BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES

COURSE HANDOUT

Part A: Content Design

Course Title	Mathematical Foundations for Machine Learning
Course No(s)	AIMLC ZC416
Credit Units	4
Course Author	Srinath Naidu
Version No	3
Date	01.08.2022
Lead Instructor	Srinath Naidu

Course Description

Vector and matrix algebra, systems of linear algebraic equations and their solutions; Eigenvalues, eigenvectors and diagonalization of matrices; multivariate calculus, vector calculus, Jacobian and Hessian, multivariate Taylor series, gradient descent, unconstrained optimization, constrained optimization, nonlinear optimization, stochastic gradient descent, dimensionality reduction and PCA, optimization for support vector machines.

Course Objectives

No	Objective- The course aims to
CO1	Introduce concepts in linear algebra and to use it as a platform for machine learning.
CO2	Provide techniques for analytical and numerical solutions of linear equations.
CO3	Introduce concepts in multivariate calculus and vector calculus, and introduce multivariate Taylor series.
CO4	Introduce concepts of gradient descent, stochastic gradient descent and the concepts of unconstrained, constrained and nonlinear optimization.
CO5	Introduce the concepts of dimensionality reduction, principal components analysis, and optimization for support vector machines.

Text Book(s)

No	Author(s), Title, Edition, Publishing House
T1	M.P. Diesenroth, A. Aldo Faisal, Cheng Soon Ong, Mathematics for Machine Learning, Cambridge University Press, 2020.
T2	Charu C. Aggarwal, Linear Algebra and Optimization, Springer Nature Switzerland AG, 2020

Reference Book(s) & other resources

No	Author(s), Title, Edition, Publishing House
R1	K Hoffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.

R2	Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 10th Edition, 2015 (earlier editions are also okay)

Content Structure

No	Title of the module	References
M1	Solution of linear systems – systems of linear equations, matrices, solving systems of linear equations.	T1: Sec 2.1, 2.2, 2.3
M2	Vectors Spaces - linear independence, basis and rank, affine spaces, Norms, inner products, Lengths and distances, Angles and orthogonality, Orthonormal basis	T1: Sec 2.4, 2.5, 2.6, 2.7, 2.8, 3.1, 3.2, 3.3, 3.4, 3.5
M3	Matrix Decomposition methods - Determinant and Trace, Eigenvalues and Eigenvectors, Cholesky decomposition, Eigen-decomposition and diagonalization, singular value decomposition, matrix approximation	T1: Sec 4.1, 4,2, 4.3, 4.4, 4.5, 4.6
M4	Vector Calculus - Differentiation of univariate functions, Partial differentiation and gradients, Gradients of vector-valued functions , Gradients of matrices, Some useful identities for computing gradients, Backpropagation and automatic differentiation	T1: Sec 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8
M5	Continuous Optimization - Optimization using gradient descent, Constrained optimization and Lagrange multipliers, Convex optimization	T1: 7.1, 7.2, 7.3
M6	Nonlinear Optimization - Minutiae of Gradient Descent – learning rate decay, initialization, Properties of optimization in learning – typical objective functions, stochastic gradient descent, how optimization in machine learning is different, tuning hyperparameters, importance of feature pre-processing, Challenges in Gradient-based optimization, local optima and flat regions, differential curvature, examples of difficult topologies like cliffs and valleys, adjusting first-order derivatives for descent, momentum-based learning, AdaGrad, RMSProp, Adam	T2: 4.4, 4.5, 5.2, 5.3
M7	Dimensionality reduction and PCA – problem setting, maximum variance perspective, projection perspective, eigenvector and low-rank approximations, PCA in high dimensions, key steps of PCA in practice, latent variable perspective, Mathematical preliminaries of SVM, primal/dual perspective for SVM, nonlinear SVM - kernels.	T1: 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, 10.7, 12.1, 12.2, 12.3, 12.4, 12.5, 12.6, T2: Sec 6.4 Class Notes

Learning Outcomes:

No	Learning Outcomes
LO1	Students will be able to effectively use matrix algebra tools to analyze and solve systems of linear equations.
LO2	Students will be able to learn concepts of linear algebra that form the foundation of data science problems.

LO3	Students would be able to learn the basic techniques of constrained and unconstrained optimization and understand how these are applied in the context of support vector machines.
LO4	Students will be able to learn the concepts of dimensionality reduction, and principal components analysis.

Part B: Contact Session Plan

Academic Term	Second Semester 2022-2023
Course Title	Mathematical Foundations for Machine Learning
Course No	AIMLC ZC416
Lead Instructor	Srinath Naidu

Course Contents

Contact Hours	List of Topic Title	Text/Ref Book/external resource
1	Solution of linear systems – systems of linear equations, matrices, solving systems of linear equations	T1: Sec 2.1, 2.2, 2.3
2	Vectors Spaces, linear independence, basis and rank, affine spaces	T1: Sec 2.4, 2.5, 2.6, 2.8
3	Analytic Geometry – norms, inner products, lengths and distances, angles and orthogonality, orthonormal basis	T1: Sec 3.1, 3.2, 3.3, 3.4, 3.5
4	Matrix Decomposition – I Determinant and Trace Eigenvalues and Eigenvectors Cholesky Decomposition	T1: Sec 4.1, 4.2, 4.3

5	Matrix Decomposition – II Eigen-decomposition and Diagonalization Singular Value Decomposition Matrix Approximation	T1: Sec 4.4, 4.5, 4.6
6	Vector Calculus – I Differentiation of univariate functions Partial differentiation and gradients Gradients of vector-valued functions	T1: Sec 5.1, 5.2, 5.3
7	Vector Calculus – II Gradients of matrices Some useful identities for computing gradients Backpropagation and automatic differentiation	T1: Sec 5.4, 5.5, 5.6
8	Vector Calculus – III Higher-order derivatives Linearization and multivariate Taylor's series Computing maxima and minima for unconstrained optimization	T1: 5.7, 5.8 Class Notes
9	Continuous Optimization Optimization using gradient descent Constrained optimization and Lagrange multipliers Convex optimization	T1: Sec 7.1, 7.2, 7.3,
10	Nonlinear Optimization- I Minutiae of Gradient Descent – learning rate decay, initialization Properties of optimization in learning – typical objective functions, stochastic gradient descent, how optimization in machine learning is different, tuning hyperparameters, importance of feature pre-processing	T2: Sec 4.4, Sec 4.5
11	Nonlinear Optimization- II Challenges in Gradient-based optimization, local optima and flat regions, differential curvature, examples of difficult topologies like cliffs and valleys. Adjusting first-order derivatives for descent, Momentum-based learning, AdaGrad, RMSProp, Adam	T2: Sec 5.2, 5.3

12	Dimensionality Reduction and PCA – I Problem setting Maximum variance perspective Projection perspective	T1: Sec 10.1, 10.2, 10.3
13	Dimensionality Reduction and PCA – II Eigenvector Computation and low-rank approximation PCA in high dimensions Key steps of PCA in practice Latent variable perspective	T1: 10.4, 10.5, 10.6, 10.7
14	Mathematical preliminaries for SVM Karash-Kuhn-Tucker conditions	T2: Sec 6.4
15	Primal/dual perspective for linear SVM	T1: 12.1, 12.2, 12.3, 12.4
16	Nonlinear SVM (Expert lecture) kernels examples	Class Notes

 # The above contact hours and topics can be adapted for non-specific and specific WILP programs depending on the requirements and class interests.

