絶対値の近い数の加減算による桁落ち

絶対値が極近い 2 数を足したり、引いたりして[1]結果の絶対値が小さくなるような計算をすると有効数字が減る. このような現象を桁落ちという.

桁落ちの例 1　ごく近い 2 数の引き算での桁落ち

例えば絶対値のごく小さな x が

$$x = 0.0031834 = 0.31834 \times 10^{-2} \tag{1}$$

という数として

$$1 - \frac{1}{\sqrt{1+x}} = \frac{\sqrt{1+x} - 1}{\sqrt{1+x}} \tag{2}$$

の値を計算せよと言われたとする. このとき左辺と右辺で値が変わってしまう. x と同じ有効数字 5 桁で 6 桁めで四捨五入するとき, (2) 式の左辺の値は

$$\begin{aligned} 1 - \frac{1}{\sqrt{1+x}} &\approx 1 - \frac{1}{\sqrt{1.0032}} \\ &\approx 1 - \frac{1}{1.0016} \\ &\approx 1 - 0.99840 \\ &\approx 0.00160 \end{aligned} \tag{3}$$

[1] 異符号のときは足す, 同符号のときは引く.

となり, 最後の引き算で有効数字が 3 桁まで落ちてしまった. それに対して右辺の値は

$$
\begin{aligned}
\frac{\sqrt{1+x}-1}{\sqrt{1+x}} &\approx \frac{\sqrt{1.0032}-1}{\sqrt{1.0032}} \\
&\approx \frac{1.0016-1}{1.0016} \\
&\approx \frac{0.0016}{1.0016} \\
&\approx 0.0015\underline{974}
\end{aligned} \tag{4}
$$

となる. しかし, 下線がついた有効数字 3 桁以降は誤差を含んでしまっている[2]. そのため有効数字は 2 桁となってしまっている. 有効数字が 2 桁になってしまったのは $1.0016-1$ の計算で有効数字が 2 桁になってしまったからである.

結局左辺のまま計算しても, 右辺に変形しても有効数字は桁落ちを起こしてしまう. しかし, (2) 式は, 分子の有理化ともいうべき形に変形すると全桁正しく計算できる. (2) 式の右辺を次のように変形する.

$$
\begin{aligned}
\frac{\sqrt{1+x}-1}{\sqrt{1+x}} &= \frac{x}{\sqrt{1+x}(\sqrt{1+x}+1)} \\
&= \frac{x}{(1+x)+\sqrt{1+x}}.
\end{aligned} \tag{5}
$$

このように変形してから計算すると,

$$
\begin{aligned}
\frac{x}{(1+x)+\sqrt{1+x}} &\approx \frac{0.0031834}{1.0032+\sqrt{1.0032}} \\
&\approx \frac{0.0031834}{1.0032+1.0016} \\
&\approx \frac{0.0031834}{2.0048} \\
&= 0.0015879
\end{aligned} \tag{6}
$$

となって, 有効数字は 5 桁のままになる[3].

(5) 式の例は「式の変形で桁落ちを回避できることがある」ことを示している. 数値計算の観点からみてみるためにそれぞれの計算に必要な演算の回数を計算する. (2) 式の右辺

[2] 3 桁め以降に誤差が含まれていることは, 割り算を筆算で行うことで確認できる. 筆算をする際に被除数に 0 を補って筆算を行うが, それは不正確な数なので, その数字が計算の過程に含まれる数字には誤差が含まれることになる.

[3] 途中で引き算が入らないので桁落ちが起こらない.

$\dfrac{\sqrt{1+x}-1}{\sqrt{1+x}}$ はまず $\sqrt{1+x}$ の計算をやる. この計算では和算を 1 回, 開平を 1 回行っている. $\sqrt{1+x}$ の計算した結果を変数に格納しておく. さらに $\sqrt{1+x}$ の計算結果から 1 を引き, 格納しておいた変数で割る. この計算では減算を 1 回, 除算を 1 回行っている. よって, (2) 式の計算をまとめると

$+$	$-$	$\div$	$\sqrt{\ }$
1 回	1 回	1 回	1 回

となる[4] . それに対して (5) 式の右辺 $\dfrac{x}{(1+x)+\sqrt{1+x}}$ の計算は $1+x$ の値を先に計算して $1+x$ の計算結果を変数に格納しておく. この計算に和算を 1 回. 格納しておいた変数の開平をし, 格納した変数と足し合わせる. この計算で開平を 1 回, 和算を 1 回. 最後に x を $(1+x)+\sqrt{1+x}$ の計算した値で割る. よって除算を 1 回である. (5) 式の計算をまとめると

$+$	$\div$	$\sqrt{\ }$
2 回	1 回	1 回

となる. (2) 式も (5) 式も演算の種類の違いはあるが全体の計算の数は変わらない. しかし計算精度の観点に立つと (2) 式と (5) 式でははっきり優劣がでてくる.

