KUTAISI
INTERNATIONAL
UNIVERSITY

Numerical Linear Algebra

Ramaz Botchorishvili

Kutaisi International University

October 19, 2022

Numerical Linear Algebra

Ramaz Botchorishvili

Kutaisi International University

October 19, 2022

Well posed problem, ill conditioned problem, condition Number

- Recap of Previous Lecture
- Perturbations in right hand side and coefficients
- Error sources
- Number systems
- Floating point
- Q & A

Recap of Previous Lecture

- Ill conditioned linear system and matrix properties
- Condition number of a matrix
- Properties of $Cond(A)$
- Perturbations in right hand side
- Perturbations in coefficients

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

- *A is invertible*

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

- *A is invertible*
- *$Ax = b$*

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

- *A is invertible*
- $Ax = b$
- *δb is perturbation of b*

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

- *A is invertible*
- $Ax = b$
- *δb is perturbation of b*
- *δx is pertubation caused by δb*

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

- *A is invertible*
- *$Ax = b$*
- *δb is perturbation of b*
- *δx is pertubation caused by δb*

Then the following holds true:

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}$$

Perturbations in linear system $Ax = b$, case - RHS b

Theorem 5.1

(Right perturbation theorem)
Suppose

- *A is invertible*
- *$Ax = b$*
- *δb is perturbation of b*
- *δx is pertubation caused by δb*

Then the following holds true:

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}$$

$$\frac{1}{cond(A)}\frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$$

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- *A is invertible*

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- *A is invertible*
- $Ax = b$

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- *A is invertible*
- *$Ax = b$*
- *δA is perturbation of A*

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- *A is invertible*
- $Ax = b$
- *δA is perturbation of A*
- *δx is pertubation caused by δA*

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- *A is invertible*
- $Ax = b$
- *δA is perturbation of A*
- *δx is pertubation caused by δA*
- $\|\delta A\| < 1/\|A^{-1}\|$

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- *A is invertible*
- *$Ax = b$*
- *δA is perturbation of A*
- *δx is pertubation caused by δA*
- *$\|\delta A\| < 1/\|A^{-1}\|$*

Then the following holds true:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\delta A\|}{1 - \|A^{-1}\|\|\delta A\|}$$

Perturbations in linear system $Ax = b$, case - matrix A

Theorem 5.2

(Left perturbation theorem)
Suppose

- A *is invertible*
- $Ax = b$
- δA *is perturbation of* A
- δx *is pertubation caused by* δA
- $\|\delta A\| < 1/\|A^{-1}\|$

Then the following holds true:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\delta A\|}{1 - \|A^{-1}\|\|\delta A\|}$$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)\frac{\|\delta A\|}{\|A\|}}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}$$

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- *A is invertible, $b \neq 0$*

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- *A is invertible, $b \neq 0$*
- *$Ax = b$*
- *δA is perturbation of A*
- *δb is perturbation of b*
- *δx is pertubation caused by δA and δb*

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b
- δx *is pertubation caused by* δA *and* δb
- $\|\delta A\| < 1/\|A^{-1}\|$

Perturbations in linear system $Ax = b$, case - matrix A and RHS b

Theorem 5.3

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b
- δx *is pertubation caused by* δA *and* δb
- $\|\delta A\| < 1/\|A^{-1}\|$

Then the following holds true:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A) \frac{\|\delta b\|}{\|b\|}$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A) \frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A) \frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$
  - $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$
  - $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$
  - required accuracy is 10^{-r}

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$
  - $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$
  - required accuracy is 10^{-r}
- Required accuracy is ensured if $10^{-r} \geq 10^c 10^{-p} \Rightarrow -r \geq c - p$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$
  - $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$
  - required accuracy is 10^{-r}
- Required accuracy is ensured if $10^{-r} \geq 10^c 10^{-p} \Rightarrow -r \geq c - p$
- the problem is well-conditioned if $c \leq p - r$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$
  - $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$
  - required accuracy is 10^{-r}
- Required accuracy is ensured if $10^{-r} \geq 10^c 10^{-p} \Rightarrow -r \geq c - p$
- the problem is well-conditioned if $c \leq p - r$
- the problem is ill-conditioned if $c > p - r$

Perturbation theorems and conditioning of linear systems

- Q: How large condition number should be for ...
  - classifying problem as ill-conditioned
  - classifying problem as well-conditioned
- A: depends what is required accuracy

Example 5.4

- Right perturbation theorem $\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$
- suppose
  - $cond(A) = 10^c$
  - $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$
  - required accuracy is 10^{-r}
- Required accuracy is ensured if $10^{-r} \geq 10^c 10^{-p} \Rightarrow -r \geq c - p$
- the problem is well-conditioned if $c \leq p - r$
- the problem is ill-conditioned if $c > p - r$
- the same condition number can be indicator of ..?(ill -/ well-conditioning)

Perturbation theorems and conditioning of linear systems

- How large condition number should be for ...
  - Q: classifying problem as ill-conditioned
  - Q: classifying problem as well-conditioned

Perturbation theorems and conditioning of linear systems

- How large condition number should be for ...
  - Q: classifying problem as ill-conditioned
  - Q: classifying problem as well-conditioned
  - A: depends what is required accuracy

Perturbation theorems and conditioning of linear systems

- How large condition number should be for ...
  - Q: classifying problem as ill-conditioned
  - Q: classifying problem as well-conditioned
  - A: depends what is required accuracy

Example 5.5

- Formula (gen.perturbation theorem)

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

Perturbation theorems and conditioning of linear systems

- How large condition number should be for ...
  - Q: classifying problem as ill-conditioned
  - Q: classifying problem as well-conditioned
  - A: depends what is required accuracy

Example 5.5

- Formula (gen.perturbation theorem)

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

- max. amplification factor for relative errorrs

$$\frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}} \geq cond(A)$$

Perturbation theorems and conditioning of linear systems

- How large condition number should be for ...
  - Q: classifying problem as ill-conditioned
  - Q: classifying problem as well-conditioned
  - A: depends what is required accuracy