Lab Details

Title	Access URL
Lab Setup Instructions	Not applicable
Lab Capsules	Not applicable
Additional References	Not applicable

Select Topics and Case Studies from business for experiential learning

Topic No.	Select Topics in Syllabus for experiential learning	Access URL
1	Assignment - linear algebra topics	
2	Assignment- Vector Calculus/Optimization based topics	

Evaluation Scheme

Legend: EC = Evaluation Component

No	Name	Type	Duration	Weight	Day, Date, Session, Time
1	Assignment 1	Online		10%	To be announced
2	Assignment 2	Online		10%	To be announced
3	Quiz 1	Online	*	5%	To be announced
4	Quiz 2	Online	*	5%	To be announced
5	Mid-Semester Exam	Closed book	90 min	30%	To be announced
6	Comprehensive Exam	Open book	150 min	40%	To be announced

Important Information

Syllabus for Mid-Semester Test (Closed Book): Topics in Weeks 1-8

Syllabus for Comprehensive Exam (Open Book): All topics (in sessions 1 to 16) given in plan of study

Evaluation Guidelines:

1. EC-1 consists of two Assignments and two Quizzes (best two out of the three would be taken for grading). Announcements regarding the same will be made in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
3. For Open Book exams: Use of prescribed and reference text books, in original (not photocopies) is permitted. Class notes/slides as reference material in filed or bound form is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.

4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam. The genuineness of the reason for absence in the Regular Exam shall be assessed prior to giving permission to appear for the Make-up Exam. Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course handout, attend the lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.

Lecture 1

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

What is linear algebra?

- Linear algebra is the study of vectors and rules to manipulate vectors.
- Vectors are not only the familiar geometric vectors from high school (points in 2D/3D space) but any special objects which can be added together and multiplied by scalar values to produce another object of the same kind. For example, polynomials can also be treated as vectors.
- We shall deal with vectors in the space $\mathbb{R}^n$

Idea of closure

- Let's say we have a bunch of mathematical objects and we perform some operations on them. Do we get back similar objects?
- This leads to the idea of a vector space which underlies much of machine learning.

Systems of linear equations

- Systems of linear equations form a central part of linear algebra.
- Many problems can be formulated as systems of linear equations.
- Tools of linear algebra can be used to solve such problems.

Motivating problem

Consider the following problem A company produces products $N_1, N_2, ...N_n$ for which resources $R_1, R_2, ...R_m$ are required. To produce a unit of product N_i , a_{ij} units of resource R_i are needed, where $1 \leq i \leq n, 1 \leq j \leq m$. Find an optimal production plan where x_j units of product N_j are produced if a total b_i units of resource R_i are available, and no resources are left over.

Formulation

If we produce $x_1, x_2, \ldots x_n$ units of the products $N_1, N_2 \ldots N_n$ we need a total of $a_{i1}x_1 + a_{i2}x_2 + a_{in}x_n$ units of resource R_i. Thus we set up the equation:

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = b_i$$

We can similarly set up the following set of linear equations in n unknowns, $x_1, x_2 \ldots x_n$.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

Does a linear system always have solutions?

- A linear system has zero, one or infinitely many solutions
- Linear regression, a Machine Learning technique, provides an approximate solution to an overconstrained linear system, i.e one with no solution

Example

Consider the folllowing system of linear equations

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 + 2x_3 &= 2 \\ 2x_1 + 3x_3 &= 1 \end{aligned}$$

Adding the first and second equations gives $2x_1 + 3x_3 = 5$ which contradicts the third equation. Thus there is no set of values for the variables x_1, x_2, x_3 such that the equations above are simultaneously satisfied.

Modified example

Consider a slightly modified example

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 + 2x_3 &= 2 \\ x_2 + x_3 &= 2 \end{aligned}$$

In this case we can see from the first and third equations that $x_1 = 1$. Substituting this value of x_1 into equation (2), we get $-x_2 + 2x_3 = 1$. Adding this equation to equation (3), we get $3x_3 = 3$ which means $x_3 = 1$. Substituting $x_3 = 1$ into equation (3) shows $x_2 = 1$, so the overall solution is $x_1 = x_2 = x_3 = 1$. This is the unique solution to the problem

Infinite solutions

Now consider another modification to the original set of equations

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 + 2x_3 &= 2 \\ 2x_1 + 3x_3 &= 5 \end{aligned}$$

Adding the first and second equations gives $2x_1 + 3x_3 = 5$ which is the same as the third equation. Thus the solution to the three equations is any tuple x_1, x_2, x_3 which satisfies $2x_1 + 3x_3 = 5$, and there are infinite solutions. We now express these solutions in a way whose motivation will become clear later: adding equations (1) and (2) above we get $2x_1 = 5 - 3x_3$.

Infinite solutions

- Subtracting equation (2) from (1) we get $2x_2 - x_3 = 1$, so we can write

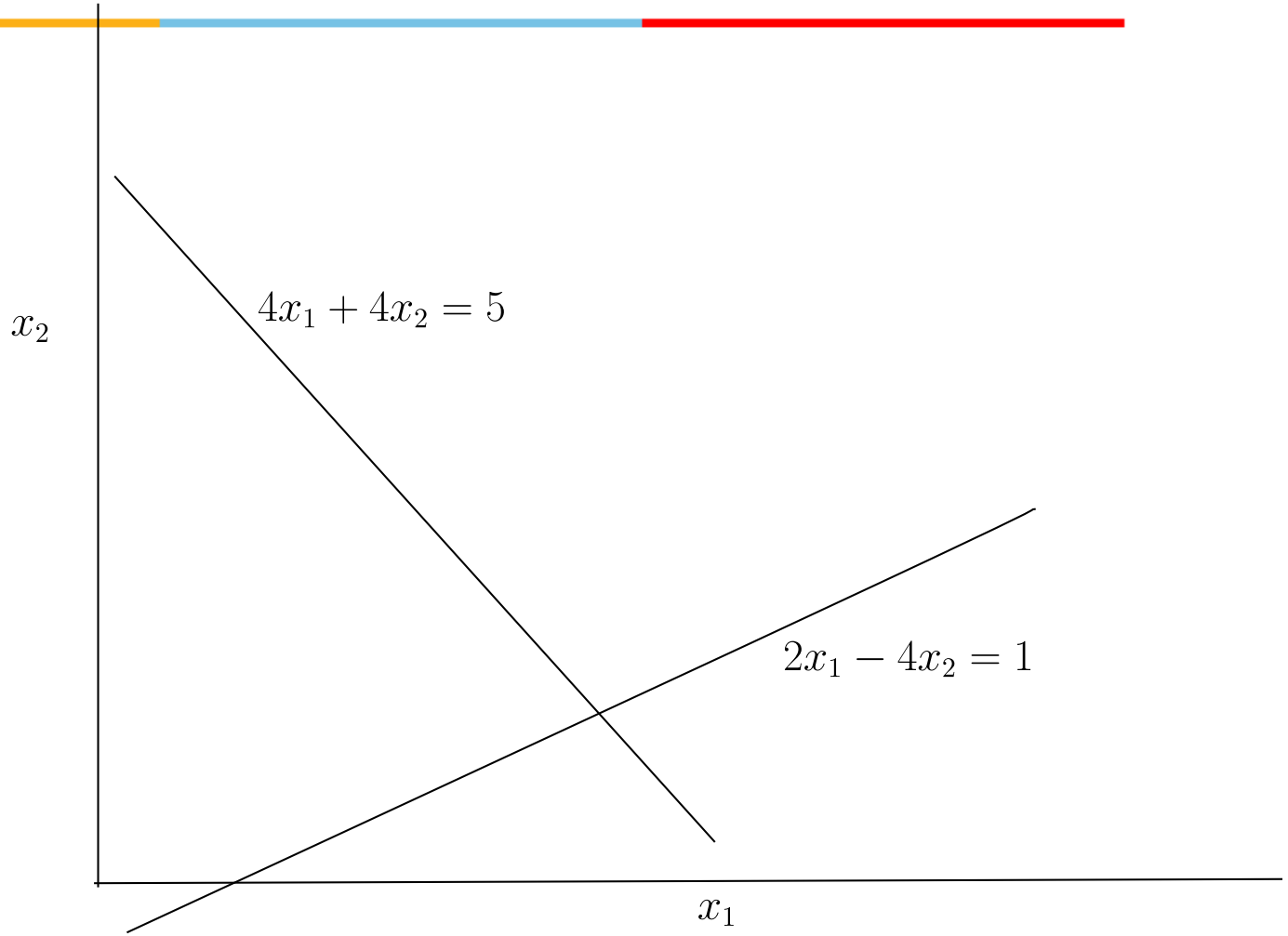
$$\begin{aligned} x_1 &= \frac{5}{2} - \frac{3}{2}x_3 \\ x_2 &= \frac{1}{2} + \frac{x_3}{2} \end{aligned}$$

