Background and basics

Daniel P. Robinson
Department of Applied Mathematics and Statistics
Johns Hopkins University

September 1, 2020

Outline

1 Computer arithmetic
   • Floating-point (real) numbers
   • Floating-point (real) arithmetic

2 Linear systems, norms, and condition numbers
   • Review and motivation
   • Norms
   • Condition number of a linear system
   • Accuracy analysis

3 Some coding tips

4 Useful calculus facts and approximations
Floating-point (real) numbers

Modern computers store real numbers as

\[ x = \pm \left( d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right) \beta^E \]  

(4.362781 \times 10^{-08})

- **base:** \( \beta \) (e.g., 2)
- **precision:** \( p \) (e.g., 24 (SP), 53 (DP))
- **exponent:** \( E \in [L, U] \) (e.g., \([-126, 127]\) (SP), \([-1022, 1023]\) (DP))
- \( d_i \in [0, \beta - 1] \) for \( i = 0, \ldots, p - 1 \)
- the floating-point system is completely characterized by the four integers \( \beta, p, L, \) and \( U \)
- **mantissa:** \( d_0d_1\ldots d_{p-1} \)
- **fraction:** \( d_1 \ldots d_{p-1} \)
- floating-point system is **normalized** if \( d_0 \) is always nonzero unless the number represented is zero
- we will only consider normalized floating-point systems

**Example (Floating-point system)**

\( \beta = 10, \ p = 4, \ L = -99, \) and \( U = 99 \)

- some numbers
  - \( 1 = 1.000 \times 10^{00} \)
  - \( 34.67 = 3.467 \times 10^{01} \)
  - \( 0.0346 = 3.460 \times 10^{-02} \)
- smallest positive number: \( 1.000 \times 10^{-99} \) (underflow level)
- largest number: \( 9.999 \times 10^{99} \) (overflow level)
Facts about floating-point systems

- It is finite, i.e., not all real numbers can be stored.
- Machine numbers are those real numbers that may be exactly represented.
- Total number of normalized floating-point numbers is
  \[2(\beta - 1)\beta^{-1}(U - L + 1) + 1\]

- Smallest positive number: \(UFL = \beta^L\) (underflow level)
  - numbers smaller than \(UFL\) stored as zero
  - often not serious, because zero is a good approximation

- Largest number: \(OFL = \beta^{U+1}(1 - \beta^{-p})\) (overflow level)
  - numbers larger than \(OFL\) may not be stored
  - serious problem, compilers typically terminate

Rounding

When a real number \(x\) is not exactly representable, it is approximated by a “nearby” floating-point number \(\hat{x}(x)\). This process is called **rounding** and the error that is introduced is called **rounding error**.

- Common rounding strategies
  - chopping: \(\hat{x}(x)\) is obtained by truncating the expansion of \(x\) after \(d_{p-1}\). Also called round-to-zero.
  - round-to-nearest: \(\hat{x}(x)\) is the closest floating-point number to \(x\). In case of a tie, use the floating-point number whose last stored digit is even. Also called round-to-even.

- We will assume round-to-nearest since it is the most accurate and the default rounding rule on machines that abide by the IEEE standards

**Question:** How bad can the rounding error be?

**Answer:** Involves the concept of machine precision
Example (Motivation of machine precision)
Consider the following numbers $x$ and their nearest neighbor to the “right” $x_r$ (using $\beta = 10$ and $p = 4$)

\[
\begin{align*}
  x &= 1.000 \times 10^{00} \\
  x_r &= 1.001 \times 10^{00} \\
  x &= 1.000 \times 10^{06} \\
  x_r &= 1.001 \times 10^{06}
\end{align*}
\]

- Distance is $10^{-03}$
- Relative distance of both is $10^{-03}$
- Largest error in a number that is stored as 1 is $\frac{1}{2}10^{-03} = \frac{1}{2}\beta^{-p}$

Machine precision assuming round-to-nearest

$\epsilon_{\text{mach}} \overset{\text{def}}{=} \frac{1}{2}\beta^{1-p}$

bounds the relative error in storing a floating-point number:

\[
\frac{|\text{fl}(x) - x|}{|x|} \leq \epsilon_{\text{mach}}
\]

Definition (Machine precision)
The following three definitions are (roughly) equivalent. The machine precision is equal to

- the smallest number $\epsilon$ such that $\text{fl}(1 + \epsilon) > 1$
- the largest number $\epsilon$ such that $\text{fl}(1 + \epsilon) = 1$
- half the distance between 1 and the nearest floating-point number

Note: also called machine epsilon and unit-round-off.

