Mathematics for Machine Learning

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Maxim

- Matrix -- *The mother of all data structures. The nonmathematical uses of the word `matrix' reflect its Latin origins in `mater', or mother.... The word has two meanings -- a representation of a linear mapping and the basis for all our existence.*

---Cleve Moler

Chapter 0. Vectors and Matrices

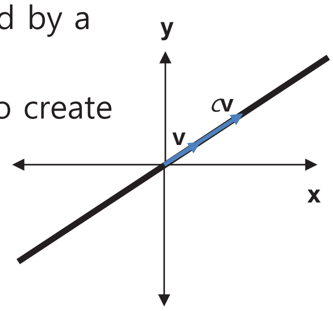
Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Reading

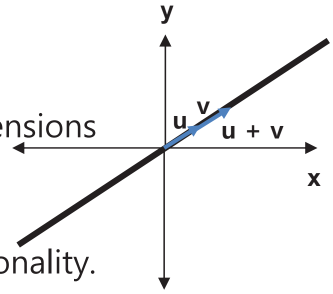
- [Strang. (2006), Chapter 1 Matrices and Gaussian Elimination]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.
  - *Linear Algebra has become as basic and as applicable as calculus, and fortunately it is easier.* --Gilbert Strang, MIT
- Prof. Gilbert Strang's course videos:
  - http://ocw.mit.edu/OcwWeb/Mathematics/18-06Spring-2005/VideoLectures/index.htm
- Borrows some slides from S. Kalyanaraman, *Linear Algebra A gentle introduction*.

What is “Linear” & “Algebra”?

- Properties satisfied by a *line through the origin* (“one-dimensional case”.
  - A directed arrow from the origin (**v**) on the line, when scaled by a constant (*c*) remains on the line
  - Two directed arrows (**u** and **v**) on the line can be “added” to create a longer directed arrow (**u** + **v**) in the same line.



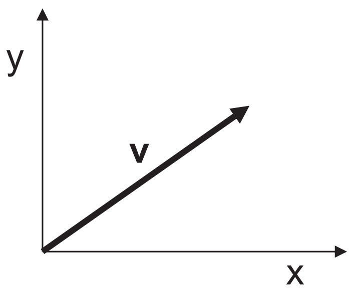
- Wait a minute! This is nothing but *arithmetic with symbols*!
  - “Algebra”: generalization and extension of arithmetic.
  - “Linear” operations: addition and scaling.
- Abstract and Generalize !
  - “Line” ↔ **vector space** having N dimensions
  - “Point” ↔ **vector** with N components in each of the N dimensions (***basis*** vectors).
    - Vectors have: “Length” and “Direction”.
    - Basis vectors: “span” or define the space & its dimensionality.
  - Linear function transforming vectors ↔ **matrix**.
    - The function acts on each vector component and scales it
    - Add up the resulting scaled components to get a new vector!
    - In general: f(c**u** + d**v**) = cf(**u**) + df(**v**)



What is a Vector ?

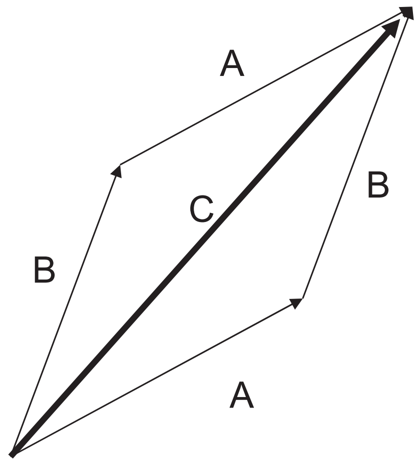
- Think of a vector as a *directed line segment in N-dimensions*! (has “length” and “direction”)
- Basic idea: convert geometry in higher dimensions into algebra!
  - Once you define a “nice” *basis* along each di mension: x-, y-, z-axis …
  - Vector becomes a 1 x N matrix!
  - **v** = [a b c]^T
  - Geometry starts to become linear algebra on vectors like **v**!

$$\vec{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$



Vector Addition: A+B

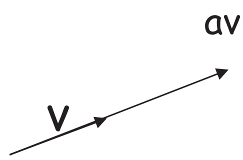
$$\mathbf{A+B} = (x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$$



A+B = C
(use the head-to-tail method to combine vectors)

Scalar Product: $a\mathbf{v}$

$$a\mathbf{v} = a(x_1, x_2) = (ax_1, ax_2)$$



Change only the length ("scaling"), but keep *direction fixed*.

Sneak peek: matrix operation (**Av**) can change *length, direction and also dimensionality*!

Vectors: Dot Product

$$A \cdot B = A^T B = \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} d \\ e \\ f \end{bmatrix} = ad + be + cf$$

Think of the dot product as a matrix multiplication

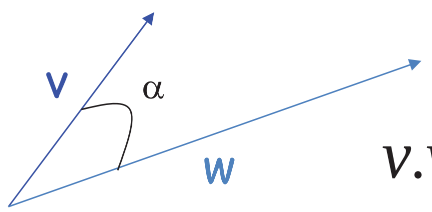
$$\|A\|^2 = A^T A = aa + bb + cc$$

The magnitude is the dot produ ct of a vector with itself

$$A \cdot B = \|A\| \, \|B\| \cos(\theta)$$

The dot product is also related to the angle between the two vectors

Inner (dot) Product: v.w or w^{T}v



$$v.w = (x_1, x_2).(y_1, y_2) = x_1 y_1 + x_2 . y_2$$

The inner product is a **SCALAR!**

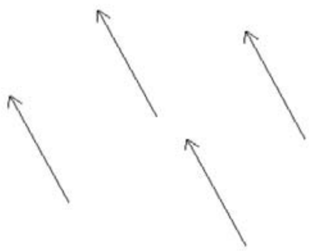
$$v.w = (x_1, x_2).(y_1, y_2) = \| v \| \cdot \| w \| \cos\alpha$$

$$v.w = 0 \Leftrightarrow v \perp w$$

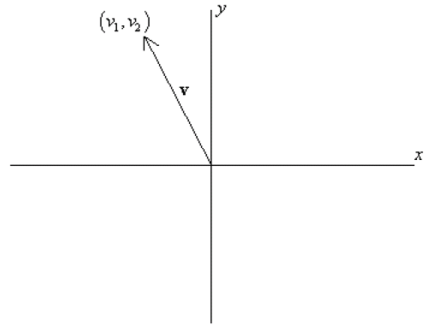
If vectors **v, w** are "columns", then dot product is $\mathbf{w}^T\mathbf{v}$

Bases & Orthonormal Bases

- Basis (or axes): frame of reference



VS



Basis: a space is totally defined by a set of vectors – any point is a *linear combination* of the basis

Ortho-Normal: orthogonal + normal

[**Sneak peek**:
Orthogonal: dot product is zero
Normal: magnitude is one]

$$x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \qquad x \cdot y = 0$$
$$y = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \qquad x \cdot z = 0$$
$$z = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \qquad y \cdot z = 0$$

What is a Matrix?

- A matrix is a set of elements, organized into rows and columns

rows

columns

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Special matrices

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$ diagonal

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$$ upper-triangular

$$\begin{pmatrix} a & b & 0 & 0 \\ c & d & e & 0 \\ 0 & f & g & h \\ 0 & 0 & i & j \end{pmatrix}$$ tri-diagonal

$$\begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix}$$ lower-triangular

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$ I (identity matrix)

Matrices

Matrix locations/size defined as rows x columns (R x C)

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix}$$

d_{ij} : i^{th} row, j^{th} column

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

Square (3 x 3)

$$\begin{bmatrix} x11 & x12 & x13 \\ x21 & x22 & x23 \\ x31 & x32 & x33 \end{bmatrix}$$

3 dimensional (3 x 3 x 5)

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Rectangular (3 x 2)

Basic Matrix Operations

- Addition, Subtraction, Multiplication: creating new matrices (or functio ns)

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

Just add elements

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a-e & b-f \\ c-g & d-h \end{bmatrix}$$

Just subtract elements

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

Multiply each row b y each column

Example: Matrix Calculations

- Addition
  - Commutative: A+B=B+A
  - Associative: (A+B)+C=A+(B+C)

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} 2 & 4 \\ 2 & 5 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 2+1 & 4+0 \\ 2+3 & 5+1 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$$

- Subtraction
  - By adding a negative matrix

$$\mathbf{A}-\mathbf{B} = \begin{bmatrix} 2 & 4 \\ 5 & 3 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 5 & 3 \end{bmatrix} + \begin{bmatrix} -1 & -2 \\ -3 & -4 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$$

Basic Matrix Operations

- Transpose: You can think of it as
  - "flipping" the rows and columns

    OR
  - "reflecting" vector/matrix on line

e.g.

$$\begin{pmatrix} a \\ b \end{pmatrix}^T = \begin{pmatrix} a & b \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A+B)^T = A^T + B^T$

Example: Transposition

$$\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \qquad \mathbf{b}^T = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix} \qquad \mathbf{d} = \begin{bmatrix} 3 & 4 & 9 \end{bmatrix} \qquad \mathbf{d}^T = \begin{bmatrix} 3 \\ 4 \\ 9 \end{bmatrix}$$

column → row row → column

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 4 & 1 \\ 6 & 7 & 4 \end{bmatrix} \qquad \mathbf{A}^T = \begin{bmatrix} 1 & 5 & 6 \\ 2 & 4 & 7 \\ 3 & 1 & 4 \end{bmatrix}$$

Matrix Multiplication

*"When **A** is a **mxn** matrix & **B** is a **kxl** matrix, **AB** is only possible if **n=k**. The result will be an **mxl** matrix"*

Simply put, can ONLY perform A*B IF:

Number of columns in A = Number of rows in B

$$\begin{array}{cc} & \overrightarrow{\quad \mathbf{n} \quad} \\ \mathbf{m}\downarrow & \begin{bmatrix} A_1 & A_2 & A_3 \\ A_4 & A_5 & A_6 \\ A_7 & A_8 & A_9 \\ A_{10} & A_{11} & A_{12} \end{bmatrix} \end{array} \times \begin{array}{cc} \overrightarrow{\quad \mathbf{l} \quad} & \\ \begin{bmatrix} B_{13} & B_{14} \\ B_{15} & B_{16} \\ B_{17} & B_{18} \end{bmatrix} & \downarrow \mathbf{k} \end{array} = \mathbf{m} \times \mathbf{l} \text{ matrix}$$

Matrix Times Matrix

$$\mathbf{L} = \mathbf{M} \cdot \mathbf{N}$$

$$\begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix} = \begin{bmatrix} \cancel{m_{11}} & \cancel{m_{12}} & \cancel{m_{13}} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \cdot \begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ n_{31} & n_{32} & n_{33} \end{bmatrix}$$

$$l_{12} = m_{11}n_{12} + m_{12}n_{22} + m_{13}n_{32}$$

Multiplication

- Is AB = BA? Maybe, but maybe not!

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & \dots \\ \dots & \dots \end{bmatrix} \qquad \begin{bmatrix} e & f \\ g & h \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} ea+fc & \dots \\ \dots & \dots \end{bmatrix}$$

- Matrix multiplication AB: apply transformation B first, and then again transfor m using A!
- Heads up: multiplication is NOT commutative!

- **Note**: If A and B both represent either pure "*rotation*" or "*scaling*" they can be interchanged (i.e. AB = BA)

Matrix multiplication

- Matrix multiplication is NOT commutative i.e the order matters!
  - $AB \neq BA$

- Matrix multiplication IS associative
  - A(BC)=(AB)C

- Matrix multiplication IS distributive
  - A(B+C)=AB+AC
  - (A+B)C=AC+BC

Identity matrix

- Identity matrix
  - A special matrix which plays a similar role as the number 1 in number multiplication?

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

For any nxn matrix $\boldsymbol{A}$, we have $\boldsymbol{A}\,\boldsymbol{I}_n = \boldsymbol{I}_n\,\boldsymbol{A} = \boldsymbol{A}$

For any nxm matrix $\boldsymbol{A}$, we have $\boldsymbol{I}_n\,\boldsymbol{A} = \boldsymbol{A}$, and $\boldsymbol{A}\,\boldsymbol{I}_m = \boldsymbol{A}$ (so 2 possible matrices)

If the answers always A, why use an identity matrix?

Can't divide matrices, therefore to solve may problems have to use the inverse. The identity is important in these types of calculations.

Example: Identity matrix

Worked example $\boldsymbol{A}\,\boldsymbol{I}_3 = \boldsymbol{A}$ for a 3x3 matrix:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1+0+0 & 0+2+0 & 0+0+3 \\ 4+0+0 & 0+5+0 & 0+0+6 \\ 7+0+0 & 0+8+0 & 0+0+9 \end{bmatrix}$$

Inverse of a matrix

- Inverse of a square matrix A, denoted by A^{-1} is the unique matrix s.t.
  - $AA^{-1} = A^{-1}A = I$ (identity matrix)
- If A^{-1} and B^{-1} exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$,
  - $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$
  - $(A^T)^{-1} = (A^{-1})^T$
- For orthonormal matrices $\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T}$
- For diagonal matrices

$$\mathbf{D}^{-1} = \mathrm{diag}\{d_1^{-1}, \ldots, d_n^{-1}\}$$

Mathematics for Machine Learning

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

The Best of the 20th Century: Top 10 Algorithms

- Top 10 Algorithms
  - 1946: The Metropolis Algorithm for Monte Carlo.
  - 1947: Simplex Method for Linear Programming. (→ OR1)
  - 1950: Krylov Subspace Iteration Method.
  - 1951: The Decompositional Approach to Matrix Computations.
  - 1957: The Fortran Optimizing Compiler.
  - 1959: QR Algorithm for Computing Eigenvalues.
  - 1962: Quicksort Algorithms for Sorting.
  - 1965: Fast Fourier Transform.
  - 1977: Integer Relation Detection.
  - 1987: Fast Multipole Method.

Source: The Best of the 20th Century: Editors Name Top 10 Algorithms, B. A. Cipra, SIAM News

Chapter 1. Linear Equations

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Maxim

- In truth, it is not knowledge, but learning, not possessing, but production, not being there, but traveling there, which provides the greatest pleasure. When I have completely understood something, then I turn away and move on into the dark; indeed, so curious is the insatiable man, that when he has completed one house, rather than living in it peacefully, he starts to build another.

--- C. F. Gauss

Reading

- [Strang. (2006), Chapter 1 Matrices and Gaussian Elimination]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.

Introduction

The Geometry of Linear Equations

- The central problem of linear algebra is the solution of linear equations.
- Example,

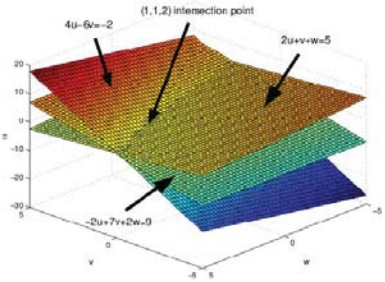
$$2u+v+w=5$$
$$4u-6v=-2$$
$$-2u+7v+2w=9$$

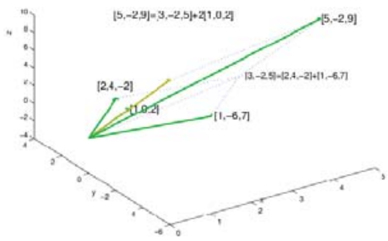
$$u\begin{bmatrix}2\\4\\-2\end{bmatrix}+v\begin{bmatrix}1\\-6\\7\end{bmatrix}+w\begin{bmatrix}1\\0\\2\end{bmatrix}=\begin{bmatrix}5\\-2\\9\end{bmatrix}$$

$$Ax=\begin{bmatrix}2&1&1\\4&-6&0\\-2&7&2\end{bmatrix}\begin{bmatrix}u\\v\\w\end{bmatrix}=\begin{bmatrix}5\\-2\\9\end{bmatrix}=b$$

- There are two ways to look at that system.

1. **Row picture:** Intersection of n planes. The first approach concentrates on the separate equations on the rows. Each equation describes a plane in three dimensions.
2. **Column picture:** The right side b is a combination of the column vectors. The second approach looks at the columns of the linear system. The three separate equations are





Linear Systems

Linear system of algebraic equations

$$a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1$$
$$a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2$$
$$\ldots\ldots\ldots\ldots$$
$$a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=b_n$$

... where the x_1, x_2, ... , x_n are the unknowns ...
in matrix form

$$\mathbf{Ax}=\mathbf{b}$$

Matrix Algebra – Linear Systems

$$\boxed{\mathbf{Ax} = \mathbf{b}}$$ where

$$\mathbf{A} = [\mathrm{a_{ij}}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{11} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \qquad \mathbf{x} = \{x_i\} = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix}$$

$$\mathbf{b} = \{b_i\} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{Bmatrix}$$

A is a *nxn* (square) matrix, and **x** and **b** are column vectors of dimension *n*

Matrix Algebra – Vectors

Row vectors

$$\mathbf{v} = [v_1 \quad v_2 \quad v_3]$$

Column vectors

$$\mathbf{w} = \begin{Bmatrix} w_1 \\ w_2 \\ w_3 \end{Bmatrix}$$

Matrix addition and subtraction

$$\mathbf{C} = \mathbf{A} + \mathbf{B} \quad \text{with} \quad c_{ij} = a_{ij} + b_{ij}$$

$$\mathbf{D} = \mathbf{A} - \mathbf{B} \quad \text{with} \quad d_{ij} = a_{ij} - b_{ij}$$

Matrix multiplication

$$\mathbf{C} = \mathbf{AB} \quad \text{with} \quad c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$$

where **A** (size *lxm*) and **B** (size *mxn*) and *i=1,2,…,l* and *j=1,2,…,n*.
Note that in general AB≠BA but (AB)C=A(BC)

Matrix Algebra – Special

Transpose of a matrix

$$\mathbf{A} = [a_{ij}] \qquad \mathbf{A}^{\mathsf{T}} = [a_{ji}]$$

$$(\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$$

Symmetric matrix

$$\mathbf{A} = \mathbf{A}^{\mathsf{T}}$$

$$a_{ij} = a_{ji}$$

Identity matrix

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

with **AI=A, Ix=x**

Introduction

- Matrix Notation and Matrix Multiplication
  - If $A \in \Re^{m\times p}$ and $B \in \Re^{p\times n}$, then the product $C = AB$ is defined by

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

$$C = AB = \begin{bmatrix} A(1,:) \\ A(2,:) \\ \vdots \\ A(m,:) \end{bmatrix} \begin{bmatrix} | & | & & | \\ B(:,1) & B(:,2) & \cdots & B(:,n) \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ A(:,1) & A(:,2) & \cdots & A(:,p) \\ | & | & & | \end{bmatrix} \begin{bmatrix} B(1,:) \\ B(2,:) \\ \vdots \\ B(p,:) \end{bmatrix} = \sum_{k=1}^{p} A(:,k)B(k,:)$$

LU Decomposition

Gaussian Elimination

- Elementary matrix: The matrix that leaves every vector unchanged is the identity matrix I, with 1's on the diagonal and 0's everywhere else. The matrix that subtracts a multiple l of row from row j is the elementary matrix i with 1's on the diagonal and the number $E_{ij}(l)$ in row i, column j.

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad and \quad E_{31}(l) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -l & 0 & 1 \end{bmatrix}$$

- The inverse matrix of $E_{ij}(l)$ is $E_{ij}(-l)$
- **Forward Elimination:**

$$\begin{aligned} 2u + v + w &= 5 \\ 4u - 6v &= -2 \\ -2u + 7v + 2w &= 9 \end{aligned} \quad \Rightarrow \begin{bmatrix} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & 7 & 2 & 9 \end{bmatrix} = [A;b]$$

  - subtract 2 times the first equation from the second.

LU Decomposition

Gaussian Elimination

- Elementary matrix: The matrix that leaves every vector unchanged is the identity matrix I, with 1's on the diagonal and 0's everywhere else. The matrix that subtracts a multiple l of row from row j is the elementary matrix i with 1's on the diagonal and the number $E_{ij}(l)$ in row i, column j.

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad and \quad E_{31}(l) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -l & 0 & 1 \end{bmatrix}$$

  - The inverse matrix of $E_{ij}(l)$ is $E_{ij}(-l)$

Forward Elimination:

$$\begin{aligned} 2u + v + w &= 5 \\ 4u - 6v &= -2 \\ -2u + 7v + 2w &= 9 \end{aligned} \quad \Rightarrow \begin{bmatrix} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & 7 & 2 & 9 \end{bmatrix} = [A;b]$$

$$\begin{aligned} 2u + v + w &= 5 \\ -8v - 2w &= -12 \\ -2u + 7v + 2w &= 9 \end{aligned} \xrightarrow{R_2 - (2)R_1} \begin{bmatrix} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ -2 & 7 & 2 & 9 \end{bmatrix} = E_{21}(2)[A;b]$$

LU Decomposition

- Forward Elimination:

$$\begin{aligned} 2u+v+w&=5 \\ -8v-2w&=-12 \\ 8v+3w&=14 \end{aligned} \quad \xleftarrow{R_3-(-1)R_1} \quad \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 8 & 3 & -14 \end{array}\right] = E_{31}(-1)E_{21}(2)[A;b]$$

equivalent but simpler system, with an upper triangular matrix U

$$\begin{aligned} 2u+v+w&=5 \\ -8v-2w&=-12 \\ w&=2 \end{aligned} \quad \xleftarrow{R_3-(-1)R_2} \quad \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 0 & 1 & 2 \end{array}\right] = E_{32}(-1)E_{31}(-1)E_{21}(2)[A;b] = [L^{-1}A;L^{-1}b] = [U;y]$$

$$L = (E_{32}(-1)E_{31}(-1)E_{21}(2))^{-1} = E_{21}(-2)E_{31}(1)E_{32}(1) = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix}$$

$$Ax = \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix} = b \Rightarrow Ux = \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 5 \\ -12 \\ 2 \end{bmatrix} = y$$

$$\boxed{A = LU}$$

Forward elimination & Back-substitution

- LU decomposition: $LU\mathbf{x} = \mathbf{b} \;\Rightarrow\; U\mathbf{x} = \mathbf{y},\;\; L\mathbf{y} = \mathbf{b}$

- Forward elimination

$$\begin{aligned} 2u+v+w&=5 & w&=2 \\ -8v-2w&=-12 \Rightarrow & v&=(-12+2w)/(-8)=1 \\ w&=2 & u&=(5-v-w)/2=1 \end{aligned}$$

$$Ly = b \Rightarrow \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{m1} & l_{m2} & \cdots & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = b \qquad y_j = (c_j - \sum_{i=1}^{j-1} l_{ji}y_i)$$

- Back-substitution:

$$U\mathbf{x} = \mathbf{y} \Rightarrow \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ & u_{22} & \cdots & u_{2m} \\ & & \ddots & \vdots \\ & & & u_{mm} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \mathbf{y}. \qquad x_j = \frac{1}{u_{jj}}(y_i - \sum_{k=j+1}^{n} u_{jk}x_k)$$

- The total cost of Forward elimination & Back-substitution $\sim m^2 + m^2 = 2m^2$ flops.

ALGORITHM: Solve Ax = b by using A= LU.

- MATLAB code

```
function x = plusol(A, b)
[L, U, P] = lu(A);
[m, n] = size(A);
% Forward elimination to solve L*y = Pb=c.
y = zeros(m, 1);c=P*b;
for j = 1:m
y(j) = c(j) - L(j,1:j-1)*y(1:j-1);
end
% Back substitution to solve U*x = y.
x = zeros(n, 1);
for j= n:-1:1
x(j) = (y(j) - U(j,j+1:n)*x(j+1:n))/U(j, j);
end
```


LU Decomposition

- Theorem (Triangular factorization A = LU)
  - If no row exchanges are required, the original matrix A can be uniquely written as a product A = LU. The matrix L is lower triangular, with 1's on the diagonal and the multipliers (taken from elimination) below the diagonal. U is the upper trianglar matrix which appears after forward elimination and before back-substitution; its diagonal entries are the pivots.
  - The cost of elimination for Gaussian elimination $\sim \frac{2}{3}m^3$ flops.
- Gaussian Elimination with Pivoting

$$A=\begin{bmatrix}\delta & 1\\ 1 & 1\end{bmatrix}=\begin{bmatrix}1 & 0\\ \frac{1}{\delta} & 1\end{bmatrix}\begin{bmatrix}\delta & 1\\ 0 & 1-\frac{1}{\delta}\end{bmatrix}=LU \xrightarrow{\delta<\varepsilon_{machine}\approx 10^{-16}} \tilde{L}=\begin{bmatrix}1 & 0\\ \frac{1}{\delta} & 1\end{bmatrix},\tilde{U}=\begin{bmatrix}\delta & 1\\ 0 & -\frac{1}{\delta}\end{bmatrix}\Rightarrow \tilde{L}\tilde{U}=\begin{bmatrix}\delta & 1\\ 1 & 0\end{bmatrix}$$

$$b=(1,0)^T \qquad Ax=b \qquad \tilde{L}\tilde{U}x=b$$

$$x=(-1,1)^T \qquad \tilde{x}=(0,1)^T$$

  - Such instability can be controlled by permuting the order of the rows of the matrix being operated on, a strategy called pivoting.