桁落ちの例 2　倍角・半角の公式での桁落ち

似たようなことは三角関数を含んだ式の場合にも起こる. よく知られた倍角・半角の公式

$$\sin^2\frac{\theta}{2}=\frac{1}{2}(1-\cos\theta) \tag{7}$$

において, θ が小さいとき, 例えば

$$\theta = 1.23456^\circ \tag{8}$$

[4] 開平は他の演算よりも計算量が多い. 具体的には, 次の漸化式 $a_{n+1}=\frac{x/a_n+a_n}{2}$ を用いたニュートン法による反復計算を用いて計算しているようである.

のとき,

$$\begin{aligned}\sin^2\frac{\theta}{2} &= \sin^2 0.61728^\circ \\ &\approx (0.0107734)^2 \\ &\approx 1.16066\times 10^{-4} \end{aligned} \tag{9}$$

となるが, 右辺を計算すると

$$\begin{aligned}\frac{1}{2}(1-\cos\theta) &= \frac{1}{2}(1-\cos 1.23456^\circ) \\ &\approx \frac{1}{2}(1-0.999768) \\ &\approx 1.16\times 10^{-4}. \end{aligned} \tag{10}$$

このように $\cos\theta$ が 1 にごく近いとき $1-\cos\theta$ は桁落ちを起こすので $1-\cos\theta$ を計算するときは $2\sin^2(\theta/2)$ で, 計算した方が良い.

桁落ちの例 3 解の公式の桁落ち

二次方程式でも同種の問題がある. 例えば解の公式をつかって

$$2.718282x^2 - 684.4566x + 0.3161592 = 0 \tag{11}$$

の解を 10 進 7 桁の精度で計算してみる. まず判別式は

$$\begin{aligned} D &= (684.4566)^2 - 4\times 2.718282\times 0.3161592 \\ &\approx 468480.8 - 3.437639 \\ &\approx 46847.44 \end{aligned} \tag{12}$$

となる. ゆえに

$$\sqrt{D}\approx 684.4541 \tag{13}$$

で,

$$\begin{aligned} x_1 &= \frac{684.4566+684.4541}{2\times 2.718282} \\ &\approx \frac{1368.911}{5.436564} \\ &\approx 251.7970, \end{aligned} \tag{14}$$

$$
\begin{aligned}
x_2 &= \frac{684.4566 - 684.4541}{2 \times 2.718282} \\
&\approx \frac{0.0025}{5.436564} \\
&\approx 0.0004\underline{598493}.
\end{aligned}
\tag{15}
$$

(15) 式の分子で非常に近い 2 数の引き算をしているので有効数字が 2 桁まで落ちてしまっている.

(2) 式を分子の有理化によって全桁の計算をしたように解の公式においても全桁に近い計算をする方法がある. 二次方程式

$$
ax^2 + bx + c = 0 \tag{16}
$$

が 2 つの実解をもつとき, その解の公式

$$
x_{1,2} = \frac{-b \pm \sqrt{D}}{2a} \qquad (D \equiv b^2 - 4ac) \tag{17}
$$

において, $-b \pm \sqrt{D}$ の "$\pm$"は

「$b > 0$ なら $-$, $b < 0$ なら $+$」

として, 二つの解のうち, 一つだけを解の公式から求める. そのときの解を x_1 と定めて, もう一方の解 x_2 は「解と係数の関係」

$$
x_2 = \frac{\frac{c}{a}}{x_1} \tag{18}
$$

により計算すると良い.

(11) 式では

$$
\begin{aligned}
x_2 &\approx \frac{\left(\frac{0.3161592}{2.718282}\right)}{251.7970} \\
&\approx \frac{0.1163085}{251.7970} \\
&\approx 0.0004619138
\end{aligned}
\tag{19}
$$

となり, 最後の桁に「2」の狂いがあるだけである[5]. これは解の公式の分子を有理化した

[5] $x_2 = \frac{c}{ax_1}$ で計算すると全桁正しく計算できる.

$$
x_2 \approx \frac{0.3161592}{2.718282 \times 251.7970} \approx \frac{0.3161592}{684.4553} \approx 0.0004619136.
$$

ものを使っているともみなせる.

