§4 计算机算术

§4.2 取整误差分析

§4.2.1 单个数取整

Def 4.24 舍入 $fl: \mathbb{R} \to F \cup \{+\infty, -\infty, NaN\}$，默认舍入到最近整数；距离相等情况下选择舍入到最近偶数.

Def 4.25 若 $|x| > OFL(F)$，$fl(x) = NaN$，称上溢；若 $|x| \in (0, UFL(F))$，$fl(x) = 0$，称下溢.

Def 4.26 F 的单位舍入 $\epsilon_u = \frac{1}{2}\epsilon_M = \frac{1}{2}\beta^{1-p}$

Thm 4.27 对 $x \in \mathcal{R}(F)$，有 $fl(x) = x(1+\delta)$，$|\delta| < \epsilon_u$

PF: $x_L, x_R \in F$，$x_L \le x \le x_R$ 为 x 相邻 2 个浮点数，$|fl(x) - x| \le \frac{1}{2}|x_R - x_L| \le \epsilon_u \min(|x_L|, |x_R|) < \epsilon_u |x|$（由 Lem 4.23）

Thm 4.28 对 $x \in \mathcal{R}(F)$，有 $fl(x) = \frac{x}{1+\delta}$，$|\delta| \le \epsilon_u$

PF: $|fl(x) - x| \le \frac{1}{2}|x_R - x_L| \le \epsilon_u \min(|x_L|, |x_R|) \le \epsilon_u |fl(x)|$

Ex 4.29 $\frac{2}{3} = (0.101010\ldots)_2 = (1.010101\ldots)_2 \times 2^{-1}$

$x_L = (1.010\ldots10)_2 \times 2^{-1}$　　$x - x_L = \frac{2}{3} \times 2^{-24}$,　　$\Rightarrow fl(x) = x_R$

$x_R = (1.010\ldots11)_2 \times 2^{-1}$　　$x_R - x_L = 2^{-24}$

§4.2.2 二进制浮点数运算

Def 4.30（加减）对 $a, b \in F$，$a = M_a \times \beta^{e_a}$，$b = M_b \times \beta^{e_b}$，$M_a = \pm m_a$，$M_b = \pm m_b$，不妨设 $|a| > |b|$，在精度至少为 $2p$ 的寄存器中计算 $c = fl(a+b) \in F$.

① 比较指数：若 $e_a - e_b > p+1$，相差太大，返回 $C = a$；否则 $e_c \leftarrow e_a$，$M_b \leftarrow M_b / \beta^{e_a - e_b}$（对 M_b 右移）

② 加法：$M_c \leftarrow M_a + M_b$，舍入后精度为 $2p$

③ 正规化：若 $M_c = 0$，返回 0；移位使 $M_c \in [1, \beta)$，同时改 e_c：$|M_c| \in (\beta, \beta^2)$，$M_c \leftarrow \frac{M_c}{\beta}$；$|M_c| \in (0, \beta - \epsilon_u(p))$，一直 $M_c \leftarrow M_c p$ 直到 $|M_c| \in [1, \beta)$；若 $|M_c| \in [\beta - \epsilon_u(p), \beta]$，$|M_c| \leftarrow 1.0$.

④ 检查范围 (e_c)：NaN / 0

⑤ 舍入到精度 p　⑥ $C \leftarrow M_c \times \beta^{e_c}$

Lem 4.34 对 $a, b \in F$，$a+b \in \mathcal{R}(F)$ 有 $fl(a+b) = (a+b)(1+\delta)$，$|\delta| < \epsilon_u$

PF: 倒数第 2 步（⑤）的舍入误差占主要，由 Thm 4.27 即证.

Def 4.35（乘）对 $a, b \in F$，$a = M_a \times \beta^{e_a}$，$b = M_b \times \beta^{e_b}$，$M_a = \pm m_a$，$M_b = \pm m_b$，在精度至少为 $2p$ 的寄存器计算 $fl(ab)$

① $e_c \leftarrow e_a + e_b$　② $M_c \leftarrow M_a M_b$ 在寄存器中舍入.

