \left(\frac{\frac{||\delta \mathbf{f}||}{||\mathbf{f}||}}{\frac{||\delta \mathbf{x}||}{||\mathbf{x}||}} \right) = \max_{\delta x} \left(\frac{\frac{||\delta \mathbf{f}||}{||\delta \mathbf{x}||}}{\frac{||\mathbf{f}(x)||}{||\mathbf{x}||}} \right) = \frac{||J(x)||}{||\frac{f(x)}{x}||} \tag{7}$$

Examples:

1. f(x) = x/2 , $x \in \mathbb{R}$
   Input: x, Output: x/2, J = $\frac{df}{dx}$ = 1/2

$$K' = \frac{||J||}{\frac{||f(x)||}{||x||}} = \frac{1/2}{|\frac{x/2}{x}|}$$

2. $f(x) = x_1 - x_2$, where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$J = \begin{pmatrix} \frac{\delta \mathbf{f}}{\delta \mathbf{x}_1} & \frac{\delta \mathbf{f}}{\delta \mathbf{x}_2} \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix}$$

$$K^R = \frac{||J||_\infty}{||\mathbf{f}(x)||_\infty / ||x||_\infty} = \frac{2}{\frac{||f(x)||}{\max\{|x_1|,|x_2|\}}} = \frac{2\max\{|x_1|, |x_2|\}}{|x_1 - x_2|}$$

if $|x_1 - x_2|$ is small ≈ 0, K^R is large and is not well conditioned

————————Missing—

$$K^R = \frac{||J||_1}{||\mathbf{f}(x)||_1 / ||x||_1} = \frac{2}{\frac{||f(x)||}{\max\{|x_1|,|x_2|\}}} = \frac{2\max\{|x_1|, |x_2|\}}{|x_1 - x_2|}$$

Eigen values of a matrix

Input: A
Output: Eigenvalues λ of $\underset{\sim}{A}$

Consider a symmetric matrix $\mathbf{A} = \mathbf{A}^T$ λ and $\lambda + \delta\lambda$ are corresponding eigen values of $\mathbf{A}$ and $\mathbf{A} + \delta\mathbf{A}$ then

$$|\delta\lambda| \leq ||\delta\mathbf{A}||_2 \tag{8}$$

Relative condition number:

$$\begin{aligned} &= \max_{\delta\mathbf{A}} \frac{\left(\frac{|\delta\lambda|}{|\lambda|}\right)}{\frac{||\delta\mathbf{A}||_2}{||\mathbf{A}||}} = \max_{\delta\mathbf{A}} \frac{|\delta\lambda|}{|\lambda|}.\frac{||\mathbf{A}||}{||\delta\mathbf{A}||} \\ \text{Relative condition number}(k) &= \frac{||A||_2}{|\lambda|} \end{aligned} \tag{9}$$

3 Conditioning of a matrix-vector multiplicatoin

Fixed A: Input: x, Output: Ax y = Ax

$$\hat{K} = \frac{||A||.||x||}{||Ax||}$$

A is non singular

$$\begin{aligned} => x = AA^{-1}x \implies ||x|| &= ||AA^{-1}x|| \\ &= ||A^{-1}Ax|| \\ &\leq ||A^{-1}||.||Ax|| \end{aligned}$$

Compute $A^{-1}b$ for a given input b Input: b Output: $A^{-1}b = x$

$$\begin{aligned} \hat{k} &= \frac{||A^{-1}||.||b||}{||A^{-1}b||} = \frac{||A^{-1}||.||b||}{||x||} \\ \implies \hat{k} &\leq \|A^{-1}\|\|A\| \end{aligned}$$

Result: $A \in \mathbb{R}^{m \times n}$ and non-singular and consider $Ax = b$, the problem of computing for an input x

$$\hat{k} = \frac{||A||.||x||}{||Ax||} \leq \|A\|\|A^{-1}\|$$

The problem of computing of given input b has condition number

3.1 Condition number of a matrix

$$k(A) = \|A\|\|A^{-1}\|$$

if $k(A)$ is small, A is said to be well conditioned

$$k(A) = \|A\|_2\|A^{-1}\|_2$$

$\|A\|_2 \rightarrow \sigma_1$ (max singular matrix value of A) $\|A^{-1}\|_2 \rightarrow \frac{1}{\sigma_m}$ (min singular matrix value of A)

$$\implies k(A) = \frac{\sigma_1}{\sigma_m}$$

Non zero singular values of A are square roots of non-zero eigen values of A^TA or AA^T

$$k(A) = \frac{\sqrt{\lambda_{max}(A^TA)}}{\sqrt{\lambda_{min}(A^TA)}}$$

If A is a symmetrix matrix

$$k(A) = \frac{|\lambda_{max}(A)|}{|\lambda_{min}(A)|}$$

If $A \in \mathbb{R}^{m\times n}$ $(m > n)$, and $A^+ = (A^TA)^{-1}A^T$ (pseudo inverse of A)