「解と係数の関係」の復習

本文に出てきた「解と係数の関係」の復習をし, 解の公式の分子を有理化したものを使っているとみなせることを確認する.

二次方程式

$$(px-q)(rx-u)=0 \tag{20}$$

を考えて, 解と係数の関係を導く. (20) 式を展開すると,

$$prx^2-(ps+qr)x+qs=0 \tag{21}$$

また, (16) と (21) を比較すると,

$$\begin{aligned} a&=pr, \\ b&=ps+qr, \\ c&=qs \end{aligned} \tag{22}$$

となる. さらに (20) 式の解をそれぞれ x_1,x_2 とおくと,

$$x_1=\frac{q}{p}, \quad x_2=\frac{s}{r} \tag{23}$$

となる. (22) 式と (23) 式から,

$$\frac{c}{a}=x_1x_2, \tag{24}$$

$$x_2=\frac{c}{ax_1} \tag{25}$$

である.

解の公式の分子を有理化する. 本文に沿って x_1 と x_2 を定義して,

$$x_1=\frac{-b\mp\sqrt{D}}{2a}, \tag{26}$$

$$x_2=\frac{-b\pm\sqrt{D}}{2a} \tag{27}$$

とする.

$$\begin{aligned} x_2 &= \frac{-b \pm \sqrt{D}}{2a} \\ &= \frac{(-b \pm \sqrt{D})(-b \mp \sqrt{D})}{2a(-b \mp \sqrt{D})} \\ &= \frac{b^2 - b^2 + 4ac}{2a(-b \mp \sqrt{D})} \\ &= \frac{2c}{(-b \mp \sqrt{D})} \end{aligned} \tag{28}$$

となる. 一方 (18) 式の x_1 に解の公式を当てはめると,

$$\begin{aligned} x_2 &= \frac{\frac{c}{a}}{x_1} \\ &= \frac{c}{a \times \frac{-b \mp \sqrt{D}}{2a}} \\ &= \frac{2c}{(-b \mp \sqrt{D})} \end{aligned} \tag{29}$$

となって同じになる.

桁落ちの例 4　重解の桁落ち

重解（もしくは二つの解が非常に近い）の場合も桁落ちが起こる. 例えば,

$$2.718282x^2 - 1.854089x + 0.3161592 = 0 \tag{30}$$

の判別式は, 10 進 7 桁の計算で

$$\begin{aligned} D &= (1.854089)^2 - 4 \times 2.718282 \times 0.3161592 \\ &\approx 3.437646 - 3.437639 \approx 0.000007 \end{aligned} \tag{31}$$

になり, D が 0 に近い[6]ということは非常に近い数字の引き算をすることになるので有効数字も少なくなる.

次に $\sqrt{D}$ を求める.

$$\sqrt{D} \approx 0.0026\underline{4575} \tag{32}$$

[6] D が 0 のとき重解になるので今回は 0 に近くなる

0 以外の数が格納された桁が増えたように見えるが、この場合 D の有効数字が 1 桁で $\sqrt{D}$ の有効数字も 1 桁になり, 下線以下には意味がない. この値を解の公式に入れると,

$$\begin{aligned} x_1 &\approx \frac{1.854089 + 0.002\underline{64575}}{2 \times 2.718282} \\ &\approx \frac{1.856\underline{735}}{2 \times 2.718282} \\ &\approx 0.3415\underline{273}, \end{aligned} \tag{33}$$

$$\begin{aligned} x_2 &\approx \frac{1.854089 - 0.002\underline{64575}}{2 \times 2.718282} \\ &\approx \frac{1.851\underline{443}}{2 \times 2.718282} \\ &\approx 0.3405\underline{539}. \end{aligned} \tag{34}$$

ただし, 有効数字は 3 桁で下線には意味がない.

重解の場合は計算手順をいくら工夫しても桁落ちを避けることはできない. 計算に伴う誤差がグラフを描いたときの「線」の太さを増加させるようなものだと考えれば, 有効数字を増やすということはグラフと軸との交点の大きさを小さくしていくことになる. しかし, 今回の場合は線が太くなっているのでグラフと軸の交点の幅も大きくなる. しかも, 重解なのでグラフと軸の接している長さが長くなる. その分点がより大きくなってしまうので誤差を減らすことが難しくなる (図 2.1 参照).

