Operações de números reais no sistema de Aritmética do Ponto Flutuante e Estabilidade dos Algoritmos

MS211 – Cálculo Numérico – Turma C

Giuseppe Romanazzi

Agosto 2023

Conteúdo

1. Operações na Aritmética FP
2. Problemas bem postos
3. Estabilidade dos métodos numéricos e algoritmos

Sistema de Representação dos reais no sistema de virgula móvel (ou ponto flutuante)

Dado um x número real, a sua representação no sistema $FP(\beta, t, \ell, u)$ é $\bar{x} = \pm m\beta^e$ (denotada também com $fl(x)$)

- β é a base do sistema
- $m = 0.d_1 d_2 \ldots d_t$ é chamada mantissa do número $\bar{x}$
- t é o número de dígitos da mantissa
- e é o expoente, que satisfaz $e \in [\ell, u]$.

O sistema de representação $FP(\beta, t, \ell, u)$ é chamado sistema de representação em virgula móvel, ou em ponto flutuante. Pode ser chamado Aritmética de Ponto Flutuante uma vez que consideramos as operações aritméticas $+, -, *, /$ entre tais números do sistema FP.

Operações na Aritmética Finita

A operação $x\ op\ y$ onde "op" é uma operação aritmética (adição, subtração, multiplicação ou divisão) é feita em três passos na aritmética finita (ou sistema de ponto flutuante) $FP(\beta, t, \ell, u)$:

Passo 1 Representação de x e y em FP, ou seja determinar $\bar{x}, \bar{y} \in FP$ usando o arredondamento ou truncamento;

Passo 2 Efetuar a operação $\bar{x}\ op\ \bar{y}$. Este passo pode necessitar um alinhamento, como discutido a seguir no caso da soma;

Passo 3 Representação de $\bar{x}\ op\ \bar{y}$ em FP, obtemos $\overline{\bar{x}\ op\ \bar{y}}$

Sendo que $\overline{\bar{x}\ op\ \bar{y}}$ é o resultado final da operação $x + y$ em FP, ele pode ser indicado também com $\overline{x\ op\ y}$.

Alinhamento na Soma

Dados dois números na *FP*:

$$\bar{x} = m_x \beta^{e_x}, \quad \bar{y} = m_y \beta^{e_y}$$

a soma das mantissas m_x, m_y é exatamente a mantissa de $\bar{x} + \bar{y}$ somente se os exponentes foram iguais $e_x = e_y = e$
Neste caso $\bar{x} + \bar{y} = (m_x + m_y)\beta^e$.
Note que se for $m_x + m_y$ maior de 1 temos de mudar e.
Exemplo:
$\bar{x} = 0.5 \cdot 10^2, \bar{y} = 0.7 \cdot 10^2 \rightarrow \bar{x} + \bar{y} = 1.2 \cdot 10^2 = 0.12 \cdot 10^3$

No caso $e_x \neq e_y$ usamos o seguinte alinhamento (obtendo números com o mesmo exponente) e depois somamos os número alinhados:

$$\bar{x} + \bar{y} = \begin{cases} (m_x + m_y \beta^{e_y - e_x})\beta^{e_x} & \text{se } e_x > e_y \\ (m_x \beta^{e_x - e_y} + m_y)\beta^{e_y} & \text{se } e_y > e_x \end{cases}$$

Exemplo de adição em $FP(10, 4, -99, 99)$

Sejam $x = 93.702 * 10^2$, $y = 1.2723 * 10^1$, queremos calcular a soma $x + y$ na $FP = FP(10, 4, -99, 99)$ com truncamento e arredondamento

Passo 1 Representação de x e y em FP: $\bar{x} = 0.937 * 10^4$, $\bar{y} = 0.1272 * 10^2$