$$k(A) = \|A\|\|A^+\|$$

3.2 Conditioning of a system of equations

Fix b: Consider $f : A \rightarrow x = A^{-1}b$ Input: A Output: x

$$\begin{aligned}
&(A + \delta A)(x + \delta x) = b \\
\implies &Ax + A\delta x + \delta Ax + \delta A\delta x = b \\
\implies &A\delta x + \delta Ax = 0 \\
\implies &\delta x = -A^{-1}(\delta A)x \\
\implies &\|\delta x\| = \| - A^{-1}(\delta A)x\| \leq \|A^{-1}\|\|\delta Ax\| \\
\implies &\boxed{\|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x\|}
\end{aligned}$$

$$\hat{k} = \max_{\delta A} \frac{\frac{\|\delta x\|}{\|x\|}}{\frac{\|\delta A\|}{\|A\|}} \implies \frac{\frac{\|\delta x\|}{\|x\|}}{\frac{\|\delta A\|}{\|A\|}} \leq \|A\|\|A^{-1}\|$$

Perturbation of δA exists, which makes above inequality an equality

$$\hat{k} = \|A\|\|A^{-1}\| = k(A)$$

4 Stability of Algorithms

Getting the best answer for a given problem though it is not an exact answer for the problem

Algorithm: $f : X \to Y$, where X is the vector space of data, and Y is the vector space of solution

$y = f(x)$, where $x \in X, y \in Y$ An algorithm can be viewed as a function $\tilde{f}$ which takes the same input $x \in X$ and maps it to a result which is a collection of floating point numbers that belongs to Y

Accuracy: A good algorithms $\tilde{f}$ should be designed in a way such that it closely approximates the underlying problem f.

Absolute error of computation: $\|\tilde{f}(x) - f(x)\|$ Relative error of computation: $\frac{\|\tilde{f}(x)-f(x)\|}{\|f(x)\|}$

We say that $\tilde{f}$ is an accurate algorithm for f for all relevant input x

$$2\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = O(\epsilon_M)$$

Forward relative error: If f is ill-conditioned

$$\max_{\delta x} \frac{\frac{\|\delta f\|}{\|f\|}}{\frac{\|\delta x\|}{\|x\|}} = \hat{k} \text{ is very large}$$

Since $\frac{\|\delta x\|}{\|x\|} = O(\epsilon_M)$, $\frac{\|\delta f\|}{\|f\|} \leq \hat{k} O(\epsilon_M)$

We can say an algorithm $\tilde{f}$ for solving a problem f is stable for all input data x if

$$\frac{||\tilde{f}(x) - f(\tilde{x})||}{||f(\tilde{x})||} = O(\epsilon_M)$$

for some $\tilde{x}$ satisfying

$$\frac{||\tilde{x} - x||}{||x||} = O(\epsilon_M)$$

A stable algorithm gives nearly right answer to nearly right question. $||\tilde{f}(x) - f(\tilde{x})||$ is called backward error

4.1 Backward Stability

$\tilde{f}$ for a problem f such that $\boxed{\tilde{f}(x) = f(\tilde{x})}$. That is exactly right answer for n nearly right question

4.1.1 Stability of floating point arithmetic operation

$$\begin{aligned} f'(x) &= x(1+\epsilon), && \text{where } |\epsilon| < \epsilon_M \\ x * y &= x * y(1+\epsilon), && \text{where } |\epsilon| < \epsilon_M \end{aligned}$$

Example : Foating point arithmetic for $-$

$$
\begin{aligned}
&f(\underset{\sim}{x}) = x_1 - x_2 \\
&x_1 \to f'(x_1), x_2 \to f'(x_2) \\
&f'(x_1) = x_1(1+\epsilon_1), f'(x_2) = x_2(1+\epsilon_2)
\end{aligned}
$$

Algorithm:

$$
\begin{aligned}
f'(x_1) -' f'(x_2) &= \tilde{f} \\
&= x_1(1+\epsilon_1) -' x_2(1+\epsilon_2) \\
&= (x_1(1+\epsilon_1) - x_2(1+\epsilon_2))(1+\epsilon_3) \\
&= x_1(1+\epsilon_1)(1+\epsilon_3) - x_2(1+\epsilon_2)(1+\epsilon_3) \\
&= x_1(1+\epsilon_1)(1+\epsilon_3) - x_2(1+\epsilon_2)(1+\epsilon_3) \\
&= x_1(1+\epsilon_1+\epsilon_3) - x_2(1+\epsilon_2+\epsilon_3) \\
&= x_1(1+\epsilon_4) - x_2(1+\epsilon_5) \\
\implies f'(x_1) - f'(x_2) &= x_1(1+\epsilon_4) - x_2(1+\epsilon_5) \\
&= \tilde{x}_1 - \tilde{x}_2 \\
&= f(\tilde{x}_1, \tilde{x}_2) = f(\tilde{x})
\end{aligned}
$$