LU Decomposition

Theorem 2 (PA = LU)

- To any m x m matrix A, there correspond a permutation matrix P, a lower triangular matrix L with unit diagonal, and an m x m upper triangular matrix U, such that PA = LU.
  - The purpose of pivoting is to make Gaussian elimination applicable to all matrices and stable.

$$L = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{m1} & l_{m2} & \cdots & 1 \end{bmatrix}, \quad U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ & u_{22} & \cdots & u_{2m} \\ & & \ddots & \vdots \\ & & & u_{mm} \end{bmatrix}$$

- Using the LU decomposition, Ax = b can be solved as follows.

$$LU\mathbf{x} = P\mathbf{b} \quad \Rightarrow \quad U\mathbf{x} = \mathbf{y}, \quad L\mathbf{y} = P\mathbf{b} = \mathbf{c}$$

LU Decomposition

Point of view of stability

- pivoting typically ensure that $\| L \|$ is of order 1 and $\| U \|$ is of the order of $\| A \|$. However, for certain matrices A, $\| U \| / \| A \|$ turns out to be huge.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

- The pattern can be continued to matrices of arbitrary dimension m with $u_{mm} = 2^{m-1}$. Yet despite examples like this, partial pivoting is resoundingly stable in practice. If you pick a billion matrices A at random, you will almost certainly not see behavior like this.

LU Decomposition

Inverses and Gauss-Jordan Method

- Never explicitly invert a matrix numerically.

$$[A\,|\,e_1\,|\,e_2\,|\,e_3] = \begin{bmatrix} 2 & 1 & 1 & 1 & 0 & 0 \\ 4 & -6 & 0 & 0 & 1 & 0 \\ -2 & 7 & 2 & 0 & 0 & 1 \end{bmatrix} \to \begin{bmatrix} 2 & 1 & 1 & 1 & 0 & 0 \\ 0 & -8 & -2 & -2 & 1 & 0 \\ 0 & 8 & 3 & 1 & 0 & 1 \end{bmatrix}$$

$$\to \begin{bmatrix} 2 & 1 & 1 & 1 & 0 & 0 \\ 0 & -8 & -2 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{bmatrix} = [\,U\ L^{-1}\,]$$

$$\to \begin{bmatrix} 2 & 1 & 0 & 2 & -1 & -1 \\ 0 & -8 & 0 & -4 & 3 & 2 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{bmatrix} \to \begin{bmatrix} 2 & 0 & 0 & 12/8 & -5/8 & -6/8 \\ 0 & -8 & 0 & -4 & 3 & 2 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{bmatrix}$$

$$\to \begin{bmatrix} 1 & 0 & 0 & 12/16 & -5/16 & -6/16 \\ 0 & 1 & 0 & 4/8 & -3/8 & -2/8 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{bmatrix} = [\,I\ A^{-1}\,]$$

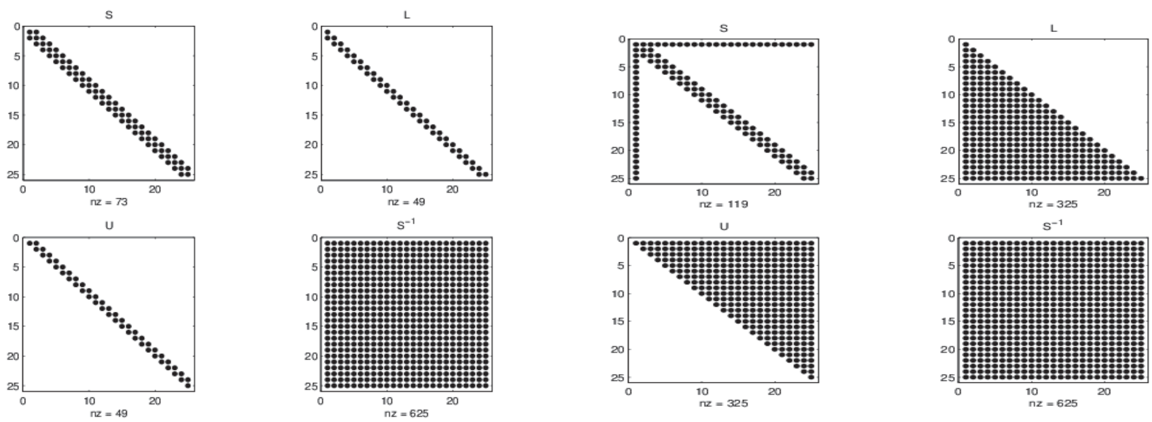
- The final operation count for computing A^{-1} $\sim \frac{n^3}{6} + \frac{n^3}{3} + n(\frac{n^2}{2}) = n^3$

Sparse Matrix

Sparse

- For many matrices that arise in practice, the ratio is small

$$\frac{\text{Number of Nonzero entries}}{\text{Number of Zero entries}}$$



Summary for LU decomposition

- LU decomposition is often very efficient to deal with sparse matrices.
- LU decomposition provides a way to derive a matrix inversion formula

Reading

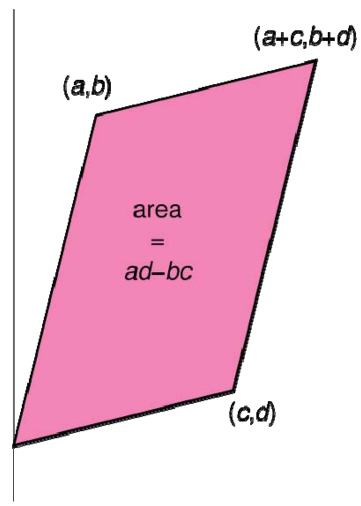
- [Strang. (2006), Chapter 4 Determinants, Chapter 6 Positive Definite Matrices]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.

Determinant of a Matrix

- The determinant gives an idea of the 'volume' occupied by the matrix in vector space
- Used for inversion
- A matrix A has an inverse matrix A^{-1} if and only if **det(A)≠0.**

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \det(A) = ad - bc$$

$$A^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$



Determinant of a Matrix

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - afh - bdi - ceg$$

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}$$

Sum from left to right
Subtract from right to left
Note: N! terms

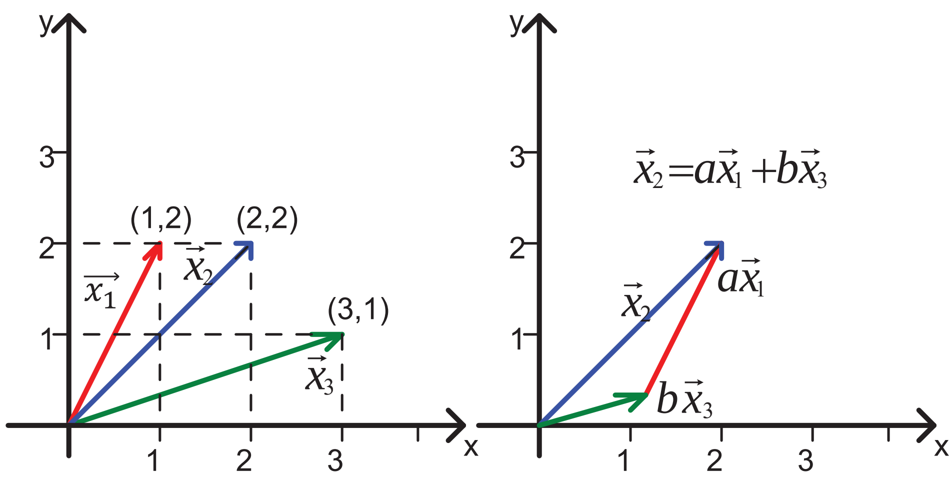
Example: Determinants

- Determinants
  - In a vectorial space of n dimensions, there will be no more than n linearly independent vectors.
  - If 3 vectors (2×1) $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are represented by a matrix X:
  - Graphically, we have:

Here x3 can be expressed by a *linear combination* of x1 and x2.

The *determinant* of the matrix X' will thus be zero.

The largest square sub-matrix with a non-zero determinant will be a matrix of 2x2 => the *rank* of the matrix is 2.



Determinants

Determinant

- The determinant of A is a combination of row i and the cofactors of row i.

$$\det(A) = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in} \qquad A_{ij} = (-1)^{i+j}\det(M_{ij})$$

where M_{ij} is formed by deleting row i and column j of A.

Properties of Determinant

- If two rows of A are equal, then det(A) = 0.
- The elementary operation of subtracting a multiple of one row from another row leaves the determinant unchanged.
- If A has a zero row or zero column, then det(A) = 0.
- If A is triangular, then $\det(\mathrm{A}) = a_{11}a_{22}\cdots a_{nn}$. In particular, $\det(I_n) = 1$.
- If A is singular, then det(A) = 0. If A is invertible, then $\det(A) \neq 0$.
- $\det(AB) = \det(A)\det(B)$; $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
- $\det(A^T) = \det(A)$.
- $\det(\mathbf{I}_N + \mathbf{A}\mathbf{B}^T) = \det(\mathbf{I}_M + \mathbf{A}^T\mathbf{B}) \qquad where \quad \mathbf{A}, \mathbf{B} \in \Re^{N\times M}$

Determinants

Formulas of the determinant

- If A is nonsingular, then $A = P^{-1}LU$, and

$$\det(A) = \det(P^{-1}LU) = \pm(\textit{products of the pivots})$$

where ± 1 is the determinant of P and depends on whether the number of row exchanges is even or odd. The triangular factors have

$$\det L = 1 \quad and \quad \det U = d_1 \ldots d_n$$

- If A is invertible, then $\det(A) \neq 0$ and $A^{-1} = \dfrac{\mathrm{adj}(A)}{\det(A)}$
  - Sketch of Proof:

$$A\cdot\mathrm{adj}(A) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} = \begin{bmatrix} \det(A) & 0 & \cdots & 0 \\ 0 & \det(A) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \det(A) \end{bmatrix} = \det(A)I$$

To prove that we get zeros everywhere off the diagonal, let B be the same as A except in row j 6= i where the i-th row is copied into the j-th row of B. Then

$$\det(B) = 0 = a_{i1}A_{j1} + a_{i2}A_{j2} + \cdots + a_{in}A_{jn}, \forall i \neq j$$

Determinants

Applications of Determinant

- It gives a test for invertibility. If the determinant of A is zero, then A is singular. If $\det(A) \neq 0$, then A is invertible.
- The determinant of A equals the volume of a parallelepiped P in n-dimensional space, provided the edges of P come from the rows of A.
- The determinant gives formula for the pivots. From the formula $determinant = \pm\,(products\ of\ the\ pivots)$, it follows that regardless of the order of elimination, the product of the pivots remains the same apart from sign.
- (Cramer's rule) The determinant measures the dependence of $A^{-1}b$ on each element of b. If one parameter is changed in an experiment, or one observation is corrected, the "influence coefficient" on $x = A^{-1}b$ is a ratio of determinants.

$$x_j = \frac{\det(B_j)}{\det(A)},\ where\ B_j = \begin{bmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ a_{21} & \cdots & b_2 & \cdots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & b_n & \cdots & a_{nn} \end{bmatrix}$$

Positive definite matrix

Symmetric positive definite matrix

- A matrix $A = (a_{ij}) \in \Re^{n \times n}$ is symmetric if A = A^T
- $A = (a_{ij}) \in \Re^{n \times n}$ is positive definite (or positive semi-definite) if $x^T A x > 0$ (or $x^T A x \geq 0$) for all nonzero $x \in \Re^n$, denoted by $A \succ 0$ (or A $\succsim 0$)
- If $C \in \Re^{n \times k}$ has full rank and A = C^{T}C, then A is SPD.

$$x^T A x = x^T C^T C x = \|Cx\|^2 > 0$$

- Correlation matrix is SPD.

Tests for Positive Definiteness

- All the eigenvalues of A satisfy $\lambda_i > 0$.
- All the upper left submatrices A_k have positive determinants.
- 2×2-matrix

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} \quad \text{is positive definite when} \quad a > 0 \quad \text{and} \quad ac - b^2 > 0.$$

Examples

- Example 1:

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \quad \text{is positive } semi\text{definite,}$$

(I′) $x^{\mathrm{T}}Ax = (x_1 - x_2)^2 + (x_1 - x_3)^2 + (x_2 - x_3)^2 \geq 0$ (zero if $x_1 = x_2 = x_3$).

(II′) The eigenvalues are $\lambda_1 = 0$, $\lambda_2 = \lambda_3 = 3$ (a zero eigenvalue).

(III′) $\det A = 0$ and smaller determinants are positive.

- Example 2:
  - For what range of numbers a and b are the matrices A and B positive definite?

$$A = \begin{bmatrix} a & 2 & 2 \\ 2 & a & 2 \\ 2 & 2 & a \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 2 & 4 \\ 2 & b & 8 \\ 4 & 8 & 7 \end{bmatrix}$$

A is positive definite for $a > 2$. B is never positive definite: notice $\begin{bmatrix} 1 & 4 \\ 4 & 7 \end{bmatrix}$

The Cholesky factorization

- Theorem 4 (Cholesky factorization)
  - Every symmetric positive definite matrix $A = (a_{ij}) \in \Re^{n \times n}$ has a unique Cholesky factorization

$$A = R^T R, \qquad r_{ii} > 0$$

where $R = (r_{ij})$ is an n ×n upper-triangular matrix with positive diagonal entries.

$$A = \begin{bmatrix} a_{11} & v^T \\ v & K \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha & 0 \\ v/\alpha & I \end{bmatrix}}_{R_1^T} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & K - vv^T/a_{11} \end{bmatrix}}_{A_1} \underbrace{\begin{bmatrix} \alpha & v^T/\alpha \\ 0 & I \end{bmatrix}}_{R_1} \Rightarrow A = \underbrace{R_1^T R_2^T \cdots R_n^T}_{R^T} \underbrace{R_n \cdots R_2 R_1}_{R}$$

$\alpha = \sqrt{a_{11}}$

- Cholesky factorization
  - The system of equations Ax = b where A is SPD can be solved using the Cholesky factorization A = R^TR via

$$R^T y = b, \qquad Rx = y$$

  - Least squares solutions of $\min\|Ax - b\|$ can be solved using the Cholesky factorization A^TA = R^TR via $(A^T A)x = A^T b \;\Rightarrow R^T y = A^T b, \qquad Rx = y$

Matrix Inversion Formula

- Schur Complement

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \Rightarrow \quad A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}$$

where $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is known as the **Schur complement** of A_{11}.

- Sherman-Woodbury-Morrison identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

Proof. (1) First verify the formula

$$\begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

for "elimination" of the block A_{21}. Then use the Gaussian elimination again.

(2) $(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})x = b$ has the same solution as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & -\mathbf{D} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

Then use the Shur complement.

Determinant Formula

- Trace and determinant of products:

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$$

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$$

- $A \in \Re^{n \times n}$, $n \times n$ matrix

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}, \qquad \frac{\partial}{\partial x}\ln|\mathbf{A}| = \mathrm{tr}(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}),$$

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \mathbf{B}, \qquad \frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = (\mathbf{A}^{-1})^T$$

Chapter 2. Vectors and Norms

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Maxim

- "There is nothing more practical than a good theory."

  \- James C. Maxwell

Reading

- [Strang. (2006), Chapter 2 Vector Spaces]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.

Linear Algebra

- Vector Spaces
  - A vector space V over a set of scalars R is a collection of objects known as "vectors", together with an additive operation + and a scalar multiplication operation, that satisfy the following properties: for all $x, y, z \in V$ and $h, k \in R$

    A1. $x + y \in V$.

    A2. $x + y = y + x$.

    A3. $(x + y) + z = x + (y + z)$.

    A4. There is an element $0 \in V$, such that $x + 0 = 0 + x = x$ for all $x \in V$.

    A5. For each $x \in V$, there is an element $-x \in V$ such that $x + (-x) = (-x) + x = 0$.

    A6. $kx \in V$.

    A7. $k(x + y) = kx + ky$.

    A8. $(h + k)x = hx + kx$.

    A9. $(hk)x = h(kx) = k(hx)$.

    A10. $1x = x$.
  - Example:
    - the Euclidean space R^n, the infinite dimensional space R^∞,
    - the space of m×n matrices,
    - the space of real valued functions $f: [a; b] \to R$.

Subspaces

- Subspaces
  - Let $S \in V$ be a subset of a vector space V such that S is itself a vector space. Then V is said to be subspace of V.
  - A subspace S of a vector space is a nonempty subset that satisfies two requirements:

    I. If $x, y \in S$, then $x + y \in S$ and

    II. $kx \in S$.

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $v_1,\ldots,v_k$ are linearly independent if $c_1v_1+\ldots+c_kv_k = 0$ implies $c_1=\ldots=c_k=0$

$$\begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

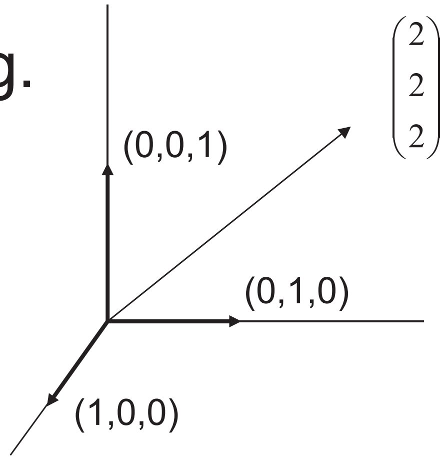
e.g. $\begin{pmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ (u,v)=(0,0), i.e. the columns are linearly in dependent.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix} \qquad x_3 = -2x_1 + x_2$$

Span of a vector space

- If all vectors in a vector space may be expressed as linear combinations of a set of vectors $v_1,\ldots,v_k$, then $v_1,\ldots,v_k$ spans the space.
- The cardinality of this set is the dimension of the vector space.

e.g.



$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 2\begin{pmatrix}1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\end{pmatrix}$$

- A basis is a maximal set of linearly independent vectors and a minimal set of spanning vectors of a vector space

Rank of a Matrix

- rank(A) (the rank of a m-by-n matrix A) is

  The maximal number of linearly independent columns

  =The maximal number of linearly independent rows

  =The dimension of col(A)

  =The dimension of row(A)

- If A is n by m, then
  - rank(A)<= min(m,n)
  - If n=rank(A), then A has full row rank
  - If m=rank(A), then A has full column rank

$$\begin{pmatrix}1 & 1\\0 & 1\end{pmatrix} \quad \begin{pmatrix}2 & 1\\4 & 2\end{pmatrix}$$

Example: Linear dependency and rank

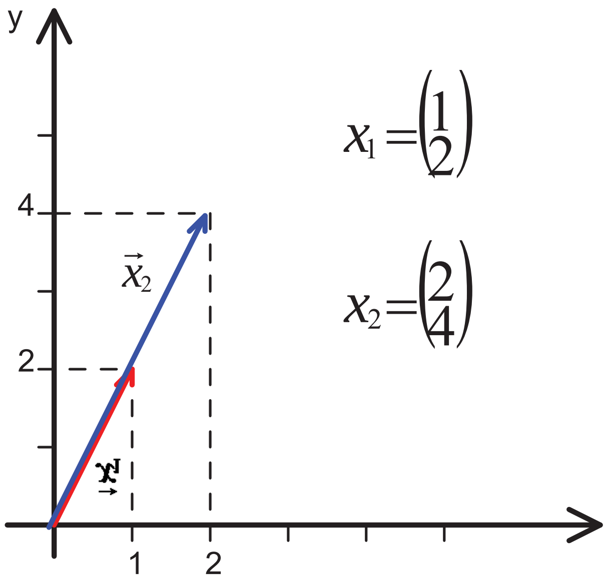
Linear dependency and rank

- If one can find a *linear relationship* between the lines or columns of a matrix, then the *rank* of the matrix (number of dimensions of its vectorial space) will not be equal to its number of column/lines – the matrix will be said to be *rank-deficient*.
- Example

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

When representing the vectors, we see that x1 and x2 are superimposed. If we look better, we see that we can express one by a *linear combination* of the other: $x_2 = 2x_1$.

The *rank* of the matrix will be 1.
In parallel, the *vectorial space* defined will has only one dimension.



$$x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$x_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

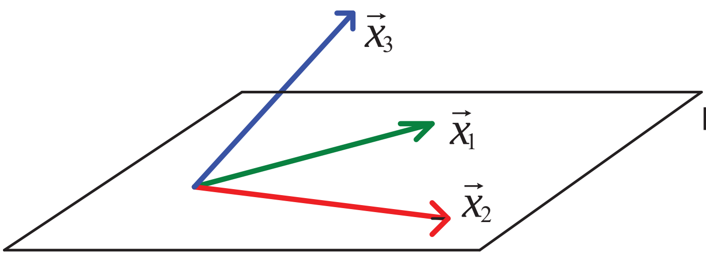
Example: Linear dependency and rank

Linear dependency and rank

- The *rank of a matrix* corresponds to the *dimensionality* of the vectorial space defined by this matrix. It corresponds to the number of vectors defined by the matrix that are linearly independents from each other.
- *Linealy independent* vectors are vectors defining each one one more dimension in space, compared to the space defined by the other vectors. They cannot be expressed by a linear combination of the others.
- Note. *Linearly independent* vectors are not necessarily *orthogonal* (perpendicular).

Example: take 3 linearly independent vectors x_1, x_2, x_3.

- Vectors x_1, x_2 define a plane (x,y) and vector x_3 has an additional non-zero component in the z axis. But x_3 is not perpendicular to $x_1, or\ x_2$.



Linear dependency and basis

- Definition
  - The vectors $v_1, v_2, \dots, v_k$ are linearly independent if $c_1 v_1 + \cdots + c_k v_k = 0$ only happens when $c_1 = \cdots = c_k = 0$. Otherwise, they are linearly dependent, and one of them is a linear combination of the others.
  - The vectors $w_1, w_2, \dots, w_k$ span the vector space V if for every vector $v \in V$, $v = c_1 v_1 + \cdots + c_k v_k$ for some coefficients c_i.
  - The set of vectors $v_1, v_2, \dots, v_k$ is a basis of V if (i) it is linearly independent and (ii) it spans the space V .
  - The dimension of a vector space V is the number of vectors in a basis of V .

Vector Products

Two vectors:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Inner product = scalar

Inner product X^TY is a scalar

(1xn) (nx1)

$$\mathbf{x}^T\mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 = \sum_{i=1}^{3} x_i y_i$$

Outer product = matrix

$$xy^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{bmatrix}$$

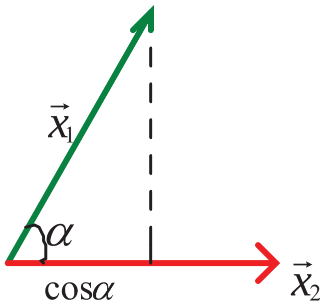
Outer product XY^T is a matrix

(nx1) (1xn)

Inner product of vectors

- Inner product of vectors

Calculate the *scalar product* of two vectors is equivqlent to make the *projection* of one vector on the other one. One can indeed show that $\vec{x}_1 \bullet \vec{x}_2 = |\vec{x}_1| \,.\, |\vec{x}_2| \,.\, \cos\alpha$ *where* α is the angle that separates two vectors when they have both the same origin.



$$x_1 \bullet x_2 = |\vec{x}_1| \,.\, |\vec{x}_2| \,.\, \cos\alpha$$

$$|\langle x_1, x_2 \rangle| \le \| x_1 \| \cdot \| x_2 \|$$ Cauchy-Schwarz Inequality

In parallel, if two vectors are orthogonal, their scalar product is zero: the projection of one onto the other will be zero.

Inner products

- Inner Product Space
  - A vector space X over the reals $\mathbb{R}$ is an inner product space if there exists a real-valued symmetric bilinear map $<.,.>$ that satisfies

$$\langle x, x \rangle \ge 0$$

  - This bilinear map is known as the inner, dot or scalar product and we will say the inner product is strict if

$$\langle x, x \rangle = 0 \text{ if and only if } \quad x = 0$$

- A Hilbert Space
  - A **Hilbert Space** $\mathcal{F}$ is a strict inner product space with the additional properties that is is separable and complete. Completeness refers to the property that every Cauchy sequence $\{h_n\}_{n\ge1}$ of elements of $\mathcal{F}$ converges to a element $h \in \mathcal{F}$, where a Cauchy sequence is one satisfying the property that

$$\sup_{m>n} ||h_n - h_m|| \to 0, \text{ as } \quad n \to \infty$$

Inner products

- Example (Hilbert Space)
  - **[ℓ^2 Space]** Let X be the set of all countable sequences of real numbers $x = (x_1, x_2, \dots, x_n, \dots)$, such that the sum $\sum_{i=1}^{\infty} x_i^2 < \infty$ with the inner product between two sequences x and y defined by

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i .$$

  - **[Inner Product in Function Space]** Let $\mathcal{F} = L^2(X)$ be the vector space of square integrable functions on a compact subset X of $\mathbb{R}^n$ with the obvious definitions of addition and scalar multiplication, that is $L^2(X) = \{f: \int_X f(x)^2 dx < \infty\}$. For $f, g \in X$, define the inner product by

$$\langle f, g \rangle = \int_X f(x) g(x) dx.$$

- Cauchy-Schwarz Inequality
  - In an inner product space

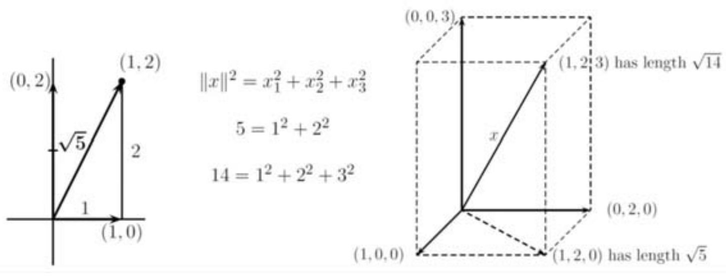
$$|\langle x, z \rangle| \le \| x \| \cdot \| z \| .$$

and the equality sign holds in a strict inner product space if and only if x and z are rescalings of the same vector.