- For the previous problem we can express the set of infinite solutions in terms of the free variable x_3.
- Once x_3 is fixed, the other two variables have to take on specific values - they are known as pivot variables.
- We will show later how to identify pivot and free variables using Gaussian Elimination

Geometrical interpretation

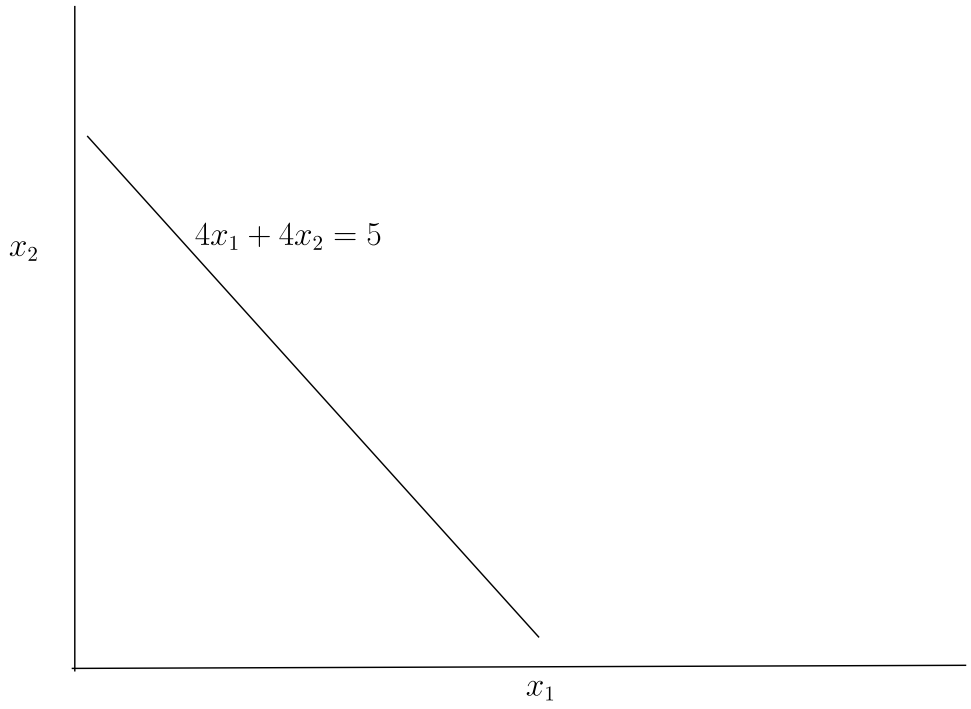
The solution to linear equations can be given a geometrical interpretation. When equations are given in terms of two variables, the solution to two equations in two variables could be a point (unique solution), a line (infinite solutions) or no solution (parallel lines). The first case is shown in the next slide.

Geometrical Interpretation



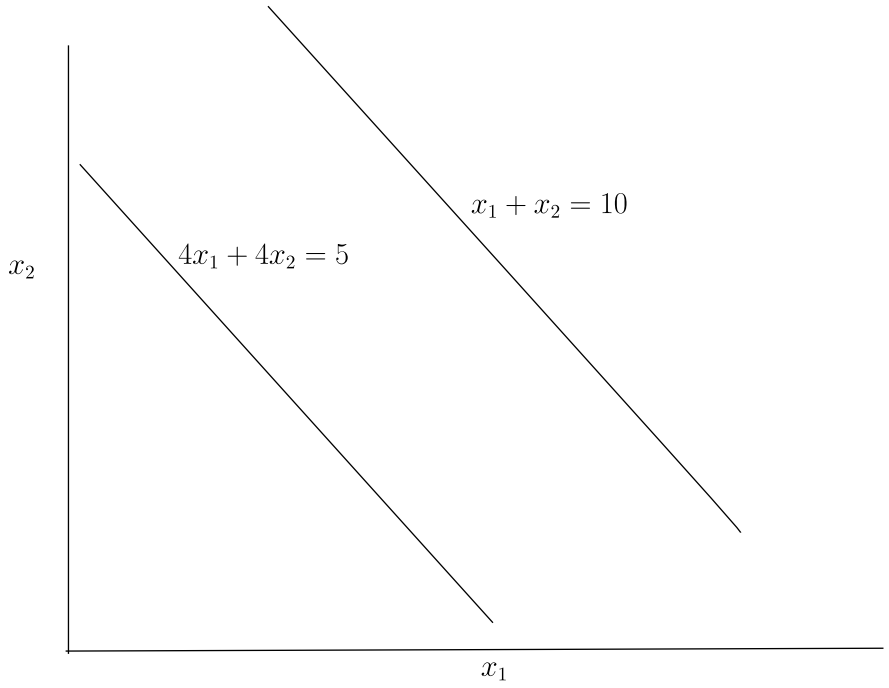
Geometrical interpretation

In the second case both constraints are the same, so there are an infinite number of solutions:



Geometrical interpretation

In the third case the constraints are mutually incompatible, so there is no assignment to x_1, x_2 which satisfies both constraints. The graph of both constraints shows a pair of parallel lines:



Higher dimensions

- In 3D each constraint is a plane.
- The intersection of two planes is a line.
- The intersection of the third plane with the first two planes will be a point on the line in case of a unique solution, or it may lead to pars of parallel lines (constraint 1 intersection constraint 2 gives one line, constraint 1 intersection constraint 3 gives parallel line, constraint 2 intersection constraint gives parallel line) which means there is no solution.
- All three constraints or planes may intersect in the same line which means infinite solutions.

Matrices

An (m, n) matrix A is a mn tuple of elements a_{ij} arranged in m rows and n columns as below:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Note that row and column vectors are also matrices - a row vector is a $(1, n)$ matrix and a column vector is a $(m, 1)$ matrix. $R^{m \times n}$ is the set of all (m, n) matrices consisting of real numbers as elements

Matrix addition and multiplication

The sum of two matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m \times n}$ is defined as an element-wise sum of the elements of the two matrices:

$$A = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}$$

Matrix addition and multiplication

- The product of two matrices $A \in \mathbb{R}^{m \times k}, B \in \mathbb{R}^{k \times n}$ is defined as the $m \times n$ matrix $C = AB$ where the elements c_{ij} are calculated as follows: $c_{ij} = \sum_{l=1}^{l=k} a_{il} b_{lj}$
- Essentially we are mutliplying the elements of the ith row of A with the jth column of B.
- The product BA is not defined if $n \neq m$

Matrix properties

- Associativity:
  $\forall A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}, C \in \mathbb{R}^{p\times k}, \;\; (AB)C = A(BC)$
- Distributivity: $\forall A, B \in \mathbb{R}^{m\times n}$,
  $\forall C, D \in \mathbb{R}^{n\times p}, (A+B)C = AC + BC,\; A(C+D) = AC + AD$
- Multiplication with the identity matrix: $\forall A \in \mathbb{R}^{m\times n}$,
  $I_m A = A I_n = A$ where I_m is the $m \times m$ identity matrix and I_n is the $n \times n$ matrix.