Example 5.5

- Formula (gen.perturbation theorem)

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

- max. amplification factor for relative errorrs

$$\frac{cond(A)}{1-cond(A)\frac{\|\delta A\|}{\|A\|}} \geq cond(A)$$

- Analysis similar to right perturbation theorem on previous slide

Perturbation theorems and conditioning of linear systems

- How large condition number should be for ...
  - Q: classifying problem as ill-conditioned
  - Q: classifying problem as well-conditioned
  - A: depends what is required accuracy

Example 5.5

- Formula (gen.perturbation theorem)

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

- max. amplification factor for relative errorrs

$$\frac{cond(A)}{1-cond(A)\frac{\|\delta A\|}{\|A\|}} \geq cond(A)$$

- Analysis similar to right perturbation theorem on previous slide
- if $cond(A) = 10^c$ then perturbations are amplified by at least 10^c

Perturbation theorems and conditioning of linear systems

Learned from perturbation theorems:
if $cond(A) = 10^c$ then perturbations are amplified by at leas 10^c

Perturbation theorems and conditioning of linear systems

Learned from perturbation theorems:
if $cond(A) = 10^c$ then perturbations are amplified by at leas 10^c

Example 5.6

Some ill-conditioned matrices

Perturbation theorems and conditioning of linear systems

Learned from perturbation theorems:
if $cond(A) = 10^c$ then perturbations are amplified by at leas 10^c

Example 5.6

Some ill-conditioned matrices

- Hilbert matrix

$$a_{ij} = \frac{1}{i+j-1}, \quad cond_2(A_{10\times 10}) = 1.6025 \cdot 10^{13}$$

Perturbation theorems and conditioning of linear systems

Learned from perturbation theorems:
if $cond(A) = 10^c$ then perturbations are amplified by at leas 10^c

Example 5.6

Some ill-conditioned matrices

- Hilbert matrix
$$a_{ij} = \frac{1}{i+j-1}, \quad cond_2(A_{10\times 10}) = 1.6025 \cdot 10^{13}$$
- Pei matrix
$$a_{ii} = \alpha, a_{ij,i\neq j} = 1, \quad cond_2(A_{5\times 5,\alpha=0.9999} = 5 \cdot 10^4)$$

Perturbation theorems and conditioning of linear systems

Learned from perturbation theorems:
if $cond(A) = 10^c$ then perturbations are amplified by at leas 10^c

Example 5.6

Some ill-conditioned matrices

- Hilbert matrix

$$a_{ij} = \frac{1}{i+j-1}, \quad cond_2(A_{10\times 10}) = 1.6025 \cdot 10^{13}$$

- Pei matrix

$$a_{ii} = \alpha, a_{ij,i\neq j} = 1, \quad cond_2(A_{5\times 5,\alpha=0.9999} = 5 \cdot 10^4)$$

- Vandermonde matrix

$$a_{ij} = v_i^{n-j}, v \in \mathbb{R}, \quad cond_2(A_{5\times 5,v_i=i}) = 2.617 \cdot 10^4$$

General perturbation theorem, 1

Theorem 5.7

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b
- δx *is pertubation caused by* δA *and* δb
- $\|\delta A\| < 1/\|A^{-1}\|$

Then the following holds true:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)
Suppose

- *induced matrix norm* $\| \cdot \| : \mathcal{C}^{n \times n} \rightarrow R$

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)
Suppose

- *induced matrix norm* $\|\cdot\| : \mathcal{C}^{n\times n} \to R$
- *vector norm* $\|\cdot\| : \mathcal{C}^{n} \to R$

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)
Suppose

- *induced matrix norm* $\|\cdot\| : \mathcal{C}^{n\times n} \to R$
- *vector norm* $\|\cdot\| : \mathcal{C}^{n} \to R$
- $M \in \mathcal{C}^{n\times n}, \|M\| < 1$

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)
Suppose

- *induced matrix norm* $\|\cdot\| : \mathcal{C}^{n\times n} \to R$
- *vector norm* $\|\cdot\| : \mathcal{C}^{n} \to R$
- $M \in \mathcal{C}^{n\times n}, \|M\| < 1$

$$\Downarrow$$

1. $(I - M)^{-1}$ *exists*

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)
Suppose

- *induced matrix norm* $\|\cdot\| : \mathcal{C}^{n\times n} \to R$
- *vector norm* $\|\cdot\| : \mathcal{C}^{n} \to R$
- $M \in \mathcal{C}^{n\times n}, \|M\| < 1$

$$\Downarrow$$

1. $(I - M)^{-1}$ *exists*
2. $\|(I - M)^{-1}\| \leq \frac{1}{1-\|M\|}$

General perturbation theorem, 2

Theorem 5.8

(Estimates for $I - M$)
Suppose

- *induced matrix norm* $\|\cdot\| : \mathcal{C}^{n\times n} \to R$
- *vector norm* $\|\cdot\| : \mathcal{C}^{n} \to R$
- $M \in \mathcal{C}^{n\times n}, \|M\| < 1$

$$\Downarrow$$

1. $(I - M)^{-1}$ *exists*
2. $\|(I - M)^{-1}\| \leq \frac{1}{1-\|M\|}$
3. $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$

General perturbation theorem, 3

Proof.

THM [**?**].1 $(I - M)^{-1}$ exists:

General perturbation theorem, 3

Proof.

THM [**?**].1 $(I - M)^{-1}$ exists:

- $\|(I - M)x\| = \|x - Mx\|$

General perturbation theorem, 3

Proof.

THM [?].1 $(I - M)^{-1}$ exists:

- $\|(I - M)x\| = \|x - Mx\|$
- $\|x - Mx\| \geq |\|x\| - \|Mx\||$

General perturbation theorem, 3

Proof.