③ 正规化：① $|M_c| \in (\beta, \beta^2)$，$M_c \leftarrow \frac{M_c}{\beta}$，$e_c \leftarrow e_c + 1$　② $|M_c| \in [\beta - \epsilon_u(p), \beta]$，$|M_c| \leftarrow 1.0$，$e_c \leftarrow e_c + 1$

④ 检查范围 NaN / 0　⑤ 舍入到精度 p　⑥ $C \leftarrow M_c \times \beta^{e_c}$

Lem 4.37 对 $a, b \in F$，$|ab| \in \mathcal{R}(F)$ 有 $fl(ab) = (ab)(1+\delta)$，$|\delta| < \epsilon_u$

PF: 同样只有⑤误差占主要地位.

Def 4.38 对 $a,b\in\mathcal{F}$, $a=M_a\times\beta^{e_a}$, $b=M_b\times\beta^{e_b}$, $M_a=\pm m_a$, $M_b=\pm m_b$. 在精度至少为 $2p+1$ 寄存器中计算 $fl(\frac{a}{b})\in\mathcal{F}$.

① 若 $m_b=0$. 返回 NaN; ② 否则 $e_c\leftarrow e_a-e_b$; ③ $M_c\leftarrow\frac{M_a}{M_b}$, 在寄存器中舍入

③ Normalization: 若 $|M_c|<1$, $M_c\leftarrow M_c\cdot\beta$, $e_c\leftarrow e_c-1$ ④ 检查范围 (e_c): NaN/0

⑤ 舍入到精度 p; ⑥ 返回 $c\leftarrow M_c\times\beta^{e_c}$

Lem 4.39 对 $a,b\in\mathcal{F}$, $\frac{a}{b}\in\mathcal{R}(\mathcal{F})$, 有 $fl(\frac{a}{b})=\frac{a}{b}(1+\delta)$, $|\delta|<\epsilon_u$

PF: 当 $|M_a|=|M_b|$ 时, 没有误差; 考虑 $|M_a|>|M_b|$, $|M_a|,|M_b|\in[1,\beta)$, $\left|\frac{M_a}{M_b}\right|\geq\frac{\beta-\epsilon_M}{\beta-2\epsilon_M}>1+\beta^{-1}\epsilon_M$

此时③不用进行, 设有 $p+k$ 精度, 则舍入单位 $\frac{1}{2}\beta^{1-p-k}=\frac{1}{2}\beta^{1-p}\beta^{1-p}\beta^{p-1-k}=\epsilon_u\epsilon_M\beta^{p-1-k}$

令 $k=p+1$, 则舍入单位为 $\epsilon_u\epsilon_M\beta^{-2}$, 记第②步结果为 M_{c_1}, ⑤结果为 M_{c_2}, 从而:

$$M_{c_2}=M_{c_1}+\delta_2\ (|\delta_2|<\epsilon_u)=\frac{M_a}{M_b}+\delta_1+\delta_2\ (|\delta_1|<\epsilon_u\epsilon_M\beta^{-2})=\frac{M_a}{M_b}(1+\delta)$$

$$|\delta|=\left|\frac{\delta_1+\delta_2}{M_a/M_b}\right|<\frac{\epsilon_u(1+\epsilon_M\beta^{-2})}{1+\epsilon_M\beta^{-1}}<\epsilon_u$$

考虑 $|M_a|<|M_b|$, 类似有 $\left|\frac{M_a}{M_b}\right|\leq\frac{\beta-2\epsilon_M}{\beta-\epsilon_M}=1-\frac{\epsilon_M}{\beta-\epsilon_M}<1-\beta^{-1}\epsilon_M$, 此时③一定会进行

$$M_{c_1}=M_a/M_b+\delta_1,\ |\delta_1|<\epsilon_u\epsilon_M\beta^{-2},\quad M_{c_2}=\beta M_{c_1}+\delta_2=\beta\frac{M_a}{M_b}\left(1+\frac{\beta\delta_1+\delta_2}{\beta M_a/M_b}\right),\ |\delta_2|<\epsilon_u$$

$$\beta\left|\frac{M_a}{M_b}\right|\geq\frac{\beta}{\beta-\epsilon_M}=1+\frac{\epsilon_M}{\beta-\epsilon_M}>1+\beta^{-1}\epsilon_M,\ \text{代入得到}\ |\delta|=\left|\frac{\beta\delta_1+\delta_2}{\beta M_a/M_b}\right|<\frac{\epsilon_u\epsilon_M\beta^{-1}+\epsilon_u}{1+\beta^{-1}\epsilon_M}=\epsilon_u$$