Example 2 : Outer product between 2 vectors $\underset{\sim}{x} \in \mathbb{R}^m$, $\underset{\sim}{y} \in \mathbb{R}^m$ and $\underset{\sim}{A} = \underset{\sim}{x}\underset{\sim}{y}^T$ i.e., $A_{ij} = x_i y_j$

$$
\begin{aligned}
\tilde{f}(\underset{\sim}{x}, \underset{\sim}{y}) = \tilde{A}_{ij} &= f'(x_i) \times f'(y_j) \\
&= f'(x_i) \times f'(y_j) \times (1+\epsilon_3^{ij}) \\
&= x_i(1+\epsilon_1^i) \times y_j(1+\epsilon_2^j) \times (1+\epsilon_3^{ij}) \\
&= x_i y_j (1+\epsilon_1^i)(1+\epsilon_2^j)(1+\epsilon_3^{ij}) \\
&= x_i y_j (1+\epsilon_1^i+\epsilon_2^j)(1+\epsilon_3^{ij}) \\
&= x_i y_j (1+\epsilon_4^{ij})(1+\epsilon_3^{ij})
\end{aligned}
$$

Verify: $\tilde{f}(\underset{\sim}{x}, \underset{\sim}{y}) = f(\tilde{x}, \tilde{y}) = \underset{\sim}{\tilde{x}}\underset{\sim}{\tilde{y}}^T = (\underset{\sim}{x} + \delta \underset{\sim}{x})(\underset{\sim}{y} + \delta \underset{\sim}{y})^T$

Exercise: Adding 1 to a real number i.e., $f(x) = x + 1$, $x \in \mathbb{R}$

$$\begin{aligned}\tilde{f}(x) &= f'(x) +' 1 \\ &= (f'(x) + 1)(1 + \epsilon_1) \\ &= (x(1 + \epsilon_2) + 1)(1 + \epsilon_1) \\ &= (x + x\epsilon_2 + 1)(1 + \epsilon_1) \\ &= x + x\epsilon_2 + 1 + x\epsilon_1 + x\epsilon_1\epsilon_2 + \epsilon_1 \\ &= (1 + \epsilon_1) + x(1 + \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2)\end{aligned}$$

Is it stable?

4.1.2 Unstable Algorithms

Computing eigen values of symmetric matrix Algorithm:

- Find the coefficients of $p(\lambda) = det(\underset{\sim}{A} - \lambda \underset{\sim}{I})$
- Roots of $p(\lambda)$

Example $\underset{\sim}{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$,

$$\implies \lambda^2 - p\lambda + 1;\ \text{Roots} \to \frac{p \pm \sqrt{p^2 - 4}}{2};\ \text{where } \tilde{p} = p(1 + \epsilon), |\epsilon| < \epsilon_M$$
$$\implies \text{Roots: } \frac{p(1 + \epsilon) \pm \sqrt{(p(1 + \epsilon)^2 - 4}}{2}$$

If $p = 2$, then the roots are $(1 + \epsilon) \pm \sqrt{2\epsilon}$. This implies that $error \approx ()(\sqrt{\epsilon}) > O(\epsilon_M)$

4.1.3 Accuracy of a backward stable algorithm

If a backward stable algorithm is applied to solve a problem f with condition number κ, the relative forward errors satisfy

$$\frac{||\tilde{f}(x) - f(x)||}{||f(x)||} = O(\kappa\epsilon_M)$$

Proof: Since f is backward stable, $\tilde{f}(x) = f(\tilde{x})$ where $\frac{||\underset{\sim}{x}-\underset{\sim}{\tilde{x}}||}{||\underset{\sim}{x}||} = O(\epsilon_M)$

$$\kappa(x) = \max_{\delta x} \frac{\frac{||\delta f||}{||f||}}{\frac{||\delta x||}{||x||}}$$

$$\implies \frac{\frac{||f(\tilde{x})-f(x)||}{||f(x)||}}{\frac{||\tilde{x}-x||}{||x||}} \leq \kappa(x)$$

$$\implies \frac{||f(\tilde{x}) - f(x)||}{||f(x)||} \leq \kappa(x).\frac{||\tilde{x} - x||}{||x||}$$

$$\implies \frac{||f(\tilde{x}) - f(x)||}{||f(x)||} \leq O(\kappa\epsilon_M)$$

5 Singular Value Decomposition (SVD)

5.1 Geometric Intuition

$\underset{\sim 1}{u}, \underset{\sim 2}{u}$ are the principal semi-axes of an ellipse with lengths σ_1, σ_2. $\tilde{v}_1, \tilde{v}_2$ are the pre-image vectors generating $\tilde{u}_1, \tilde{u}_2$ as the axes of ellipse

$$\implies \tilde{A}\tilde{v}_1 = \sigma_1\tilde{u}_1, \tilde{A}\tilde{v}_1 = \sigma_1\tilde{u}_1$$

$$\implies \tilde{A}[\tilde{v}_1\, \tilde{v}_2]$$