THM [**?**].1 $(I - M)^{-1}$ exists:

- $\|(I - M)x\| = \|x - Mx\|$
- $\|x - Mx\| \geq |\|x\| - \|Mx\||$
- $|\|x\| - \|Mx\|| = |(1 - \frac{\|Mx\|}{\|x\|})\|x\||$

General perturbation theorem, 3

Proof.

THM [**?**].1 $(I - M)^{-1}$ exists:

- $\|(I - M)x\| = \|x - Mx\|$
- $\|x - Mx\| \geq |\|x\| - \|Mx\||$
- $|\|x\| - \|Mx\|| = |(1 - \frac{\|Mx\|}{\|x\|})\|x\||$
- $|(1 - \frac{\|Mx\|}{\|x\|})| \geq 1 - \|M\|$

General perturbation theorem, 3

Proof.

THM [?].1 $(I - M)^{-1}$ exists:

- $\|(I-M)x\| = \|x - Mx\|$
- $\|x - Mx\| \geq |\|x\| - \|Mx\||$
- $|\|x\| - \|Mx\|| = |(1 - \frac{\|Mx\|}{\|x\|})\|x\||$
- $|(1 - \frac{\|Mx\|}{\|x\|})| \geq 1 - \|M\|$

$$\Downarrow$$

- $\|(I-M)x\| \geq (1 - \|M\|)\|x\|$

General perturbation theorem, 3

Proof.

THM [?].1 $(I - M)^{-1}$ exists:

- $\|(I-M)x\| = \|x - Mx\|$
- $\|x - Mx\| \geq |\|x\| - \|Mx\||$
- $|\|x\| - \|Mx\|| = |(1 - \frac{\|Mx\|}{\|x\|})\|x\||$
- $|(1 - \frac{\|Mx\|}{\|x\|})| \geq 1 - \|M\|$

$$\Downarrow$$

- $\|(I-M)x\| \geq (1 - \|M\|)\|x\|$

$$\Downarrow$$

- $(I - M)x = 0 \Rightarrow x = 0$

General perturbation theorem, 3

Proof.

THM [?].1 $(I - M)^{-1}$ exists:

- $\|(I - M)x\| = \|x - Mx\|$
- $\|x - Mx\| \geq |\|x\| - \|Mx\||$
- $|\|x\| - \|Mx\|| = |(1 - \frac{\|Mx\|}{\|x\|})\|x\||$
- $|(1 - \frac{\|Mx\|}{\|x\|})| \geq 1 - \|M\|$

$$\Downarrow$$

- $\|(I - M)x\| \geq (1 - \|M\|)\|x\|$

$$\Downarrow$$

- $(I - M)x = 0 \Rightarrow x = 0$

$$\Downarrow$$

- $(I - M)$ is invertible

$\square$

General perturbation theorem, 4

Proof.

THM [**?**].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

- $1 = \|I\| = \|(I-M)(I-M)^{-1}\| =$

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

- $1 = \|I\| = \|(I-M)(I-M)^{-1}\| =$
- $= \|(I-M)^{-1} - M(I-M)^{-1}\| \geq$

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

- $1 = \|I\| = \|(I-M)(I-M)^{-1}\| =$
- $= \|(I-M)^{-1} - M(I-M)^{-1}\| \geq$
- $|\|(I-M)^{-1}\| - \|M(I-M)^{-1}\|| \geq$

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I - M)^{-1}\| \le \frac{1}{1 - \|M\|}$:

- $1 = \|I\| = \|(I - M)(I - M)^{-1}\| =$
- $= \|(I - M)^{-1} - M(I - M)^{-1}\| \ge$
- $|\|(I - M)^{-1}\| - \|M(I - M)^{-1}\|| \ge$
- $|\|(I - M)^{-1}\| - \|M\|\|(I - M)^{-1}\||$

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

- $1 = \|I\| = \|(I-M)(I-M)^{-1}\| =$
- $= \|(I-M)^{-1} - M(I-M)^{-1}\| \geq$
- $|\|(I-M)^{-1}\| - \|M(I-M)^{-1}\|| \geq$
- $|\|(I-M)^{-1}\| - \|M\|\|(I-M)^{-1}\||$

$$\Downarrow$$

- $1 \geq |\|(I-M)^{-1}\| - \|M\|\|(I-M)^{-1}\||$

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

- $1 = \|I\| = \|(I-M)(I-M)^{-1}\| =$
- $= \|(I-M)^{-1} - M(I-M)^{-1}\| \geq$
- $|\|(I-M)^{-1}\| - \|M(I-M)^{-1}\|| \geq$
- $|\|(I-M)^{-1}\| - \|M\|\|(I-M)^{-1}\||$

$$\Downarrow$$

- $1 \geq |\|(I-M)^{-1}\| - \|M\|\|(I-M)^{-1}\||$

$$\Downarrow$$

- $1 \geq (1-\|M\|)\|(I-M)^{-1}\|$

General perturbation theorem, 4

Proof.

THM [?].2 $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$:

- $1 = \|I\| = \|(I-M)(I-M)^{-1}\| =$
- $= \|(I-M)^{-1} - M(I-M)^{-1}\| \geq$
- $|\|(I-M)^{-1}\| - \|M(I-M)^{-1}\|| \geq$
- $|\|(I-M)^{-1}\| - \|M\|\|(I-M)^{-1}\||$

$$\Downarrow$$

- $1 \geq |\|(I-M)^{-1}\| - \|M\|\|(I-M)^{-1}\||$

$$\Downarrow$$

- $1 \geq (1-\|M\|)\|(I-M)^{-1}\|$

$$\Downarrow$$

- $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$

□

General perturbation theorem, 5

Proof.

THM [**?**].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

General perturbation theorem, 5

Proof.

THM [**?**].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$

General perturbation theorem, 5

Proof.

THM [**?**].3 $(I-M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$
- $S_j(I-M) = \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1}$

General perturbation theorem, 5

Proof.

THM [**?**].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$
- $S_j(I - M) = \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1}$
- $\sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1} = I - M^{j+1}$

General perturbation theorem, 5

Proof.