Length

- The length of $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$

$\|\boldsymbol{x}\|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 = \boldsymbol{x}^T \boldsymbol{x}$

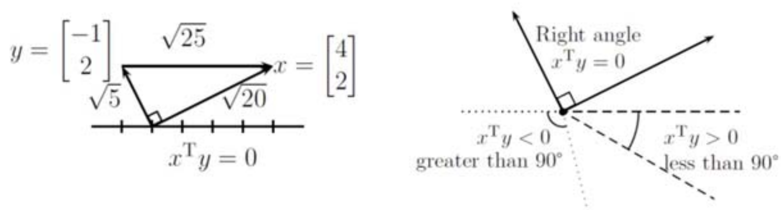


- Inner product of $\boldsymbol{x} = (x_1, \dots, x_n)$ and $\boldsymbol{y} = (y_1, \dots, y_n)$

$\boldsymbol{x}^T \boldsymbol{y} = x_1 y_1 + \cdots + x_n y_n$

- Orthogonal vectors

$\boldsymbol{x}^T \boldsymbol{y} = x_1 y_1 + \cdots + x_n y_n = 0$



Norms

- How can we compare the size of vectors, matrices (and functions!)?
  - For scalars it is easy (absolute value). The generalization of this concept to vectors, matrices and functions is called a norm. Formally the norm is a function from the space of vectors into the space of scalars denoted by

$$\boxed{\|(.)\|}$$

- Definition (Norms)
  - Let S be a vector space with elements x. A real-valued function $\|x\|$ is said to be a norm if $\|x\|$ satisfies the following properties.

  1. $\|x\| \geq 0$ for any $x \in S$.
  2. $\|x\| = 0$ if and only if x = 0.
  3. $\|\alpha x\| = |\alpha|\|x\|$, where α is an arbitrary scalar.
  4. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

Norms

- Vector Norms
  - (Vector p-norm) $$\boxed{\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}}$$
  - Examples:
    - for $p = 2$ we have the ordinary euclidian norm: $\|x\|_2 = \sqrt{x^T x}$
    - for $p = \infty$ the definition is $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$

- Matrix Norms
  - (Matrix p-norm) $$\|A\|_p = \sup_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|=1}\|Ax\|_p$$
  - (Frobenius norm) $$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}} = \sqrt{\mathrm{tr}(A^T A)}$$

Norms

- Properties of vector and matrix norms $A \in \Re^{m\times n}$
  - $\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|$ the largest column sum
  - $\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|$ the largest row sum
  - $\exists\, z \in \Re^n$ such that $A^T A z = \mu^2 z$ where $\mu = \|A\|_2$. In particular, $\|A\|_2$ is the square root of the largest eigenvalues of $A^T A$.

- Spectral norm
  - For an $n \times n$ matrix A with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, if the **spectral radius** $\rho(A)$ is defined as $\rho(A) = \max_i |\lambda_i|$, then,

$$\| A \|_2 = \sqrt{\rho(A^T A)}$$

  - In case A is Hermitian,

$$\| A \|_2 = \rho(A)$$

  - The l_2 norm is also called the **spectral norm**.

Four Fundamental Subspaces

- Definition (The four fundamental subspaces)
  - Let A be a m×n matrix.
    - $R(A)$= column space of A consists of all linear combinations of the columns of A.
    - $N(A)$= nullspace of A consists of all vectors x such that $Ax = 0$.
    - $R(A^T)$= row space of A consists of all linear combinations of the rows of A.
    - $N(A^T)$= left nullspace of A consists of all vectors y such that $A^T y = 0$.
  - The dimension of the column space(or the row space) of the matrix A is the rank of the matrix.

- Theorem (Fundamental Theorem of Linear Algebra-I)
  - Let A be a m×n matrix.
    - $R(A) \subset R^m$ and $dim R(A) = r$.
    - $N(A) \subset R^n$ and $dim N(A) = n - r$.
    - $R(A^T) \subset R^n$ and $dim R(A^T) = r$.
    - $N(A^T) \subset R^m$ and $dim N(A^T) = m - r$.

Four Fundamental Subspaces

Orthogonality

- Vectors x and y are said to be **orthogonal** if $x^T y = 0$. Notationally, this is denoted as $x \perp y$.
- For a subset V of an inner product space S, he space of all vectors orthogonal to V is called the orthogonal complement of V . This is denoted as $V^{\perp}$.

Theorem (Fundamental Theorem of Linear Algebra-II)

- Let A be a $m \times n$ matrix.
  - The left nullspace is the orthogonal complement of the column space in R^m, i.e.,

$$N(A^T)^{\perp} = R(A)$$

  - The nullspace is the orthogonal complement of the row space in R^n, i.e.,

$$N(A)^{\perp} = R(A^T)$$

Four Fundamental Subspaces

Theorem (Fredholm alternative theorem)

- Let A be a m×n matrix. The equation $Ax = b$ has a solution if and only if $v^T b = 0$ for every vector v such that $A^T v = 0$. More succinctly

$$b \in R(A) \Leftrightarrow b \perp N(A^T).$$

Application of Fredholm alternative theorem

- A typical vector x has a "row space component" and a "nullspace component",

$$x = x_r + x_n, \qquad x_r \in R(A^T), x_n \in N(A)$$

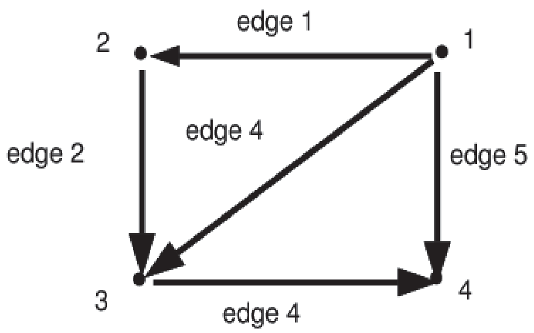
  where the nullspace component goes to zero: $Ax_n = 0$ and the row space component goes to the column space: $Ax = Ax_r$.
- The mapping from row space to column space is actually invertible.

$$R(A^T) \Leftrightarrow R(A)$$

- Every vector b in the column space comes from one and only one vector xr in the row space. Therefore, the solution to Ax = b (if it exists) is unique if and only if the only solution of Ax = 0 is x = 0, that is, if $N(A) = \{0\}$.

Example: Graph

- Example: Graph from section 2.5 in [Strang (2006)]
  - A graph has two ingredients: a set of vertices or “nodes", and a set of arcs or “edges" that connect them.
  - This graph introduces the edge-node incidence matrix by A where if the edge goes from node j to node k, then that row has -1 in column j and $+1$ in column k. If we think of the components x_1, x_2, x_3, x_4 as the potentials at the nodes, then the vector Ax gives the potential differences.



$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$

Example: Graph

- Fundamental Subspaces
  - Nullspace: dimension 1 and since the columns add up to the zero column

$$\mathcal{N}(A) = \text{span}\{(1,1,1,1)^T\}$$

  - Column space: dimension 3 and the test for $b = Ax$ to be in the column space is Kirchhoff’s Voltage Law: *The sum of potential differences around a loop must be zero.*

$$b_1 + b_2 - b_3 = 0 \quad and \quad b_3 + b_4 - b_5 = 0$$

  - Left nullspace: dimension 2 and since the vectors in the left null-space correspond to loops in the graph. ($y^T b = 0$ for $b \in \mathcal{R}(A)$.)

$$\mathcal{N}(A^T) = \text{span}\{(1,1,-1,0,0)^T, (0,0,1,1,-1)^T\}$$

  - Row space: dimension 3 and the test for $f = A^T y$ to be in the row space is Kirchhoff’s Current Law: *The net current into every node is zero.*
    - This law can only be satisfied if the total current entering the nodes from outside is $f_1 + f_2 + f_3 + f_4 = 0$ since if $f = (f_1, f_2, f_3, f_4)$ is in the row space and x is in the nullspace, then $f^T x = 0$. (Note the numbers f_1, f_2, f_3, f_4 are current sources at the nodes. For example, the source f_1 balance $-y_1 - y_3 - y_5$, which is the flow leaving node 1 along edges 1,3,5.)

m Equations in n Unknowns

The Solution of m Equations in n Unknowns

$$Ax = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 5 \end{bmatrix} = b$$

$$[A;b] = \begin{bmatrix} 1 & 3 & 3 & 2 & 1 \\ 2 & 6 & 9 & 5 & 5 \\ -1 & -3 & 3 & 0 & 5 \end{bmatrix} \xleftarrow{R_2 - (2)R_1} \quad E_{21}(2)[A;b] = \begin{bmatrix} 1 & 3 & 3 & 2 & 1 \\ 0 & 0 & 3 & 1 & 3 \\ -1 & -3 & 3 & 0 & 5 \end{bmatrix}$$

$$\xleftarrow{R_3 - (-1)R_1} \quad E_{31}(-1)E_{21}(2)[A;b] = \begin{bmatrix} 1 & 3 & 3 & 2 & 1 \\ 0 & 0 & 3 & 1 & 3 \\ 0 & 0 & 6 & 2 & 6 \end{bmatrix} \quad \xleftarrow{R_3 - (2)R_2} \quad E_{32}(2)E_{31}(-1)E_{21}(2)[A;b] = \begin{bmatrix} 1 & 3 & 3 & 2 & 1 \\ 0 & 0 & 3 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$Ax = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 5 \end{bmatrix} = b \Rightarrow Ux = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} = c \qquad L = E_{21}(-2)E_{31}(1)E_{32}(2) = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix}$$

m Equations in n Unknowns

The Solution of m Equations in n Unknowns

- The elimination process can continue until the matrix is in a still simpler reduced row echelon form (rref) matrix R

$$U = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow R = \begin{bmatrix} 1 & 3 & 0 & 1 \\ 0 & 0 & 1 & 1/3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} \otimes & * & * & * & * & * & * & * & * \\ 0 & \otimes & * & * & * & * & * & * & * \\ 0 & 0 & 0 & \otimes & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \otimes \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 & * & 0 & * & * & * & * & 0 \\ 0 & 1 & * & 0 & * & * & * & * & 0 \\ 0 & 0 & 0 & 1 & * & * & * & * & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Solution to a rectangular system Ax = b

$$x = \begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \end{bmatrix} + v \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + y \begin{bmatrix} -1 \\ 0 \\ -1/3 \\ 1 \end{bmatrix} = x_{partic} + x_{\text{hom}} \qquad x_{\text{hom}} = v \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + y \begin{bmatrix} -1 \\ 0 \\ -1/3 \\ 1 \end{bmatrix}$$

Solution to a homogeneous system Ax = 0

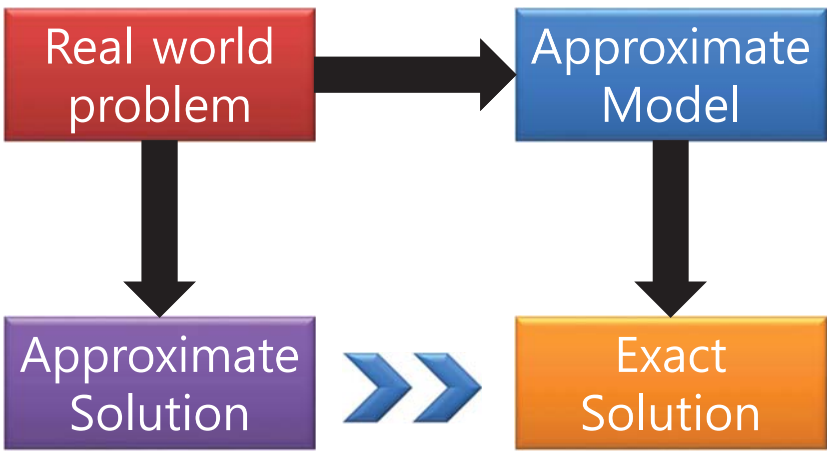
Chapter 3. Least Squares and QR factorization

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Maxim

- An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

---John Tukey



Reading

- [Strang. (2006), Chapter 3 Orthogonality]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.

Matrix operating on vectors

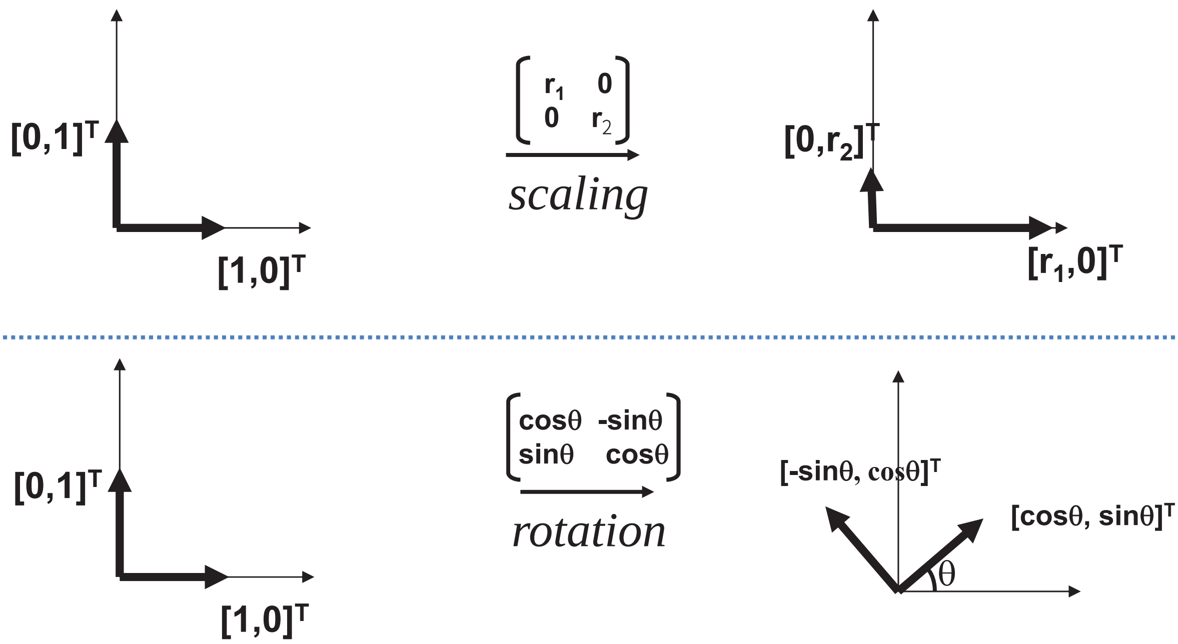
- Matrix is like a <u>function</u> that <u>transforms the vectors on a plane</u>
- Matrix operating on a general point => transforms x- and y-components
- *System of linear equations*: matrix is just the bunch of coeffs !

- $x' = ax + by$
- $y' = cx + dy$

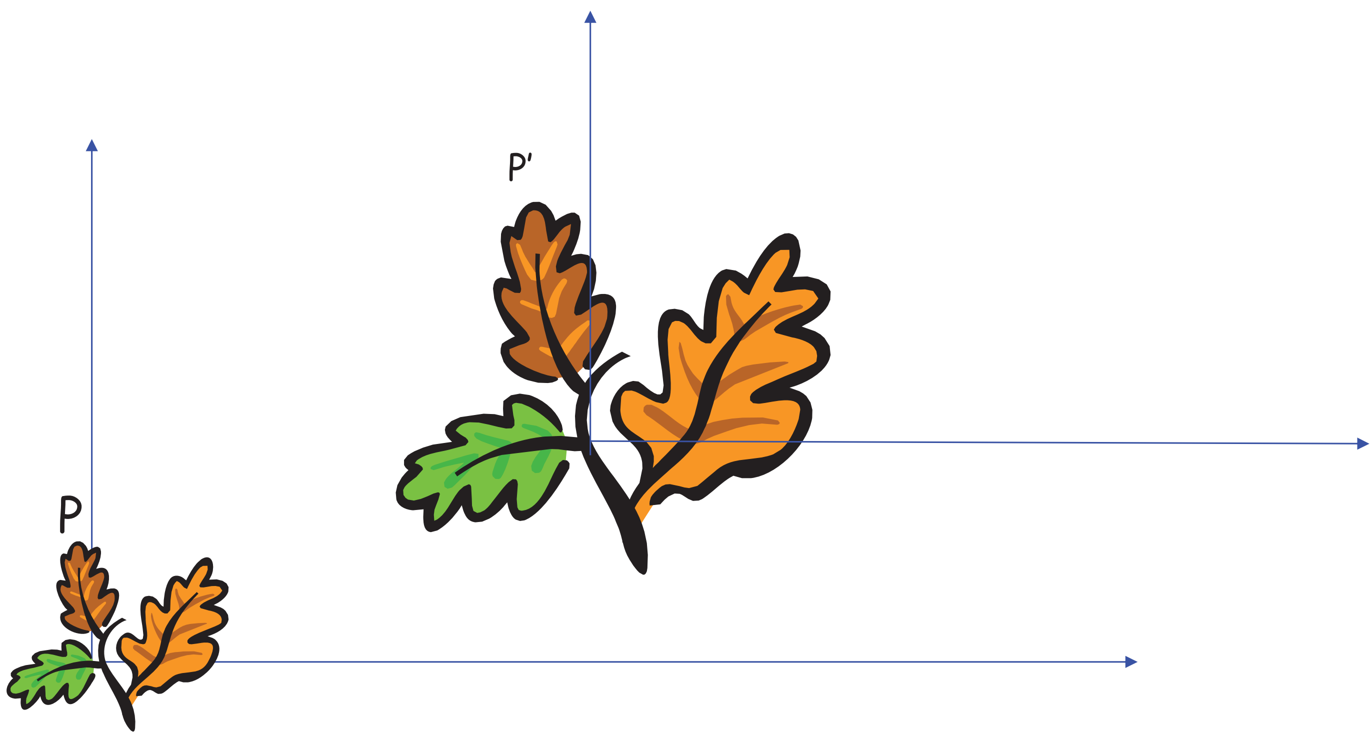
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}$$

Matrices: Scaling, Rotation, Identity

- Pure scaling, no rotation => "**diagonal** matrix" (note: x-, y-axes could be scaled differently!)
- Pure rotation, no stretching => "**orthogonal** matrix" **O**
- **Identity** ("do nothing") matrix = unit scaling, no rotation!



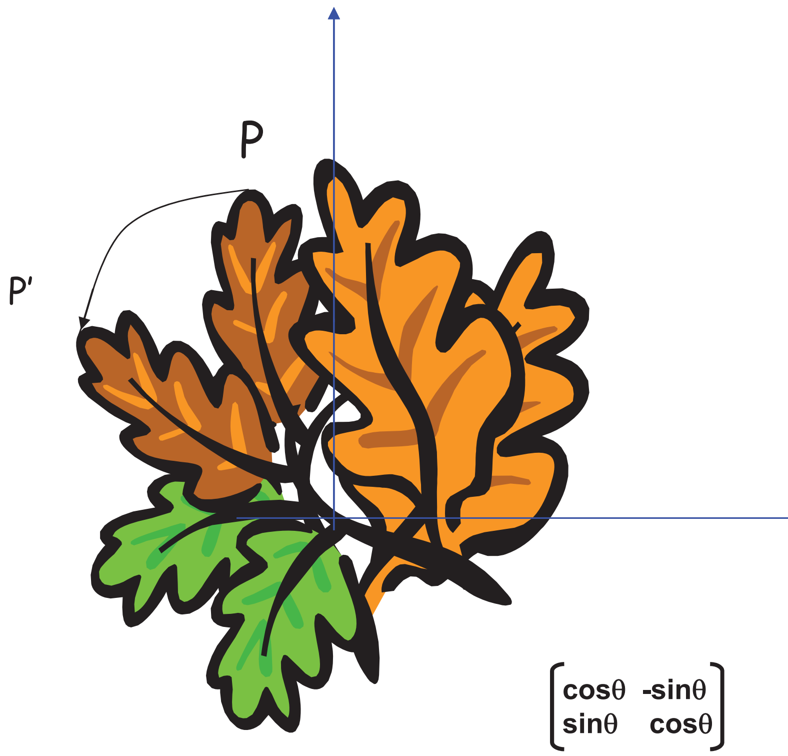
Scaling



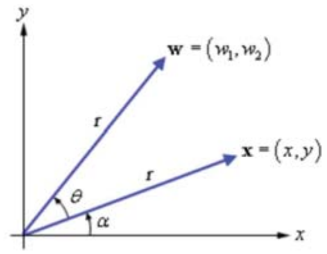
$$\begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}$$

a.k.a: dilation ($r > 1$),
contraction ($r < 1$)

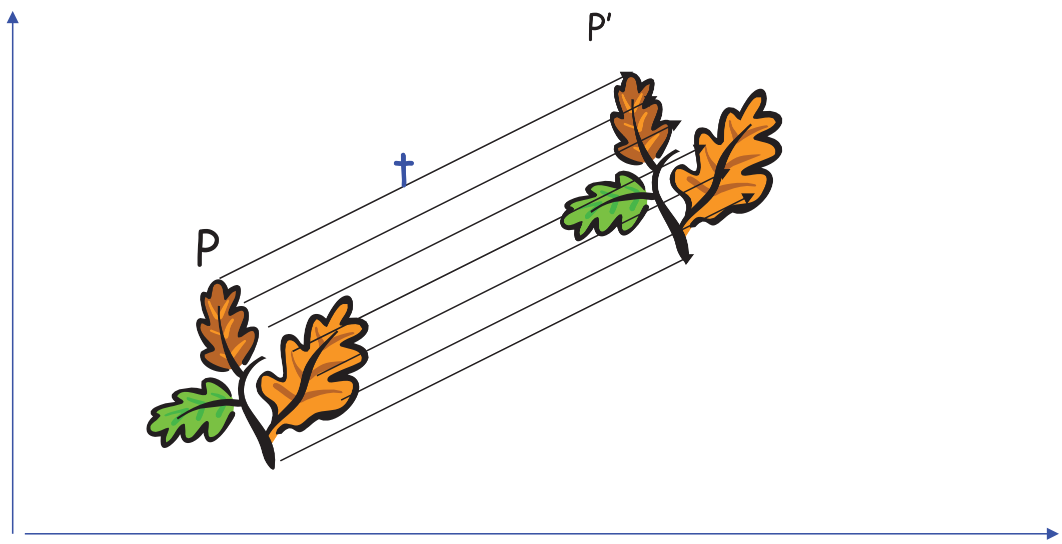
Rotation



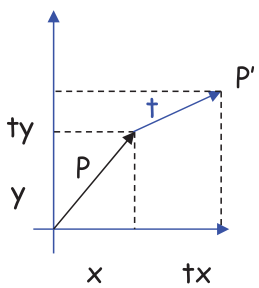
$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$



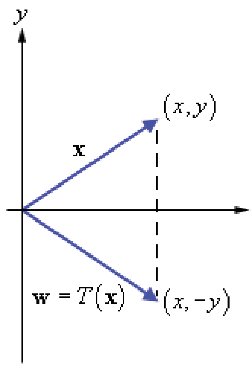
2D Translation



$$\mathbf{P}' = (x + t_x, y + t_y) = \mathbf{P} + \mathbf{t}$$

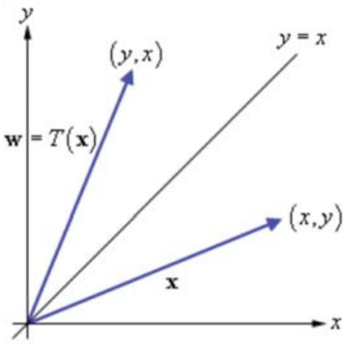


Reflections



- Reflection can be about any line or point.
- Complex Conjugate: reflection about x-axis (i.e. flip the phase θ to -θ)
- Reflection => two times the projection distance from the line.
- Reflection does not affect magnitude

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



$w_1 = y$
$w_2 = x$

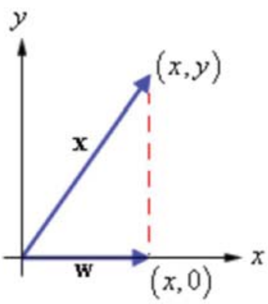
Induced Matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

