

10/31, 11/28 休, 1/31 休

### IEEE 標準の浮動小数点数(1)

- 倍精度の場合(単精度はお勧めしない)
  - 64 bit: 符号部 1 bit, 指数部 11 bit, 仮数部 52 bit
  - 値:  $y = (-1)^s 2^e f$ 

s	e'	f
---	----	---
  - 指数部 e は biased exponent ( $e = e' - 1023$ )
    - 正規化数の指数範囲  $-1022 \leq e \leq 1023$
  - 仮数部 f は頭の1を省略 ( $f = 1.f$ ) 実質 53 bit
    - マシンイプシロン  $\epsilon: \epsilon > 0, \Omega(1.0 + \epsilon) = 1.0 + \epsilon$
    - 丸め単位  $u = \epsilon/2: \Omega(x \text{ op } y) = (x \text{ op } y)(1 + \delta)$  with  $|\delta| < u$
    - $\epsilon = 2^{-52} \approx 2.22 \times 10^{-16}, u = 2^{-53} \approx 1.11 \times 10^{-16}$

単精度: 指数部 8 bit ( $e = e' - 127$ ), 仮数部 23 bit

### IEEE 標準の浮動小数点数(2)

- 絶対値が小さくなると... underflow
  - 非正規化数: 指数部  $e' = 0$  のとき,  $f = 0.f$  とする
    - 相対精度が低下する
  - 0: すべてのビットが 0 のときだが, 符号はつけられる
- 絶対値が大きくなると... overflow
  - 無限大: やはり符号付き
- 特殊な数
  - NaN (Not a Number): 0/0, sqrt(-1) など

### 二次方程式の解の公式(1)

- $0.02x^2 + 802x + 1.01 = 0$  を, 公式

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

を使って求める

- 倍精度:  $-4.0099998e+4$  &  $-1.2593517e-3$
- 単精度:  $-4.0100000e+4$  &  $-9.7412721e-4$
- 相対精度が恐ろしく悪い
  - 単精度がいかに使えないか...

符号反カシいては  
後ろろ

### 二次方程式の解の公式(2)

- 単精度と倍精度の計算を逐一比較すると

	単精度	倍精度
b	8.02	8.02
sqrt(d)	8.01999961	8.01999950
-b + sqrt(d)	0.00000039	0.00000050

- ほぼ同じ大きさの 2 つの数の引き算において, 相対精度が極端に下がる  
「桁落ち」(cancellation) という

### 無限和の近似計算(1)

- 無限和

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \approx 1.64493407$$

を単精度で計算すると,  $k = 4096$  で変化がなくなり, 和の値は  $1.64472532$

- 有効数字が 4 桁しかない!
- 大きい「部分和」に小さい「項」が足されるときに, 情報落ちが発生している

Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, SIAM

大+小 → 情報落ち

### 無限和の近似計算(2)

- 倍精度で計算すると  $1.64469002$
- 単精度で  $k = 1, 2, \dots, 4069$  の順:  $1.64472532$
- 単精度で  $k = 4096, 4095 \dots$  の順:  $1.64469004$ 
  - 逆順に計算した  $1.64469002$  すら

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \approx 1.64493407$$

- にはあまり近くない
  - 変化がなくなっても, 真値に近いとは限らない!
  - 真値に近いかどうかは, 別の方法で確認すべし



## 数値微分(1)

- 関数  $f(x)$  の微分の計算
  - 原理は簡単だが、関数が複雑だと長い式になる
    - 当然計算量が増える
- 微分の「定義」に戻って近似する

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

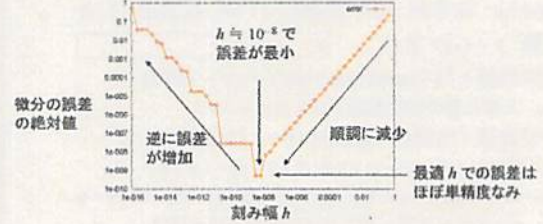
「数値微分」と呼ぶ

- たった2倍の計算量; デバッグにも使える
- $h$  はどうしたらよいか?

7

## 数値微分(2)

- 実験:  $f(x) = \sin(x)$  で,  $x = 1.0$  とした



8

## 一読の価値がある参考書

- 伊理正夫・藤野和建: 数値計算の常識, 共立
- 二宮市三ほか: 数値計算のつぼ, 共立
- いずれも, アルゴリズム以前に重要な, 数値計算の【センス】や【考え方】に詳しい
- アルゴリズムを集める(知っている)だけでは, 適切な数値計算はできない!