THM [**?**].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$
- $S_j(I - M) = \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1}$
- $\sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1} = I - M^{j+1}$
- $\| - M\| < 1 \Rightarrow \|M^{j+1}\| \rightarrow_{j \to \infty} 0$

General perturbation theorem, 5

Proof.

THM [**?**].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$
- $S_j(I - M) = \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1}$
- $\sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1} = I - M^{j+1}$
- $\| - M\| < 1 \Rightarrow \|M^{j+1}\| \to_{j\to\infty} 0$

$$\Downarrow$$

- $\|S_j(I - M) - I\| = \|M^{j+1}\| \to_{j\to\infty} 0$

General perturbation theorem, 5

Proof.

THM [?].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$
- $S_j(I - M) = \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1}$
- $\sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1} = I - M^{j+1}$
- $\| - M\| < 1 \Rightarrow \|M^{j+1}\| \to_{j\to\infty} 0$

$$\Downarrow$$

- $\|S_j(I - M) - I\| = \|M^{j+1}\| \to_{j\to\infty} 0$

$$\Downarrow$$

- $\lim_{j\to\infty} S_j(I - M) = I$

General perturbation theorem, 5

Proof.

THM [?].3 $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$:

- $S_j = \sum_{k=0}^{j} M^k$
- $S_j(I - M) = \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1}$
- $\sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1} = I - M^{j+1}$
- $\| - M\| < 1 \Rightarrow \|M^{j+1}\| \to_{j\to\infty} 0$

$$\Downarrow$$

- $\|S_j(I - M) - I\| = \|M^{j+1}\| \to_{j\to\infty} 0$

$$\Downarrow$$

- $\lim_{j\to\infty} S_j(I - M) = I$

$$\Downarrow$$

- $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$

□

General perturbation theorem, 6

Corollary 5.9

1. $\|(I + M)^{-1}\| \leq \frac{1}{1-\|M\|}$

General perturbation theorem, 6

Corollary 5.9

1. $\|(I+M)^{-1}\| \leq \frac{1}{1-\|M\|}$
2. $\|S^{-1}\|\|S-T\| < 1 \Rightarrow T$ *is invertible*

General perturbation theorem, 6

Corollary 5.9

1. $\|(I+M)^{-1}\| \leq \frac{1}{1-\|M\|}$
2. $\|S^{-1}\|\|S-T\| < 1 \Rightarrow T$ *is invertible*

Proof.

- $\#1: \|M\| = \|-M\| \Rightarrow 1$
- $\#2: T = S(I-(I-S^{-1}T))$

General perturbation theorem, 6

Corollary 5.9

1. $\|(I+M)^{-1}\| \leq \frac{1}{1-\|M\|}$
2. $\|S^{-1}\|\|S-T\| < 1 \Rightarrow T$ *is invertible*

Proof.

- $\#1: \|M\| = \|-M\| \Rightarrow 1$
- $\#2: T = S(I-(I-S^{-1}T))$
- $M = I - S^{-1}T \Rightarrow M = S^{-1}(S-T)$

General perturbation theorem, 6

Corollary 5.9

1. $\|(I+M)^{-1}\| \leq \frac{1}{1-\|M\|}$
2. $\|S^{-1}\|\|S-T\| < 1 \Rightarrow T$ *is invertible*

Proof.

- $\#1 : \|M\| = \|-M\| \Rightarrow 1$
- $\#2 : T = S(I - (I - S^{-1}T))$
- $M = I - S^{-1}T \Rightarrow M = S^{-1}(S-T)$

$$\Downarrow$$

- $\|M\| = \|S^{-1}(S-T)\| \leq \|S^{-1}\|\|(S-T)\| < 1$

General perturbation theorem, 6

Corollary 5.9

1. $\|(I+M)^{-1}\| \leq \frac{1}{1-\|M\|}$
2. $\|S^{-1}\|\|S-T\| < 1 \Rightarrow T$ *is invertible*

Proof.

- $\#1: \|M\| = \|-M\| \Rightarrow 1$
- $\#2: T = S(I-(I-S^{-1}T))$
- $M = I - S^{-1}T \Rightarrow M = S^{-1}(S-T)$

$$\Downarrow$$

- $\|M\| = \|S^{-1}(S-T)\| \leq \|S^{-1}\|\|(S-T)\| < 1$

$$\Downarrow$$

- $(I-M)$ invertible

General perturbation theorem, 6

Corollary 5.9

1. $\|(I+M)^{-1}\| \leq \frac{1}{1-\|M\|}$
2. $\|S^{-1}\|\|S-T\| < 1 \Rightarrow T$ *is invertible*

Proof.

- $\#1: \|M\| = \|-M\| \Rightarrow 1$
- $\#2: T = S(I-(I-S^{-1}T))$
- $M = I - S^{-1}T \Rightarrow M = S^{-1}(S-T)$

$$\Downarrow$$

- $\|M\| = \|S^{-1}(S-T)\| \leq \|S^{-1}\|\|(S-T)\| < 1$

$$\Downarrow$$

- $(I-M)$ invertible

$$\Downarrow$$

- $T = S(I-M)$ invertible

□

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- *A is invertible,* $b \neq 0$

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- *A is invertible, $b \neq 0$*
- *$Ax = b$*
- *δA is perturbation of A*
- *δb is perturbation of b*

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b
- δx *is pertubation caused by* δA *and* δb

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b
- δx *is pertubation caused by* δA *and* δb
- $\|\delta A\| < 1/\|A^{-1}\|$

General perturbation theorem, 7

Theorem 5.10

(General perturbation theorem)
Suppose

- A *is invertible,* $b \neq 0$
- $Ax = b$
- δA *is perturbation of* A
- δb *is perturbation of* b
- δx *is pertubation caused by* δA *and* δb
- $\|\delta A\| < 1/\|A^{-1}\|$

