IEEE Floating Point Arithmetic

Emre Mengi

Department of Mathemtics
Koç University
Istanbul, Turkey

November 14, 2011

- From here on we will analyze the accuracy of the algorithms in the presence of rounding errors.
- The numbers are stored in computers using a floating point arithmetic whose standards are set by IEEE.
- The rounding errors are consequences of the IEEE floating point arithmetic.

- From here on we will analyze the accuracy of the algorithms in the presence of rounding errors.
- The numbers are stored in computers using a floating point arithmetic whose standards are set by IEEE.
- The rounding errors are consequences of the IEEE floating point arithmetic.

- From here on we will analyze the accuracy of the algorithms in the presence of rounding errors.
- The numbers are stored in computers using a floating point arithmetic whose standards are set by IEEE.
- The rounding errors are consequences of the IEEE floating point arithmetic.

- 64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2}$$

  - 52 bits for the significand (mantissa)
  - 11 bits for the exponent
  - 1 bit for the sign

e.g.

$$(1.\underbrace{1}_{b_1}0\dots0\underbrace{1}_{b_{52}})_2 \times 2^{(00\dots010)_2} = \left(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-52}\right) \times 2^2$$

- 64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2}$$

  - 52 bits for the significand (mantissa)
  - 11 bits for the exponent
  - 1 bit for the sign

e.g.

$$(1.\underbrace{1}_{b_1}0\dots0\underbrace{1}_{b_{52}})_2 \times 2^{(00\dots010)_2} = \left(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-52}\right) \times 2^2$$

- 64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2}$$

  - 52 bits for the significand (mantissa)
  - 11 bits for the exponent
  - 1 bit for the sign

e.g.

$$(1.\underbrace{1}_{b_1}0\dots0\underbrace{1}_{b_{52}})_2 \times 2^{(00\dots010)_2} = \left(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-52}\right) \times 2^2$$

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.

- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).

- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

- Let x be any floating point number in double precision.

$$
\begin{array}{rcl}
-(1.11\ldots1)_2 \times 2^{1023} & \leq x \leq & (1.11\ldots1)_2 2^{1023} \\
-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} & \leq x \leq & ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\
\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} & \leq x \leq & \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}
\end{array}
$$

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.

- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).

- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

- Let x be any floating point number in double precision.

$$
\begin{array}{rcl}
-(1.11\ldots1)_2 \times 2^{1023} & \leq x \leq & (1.11\ldots1)_2 2^{1023} \\
-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} & \leq x \leq & ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\
\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} & \leq x \leq & \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}
\end{array}
$$

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.
- Let x be any floating point number in double precision.

$$
\begin{aligned}
-(1.11\ldots1)_2 \times 2^{1023} &\leq x \leq (1.11\ldots1)_2 2^{1023} \\
-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} &\leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\
\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} &\leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}
\end{aligned}
$$

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.
- Let x be any floating point number in double precision.

$$
\begin{aligned}
-(1.11\ldots1)_2 \times 2^{1023} &\leq x \leq (1.11\ldots1)_2 2^{1023} \\
-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} &\leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\
\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} &\leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}
\end{aligned}
$$

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.
- Let x be any floating point number in double precision.

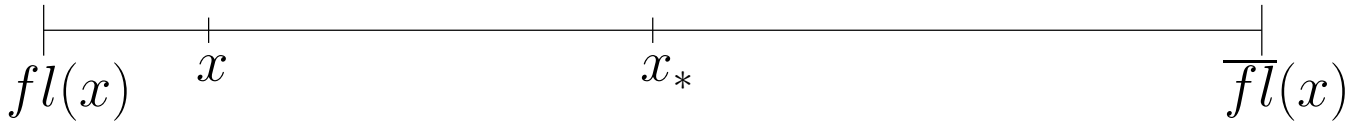
$$
\begin{aligned}
-(1.11\ldots1)_2 \times 2^{1023} &\leq x \leq (1.11\ldots1)_2 2^{1023} \\
-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} &\leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\
\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} &\leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}
\end{aligned}
$$

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.
- Let x be any floating point number in double precision.

$$
\begin{array}{rcl}
-(1.11\ldots1)_2 \times 2^{1023} & \leq x \leq & (1.11\ldots1)_2 2^{1023} \\
-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} & \leq x \leq & ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\
\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} & \leq x \leq & \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}
\end{array}
$$

ϵ_{mach} (machine precision or unit round-off error)

Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E$

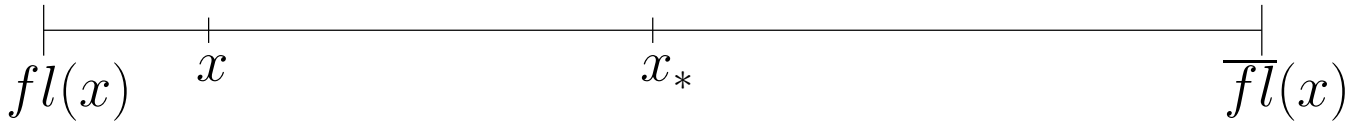
$x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \frac{\hat{s}\times 2^E + (\hat{s}+2^{-52})\times 2^E}{2} = (\hat{s} + 2^{-53}) \times 2^E$

- Relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 10^{-16} \quad (|s| \geq 1)$$

ϵ_{mach} (machine precision or unit round-off error)

Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E$

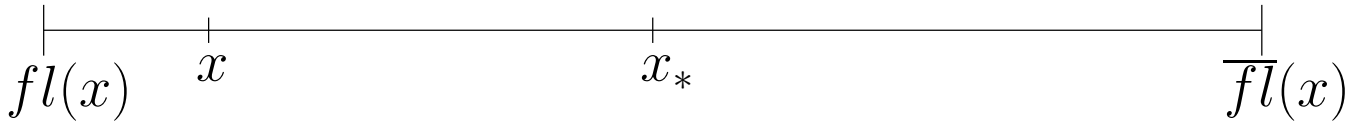
$x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \frac{\hat{s}\times 2^E + (\hat{s}+2^{-52})\times 2^E}{2} = (\hat{s} + 2^{-53}) \times 2^E$

- Relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 10^{-16} \quad (|s| \geq 1)$$

ϵ_{mach} (machine precision or unit round-off error)

Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E$

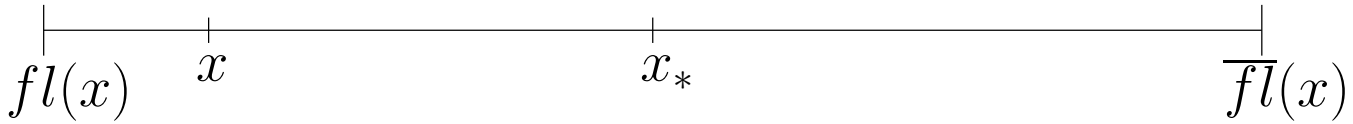
$x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \frac{\hat{s}\times 2^E + (\hat{s}+2^{-52})\times 2^E}{2} = (\hat{s} + 2^{-53}) \times 2^E$

- Relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 10^{-16} \quad (|s| \geq 1)$$

ϵ_{mach} (machine precision or unit round-off error)

Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E$

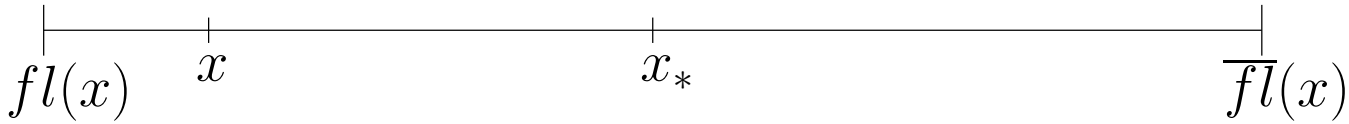
$x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \frac{\hat{s}\times 2^E + (\hat{s}+2^{-52})\times 2^E}{2} = (\hat{s} + 2^{-53}) \times 2^E$

- Relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 10^{-16} \quad (|s| \geq 1)$$

ϵ_{mach} (machine precision or unit round-off error)

Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E$

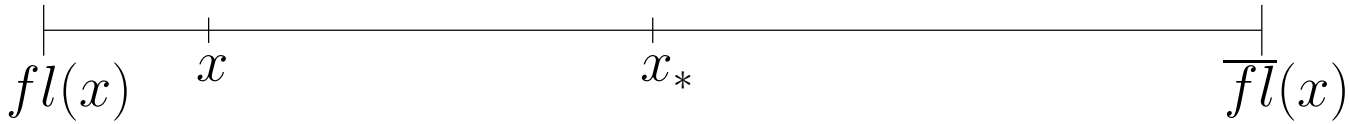
$x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \frac{\hat{s}\times 2^E + (\hat{s}+2^{-52})\times 2^E}{2} = (\hat{s} + 2^{-53}) \times 2^E$

- Relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 10^{-16} \quad (|s| \geq 1)$$

ϵ_{mach} (machine precision or unit round-off error)

Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E$

$x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \frac{\hat{s}\times 2^E + (\hat{s}+2^{-52})\times 2^E}{2} = \left(\hat{s} + 2^{-53}\right) \times 2^E$

- Relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 10^{-16} \quad (|s| \geq 1)$$

- Smallest non-zero number in absolute value
  - When $(a_1 a_2 \dots a_{11})_2 = 0$ the floating point number is in the (subnormalized) form

$$(0.b_1 \dots b_{52})_2 \times 2^{-1022}$$

  - The smallest number

$$(0.0 \dots 01)_2 \times 2^{-1022} = 2^{-52} \times 2^{-1022} = 2^{-1074} \approx 4.94 \times 10^{-324}$$