Example (Understanding the definition of $\epsilon_{\text{mach}}$)
Using round-to-nearest, $p = 4$, and $\beta = 10$, we have

\[
\begin{align*}
  1.000 + 0.0005 &= 1.0005 \overset{\text{comp}}{=} 1.000 \\
  1.000 + 0.00051 &= 1.00051 \overset{\text{comp}}{=} 1.001
\end{align*}
\]

$\Rightarrow \epsilon_{\text{mach}} = 0.0005 = 5 \times 10^{-04} = \frac{1}{2} \times 10^{-03} = \frac{1}{2}\beta^{1-p}$

Comment: Generally, $0 < \text{UFL} < \epsilon_{\text{mach}} < \text{OFL}$
Exceptional values in the floating-point system

IEEE standard allows for the following exceptional values:

- **Inf**: represents “infinity” and results from dividing a finite number by zero
- **NaN**: stands for “not a number” and results from undefined or not well-defined operations (e.g., \(0/0\), \(0 \times \infty\), \(\infty/\infty\))

---

The basic idea (simplified)

- **Multiplication of two floating-point numbers (similar for division)**
  - exponents are summed and mantissas multiplied
  - product of two \(p\) digit mantissas is generally \(2p\) digits (must round)
  - example:
    \[
    x \times y = (4.452 \times 10^{62}) \times (6.436 \times 10^{-61}) = 28.653072 \times 10^{61} \\
    = 2.8653072 \times 10^{62} \approx 2.865 \times 10^{62}
    \]
- **Addition of two floating-point numbers (similar for subtraction)**
  - shift so that exponents are the same, add, then re-normalize
  - example:
    \[
    x + y = (4.452 \times 10^{62}) + (6.436 \times 10^{-61}) \\
    = 4.452 \times 10^{62} + 0.006436 \times 10^{62} \\
    = 4.458436 \times 10^{62} \approx 4.458 \times 10^{62}
    \]
- generally, trailing digits of smaller (in magnitude) number are lost

\[x = 4.452 \times 10^{62} \text{ and } y = 6.436 \times 10^{-61}\]
Example (Motivate concept of catastrophic cancellation)
Consider computing for some $a$ and $b$ the following:

$$z = 333.75b^6 + a^2(11a^2b^2 - b^5 - 121b^4 - 2)$$
$$x = 5.5b^8$$
$$y = z + x + a/(2b)$$

If $a = 77617$ and $b = 33096$ then

$$z = -7917111340668961361101134701524942850$$
$$x = 7917111340668961361101134701524942848$$
$$z + x = -2 \implies y = -2 + a/(2b) = -0.827396\ldots$$

But, if precision $p \leq 35$, then

$$z + x \text{ comp} = 0 \implies y \text{ comp} (a/2b) = 1.1726\ldots$$

Not even the correct sign!

The problem of interest
Given data input $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$, solve the linear system

$Ax = b$

- Let $a_{ij}$ denote the element of $A$ in row $i$ and column $j$
- Can consider questions of existence and uniqueness of solutions
- Conditioning (sensitivity of the solution) solution is $x = A^{-1}b$

Example (System of equations)

$$\begin{align*}
   x_1 + 3x_2 &= 5 \\
   2x_1 + 7x_2 &= 3
\end{align*}$$

where

$$n = 2, \quad A = \begin{pmatrix} 1 & 3 \\ 2 & 7 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

Question: Is the solution unique?
Definition (Nonsingular case)
A square matrix $A \in \mathbb{R}^{n \times n}$ is said to be nonsingular if any of the following equivalent conditions are satisfied:

- the inverse matrix $A^{-1}$ exists
- $\det(A) \neq 0$
- $\text{rank}(A) = n$
- $Az = 0 \implies z = 0$
- $z \neq 0 \implies Az \neq 0$

Result
If $A$ is nonsingular, then $Ax = b$ has a unique solution for any $b$

Singular case
If the square matrix $A \in \mathbb{R}^{n \times n}$ is singular (inverse does not exist), then

- if $b \in \text{span}(A)$, then infinitely many solutions exist
- if $b \notin \text{span}(A)$, then no solutions exist

Example (Singular $A$)

$$
\begin{pmatrix}
1 & 2 \\
3 & 6
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
=
\begin{pmatrix}
b_1 \\
b_2
\end{pmatrix}
$$

- $\det(A) = 1 \cdot 6 - 3 \cdot 2 = 0 \implies A$ is singular
- $b = \begin{pmatrix}4 \\ 8\end{pmatrix} \implies x = \begin{pmatrix}1 \\ 1\end{pmatrix}$, or ... infinitely many solutions
- $b = \begin{pmatrix}1 \\ 1\end{pmatrix} \implies$ no solutions
“Known” material for square $A$
- general $A$: solve $Ax = b$ using $A = LU$ factorization (Gaussian elimination)
- positive-definite $A$: solve $Ax = b$ using $A = LL^T$ factorization (Cholesky factorization)

New material for square $A$
- Conditioning: how sensitive is the solution $x$ to the system $Ax = b$ to the input data $A$ and $b$?
- To understand conditioning, we will introduce the condition number of a matrix $A$
  \[
  \text{cond}(A) \overset{\text{def}}{=} \|A\| \|A^{-1}\|
  \]
- This requires us to understand matrix norms $\|A\|$.
- Which requires us to understand vector norms (next section)

<Matlab demo 1>

### Vector norms

There are many vector norms
- $\|x\|_2 = \sqrt{x_1^2 + x_2^2 \ldots x_n^2}$ (2-norm)
- $\|x\|_1 = |x_1| + |x_2| \ldots |x_n| = \sum_{i=1}^{n} |x_i|$ (1-norm)
- $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ ($\infty$-norm)

### Example (Some vector norms)

$x = \begin{pmatrix} -12 \\ 3 \\ 4 \end{pmatrix}$
- $\|x\|_2 = 13$
- $\|x\|_1 = 19$
- $\|x\|_\infty = 12$
Sometimes a specific norm may be better than another

Suppose we have accumulated data as the result of a carefully designed experiment, and then obtained a model of the data.

If we store the error of each data point in the vector $x$ then

- $x = (10^{-43} \ 10^{-42} \ \ldots \ 3 \ \ldots \ 10^{-43})^T$
- $\|x\|_\infty = 3$ because of the outlier
- Maybe better to use $\|x\|_2/n$?

The geometry of vector norms

Some results

The following hold:

- $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$
- $\|x\|_1 \leq \sqrt{n} \|x\|_2$
- $\|x\|_2 \leq \sqrt{n} \|x\|_\infty$
- $\|x\|_1 \leq n \|x\|_\infty$
Definition (Vector norm)
A vector norm is any real-valued function $\| \cdot \|$ of a vector that satisfies the following properties:

1. If $x \neq 0$ then $\|x\| > 0$
2. $\|\alpha x\| = |\alpha|\|x\|$ for any $\alpha \in \mathbb{R}$
3. $\|x + y\| \leq \|x\| + \|y\|$ (triangle-inequality)

Using the above properties, it may be shown that

- $\|x\| = 0$ if and only if $x = 0$
- $\|x\| - \|y\| \leq \|x - y\|$ (reverse triangle-inequality)

We have already seen some examples

- $\|x\| \overset{\text{def}}{=} \|x\|_2$
- $\|x\| \overset{\text{def}}{=} \|x\|_1$
- $\|x\| \overset{\text{def}}{=} \|x\|_\infty$

Definition (Induced matrix norm)
Given a vector norm $\|x\|$, we define the induced matrix norm as

$$\|A\| \overset{\text{def}}{=} \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