Then the following holds true:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$
- $= (I + A^{-1}\delta A)^{-1}A^{-1}(b + \delta b)$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$
- $= (I + A^{-1}\delta A)^{-1}A^{-1}(b + \delta b)$
- $\delta x = (I - M)^{-1}A^{-1}(b + \delta b) - x =$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \le \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$
- $= (I + A^{-1}\delta A)^{-1}A^{-1}(b + \delta b)$
- $\delta x = (I - M)^{-1}A^{-1}(b + \delta b) - x =$
- $(I - M)^{-1}A^{-1}(b + \delta b) - A^{-1}b =$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$
- $= (I + A^{-1}\delta A)^{-1}A^{-1}(b + \delta b)$
- $\delta x = (I - M)^{-1}A^{-1}(b + \delta b) - x =$
- $(I - M)^{-1}A^{-1}(b + \delta b) - A^{-1}b =$
- $= (I - M)^{-1}(A^{-1}(b + \delta b) - (I - M)A^{-1}b) =$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \le \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$
- $= (I + A^{-1}\delta A)^{-1}A^{-1}(b + \delta b)$
- $\delta x = (I - M)^{-1}A^{-1}(b + \delta b) - x =$
- $(I - M)^{-1}A^{-1}(b + \delta b) - A^{-1}b =$
- $= (I - M)^{-1}(A^{-1}(b + \delta b) - (I - M)A^{-1}b) =$
- $= (I - M)^{-1}(A^{-1}\delta b - MA^{-1}b)$

General perturbation theorem, 8

Proof.

(General perturbation theorem)

- $(A + \delta A)(x + \delta x) = b + \delta b$
- $M = -A^{-1}\delta A$, $x = A^{-1}b$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $x + \delta x = (A + \delta A)^{-1}(b + \delta b) = (A(I + A^{-1}\delta A))^{-1}(b + \delta b) =$
- $= (I + A^{-1}\delta A)^{-1}A^{-1}(b + \delta b)$
- $\delta x = (I - M)^{-1}A^{-1}(b + \delta b) - x =$
- $(I - M)^{-1}A^{-1}(b + \delta b) - A^{-1}b =$
- $= (I - M)^{-1}(A^{-1}(b + \delta b) - (I - M)A^{-1}b) =$
- $= (I - M)^{-1}(A^{-1}\delta b - MA^{-1}b)$
- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$

$\square$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $\|\delta x\| \leq \frac{1}{1-cond(A)\frac{\|\delta A\|}{\|A\|}}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $\|\delta x\| \leq \frac{1}{1-cond(A)\frac{\|\delta A\|}{\|A\|}}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
    - $\|MA^{-1}b\| \leq \|M\|\|x\| \leq cond(A)\frac{\|\delta A\|}{\|A\|}\|x\|$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $\|\delta x\| \leq \frac{1}{1-cond(A)\frac{\|\delta A\|}{\|A\|}}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
  - $\|MA^{-1}b\| \leq \|M\|\|x\| \leq cond(A)\frac{\|\delta A\|}{\|A\|}\|x\|$
  - $\|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|\frac{\|Ax\|}{\|Ax\|} \leq K(A)\frac{\|\delta b\|}{\|b\|}\|x\|$

General perturbation theorem, 9

Proof.

(General perturbation theorem)

- $\|\delta x\| \leq \frac{1}{1-\|M\|}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
- $\|M\| \leq \|A^{-1}\|\|\delta A\| < 1$
- $\|A^{-1}\|\|\delta A\| = cond(A)\frac{\|\delta A\|}{\|A\|}$
- $\|\delta x\| \leq \frac{1}{1-cond(A)\frac{\|\delta A\|}{\|A\|}}(\|A^{-1}\delta b\| + \|MA^{-1}b\|)$
  - $\|MA^{-1}b\| \leq \|M\|\|x\| \leq cond(A)\frac{\|\delta A\|}{\|A\|}\|x\|$
  - $\|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|\frac{\|Ax\|}{\|Ax\|} \leq K(A)\frac{\|\delta b\|}{\|b\|}\|x\|$

$$\Downarrow$$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

□

Sources of errors, 1

Main sources of errors in numerical computations

- Rounding errors (arithmetic errors)
  - consequence of finite precision arithmetic
  - unavoidable
- Uncertainty in data, may arise in several ways:
  - errors in measuring physical quantities
  - errors from earlier computations
  - from wrong mathematical models of reality
  - ...
- Truncation errors (discretization errors, approximation errors)

Sources of errors, 2

- Rounding errors (arithmetic errors)
  - consequence of finite precision arithmetic
  - unavoidable

Example 5.11

-
```
In [14]: x=0.001

In [15]: print((1-math.cos(x))/(x*x))
0.49999995832550326

In [16]: x=0.000001

In [17]: print((1-math.cos(x))/(x*x))
0.5000444502911705

In [18]: x=0.000000000001

In [19]: print((1-math.cos(x))/(x*x))
0.0
```

Figure: Arithmetic error, true value is 0.5

Number systems, 1

- Binary, base $= 2$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

- Base: 10

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

- Base: 10
- Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

- Base: 10
- Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$
- Integer: $2021 = 2 * 10^3 + 0 * 10^2 + 2 * 10^1 + 1 * 10^0$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

- Base: 10
- Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$
- Integer: $2021 = 2 * 10^3 + 0 * 10^2 + 2 * 10^1 + 1 * 10^0$
- Real: $10.19 = 1 * 10^1 + 0 * 10^0 + 1 * 10^{-1} + 9 * 10^{-2}$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

- Base: 10
- Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$
- Integer: $2021 = 2 * 10^3 + 0 * 10^2 + 2 * 10^1 + 1 * 10^0$
- Real: $10.19 = 1 * 10^1 + 0 * 10^0 + 1 * 10^{-1} + 9 * 10^{-2}$
- n-digit number:
  $d_{n-1}d_{n-2}...d_1d_0 = d_{n-1} * 10^{n-1} + d_{n-2} * 10^{n-2} + ... + d_1 * 10^1 + d_0 * 10^0$

