Background and basics

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

September 1, 2020

Notes

Outline

Notes

Floating-point (real) numbers

Modern computers store real numbers as

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right) \beta^E \qquad (4.362781 * 10^{-06})$$

- base: β (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 β, p, L, and U
- mantissa: $d_0 d_1 \ldots d_{p-1}$
- 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

Notes

Example (Floating-point system)

$$\beta = 10, \quad p = 4, \quad L = -99, \quad \text{and} \quad U = 99$$

- some numbers
  - $1 = 1.000 * 10^{00}$
  - $34.67 = 3.467 * 10^{01}$
  - $0.0346 = 3.460 * 10^{-02}$
- smallest positive number: $1.000 * 10^{-99}$ (underflow level)
- largest number: $9.999 * 10^{99}$ (overflow level)

Notes

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

$$2(\beta - 1)\beta^{p-1}(U - L + 1) + 1$$

- Smallest positive number: $\mathbf{UFL} = \beta^L$ (underflow level)
  - numbers smaller than $\mathbf{UFL}$ stored as zero
  - often not serious, because zero is a good approximation
- Largest number: $\mathbf{OFL} = \beta^{U+1}(1 - \beta^{-p})$ (overflow level)
  - numbers larger than $\mathbf{OFL}$ may not be stored
  - serious problem, compilers typically terminate

Notes

Rounding

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

- Common rounding strategies
  - chopping: $\mathbf{fl}(x)$ is obtained by truncating the expansion of x after d_{p-1}. Also called round-to-zero.
  - round-to-nearest: $\mathbf{fl}(x)$ 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

Notes

Example (Motivation of machine precision)

Consider the following numbers x and their nearest neighbor to the "right" xr (using $\beta = 10$ and $p = 4$)

$$\left.\begin{aligned} x &= 1.000 * 10^{00} \\ xr &= 1.001 * 10^{00} \end{aligned}\right\} \text{ Distance is } 10^{-03}$$

$$\left.\begin{aligned} x &= 1.000 * 10^{06} \\ xr &= 1.001 * 10^{06} \end{aligned}\right\} \text{ Distance is } 10^{+03}$$

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

Machine precision assuming round-to-nearest

$$\varepsilon_{\text{mach}} \stackrel{\text{def}}{=} \frac{1}{2}\beta^{1-p}$$

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

$$\frac{|\mathbf{fl}(x) - x|}{|x|} \leq \varepsilon_{\text{mach}}$$

Notes

Definition (Machine precision)

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

- the smallest number ε such that $\mathbf{fl}(1 + \varepsilon) > 1$
- the largest number ε such that $\mathbf{fl}(1 + \varepsilon) = 1$
- 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 $\varepsilon_{\text{mach}}$)

Using round-to-nearest, $p = 4$, and $\beta = 10$, we have

$$1.000 + 0.0005 = 1.0005 \stackrel{\text{comp}}{=} 1.000$$
$$1.000 + 0.00051 = 1.00051 \stackrel{\text{comp}}{=} 1.001$$

$$\Longrightarrow \quad \varepsilon_{\text{mach}} = 0.0005 = 5 * 10^{-04} = \frac{1}{2} * 10^{-03} = \frac{1}{2}\beta^{1-p}$$

Comment: Generally, $0 < \text{UFL} < \varepsilon_{\text{mach}} < \text{OFL}$

Notes

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., $0/0$, $0 * \infty$, ∞/∞)

Notes

$$x = 4.452 * 10^{02} \text{ and } y = 6.436 * 10^{-01}$$

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:

$$\begin{aligned} x * y &= (4.452 * 10^{02}) * (6.436 * 10^{-01}) = 28.653072 * 10^{01} \\ &= 2.8653072 * 10^{02} \overset{\text{comp}}{=} 2.865 * 10^{02} \end{aligned}$$

- Addition of two floating-point numbers (similar for subtraction)
  - shift so that exponents are the same, add, then re-normalize
  - example:

$$\begin{aligned} x + y &= (4.452 * 10^{02}) + (6.436 * 10^{-01}) \\ &= (4.452 * 10^{02}) + (0.006436 * 10^{02}) \\ &= 4.458436 * 10^{02} \overset{\text{comp}}{=} 4.458 * 10^{02} \end{aligned}$$

  - generally, trailing digits of smaller (in magnitude) number are lost

Notes

Example (Motivate concept of catastrophic cancellation)

Consider computing for some a and b the following:

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

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

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

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

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

Not even the correct sign!