- Measures the maximum amount of "elongation" resulting from multiplication by $A$
- It can be shown that
  - $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|$ (maximum absolute column sum)
  - $\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$ (maximum absolute row sum)

Example (Matrix norms)

$$A = \begin{pmatrix} -7 & 4 & 3 & 1 \\ 8 & -5 & 6 & 0 \\ -1 & -3 & 7 & 4 \\ 5 & 0 & 0 & -5 \end{pmatrix}$$

- $\|A\|_1 = 21$ and $\|A\|_\infty = 19$
Definition (Matrix norm)

A matrix norm is any real-valued function $\| \cdot \|$ of a matrix that satisfies the following properties:

1. if $A \neq 0$ then $\|A\| > 0$
2. $\|\alpha A\| = |\alpha|\|A\|$ for any $\alpha \in \mathbb{R}$
3. $\|A + B\| \leq \|A\| + \|B\|$ (triangle-inequality)

Using the above properties, it may be shown that

- $\|A\| = 0$ if and only if $A = 0$
- $\|A\| - \|B\| \leq \|A - B\| \leq \|A\| - \|B\|$ (reverse triangle-inequality)

We have already seen some examples

- $\|A\| \overset{\text{def}}{=} \|A\|_1$
- $\|A\| \overset{\text{def}}{=} \|A\|_\infty$

Induced matrix norms (not all norms) are consistent, i.e., satisfy

- $\|AB\| \leq \|A\| \|B\|$ for any $x$
- $\|Ax\| \leq \|A\| \|x\|$ for any $x$

Definition (condition number)

We define the condition number of a square matrix $A$ as

$$\text{cond}(A) = \begin{cases} 
\|A\| \|A^{-1}\| & \text{if } A \text{ is nonsingular} \\
\infty & \text{if } A \text{ is singular}
\end{cases}$$

- large condition number $\implies A$ is nearly singular
- geometric interpretation: the condition number is the ratio of the largest stretching over the smallest shrinking caused by multiplication by $A$
- the residual $r = b - Ax$ is not a reliable measure of accuracy
- for well-conditioned problems, the relative residual is reliable:

$$\frac{\|b - Ax\|}{\|x\| \|A\|}$$

- fact: backward stable algorithms produce small relative residuals

<Matlab demo 2>
\[ \text{cond}(A) = \begin{cases} \|A\| \|A^{-1}\| & \text{if } A \text{ is nonsingular} \\ \infty & \text{if } A \text{ is singular} \end{cases} \]

**Properties of the condition number**

If the condition number is defined by any induced matrix norm, then
- \( \text{cond}(I) = 1 \)
- \( \text{cond}(A) \geq 1 \)
- \( \text{cond}(\alpha A) = \text{cond}(A) \) for all \( \alpha \neq 0 \)
- If \( D \) is a diagonal matrix, then
  \[ \text{cond}(D) = \frac{\max_{1 \leq i \leq n} |d_{ii}|}{\min_{1 \leq i \leq n} |d_{ii}|} \]

**Example (Condition number of a diagonal matrix)**

\[ D = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -0.01 \end{pmatrix} \implies \text{cond}(D) = 5000 \]

**Computing the condition number**

- Computing \( \|A\| \) is computationally cheap
- Computing \( A^{-1} \) is very computationally expensive
- It is more expensive to compute \( A^{-1} \) than it is to solve \( Ax = b \)
- Some software cheaply estimates \( \text{cond}(A) \) while solving \( Ax = b \)
  - LINPACK \( \rightarrow \) sgeco
  - LAPACK \( \rightarrow \) sgecon
  - NAG \( \rightarrow \) f07agf
  - Matlab \( \rightarrow \) condest
Accuracy analysis

Suppose we are given $A$, $b$ and a perturbed right-hand-side

$$\hat{b} = b + \Delta b.$$ 

Let $x$ and $\hat{x}$ satisfy

$$Ax = b \implies \|b\| = \|Ax\| \leq \|A\|\|x\| \quad \text{(consistency)} \quad (4)$$