Thm 4.40 记 $\mathcal{F}$ 为精度 p 的正规 FPN 系统, 则 $\forall\odot=+,-,\times,/$, $\forall a,b\in\mathcal{F}$ $a\odot b\in\mathcal{R}(\mathcal{F})$, $fl(a\odot b)=(a\odot b)(1+\delta)$, $|\delta|<\epsilon_u$

当且仅当在 $2p+1$ 精度的寄存器中运算.

§4.2.3 舍入误差的传递

Thm 4.41 若 $\forall i=0,1,\dots,n$, $a_i\in\mathcal{F}$, $a_i>0$, 则 $fl(\sum_{i=0}^{n}a_i)=(1+\delta_n)\sum_{i=0}^{n}a_i$, 其中 $|\delta_n|<(1+\epsilon_u)^n-1\approx n\epsilon_u$

PF: 令 $S_k=\sum_{i=0}^{k}a_i$, $S_0^*=a_0$, $S_{k+1}^*=fl(S_k^*+a_{k+1})$ $\delta_k=\frac{S_k^*-S_k}{S_k}$, $\epsilon_k=\frac{S_{k+1}^*-(S_k^*+a_{k+1})}{S_k^*+a_{k+1}}$

由归纳法, 则有 $\delta_{k+1}=\frac{S_{k+1}^*-S_k}{S_{k+1}}=\frac{(S_k^*+a_{k+1})(1+\epsilon_k)-S_{k+1}}{S_{k+1}}=\frac{(S_k(1+\delta_k)+a_{k+1})(1+\epsilon_k)-S_k-a_{k+1}}{S_{k+1}}=\frac{(\epsilon_k+\delta_k+\epsilon_k\delta_k)S_k+\epsilon_k a_{k+1}}{S_{k+1}}$

$=\frac{\epsilon_k S_{k+1}+\delta_k(1+\epsilon_k)S_k}{S_{k+1}}=\epsilon_k+\delta_k(1+\epsilon_k)\frac{S_k}{S_{k+1}}$, $\because S_k<S_{k+1}$, $|\epsilon_k|<\epsilon_u$ $\therefore|\delta_{k+1}|<|\epsilon_k|+|\delta_k|(1+\epsilon_u)<\epsilon_u+|\delta_k|(1+\epsilon_u)$

$\therefore\forall k\in\mathbb{N}$, $|\delta_{k+1}|<\epsilon_u\sum_{i=0}^{k}(1+\epsilon_u)^i=\epsilon_u\frac{(1+\epsilon_u)^{k+1}-1}{1+\epsilon_u-1}=(1+\epsilon_u)^{k+1}-1$

Ex 4.42 对 $a_i>0$, 按从小到大进行加法, 可以最小化舍入误差.

Ex 4.43 类比 $fl(\sum_i\prod_j a_{ij})$, $fl(a_1b_1+a_2b_2+a_3b_3)=(1+\delta)$ Ans 其中 $|\delta|<(1+\epsilon_u)^6-1$

Thm 4.44 对给定 $\mu\in\mathbb{R}^+$, 及正整数 $n\leq\lfloor\frac{\ln 2}{\mu}\rfloor$, 设 $|\delta_i|\leq\mu$, $i=1,2,\dots,n$. 则 $1-n\mu\leq\prod_{i=1}^{n}(1+\delta_i)\leq1+n\mu+(n\mu)^2$

或者等价地对 $I_n=[-\frac{1}{1+n\mu},1]$ 有, $\exists\theta\in I_n$, s.t. $\prod_{i=1}^{n}(1+\delta_i)=1+\theta(n\mu+n^2\mu^2)$

PF: 由 $|\delta_i|\leq\mu$, $i=1,2,\dots,n$ 可得 $(1-\mu)^n\leq\prod_{i=1}^{n}(1+\delta_i)\leq(1+\mu)^n\leq e^{n\mu}$ $(1+\mu\leq e^\mu)$

将 $f(\mu)=(1-\mu)^n$ 在 $\mu=0$ 展开 $(1-\mu)^n=1-n\mu+\frac{1}{2}n(n-1)\theta^2\geq1-n\mu$

而 $e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots\leq1+x+\frac{x^2}{2}e^x$ 令 $x=n\mu\leq\ln 2$ 则有 $e^{n\mu}\leq1+n\mu+(n\mu)^2$

由于 $\prod_{i=1}^{n}(1+\delta_i)$ 在连续函数 $f(\tau)=1+\tau(1+n\mu)n\mu$, $\tau\in I_n$ 的值域中, 由中值定理即证.