- Smallest non-zero number in absolute value
  - When $(a_1 a_2 \dots a_{11})_2 = 0$ the floating point number is in the (subnormalized) form

$$(0.b_1 \dots b_{52})_2 \times 2^{-1022}$$

  - The smallest number

$$(0.0 \dots 01)_2 \times 2^{-1022} = 2^{-52} \times 2^{-1022} = 2^{-1074} \approx 4.94 \times 10^{-324}$$

- Smallest non-zero number in absolute value
  - When $(a_1 a_2 \dots a_{11})_2 = 0$ the floating point number is in the (subnormalized) form

$$(0.b_1 \dots b_{52})_2 \times 2^{-1022}$$

  - The smallest number

$$(0.0 \dots 01)_2 \times 2^{-1022} = 2^{-52} \times 2^{-1022} = 2^{-1074} \approx 4.94 \times 10^{-324}$$

- Floating point operations or flops ($\oplus, \otimes, \ominus, \oslash$) in IEEE double precision
  - IEEE standards require the flops to satisfy

$$x \oplus y = fl(x + y)$$
$$x \ominus y = fl(x - y)$$
$$x \otimes y = fl(x \times y)$$
$$x \oslash y = fl(x / y)$$

where x and y are floating point numbers.

Examples in double precision

1. $1 \oplus 2^{-52} = 1 + 2^{-52}$, but $1 \oplus 2^{-53} = 1$
2. $$\begin{aligned}(1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl((1 + 2^{-52} + 2^{-52} + 2^{-104}) \times 2) \\ &= 2(1 + 2^{-51})\end{aligned}$$

- Floating point operations or flops $(\oplus, \otimes, \ominus, \oslash)$ in IEEE double precision
  - IEEE standards require the flops to satisfy

$$x \oplus y = fl(x + y)$$
$$x \ominus y = fl(x - y)$$
$$x \otimes y = fl(x \times y)$$
$$x \oslash y = fl(x / y)$$

where x and y are floating point numbers.

Examples in double precision

1. $1 \oplus 2^{-52} = 1 + 2^{-52}$, but $1 \oplus 2^{-53} = 1$
2. $$\begin{aligned}(1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl((1 + 2^{-52} + 2^{-52} + 2^{-104}) \times 2) \\ &= 2(1 + 2^{-51})\end{aligned}$$

- Floating point operations or flops $(\oplus, \otimes, \ominus, \oslash)$ in IEEE double precision
  - IEEE standards require the flops to satisfy

$$x \oplus y = fl(x + y)$$
$$x \ominus y = fl(x - y)$$
$$x \otimes y = fl(x \times y)$$
$$x \oslash y = fl(x/y)$$

where x and y are floating point numbers.

Examples in double precision

1. $1 \oplus 2^{-52} = 1 + 2^{-52}$, but $1 \oplus 2^{-53} = 1$
2. $$\begin{aligned}(1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl((1 + 2^{-52} + 2^{-52} + 2^{-104}) \times 2) \\ &= 2(1 + 2^{-51})\end{aligned}$$

- Floating point operations or flops $(\oplus, \otimes, \ominus, \oslash)$ in IEEE double precision
  - IEEE standards require the flops to satisfy

$$x \oplus y = fl(x + y)$$
$$x \ominus y = fl(x - y)$$
$$x \otimes y = fl(x \times y)$$
$$x \oslash y = fl(x/y)$$

where x and y are floating point numbers.

Examples in double precision

1. $1 \oplus 2^{-52} = 1 + 2^{-52}$, but $1 \oplus 2^{-53} = 1$
2. $$\begin{aligned}(1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl((1 + 2^{-52} + 2^{-52} + 2^{-104}) \times 2) \\ &= 2(1 + 2^{-51})\end{aligned}$$

The rounding errors may be amplified by the sensitivity of the problems.

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}$$

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b} \implies x(A) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

The rounding errors may be amplified by the sensitivity of the problems.

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}$$

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b} \implies x(A) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

The rounding errors may be amplified by the sensitivity of the problems.

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}$$

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b} \implies x(A) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

Slightly perturbed system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.05 \end{bmatrix}}_{A+\delta A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}$$

with the solution $x(A+\delta A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

The error in the solution is bigger than the perturbation δA.

$$x(A+\delta A) - x(A) = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \quad \text{and} \quad \delta A = \begin{bmatrix} 0 & 0 \\ 0 & 0.05 \end{bmatrix}$$

Slightly perturbed system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.05 \end{bmatrix}}_{A+\delta A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}$$

with the solution $x(A+\delta A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

The error in the solution is bigger than the perturbation δA.

$$x(A+\delta A) - x(A) = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \quad \text{and} \quad \delta A = \begin{bmatrix} 0 & 0 \\ 0 & 0.05 \end{bmatrix}$$

Slightly perturbed system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.05 \end{bmatrix}}_{A+\delta A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}$$

with the solution $x(A+\delta A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

The error in the solution is bigger than the perturbation δA.

$$x(A+\delta A) - x(A) = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \quad \text{and} \quad \delta A = \begin{bmatrix} 0 & 0 \\ 0 & 0.05 \end{bmatrix}$$