Number systems, 1

- Binary, base $= 2$
- Octal, base $= 8$
- Decimal, base $= 10$
- Hexadecimal, base $= 16$

Example 5.12

Decimal system

- Base: 10
- Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$
- Integer: $2021 = 2 * 10^3 + 0 * 10^2 + 2 * 10^1 + 1 * 10^0$
- Real: $10.19 = 1 * 10^1 + 0 * 10^0 + 1 * 10^{-1} + 9 * 10^{-2}$
- n-digit number:
  $d_{n-1}d_{n-2}...d_1d_0 = d_{n-1} * 10^{n-1} + d_{n-2} * 10^{n-2} + ... + d_1 * 10^1 + d_0 * 10^0$
- n-digit integral and m-digit fractional part:
  $d_{n-1}d_{n-2}...d_1d_0.d_{-1}d_{-2}...d_{-m} = d_{n-1} * 10^{n-1} + ... + d_1 * 10^1 + d_0 * 10^0 + d_{-1} * 10^{-1} + d_{-2} * 10^{-2} + ... + d_{-m} * 10^{-m}$

Number systems, 2

Example 5.13

Binary system

Number systems, 2

Example 5.13

Binary system

- Base: 2

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$
- Real: $10.10_2 = 1 * 2^1 + 0 * 2^0 + 1 * 2^{-1} + 0 * 2^{-2} = 2.5_{10}$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$
- Real: $10.10_2 = 1 * 2^1 + 0 * 2^0 + 1 * 2^{-1} + 0 * 2^{-2} = 2.5_{10}$
- n-digit number:
  $(d_{n-1}d_{n-2}...d_1d_0)_2 = (d_{n-1}*2^{n-1}+d_{n-2}*2^{n-2}+...+d_1*2^1+d_0*2^0)_{10}$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$
- Real: $10.10_2 = 1 * 2^1 + 0 * 2^0 + 1 * 2^{-1} + 0 * 2^{-2} = 2.5_{10}$
- n-digit number:
  $(d_{n-1} d_{n-2} ... d_1 d_0)_2 = (d_{n-1} * 2^{n-1} + d_{n-2} * 2^{n-2} + ... + d_1 * 2^1 + d_0 * 2^0)_{10}$
- n-digit integral and m-digit fractional part:
  $(d_{n-1} d_{n-2} ... d_1 d_0 . d_{-1} d_{-2} ... d_{-m})_2 =$
  $(d_{n-1} * 2^{n-1} + ... + d_1 * 2^1 + d_0 * 2^0 + d_{-1} * 2^{-1} + d_{-2} * 2^{-2} + ... + d_{-m} * 2^{-m})_{10}$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$
- Real: $10.10_2 = 1 * 2^1 + 0 * 2^0 + 1 * 2^{-1} + 0 * 2^{-2} = 2.5_{10}$
- n-digit number:
  $(d_{n-1} d_{n-2} ... d_1 d_0)_2 = (d_{n-1} * 2^{n-1} + d_{n-2} * 2^{n-2} + ... + d_1 * 2^1 + d_0 * 2^0)_{10}$
- n-digit integral and m-digit fractional part:
  $(d_{n-1} d_{n-2} ... d_1 d_0 . d_{-1} d_{-2} ... d_{-m})_2 =$
  $(d_{n-1} * 2^{n-1} + ... + d_1 * 2^1 + d_0 * 2^0 + d_{-1} * 2^{-1} + d_{-2} * 2^{-2} + ... + d_{-m} * 2^{-m})_{10}$

- Limitation: can only exactly represent numbers $x/2^k$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$
- Real: $10.10_2 = 1 * 2^1 + 0 * 2^0 + 1 * 2^{-1} + 0 * 2^{-2} = 2.5_{10}$
- n-digit number:
  $(d_{n-1} d_{n-2} ... d_1 d_0)_2 = (d_{n-1} * 2^{n-1} + d_{n-2} * 2^{n-2} + ... + d_1 * 2^1 + d_0 * 2^0)_{10}$
- n-digit integral and m-digit fractional part:
  $(d_{n-1} d_{n-2} ... d_1 d_0 . d_{-1} d_{-2} ... d_{-m})_2 =$
  $(d_{n-1} * 2^{n-1} + ... + d_1 * 2^1 + d_0 * 2^0 + d_{-1} * 2^{-1} + d_{-2} * 2^{-2} + ... + d_{-m} * 2^{-m})_{10}$

- Limitation: can only exactly represent numbers $x/2^k$
- $0.2_{10} = 0.00110011[0011]..._2$

Number systems, 2

Example 5.13

Binary system

- Base: 2
- Symbols: $0, 1$
- Integer: $1010_2 = 1 * 2^3 + 0 * 2^2 + 2^1 + 1 * 2^0 = 10_{10}$
- Real: $10.10_2 = 1 * 2^1 + 0 * 2^0 + 1 * 2^{-1} + 0 * 2^{-2} = 2.5_{10}$
- n-digit number:
  $(d_{n-1} d_{n-2} ... d_1 d_0)_2 = (d_{n-1} * 2^{n-1} + d_{n-2} * 2^{n-2} + ... + d_1 * 2^1 + d_0 * 2^0)_{10}$
- n-digit integral and m-digit fractional part:
  $(d_{n-1} d_{n-2} ... d_1 d_0 . d_{-1} d_{-2} ... d_{-m})_2 =$
  $(d_{n-1} * 2^{n-1} + ... + d_1 * 2^1 + d_0 * 2^0 + d_{-1} * 2^{-1} + d_{-2} * 2^{-2} + ... + d_{-m} * 2^{-m})_{10}$

- Limitation: can only exactly represent numbers $x/2^k$
- $0.2_{10} = 0.00110011[0011]..._2$

Number systems, 3

- - Octal system

Number systems, 3

- - Octal system
  - Base: 8

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?
- - Hexadecimal system

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?
- - Hexadecimal system
  - Base: 16

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?
- - Hexadecimal system
  - Base: 16
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F$

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?
- - Hexadecimal system
  - Base: 16
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F$
  - Representing numbers and conversion to other number system?

