数値計算　第３回
今回の講義の主題

- テーマ：　計算機による数値の表現と演算はどうなっているか
  - 有限精度しかない計算機メモリ
  - 全ての数はどう表現できるか？

数学では考えなくてもいいことが
数値計算ではおきることを学ぶ

今回の演習問題

- 演習3−1：　10進数を2進数に変換
- 演習3−2：　計算機で扱う最小の絶対値を計算
- 演習3−3：　計算機で扱う最大の絶対値を計算
- 演習3−4：　1+x=1になるxを計算

1.1　計算機と数値計算

電卓での計算の仕組み

- 1）データと計算手順書をもらいノートに記録
- 2）計算手順書にしたがい，途中結果をノートに記録しつつ電卓で計算
- 3）最終結果をノートから読み上げて報告

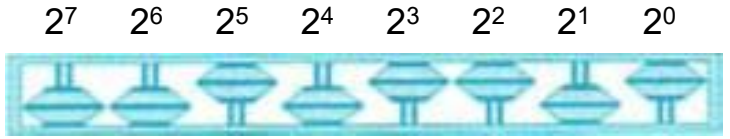
計算機による計算の仕組みも同じ

- 計算手順書　⇒　プログラム
- ノート　⇒　メモリ
- 電卓　⇒　演算部

教科書1ページ

メモリと数値表現

8桁の2進数をそろばんで表現する
8ビット（0 から　255）



2^5　　$+2^3+2^2$　　$+ 2^0 = 45$

パソコンは，4Gバイト（32Gビット）のメモリ空間
⇒　320億桁（32×10^9）の2進数が扱える

MATLAB での演算1

```
% リスト3-1
% 2進数の表現
clear
f=[ 0 0 1 0 1 1 0 1]
%2のベキ乗のデータ列
bp= 2.^(length(f)-1:-1:0)
%各桁を2のベキ乗する
f2=f.*bp
% 0.  0.   32.   0.   8.   4.   0.   1.
%全桁を足し合わせる
sum(f2)
```


計算法

- 計算コスト
  - 計算時間
  - メモリ量
- 計算機による制約
  - 有限精度で計算　誤差が避けられない
  - 誤差　なるべく小さく
- 同じ結果を得るために複数の計算法
  - なるべく計算コストが少なく，誤差の小さい計算法がよい

■例1■　3つの例

教科書2ページ

		1)	2)	3)
真値	x	1.232	1000	1000
近似値	x’	1.234	1001	999
誤差	x’-x	2×10^{-3}	1	-1
絶対誤差	|x’-x|	2×10^{-3}	1	1
相対誤差	|(x’-x)/x|	1.6×10^{-3}	10^{-3}	10^{-3}
10進有効桁数	$-\log_{10}\|(x'-x)/x\|$	2.8	3	3

MATLAB での演算例

デフォルト

```
% リスト 3-2 例1の表を計算
format short  % 固定小数点書式
              % 小数点以下4桁
x      =[1.232 1000 1000];  % 真値
xdash  =[1.234 1001 999];   %近似値
error  =xdash-x  % 誤差
aerror  =abs(xdash-x)  % 絶対誤差
rerror  =abs(xdash-x)./abs(x)   % 相対誤差
ddigit  =-log10(abs(xdash-x)./abs(x))   % 10進有効桁
```

参考　MATLABの表示形式 short long

入力 例	表示形式
format short; pi % 固定小数点(デフォルト)	3.1416
format shortE; pi % 指数付きshort	3.1416e+00
format long; pi % 固定小数点(倍精度)	3.141592653589793
format longE; pi % 指数付きlong	3.141592653589793e+00

1.2　実数の記録と演算

国際規格IEEE754

- 機械演算
  - 加減乗除と平方根のみハードウェアで演算
  - 残りの演算は？
- 機械実数
  - 正規数
  - 0
  - 副正規数 (正規数より絶対値が小さい数)
  - ±∞
  - NaN (Not a Number, 非数，不正な計算結果)