Notes

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)

$$\left.\begin{aligned} x_1 + 3x_2 &= 5 \\ 2x_1 + 7x_2 &= 3 \end{aligned}\right\} \Rightarrow Ax = b$$

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?

Notes

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
- $\det(A) \neq 0$
- $\operatorname{rank}(A) = n$
- $Az = 0 \implies z = 0$
- $z \neq 0 \implies Az \neq 0$

Result

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

Notes

Singular case

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

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

Example (Singular A)

$$\overbrace{\begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}}^{A} \overbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}^{x} = \overbrace{\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}}^{b}$$

- $\det(A) = 1 * 6 - 3 * 2 = 0 \implies A$ is singular
- $b = \begin{pmatrix} 4 \\ 8 \end{pmatrix} \implies x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, x = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$, or ... infinitely many solutions
- $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \implies$ no solutions

Notes

"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

$$\mathbf{cond}(A) \stackrel{\text{def}}{=} ||A||\,||A^{-1}||$$

- This requires us to understand matrix norms $||A||$
- Which requires us to understand vector norms (next section)

<Matlab demo 1>

Notes

Vector norms

There are many vector norms

- $||x||_2 = \sqrt{x_1^2 + x_2^2 \dots x_n^2} = \sqrt{\sum_{i=1}^n x_i^2}$ (2-norm)
- $||x||_1 = |x_1| + |x_2| \dots |x_n| = \sum_{i=1}^n |x_i|$ (1-norm)
- $||x||_\infty = \max_{1 \le i \le n} |x_i|$ (∞-norm)

Example (Some vector norms)

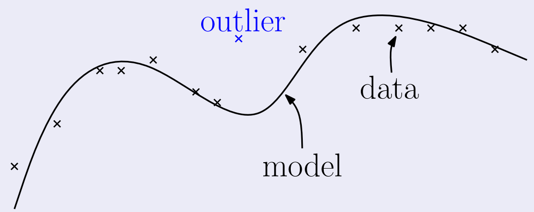
$$x = \begin{pmatrix} -12 \\ 3 \\ 4 \end{pmatrix}$$

- $||x||_2 = 13$
- $||x||_1 = 19$
- $||x||_\infty = 12$

Notes

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.

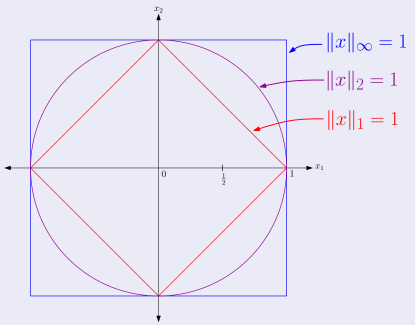


If we store the error of each data point in the vector $\boldsymbol{x}$ then

- $\boldsymbol{x} = \begin{pmatrix} 10^{-03} & 10^{-02} & \cdots & 3 & \cdots & 10^{-03} \end{pmatrix}^T$
- $\|\boldsymbol{x}\|_\infty = 3$ because of the outlier
- Maybe better to use $\|\boldsymbol{x}\|_2/n$?