一般に「m 重解の有効数字の桁数は, 計算桁数の $\frac{1}{m}$ になる」といわれている[7].

大まかな誤差の求め方

多項式 $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ において特定の x に対して計算するときに生じる誤差 δ を求める. x に含まれる誤差をだいたい ϵ としたら, 多項式の表現誤差 δ は,

$$\delta = P(x + \epsilon) - P(x) \tag{35}$$

[7] 「ニュートン法」[http://www.ep.sci.hokudai.ac.jp/~gfdlab/comptech/y2024/resume/070_newton/2024_0626-yoshis.pdf を参照されたい.

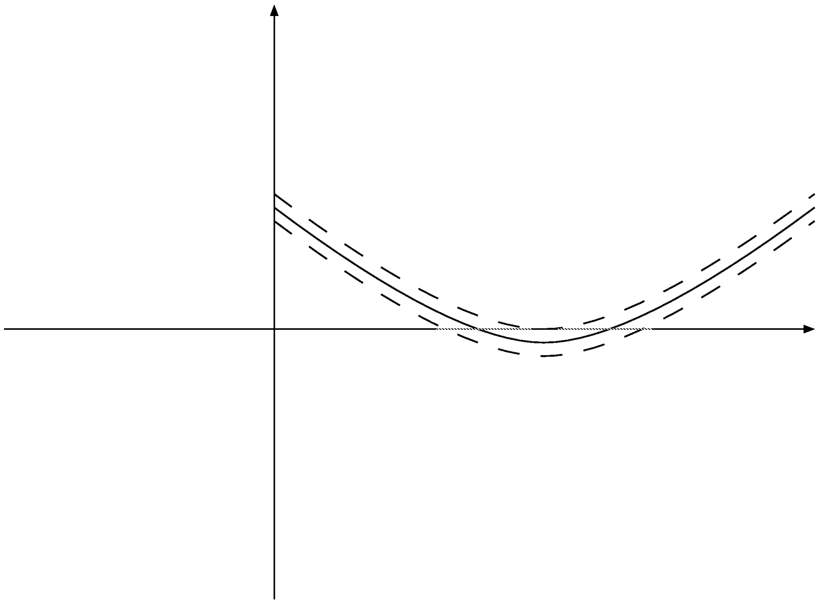

図 2.1　重解の誤差の概要. 点線は線の太さ

となる[8]. よって,

$$
\begin{aligned}
\delta &= a_0(x+\epsilon)^n + \cdots + a_{n-k}(x+\epsilon)^k + \cdots + a_{n-1}(x+\epsilon) + a_n \\
&\quad - \left(a_0x^n + \cdots + a_{n-k}x^k + a_{n-1}x + a_n\right) \\
&= a_0x^n\left(1+\frac{\epsilon}{x}\right)^n + \cdots + a_{n-k}x^k\left(1+\frac{\epsilon}{x}\right)^k + \cdots + a_{n-1}x\left(1+\frac{\epsilon}{x}\right) + a_n \\
&\quad - \left(a_0x^n + \cdots + a_{n-k}x^k + \cdots + a_{n-1}x + a_n\right). \qquad (36)
\end{aligned}
$$

$\dfrac{\epsilon}{x} \ll 1$ なのでそれぞれの項をマクローリン展開し, 1 次まで求めるとすると,

$$
\begin{aligned}
\delta &\approx a_0x^n\left(1+n\frac{\epsilon}{x}\right) + \cdots + a_{n-k}x^k\left(1+k\frac{\epsilon}{x}\right) + \cdots + a_{n-1}x\left(1+\frac{\epsilon}{x}\right) + a_n \\
&\quad - \left(a_0x^n + \cdots + a_{n-k}x^k + \cdots + a_{n-1}x + a_n\right) \\
&= n\frac{\epsilon}{x}a_0x^n + \cdots + k\frac{\epsilon}{x}a_{n-k}x^k + \cdots + \frac{\epsilon}{x}a_{n-1}x. \qquad (37)
\end{aligned}
$$

k 次の項の絶対値が他の項の絶対値に比べて大きいとする. そのとき, 多項式の計算の誤差は k 次の項によって大まかに見積もられる. よって,

$$
\delta \approx \left|k\frac{\epsilon}{x}a_{n-k}x^k\right|. \qquad (38)
$$

[8] この計算ではそれぞれの項の掛け算によって出る誤差を計算しているが, それぞれの項を足し合わせるときに出る誤差は無視している. また, それぞれの項を求める際に行う乗算 1 回あたりの誤差もだいたい同じものであるとしている.