教科書3ページ

２進浮動小数点数

- 計算機での実数の表現方法
  - ２進　　２進数を使う
  - 浮動小数点　　小数点の位置が動く

$$10101 \Rightarrow 1.0101\times 2^4 \Rightarrow (-1)^0 1.0101\times 2^4$$

$$-0.011 \Rightarrow -1.1\times 2^{-2} \Rightarrow (-1)^1 1.1\times 2^{-2}$$

符号　仮数　指数

教科書3ページ

２進浮動小数点数　つづき

$$a = \pm 2^e f = \pm 2^e (f_0 . f_1 f_2 \cdots f_n)_2 = \pm 2^e \sum_{k=1}^{n} 2^{-k} f_k$$

- 符号　$\pm$
- 指数　e
- 仮数　$(f_0 f_1 f_2 \cdots f_n)_2$

メモリ上の表記

符号部	指数部	仮数部
s	$e_m \cdots e_1 e_0$	$f_1 f_2 \cdots f_n$

1から始める（正規数）

$$a = (-1)^s 2^e f = (-1)^s 2^e (1.f_1 f_2 \cdots f_n)_2 = (-1)^s 2^e \sum_{k=1}^{n} 2^{-k} f_k$$

単精度（32ビット）：符号部１ビット，指数部m＝7，仮数部n＝23
倍精度（64ビット）：符号部１ビット，指数部m＝10，仮数部n＝52

演習3－1

次の10進数は2進数でどう表現されるか
（MATLABで計算）
(1)　7
(2)　200
(3)　1024

ヒント：　MATLABの関数dec2binを使う

正規数の絶対値の限界

- 正規数の定義式

$$e_{\min} \le e \le e_{\max}$$

$$a = (-1)^s 2^e (1+f) = (-1)^s 2^e (1.f_1 f_2 \cdots f_n)_2$$

- 絶対値の限界

$$A_{\max} = 2^{e_{\max}} (1.11\cdots1)_2 = 2^{e_{\max}} (2 - 2^{-n}) \approx 2^{e_{\max}+1} = 2^{128}$$ （単精度）

$$A_{\min} = 2^{e_{\min}} (1.00\cdots0)_2 = 2^{e_{\min}} (1) = 2^{e_{\min}} = 2^{-126}$$ （単精度）

  – 絶対値の限界を超えたときの表現
    - 大きい場合：オーバーフロー　$\pm\infty$
    - 小さい場合：アンダーフロー　副正規数または0

IEEE 754 単精度と倍精度の定数

	ビット長	仮数部ビット長	Emin	Emax	相対誤差 u	Amin	Amax
単精度	32	23	-126	127	6.0×10^{-8}	1.2×10^{-38}	3.4×10^{38}
倍精度	64	52	-1022	1023	1.1×10^{-16}	2.2×10^{-308}	1.8×10^{308}

教科書5ページ

機械演算と丸め

- 丸め：計算した後に桁数を正規数の表現に収まるよう計算値を修正すること
- 教科書の丸め計算の例(仮数部は3桁)

$$2^{e}(1.11)_2 + 2^{e-1}(1.01)_2$$
$$\rightarrow 2^{e+1}(1.01)_2$$

実際に手計算をして答えが合うかやってみる

教科書5ページ

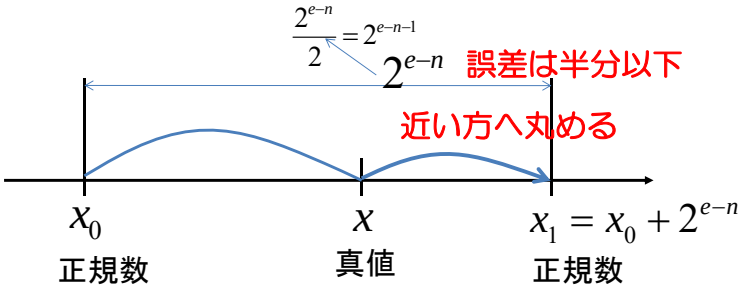
丸め演算の手計算

$$2^{e}(1.11)_2 + 2^{e-1}(1.01)_2$$
$$= 2^{e}(1.11)_2 + 2^{e}(0.101)_2$$ 指数部桁揃え
$$= 2^{e}(10.011)_2 = 2^{e+1}(1.0011)_2$$
$$\rightarrow 2^{e+1}(1.01)_2$$ 仮数部の小数点合わせ