9

## 浮動小数の誤差モデル

Nicholas J. Higham: Accuracy and Stability of Numerical Algorithms, SIAM

- 四則演算と平方根に対して
 
$$fl(x \text{ op } y) = (x \text{ op } y) (1 + \delta) \quad |\delta| \leq u = 2^{-53}$$
- つぎのようにも書ける  $u = \sup\{|\delta|\}$  ではない

$$fl(x \text{ op } y) = \frac{x \text{ op } y}{1 + \delta}$$

- さらに,  $|\delta_i| \leq u, \rho_i = \pm 1, nu < 1$  のとき

$$\prod_{i=1}^n (1 + \delta_i)^{\rho_i} = 1 + \theta_n \quad |\theta_n| \leq \frac{nu}{1 - nu} =: \gamma_n$$

帰納法で丁寧に証明する

10

## 和・積を和に変換するアルゴリズム

- Möller-Knuth のアルゴリズム
  - $z = x + y; w = z - x; zz = (y - w) - ((z - w) - x)$
  - $x + y = z + zz$  が厳密に成り立つ
- Dekker のアルゴリズム Numerische Mathematik, 18, 224-242 (1971)
  - $p = x * c; hx = (x - p) + p; tx = x - hx;$
  - $p = y * c; hy = (y - p) + p; ty = y - hy;$
  - $p = hx * hy; q = hx * ty + tx * hy;$
  - $z = p + q; zz = ((p - z) + q) + tx * ty$
  - $x * y = z + zz$  が厳密に成り立つ

$cx$  は  $x$  の上 26 ビット (27 ビット目を右捨五入)

$$c = 2^{27} + 1$$

11

## 区間演算

- 数  $x$  の近似を, 区間で表す  $x \in [x, \bar{x}]$
- 区間どうしの演算を定義

$$[x, \bar{x}] + [y, \bar{y}] = [x + y, \bar{x} + \bar{y}]$$

$$[x, \bar{x}] - [y, \bar{y}] = [x - \bar{y}, \bar{x} - y]$$

$$[x, \bar{x}] * [y, \bar{y}] = [\min\{xy, x\bar{y}, \bar{x}y, \bar{x}\bar{y}\}, \max\{xy, x\bar{y}, \bar{x}y, \bar{x}\bar{y}\}]$$

$$[x, \bar{x}] / [y, \bar{y}] = [x, \bar{x}] * [1/\bar{y}, 1/y]$$

- これは厳密な計算での話
- 実際には, 浮動小数の丸めも考慮する

12

## 連続系アルゴリズム レポート課題3

### • 宿題(レポート)

- 提出状況に応じて、期末試験の点に最大 50 点を加算します
  - まったく提出しなくても、期末試験の点で成績が出ます
  - ただし、試験を受けなければ単位は出ません
- プログラムだけはメールの添付ファイルで reiji@is.s.u-tokyo.ac.jp に送付してください
  - 所属学科、学年、学籍番号、氏名をメール本文に明記すること
  - メールの題名(Subject)は「連続系レポート課題第3回 提出者氏名」とすること
- 手計算をするものや、計算結果と考察など、プログラム以外は A4 の紙 1 枚(裏もつかってよい)にまとめてください
  - 所属学科、学年、学籍番号、氏名をレポートの最初に明記すること
  - 次回の講義の前に集めます
  - あるいは PDF で A4 サイズ 2 ページでもよい(プログラムと一緒にメールで送付してください)

### • 問題1

- $a$  を実数定数として、以下の関数  $f(x)$  に対して、 $f(x) = 0$  の根を求める Newton 法の式を導け。

(1) 
$$f(x) = \frac{1}{x} - a$$

(2) 
$$f(x) = \frac{1}{x^2} - a$$

### • 問題2

- Newton 法により次の値を求めるプログラムを書け

(1)  $[1, 2]$  の実数  $x$  に対する逆数  $1/x$

(2)  $[1, 4]$  の実数  $x$  に対する平方根の逆数  $x^{-1/2}$

- ただし、使う浮動小数演算は加減算と乗算のみとする。反復回数は固定でもよい。

- いくつか値を与えて、精度を調べよ

### おまけ

- $z^2 - 1 = 0$  の根を求める複素 Newton 法のプログラムを書け
- $z = x + iy$  ( $-1 \leq x \leq 1, -1 \leq y \leq 1$ ) を初期値として上記のプログラムの反復回数を得よ
- $z$  に対して反復回数をプロットせよ