Inverse and transpose

Consider a square matrix $A \in \mathbb{R}^{n \times n}$. Let matrix $B \in \mathbb{R}^{n \times n}$ exist such that $AB = I_n$. Does every matrix $A \in \mathbb{R}^{n \times n}$ possess an inverse?

Let us take a simple 2×2 example $\rightarrow$ under what circumstances does it possess an inverse?

2×2 case

Define a 2×2 matrix A as follows:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Define matrix B to be

$$B = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

The product AB is

$$AB = \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & a_{12}a_{11} - a_{11}a_{12} \\ -a_{21}a_{22} - a_{22}a_{21} & -a_{21}a_{12} + a_{11}a_{22} \end{bmatrix}$$

2×2 case

$$AB = \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix} = (a_{11}a_{22} - a_{12}a_{21})I_2$$

$AB = (a_{11}a_{22} - a_{12}a_{21})I_2$. We can define

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

whenever $a_{11}a_{22} - a_{12}a_{21} \neq 0$.
Bottomline is that the inverse of a square matrix exists if and only of its determinant is non-zero.

Transpose

For $A \in \mathbb{R}^{m \times n}$, the matrix $B \in \mathbb{R}^{n \times m}$ with $b_{ij} = a_{ji}$ is called the transpose of A. We say $B = A^T$. Transposition means writing the rows of one matrix as the columns of the other. The following properties can be shown:

$$
\begin{aligned}
AA^{-1} = A^{-1}A &= I \\
(AB)^{-1} &= B^{-1}A^{-1} \\
(A+B)^{-1} &\neq B^{-1} + A^{-1} \\
(A^T)^T &= A \\
(A+B)^T &= A^T + B^T \\
(AB)^T &= B^T A^T
\end{aligned}
$$

Proof of $(AB)^T = B^T A^T$

- How do we show these properties? Let us look at the last one for example ...
- The (i, j)th entry of AB is obtained by taking the inner product of the ith row of A with the jth column of B.
- But the ith row of A is the ith column of A^T, and the jth column of B is the jth row of B^T.
- Thus we can take the inner product of the jth row of B^T with the ith column of A^T to get the same value for the $(i, j)th$ entry of AB. Thus when when we compute $C = B^T A^T$, we find that $C_{ji} = (AB)_{ij}$, so $C = (AB)^T$.

Multiplication by a scalar

- Let $A \in \mathbb{R}^{m \times n}$ and λ be a scalar. The $\lambda A = B$ such that $B_{ij} = \lambda A_{ij}$, i.e every element in A is scaled by λ to get the corresponding element in the scaled matrix B.
- For $\lambda, \psi \in \mathbb{R}$ we have
  - associativity: $(\lambda\psi)A = \lambda(\psi A), A \in \mathbb{R}^{m \times n}$ and $\lambda(AB) = (\lambda A)B = A(\lambda B)$ where $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times k}$.
- distributivity:
  - $(\lambda A)^T = A^T \lambda^T = A^T \lambda = \lambda A^T$ since $\lambda = \lambda^T$ for all $\lambda \in \mathbb{R}$.
  - $(\lambda + \psi)A = \lambda A + \psi A$
  - $\lambda(A + B) = \lambda A + \lambda B$

Compact representations of systems of equations

Consider the following system of equations:

$$\begin{aligned} 2x_1 + 3x_2 + 5x_3 &= 1 \\ 4x_1 - 2x_2 - 7x_3 &= 8 \\ 9x_1 + 5x_2 - 3x_3 &= 2 \end{aligned}$$

In matrix terms we can write this set of equations as

$$\begin{bmatrix} 2 & 3 & 5 \\ 4 & -2 & -7 \\ 9 & 5 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix}$$

Solving a system of equations

- From the previous slide we see that a system of equations can be expressed as $Ax = b$ where $A \in \mathbb{R}^{m \times n}$, x is a $n \times 1$ matrix and b is a $m \times 1$ matrix.
- We now look at how to obtain a particular and general solution for a system of equations
- Let us first look at an example.

Example of system of equations

$$\begin{bmatrix} 1 & 0 & 8 & -4 \\ 0 & 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \end{bmatrix}$$

- This system has two equations and four unknowns, so it is underconstrained. We expect an infinity of solutions.
- Is there a special way in which to express the solutions to this system?
- Let us examine the structure of the given problem matrix.

Structure of the solution

- Looking at the previous slide we can see that a linear combination of columns of the matrix will give the right hand side.
- The ith column vector in the matrix appears in the linear combination, scaled by the corresponding x_i as below.

$$x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} 8 \\ 2 \end{bmatrix} + x_4 \begin{bmatrix} -4 \\ 12 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \end{bmatrix}$$

Particular solution to the example

- A closer look at the linear combination to give the right hand side shows that we can take $x_1 = 42$, $x_2 = 8$, $x_3 = 0$, $x_4 = 0$ since the first two columns are $(1, 0)^T$ and $(0, 1)^T$ respectively.
- Therefore a solution is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \\ 0 \\ 0 \end{bmatrix}$$

- This solution is called the particular solution

Any other solutions possible?

- We can generate other solutions than the particular solution, by adding the vector $\mathbf{0}$ to the particular solution
- But isn't this the same as the particular solution as any vector $+\ \mathbf{0}$ is that vector itself?
- The trick is to express $\mathbf{0}$ in terms of the linear combination of some vectors.
- Describing $\mathbf{c_1}, \mathbf{c_2}, \mathbf{c_3}, \mathbf{c_4}$ as the four column vectors associated with the given matrix in the example we can see that $8\mathbf{c_1} + 2\mathbf{c_2}$ - $1\mathbf{c_3} + 0\mathbf{c_4} = \mathbf{0}$.

Any other solutions possible?

- Writing the linear combination in terms of a matrix-vector product we have

$$\begin{bmatrix} 1 & 0 & 8 & -4 \\ 0 & 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} 8 \\ 2 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Any vector $\lambda(8, 2, -1, 0)^T, \lambda \in R$ will also produce the $\mathbf{0}$ vector

Any other solutions possible?

- We can add the vector $(8, 2, -1, 0)^T$ to the original particular solution $(42, 8, 0, 0)^T$ to get another solution since

$$A(\begin{bmatrix} 42 \\ 8 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 8 \\ 2 \\ -1 \\ 0 \end{bmatrix}) = \begin{bmatrix} 42 \\ 8 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \end{bmatrix}$$

Any other solutions possible?