Passo 2 Precisamos de alinhar $\bar{x}, \bar{y}$:
$\bar{x} = 0.937 * 10^4, \bar{y} = (0.1272 * 10^{2-4}) * 10^4 = 0.001272 * 10^4$
então podemos realizar a soma
$\bar{x} + \bar{y} = (0.937 + 0.001272) * 10^4 = 0.938272 * 10^4$

Passo 3 Aproximamos a soma obtida no passo 2 em FP :

$$\overline{\bar{x} + \bar{y}} = \begin{cases} 0.9382 * 10^4 & \text{com truncamento} \\ 0.9383 * 10^4 & \text{com arredondamento} \end{cases}$$

Resultado final é $\overline{\bar{x} + \bar{y}}$ que pode ser denotado também com $\overline{x + y}$.

Erro absoluto da adição em *FP*

Por definição o erro absoluto da adição(ou da soma) é

$$E_A(x+y) := |x+y-\overline{x+y}|.$$

Suponhamos no seguinte que $x > \bar{x}$; $y > \bar{y}$; $x+y > \bar{x}+\bar{y} > \overline{x+y}$.
Mesmas conclusões são obtidas nos outros casos.
Sendo que $E_Ax := E_A(x) = x - \bar{x}$ e $E_Ay = y - \bar{y}$ obtemos que

$$x+y = \bar{x}+\bar{y}+E_Ax+E_Ay.$$

Se $\bar{x}+\bar{y}$ estiver já no sistema *FP* então $\overline{x+y} = \overline{\bar{x}+\bar{y}} = \bar{x}+\bar{y}$ e então

$$E_A(x+y) = E_Ax+E_Ay.$$

Em vez, se for $\bar{x}+\bar{y} \notin FP$

$$E_A(x+y) = x+y-(\bar{x}+\bar{y})+(\bar{x}+\bar{y})-\overline{x+y} = E_Ax+E_Ay+E_A(+)$$

onde $E_A(+) := (\bar{x}+\bar{y})-\overline{\bar{x}+\bar{y}}$ é o erro absoluto da aproximação do resultado da soma $\bar{x}+\bar{y}$ em *FP*.
O erro absoluto da soma é então a soma dos erros absolutos de representação de cada termo mais o erro absoluto da aproximação final.

Erro absoluto do resultado da operação genérica $\bar{x}\ op\ \bar{y}$ em $FP(\beta, t, \ell, u)$

Note que o erro absoluto $E_A(op) := |(\bar{x} op \bar{y}) - \overline{\bar{x} op \bar{y}}|$ é majorado como todos os erros absolutos da representação de um número (ver slide seguinte) ou seja vale que: se $\bar{x} op \bar{y} = m\beta^e$ com $0 < m < 1$ qualquer (que pode ter um numero qualquer de digitos)

$$E_A(op) \leq \begin{cases} \beta^{e-t} & \text{com truncamento} \\ \\ \frac{1}{2}\beta^{e-t} & \text{com arredondamento} \end{cases}$$

Majorantes erros absolutos e relativos em $FP(\beta, t, \ell, u)$

Note que dado um valor $z = m\beta^e$ com $0 < m < 1$ qualquer, obtemos os seguintes majorantes:
para o erro absoluto na representação de z

$$E_A z = |z - \bar{z}| \leq \begin{cases} \beta^{e-t} & \text{com truncamento} \\ \frac{1}{2}\beta^{e-t} & \text{com arredondamento} \end{cases}$$

para o erro relativo na representação de z

$$E_R z = \frac{E_A z}{|\bar{z}|} \leq u \equiv \begin{cases} \beta^{1-t} & \text{com truncamento} \\ \frac{1}{2}\beta^{1-t} & \text{com arredondamento} \end{cases}$$

u é chamada precisão da maquina ou unidade de arredondamento.
Note então que $E_A z$ e $E_A(op)$ tem os mesmos majorantes porque são associados a uma aproximação de um valor! O mesmo vale entre $E_R z$ e $E_R(op)$.