## C 言語補足

### • printf の書式いろいろ

- 科学技術形式

```
printf("%e\n", x);
```

- 3.1415e+00 のような出力になる(指数部の桁数はシステムや表現する数によって3桁のことも)
- これは  $3.1415 \times 10^0$  の意味
- 入力でも使える: 3e+12 とか

- 有効数字の設定

```
printf("%14.6e\n", x);
```

- 全体で 14 文字、小数点以下 6 桁
- 3.141592e+00 のような出力になる
- 符号、整数部、小数点、指数部の文字数をよく覚えて、十分な文字数を配置する
  - 足りない時にはいるんことがおきる

- 整数を出力

```
printf("%8d\n", i);
```

- これだと 8 桁で表示

### • 関数

- 似たような計算を繰り返す時には関数を使う
- まずは関数の「宣言」

```
double y(double x) {
    return 3.0 * x + 1.0;
}
```

- 最初の double は「返り値」の型
- つぎの y は関数の名前
- つづく丸カッコ内 double x は「引数」の型と名前
  - これは関数本体で変数同様に扱われる
- 最後の波カッコ内は関数の「本体」(計算内容)
  - ここでは return 文しかない
- return 文では、関数が返す値を示すとともに、そこで関数の計算が終わることを意味する

- 次に関数の呼び出し

- 別の関数などにおいて式の一部として書ける

```
dy = (y(1.1) - y(1.0)) / 0.1;
```

- これは  $((3.0 * 1.1 + 1.0) - (3.0 * 1.0 + 1.0)) / 0.1$  が  $dy$  に代入されることになる

## 関数のつづき

- 関数内で、引数は変数と同じ扱い

- 代入もできるが、呼び出し元に影響はない

```
double f(double x, double y) {
    x *= y;
    return x * x;
}
```

```
int main(void) {
    double a = 2.0, b = 2.0, c;
    c = f(a, b);
    printf("%f %f %f\n", a, b, c);
    return 0;
}
```

- 出力は 2.00 2.00 16.00 である
  - つまり、main 中の a は変更されない
- 関数 f を実行する際には、main 中の a の値が、f 中の x にコピーされる

- おまけ

- 関数 f の宣言は、引数が複数ある例にもなっている
- x \*= y は x = x \* y と同じこと
- main にあるように変数の宣言時に初期値を指定できる
  - 初期値を指定しないと、無意味な値が入っているので、使うときにはまっとうな値を代入してやらないといけない

- 関数ごとに変数は別物

```
void f(void) {
    int a;
    a = 1;
}
```

```
int main(void) {
    int a = 0;
    f();
    return a;
}
```

- f は引数のない関数、返り値のない関数の例
  - 呼び出しかたも確認してほしい(fortran と違う)
- main 内でも f 内でも、いずれも a という変数が宣言されているが、これらは「別物」
  - それぞれに違う値をしまっておける
- 上記の例では main は 0 を返す(1 ではない)

## 制御構文一式

- 文

- C 言語における実行の単位
- 例: 代入文     x = 0;
- 例: 関数呼び出し   printf(ほにやら);
- 例: 空文         ;

- 複文

- { と } の間に、複数の文がある
- ( から ) までの中身が、1 つの文と同等になる

- if 文

- 書き方1: if(条件) 文
- 書き方2: if(条件) 文 else 文
- 入れ子に注意

```
if(条件1)
    if(条件2)
        文1
    else
        文2
```

```
if(条件1) {
    if(条件2)
        文1
    } else
        文2
```

- 条件 (a <= 0 のたぐい)

- C 言語の条件は、0 が偽で、それ以外が真
- if(1) 文 | else 文2 は、常に文1 が実行される

- while 文

- 書き方: while(式) 文
- 「式」が成り立っている限り、「文」が繰り返される
- do 文というものもある

- 無限ループ

- while(1) 文
- for(;;) 文

- break 文

- 書き方: break;
- 最内側 for ループや while ループから抜け出す

```
while(1) {
    前半の計算
    if(終了条件)
        break;
    後半の計算
}
```

```
for(c=0; c<MAX; c++) {
    前半の計算
    if(終了条件)
        break;
    後半の計算
}
```