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?
- - Hexadecimal system
  - Base: 16
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F$
  - Representing numbers and conversion to other number system?
- Can you define number system with arbitrary base r?

Number systems, 3

- - Octal system
  - Base: 8
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7$
  - Representing numbers and conversion to other number system?
- - Hexadecimal system
  - Base: 16
  - Symbols: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F$
  - Representing numbers and conversion to other number system?
- Can you define number system with arbitrary base r?
- Is conversion to and from decimal system possible?

Floating point system numbers, 1

- Most computers use binary number system

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i} \beta^{-i}) \beta^e$

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i}\beta^{-i}$ - fraction, called mantissa

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i}\beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i}\beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Definition 5.15

- Floating point number is called normalized if $d_{-1} > 0$

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i}\beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Definition 5.15

- Floating point number is called normalized if $d_{-1} > 0$
- Number of digits in mantissa is called precision

Floating point system numbers, 1

- Most computers use binary number system
- Numeric computations on a computer are performed using floating point operations

Definition 5.14

Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i}\beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Definition 5.15

- Floating point number is called normalized if $d_{-1} > 0$
- Number of digits in mantissa is called precision

Example 5.16

$0.1 \cdot 10^{-2}$ normalized, $0.01 \cdot 10^{-1}$ - unnormalized

Floating point system numbers, 2

Definition 5.17

Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i}\beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Floating point system numbers, 2

Definition 5.17

Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^m d_{-i}\beta^{-i})\beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^m d_{-i}\beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Definition 5.18

IEEE floating point standard

- Single precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 23 & 8 \end{pmatrix}$

Floating point system numbers, 2

Definition 5.17

Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i} \beta^{-i}) \beta^e$

- β - base
- s -sign, $d_{-i} \in \{0, ..., \beta - 1\}, i = 1, ..., m$
- e - exponent, $e \in \{e_{min}, ..., e_{max}\}$
- $\sum_{i=1}^{m} d_{-i} \beta^{-i}$ - fraction, called mantissa
- m - mantissa length

Definition 5.18

IEEE floating point standard

- Single precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 23 & 8 \end{pmatrix}$
- Double precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 52 & 11 \end{pmatrix}$

Floating point system numbers, 3

- Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i} \beta^{-i}) \beta^e$

Floating point system numbers, 3

- Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$
- IEEE floating point single precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 23 & 8 \end{pmatrix}$

Floating point system numbers, 3

- Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$
- IEEE floating point single precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 23 & 8 \end{pmatrix}$
- IEEE floating point single precision, accuracy: $2^{-23} \approx 1.2 \cdot 10^{-7}$

Floating point system numbers, 3

- Floating point number $\tilde{x} = (-1)^s(\sum_{i=1}^{m} d_{-i}\beta^{-i})\beta^e$
- IEEE floating point single precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 23 & 8 \end{pmatrix}$
- IEEE floating point single precision, accuracy: $2^{-23} \approx 1.2 \cdot 10^{-7}$
- IEEE floating point double precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 52 & 11 \end{pmatrix}$

Floating point system numbers, 3

- Floating point number $\tilde{x} = (-1)^s (\sum_{i=1}^{m} d_{-i} \beta^{-i}) \beta^e$
- IEEE floating point single precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 23 & 8 \end{pmatrix}$
- IEEE floating point single precision, accuracy: $2^{-23} \approx 1.2 \cdot 10^{-7}$
- IEEE floating point double precision: $\begin{pmatrix} \text{sign} & \text{mantissa} & \text{exponent} \\ 1 & 52 & 11 \end{pmatrix}$
- IEEE floating point double precision, accuracy: $2^{-52} \approx 2.2 \cdot 10^{-16}$

Floating point system, 4

Definition 5.19

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Floating point system, 4

Definition 5.19

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Definition 5.20

Underflow: caused by a floating point number whose exponent is smaller then permissible range e_{min}

Floating point system, 4

Definition 5.19

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Definition 5.20

Underflow: caused by a floating point number whose exponent is smaller then permissible range e_{min}

Example 5.21

Underflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$

Floating point system, 4

Definition 5.19

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Definition 5.20

Underflow: caused by a floating point number whose exponent is smaller then permissible range e_{min}

Example 5.21

Underflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.3 \cdot 10^{-3}, b = 0.2 \cdot 10^{-3}, a \cdot b = 0.3 \cdot 10^{-3} \cdot 0.2 \cdot 10^{-3} = 0.06 \cdot 10^{-6} = 0.6 \cdot 10^{-7}$

Floating point system, 4

Definition 5.19

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Definition 5.20

Underflow: caused by a floating point number whose exponent is smaller then permissible range e_{min}

Example 5.21

Underflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.3 \cdot 10^{-3}, b = 0.2 \cdot 10^{-3}, a \cdot b = 0.3 \cdot 10^{-3} \cdot 0.2 \cdot 10^{-3} = 0.06 \cdot 10^{-6} = 0.6 \cdot 10^{-7}$
- Exponent out of range: $-7 < -3$

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Example 5.23

Overflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Example 5.23

Overflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.4 \cdot 10^2, b = 0.3 \cdot 10^2, a \cdot b = 0.4 \cdot 10^2 \cdot 0.3 \cdot 10^2 = 0.12 \cdot 10^4 = 0.12 \cdot 10^4$

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Example 5.23

Overflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.4 \cdot 10^2, b = 0.3 \cdot 10^2, a \cdot b = 0.4 \cdot 10^2 \cdot 0.3 \cdot 10^2 = 0.12 \cdot 10^4 = 0.12 \cdot 10^4$
- Exponent out of range: $4 > 3$