Exemplo (limite) em FP(10,4,-9,9,T)

$x = 0.111299\ldots9\ldots * 10^e$ é truncado como $\bar{x} = 0.1112$.

$|x - \bar{x}| = 0.99\ldots * 10^{-4} * 10^e \approx 10^{e-4}$.

Note como este exemplo é um caso limite e em geral temos que $|z - \bar{z}| < \beta^{e-t}$.

Exemplo (limite) em FP(10,4,-9,9,A)

$x = 0.1112499\ldots9\ldots10^e$ é aproximado com $\bar{x} = 0.1112$ na FP.

$|x - \bar{x}| = 0.499\ldots * 10^{-4} * 10^e \approx 0.510^{e-4}$.

Note como este exemplo é um caso limite e em geral temos que $|z - \bar{z}| < 0.5\beta^{e-t}$ quando usamos o arredondamento.

Erro relativo da adição em *FP*

Lembrando que $\overline{x+y} := \overline{\bar{x}+\bar{y}}$,

Se for $\overline{x+y} = \bar{x}+\bar{y}$ obtemos

$$E_R(x+y) = \frac{x+y-\overline{x+y}}{\overline{x+y}} = E_R(x)\frac{\bar{x}}{\bar{x}+\bar{y}} + E_R(y)\frac{\bar{y}}{\bar{x}+\bar{y}}$$

Se for $\overline{x+y} \neq \bar{x}+\bar{y}$

$$\begin{aligned} E_R(x+y) &= \frac{x+y-\overline{x+y}}{\overline{x+y}} = E_R(x)\frac{\bar{x}}{\overline{x+y}} + E_R(y)\frac{\bar{y}}{\overline{x+y}} + \frac{E_A(+)}{\overline{x+y}} \\ &= E_R(x)\frac{\bar{x}}{\overline{x+y}} + E_R(y)\frac{\bar{y}}{\overline{x+y}} + E_R(+) \end{aligned}$$

Notamos que o erro relativo da soma pode resultar menor da soma dos erros relativos se $x >> y$ ou viceversa ($x << y$).

Erros na Subtração

No seguinte vamos supor que $\overline{x-y} = \bar{x} - \bar{y}$. No caso $\overline{x-y} \neq \bar{x} - \bar{y}$ sabemos já que temos de adicionar ao erro absoluto o termo $E_A(op)$ e ao erro relativo o termo $E_R(op)$.
Se for $x - y > \bar{x} - \bar{y}$ e $x > \bar{x}, y > \bar{y}$ obtemos

$$E_A(x-y) = E_A(x) - E_A(y).$$

Nos outros casos podemos ter uma mudança de sinal do tipo $-E_A(x) + E_A(y)$ ou $E_A(x) + E_A(y)$. Podemos dizer que o erro absoluto da subtração é da mesma ordem dos erros absolutos da representação de x ou de y (ou do maior destes erros).
O erro relativo da subtração é em vez

$$E_R(x-y) = E_R(x)\frac{\bar{x}}{\bar{x}-\bar{y}} - E_R(y)\frac{\bar{y}}{\bar{x}-\bar{y}} = \frac{E_A x}{\bar{x}-\bar{y}} - \frac{E_A y}{\bar{x}-\bar{y}}.$$

Aqui há uma observação importante para fazer . . .

Cancelamento subtrativo

Se $\bar{x} \approx \bar{y}$ o erro relativo da subtração $x - y$ pode ser muito grande!
Este acontece normalmente quando x e y são muito perto, por isso neste caso diz se que:
temos um **erro de "cancelamento subtrativo".**

Uma boa prática nos algoritmos e métodos é de evitar subtrações de valores muitos próximos na aritmética finita.

Exemplo de erro de cancelamento subtrativo

Sejam dados $x = 0.43787 * 10^{-2}$ e $y = 0.43783 * 10^{-2}$ queremos computar na aritmética $FP(10, 4, -9, 9, A)$ a subtração $x - y$.