- Following the same line of reasoning as before, we can create the $\mathbf{0}$ vector by expressing the fourth column of the matrix $\boldsymbol{A}$ in terms of the first two columns - note that the first two columns appear capable of generating any two-dimensional vector!
- We can see that $-4\mathbf{c_1} + 12\mathbf{c_2} + 0\mathbf{c_3}$ - $1\mathbf{c_4} = \mathbf{0}$.
- Thus we have

$$\begin{bmatrix} 1 & 0 & 8 & -4 \\ 0 & 1 & 2 & 12 \end{bmatrix} (\lambda_2 \begin{bmatrix} -4 \\ 12 \\ 0 \\ -1 \end{bmatrix}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Putting things together

- We obtain the following general solution as the sum of the particular solution and a linear combination of solutions to the equation $\mathbf{Ax} = \mathbf{0}$ as follows:

$$\{\mathbf{x} \in \mathbb{R}^4 : \mathbf{x} = \begin{bmatrix} 42 \\ 8 \\ 0 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 8 \\ 2 \\ -1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} -4 \\ 12 \\ 0 \\ -1 \end{bmatrix}\}$$

- The general approach consisted of finding a particular solution to $\mathbf{Ax} = \mathbf{b}$, finding all solutions to $\mathbf{Ax} = \mathbf{0}$ and combining the particular and general solutions.
- Neither the particular nor general solutions are unique $\rightarrow$ why?

Algorithmic way of solving equations

- The system of equations in our example was easy to solve because of the special structure of the matrix - we could guess the solution without much difficulty.
- Can we develop an algorithmic way of solving a general system of equations?
- The answer is yes $\rightarrow$ we call the procedure Gaussian elimination

Elementary transformations

- The key idea is to take a complex looking matrix and transform it using elementary row operations to a simple looking matrix like the one we just handled, for which solutions could be obtained essentially by inspection.
- To make this work we need to preserve solutions of the original system of equations, i.e ensure that elementary transformations of the original matrix do not change its solutions.
- Do such elementary transformations exist?

What are the elementary operations?

- Exchange of rows
- Multiplying a row by a constant $\lambda \in R \backslash \{0\}$
- Adding a a row to another row
- Question $\rightarrow$ why must any multiplier to a row be non-zero?

Example to illustrate elementary operations

Consider the folliowing system where we seek all solutions for some $a \in R$.

$$\begin{aligned} -2x_1 + 4x_2 - 2x_3 - x_4 + 4x_5 &= -3 \\ 4x_1 - 8x_2 + 3x_3 - 3x_4 + x_5 &= 2 \\ x_1 - 2x_2 + x_3 - x_4 + x_5 &= 0 \\ x_1 - 2x_2 - 3x_4 + 4x_5 &= a \end{aligned}$$

Compact representation

Let us take the preceding equations and express them compactly in matrix form:

$$\left[\begin{array}{ccccc|c} -2 & 4 & -2 & -1 & 4 & -3 \\ 4 & -8 & 3 & -3 & 1 & 2 \\ 1 & -2 & 1 & -1 & 1 & 0 \\ 1 & -2 & 0 & -3 & 4 & a \end{array}\right]$$

This matrix is called the augmented matrix. It is on this matrix that we will perform the elementary row operations.

Swap rows

Now swap rows 1 and 3 in the augmented matrix to get

$$\left[\begin{array}{ccccc|c} 1 & -2 & 1 & -1 & 1 & 0 \\ 4 & -8 & 3 & -3 & 1 & 2 \\ -2 & 4 & -2 & -1 & 4 & -3 \\ 1 & -2 & 0 & -3 & 4 & a \end{array}\right]$$

Does this change the system of equations? No, because we are swapping **both left and right hand sides of the equality sign, so we are still dealing with the same set of equations.**

Subtract rows

$$\left[\begin{array}{ccccc|c} 1 & -2 & 1 & -1 & 1 & 0 \\ 4 & -8 & 3 & -3 & 1 & 2 \\ -2 & 4 & -2 & -1 & 4 & -3 \\ 1 & -2 & 0 & -3 & 4 & a \end{array}\right] \begin{array}{c} \\ -4R_1 \\ +2R_1 \\ +R_1 \end{array}$$

The notation above is used to convey that we could like to add $-4\times$ first row to the second row, $2\times$ the first row to the third row, and $1\times$ the first row to the fourth row to get a new augmented matrix.

New Augmented Matrix

$$\left[\begin{array}{ccccc|c} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 & -3 & 2 \\ 0 & 0 & 0 & -3 & 6 & -3 \\ 0 & 0 & -1 & -2 & 3 & a \end{array}\right] \quad -R_2 - R_3$$

Note that the augmented matrix shown is obtained by performing the operations shown on the previous slide. To get the next augmented matrix we subtract the second and third rows of this augmented matrix from the last row.

New Augmented Matrix

$$\left[\begin{array}{ccccc|c} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 & -3 & 2 \\ 0 & 0 & 0 & -3 & 6 & -3 \\ 0 & 0 & 0 & 0 & 0 & a+1 \end{array}\right] \begin{array}{c} \\ -1 \\ 1/3 \\ \\ \end{array}$$

Now multiply the second row by -1 and the third row by $\frac{1}{3}$ to get the augmented matrix in its final form, known as the row-echelon form.

Row-echelon form

$$\left[\begin{array}{ccccc|c} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 3 & -2 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & a+1 \end{array}\right]$$

We can revert to the set of equations represented by the augmented matrix as follows. These equations are equivalent to the original set of equations:

$$\begin{aligned} x_1 - 2x_2 + x_3 - x_4 + x_5 &= 0 \\ x_3 - x_4 + 3x_5 &= -2 \\ x_4 - 2x_5 &= 1 \\ 0 &= a+1 \end{aligned}$$

Existence of solution and particular solution

- The preceding set of equations cannot be solved when $a \neq -1$.
- The last equation is consistent only for $a = -1$.
- A particular solution is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

General Solution

$$x \in \mathbb{R}^5 : x = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ -2 \\ 1 \\ 1 \end{bmatrix}, \lambda_1, \lambda_2 \in R$$

Row-echelon form definition

- All rows that contain only zeros are at the bottom of the matrix.
- All rows that contain at least one nonzero element are on top of rows that contain only zeros.
- Considering only the non-zero rows, the first non-zero element in a given row is called the pivot and is always to the right of the pivot in the row above it.
- The positions of the pivots in the non-zero rows give rise to a staircase pattern.
- The variables corresponding to the pivot variables are called basic variables and those corresponding to the non-pivot positions correspond to the free variables.

Finding the particular solution

- The row echelon form makes finding a particular solution easy
- Remember that the idea is that a linear combination of the pivot columns must give the right hand side.
- In the example above this means that

$$\lambda_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \\ 1 \\ 0 \end{bmatrix}$$

- This looks like any regular linear combination for which we need to find the coefficients $\lambda_1, \lambda_2, \lambda_3$, so how is this really different from the original problem $\mathbf{Ax} = \mathbf{b}$?

Finding a particular solution

- The linear combination from the previous slide is easily solved.
- Start with finding the value of λ_3. We can see that the third equation establishes $\lambda_3 = 1$.
- The second equation involves only λ_2 and λ_3. Plugging the just discovered value of λ_3 into the second equation, we can find $\lambda_2 = -1$.
- Now we can plug the values of λ_2, λ_3 into the first equation to get $\lambda_1 = 2$

Reduced row-echelon form

- We can convert the row-echelon form into a simpler form called the reduced row-echelon form.
- In reduced row-echelon form, every pivot is equal to 1.
- The pivot is the only non-zero entry in its column
- Therefore the pivot columns look like canonical basis vectors of $\mathbb{R}^m$ where the original given matrix A is a $\mathbb{R}^{m \times n}$ matrix.

Example

- Consider the following matrix in reduced row-echelon form.