§4.3 精确性和稳定性.

§4.3.1 Avoiding catastrophic cancellation

$$\begin{tikzcd} \hat{x} \arrow[r, "f"] \arrow[d, "\Delta x"'] & \hat{y} \arrow[d, "\Delta y"] \\ x \arrow[r, "f"'] \arrow[ur, "\hat{f}"] & y \end{tikzcd}$$

Def 4.45 设 $\hat{x}$ 为 $x\in\mathbb{R}$ 的近似，绝对误差为 $E_{abs}(\hat{x})=|x-\hat{x}|$，相对误差为 $E_{rel}(\hat{x})=\frac{|\hat{x}-x|}{|x|}$

Def 4.46 通过 $\hat{y}=\hat{f}(x)$ 近似 $y=f(x)$，向前误差为 $\hat{y}$ 近似 y 的相对误差；向后误差是用 $f(\hat{x})=\hat{f}(x)$ 中 $\hat{x}$ 近似 x 的最小相对误差.

Def 4.47 计算 $y=f(x)$ 的算法 $\hat{y}=\hat{f}(x)$ 是精确的 如果存在 $c>0$ 且很小，$\forall x\in dom(f)$，$E_{rel}(\hat{f}(x))=\left|\frac{\hat{f}(x)-f(x)}{f(x)}\right|\le c\epsilon_u$

Ex 4.48 (Catastrophic cancellation)，$x,y\in\mathcal{R}(\mathcal{F})$，由 Thm 4.27/40. $fl(fl(x)\odot fl(y))=((x)(1+\delta_1)\odot y(1+\delta_2))(1+\delta_3)$，$|\delta_i|\le\epsilon_u$

由 Thm 4.40/44 $fl(fl(x)\times fl(y))=xy(1+\delta_1)(1+\delta_2)(1+\delta_3)=xy[1+\theta(3\epsilon_u+3\epsilon_u^2+\epsilon_u^3)]$，$\theta\in[-1,1]$

类似的，$fl\left(\frac{fl(x)}{fl(y)}\right)=\frac{x(1+\delta_1)}{y(1+\delta_2)}(1+\delta_3)=\frac{x}{y}(1+\delta_1)(1-\delta_2+\delta_2^2-\cdots)(1+\delta_3)\approx\frac{x}{y}(1+\delta_1)(1-\delta_2)(1+\delta_3)$

然而 加/减 可能不精确：$fl(fl(x)+fl(y))=(x(1+\delta_1)+y(1+\delta_2))(1+\delta_3)=(x+y)\left(1+\delta_3+\frac{x\delta_1+y\delta_2}{x+y}+\delta_3\frac{x\delta_1+y\delta_2}{x+y}\right)$

当 $x+y\to 0$ 时，相对误差可以任意大

Thm 4.49 设 $x,y\in\mathcal{F}$，$x>y>0$ 且 $\beta^{-t}\le 1-\frac{y}{x}\le\beta^{-s}$，在计算 $x-y$ 时最大有效数字的位数至多丢失 t，至少丢失 s.

PF: 令 $x=m_x\times\beta^n$，$y=m_y\times\beta^m$，$m_x,m_y\in[1,\beta)$，首先对 y 移位 $y=(m_y\times\beta^{m-n})\times\beta^n\Rightarrow x-y=(m_x-m_y\times\beta^{m-n})\times\beta^n$

$\Rightarrow m_{x-y}>m_x-m_y\times\beta^{m-n}=m_x\left(1-\frac{m_y\times\beta^m}{m_x\times\beta^n}\right)=m_x\left(1-\frac{y}{x}\right)\Rightarrow\beta^{-t}\le m_{x-y}<\beta^{1-s}$，正规化时左移至多 t 位，至少 s 位.