Notes

The geometry of vector norms



Some results

The following hold:

- $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1$
- $\|\boldsymbol{x}\|_1 \leq \sqrt{n}\|\boldsymbol{x}\|_2$
- $\|\boldsymbol{x}\|_2 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty$
- $\|\boldsymbol{x}\|_1 \leq n\|\boldsymbol{x}\|_\infty$

Notes

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 $x = 0$
  - $||x|| - ||y|| \leq |||x|| - ||y||| \leq ||x - y||$ (reverse triangle-inequality)
- We have already seen some examples

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

Notes

Definition (Induced matrix norm)

Given a vector norm $||x||$, we define the induced matrix norm as

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

- 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 $||A||_\infty = 19$

Notes

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

$$||A|| \stackrel{\text{def}}{=} ||A||_1$$

$$||A|| \stackrel{\text{def}}{=} ||A||_\infty$$

- Induced matrix norms (not all norms) are consistent, i.e., satisfy
  - $||AB|| \leq ||A||||B||$
  - $||Ax|| \leq ||A||||x||$ for any x

Notes

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 $r = b - A\hat{x}$ is not a reliable measure of accuracy
- for well-conditioned problems, the relative residual is reliable:

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

- fact: backward stable algorithms produce small relative residuals

<Matlab demo 2>

Notes

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

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)

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

Notes

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

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 $\longrightarrow$ sgeco
  - LAPACK $\longrightarrow$ sgecon
  - NAG $\longrightarrow$ f07agf
  - Matlab $\longrightarrow$ condest

Notes

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

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

Define

$$\Delta x \overset{\text{def}}{=} \hat{x} - x$$

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

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

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

This is precisely

$$\frac{\|\hat{x} - x\|}{\|x\|} \le \mathbf{cond}(A)\frac{\|\hat{b} - b\|}{\|b\|}$$

Notes

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

$$\frac{1}{\mathbf{cond}(A)}\frac{\|\hat{b} - b\|}{\|b\|} \le \frac{\|\hat{x} - x\|}{\|x\|} \le \mathbf{cond}(A)\frac{\|\hat{b} - b\|}{\|b\|}$$

A similar analysis shows the following.

Theorem (Error bound for linear systems)

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

$$\frac{\|\hat{x} - x\|}{\|x\|} \le \frac{\mathbf{cond}(A)}{1 - \mathbf{cond}(A)\frac{\|\hat{A} - A\|}{\|A\|}}\frac{\|\hat{A} - A\|}{\|A\|}$$

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

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

Notes

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 $\widehat{b} = \mathbf{fl}(b)$ so that $\widehat{b}$ satisfies

$$\frac{\|\widehat{b} - b\|}{\|b\|} = \frac{\|\mathbf{fl}(b) - b\|}{\|b\|} \leq \varepsilon_{\text{mach}}$$

3. We solve $A\widehat{x} = \widehat{b}$
4. From result on previous slide we know that

$$\frac{\|\widehat{x} - x\|}{\|x\|} \leq \mathbf{cond}(A)\frac{\|\widehat{b} - b\|}{\|b\|} \leq \mathbf{cond}(A)\varepsilon_{\text{mach}}$$

<Matlab demo 3>

Notes

Theorem (Geometric interpretation of the condition number)

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

Thus, the reciprocal of the condition number measures the (normalized) distance to the closest singular matrix.

Notes

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

Notes

Solving linear systems

- In Matlab, you can compute the inverse of a matrix $\boldsymbol{A}$ with

```
Ainv = inv(A);
```

- DO NOT DO THIS!
- When Matlab computes $\boldsymbol{A}^{-1}$ it
  - is creating numerical error
  - is very costly
- It is much better to solve the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ by typing

```
x = A\b;
```

  so that Matlab may use a stable, fast, direct method (i.e., a factorization of $\boldsymbol{A}$)
- Due to ill-conditioning, however, do not always assume that the results are accurate!

Notes

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

$$\|F(x_k)\| \leq \varepsilon \quad \text{for some } \varepsilon \geq 0$$

  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?