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \end{bmatrix}$$

- To find solutions for $\mathbf{Ax} = \mathbf{0}$ we need to look at non-pivot columns and note that the pivot columns are "strong enough" to generate the non-pivot columns.
- Our strategy to find solutions to $\mathbf{Ax} = \mathbf{0}$ is to find linear combinations of the pivot columns to the left of a non-pivot column to cancel out the non-pivot column, while setting all other coefficients to zero.

Example continued

- Thus we note that the second column is a non-pivot column which can be expressed as a multiple of the first column such that 3 times the first column + -1 * second column is equal to zero. This gives us our first solution.

$$\begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Example continued

- Similarly we note that 3 $\times$ the first column $+$ 9 $\times$ the third column $+$ -4 $\times$ the fourth column $+ -1\times$ the fifth column is equal to zero. This gives us our second solution:

$$\begin{bmatrix} 3 \\ 0 \\ 9 \\ -4 \\ -1 \end{bmatrix}$$

General solution

- If $\mathbf{x_1}$ and $\mathbf{x_2}$ are solutions to $\mathbf{Ax} = \mathbf{0}$, then any linear combination $\lambda_1\mathbf{x_1} + \lambda_2\mathbf{x_2}$, $\lambda_1, \lambda_2 \in R$ is also a solution
- Thus the general solution to the problem is

$$x \in \mathbb{R}^5 : x = \lambda_1 \begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 3 \\ 0 \\ 9 \\ -4 \\ -1 \end{bmatrix}, \lambda_1, \lambda_2 \in R$$

Gaussian elimination

- Consider the system of equations $\mathbf{Ax} = \mathbf{b}$ where A is a $n \times n$ matrix.
- If A is invertible, it means that A^{-1} exists such that $AA^{-1} = A^{-1}A = I_n$
- In such a case the row-reduced echelon form of A is I_n, i.e every column is a pivot column where the pivot is 1.
- The process of converting A to I_n that we have discussed above is called Gaussian Elimination

Gaussian elimination

- In Gaussian Elimination we use multiples of the first row to eliminate the entries in the first column below the first row.
- Then we use multiples of the second row to eliminate entries in the second column below the second row and so on until we get an upper-triangular matrix.
- This process is shown diagramatically in the next slide.
- Then we take multiples of the last row to eliminate non-zero entries in the last column above the last entry, followed by multiples of the last but one row to eliminate non-zero entries in the last but one column and so on. This gives us a diagonal matrix.

Gaussian elimination diagram

$$\begin{tikzcd}
\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \arrow[r] & \begin{bmatrix} a & b & c \\ 0 & e' & f' \\ 0 & h' & i' \end{bmatrix} \arrow[r] & \begin{bmatrix} a & b & c \\ 0 & e' & f' \\ 0 & 0 & i'' \end{bmatrix} \arrow[r] & \begin{bmatrix} a & 0 & c' \\ 0 & e' & f' \\ 0 & 0 & i'' \end{bmatrix} \arrow[d] \\
 & & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} a & 0 & 0 \\ 0 & e' & 0 \\ 0 & 0 & i'' \end{bmatrix} \arrow[l]
\end{tikzcd}$$

Calculating inverse

- Can the Gaussian elimination procedure calculate the inverse of a matrix?
- For example, let A be a $n \times n$ matrix whose inverse A^{-1} exists. We would like to compute its inverse using Gaussian elimination. Is this possible?
- Yes we can compute the inverse in the following way: we simply set up n linear systems of the form $\mathbf{Ax} = \mathbf{e_i}$, $1 \leq i \leq n$ where $\mathbf{e_i}$ is the ith canonical basis vector and find their solutions $\mathbf{x}$. Each solution vector constitutes a column in A^{-1}. Why is this true?

Calculating Inverse

- Consider the linear system $\mathbf{Ax} = \mathbf{e_i}$.
- Gaussian elimination will convert this sytem to the equivalent system $\mathbf{I_n x} = \mathbf{c_i}$ whose solution is $\mathbf{x} = \mathbf{c_i}$.
- On the other hand, the solution to $\mathbf{Ax} = \mathbf{e_i}$ is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{e_i}$.
- Since the two systems are equivalent they have the same solution, so $\mathbf{x} = \mathbf{c_i} = \mathbf{A}^{-1}\mathbf{e_i}$ which means $\mathbf{c_i}$ is the ith column of $\mathbf{A}^{-1}$.
- Thus when we create the augmented matrix $[Ae_i]$, Gaussian elimination will convert it into $[I_n c_i]$.
- We can solve n linear systems at once by letting the augmented matrix be $[\ \mathrm{A} \ \big|\ I_n\ ]$ which will become $[I_n A^{-1}]$.

Lecture 2

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Vector Spaces

- We would like to shift focus from systems of linear equations and their solutions to vector spaces.
- Vector spaces are structured spaces in which vectors live.
- What do we mean by "structure" here?

Groups

- We already have some notion about the structure of a vector space, i.e adding two vectors returns a vector, multiplying a vector by a scalar also returns a vector.
- To formalize these notions, we need the concept of a *Group*
- *A Group is a set G and an operation $\otimes : G \times G \rightarrow G$ defined on G. Then $(G, \otimes)$ is called a group if the following properties hold:*
  - *Closure of G under $\otimes$:* $\forall x, y \in G, x \otimes y \in G$
  - *Associativity:* $\forall x, y, z \in G, (x \otimes y) \otimes z = x \otimes (y \otimes z)$
  - *Neutral or Identity element:* $\ni e \in G, \forall x \in G \; x \otimes e = x$
  - *Inverse element:* $\forall x \in G, \ni y \in G, x \otimes y = y \otimes x = e$
- *If in addition to all the above properties we have $\forall x, y \in G, x \otimes y = y \otimes x$ then $(G, \otimes)$ is an Abelian group*

Some examples

- $(Z, +)$ is an Abelian group. We can see that all the aforementioned properties hold.
- What about $(N_0, +)$? This is not a group since there is no inverse element for an arbitrary element in it.
- $(Z, .)$ where Z is the set of integers, and . is product. It has the identity element but there exist elements in it that don't have an inverse.

Vector Spaces

- We will now consider sets that are just like groups in terms of its properties with respect to an inner operation $+$, and also have an outer operation called . which denotes the multiplication of a vector $x \in G$ by a scalar $\lambda \in R$.
- The inner operation can be viewed as a form of addition while the outer operation can be viewed as a form of scaling.
- A real-valued vector space is a set $V = (\mathcal{V}, +, .)$ with operations $+, .$ such that $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ and $. : R \times \mathcal{V} \to \mathcal{V}$ where $(\mathcal{V}, +)$ is an Abelian group, and the following properties hold:
  - Distributivity
  - associativity with respect to the outer operation.
  - there exists a neutral or identity element with respect to the outer operation.

Vector Spaces

Let us look at the properties of a Vector Space more carefully:

- Distributivity: $\forall \lambda \in R, \mathbf{x}, \mathbf{y} \in \mathcal{V}, \lambda.(\mathbf{x}+\mathbf{y}) = \lambda.\mathbf{x} + \lambda.\mathbf{y}$ and $\forall \lambda, \psi \in R, (\lambda + \psi).\mathbf{x} = \lambda.\mathbf{x} + \psi.\mathbf{x}$
- Associativity with respect to the outer operation: $\forall \lambda, \psi \in R, \mathbf{x} \in \mathcal{V}, \lambda.(\psi.\mathbf{x}) = (\lambda\psi).\mathbf{x}$
- Neutral element with respect to the outer operation: $\forall \mathbf{x} \in \mathcal{V}$, $1.\mathbf{x} = \mathbf{x}$
- The elements $\mathbf{x} \in \mathcal{V}$ are called vectors.
- The neutral element with respect to $(\mathcal{V}, +)$ is the zero vector $[0, 0, \ldots 0]^T$ and the inner operation is called vector addition.