Define

$$\Delta x \equiv \hat{x} - x \quad A \Delta x = A(\hat{x} - x) = A\hat{x} - Ax = \hat{b} - b = \Delta b \implies \Delta x = A^{-1} \Delta b$$

Using the previous equality, the consistency property, and (4) we have

$$\implies \frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1} \Delta b\|}{\|x\|} \leq \frac{\|A\|\|A^{-1}\|\|\Delta b\|}{\|b\|}$$

This is precisely

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\hat{b} - b\|}{\|b\|}$$

With a little more work, we obtain the following perturbation result

**Theorem (Error bound for linear systems)**

If $A$ is nonsingular, $Ax = b$, and $A\hat{x} = \hat{b}$, then

$$\frac{1}{\operatorname{cond}(A)} \frac{\|\hat{b} - b\|}{\|b\|} \leq \frac{\|\hat{x} - x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\hat{b} - b\|}{\|b\|}$$

A similar analysis shows the following.

**Theorem (Error bound for linear systems)**

If $A$ is nonsingular, $Ax = b$, and $A\hat{x} = b$, then

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

provided $\|\hat{A} - A\| \leq 1/\|A^{-1}\|$.

- Similar result holds when $A$ and $b$ are perturbed simultaneously.
- What does this mean in terms of computer representation?
What does this mean in terms of computer representation?

1. We give the computer \( A \) and \( b \) and want to find \( x \) such that \( Ax = b \). We assume that \( A \) is exactly representable, but that \( b \) is not.

2. Define \( \hat{b} = \Phi(b) \) so that \( \hat{b} \) satisfies
   \[
   \frac{\|\hat{b} - b\|}{\|b\|} = \frac{\|\Phi(b) - b\|}{\|b\|} \leq \epsilon_{\text{mach}}
   \]

3. We solve \( A\hat{x} = \hat{b} \)

4. From result on previous slide we know that
   \[
   \frac{\|\hat{x} - x\|}{\|x\|} \leq \text{cond}(A) \frac{\|\hat{b} - b\|}{\|b\|} \leq \text{cond}(A) \epsilon_{\text{mach}}
   \]

<Matlab demo 3>

Theorem (Geometric interpretation of the condition number)

\[
\frac{1}{\text{cond}(A)} = \inf \left\{ \frac{\|A - B\|}{\|A\|} : B \text{ is not invertible} \right\}
\]

Thus, the reciprocal of the condition number measures the (normalized) distance to the closest singular matrix.
If computer arithmetic was exact, writing programs would be “easy”
- prove that an algorithm works
- implement the algorithm verbatim
- watch it solve every problem that it ever encounters!

Computer arithmetic is not exact
- writing good code is a combination of
  ▶ science
  ▶ attention to detail
  ▶ organization
  ▶ experience
  ▶ black art
- You will likely run into numerical issues while writing Matlab code for this course, but with some tricks/techniques you can avoid unnecessary trouble

Solving linear systems

- In Matlab, you can compute the inverse of a matrix $A$ with
  $$A_{inv} = \text{inv}(A);$$

- DO NOT DO THIS!
- When Matlab computes $A^{-1}$ it
  ▶ is creating numerical error
  ▶ is very costly
- It is much better to solve the linear system $Ax = b$ by typing
  $$x = A\backslash b;$$
  so that Matlab may use a stable, fast, direct method (i.e., a factorization of $A$)
- Due to ill-conditioning, however, do not always assume that the results are accurate!
Termination tests

- Numerical algorithms require a termination test to know when to stop
- **Example:** for finding $x$ such that $F(x) = 0$, we may choose to stop when
  $$\|F(x_k)\| \leq \varepsilon$$
  for some $\varepsilon \geq 0$

  where $\{x_k\}_{k \geq 0}$ are the iterates generated by the algorithm
- If you choose $\varepsilon = 0$, your code will typically **never** stop in practice
- If you choose $\varepsilon = 10^{-15}$, your code will **rarely** stop in practice
- A good choice is something like
  $$\varepsilon = 10^{-6}\|F(x_0)\|$$
  so that the algorithm terminates when the relative tolerance
  $$\frac{\|F(x_k)\|}{\|F(x_0)\|} \leq 10^{-6}$$
  is satisfied
- Why not stop when $\|x_k - x_{k+1}\|$ is small?