- Valor exato $x - y = 0.4 * 10^{-6}$,
- $\bar{x} = 0.4379 * 10^{-2}$, $\bar{y} = 0.4378 * 10^{-2}$, $\bar{x} - \bar{y} = 0.1 * 10^{-5}$
- $E_A(x - y) = |x - y - (\bar{x} - \bar{y})| = 0.6 * 10^{-6}$
- $E_R(x - y) = \frac{E_A(x-y)}{|\bar{x}-\bar{y}|} = \frac{0.6*10^{-6}}{1*10^{-6}} = 0.6 = \frac{60}{100}$

O erro relativo entre $x - y$ e o valor $\bar{x} - \bar{y}$ obtido na FP é de 60%.
Erro relativo grande!

Erros na Multiplicação

Sejam $x > \bar{x}$, $y > \bar{y}$, e tais que $\overline{xy} = \bar{x}\bar{y}$ está já na FP considerada. Então temos $E_A(xy) = xy - \bar{x}\bar{y}$ e

$$E_A(xy) = (\bar{x} + E_A(x))(\bar{y} + E_A(y)) - \bar{x}\bar{y} \approx \bar{x}E_A(y) + \bar{y}E_A(x)$$

Esta aproximação do erro é obtida desconsiderando o termo $E_A(x)E_A(y)$ que é esperado ser pequeno, veja slide 8.
Analogamente obtemos

$$E_R(xy) \approx \frac{\bar{x}E_A(y) + \bar{y}E_A(x)}{\bar{x}\bar{y}} = E_R(x) + E_R(y)$$

Nos outros casos ($x < \bar{x}$ ou $y > \bar{y}$) pode mudar o sinal dos adendos da soma final.
A multiplicação então como a soma não deixa o erro final relativo ou absoluto crescer muito, é uma ótima operação para ser usada nos métodos e algoritmos.

Erros na Divisão

Valem a seguintes formulas (a menos de troca de sinal) e se $\frac{\bar{x}}{\bar{y}}$ está na FP considerada

$$E_A(\frac{x}{y}) \approx \frac{\bar{y}E_A(x) - \bar{x}E_A(y)}{\bar{y}^2}.$$

O erro absoluto da divisão pode ser enorme se $\bar{y}$ é perto de zero: dividir para números pequenos leva a valores grandes e pode levar a erros ainda maiores!
Mas o erro relativo da divisão é sempre limitado porque vale

$$E_R(\frac{x}{y}) \approx E_R(x) - E_R(y).$$

Conclusão da análise dos erros das operações em *FP*

Nos métodos e algoritmos temos de tomar cuidado quando:

- usamos a subtração entre termos de valores similares
- e nas divisões com denominadores pequenos.

Se possível temos de evitar estes operações, e tentar de usar métodos com adições, multiplicações e divisões com denominadores não "pertos" de zero.

Problema bem posto

Um Problema P, definido num conjunto de dados de input Ω_I, é dito bem posto, se satisfaz o seguinte:

- Por cada dado de input $x \in \Omega_I$, existe uma única solução do problema P que será indicada com $P(x)$
- Cada vez que damos o mesmo dado de input x é esperado sempre a mesma solução $P(x)$
- A solução depende em maneira contínua dos dados de input. Este significa que pequenas variações dos dados de input:
  x, $x + \delta$, com $|\delta|$ pequeno
  levam a pequenas variações nas soluções:
  $|P(x + \delta) - P(x)|$ resulta ser pequena
  (ouseja é menor ou da mesma dimensão de δ)

Exemplo de problema bem posto

- Problema 1:
  Dado o valor x consideramos o problema de computar $P(x) = x + 3$.
  Se em vez de x for dado $x + \delta$, com $\delta > 0$ pequeno, obtemos a diferença (erro) na solução
  $|P(x + \delta) - P(x)| = |x + \delta + 3 - (x + 3)| = |\delta| = \delta$.
  Portanto se em vez de $x + 3$ computamos $x + \delta + 3$ cometemos um erro pequeno. **O problema é bem posto!**