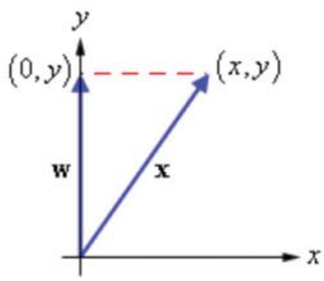
Reflection about x-axis in $\mathbb{R}^2$	$w_1 = x$ $w_2 = -y$	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
Reflection about y-axis in $\mathbb{R}^2$	$w_1 = -x$ $w_2 = y$	$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$
Reflection about line $x = y$ in $\mathbb{R}^2$	$w_1 = y$ $w_2 = x$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
Reflection about origin in $\mathbb{R}^2$	$w_1 = -x$ $w_2 = -y$	$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$

Orthogonal Projections: Matrices

Orthogonal projection on x-axis



Orthogonal projection on y-axis

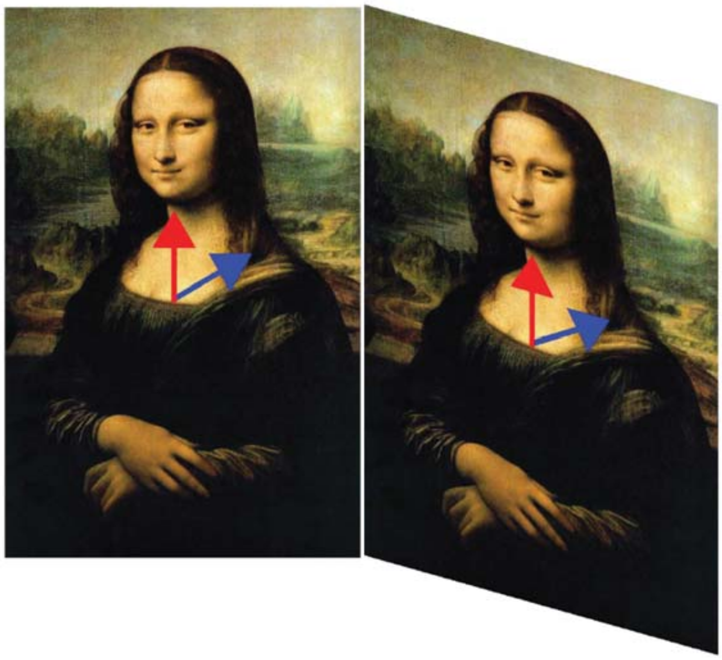


Orthogonal Projection	Equations	Induced Matrix
Projection on x-axis in $\mathbb{R}^2$	$w_1 = x$ $w_2 = 0$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
Projection on y-axis in $\mathbb{R}^2$	$w_1 = 0$ $w_2 = y$	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

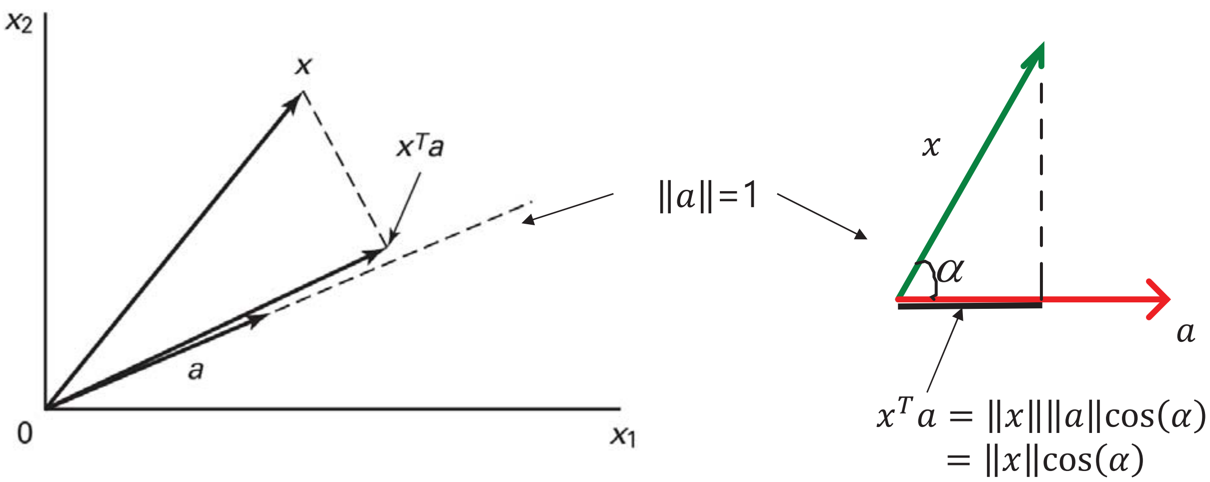
Shear Transformations

- Hold one direction constant and transform ("pull") the other direction

$$\begin{bmatrix} 1 & 0 \\ -0.5 & 1 \end{bmatrix}$$



Projection: Using Inner Products (I)

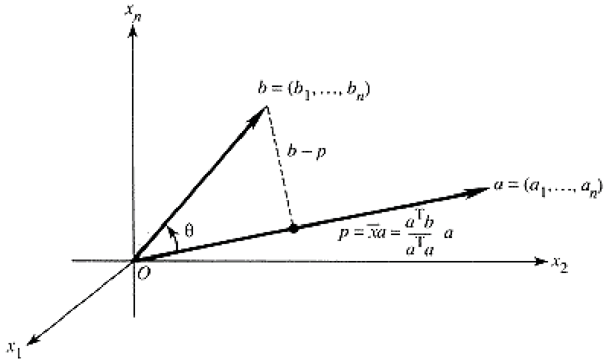


Projection of x along the direction **a** ($\|\mathbf{a}\| = 1$).

$$\mathbf{p} = \mathbf{a}\,(\mathrm{a}^{\mathrm{T}}\mathrm{x})$$
$$\|\mathrm{a}\| = \mathrm{a}^{\mathrm{T}}\mathrm{a} = 1$$

Projection: Using Inner Products (II)

p = **a** ($a^T b$)/ ($a^T a$)



Note: the "error vector" **e** = **b-p** is orthogonal (perpendicular) to **p**.
i.e. Inner product: **(b-p)^{T}p = 0**

<u>"Orthogonalization" principle:</u> after projection, the difference or "error" is orthogonal to the projection

<u>**Sneak peek :**</u> we use this idea to find a *"least-squares"* line that minimizes the sum of squared errors (i.e. min **Σe^{T}e**).

The projection theorem

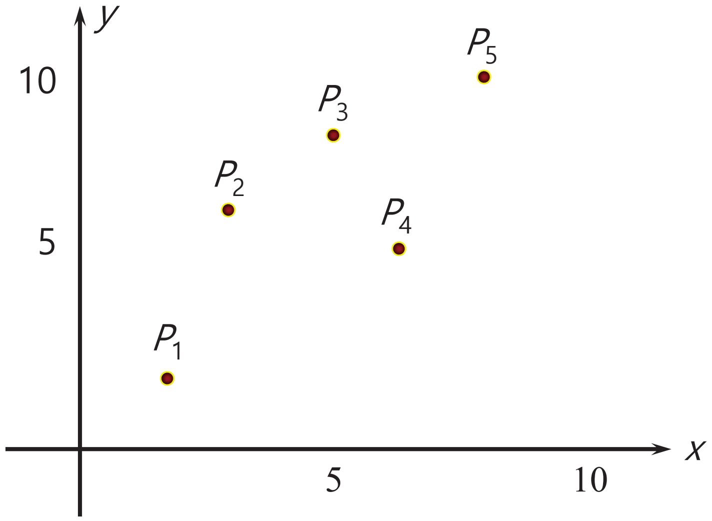
- Projections and Orthogonal Projections $P \in \Re^{n \times n}$
  - P is said to be a projection if $P^2 = P$
  - P is said to be an orthogonal projection if $P = P^T$, i.e. $R(P) \perp N(P)$
- Properties of Projections (Let P be an n×n projection matrix)
  - P an orthogonal projection matrix if and only if $P = P^T$.
  - $I - P$ is a projection, $R(I-P) = N(P)$, $R(P) = N(I-P)$.
  - If P is an orthogonal projection, then $I - P$ is also an orthogonal projection and $R(I-P) \perp R(P)$.
  - Let $S_1 = R(P)$, $S_2 = N(P)$ and $v \in R^n$. If $v_1 \in S_1, v_2 \in S_2$ such that $v_1 + v_2 = v$, then $v_1 = Pv$ and $v_2 = (I-P)v$.
- Theorem (The projection theorem)
  - Let S be a Hilbert space and let V be a closed subspace of S. For any vector $x \in S$, there exists a unique vector $v_0 \in V$ closest to x, i.e. $\|\mathbf{x} - \mathbf{v}_0\| \le \|\mathbf{x} - \mathbf{v}\|, \quad \forall \mathbf{v} \in S$ Furthermore, the point v_0 is the minimizer of $\|x - v\|$ if and only if $x - v_0$ is orthogonal to V.

The Method of Least Squares

- Suppose we are given the data points

$$P_1(x_1, y_1),\ P_2(x_2, y_2),\ P_3(x_3, y_3),\ P_4(x_4, y_4), \text{ and } P_5(x_5, y_5)$$

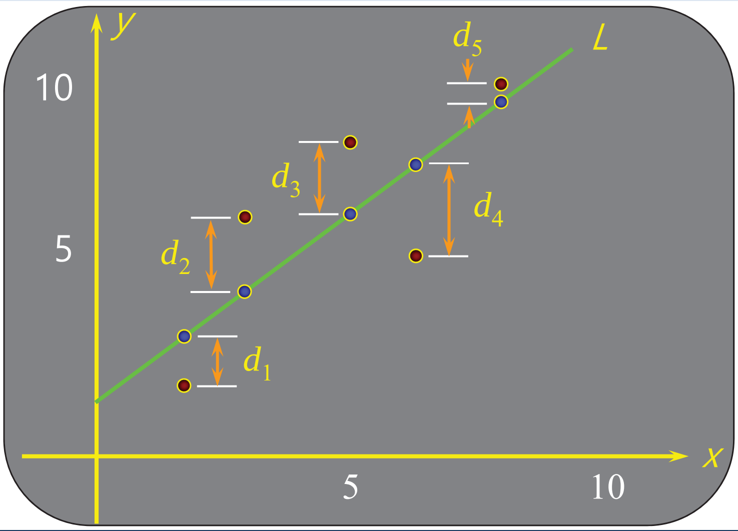
  that describe the relationship between two variables x and y.
- By plotting these data points, we obtain a scatter diagram:



The Method of Least Squares

- Suppose the regression line L is $y = f(x) = mx + b$, where m and b are to be determined.
- The principle of least squares states that the straight line L that fits the data points best is the one chosen by requiring that the sum of the squares of d_1, d_2, d_3, d_4, and d_5, that is be made as small as possible where the distances d_1, d_2, d_3, d_4, and d_5, represent the errors the line L is making in estimating these points

$$f(m,b) = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 = (mx_1 + b - y_1)^2 + (mx_2 + b - y_2)^2 + (mx_3 + b - y_3)^2 + (mx_4 + b - y_4)^2 + (mx_5 + b - y_5)^2$$



Least Squares

Theorem (Least Squares Solution)

- The least squares solution to an inconsistent system Ax = b of m equations in n unknowns, i.e.,

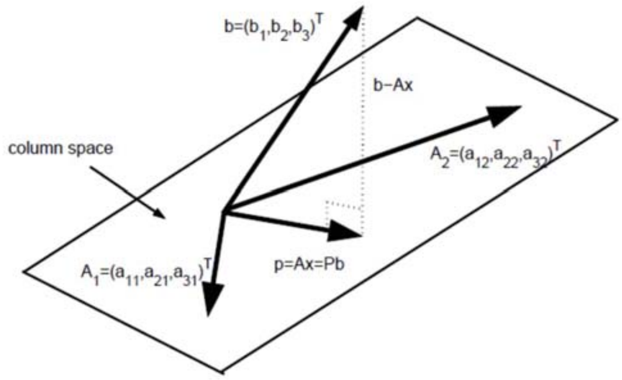
$$\min \|Ax - b\|$$

satisfies $A^T A\bar{x} = A^T b$: *normal equations*

If p is the projection of b onto the column space of A, then $p = A\bar{x} = Pb$

where P is an orthogonal projection matrix. Moreover, if the columns of A are linearly independent, then A^TA is invertible and

$$\bar{x} = (A^T A)^{-1} A^T b \qquad P = A(A^T A)^{-1} A^T$$



Orthogonal matrices

Orthogonal matrices

- If Q has real elements and $Q^TQ = I$, then Q is said to be an orthogonal matrix.
- If the columns of Q are orthonormal then

$$Q^TQ = \begin{bmatrix} -- & q_1^T & -- \\ -- & q_2^T & -- \\ & \vdots & \\ -- & q_n^T & -- \end{bmatrix} \begin{bmatrix} | & | & & | \\ q_1 & q_2 & \cdots & q_n \\ | & | & & | \end{bmatrix} = I \quad \Longrightarrow \quad Q^T = Q^{-1}$$

- Orthogonal matrix Q preserves lengths, inner products, and angles.

$$\|Qx\| = \|x\| \qquad (Qx)^T(Qy) = x^T y \qquad \|Y\|_F = \|X\|_F$$

Lemma

- Let $Q \in \Re^{l\times m}$, $l > m$, and $Z \in \Re^{n\times r}$, $n < r$, be orthogonal, then for any $A \in \Re^{m\times n}$,

$$\| QAZ \|_2 = \| A \|_2 \qquad and \qquad \| QAZ \|_F = \| A \|_F$$

Least Squares

- Theorem (Orthogonal matrix)
  - If the columns of $Q = [q_1, \dots, q_r] \in R^{n\times r}$ are an orthonomal basis for a subspace S, then the least squares problem $\min \|Qx-b\|$ becomes easy

    $$Q^T Q\bar{x} = Q^T b \Rightarrow \bar{x} = Q^T b$$

    The projection of b onto the column space is $p = Q\bar{x} = QQ^T b$ and

    $$P = QQ^T = \sum_{i=1}^{r} q_i q_i^T$$

    is the unique orthogonal projection onto S. In particular, if $v \in R^n$, then $P_v = vv^T / v^T v$ is the orthogonal projection onto $S = \mathrm{span}\{v\}$ and
    $P_v^{\perp} = I - vv^T / v^T v$ is the orthogonal projection onto $S^{\perp}$.
  - If the columns of $Q = [q_1, \dots, q_n] \in R^{n\times n}$ are an orthonomal basis, then b can be written as

    $$b = x_1 q_1 + \cdots + x_n q_n = Qx$$

    $$x = Q^T b = \begin{bmatrix} -- & q_1^T & -- \\ -- & q_2^T & -- \\ & \vdots & \\ -- & q_n^T & -- \end{bmatrix} [b] = \begin{bmatrix} q_1^T b \\ q_2^T b \\ \vdots \\ q_n^T b \end{bmatrix}$$

    $$b = QQ^T b = (q_1^T b) q_1 + \cdots + (q_n^T b) q_n$$

QR factorization

- Theorem(Gram-Schmidt Orthogonalization)
  - The Gram-Schmidt process starts with independent vectors $a_1, \dots, a_n$ and ends with orthonormal vectors $q_1, \dots, q_n$. At step k, it subtracts from a_k its components in the direction that are already settled:

    $$\mathbf{a}'_k = \mathbf{a}_k - \sum_{i=1}^{k-1} \langle \mathbf{a}_k, \mathbf{q}_i \rangle \mathbf{q}_i \qquad \mathbf{q}_k = \frac{\mathbf{a}'_k}{\|\mathbf{a}'_k\|_2}$$

    $$A = QR$$

    $$= \begin{bmatrix} | & & | \\ q_1 \,| & \cdots & |\, q_n \\ | & & | \end{bmatrix} \begin{bmatrix} \|a_1\| & q_1^T a_2 & \cdots & q_1^T a_n \\ & \|a_2'\| & \cdots & q_2^T a_n \\ & & \ddots & \vdots \\ & & & \|a_n'\| \end{bmatrix}$$

  - Operation Count of the Gram-Schmidt algorithm: ~ $2mn^2$ flops.

QR factorization

- Classical Gram-Schmidt Algorithm
  - computes a single orthogonal projection of rank $m-(k-1)$:

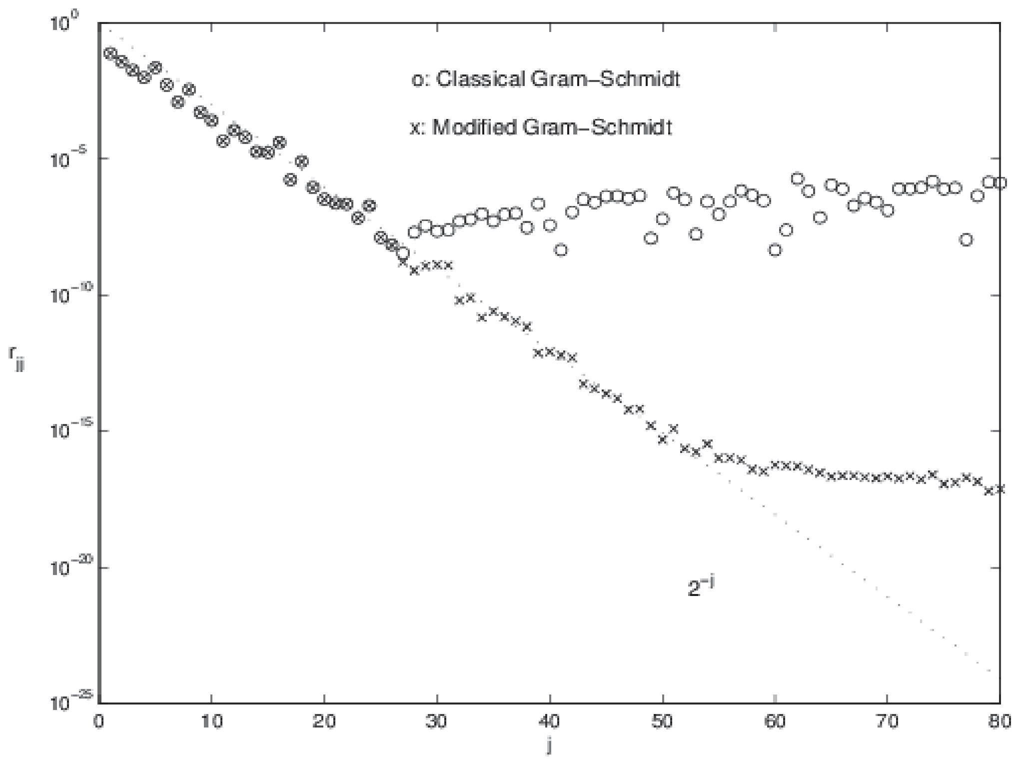
$$\mathbf{a}'_k = \left(I - \sum_{i=1}^{k-1} \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{a}_k = (I - P_{k-1})\mathbf{a}_k$$

- Modified Gram-Schmidt algorithm
  - computes the same result by a sequence of $k-1$ orthogonal projections of rank $m-1$:

$$\mathbf{a}'_k = \left(\prod_{i=1}^{k-1} (I - \mathbf{q}_i \mathbf{q}_i^T) \right) \mathbf{a}_k = \left(\prod_{i=1}^{k-1} (I - P_{q_i}) \right) \mathbf{a}_k$$

  - Mathematically Eq. (3) and Eq. (4) are equivalent, but Eq. (4) introduces smaller error than Eq. (3).

Classical v.s. Modified Gram-Schmidt Algorithm



QR factorization

- **Theorem(QR decomposition)**
  - Every m×n matrix A with full rank and $m \geq n$ can be factored into

$$\boxed{A = QR}$$

where Q is an orthogonal m×n matrix and R is an upper triangular n×n. When m = n and all matrices are square, Q becomes an orthogonal matrix.
  - Operation Count of the Householder algorithm: ~ 4/3mn² flops.

- **Least Squares Solution**
  - The solution of the least squares problem $\min\|Ax - b\|$

$$A^T Ax = A^T b \Rightarrow R^T R\overline{x} = R^T Q^T b \longrightarrow R\overline{x} = Q^T b$$

1. Compute a QR factorization A = QR.
2. Compute y = Q^T b.
3. Solve Rx = y for x.

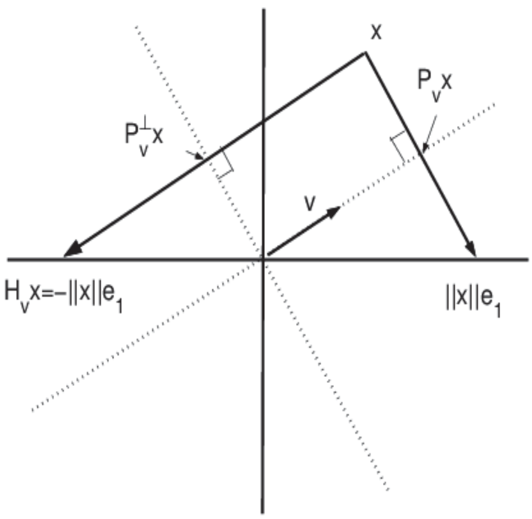
Householder reflector

- **Recall projection matrix**
  - Orthogonal projection matrix onto span($\mathbf{v}$)

$$P_v = \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$$

  - Projection matrix orthogonal to P_v

$$P_v^{\perp} = I - P_v = I - \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$$



- **Householder reflector** matrix with respect to a nonzero vector $\mathbf{v}$:

$$H_v = I - 2\frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}} = I - 2P_v$$

  - Note that the projector $P_v^{\perp}$ is of rank $(m-1)$ and the Householder reflector H_v is an orthogonal matrix with full rank ($H_v^T H_v = I$).

Householder reflection

- Finding Householder reflection

  - used to zero out all the elements of a vector except for one component; for a given vector $\mathbf{x} = [x_1 x_2 \dots x_n]^T$, we want to find a vector $\mathbf{v}$ such that

$$x = \begin{bmatrix} \times \\ \times \\ \vdots \\ \times \end{bmatrix} \rightarrow H_v\mathbf{x} = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \alpha \mathbf{e}_1$$

$$H_v\mathbf{x} = \mathbf{x} - 2\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}\mathbf{v}^T}\mathbf{v} = \pm \| \mathbf{x} \|_2 \ \mathbf{e}_1 \Rightarrow \left(2\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}\mathbf{v}^T}\right)\mathbf{v} = \mathbf{x} \mp \| \mathbf{x} \|_2 \ \mathbf{e}_1 \Rightarrow \mathbf{v} = \mathbf{x} \mp \| \mathbf{x} \|_2 \ \mathbf{e}_1$$

  - If $\mathbf{x}$ is close to a multiple of $\mathbf{e}_1$, $\mathbf{v} = \mathbf{x} - \text{sign}(x_1) \| \mathbf{x} \|_2 \ \mathbf{e}_1$ has a small norm, which could lead to a large relative error. We choose therefore

$$\mathbf{v} = \mathbf{x} + \text{sign}(x_1) \| \mathbf{x} \|_2 \ \mathbf{e}_1$$

Householder QR factorization

- Householder QR factorization

  - At iteration 1,

$$A = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \Rightarrow A^{(1)} = Q_1 A = \begin{bmatrix} \alpha_1 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix}$$

where $Q_1 = H_{v_1}$ and $v_1 = A(:,1) + \text{sign}(A_{11}) \| A(:,1) \|_2 \ \mathbf{e}_1$

  - At iteration 2,

$$\begin{bmatrix} \alpha_1 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \Rightarrow A^{(2)} = Q_2 A^{(1)} = \begin{bmatrix} \alpha_1 & \times & \times \\ 0 & \alpha_2 & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix}$$

where $Q_2 = \begin{bmatrix} 1 & \mathbf{0} \\ 0 & H_2 \end{bmatrix} = I - 2\frac{\overline{\mathbf{v}}_2 \overline{\mathbf{v}}_2^T}{\overline{\mathbf{v}}_2^T \overline{\mathbf{v}}_2}, \quad \overline{\mathbf{v}}_2 = \begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}$

and $v_2 = A^{(1)}(2:m,2) + \text{sign}(A_{22}^{(1)}) \| A^{(1)}(2:m,2) \|_2 \ \mathbf{e}_2(2:m)$

Householder QR factorization (continued)

- Householder QR factorization (continued)
  - At iteration 3, $\begin{bmatrix} \alpha_1 & \times & \times \\ 0 & \alpha_2 & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix} \Rightarrow A^{(3)} = Q_3 A^{(2)} = \begin{bmatrix} \alpha_1 & \times & \times \\ 0 & \alpha_2 & \times \\ 0 & 0 & \alpha_3 \\ 0 & 0 & 0 \end{bmatrix}$

    where

$$Q_3 = \begin{bmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & 0 \\ 0 & 0 & H_3 \end{bmatrix} = I - 2\frac{\bar{\mathbf{v}}_3 \bar{\mathbf{v}}_3^{\mathbf{T}}}{\bar{\mathbf{v}}_3{}^T \bar{\mathbf{v}}_3}, \qquad \bar{\mathbf{v}}_3 = \begin{bmatrix} 0 \\ 0 \\ \mathbf{v}_3 \end{bmatrix}$$

    and $v_3 = A^{(2)}(3{:}m,3) + \text{sign}(A_{33}^{(2)}) \parallel A^{(2)}(3{:}m,3) \parallel_2 \mathbf{e}_3(3{:}m)$
  - Therefore $Q_1Q_2Q_3A = R$. Since the Q_i's are orthogonal, we have $A = QR$ by setting $Q^T = Q_1Q_2Q_3$ or $Q = Q_3Q_2Q_1$.
  - At iteration k, $Q_j = \begin{bmatrix} I & 0 \\ 0 & H_j \end{bmatrix} = I - 2\frac{\bar{\mathbf{v}}_j \bar{\mathbf{v}}_{\mathbf{j}}^{\mathbf{T}}}{\bar{\mathbf{v}}_{\mathbf{j}}{}^T \bar{\mathbf{v}}_{\mathbf{j}}}, \quad \bar{\mathbf{v}}_{\mathbf{j}} = \begin{bmatrix} 0 \\ \mathbf{v}_j \end{bmatrix}$

Chapter 4. Eigenvalues and Singular Value Decomposition

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Maxim

- The laws of Nature are expressed by differential equations, so it is useful to solve differential equations.