Arithmetic anomalies

- In your code, you may make a decision that depends on verifying whether two quantities are equal
  - **DO NOT DO THIS!**
- **Example:** if you verify the equality $3 = (\sqrt{3})^2$ at a Matlab prompt by typing
  $$3 == \text{sqrt}(3)^2$$
  it will return 0, i.e., false!
- A better strategy is something similar to
  $$(3 <= \text{sqrt}(3)^2 + 1e-12) \& (3 >= \text{sqrt}(3)^2 - 1e-12)$$
  which returns 1, i.e., true
- You may also find (e.g., in line-search methods that will be discussed later) that for three numbers $a \approx b$ and $c \approx 0$, the expression
  $$a \leq b - c$$
  may yield false, but the expression
  $$a - b \leq -c$$
  yields true! (the second is desirable in the context of line-search methods)
Other sources

- Dividing large numbers by small numbers
- Catastrophic cancellation
- Matrix-matrix, matrix-vector, or vector-vector operations
- Computing eigenvalues of a matrix \( A \) numerically
- Computing solutions of linear systems numerically
- Finding a zero of a function numerically
- Poor problem scaling, e.g., finding \( x = (x_1, x_2) \) satisfying

\[
\frac{x_1^2 - x_2}{4x_1 - 5x_2} = 0 \quad \text{versus} \quad \frac{10^4(x_1^2 - x_2)}{10^6(4x_1 - 5x_2)} = 0
\]

- Practically anything!


Notes

- For optimization theory and developing algorithms, we require tools for describing how function values change with their inputs.
- When derivatives exist, we use results from Calculus; e.g., gradients and Hessians

**Definition**

If \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is differentiable, the **gradient** of \( f \) at \( x \) is

\[
\nabla f(x) \overset{\text{def}}{=} \begin{pmatrix}
\frac{\partial f(x)}{\partial x_1} \\
\vdots \\
\frac{\partial f(x)}{\partial x_n}
\end{pmatrix}
\]

**Definition**

If \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is twice differentiable, the **Hessian** of \( f \) at \( x \) is

\[
\nabla^2 f(x) \overset{\text{def}}{=} \begin{pmatrix}
\frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\。
\vdots & \ddots & \vdots \\.
\frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2}
\end{pmatrix}
\]
Theorem (One-dimensional slices of multivariate functions)

Let \( f : \mathbb{R}^n \to \mathbb{R} \). Consider any \( x, s \in \mathbb{R}^n \) and define \( \phi : \mathbb{R} \to \mathbb{R} \) as

\[
\phi(\lambda) = f(x + \lambda s).
\]

- If \( f \) is differentiable, then so is \( \phi \) and for any \( \bar{\lambda} \in \mathbb{R} \),
  \[
  \phi'(\bar{\lambda}) = \nabla f(x + \bar{\lambda}s)^T s.
  \]
- If \( f \) is twice differentiable, then so is \( \phi \) and for any \( \bar{\lambda} \in \mathbb{R} \),
  \[
  \phi''(\bar{\lambda}) = s^T \nabla^2 f(x + \bar{\lambda}s)s.
  \]

Theorem (One-dimensional Mean Value Theorems)

Let \( \phi : \mathbb{R} \to \mathbb{R} \).

- Suppose \( \phi \) is differentiable. Then for any \( a < b \in \mathbb{R} \), there exists \( c \in (a, b) \) such that
  \[
  \phi(b) = \phi(a) + \phi'(c)(b - a).
  \]
- Suppose \( \phi \) is twice differentiable. Then for any \( a < b \in \mathbb{R} \), there exists \( c \in (a, b) \) such that
  \[
  \phi(b) = \phi(a) + \phi'(a)(b - a) + \frac{1}{2}\phi''(c)(b - a)^2.
  \]

Theorem (Higher-dimensional Mean Value Theorem)

Let \( S \) be an open subset of \( \mathbb{R}^n \) and let \( f : S \to \mathbb{R} \).