Notes

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 == sqrt(3)^2
```

  it will return 0, i.e., false!
- A better strategy is something similar to

```
(3 <= sqrt(3)^2 + 1e-12) & (3 >= 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)

Notes

Other sources

- Dividing large numbers by small numbers
- Catastrophic cancellation
- Matix-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

$$\begin{pmatrix} x_1^2 - x_2 \\ 4x_1 - 5x_2 \end{pmatrix} = 0 \quad \text{versus} \quad \begin{pmatrix} x_1^2 - x_2 \\ 10^6(4x_1 - 5x_2) \end{pmatrix} = 0$$

- Practically anything!

https://www.mathworks.com/help/matlab/matlab_prog/floating-point-numbers.html

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 $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, the gradient of f at x is

$$\nabla f(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

Definition

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, the Hessian of f at x is

$$\nabla^2 f(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}$$

Notes

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

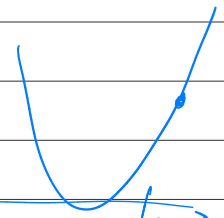
$$\phi(\lambda) = f(x + \lambda s).$$

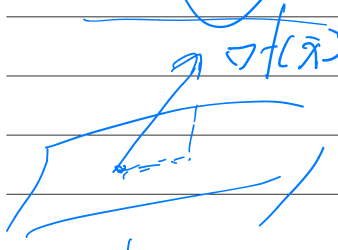
- *If f is differentiable, then so is ϕ and for any $\bar{\lambda} \in \mathbb{R}$,*
$$\phi'(\bar{\lambda}) = \nabla f(x + \bar{\lambda} s)^T s.$$
- *If f is twice differentiable, then so is ϕ and for any $\bar{\lambda} \in \mathbb{R}$,*
$$\phi''(\bar{\lambda}) = s^T \nabla^2 f(x + \bar{\lambda} s) s.$$

Notes

$\frac{1}{2}p^TAp - b^Tp$ $\quad Ap - b = 0$

$p = A^{-1}b$



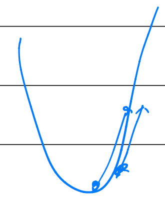


$\nabla f(\bar{x})$

$\langle \nabla f(\bar{x}), \sum_{i=1}^{k} \lambda_i s_i \rangle$

$\neq 0$

$\nabla f(\bar{x})$



projection to $x_0 + L$

Theorem (One-dimensional Mean Value Theorems)

Let $\phi : \mathbb{R} \to \mathbb{R}$.

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

Notes

$$S(\nabla f(\bar{x})) = \sum_{i=1}^{k} \frac{\langle \nabla f(\bar{x}), s_i \rangle}{\langle s_i, s_i \rangle} s_i$$

$$f(x+s) = f(x) + \nabla f(z)^T s$$

Theorem (Higher-dimensional Mean Value Theorem)

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

- *Suppose f is differentiable throughout $\mathcal{S}$. Then for any $x \in \mathcal{S}$ and $s \neq 0 \in \mathbb{R}$, such that the interval $[x, x + s] \in \mathcal{S}$, 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 $\mathcal{S}$. Then for any $x \in \mathcal{S}$ and $s \neq 0 \in \mathbb{R}$, such that the interval $[x, x + s] \in \mathcal{S}$, there exists $z \in (x, x + s)$ such that*
$$f(x + s) = f(x) + g(x)^T s + \tfrac{1}{2} s^T H(z) s$$

Definition (Lipschitz continuity)

Suppose that

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

Then

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

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

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

Notes

Theorem (Taylor approximations for real-valued functions)

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

$$|f(x+s) - m^L(x+s)| \leq \tfrac{1}{2}\gamma^L(x)\|s\|^2,$$

where

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

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

$$|f(x+s) - m^Q(x+s)| \leq \tfrac{1}{6}\gamma^Q(x)\|s\|^3,$$

where

$$m^Q(x+s) = f(x) + g(x)^T s + \tfrac{1}{2}s^T H(x)s.$$

Notes

Definition (Differential of vector-valued function)

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

$$J(x) := \nabla F(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n(x)}{\partial x_1} & \cdots & \frac{\partial F_n(x)}{\partial x_n} \end{pmatrix}$$

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

Theorem (Taylor approximation for vector-valued functions)

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

$$\|F(x + s) - M^L(x + s)\|_{\mathbb{R}^m} \leq \tfrac{1}{2}\gamma^L(x)\|s\|^2_{\mathbb{R}^n},$$

where

$$M^L(x + s) = F(x) + \nabla F(x)s.$$

Notes

Notes