さらにそれぞれの項の誤差は浮動小数点の表示による表現誤差なので, β 進 m 桁四捨五入で考えているとすると $\dfrac{\epsilon}{x}$ は「GFD ワークノート: 実数の浮動小数点表現と誤差その 2」[9) の (31) 式より

$$\frac{\epsilon}{x} \approx \frac{\beta^{-(m-1)}}{2} \tag{39}$$

である. よって,

$$\delta \approx \left| k\frac{\beta^{-(m-1)}}{2} a_{n-k} x^k \right| \tag{40}$$

となる.

代数方程式の解の計算

桁落ちは特殊な現象ではなく, 方程式を数値的に解くときなどは常に桁落ちを起こしているようなものである. 例として前述でも出てきた (11) 式

$$2.718282x^2 - 684.4566x + 0.3161592 = 0 \tag{11}$$

の解として (14) 式に近い $x = 251.7980$ を用いた場合を考える. このとき, (11) 式の左辺にホーナー (Horner) 法を用いて式を変形し代入すると,

$$\begin{aligned}
&(2.718282 \times 251.7980 - 684.4566) \times 251.7980 + 0.3161592 \\
&\approx ((684.4580 - 684.4566)) \times 251.7980 + 0.3161592 \\
&\approx 0.0014 \times 251.7980 + 0.3161592 \\
&\approx 0.35\underline{25172} + 0.3161592 \\
&\approx 0.66\underline{86764}
\end{aligned} \tag{41}$$

このように 684.4580 − 684.4566 の計算で激しい桁落ちが生じ, 結果として有効数字 2 桁の数字となる. この計算の誤差を大まかに見積もる. 10 進 7 桁四捨五入の計算をしているので相対誤差は $\dfrac{10^{-6}}{2}$. 1 次と 2 次の項の値がだいたい同じなので両方の項の誤差を考え

9) http://www.ep.sci.hokudai.ac.jp/∼gfdlab/comptech/y2025/resume/020_floatErr/2025_0513-sonodake.pdf を参照されたい.

なければならない. よって, (40) 式より誤差は,

$$\begin{aligned}\delta &\approx \left|2\frac{10^{-6}}{2}\times 2.718282\times(251.7980)^2\right| + \left|\frac{10^{-6}}{2}\times(-684.4566)\times 251.7980\right| \\ &\approx (2+1)\frac{10^{-6}}{2}\times 172345 \\ &\approx 0.26 \end{aligned} \tag{42}$$

となる. (11) 式の左辺の値が誤差とだいたい同じくらいになったらそれ以上 x の値を調節しても意味がなくなる. 一方, $x = 0.0004619157$ と (19) 式に近い値にして, (11) 式の左辺に上記と同様にホーナー法を用いた計算をすると,

$$\begin{aligned}&(2.718282\times 0.0004619157 - 684.4566)\times 0.0004619157 + 0.3161592 \\ &\approx (0.001255617 - 684.4566)\times 0.0004619157 + 0.3161592 \\ &\approx -684.4553\times 0.0004619157 + 0.3161592 \\ &\approx -0.3161606 + 0.3161592 \\ &\approx -0.0000014 \end{aligned} \tag{43}$$

となり, 今度は最後の引き算で桁落ちが生じている. さらに, 上記と同様に (11) 式の左辺の計算で生じる誤差を見積もってみる. 今回は 1 次の項が大きく 2 次の項はそれに比べてはるかに小さい. よって, (40) 式より

$$\begin{aligned}\delta &\approx \frac{10^{-6}}{2}\times(-684.4566)\times 0.0004619157 \\ &\approx \frac{10^{-6}}{2}\times 0.32 \\ &\approx 0.0000002 \end{aligned} \tag{44}$$

となる. この値より精度良くは求まらない.

この節で見たように同じ式の零点を求める場合, どのくらい零点に近くなればよいかを決めるには加減算で生じる誤差を考えなければならない.

代数方程式の解を求めるために, 解の近くで多項式の値を計算するときには, ある大きさの絶対値をもった項の和や差で, 結果が 0 に近くなるような計算をするために必ず桁落ちが生じてしまう. よって, 数値計算をするうえでは乗除算より加減算のほうがはるかに注意が必要になる.