△当 x,y 很接近时，使用减法会出现上述问题，应尽量避免.

§4.3.2 Backward stability and numerical stability

Def 4.52，计算 $y=f(x)$ 的算法 $\hat{y}=\hat{f}(x)$ 是向后稳定的，若 $\exists c>0$ s.t. $\forall x\in dom(f)$，$\exists\hat{x}\in dom(f)$，$\hat{f}(x)=f(\hat{x})\Rightarrow E_{rel}(\hat{x})\le c\epsilon_u$

Def 4.53 $\hat{y}=\hat{f}(x_1,x_2)$ 是向后稳定的，若 $\exists c_1,c_2>0$ s.t. $\forall(x_1,x_2)\in dom(f)$，$\exists(\hat{x}_1,\hat{x}_2)\in dom(f)$ s.t.

$\hat{f}(x_1,x_2)=f(\hat{x}_1,\hat{x}_2)\Rightarrow E_{rel}(\hat{x}_1)\le c_1\epsilon_u$，$E_{rel}(\hat{x}_2)\le c_2\epsilon_u$.

Lem 4.54 对 $f(x_1,x_2)=x_1-x_2$，$x_1,x_2\in\mathcal{R}(\mathcal{F})$，算法 $\hat{f}(x_1,x_2)=fl(fl(x_1)-fl(x_2))$ 是向后稳定的.

PF: $\hat{f}(x_1,x_2)=(x_1(1+\delta_1)-x_2(1+\delta_2))(1+\delta_3)=x_1(1+\delta_1+\delta_3+\delta_1\delta_3)-x_2(1+\delta_2+\delta_3+\delta_2\delta_3):=\hat{x}_1-\hat{x}_2$

Ex 4.55 $\hat{f}(x)=fl(1.0+fl(x))$ 不是向后稳定，当 $x\in(0,\epsilon_u)$ 时，$E_{rel}(\hat{x})=1$ $(\hat{x}=0)$

Def 4.56 计算 $y=f(x)$ 的算法 $\hat{y}=\hat{f}(x)$ 是稳定的(或数值稳定) $\Leftrightarrow\exists c,c_f>0$，s.t. $\forall x\in dom(f)$，$\exists\hat{x}\in dom(f)$，

s.t. $\left|\frac{\hat{f}(x)-f(\hat{x})}{f(\hat{x})}\right|\le c_f\epsilon_u$，$E_{rel}(\hat{x})\le c\epsilon_u$

Lem 4.57 如果一个算法向后稳定，则其数值稳定.

Ex 4.58 $\hat{f}(x)=fl(1.0+fl(x))$ 是稳定的.

若 $|x|<\epsilon_u$，取 $\hat{x}=x$，$\left|\frac{f(\hat{x})-\hat{f}(x)}{f(\hat{x})}\right|=\left|\frac{x}{1+x}\right|<2\epsilon_u$；若 $|x|\ge\epsilon_u$ $\hat{f}(x)=1+\delta_2+x(1+\delta_1+\delta_2+\delta_1\delta_2)$，$|\delta_1|,|\delta_2|<\epsilon_u$

令 $\hat{x}=x(1+\delta_1+\delta_2+\delta_1\delta_2)$ $E_{rel}(\hat{x})=|\delta_1+\delta_2+\delta_1\delta_2|<3\epsilon_u\Rightarrow\left|\frac{\hat{f}(x)-f(\hat{x})}{f(\hat{x})}\right|=\left|\frac{\delta_2}{1+\delta_2+x(1+\delta_1+\delta_2+\delta_1\delta_2)}\right|<|\epsilon_u|$

§4.3.3 Condition numbers: Scalar functions

Def 4.59 函数 $y=f(x)$ 的相对条件数是输入的微小变化对输出的度量 $C_f(x)=\left|\frac{xf'(x)}{f(x)}\right|$

Def 4.60 条件数小的问题称为良态，否则称为病态.