- Continue 文

- 最内側 for ループや while ループの、現在の反復の残りの処理をスキップして、次の反復に移る



## m 重根の場合の収束速度

- 同様に Taylor 展開: 重根より  $f(x^*) = f'(x^*) = 0$

$$f(x) = f(x^*) + (x-x^*)^m \frac{f^{(m)}(x^*)}{m!} + (x-x^*)^{m+1} R$$

$$f'(x) = f'(x^*) + (x-x^*)^{m-1} \frac{f^{(m)}(x^*)}{(m-1)!} + (x-x^*)^m R'$$

$$x - \frac{f(x)}{f'(x)} - x^* = x - x^* - \frac{(x-x^*)^m \frac{f^{(m)}(x^*)}{m!} + (x-x^*)^{m+1} R}{(x-x^*)^{m-1} \frac{f^{(m)}(x^*)}{(m-1)!} + (x-x^*)^m R'}$$

$$= x - x^* - (x-x^*) \frac{1}{m} \frac{f^{(m)}(x^*) + r}{f^{(m)}(x^*) + r'} \quad \frac{f^{(m)}(x^*) + r - r'}{f^{(m)}(x^*) + r'}$$

$$= \left(1 - \frac{1}{m}\right)(x-x^*) + (x-x^*)^2 \frac{1}{m} \frac{r'-r}{f^{(m)}(x^*) + r'}$$

これは 1 次収束: しかし更新量を  $m$  倍すれば 2 次収束になる

1

## Newton 法の減速

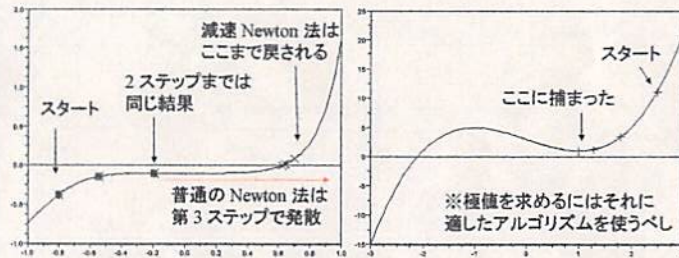
- Newton 法は遠くの解に飛んでいってしまうことがある
- 遠くに行かないように「減速」とすると有効なことがある
- 減速 Newton 法
  - 適当なパラメタ  $0 < \beta < 1$  と  $\lambda > 1$  を決めておく
  - もとの Newton 法 1 ステップを  $x^{(k+1)} = x^{(k)} + d^{(k)}$  として
 
$$\|f(x^{(k)} + \lambda^{-j} d^{(k)})\| \leq (1 - (1 - \beta) \lambda^{-j}) \|f(x^{(k)})\|$$
 となる最小の正整数  $j$  を求め、 $x^{(k+1)} = x^{(k)} + \lambda^{-j} d^{(k)}$ 
    - 条件を満たす  $j$  が必ず存在し、 $J(x^*)$  が正則なら  $x^*$  の十分近くで  $j=0$  となる (標準 Newton 法と一致する) と言える

2

## 減速 Newton 法の効果と限界

- $f(x) = \exp(x^5) - 1.1$

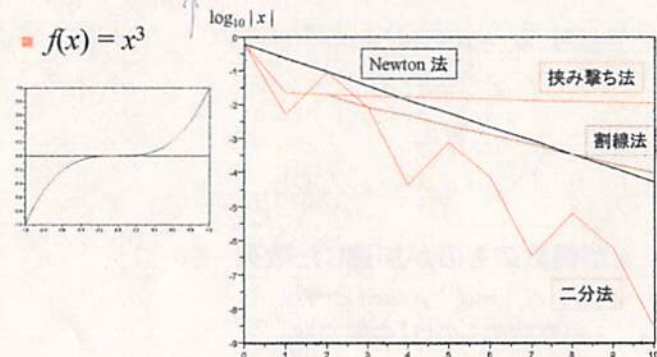
- $f(x) = x^3 - 3x + 3$



3

## 二分法のロバスト性(2)

- $f(x) = x^3$



4

反例あり





### 数列の収束

$a_1, a_2, a_3, \dots$  が  $a$  に収束するとは

1次収束. 十分大きい  $n$  に対し,

$0 < M < \epsilon$  なる  $M$  が存在し,

$$|a_{n+1} - a| \leq M |a_n - a|$$

$p$ 次収束 ( $p > 1$ )