Exemplos de problema mal posto

Problema 2: Computar dado x o valor $f(x) = e^x$. Se for dado $x + \delta$ com δ pequeno em modulo.
Usando a expansão em serie de Taylor com centro em x de uma função f diferenciável temos:

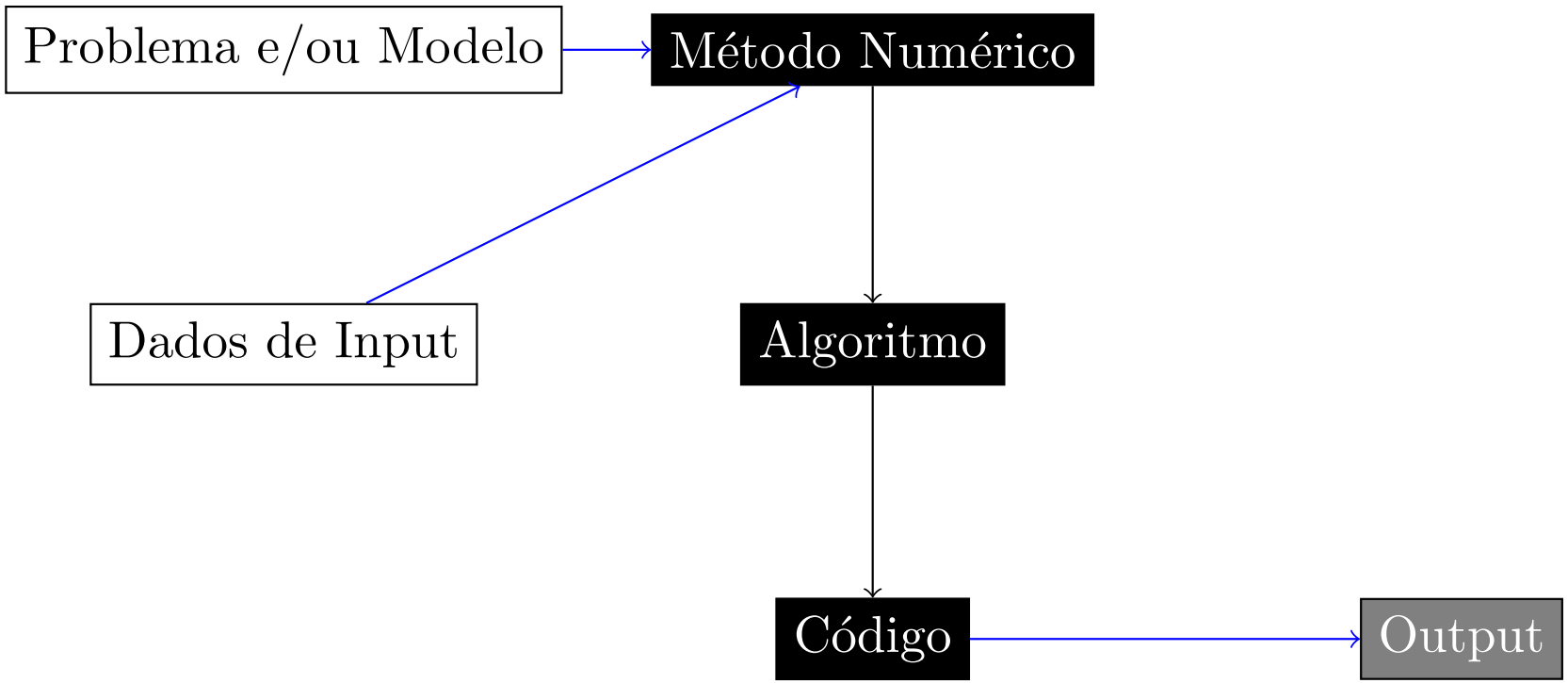
$$f(x+\delta) = f(x) + f'(x)\delta + f''(x)\frac{\delta^2}{2!} + f'''(x)\frac{\delta^3}{3!} + \ldots + f^{(k)}(x)\frac{\delta^k}{k!} + \ldots$$