- Suppose \( f \) is differentiable throughout \( S \). Then for any \( x \in S \) and \( s \neq 0 \in \mathbb{R}^n \), such that the interval \( [x, x + s] \in S \), there exists \( z \in (x, x + s) \) such that
  \[
  f(x + s) = f(x) + \nabla f(z)^T s.
  \]
- Suppose \( f \) is twice differentiable throughout \( S \). Then for any \( x \in S \) and \( s \neq 0 \in \mathbb{R}^n \), such that the interval \( [x, x + s] \in S \), there exists \( z \in (x, x + s) \) such that
  \[
  f(x + s) = f(x) + g(x)^T s + \frac{1}{2}s^T H(z)s
  \]
Definition (Lipschitz continuity)

Suppose that

- $\mathcal{X}$ and $\mathcal{Y}$ open sets
- $F : \mathcal{X} \rightarrow \mathcal{Y}$
- $\| \cdot \|_\mathcal{X}$ and $\| \cdot \|_\mathcal{Y}$ are norms

Then

- $F$ is Lipschitz continuous at $x \in \mathcal{X}$ if $\exists \gamma(x) \in \mathbb{R}$ such that

$$\|F(z) - F(x)\|_\mathcal{Y} \leq \gamma(x) \|z - x\|_\mathcal{X}$$

for all $z \in \mathcal{X}$.

- $F$ is Lipschitz continuous throughout/in $\mathcal{X}$ if $\exists \gamma \in \mathbb{R}$ such that

$$\|F(z) - F(x)\|_\mathcal{Y} \leq \gamma \|z - x\|_\mathcal{X}$$

for all $x$ and $z \in \mathcal{X}$.

Theorem (Taylor approximations for real-valued functions)

Let $S$ be an open subset of $\mathbb{R}^n$, $s \in \mathbb{R}^n$, and suppose that $f : S \rightarrow \mathbb{R}$ is continuously differentiable throughout $S$ and $g = \nabla f$ is Lipschitz continuous at $x$ with Lipschitz constant $\gamma_f(x)$ for some appropriate vector norm. It follows that if the segment $[x, x + s] \subset S$, then

$$|f(x + s) - m^f(x + s)| \leq \frac{1}{2} \gamma_f(x) \|s\|^2,$$

where

$$m^f(x + s) = f(x) + g(x)^T s.$$

If in addition, $f$ is twice continuously differentiable throughout $S$ and $H = \nabla^2 f$ is Lipschitz continuous at $x$, with Lipschitz constant $\gamma_H(x)$, then

$$|f(x + s) - m^\theta(x + s)| \leq \frac{1}{4} \gamma_H(x) \|s\|^4,$$

where

$$m^\theta(x + s) = f(x) + g(x)^T s + \frac{1}{2} s^T H(x) s.$$
Definition (Differential of vector-valued function)

If \( F : \mathbb{R}^n \to \mathbb{R}^m \) is differentiable, the Jacobian of \( F \) at \( x \) is

\[
J(x) := \nabla F(x) \overset{\text{def}}{=} \begin{pmatrix}
\frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial F_m(x)}{\partial x_1} & \cdots & \frac{\partial F_m(x)}{\partial x_n}
\end{pmatrix}
\]

where \( F_i(x), i = 1, \ldots, m \) is the \( i \)-th component of \( F(x) \).

Theorem (Taylor approximation for vector-valued functions)

Let \( S \) be an open subset of \( \mathbb{R}^n \), \( s \in \mathbb{R}^n \), and suppose that \( F : S \to \mathbb{R}^m \) is continuously differentiable throughout \( S \) and that \( \nabla F(x) \) is Lipschitz continuous at \( x \) with Lipschitz constant \( \gamma^F(x) \) for some appropriate vector norm and its induced matrix norm. It follows that if the segment \( [x, x + s] \subset S \), then

\[
\|F(x + s) - M^F(x + s)\|_{\mathbb{R}^m} \leq \frac{1}{2} \gamma^F(x) \|s\|^2_{\mathbb{R}^n},
\]

where

\[
M^F(x + s) = F(x) + \nabla F(x)s.
\]