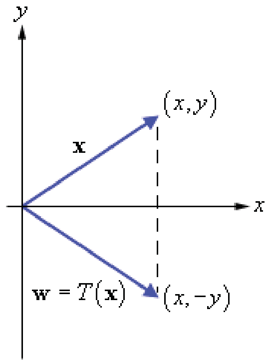
---Isaac Newton

Reading

- [Strang. (2006), Chapter 5 Eigenvalues and Eigenvectors]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.

Invariants of Matrices: Eigenvectors

- Eigenvectors
  - Consider a NxN matrix (or linear transformation) T
  - An *invariant input* x of a function T(x) is nice because it <u>does not change</u> when the function T is applied to it.
    - i.e. solve this eqn for **x**: T(**x**) = **x**
  - We <u>allow (positive or negative) scaling</u>, but want invariance w.r.t <u>direction</u>:
    - $T(\mathbf{x}) = \lambda \mathbf{x}$
  - There are multiple solutions to this equation, equal to the <u>rank</u> of the matrix T. If T is "<u>full" rank</u>, then we have a full set of solutions.
  - These invariant solution vectors **x** are *eigenvectors*, and the "<u>characteristic</u>" scaling factors associated w/ each **x** are *eigenvalues.*



$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

<u>E-vectors:</u>
- Points on the x-axis unaffected $[1\ 0]^T$
- Points on y-axis are flipped $[0\ 1]^T$
(but this is equivalent to scaling by -1!)
E-values: 1, -1 (also on diagonal of matrix)

Eigenvectors (contd)

- Eigenvectors are even more interesting because <u>any vector</u> in the domain of T can now be …
  - … viewed in a *new* coordinate system formed with the invariant "eigen" directions as a basis.
  - The operation of T(**x**) is now <u>decomposable</u> into simpler operations on **x**,
  - … which involve projecting **x** onto the "eigen" directions and applying the characteristic (eigenvalue) scaling along those directions

Eigenvalues & Eigenvectors

- **Eigenvectors** (for a square $m \times m$ matrix $\mathbf{S}$)

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v}$$

(right) eigenvector $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$

eigenvalue $\lambda \in \mathbb{R}$

Example

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- How many eigenvalues are there at most?

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v} \iff (\mathbf{S} - \lambda \mathbf{I})\,\mathbf{v} = \mathbf{0}$$

only has a non-zero solution if $|\mathbf{S} - \lambda \mathbf{I}| = 0$

this is a m-th order equation in λ which can have **at most m distinct solutions** (roots of the characteristic polynomial) – <u>can be complex even though **S** is real.</u>

Eigenvalues

- Eigenvalue decomposition $A \in \Re^{n \times n}$
  - A nonzero vector $x \in R^n$ is an eigenvector of A and the number λ is its corresponding eigenvalue of A if they satisfy

$$Ax = \lambda x$$

  - The action of a matrix A on a subspace S of R^n may sometimes mimic the action of scalar multiplication. When this happens, the special subspace S is called an eigenspace.
  - Suppose the n×n matrix A has n linearly independent eigenvectors which are chosen to be the columns of a matrix S.

$$AS = A\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \lambda_1 x_1 & \lambda_2 x_2 & \cdots & \lambda_n x_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = S\Lambda$$

$$\boxed{S^{-1}AS = \Lambda \quad or \quad A = S\Lambda S^{-1}}$$

eigenvalue decomposition of A

Eigenvalues

- Properties of eigenvalues $A \in \Re^{n\times n}$
  - λ is an eigenvalue of A if and only if it is the solution of the characteristic equation defined by $\det(A-\lambda I)=0$
  - If the eigenvectors $v_1, v_2, \dots, v_k$ correspond to different eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$, then those eigenvectors are linearly independent.
  - $\mathrm{tr}(A)=\lambda_1+\cdots+\lambda_n$, $\det(A)=\lambda_1\times\cdots\times\lambda_n$
  - The eigenvalues of A^k are $\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k$, the k-th power of the eigenvalues of A. Each eigenvector of A is still an eigenvector of A^k, and if S diagonalizes A it also diagonalizes A^k

$$A^k = S\Lambda^k S^{-1}$$

  - If A and B are diagonalizable, they share the same eigenvector matrix if and only if AB = BA.
  - A is positive definite if and only if all its eigenvalue are positive.
  - If λ is an eigenvalue of A if $\lambda+\alpha$ is an eigenvalue of $A+\alpha I$, for any real α.
  - Let λ > 0 be an eigenvalue of the matrix A^TA, $A \in \Re^{m\times n}$. Then λ is also an eigenvalue of AA^T with the same multiplicity.

Eigenvalues

- Theorem (Spectral Theorem)
  - Let $A \in \Re^{n\times n}$ be symmetric. Then A can be factored into $A = V\Lambda V^T$ where V is an orthogonal matrix and Λ is a real diagonal matrix, i.e., A has n real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, and a corresponding set of eigenvectors $v_1, v_2, \dots, v_n$ that form an orthogonal basis for R^n.

$$A = V\Lambda V^T = \lambda_1 v_1 v_1^T + \cdots + \lambda_n v_n v_n^T$$

- This is why we love *symmetric (or hermitian) matrices*: they admit nice decomposition
  - We love positive definite matrices even more: they are symmetric and all have all eigenvalues strictly positive.
  - Many linear systems are equivalent to symmetric/hermitian or positive definite transformations.

Example: Diagonal (Eigen) decomposition

- Let $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}; \lambda_1 = 1, \lambda_2 = 3.$

- The eigenvectors $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ form $U = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$

- Inverting, we have $U^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

Recall $UU^{-1} = 1$.

- Then $\boldsymbol{S=U\Lambda U^{-1}} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

- Then, S= $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

$Q = U/\sqrt{2} \longrightarrow$ ***Q*** $\qquad \boldsymbol{\Lambda} \qquad$ ***($Q^{-1} = Q^T$)***

Eigenvalues

- On computing eigenvalues
  - Eigenvalue problems can be reduced to polynomial rootfinding problems and vice versa. It is well-known (by Abel and Galois's theorem) that no formula exists for expressing the roots of an arbitrary polynomial (with a degree >5), given its coefficients. Therefore, any eigenvalue solver must be iterative.
  - In general because polynomial rootfinding is a highly ill-conditioned problem but eigenvalue problem is well-conditioned.
  - Fortunately, in practice computing eigenvalues differs from the solution of linear systems by only a small constant factor, typically closer too 1 and 10, although it is an "unsolvable" problem in principle.

Geometric View: EigenVectors

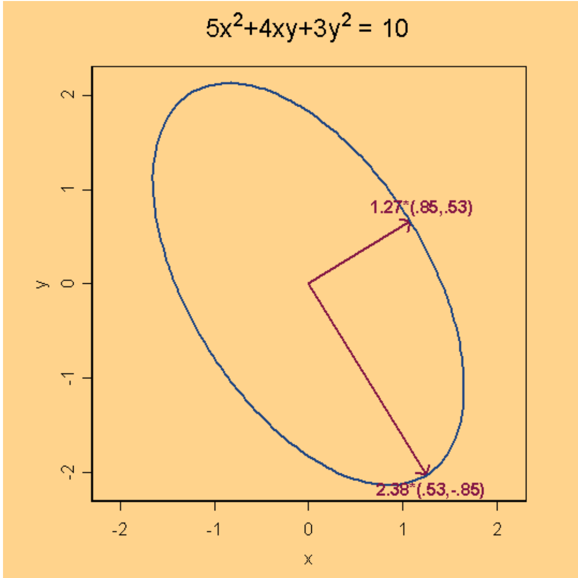
- Homogeneous (2nd order) multivariable equations: $ax^2 + 2kxy + by^2 = c$
- Represented in matrix (quadratic) form w/ symmetric matrix A:

$$\boxed{\mathbf{x}^T \mathbf{A}\mathbf{x} = c,} \quad \text{where} \quad \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} a & k \\ k & b \end{pmatrix}$$

- **Eigenvector decomposition**:

$$5x^2 + 4xy + 3y^2 = 10 \qquad \mathbf{A} = \begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}$$

$$\lambda_1 = 6.24, \mathbf{s}_1 = \begin{pmatrix} 0.85 \\ 0.53 \end{pmatrix} \qquad \lambda_2 = 1.76, s_2 = \begin{pmatrix} 0.53 \\ -0.85 \end{pmatrix}$$



- Geometry: Principal Axes of Ellipse
- Symmetric A => *orthogonal e-vectors*!
- Positive Definite A => +ve real e-values!

Why do Eigenvalues/vectors matter?

- Eigenvectors are invariants of A
  - Don't change direction when operated A
- Recall $d(e^{\lambda t})/dt = \lambda e^{\lambda t}$.
  - $e^{\lambda t}$ is an invariant function for the linear operator d/dt, with eigenvalue λ
- Pair of differential eqns:
  - $dv/dt = 4v - 5u$
  - $du/dt = 2u - 3w$
- Can be written as: $d\mathbf{y}/dt = \mathbf{Ay}$, where $\mathbf{y} = [v\ u]^T$
  - $\mathbf{y} = [v\ u]^T$ at time 0 $= [8\ 5]^T$
- Substitute $\mathbf{y} = e^{\lambda t}\mathbf{x}$ into the equation $d\mathbf{y}/dt = \mathbf{Ay}$
  - $\lambda e^{\lambda t}\mathbf{x} = \mathbf{A}e^{\lambda t}\mathbf{x}$
  - This simplifies to the eigenvalue vector equation: $\mathbf{Ax} = \lambda\mathbf{x}$
- Solutions of multivariable differential equations (the bread-and-butter in linear systems) correspond to solutions of linear algebraic eigenvalue equations!

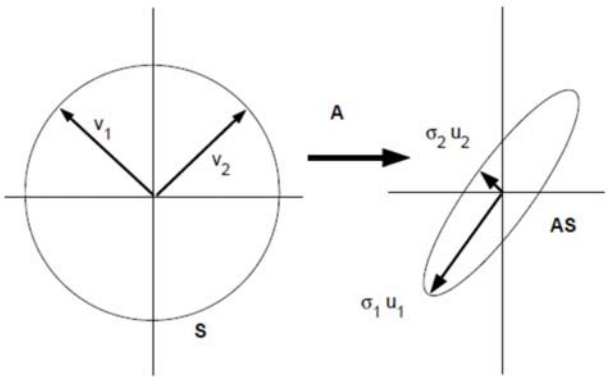
Reading

- [Strang. (2006), Chapter 6 Positive Definite Matrices]
- G. Strang *Linear Algebra And Its Applications-4th ed.* Cengage Learning, New York, 2006.

Singular Value Decomposition (SVD)

- A Geometric Observation $A \in \Re^{m \times n}$
  - The image of the unit sphere under any m×n matrix is a hyperellipse.
  - Let S be the unit sphere in R^n, and take any m×n matrix A, with $m \geq n$. Then the image AS is a hyperellipse in R^m.
    - singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ of A by the lengths of the n principal semiaxes of AS.
    - left singular vectors $u_1, u_2, \ldots, u_n$ of A by the unit vectors oriented in the n principal semiaxes of AS correspond with $\sigma_1, \sigma_2, \ldots, \sigma_n$
    - right singular vectors $v_1, v_2, \ldots, v_n$ of A by the unit vectors of S that are the preimages of the n principal semiaxes of AS

$$Av_j = \sigma_j u_j, \ 1 \leq j \leq n$$



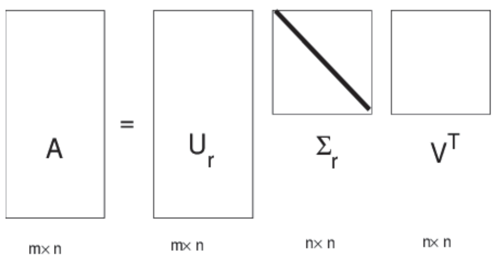
Singular Value Decomposition (SVD)

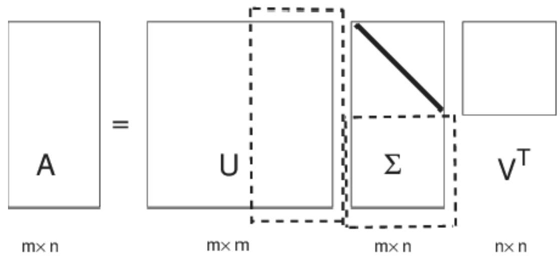
- A Geometric Observation $A \in \Re^{m\times n}$

$$Av_j = \sigma_j u_j,\ 1 \le j \le n$$

$$AV = A\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix}\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} = \hat{U}\hat{\Sigma}$$

$A = \hat{U}\hat{\Sigma}V^T$ (*Reduced SVD*) $A = U\Sigma V^T$ (*Full SVD*)





Singular Value Decomposition (SVD)

- Theorem (Singular Value Decomposition) for $A \in \Re^{m\times n}$
  - Any m×n matrix A can be factored into

$$\boxed{A = U\Sigma V^T \quad (\text{Full SVD})}$$

$U = [u_1,\ldots,u_m] \in \Re^{m\times m}$ *is orthogonal*

$V = [v_1,\ldots,v_n] \in \Re^{n\times n}$ *is orthogonal*

$\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_n) \in \Re^{m\times n}$ *is diagonal* $\sigma_1 \ge \cdots \sigma_r > \sigma_{r+1} = 0 = \cdots = \sigma_{\min\{m,n\}} = 0$

  - The cost of SVD $\sim \frac{8}{3}m^3$ flops.
  - Moreover the columns of U are eigenvectors of AA^T , the columns of V are eigenvectors of A^TA, and the r nonnegative singular values on the diagonal of Σ are the square roots of the nonzero eigenvalues of both AA^T and A^TA. Furthermore A can be factored into

$$\boxed{A = \hat{U}\hat{\Sigma}V^T \quad (\text{Reduced SVD})}$$

$$\hat{U} = U(:,1:n) = [u_1,\ldots,u_n] \in \Re^{m\times n}$$

$$\hat{\Sigma} = \Sigma(1:n,1:n) = \mathrm{diag}(\sigma_1,\ldots,\sigma_n) \in \Re^{n\times n}$$

SVD example

Let $$A=\begin{bmatrix}1 & -1\\ 0 & 1\\ 1 & 0\end{bmatrix}$$

Thus m=3, n=2. Its SVD is

$$\begin{bmatrix}0 & 2/\sqrt{6} & 1/\sqrt{3}\\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3}\\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3}\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & \sqrt{3}\\ 0 & 0\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}$$

Note: the singular values arranged in decreasing order.

Singular Value Decomposition (SVD)

- Properties of the SVD for $A \in \Re^{m\times m}$
  - $|\det(A)| = \prod_{i=1}^{m} \sigma_i$.
  - If $A = A^T$, then the singular values of A are the absolute values of the eigenvalues of A.
  - If the σ_i are distinct, the left and right singular vectors $\{u_j\}$ and $\{v_j\}$ are uniquely determined up to complex signs (i.e., complex scalar factors of absolute value 1).
- Theorem
  - Let $A = U\Sigma V^T \in \Re^{m\times n}$ is the SVD of A and that $r = \mathrm{rank}(A)$. If we let $U = [U_1, U_2]$, $V = [V_1, V_2]$ where $U_1 = [\mathbf{u}_1, \cdots, \mathbf{u}_r]$, $U_2 = [\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m]$, $V_1 = [\mathbf{v}_1, \cdots, \mathbf{v}_r]$, $V_2 = [\mathbf{v}_{r+1}, \cdots, \mathbf{v}_n]$. Then

$$\mathcal{R}(A) = \mathrm{span}(U_1), \qquad \mathcal{N}(A^T) = \mathrm{span}(U_2),$$
$$\mathcal{R}(A^T) = \mathrm{span}(V_1), \qquad \mathcal{N}(A) = \mathrm{span}(V_2)$$

  - $V_1V_1^T$ = projection onto $R(A^T)$, $V_2V_2^T$ = projection onto $N(A)$
  - $U_1U_1^T$ = projection onto $R(A)$, $U_2U_2^T$ = projection onto $N(A^T)$

Singular Value Decomposition (SVD)

- Matrix Structure from the SVD for $A \in \Re^{m\times n}$
  - Let $A = U\Sigma V^T \in \Re^{m\times n}$ is the SVD of A, where $\sigma_1 \geq \cdots \sigma_r > \sigma_{r+1} = \cdots = \sigma_{min\{m,n\}} = 0$.

    1) $\| A \|_F = \| \Sigma \|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$.

    2) $\| A \|_2 = \| \Sigma \|_2 = \sigma_1$.

    3) $\min_{x\neq 0} \| Ax \|_2 / \| x \|_2 = \sigma_n$.
  - Spectral decomposition

$$\boxed{A = \sum_{i=1}^{r} \sigma_i u_i v_i^T}$$

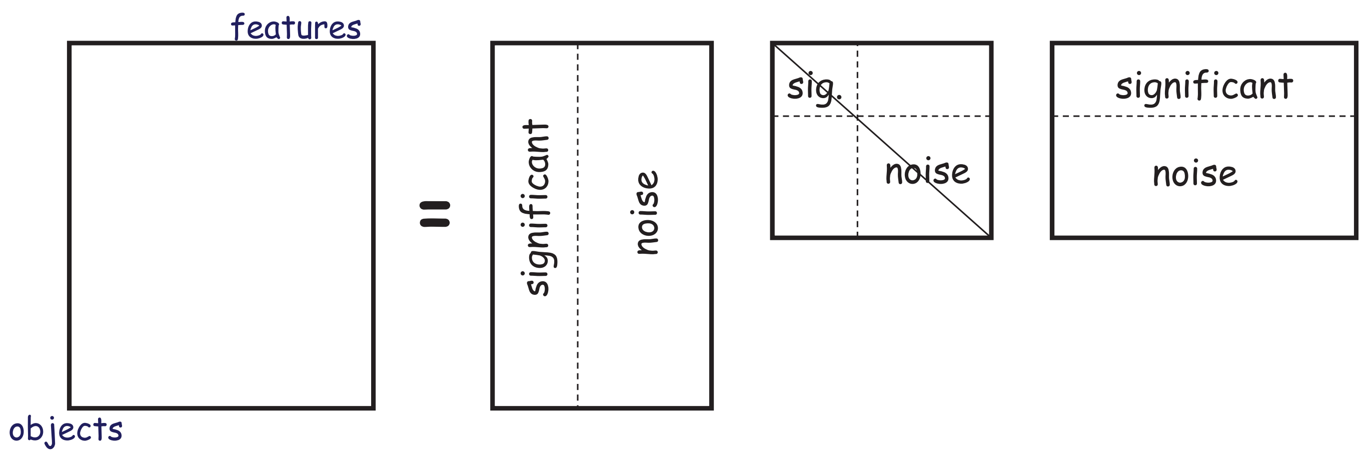
  - If $k < r = \text{rank}(A)$ and $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, then

$$\min_{\text{rank}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}$$

$$\min_{\text{rank}(B)=k} \|A - B\|_F = \|A - A_k\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}$$

Example: Low-rank Approximation w/ SVD

A = U Σ V

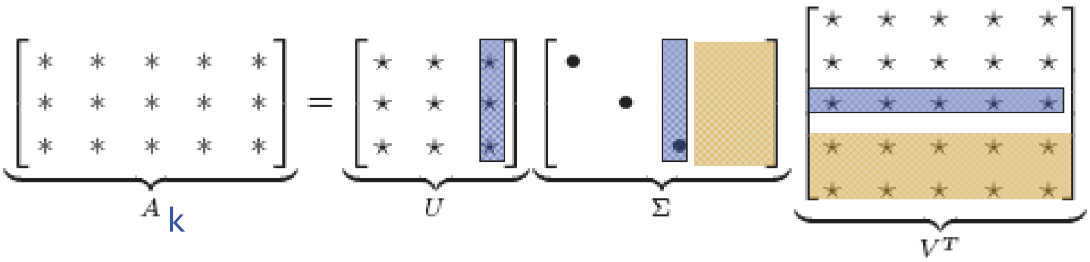


Can be used for noise rejection (compression):
aka low-rank approximation

Example: Low-rank Approximation w/ SVD

$$A_k = U\ \mathrm{diag}(\sigma_1,\ldots,\sigma_k,\underbrace{0,\ldots,0}_{})V^T$$

set smallest r-k singular values to zero



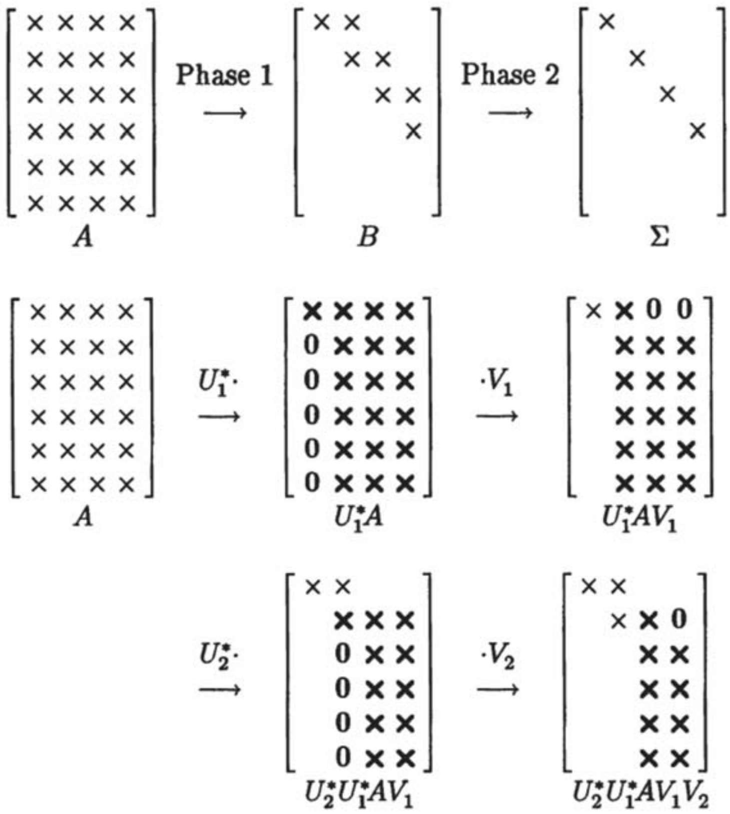
$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \longleftarrow$$

column notation: sum of rank 1 matrices

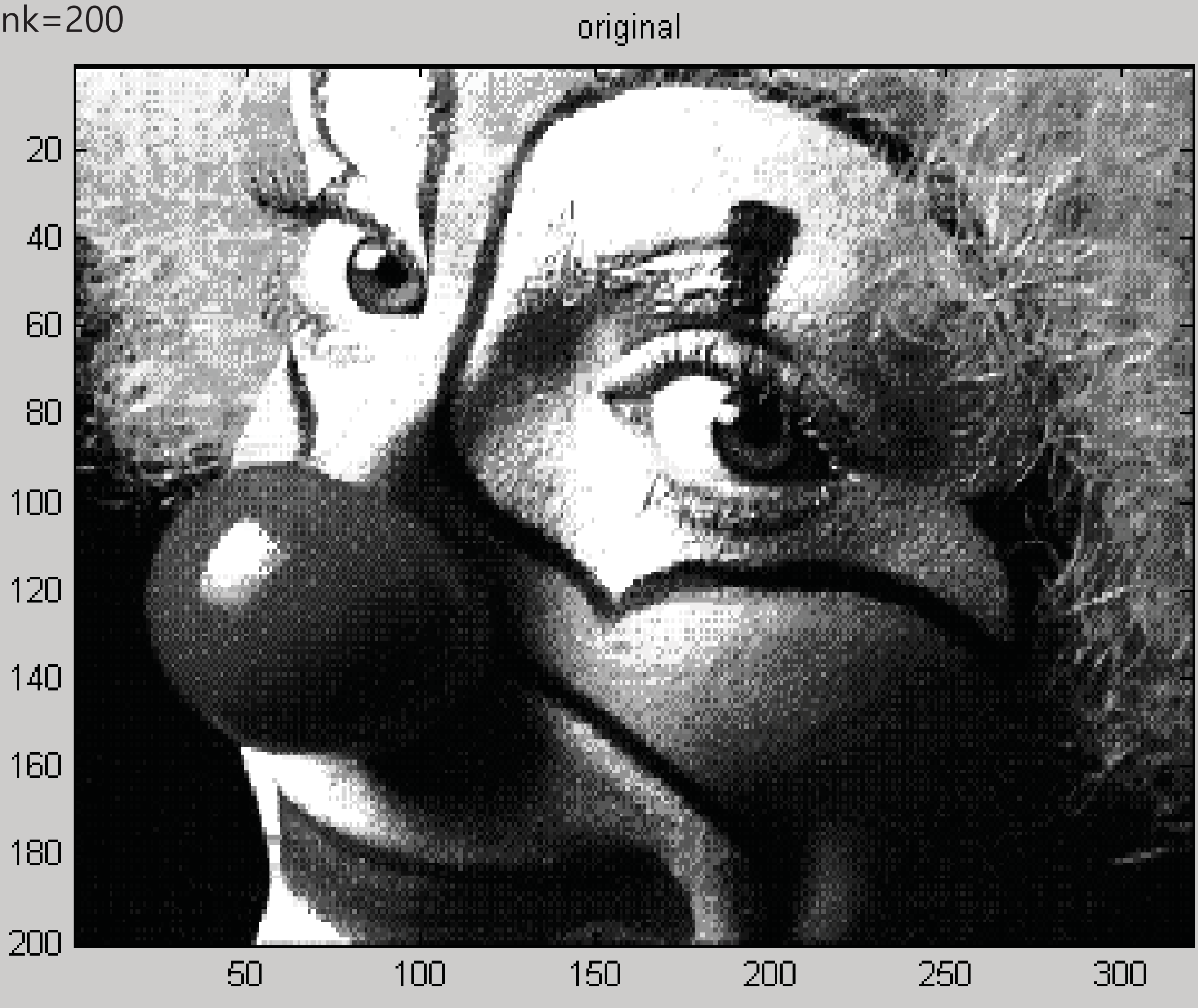
Singular Value Decomposition (SVD)