Ex 4.61，由Def 4.59，$\frac{E_{rel}(\hat{y})}{E_{rel}(\hat{x})} = \left|\frac{x(f(x)-f(\hat{x}))}{f(x)(x-\hat{x})}\right| \lessapprox C_f = \left|\frac{xf'(x)}{f(x)}\right| \Rightarrow E_{rel}(\hat{y}) \lessapprox C_f E_{rel}(\hat{x})$

可见病态问题前向误差大.

Lem 4.63 考虑对单根 r 附近求解 $f(x)=0$ 即 $f(r)=0$，$f'(r)\neq 0$，设有微小扰动 $F=f+\epsilon g$，$f, g\in C^2$，$g(r)\neq 0$ 且满足 $|\epsilon g'(r)| \ll |f'(r)|$，则 F 有根 $r+h$，其中 $h\approx -\epsilon\frac{g(r)}{f'(r)}$.

PF: $0=F(r+h)=f(r)+hf'(r)+\epsilon[g(r)+hg'(r)]+O(h^2)$ 从而有 $h\approx -\epsilon\frac{g(r)}{f'(r)+\epsilon g'(r)} \approx -\epsilon\frac{g(r)}{f'(r)}$

Ex 4.64 (Wilkinson) 定义 $f(x)=\prod_{k=1}^{p}(x-k)$，$g(x)=x^p$，根 $x=p$ 受扰动后 $h\approx -\epsilon\frac{p^p}{(p-1)!}$，当 p 很大时，几乎不可能求解高阶的根

§4.3.4 Condition numbers: vector functions

Def 4.65 向量函数 $f:\mathbb{R}^m\to\mathbb{R}^n$ 的条件数为 $\mathrm{cond}_f(x)=\frac{\|x\|\|\nabla f\|}{\|f(x)\|}$，其中 $\|\cdot\|$ 表示欧几里德范数(如1, 2, ∞)

Def 4.67 非奇异方阵 A 的条件数为 $\mathrm{cond}A:=\|A\|\|A^{-1}\|$.（由 $Au=b$ 求解 $u=A^{-1}b$ 得到）

Lem 4.68 在2范数下，非奇异方阵 A 的条件数为 $\mathrm{cond}_2A=\|A\|_2\|A^{-1}\|_2=\frac{\sigma_{max}}{\sigma_{min}}$，其中 $\sigma_{max}, \sigma_{min}$ 为最大/小奇异值.

若 A 是normal的，$\mathrm{cond}_2A=\frac{|\lambda_{max}|}{|\lambda_{min}|}$，为最大/小特征值.

Thm 4.70 A 为可逆阵，$\hat{A}$ 为 A 的一个扰动 $\|A^{-1}\|\|\hat{A}-A\|<1$，其中 $Ax=b$，$\hat{A}\hat{x}=\hat{b}$，则 $\hat{x}$ 关于 x 的相对误差满足

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

PF: $\hat{A}=A[I+A^{-1}(\hat{A}-A)]$，$\hat{A}^{-1}=[I+A^{-1}(\hat{A}-A)]^{-1}A^{-1}$ $\therefore \|\hat{A}^{-1}\|\le\frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\hat{A}-A\|}$ $\because \hat{x}-x=\hat{A}^{-1}[\hat{b}-b-(\hat{A}-A)x]$

$\Rightarrow\frac{\|\hat{x}-x\|}{\|x\|}\le\frac{\mathrm{cond}A}{1-\|\hat{A}-A\|\|A^{-1}\|}\left(\frac{\|\hat{b}-b\|}{\|A\|\|x\|}+\frac{\|\hat{A}-A\|}{\|A\|}\right)$，由 $\|A\|\|x\|\ge\|b\|$ $\Rightarrow\frac{\|\hat{x}-x\|}{\|x\|}\le\frac{\mathrm{cond}A}{1-(\mathrm{cond}A)\frac{\|\hat{A}-A\|}{\|A\|}}\left(\frac{\|\hat{b}-b\|}{\|b\|}+\frac{\|\hat{A}-A\|}{\|A\|}\right)$

[由Thm E.145 $\|(I-A)^{-1}\|\le\frac{1}{1-\|A\|}$（其中 $\|A\|<1$）]

Def 4.71 向量函数 $f:\mathbb{R}^m\to\mathbb{R}^n$ 的条件数为 $\mathrm{cond}_f(x)=\frac{\|x\|\|\nabla f\|}{\|f(x)\|}$，分量条件数为 $\mathrm{cond}_f(x)=\|A(x)\|$

其中 $A(x)=[a_{ij}(x)]$，$a_{ij}(x)=\left|\frac{x_j\frac{\partial f_i}{\partial x_j}}{f_i(x)}\right|$，也就是 a_{ij} 是 f_i 关于 x_j 的条件数.