O erro na solução é $|e^{x+\delta} - e^x| = |\delta e^x + \frac{\delta^2}{2}e^x + \frac{\delta^3}{6}e^x + \ldots| \approx const.|\delta|e^x$.
A ultima aproximação vale se for $|\delta|$ sufficientemente pequeno (com pelomenos $|\delta| < 1$) : porque vale que $|\delta| > \frac{\delta^2}{2} + \frac{|\delta|^3}{3!} + \ldots$ se $|\delta|$ for pequeno.
O erro na solução $|e^{x+\delta} - e^x| \approx |\delta|e^x$ depende de x e de δ.
Este erro é grande se x for grande independentemente se for $|\delta|$ pequeno, por isso **o problema é mal posto.**
Por exemplo, se $\delta = 0.01$ e $x = 10$, teremos um erro grande na solução $|e^{x+\delta} - e^x| \approx 0.01e^{10} = 220.2647$.
Note que $|e^{x+\delta} - e^x| = 221.3697$.

Observações

- É sempre recomendável não resolver problemas mal postos.
- Se o problema não tiver solução única por uma dado input, nunca vamos saber o erro que cometemos. Pense aos eventos/problemas como lançamento de moedas, sondagens ou todos os fenômenos sem uma solução certa, eles não vão poder ser resolvidos numericamente com um erro controlável com os métodos discutidos nesta disciplina.
- Note que a necessidade do problema ter uma variação continua na solução é muito importante porque **sempre temos variações nos dados de input** respeito os dados reais. Pense aos erros de representação ou inerentes ou random ou de modelagem, todos eles vão afetar os dados de input. Portanto se o método não for bem posto não temos a esperança de resolver ele acuradamente.
- Note que o Problema 2 anterior é um problema bem posto se x for pequeno e em vez é mal posto se x for grande, portanto **as vezes se reduzimos o domínio de input de um problema mal posto, podemos ter um problema bem posto**.

Problema e Método Numérico



Estabilidade dos métodos numéricos e algoritmos

Um **método numérico** (ou um algoritmo) diz se **estável**: se quando for aplicado a um problema bem posto, o método (ou o algoritmo) se tiver uma pequena variação nos dados de input leva a ter uma pequena variação dos dados de output. Observamos que

- se resolvemos com um método estável um problema mal posto, vamos ter grande variação no output por pequena variação no input;
- se resolvemos um problema bem posto com um algoritmo não estável podemos ter grande variação no output também;
- as vezes o mesmo método tem regiões de estabilidade onde se pegamos o input neste regiões o método resulta estável, se em vez pegamos os dados de input afora da região de estabilidade o método resulta ser instável (grande variação dos output associada a pequena variação dos dados de input).
- Dado um problema bem posto é quase sempre possível encontrar um método estável que aproxima a sua solução!

Referencia: "Métodos Numéricos" M. Cristina C. Cunha, Editora Unicamp

Solução de uma equação de segunda ordem (solução do binômio)

Resolvemos o problema bem posto $ax^2 + bx + c = 0$ com dois métodos um não estável (instavel) e um estável.

- Método 1: Computar
  $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$, $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$

- Método 2:
  Se $b < 0$ computar x_1 como feito no método 1
  e depois compute x_2 usando a relação $x_1 * x_2 = \frac{c}{a}$ obtendo assim
  $x_2 = \frac{c}{ax_1}$;

  Se $b > 0$ computar x_2 como feito no método 1
  e depois x_1 usando a relação $x_1 * x_2 = \frac{c}{a}$, $x_1 = \frac{c}{ax_2}$.