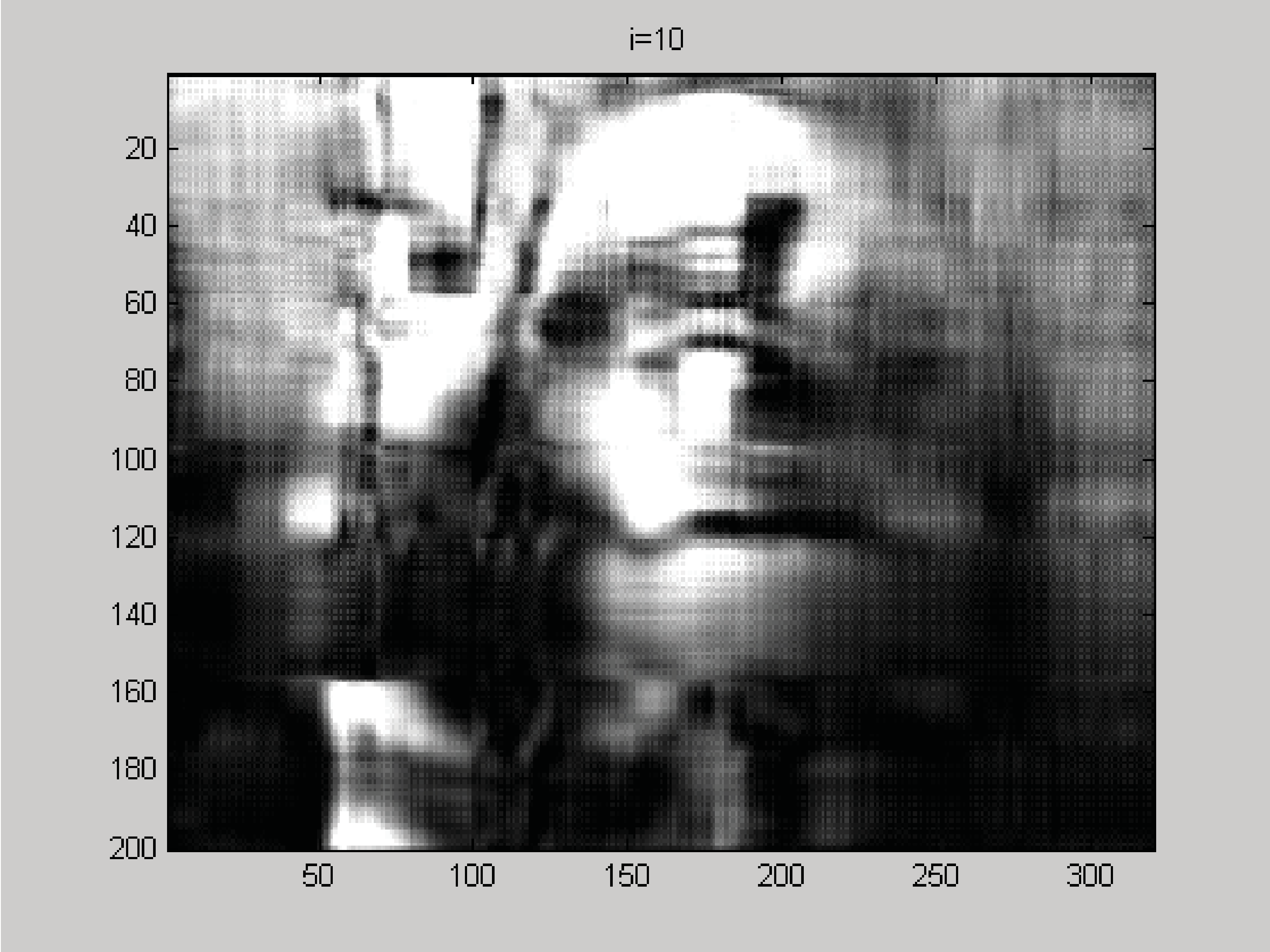
- SVD v.s. Eigenvalue decomposition
  - The SVD uses two different bases (the set of left and right singular vectors), whereas the eigenvalue decomposition uses just one (the eigenvectors).
  - The SVD uses orthonormal bases, whereas the eigenvalue decomposition uses a basis that in general is not orthogonal.
  - Not all matrices (even square ones) have an eigenvalue decomposition, but all matrices (even rectangular ones) have a SVD.
  - Conceptually, eigenvalues tend to be relevant to questions involving the behavior of iterated forms of A, such as matrix powers A^k or exponentials e^{tA}, whereas singular vectors tend to be relevant to questions involving the behavior of A itself.

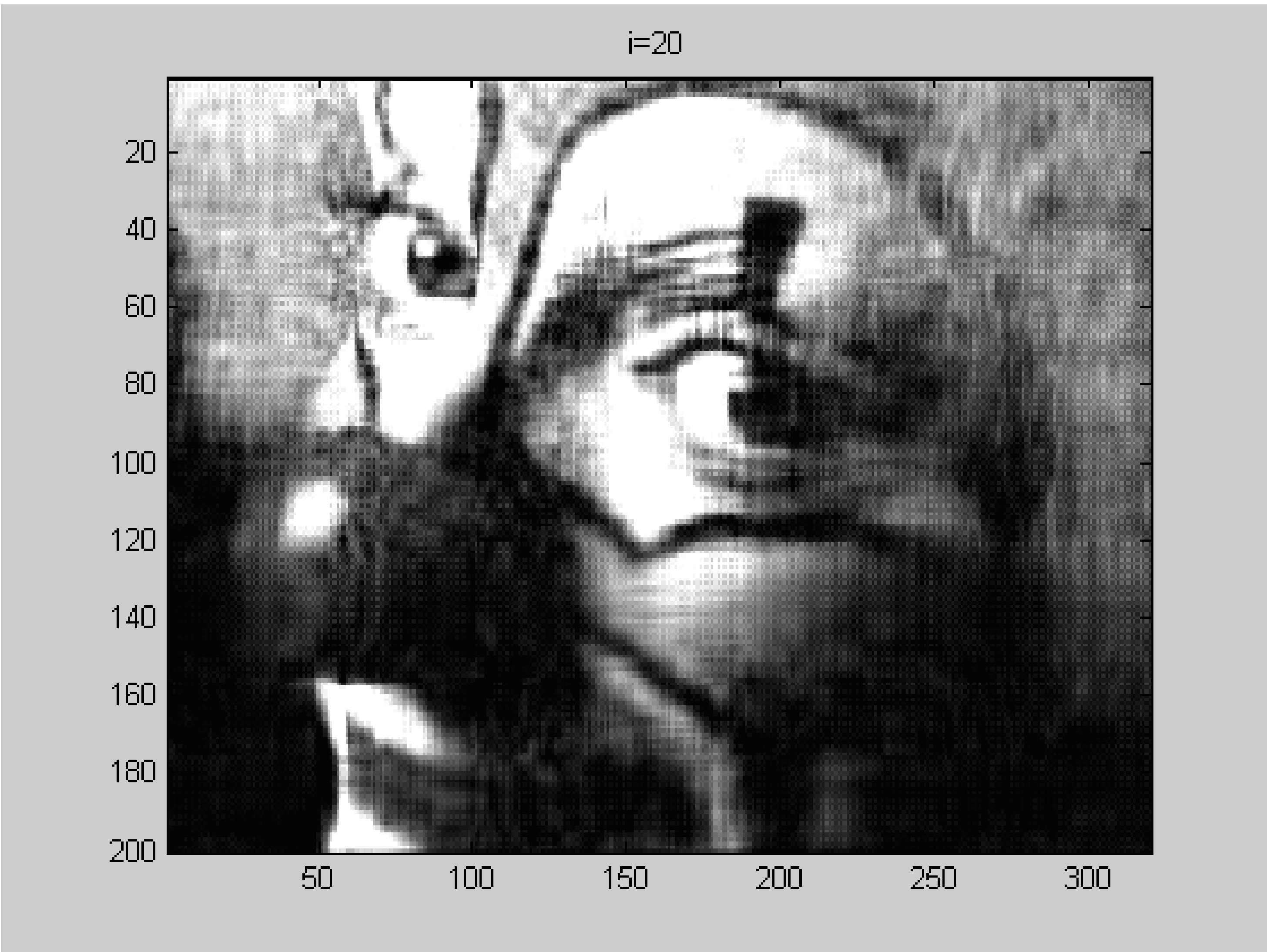
Computing the SVD

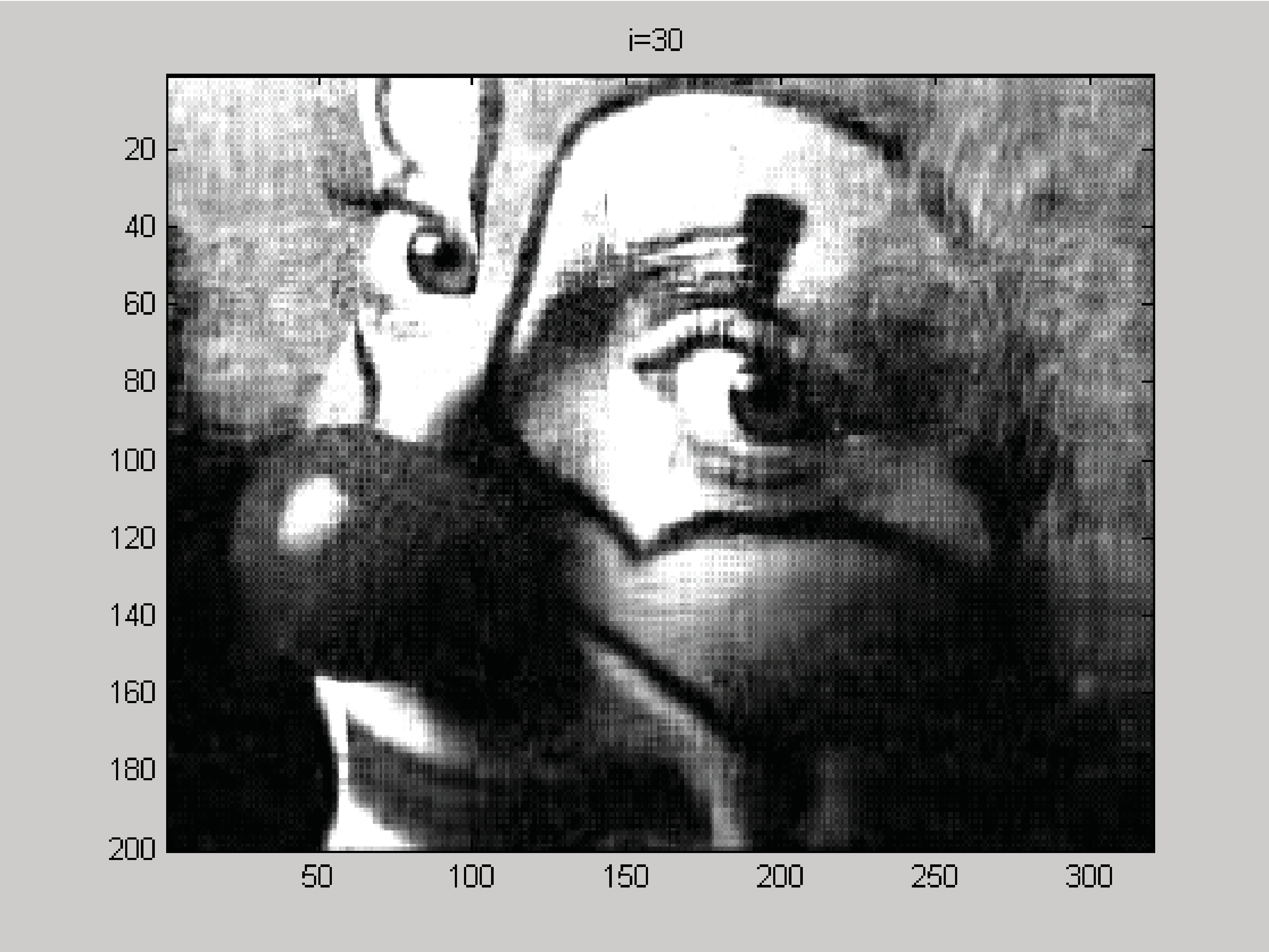


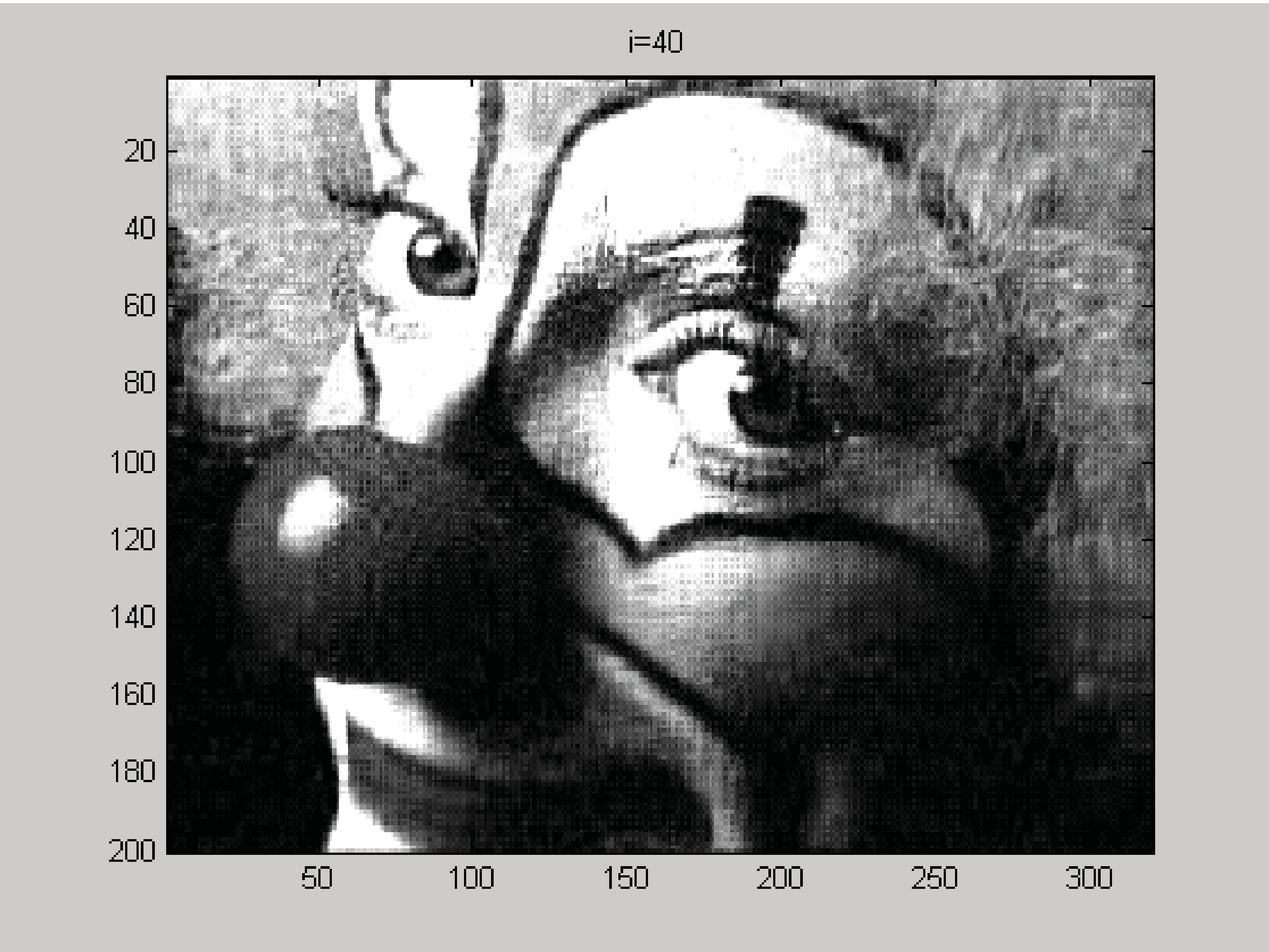
Rank=200




i=10
20
40
60
80
100
120
140
160
180
200
50
100
150
200
250
300


i=20
20
40
60
80
100
120
140
160
180
200
50
100
150
200
250
300


i=30
20
40
60
80
100
120
140
160
180
200
50
100
150
200
250
300


i=40
20
40
60
80
100
120
140
160
180
200
50
100
150
200
250
300

Chapter 5. Calculus and Convex

Prof. Jaewook Lee
Statistical Learning and Computational Finance Lab.
Department of Industrial Engineering
jaewook@snu.ac.kr
http://slcf.snu.ac.kr

Maxim

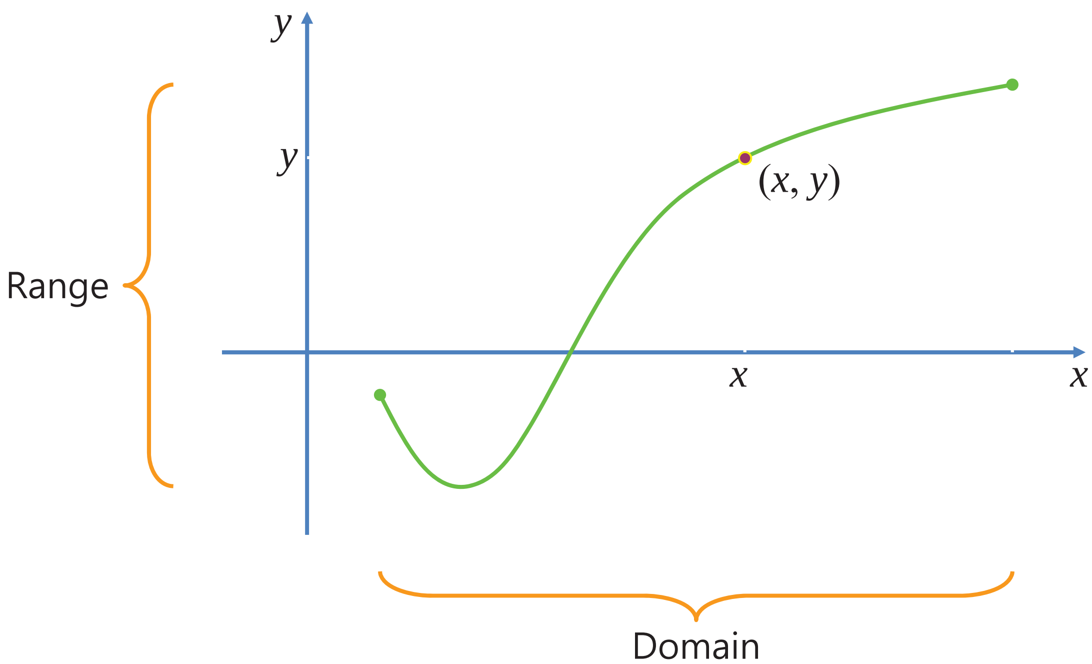
- The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it because it is beautiful. If nature were not beautiful, it would not be worth knowing, and if nature were not worth knowing, life would not be worth living.

--- Henri Poincare

CALCULUS BACKGROUNDS

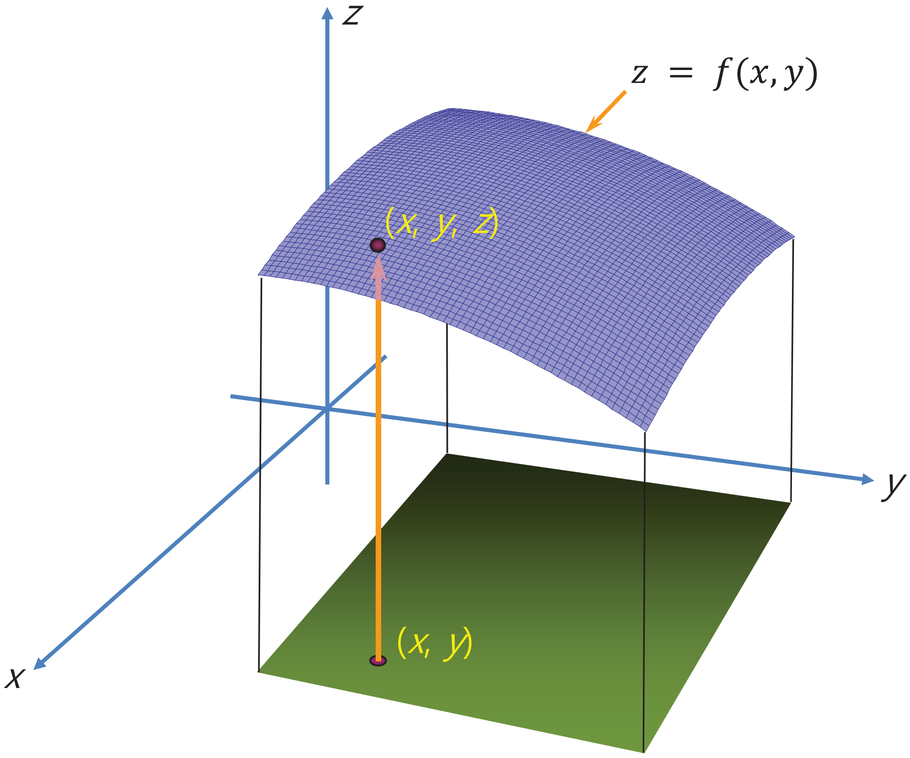
Graph of a function f

- The graph of a function *f* is shown below:

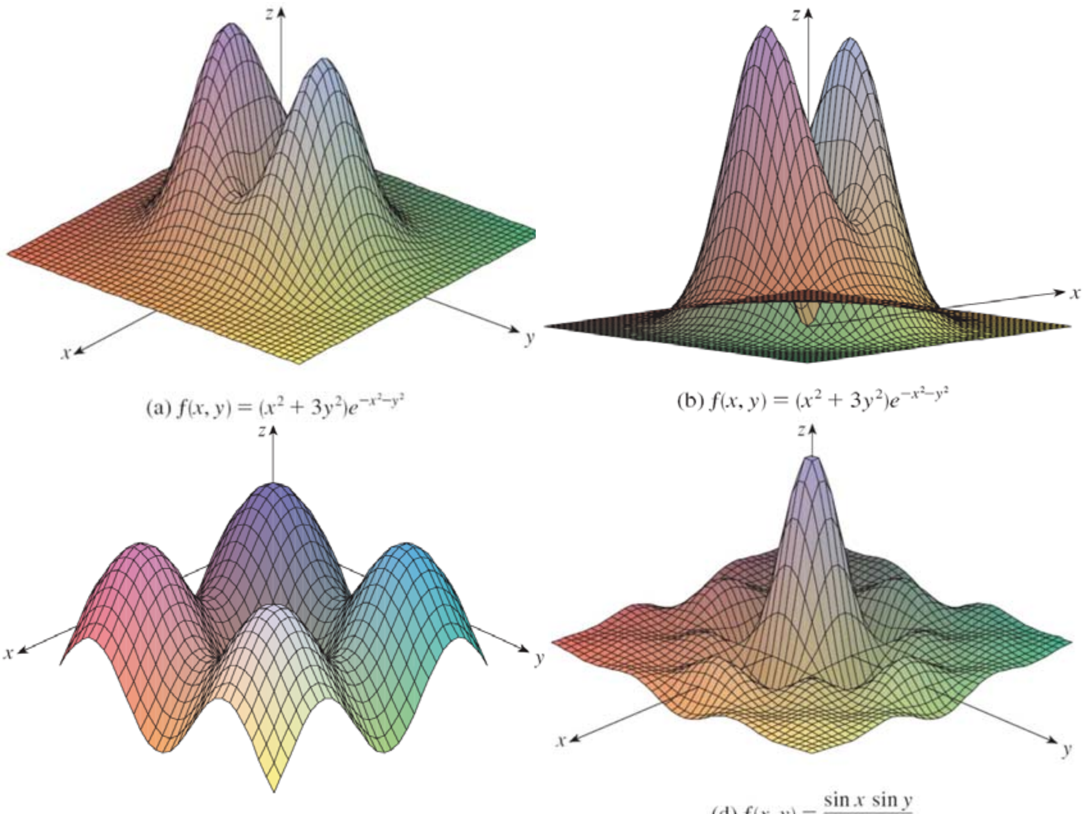


Example 5, page 53

Graphs of Functions of Two Variables



Computer-generated graphs of several functions



(a) $f(x, y) = (x^2 + 3y^2)e^{-x^2-y^2}$

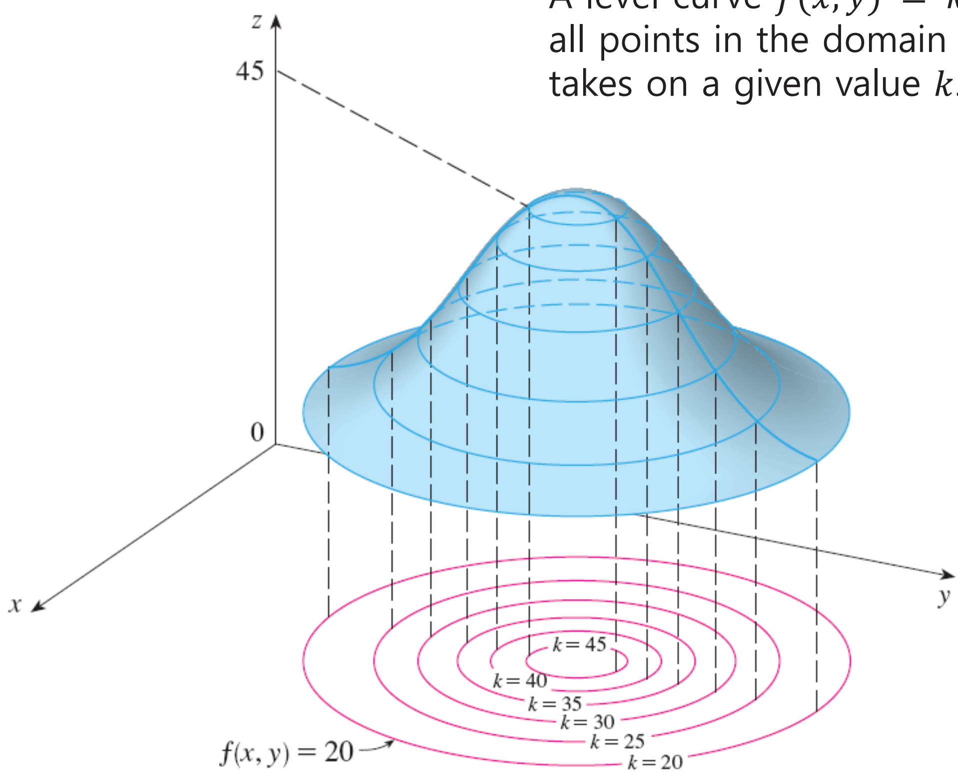
(b) $f(x, y) = (x^2 + 3y^2)e^{-x^2-y^2}$

(c) $f(x, y) = \sin x + \sin y$

(d) $f(x, y) = \dfrac{\sin x \sin y}{xy}$

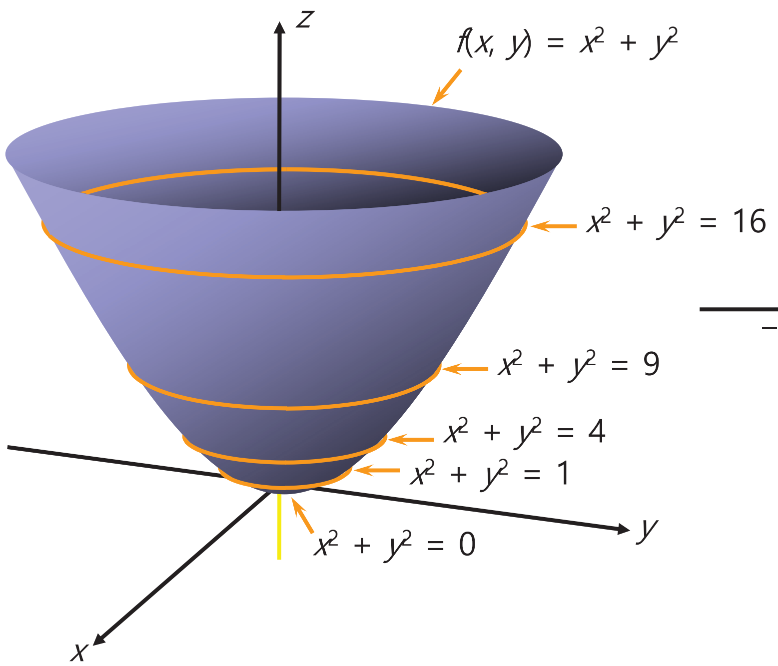
LEVEL CURVES

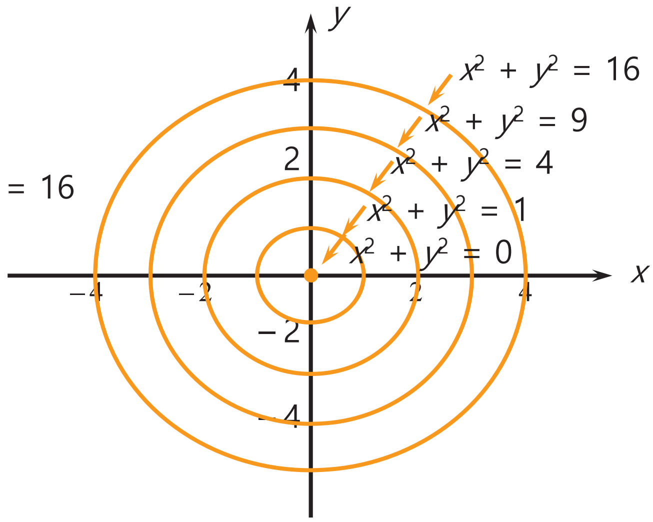
A level curve $f(x,y) = k$ is the set of all points in the domain of f at which f takes on a given value k.



Example: LEVEL CURVES

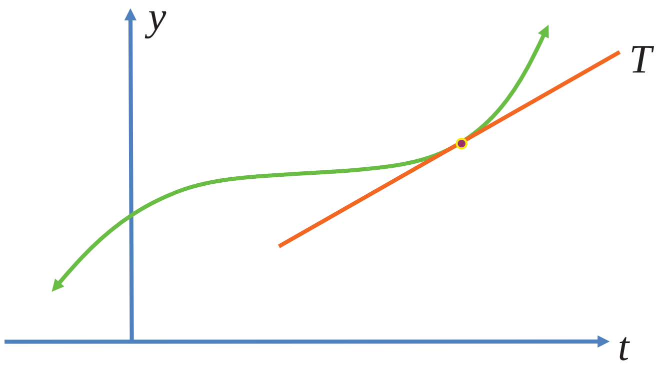
- Contour map of the function $f(x,y) = x^2 + y^2$

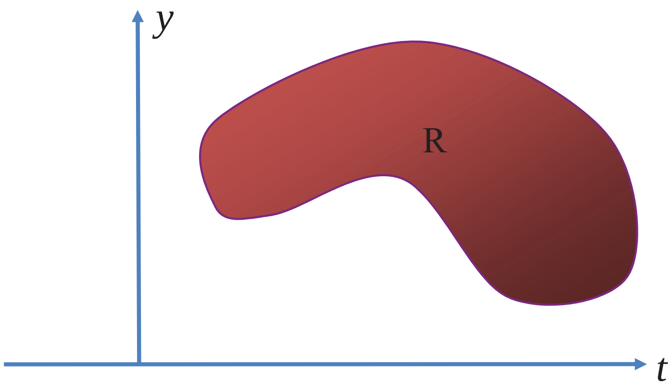




Introduction to Calculus

- Historically, the development of calculus by Isaac Newton and Gottfried W. Leibniz resulted from the investigation of the following problems:
  1. Finding the tangent line to a curve at a given point on the curve:
  2. Finding the area of planar region bounded by an arbitrary curve.



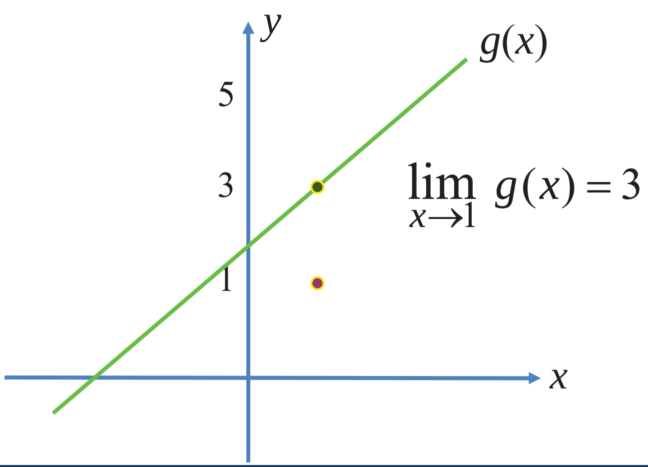


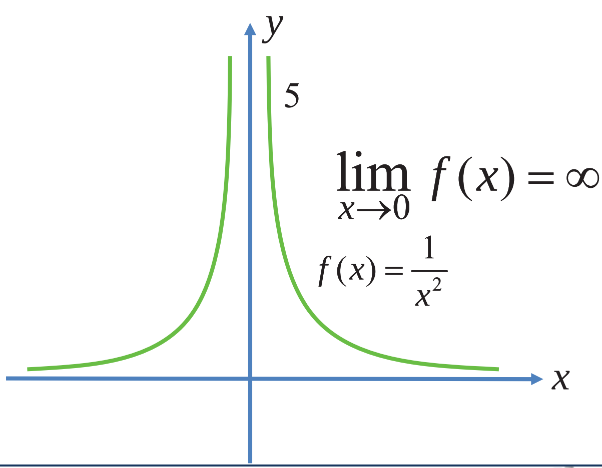
Limit of a Function

- The function f has a limit L as x approaches a, written

$$\lim_{x \to a} f(x) = L$$

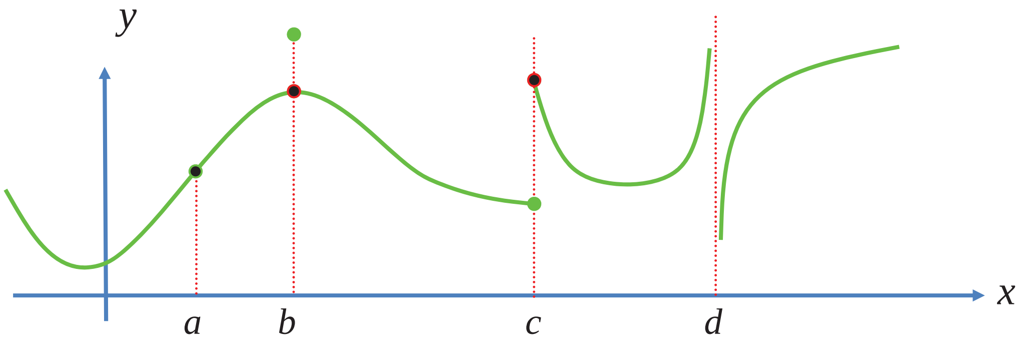
- If the value of $f(x)$ can be made as close to the number L as we please by taking x values sufficiently close to (but not equal to) a.





Continuous Functions

- Loosely speaking, a function is continuous at a given point if its graph at that point has no holes, gaps, jumps, or breaks.
- Consider, for example, the graph of f



- This function is discontinuous at the following points:
  - *At $x = a$, f is not defined ($x = a$ is not in the domain of f).*
  - *At $x = b$, $f(b)$ is not equal to the limit of $f(x)$ as x approaches b.*
  - *At $x = c$, the function does not have a limit, since the left-hand and right-hand limits are not equal.*
  - *At $x = d$, the limit of the function does not exist, resulting in a break in the graph.*

Continuity of a Function at a Number

- A function f is continuous at a number $x = a$ if the following conditions are satisfied:
  1. $f(a)$ is defined.
  2. $\lim_{x \to a} f(x)$ **exists.**
  3. $\lim_{x \to a} f(x) = f(a)$
- If f is not continuous at $x = a$, then f is said to be discontinuous at $x = a$.
- Also, f is continuous on an interval if f is continuous at every number in the interval.

Topology of the Euclidean space $\Re^n$

Limits

- We say that a sequence $\{x_k\}$ **converge** to some point $\hat{x} \in \Re^n$, written $\lim_{k\to\infty} x_k = \hat{x}$, if for any $\varepsilon > 0$, there is an index K such that

$$\| x_k - \hat{x} \| \le \varepsilon, \qquad for\ \ all\ k \ge K.$$

- We say that $\hat{x} \in \Re^n$ is a **limit point** for $\{x_k\}$

Continuity

- Let $f: D \subset \Re^n \to \Re^m$. We say that f is **continuous** at x_0 if for all $\varepsilon > 0$, there is a value $\delta > 0$ such that

$$\| x - x_0 \| < \delta \ \Rightarrow \| f(x_0) - f_0 \| < \varepsilon.$$

- For some point x_0, we write

$$\lim_{x \to x_0} f(x) = f_0$$

- The function f is said to be **Lipschitz continuous** if there is a constant $M > 0$ such that for any two points x_0, x_1 in D, we have

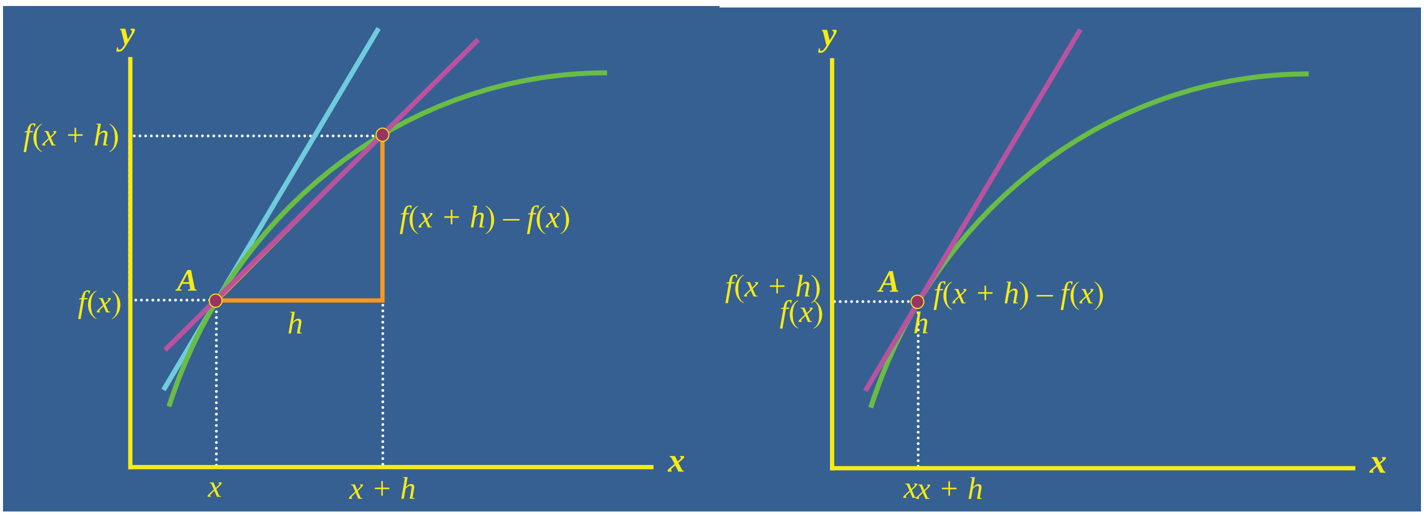
$$\| f(x_1) - f(x_0) \| \le M \| x_1 - x_0 \|.$$

Slope and Derivative

- In general, we can express the slope and the derivative as follows:

$$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{f(x+h) - f(x)}{(x+h) - x} = \frac{f(x+h) - f(x)}{h}$$

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



Derivatives

- Derivatives in $\Re$
  - Let $\phi: \Re \to \Re$ be a real-valued function of a real variable. The first derivative $\phi'(\alpha)$ is defined by

$$\frac{d\phi}{d\alpha} = \phi'(\alpha) = \lim_{\varepsilon \to 0} \frac{\phi(\alpha+\varepsilon)-\phi(\alpha)}{\varepsilon}.$$

  - The second derivative is obtained by substituting ϕ by ϕ' in this same formula;

$$\frac{d^2\phi}{d\alpha^2} = \phi''(\alpha) = \lim_{\varepsilon \to 0} \frac{\phi'(\alpha+\varepsilon)-\phi'(\alpha)}{\varepsilon}.$$

- Chain rule
  - Suppose now that α in turn depends on another quantity β. We can use the chain rule to calculate the derivative of ϕ with respect to β:

$$\frac{d\phi(\alpha(\beta))}{d\beta} = \frac{d\phi}{d\alpha}\frac{d\alpha}{d\beta} = \phi'(\alpha)\alpha'(\beta).$$

First Partial Derivatives

- First Partial Derivatives of $f(x, y)$
  - Suppose $f(x, y)$ is a function of two variables x and y.
  - Then, the first partial derivative of f with respect to x at the point (x, y) is

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}$$

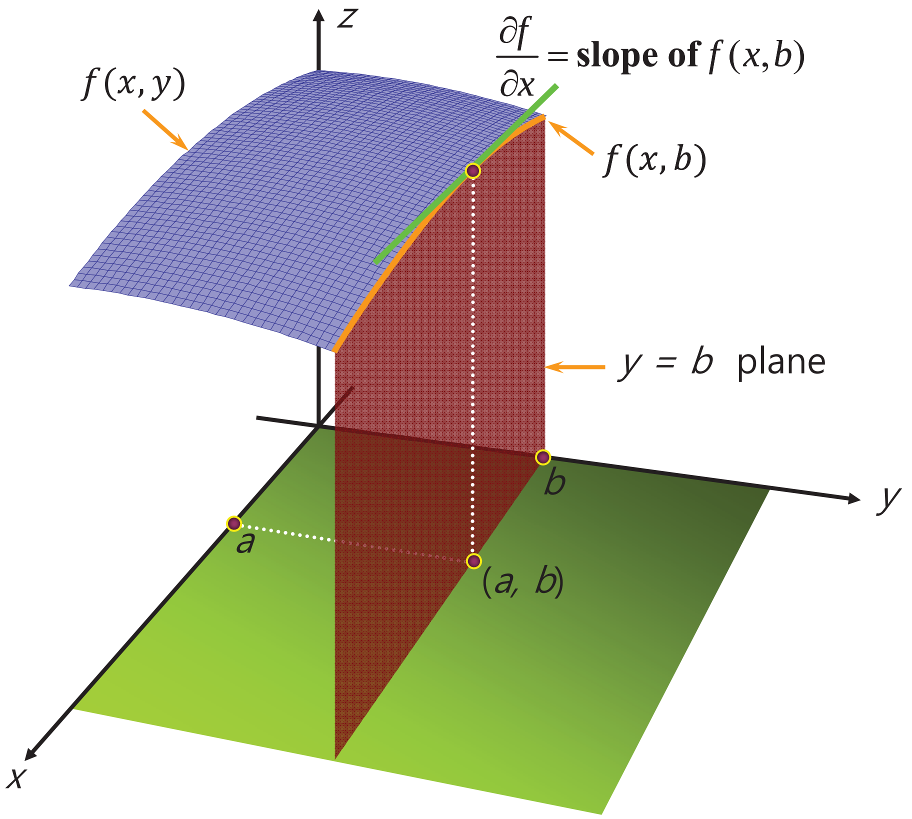
    provided the limit exists.
  - The first partial derivative of f with respect to y at the point (x, y) is

$$\frac{\partial f}{\partial y} = \lim_{k \to 0} \frac{f(x, y+k) - f(x, y)}{k}$$

    provided the limit exists.

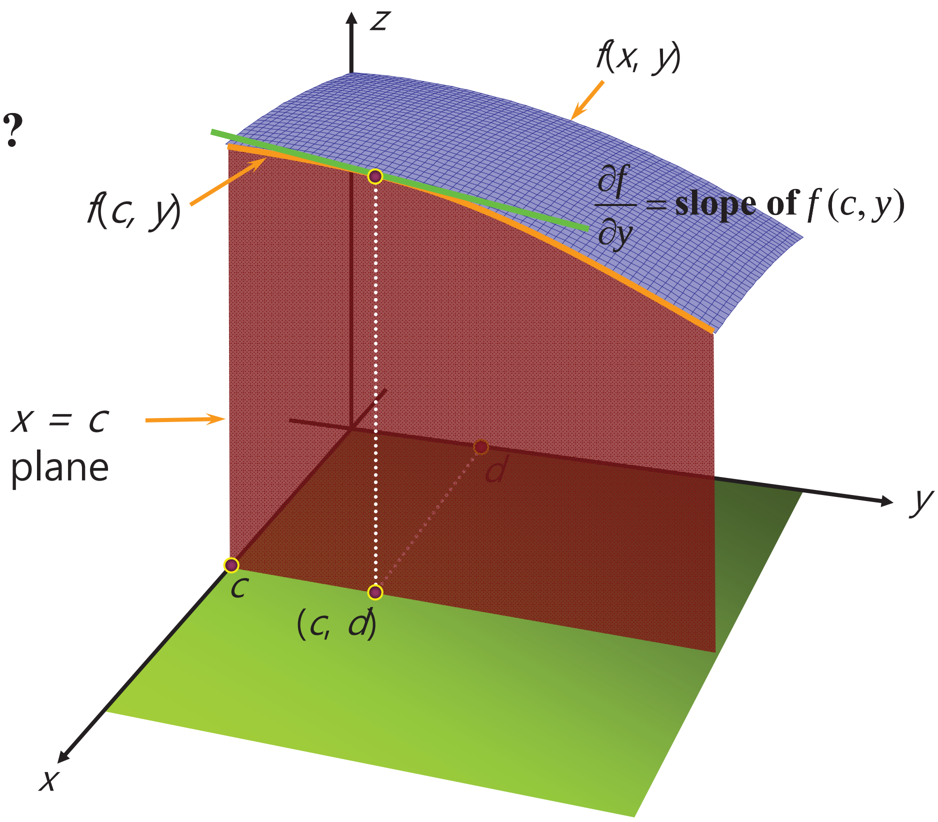
Geometric Interpretation of the Partial Derivative

What does $\dfrac{\partial f}{\partial x}$ mean?



Geometric Interpretation of the Partial Derivative

What does $\dfrac{\partial f}{\partial y}$ mean?



Example-1

- Find the partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$ of the function

$$f(x,y) = x^2 - xy^2 + y^3$$

- Use the partials to determine the rate of change of f in the x-direction and in the y-direction at the point (1, 2) .