丸め（3桁の数に修正）
ふつうは0捨1入（四捨五入）する

丸め誤差

- ０捨１入による誤差　⇒　丸め誤差
- 丸め誤差の見積り
  - ⇒　相対誤差
  - ⇒　絶対誤差 $\dfrac{|\hat{x}-x|}{|x|} \le \dfrac{2^{e-n-1}}{2^{e}} = u = 2^{-n-1}$



演習3−2

問題　浮動小数点で最小の絶対値を計算する
ヒント：　2による除算を繰り返し，アンダーフロー（0になる）する直前の数

演習3−3

問題　浮動小数点の最大の絶対値を計算してみよう
ヒント：　2による乗算を繰り返し，オーバーフロー(Inf)する直前の数

正確な最大値は規格で決まっている
教科書を参照のこと

演習3−4

問題　丸めの単位（浮動小数点の足し算で丸められてしまう数値）を計算してみよう
ヒント：　1.00に足しても値が変化しない数が丸めの単位. epsと比べてみよう

1+x=1になるx（≠0）を計算する

正確な丸めの単位は
規格で決まっている

教科書7ページ

1.4　四則演算による誤差伝搬

- 加減算での桁落ち現象
  - 二つの絶対値が近い同符号の数値の減算
  - 二つの絶対値が近い正負の数値の加算

0.123448…
-)0.123202…
0.000246…

6桁の精度が
3桁に減少

- 乗算や除算では桁落ち現象は起きない

教科書8ページ

桁落ち回避の３つの技法

1. 分子の有理化
2. 指数法則とsinh
3. 積和公式，倍角公式

例

xが小さいと桁落ちが発生

$$\sqrt{1+x}-1=\frac{(\sqrt{1+x}-1)(\sqrt{1+x}+1)}{(\sqrt{1+x}+1)}$$

$$=\frac{(1+x-1)}{\sqrt{1+x}+1}=\frac{x}{\sqrt{1+x}+1}$$

MATLABでの分子の有理化

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

$$\because (\sqrt{1+x}-1)(\sqrt{1+x}+1)$$
$$=(1+x-1)=x$$

```
% 分子の有理化
format long;
x=[1e-10 1e-15 1e-20];
y1=sqrt(1+x)-1
y2=x./(sqrt(1+x)+1)
```


MATLABでの分子の有理化の効果

	有理化なし	有理化あり
x	sqrt(1+x)-1	x/(sqrt(1+x)+1)
1.00000E-10	5.00000E-11	5.00000E-11
1.00000E-15	4.44089E-16	5.00000E-16
1.00000E-20	0.00000E+00	5.00000E-21

教科書9ページ

桁落ち回避の３つの技法

1. 分子の有理化
2. 指数法則とsinh
3. 積和公式，倍角公式

$$e^{3x}-e^{x}$$
$$=e^{2x}(e^{x}-e^{-x})$$
$$=2e^{2x}\frac{(e^{x}-e^{-x})}{2}=2e^{2x}\sinh\ x$$

教科書9ページ

桁落ち回避の３つの技法

1. 分子の有理化
2. 指数法則とsinh
3. 積和公式，倍角公式

$$\cos(a+x)-\cos a=-2\sin(a+x/2)\sin(x/2)$$
$$1-\cos 2x=2\sin^{2}x$$

今回のまとめ

- テーマ：　計算機による数値の表現と演算はどうなっているのか
  - 有限な精度しかない計算機メモリ
  - 全ての数は表現できるか？

計算機で表現できる数値
　有限精度，有限範囲
制約下で有効な計算方法を考える
　誤差の評価，精度のよい計算法

次回　関数計算

- テーマ：　四則計算（＋－×÷）しかできない計算機を使って，どのように関数を計算させるか