十分大きい  $n$  に対し,  $0 < M < \epsilon$  なる,

$$|a_{n+1} - a| \leq M |a_n - a|^p$$

### 非線形方程式

1変数  $x$  の連続関数  $f(x) = 0$

$f(a) < 0$  かつ  $f(b) > 0$

二分法

$$f(a) f(b) < 0$$

$a < c < b$  なる  $f(c) = 0$  が

1つ以上存在する。

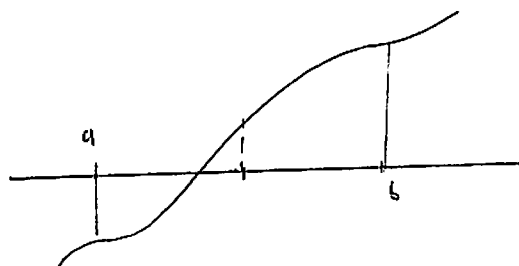
$$c := \frac{a+b}{2}$$

$f(a) f(c) < 0$  ならば  $b := c$

else  $a := c$

$\Rightarrow a, b$  が連続関数  $f(x)$  の連続区間で

最小の差  $|a-b|$  が  $\epsilon$  以下になるまで繰り返す



$M = \frac{1}{2}$  の収束.

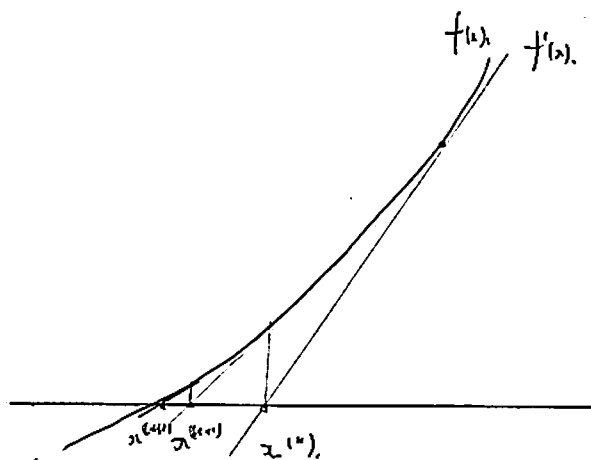
Newton (-Raphson) 法.

方程式  $f(x) = 0$

1回でも微分可能

Taylor 展開. 反復回法.

$$f(x^{(k+1)}) \doteq f(x^{(k)}) + (x^{(k+1)} - x^{(k)}) f'(x^{(k)}) = 0 \text{ として}$$



これを繰り返す.

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

多分又  $\forall \epsilon > 0$  迄行かない

これを繰り返す.

単根の場合, 根  $x^*$  を用いて

$$f(x^*) = f(x) + (x^* - x) f'(x) + \frac{(x^* - x)^2}{2} f''(\xi) = 0.$$

$$x^* - x^{(k+1)} = x^* - x^{(k)} + \frac{f(x^{(k)})}{f'(x^{(k)})}$$

$$= x^* - x^{(k)} - \frac{(x^* - x^{(k)}) f'(x^{(k)}) + \frac{(x^* - x)^2}{2} f''(\xi)}{f'(x^{(k)})}$$

$$= - \frac{(x^* - x^{(k)})^2}{2} \frac{f''(\xi)}{f'(x^{(k)})} \quad \text{--- 根  $x^*$  について  $f' = 0$  ではない ---}$$

これを繰り返す.  
2次収束である = 1/2 の収束.

(多分又  $\forall \epsilon > 0$  迄行かない)

Newton 法の収束の定理.

$$|f(x^{(k+1)})| \leq H(x^{(k+1)}) |x^{(k+1)} - x^*|^2$$

である.



m重根のとき

$$M = 1 - \frac{1}{m}$$

の1次収束である。



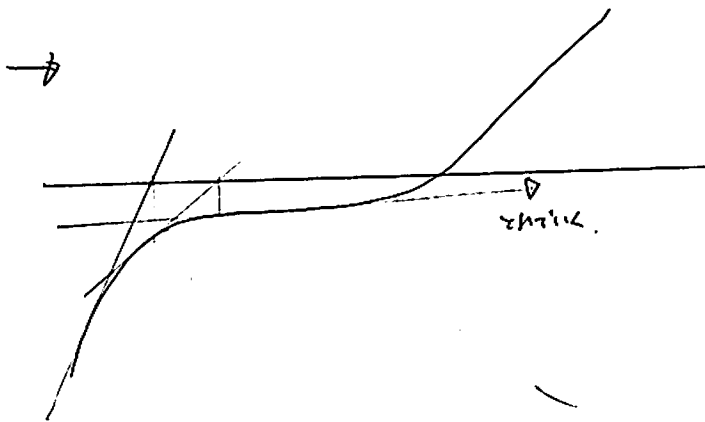
(m=2のとき)