Ex 4.72 对于向量函数 $f(x)=[\frac{1}{x_1}+\frac{1}{x_2}, \frac{1}{x_1}-\frac{1}{x_2}]$

其Jacobi矩阵为 $\nabla f=-\frac{1}{x_1^2x_2^2}\begin{bmatrix}x_2^2 & x_1^2\\ x_2^2 & -x_1^2\end{bmatrix}$，$C_c=\begin{bmatrix}\left|\frac{x_2}{x_1+x_2}\right| & \left|\frac{x_1}{x_1+x_2}\right|\\ \left|\frac{x_2}{x_1-x_2}\right| & \left|\frac{x_1}{x_1-x_2}\right|\end{bmatrix}$ 为 $\mathrm{cond}_f(x)$ 中的 $A(x)$. 可见当 $x_1\pm x_2\approx 0$ 时会导致病态. $C_1=\frac{\|x\|_1\|\nabla f\|_1}{\|f\|_1}=\frac{|x_1|+|x_2|}{|x_1x_2|}\frac{2\max(x_1^2, x_2^2)}{|x_1+x_2|+|x_1-x_2|}$，然而当 $x_1\pm x_2\approx 0$ 时 $C_1\approx 2$，看不出病态.

注：$\forall A\in\mathbb{R}^{n\times n}$，$\|A\|_1=\max_j\sum_i|a_{ij}|$

Ex 4.73 $t_1,\dots,t_n$ 在 $[-1,1]$ 中等距分布，Vandermonde阵的条件数(基于∞范数)为 $\mathrm{cond}_\infty V_n\sim\frac{1}{\pi}e^{-\pi/4}e^{n/4(\pi+2\ln 2)}$

Def 4.74 Hilbert矩阵 $H_n\in\mathbb{R}^{n\times n}$ 为 $h_{ij}=\frac{1}{i+j-1}$

Ex 4.75 希尔伯特矩阵基于2范数的条件数为 $\mathrm{cond}_2H_n\sim\frac{(\sqrt{2}+1)^{4n+4}}{2^{15/4}\sqrt{\pi n}}$

§4.3.5 Condition numbers: algorithms.

Def 4.76 向量函数 $f:\mathbb{R}^m\to\mathbb{R}^n$ 通过 $f_A:\mathcal{F}^m\to\mathcal{F}^n$ 逼近，设 $\forall x\in\mathcal{F}^m$，$\exists x_A\in\mathbb{R}^m$，s.t. $f_A(x)=f(x_A)$

则 f_A 的条件数定义为 $\mathrm{cond}_A(x)=\frac{1}{\epsilon_u}\inf_{\{x_A\}}\frac{\|x_A-x\|}{\|x\|}$

Ex 4.77 考虑算法 A 计算 $y=\ln x$，设 $\forall x>0$，计算得到 $y_A=(1+\delta)\ln x$，$|\delta|\le 5\epsilon_u$. 求 A 的条件数.

$y_A=\ln x_A$，$x_A=x^{1+\delta}$，$E_{rel}(x_A)=\left|\frac{x^{1+\delta}-x}{x}\right|=|x^\delta-1|=|e^{\delta\ln x}-1|\approx|\delta\ln x|\le 5|\ln x|\epsilon_u$

因此除了 $x\to0^+$，A 是良态的.

Thm 4.78 设光滑函数 $f:\mathbb{R}\to\mathbb{R}$ 由算法 $A:\mathcal{F}\to\mathcal{F}$，$f_A(x)=f(x)(1+\delta(x))$，$|\delta(x)|\le\varphi(x)\epsilon_u$ 逼近，若 $\mathrm{cond}_f(x)$ 有界非 0，则有 $\forall x\in\mathcal{F}$，$\mathrm{cond}_A(x)\le\frac{\varphi(x)}{\mathrm{cond}_f(x)}$.