Análise do Problema, cancelamento subtrativo

O calculo de x_1, x_2 é um problema bem posto no caso que $b^2 - 4ac >= 0$.
Vamos analisar agora o que pode acontecer numericamente numa aritmética finita.
Admitindo que "a" seja diferente de zero (e não perto de zero) a única operação que pode dar um erro (relativo) de representação grande é a subtração no calculo de x_1, x_2:

- quando subtraímos quantidades próximas temos o erro do cancelamento subtrativo (ver slides 13-14).

Isso acontece no nosso problema :

- se $b^2 \approx 4ac$
- se $b \approx \sqrt{b^2 - 4ac}$ (no calculo de x_1 quando $b > 0$)
- se $-b \approx \sqrt{b^2 - 4ac}$ (no calculo de x_2 quando $b < 0$)

Análise do Problema, cancelamento subtrativo

1. $b^2 \approx 4ac$,
2. $b \approx \sqrt{b^2 - 4ac}$ (no calculo de x_1 quando $b > 0$),
3. $-b \approx \sqrt{b^2 - 4ac}$ (no calculo de x_2 quando $b < 0$).

- Notamos que o primeiro caso pode ser raro;
- O segundo e terceiro caso não podem acontecer contemporaneamente, porque dependem do sinal de b.
- **O segundo e terceiro caso são mais frequentes e acontecem sempre que** $b^2 >> 4ac$ ou seja se b é bastante grande respeito a e c.

Exemplo, método 1

Resolvemos usando 5 dígitos significativos na $FP(10, 5, -9, 9, T)$ o problema $x^2 - 100.223x + 1.2371 = 0$.
Observamos que $b = -100.22$, $b^2 = 10044$ e $4ac = 4.9484$ na FP com 5 dígitos, sem deslocar a virgula...
Prosseguindo $b^2 - 4ac = 10039$, depois teremos $\sqrt{b^2 - 4ac} = 100.19$. Notamos logo que $b^2 >> 4ac$ de quatro ordens de magnitude, portanto era esperado que $\sqrt{b^2 - 4ac} \approx |b|$.
Se usássemos o método 1:

$$\bar{x}_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{100.22 + 100.19}{2} = 100.2$$

$$\bar{x}_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{100.22 - 100.19}{2} = 0.015$$

É esperado ter um erro grande no calculo de x_2 sendo que fizemos uma subtração de dois valores similares, este é o caso 3 da slide anterior.

Erros, método 1

Se usarmos um computador (com mais dígitos) os resultados seriam $x_1 = 100.2106550053$ e $x_2 = 0.012344994651$. Notamos que temos um erro relativamente grande para o x_2!

$$E_R(x_2) = \frac{|x_2 - \bar{x}_2|}{|\bar{x}_2|} = 0.215 \approx 21,5\%$$

Temos um erro relativo grande do $21,5\%$ **no calculo de** x_2**.**
Em vez no calculo de x_1

$$E_R(x_1) = \frac{|x_1 - \bar{x}_1|}{|\bar{x}_1|} = 1.0633 * 10^{-4} \approx 0.0106\%$$

temos um erro relativo muito pequeno.
O método 1 é instável para resolver este problema do binômio no calculo de x_2**.**

Exemplo, método 2

Sempre com cinco dígitos significativos, usando o método 2 sendo que $b = -100.22 < 0$, computamos antes a raiz x_1 que não precisa de subtrações obtemos como antes $\bar{x}_1 = 100.2$, com
$E_R(x_1) = 0.0106\%$
x_2 é aproximada usando a relação $\bar{x}_2 = \dfrac{c}{ax_1}$ do que obtemos
$\bar{x}_2 = 0.012346$.
Observamos agora que conseguimos ter um erro relativo pequeno

$$\mathbf{E_R(x_2) = \frac{|x_2 - \bar{x}_2|}{|\bar{x}_2|} = 8.1438 * 10^{-5} \approx 0.00814\%.}$$

O método 2 é estável para resolver o problema do binômio.