2分法である。

(  
 1次収束である  
 m重根の時2分法である )

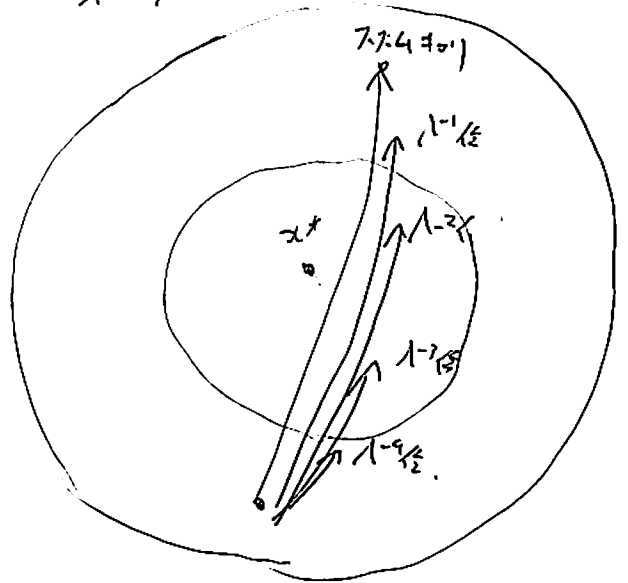
Newton法の

収束速度



Newton法の

収束速度



Newton法の

① 1次収束であるNewton法の1次収束  
(2次収束である)

② 関数が悪くない、更新が速く短時間で

$$\textcircled{3} \|f(x^{(k+1)})\| < \|f(x^{(k)})\|$$

1分法よりも速い  
2分法よりも速い

Newton法の収束速度

$$a_n = a + c_1 \lambda_1^n + c_2 \lambda_2^n + \dots$$

$$1 > |\lambda_1| > |\lambda_2| \dots$$

そこで  $n \rightarrow \infty$  とき、 $\lambda_1$  が既知の場合 ( $c_1$  は未知)

$a_{n+1}$  と  $a_n$  として  $c_1 \lambda_1^n$  を消す

$$a_{n+1} = a + c_1 \lambda_1^{n+1} + c_2 \lambda_2^{n+1} + \dots$$

$$\rightarrow \lambda_1 a_n = \lambda_1 a + c_1 \lambda_1^{n+1} + c_2 \lambda_1 \lambda_2^{n+1} \dots$$

$$a_{n+1} - \lambda_1 a_n = (1 - \lambda_1) a + c_2 (\lambda_2 - \lambda_1) \lambda_2^{n+1}$$

$$\therefore \underbrace{\frac{a_{n+1} - \lambda_1 a_n}{1 - \lambda_1}}_{:= a_n^{(1)}} = a + c_2 \underbrace{\frac{\lambda_2 - \lambda_1}{1 - \lambda_1}}_{\hat{a}_1} \lambda_2^{n+1}$$

故に右辺は  $n \rightarrow \infty$  とき  $a_n^{(1)}$  ( $u_1^{(1)}, u_2^{(1)}, \dots$ )

は  $a$  に収束し、 $a_n \neq a$  となる。

また、 $\lambda_2 \in \mathbb{R}$  と仮定する。

$$a_n^{(2)} = \frac{a_{n+1}^{(1)} - \lambda_2 a_n^{(1)}}{1 - \lambda_2}$$

Aitken 法

$\lambda_1$  が既知

$a_{n+2}, a_{n+1}, a_n$  として  $c_1$  と  $\lambda_1$  を消す

$$\lambda_1^2 \frac{a_n - a_{n-1}}{a_{n-1} - a_{n-2}}$$

これを  $R_i$  とし、 $n \rightarrow \infty$  とき  $\lambda$

$$a_n^{(1)} = a_n - \frac{(a_n - a_{n-1})^2}{a_n - 2a_{n-1} + a_{n-2}}$$