Pf: 设有 x_A 满足 $f(x_A)=f_A(x)$. 将 x_A 写作 $x(1+\epsilon_A)$，则 $f(x)(1+\delta)=f(x_A)=f(x(1+\epsilon_A))=f(x)+x\epsilon_Af'(x)+O(\epsilon_A^2)$

$\Rightarrow \epsilon_A=\frac{f(x)}{xf'(x)}\delta$. $\therefore\left|\frac{x_A-x}{x}\right|=|\epsilon_A|=\left|\frac{f(x)}{xf'(x)}\right||\delta(x)|\Rightarrow\frac{1}{\epsilon_u}\left|\frac{x_A-x}{x}\right|=\left|\frac{\delta(x)}{\epsilon_u\mathrm{cond}_f(x)}\right|\le\frac{\varphi(x)}{\mathrm{cond}_f(x)}$

Ex 4.79 假设 $\sin x$，$\cos x$ 在计算机中以舍入误差计算。(用截断 Taylor 展开)

$f_A=fl\left[\frac{fl(1-fl(\cos x))}{fl(\sin x)}\right]$ 用计算 $f(x)=\frac{1-\cos x}{\sin x}$，$x\in(0,\frac{\pi}{2})$

Sol: $\mathrm{cond}_f(x)=\frac{x}{\sin x}$，$f_A(x)=\frac{(1-\cos x(1+\delta_1))(1+\delta_2)}{\sin x(1+\delta_3)}(1+\delta_4)$ $|\delta_i|\le\epsilon_u$，忽略 $O(\delta_i^2)$，则

$f_A(x)=\frac{1-\cos x}{\sin x}\left\{1+\delta_2+\delta_4-\delta_3-\delta_1\frac{\cos x}{1-\cos x}\right\}$，则有 $\varphi(x)=3+\frac{\cos x}{1-\cos x}\Rightarrow\mathrm{cond}_A(x)=\frac{\sin x}{x}\left(3+\frac{\cos x}{1-\cos x}\right)$

因此，$x\to0$ 时 $\mathrm{cond}_A(x)$ 无界，$x\to\frac{\pi}{2}$，$\mathrm{cond}_A(x)\to\frac{\pi}{6}$

§4.3.6 Overall error of a computer solution

$\xrightarrow{x}$ Rounding $\xrightarrow{x^*}$ Algorithm $f_A\longrightarrow f_A(x^*)$

Thm 4.81 $f:\mathbb{R}^m\to\mathbb{R}^n$，$y=f(x)$. 计算机输入与输出表示为 $x^*\approx x$，$y_A^*=f_A(x^*)$

则 $E_{rel}(y_A^*)\lessapprox E_{rel}(x_A^*)\mathrm{cond}_f(x)+\epsilon_u\mathrm{cond}_f(x^*)\mathrm{cond}_A(x^*)$

Pf: $\frac{\|y_A^*-y\|}{\|y\|}=\frac{\|f_A(x^*)-f(x)\|}{\|f(x)\|}<\frac{\|f(x^*)-f(x)\|}{\|f(x)\|}+\frac{\|f_A(x^*)-f(x^*)\|}{\|f(x)\|}$

由 $E_{rel}(\hat y)\lessapprox C_f E_{rel}(\hat x)$，$\frac{\|f(x^*)-f(x)\|}{\|f(x)\|}\lessapprox\mathrm{cond}_f(x)\frac{\|x^*-x\|}{\|x\|}=E_{rel}(x^*)\mathrm{cond}_f(x)$

$\frac{\|f_A(x^*)-f(x^*)\|}{\|f(x)\|}=\frac{\|f(x_A^*)-f(x^*)\|}{\|f(x)\|}\approx\frac{\|f(x_A^*)-f(x^*)\|}{\|f(x^*)\|}\le\mathrm{cond}_f(x^*)\frac{\|x_A^*-x^*\|}{\|x^*\|}=\epsilon_u\mathrm{cond}_A(x^*)\mathrm{cond}_f(x^*)$

$E_{rel}(y_A^*)\lessapprox E_{rel}(x_A^*)\mathrm{cond}_f(x)+\epsilon_u\mathrm{cond}_f(x^*)\mathrm{cond}_A(x^*)$