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Example 5.23

Overflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.4 \cdot 10^2, b = 0.3 \cdot 10^2, a \cdot b = 0.4 \cdot 10^2 \cdot 0.3 \cdot 10^2 = 0.12 \cdot 10^4 = 0.12 \cdot 10^4$
- Exponent out of range: $4 > 3$

- Overflow is a problem: for most system result is $\pm\infty$

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Example 5.23

Overflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.4 \cdot 10^2, b = 0.3 \cdot 10^2, a \cdot b = 0.4 \cdot 10^2 \cdot 0.3 \cdot 10^2 = 0.12 \cdot 10^4 = 0.12 \cdot 10^4$
- Exponent out of range: $4 > 3$

- Overflow is a problem: for most system result is $\pm\infty$
- Underflow is a problem: for most system result is 0

Floating point system, 5

Definition 5.22

Overflow: caused by a floating point number whose exponent is larger then permissible range e_{max}

Example 5.23

Overflow:

- $\beta = 10, m = 2, e_{min} = -3, e_{max} = 3$
- $a = 0.4 \cdot 10^2, b = 0.3 \cdot 10^2, a \cdot b = 0.4 \cdot 10^2 \cdot 0.3 \cdot 10^2 = 0.12 \cdot 10^4 = 0.12 \cdot 10^4$
- Exponent out of range: $4 > 3$

- Overflow is a problem: for most system result is $\pm\infty$
- Underflow is a problem: for most system result is 0
- Example: Ariane 5 explosion due to overflow

Floating point system, 6

Rounding and Chopping

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:

  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:

  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:

  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:
  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:
  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away
- Rounding down: $d_{-m-1} < \beta/2$

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:
  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:
  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away
- Rounding down: $d_{-m-1} < \beta/2$
- Rounding up: $d_{-m-1} \geq \beta/2$

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:
  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:
  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away
- Rounding down: $d_{-m-1} < \beta/2$
- Rounding up: $d_{-m-1} \geq \beta/2$

Example 5.24

$\pi = 3.141596$

- Two-digit arithmetic $fl(\pi) = 0.31 \cdot 10^{-2}$

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:

  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:

  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away
- Rounding down: $d_{-m-1} < \beta/2$
- Rounding up: $d_{-m-1} \geq \beta/2$

Example 5.24

$\pi = 3.141596$

- Two-digit arithmetic $fl(\pi) = 0.31 \cdot 10^{-2}$
- Three-digit arithmetic $fl(\pi) = 0.314 \cdot 10^{-2}$

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:
  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:
  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away
- Rounding down: $d_{-m-1} < \beta/2$
- Rounding up: $d_{-m-1} \geq \beta/2$

Example 5.24

$\pi = 3.141596$

- Two-digit arithmetic $fl(\pi) = 0.31 \cdot 10^{-2}$
- Three-digit arithmetic $fl(\pi) = 0.314 \cdot 10^{-2}$
- Four-digit arithmetic $fl(\pi) = 0.3142 \cdot 10^{-2}$

Floating point system, 6

Rounding and Chopping

- Not all real numbers are machine representable
- Finite number of real numbers are only representable in computers
- Consider $0.d_{-1}...d_{-m}d_{-m-1}$
- **Chopping**:
  in m-digit arithmetics d_{-m-1} and all other further digits are thrown away
- **Rounding**:
  in m-digit arithmetics d_{-m} is rounded up or down, d_{-m-1} and all other further digits are thrown away
- Rounding down: $d_{-m-1} < \beta/2$
- Rounding up: $d_{-m-1} \geq \beta/2$

Example 5.24

$\pi = 3.141596$

- Two-digit arithmetic $fl(\pi) = 0.31 \cdot 10^{-2}$
- Three-digit arithmetic $fl(\pi) = 0.314 \cdot 10^{-2}$
- Four-digit arithmetic $fl(\pi) = 0.3142 \cdot 10^{-2}$

Floating point system, 7

Machine precision, significant numbers

Definition 5.25

Machine precision μ is **smallest positive number** such that

$$fl(1+\mu) > 1$$

Floating point system, 7

Machine precision, significant numbers

Definition 5.25

Machine precision μ is **smallest positive number** such that

$$fl(1+\mu) > 1$$

Definition 5.26

Suppose x is real number and $\tilde{x}$ is its approximation. $\tilde{x}$ approximates x to s significant digits if s is **largest nonnegative integer** for which relative error satisfies the inequality:

$$\frac{|x-\tilde{x}|}{|x|} < 5 \cdot 10^{-s}$$

Floating point system, 7

Machine precision, significant numbers

Definition 5.25

Machine precision μ is **smallest positive number** such that

$$fl(1+\mu) > 1$$

Definition 5.26

Suppose x is real number and $\tilde{x}$ is its approximation. $\tilde{x}$ approximates x to s significant digits if s is **largest nonnegative integer** for which relative error satisfies the inequality:

$$\frac{|x-\tilde{x}|}{|x|} < 5 \cdot 10^{-s}$$

Example 5.27

- $x = 1.31, \tilde{x} = 1.3, |x-\tilde{x}| = 0.01, \frac{|x-\tilde{x}|}{|x|} = 0.007635$

Floating point system, 7

Machine precision, significant numbers

Definition 5.25

Machine precision μ is **smallest positive number** such that

$$fl(1+\mu) > 1$$

Definition 5.26

Suppose x is real number and $\tilde{x}$ is its approximation. $\tilde{x}$ approximates x to s significant digits if s is **largest nonnegative integer** for which relative error satisfies the inequality:

$$\frac{|x-\tilde{x}|}{|x|} < 5 \cdot 10^{-s}$$

Example 5.27

- $x = 1.31, \tilde{x} = 1.3, |x-\tilde{x}| = 0.01, \frac{|x-\tilde{x}|}{|x|} = 0.007635$
- $7.635 \cdot 10^{-3} < 5 \cdot 10^{-2}$, agree up to two significant digits

Q & A