Examples of vector spaces

Consider some examples of vector spaces, for example $\mathcal{V} = \mathbb{R}^n$

- We can define addition:
  $\mathbf{x} + \mathbf{y} = (x_1, x_2, \ldots x_n) + (y_1, y_2, \ldots y_n) =$
  $(x_1 + y_1, x_2 + y_2, \ldots x_n + y_n)$
- Multiplication by scalars:
  $\lambda\mathbf{x} = \lambda(x_1, x_2, \ldots x_n) = (\lambda x_1, \lambda x_2, \ldots \lambda x_n)$, $\forall \lambda \in R, \mathbf{x} \in \mathbb{R}^{\mathbf{n}}$.

Another example

What we call a vector need not be the standard column vector that we are accustomed to treating as a vector. We can think of $m \times n$ matrices as vectors and create a vector space out of them. Thus $\mathcal{V} = \mathbb{R}^{m \times n}$ with addition and multiplication defined as below:

- Addition $\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} \dots a_{1n} + b_{1n} \\ \vdots \\ a_{n1} + b_{n1} \dots a_{nn} + b_{nn} \end{bmatrix}$ is defined element-wise for two matrices $A, B \in \mathcal{V}$.
- Multiplication by scalars: $\lambda \mathbf{A} = \begin{bmatrix} \lambda a_{11} \dots \lambda a_{1n} \\ \vdots \\ \lambda a_{n1} \dots \lambda a_{nn} \end{bmatrix}$

Vector subspaces

- Let $V = (\mathcal{V}, +, .)$ be a vector space and let $\mathcal{U} \subseteq \mathcal{V}, \mathcal{U} \neq \Phi$. then $U = (\mathcal{U}, +, .)$ is called a vector subspace of V if U is a vector space with the vector space operations $+$ and $.$ restricted to $\mathcal{U} \times \mathcal{U}$ and $R \times \mathcal{U}$.
- We use the notation $U \subseteq V$ to denote that U is a vector subspace of V.
- If $U \subseteq V$ and V is a vector space, then U naturally inherits many properties directly from V because they hold for all $\mathbf{x} \in V$, and in particular for all $\mathbf{x} \in U \subseteq V$. These properties include the Abelian group property, associativity, distributivity and the neutral element.

Establishing vector subspaces

How do we show if $(\mathcal{U}, +, .)$ is subspace of V?
We need to show that

- $\mathcal{U} \neq \Phi$
- Closure of U with respect to the outer and inner operations, i.e $\forall \lambda \in R, \forall \mathbf{x}, \in \mathcal{U}, \lambda \mathbf{x} \in \mathcal{U}$ and $\forall \mathbf{x}\ , \mathbf{y} \in \mathcal{U}, \mathbf{x} + \mathbf{y} \in \mathcal{U}$.

Examples

- Let $\mathcal{V} = \mathbb{R}^2$ and $\mathcal{U}$ be the y-axis. . Is $U = (\mathcal{U}, +, .)$ a subspace of $V = (\mathcal{V}, +, .)$? Answer: Yes, because the addition of any two vectors on the y-axis remains on the y-axis so closure with respect to the addition operation is satisfied. Also any vector on the y-axis when scaled by any real number (including 0) will yield a vector on y-axis.
- What about when we shift the y-axis one unit to the right, i.e $\mathcal{U} = \{x = 1\}$? Answer: We no longer have a vector subspace since scaling with respect to the outer operation does not have the closure property.

More examples

- What about the subset of $\mathbb{R}^2$ that represents a square around the origin , i.e $-1 \leq x \leq 1, -1 \leq y \leq 1$? This is again not a subspace.
- This subspace is of particular interest to us called the nullspace of a matrix, i.e the set of solutions to a linear system of equations $A\mathbf{x} = 0$. Why is this a vector subspace?

Linear combination and linear independence

- We know that we can stay in a vector space by adding vectors belonging to the space, and scaling them?
- We are now interested in a different question $\rightarrow$ can we come up with a set of vectors such that every vector in the vector space can be represented as a sum of these vectors with scaling as necessary?
- The answer is yes. A set of vectors capable of representing all vectors in a vector space is called a basis.
- To explore the question of finding a basis, we need to learn the concepts of linear combination and linear independence.

Linear combination

- Consider a vector space V and a set of vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_k \in V$. Then every $\mathbf{v} \in V$ that is of the form $v = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 \ldots + \lambda_k \mathbf{x}_k$ is a linear combination of the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_k$.
- Note that the $\mathbf{0}$ can be written trivially as a linear combination of the given k vectors $\mathbf{x}_i$s. We are however interested in non-trivial linear combinations of vectors to get the $\mathbf{0}$-vector.

Linear (in)dependence

- Consider a vector space V and k vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k \in V$. Then if there is a non-trivial linear combination of the given vectors such that $0 = \sum_{i=1}^{i=k} \lambda_i \mathbf{x}_i$ where at least one of the λ_is is non-zero, then the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k$ are linearly dependent.
- Put another way, if the only way we can combine vectors to get the $\mathbf{0}$-vector is by letting all the λ_i be zero, then the vectors are linearly independent.

Linear (in)dependence

- What does the concept of linear independence capture? If the vectors are linearly dependent we can write one of the vectors in terms of the others, so that vector is redundant. On the other hand, when the vectors are linearly independent, each vector brings something to the table which the other vectors collectively cannot replace.

Linear independence

- k vectors in a vector space are either linearly independent or linearly dependent. There is no third option.
- If at least one of the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_k$ is the 0-vector, then the vectors are linearly dependent. Why? Apply the definition to see that this is the case. We can choose a non-zero λ for the 0-vector and zero λs for all the other vectors to get the linear combination to be zero.
- If all the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_k$ are non-zero, then the vectors are linearly dependent if and only if one of them is a linear combination of the others.

Using Gaussian elimination to check for linear independence

- A practical way of checking whether a bunch of vectors are linearly independent is to fill out the columns of a matrix with the given vectors and then perform Gaussian elimination to get the row-echelon form
- The pivot columns indicate all the vectors that are linearly independent, and they are all on the left.
- The non-pivot columns can be expressed as a linear combination of the pivot columns

Example

- Let us start with the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 4 \end{bmatrix}$$

- Gaussian elimination of this matrix will give rise to the following row-echelon form

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

- The pivot columns are the first and the third column and we see that the second column which is a non-pivot column can be expressed as a linear combination of the pivot columns to its left, which is just twice the first column.

Another example

Consider the vectors $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ -3 \\ 4 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \end{bmatrix}$, $\mathbf{x}_3 = \begin{bmatrix} -1 \\ -2 \\ 1 \\ 1 \end{bmatrix}$

To check for linear independence we set up the equation with λs as follows: $\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \lambda_3\mathbf{x}_3 = 0$.

As before we put the given vectors into a matrix to get:

$$\begin{bmatrix} 1 & 1 & -1 \\ 2 & 1 & -2 \\ -3 & 0 & 1 \\ 4 & 2 & 1 \end{bmatrix}$$

.

Another example continued

We now perform Gaussian elimination on this matrix to identify pivot and non-pivot columns. If all columns are pivot columns then the given three vectors are linearly independent. After Gaussian elimination we get

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

.