Solution:

$$f(x,y) = x^2 - y^2x + y^3 \qquad \frac{\partial f}{\partial x} = 2x - y^2 \qquad \left.\frac{\partial f}{\partial x}\right|_{(1,2)} = 2(1) - 2^2 = -2$$

$$f(x,y) = x^2 - xy^2 + y^3 \qquad \frac{\partial f}{\partial y} = -2xy + 3y^2 \qquad \left.\frac{\partial f}{\partial y}\right|_{(1,2)} = -2(1)(2) + 3(2)^2 = 8$$

Example-2

- Find the first partial derivatives of the function

$$w(x,y) = \frac{xy}{x^2+y^2}$$

Solution:

$$w(x,y) = \frac{xy}{x^2+y^2} \qquad \frac{\partial w}{\partial x} = \frac{(x^2+y^2)y - xy(2x)}{(x^2+y^2)^2} = \frac{y(y^2-x^2)}{(x^2+y^2)^2}$$

$$w(x,y) = \frac{xy}{x^2+y^2} \qquad \frac{\partial w}{\partial y} = \frac{(x^2+y^2)x - xy(2y)}{(x^2+y^2)^2} = \frac{x(x^2-y^2)}{(x^2+y^2)^2}$$

Example-3

- Find the first partial derivatives of the function

$$g(s,t) = (s^2 - st + t^2)^5$$

Solution:

$$g(s,t) = (s^2 - st + t^2)^5$$

$$\frac{\partial g}{\partial s} = 5(s^2 - st + t^2)^4 \cdot (2s - t)$$
$$= 5(2s - t)(s^2 - st + t^2)^4$$

$$g(s,t) = (s^2 - st + t^2)^5$$

$$\frac{\partial g}{\partial t} = 5(s^2 - st + t^2)^4 \cdot (-s + 2t)$$
$$= 5(2t - s)(s^2 - st + t^2)^4$$

Examples

- Find the first partial derivatives of the function

$$h(u,v) = e^{u^2 - v^2}$$

Solution:

$$h(u,v) = e^{u^2 - v^2}$$

$$\frac{\partial h}{\partial u} = e^{u^2 - v^2} \cdot 2u = 2ue^{u^2 - v^2}$$

$$h(u,v) = e^{u^2 - v^2}$$

$$\frac{\partial h}{\partial u} = e^{u^2 - v^2} \cdot (-2v) = -2ve^{u^2 - v^2}$$

Example-4

- Find the first partial derivatives of the function

$$w = f(x,y,z) = xyz - xe^{yz} + x\ln y$$

Solution

- Here we have a function of three variables, *x*, *y*, and *z*, and we are required to compute

$$f_x \equiv \frac{\partial f}{\partial x},\quad f_y \equiv \frac{\partial f}{\partial y},\quad f_z \equiv \frac{\partial f}{\partial z}$$

$$w = f(x,y,z) = xyz - xe^{yz} + x\ln y \qquad f_x = yz - e^{yz} + \ln y$$

$$w = f(x,y,z) = xyz - xe^{yz} + x\ln y \qquad f_y = xz - xze^{yz} + \frac{x}{y}$$

$$w = f(x,y,z) = xyz - xe^{yz} + x\ln y \qquad f_z = xy - xye^{yz}$$

Second Order Partial Derivatives

- Differentiating the function f_x with respect to x leads to the second partial derivative

$$f_{xx} \equiv \frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(f_x)$$

- But the function f_x can also be differentiated with respect to y leading to a different second partial derivative

$$f_{xy} \equiv \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(f_x)$$

- Similarly, differentiating the function f_y with respect to y leads to the second partial derivative

$$f_{yy} \equiv \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(f_y)$$

- Finally, the function f_y can also be differentiated with respect to x leading to the second partial derivative

$$f_{yx} \equiv \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(f_y)$$

Second Order Partial Derivatives

- Thus, four second-order partial derivatives can be obtained of a function of two variables:

$$f \xrightarrow{\frac{\partial}{\partial x}} \frac{\partial f}{\partial x} \xrightarrow{\frac{\partial}{\partial x}} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}$$

$$\frac{\partial f}{\partial x} \xrightarrow{\frac{\partial}{\partial y}} \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x}$$

$$f \xrightarrow{\frac{\partial}{\partial y}} \frac{\partial f}{\partial y} \xrightarrow{\frac{\partial}{\partial x}} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x \partial y}$$

$$\frac{\partial f}{\partial y} \xrightarrow{\frac{\partial}{\partial y}} \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2}$$

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$$

When both are continuous

Example-1

- Find the second-order partial derivatives of the function

$$f(x, y) = x^3 - 3x^2 y + 3xy^2 + y^2$$

Solution:

$$f_x = \frac{\partial}{\partial x}(x^3 - 3x^2 y + 3xy^2 + y^2) = 3x^2 - 6xy + 3y^2$$

$$f_{xx} = \frac{\partial}{\partial x}(3x^2 - 6xy + 3y^2) = 6x - 6y$$

$$f_{xy} = \frac{\partial}{\partial y}(3x^2 - 6xy + 3y^2) = -6x + 6y$$

$$f_y = \frac{\partial}{\partial y}(x^3 - 3x^2 y + 3xy^2 + y^2) = -3x^2 + 6xy + 2y$$

$$f_{yy} = \frac{\partial}{\partial y}(-3x^2 + 6xy + 2y) = 6x + 2$$

$$f_{yx} = \frac{\partial}{\partial x}(-3x^2 + 6xy + 2y) = -6x + 6y$$

Examples

- Find the second-order partial derivatives of the function

$$f(x,y)=e^{xy^2}$$

Solution:

$$f_x=\frac{\partial}{\partial x}(e^{xy^2})=y^2e^{xy^2}$$

$$f_{xx}=\frac{\partial}{\partial x}(y^2e^{xy^2})=y^4e^{xy^2}$$

$$f_{xy}=\frac{\partial}{\partial y}(y^2e^{xy^2})=2ye^{xy^2}+2xy^3e^{xy^2}$$

$$f_y=\frac{\partial}{\partial y}(e^{xy^2})=2xye^{xy^2}$$

$$f_{yy}=\frac{\partial}{\partial y}(2xye^{xy^2})=2xe^{xy^2}+(2xy)(2xy)e^{xy^2}$$

$$f_{yx}=\frac{\partial}{\partial x}(2xye^{xy^2})=2ye^{xy^2}+2xy^3e^{xy^2}$$

Derivatives

- Partial derivative
  - Let $f:\Re^n\rightarrow\Re$. Each partial derivative $\partial f/\partial x_i$ measures the sensitivity of the function to just one of the components of x; that is,

$$\frac{\partial f}{\partial x_i}\overset{def}{=}\lim_{\varepsilon\to 0}\frac{f(x_1,\dots,x_i+\varepsilon,\dots,x_n)-f(x_1,\dots,x_n)}{\varepsilon}=\frac{f(x+\varepsilon e_i)-f(x)}{\varepsilon}$$

  - where e_i is the vector $(0,\dots,0,1,0,\dots,0)^t$, where the 1 appears in the ith position.
- Gradient & Hessian
  - $\nabla f(x)$ = the gradient of f : The first derivatives of f
  - $\nabla^2 f(x)$ = the Hessian of f : The matrix of second partial derivatives of f

$$\nabla f(x)\overset{def}{=}\begin{bmatrix}\frac{\partial f}{\partial x_1}\\ \vdots\\ \frac{\partial f}{\partial x_n}\end{bmatrix},\qquad \nabla^2 f(x)=\begin{bmatrix}\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n}\\ \frac{\partial^2 f}{\partial x_2\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 f}{\partial x_n\partial x_1} & \frac{\partial^2 f}{\partial x_n\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}\end{bmatrix}$$

Derivatives

Differentiable

- We say that f is **differentiable** if all first partial derivatives of f exist, and **continuously differentiable** if in addition these derivatives are continuous functions of x.
- Similarly, f is **twice differentiable** if all second partial derivatives of f exist and **twice continuously differentiable** if they are also continuous. Note that when f is twice continuously differentiable, the Hessian is a symmetric matrix, since

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}, \qquad \text{for all} \ \ i, j = 1, 2, \ldots, n.$$

- When the vector x in turn depends on another vector t, the chain rule for the univariate function can be extended as follows:

$$\frac{\partial f(x(t))}{\partial t} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} = \nabla f^T \frac{dx(t)}{dt}$$

Directional Derivatives

Directional Derivatives

- If f is continuously differentiable and $p \in \Re^n$, then the **directional derivative** of f in the direction p is given by

$$D(f(x); p) = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon p) - f(x)}{\varepsilon} = \nabla f(x)^T p.$$

- To verify this formula, we define the function

$$\phi(\alpha) = f(x + \alpha p) = f(y(\alpha)), \quad \text{where} \ \ y(\alpha) = x + \alpha p.$$

- Note that

$$\lim_{\varepsilon \to 0} \frac{f(x + \varepsilon p) - f(x)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{\phi(\varepsilon) - \phi(0)}{\varepsilon} = \phi'(0).$$

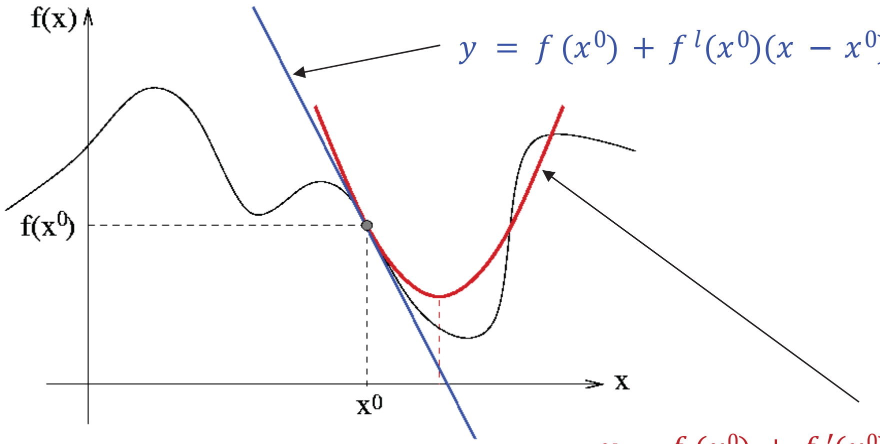
- By applying the chain rule to $f(y(\alpha))$, we obtain

$$\phi'(\alpha) = \sum_{i=1}^{n} \frac{\partial f(y(\alpha))}{\partial y_i} \nabla y_i(\alpha) = \sum_{i=1}^{n} \frac{\partial f(y(\alpha))}{\partial y_i} p_i = \nabla f(y(\alpha))^T p = \nabla f(x + \alpha p)^T p.$$

Taylor expansion in R

- For $f : R \to R$ that is C^2 (*i.e.* has a cont. 2nd order derivative), one can approximate f around any point x^0 by

$$f(x) = f(x^0) + f'(x^0)(x - x^0) + \frac{1}{2} f''(x^0)(x - x^0)^2 + o(x - x^0)^2$$



$y = f(x^0) + f'(x^0)(x - x^0)$: tangent line

$y = f(x^0) + f'(x^0)(x - x^0) + \frac{1}{2} f''(x^0)(x - x^0)^2$
: tg parabola → Can be used to find a local minimum of f.

Taylor's Theorem

- Let $f: \Re^n \to \Re$ be twice differentiable.
- First order Taylor's Theorem

$$f(x+p) \cong f(x) + \nabla f(x)^T p$$

- Second order Taylor's Theorem

$$f(x+p) \cong f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p$$

- Mean Value Theorem

$$f(x+p) = f(x) + \nabla f(x + \alpha p)^T p$$

$$f(x+p) = f(x) + \nabla f(x + \lambda p)^T p + \frac{1}{2} p^T \nabla^2 f(x + \lambda p) p,$$

- for some $\alpha \in (0,1)$ and $\lambda \in (0,1)$.

Dimensions

	Scalar	Vector	Matrix
Scalar	$\frac{\mathrm{d}y}{\mathrm{d}x}$	$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}x} = \left[\frac{\partial y_i}{\partial x}\right]$	$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}x} = \left[\frac{\partial y_{ij}}{\partial x}\right]$
Vector	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial y}{\partial x_j}\right]$	$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial y_i}{\partial x_j}\right]$	
Matrix	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \left[\frac{\partial y}{\partial x_{ji}}\right]$		

By Thomas Minka. Old and New Matrix Algebra Useful for Statistics

Examples

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T$$

$$\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B}+\mathbf{B}^T)\mathbf{x}$$

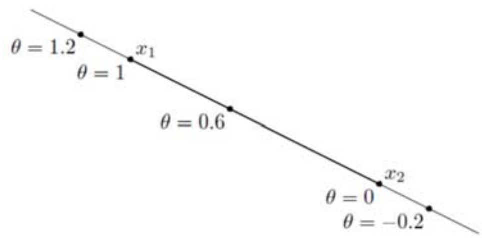
http://matrixcookbook.com/

CONVEXITY

Convex Sets

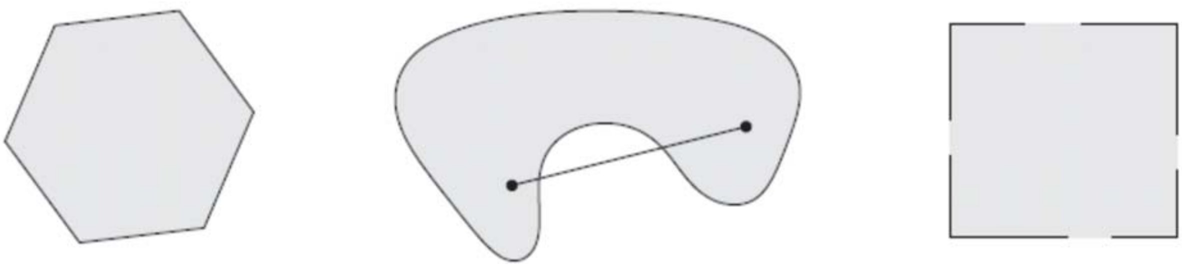
- Convex Sets
  - The convex combination of two points is the line segment between them

$$\theta x_1 + (1-\theta)x_2 \in C, \quad \theta \in [0,1]$$



  - A set C in $\Re^n$ is said to be convex, if for any $x_1, x_2 \in C$, we have

$$\lambda x_1 + (1-\lambda)x_2 \in C, \qquad \forall \lambda \in [0,1]$$

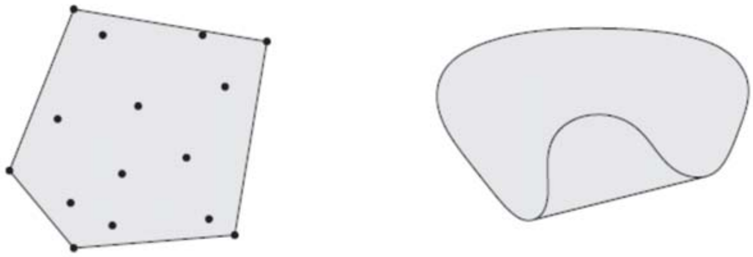


Convex Sets

Convex hull

- The convex hull of C, denoted by $\mathrm{conv}(C)$, is the collection of all convex combinations of a set $C \in \Re^n$, i.e.

$$\mathrm{conv}(C) = \{\textstyle\sum_{i=1}^{k} \lambda_i x_i \mid \sum_{i=1}^{k} \lambda_i = 1,\ \lambda_i \geq 0,\ i = 1, \ldots, k\}$$



- A subset of $\Re^n$ is convex iff it contains all the convex combinations of its elements. The smallest convex set containing a set $C \in \Re^n$ is $\mathrm{conv}(C)$. Indeed, $\mathrm{conv}(C)$ is the intersection of all convex sets containing C.

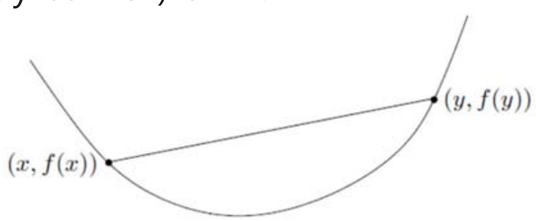
Convex Functions

Convex Functions

- Let $f: C \to \Re$, where C is a nonempty convex set in $\Re^n$. The function f is said to be convex on C if $x_1, x_2 \in C$ with $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

- The function $f: C \to \Re$ is called concave (strictly concave) on C if $(-f)$ is convex (strictly convex) on C.



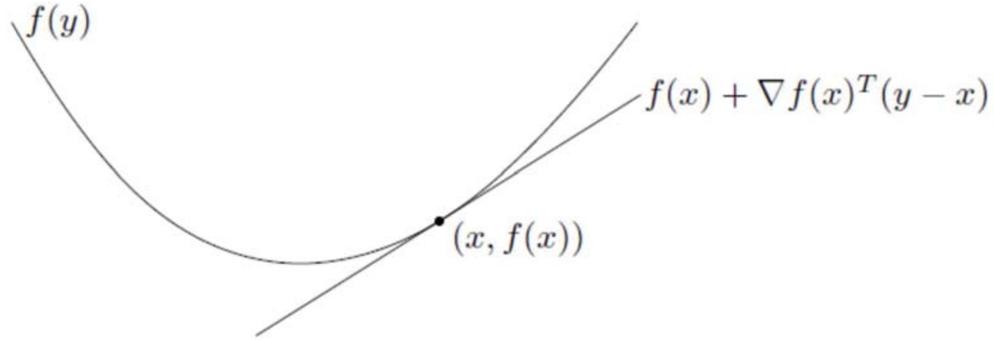
- [Jensen's inequality] For a convex function f,

$$f\left(\textstyle\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i), \quad \text{where } \sum_{i=1}^{k} \lambda_i = 1,\ \lambda_i \geq 0\ \forall i$$

- In probability, it says $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$ for a convex function f.

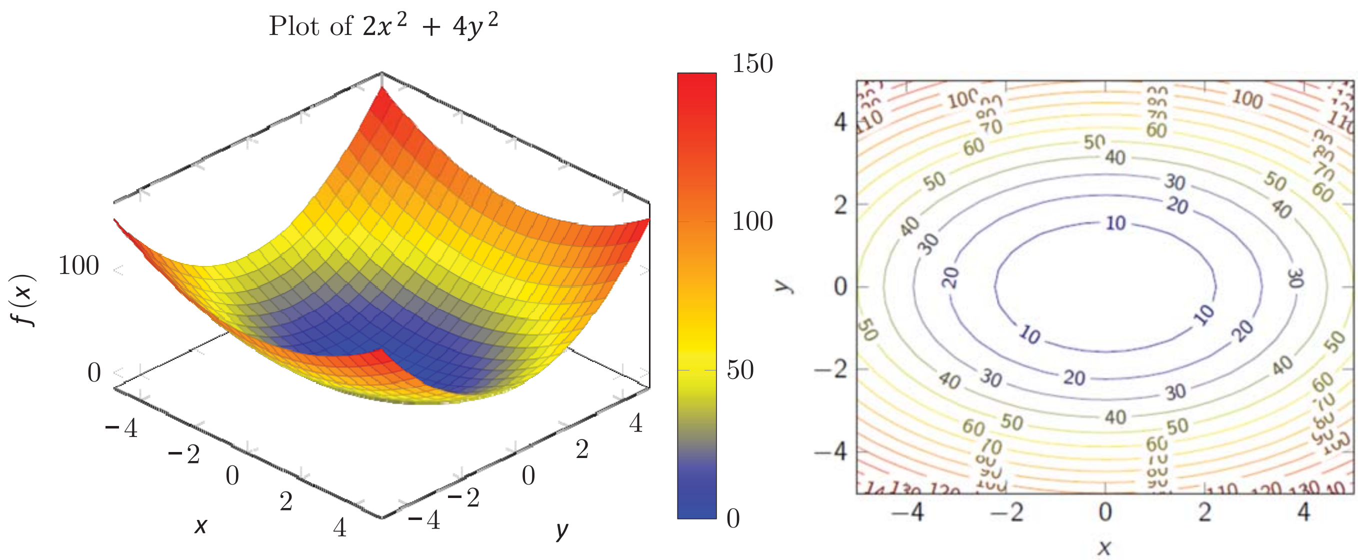
Convex Functions

- [First-order conditions]
  - Let C be a nonempty open convex set in $\Re^n$, and let $f: C \to \Re$ be differentiable on C. Then f is convex if and only if for any $x, y \in C$, we have
$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$
  - f is strictly convex if and only if for any $x, y \in C$, we have
$$f(y) > f(x) + \nabla f(x)^T (y - x).$$



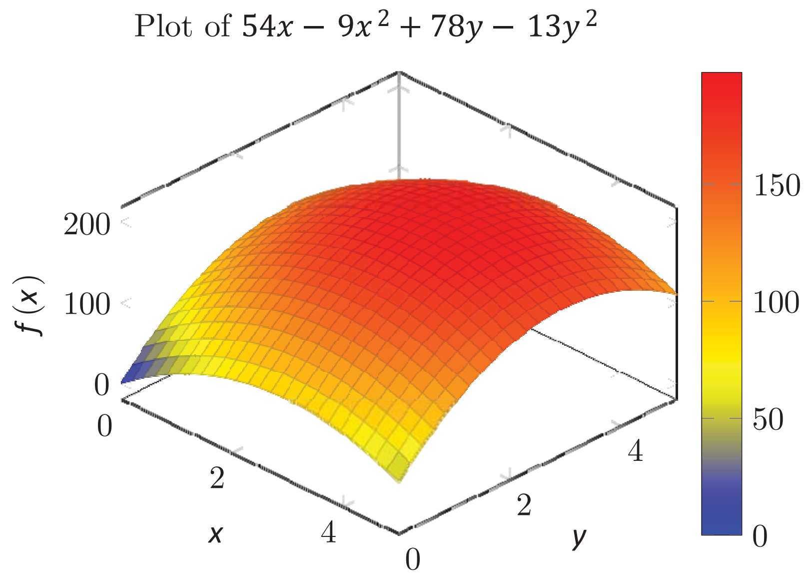
Level set

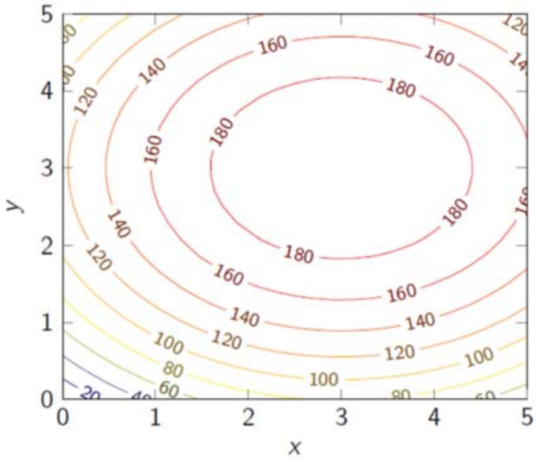
- Lower Level set: $\{\boldsymbol{x} : f(\boldsymbol{x}) \leq \beta\}$



Level set

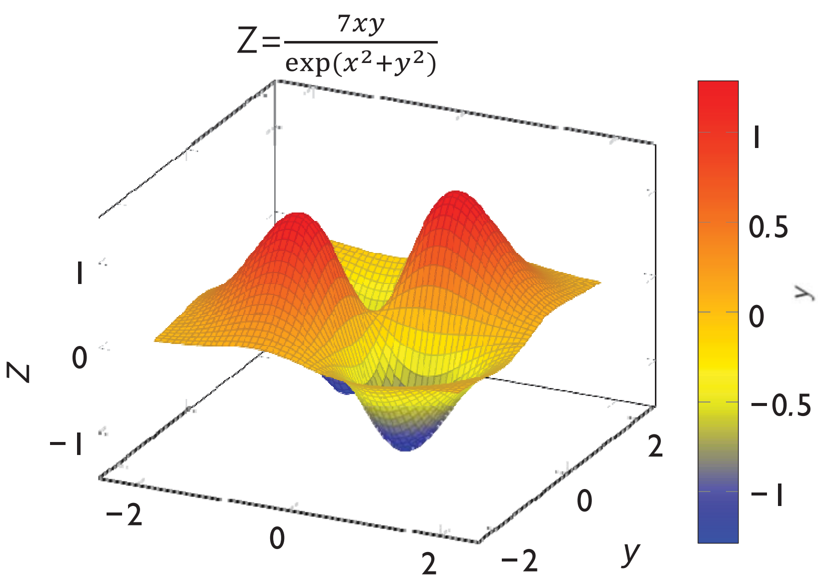
- Upper Level set: $\{\boldsymbol{x} : f(\boldsymbol{x}) \geq \beta\}$

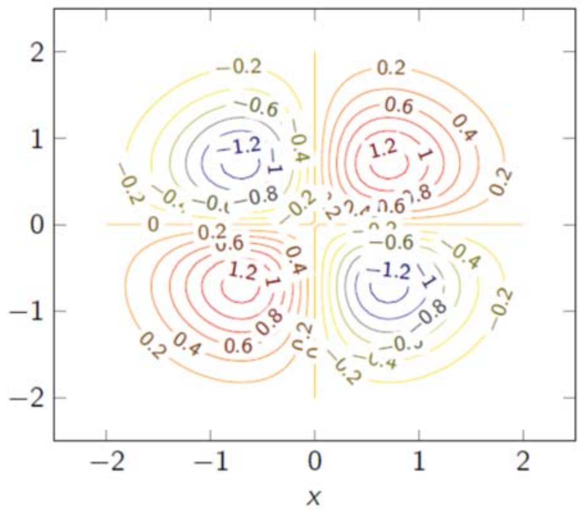




Contour set

- Example





Convexity, Concavity, and Optima

- Theorem
  - Suppose that C is convex and that $f(x)$ is convex on C for the problem

$$\min_{x \in C} f(x)$$

  - If x^* is locally minimal, then x^* is globally minimal.

- [Second-order conditions]
  - Let C be a nonempty open convex set in $\Re^n$, and let $f: C \to \Re$ be twice differentiable on C. Then
  - f is convex if and only if its Hessian matrix $\nabla^2 f(x)$ is positive semidefinite
  - If the Hessian matrix is positive definite, then f is strictly convex.
  - If f is strictly convex and quadratic, then its Hessian matrix is positive definite at each point in C.

- Example 1

$$\begin{aligned} &\text{Minimize} && -x_2 \ln x_1 + \frac{x_1}{9} + x_2^2 \\ &\text{Subject to:} && 1.0 \leq x_1 \leq 5.0 \\ & && 0.6 \leq x_2 \leq 3.6 \end{aligned}$$

Plot of $-x_2 \ln x_1 + \frac{x_1}{9} + x_2^2$