Note that all columns are pivot columns, so we conclude that the original three vectors are linearly independent. The only way to combine those vectors and get the 0-vector is to take 0s as the λs.

More on linear independence

- Let us say that we have a bunch of k independent vectors $\mathbf{b}_1$, $\mathbf{b}_2$, ... $\mathbf{b}_k$
- We now have m vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_m$, each of which is some linear combination of the vectors $\mathbf{b}_i$.
- The goal is to find under what conditions the vectors $\mathbf{x}_i$ are linearly independent.
- Can we express the criterion of linear independence of the $\mathbf{x}_i$s in terms of the way in which we combine the $\mathbf{b}_i$s to get the $\mathbf{x}_i$s?

More on linear independence

We have the following equations:

$$
\begin{aligned}
\mathbf{x}_1 &= \sum_{i=1}^{i=k} \lambda_{i1}\mathbf{b}_i; \\
\mathbf{x}_2 &= \sum_{i=1}^{i=k} \lambda_{i2}\mathbf{b}_i; \\
&\vdots \\
\mathbf{x}_m &= \sum_{i=1}^{i=k} \lambda_{im}\mathbf{b}_i;
\end{aligned}
$$

More on linear independence

- We can define a matrix $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_k]$ whose columns are the original linearly independent vectors $\mathbf{b}_1, \mathbf{b}_2 \ldots \mathbf{b}_k$ and then we see that $\mathbf{x}_j = \mathbf{B}\boldsymbol{\lambda}_j$ where $\boldsymbol{\lambda}_j$ is a vector of coefficients.
- Now to test whether $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_m$ are linearly independent, we set up the equation $\sum_{j=0}^{j=m} \psi_j \boldsymbol{x}_j = 0$ and check whether the ψ_js have to be zero. This means

$$\sum_{j=1}^{j=m} \psi_j \boldsymbol{x}_j = \sum_{j=1}^{j=m} \psi_j \boldsymbol{B}\boldsymbol{\lambda}_j = \boldsymbol{B} \sum_{j=1}^{j=m} \psi_j \boldsymbol{\lambda}_j$$

More on linear independence

Now $\sum_{j=1}^{j=m} \psi_j \boldsymbol{\lambda}_j = \boldsymbol{v}$, a vector so $\boldsymbol{Bv}$ is a linear combination of the column vectors of $\boldsymbol{B}$, and the only way a linear combination of the columns of $\boldsymbol{B}$ can be equal to zero if all the combining coefficients are zero. This means that $\boldsymbol{v} = \boldsymbol{0} = \sum_{j=1}^{j=m} \psi_j \boldsymbol{\lambda}_j$.
The fact that $\sum_{j=0}^{j=m} \psi_j \boldsymbol{x}_j = 0$ implies that $\sum_{j=1}^{j=m} \psi_j \boldsymbol{\lambda}_j = 0$ means that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_m$ are linearly independent if and only if $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots \boldsymbol{\lambda}_m$ are linearly independent.

Example

Consider the following set of equations:

$$
\begin{aligned}
\boldsymbol{x}_1 &= \boldsymbol{b}_1 - 2\boldsymbol{b}_2 + \boldsymbol{b}_3 - \boldsymbol{b}_4 \\
\boldsymbol{x}_2 &= -4\boldsymbol{b}_1 - 2\boldsymbol{b}_2 + 4\boldsymbol{b}_4 \\
\boldsymbol{x}_3 &= 2\boldsymbol{b}_1 + 3\boldsymbol{b}_2 - \boldsymbol{b}_3 - 3\boldsymbol{b}_4 \\
\boldsymbol{x}_4 &= 17\boldsymbol{b}_1 - 10\boldsymbol{b}_2 + 11\boldsymbol{b}_3 + \boldsymbol{b}_4
\end{aligned}
$$

Example

We would like to find if the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ are linearly independent. Following the theory established in the last slides we check if the vectors corresponding to the $\boldsymbol{\lambda}_j$s are linearly independent.
We therefore need to check whether the vectors

$$\begin{bmatrix} 1 \\ -2 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -4 \\ -2 \\ 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ -1 \\ -3 \end{bmatrix}, \begin{bmatrix} 17 \\ -10 \\ 11 \\ 1 \end{bmatrix} \tag{1}$$

are linearly independent.

Example

We form a matrix consisting of all these column vectors:

$$A = \begin{bmatrix} 1 & -4 & 2 & 17 \\ -2 & -2 & 3 & -10 \\ 1 & 0 & -1 & 11 \\ -1 & 4 & -3 & 1 \end{bmatrix}$$

and perform Gaussian elimination to get the reduced row-echelon form.

Example

We get the following matrix:

$$B = \begin{bmatrix} 1 & 0 & 0 & -7 \\ 0 & 1 & 0 & -15 \\ 0 & 0 & 1 & -18 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Clearly there is a non-trivial way of combining the columns of the above matrix to get the $\mathbf{0}$-vector , i.e take $7 \times$ the first column, add it to $15 \times$ the second column, add the result to $18 \times$ the third column and add the result to the fourth column to get $\mathbf{0}$. Thus the original set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ are not linearly independent.

Insight from the example

- We had a set of k vectors $\boldsymbol{b}_i$ which we combined linearly to get m vectors $\boldsymbol{x}_i$.
- Determining whether the m vectors $\boldsymbol{x}_i$ were linearly independent boiled down to checking if the column vectors of a $k \times m$ matrix were linearly independent.
- We performed Gaussian elimination on this matrix to get the reduced row-echelon form.

Insight from the example

- Here is an interesting insight: if $m > k$, we have more columns than rows and there can only be as many pivots as there are non-zero rows, so the number of pivots and therefore the number of pivot columns is less than k.
- We are therefore **guaranteed** to have non-pivot columns in this case, so that we can express the non-pivot columns in terms of the pivot columns to their left and get a non-trivial linear combination of the columns to get the $\mathbf{0}$ vector. This would make the m vectors $\boldsymbol{x}_i$ **linearly dependent**.

Basis of vector space

- Do there exist a set of vectors in a vector space which "span" the entire space?
- What we mean here is that any vector $v \in V$ can be generated as a linear combination of the vectors in question.
- Is such a set of vectors unique to the vector space?
- Are all sets capable of generating all vectors in a vector space of the same size?

Generating set and basis

- Consider a vector space $V = (\mathcal{V}, +, .)$ and a set of vectors $\mathcal{A} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k\} \subseteq V$.
- If every vector $v \in V$ can be expressed as a linear combination of the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k$, then $\mathcal{A}$ is called a **generating set** of $\mathcal{V}$.
- The set of all linear combinations of the vectors in $\mathcal{A}$ is known as the **span** of $\mathcal{A}$.
- If $\mathcal{A}$ spans the vector space V we write $V = span[\mathcal{A}]$.

Basis as smallest generating set

- Consider a vector space $V = (\mathcal{V}, +, .)$ and $\mathcal{A} \subseteq \mathcal{V}$.
- A generating set $\mathcal{A}$ of V is called minimal if there is no smaller set $\tilde{\mathcal{A}}$ such that $\tilde{\mathcal{A}} \subset \mathcal{A} \subseteq \mathcal{V}$ that spans V.
- Every linearly independent generating set of V is minimal and is called a basis of V.

Characterizations of a basis

The following statements are equivalent:

- $\mathcal{B}$ is a basis of V.
- $\mathcal{B}$ is a minimal generating set.
- $\mathcal{B}$ is a maximal linearly independent set of vectors in V such that adding a vector to it will make it a linearly dependent set.
- Every vector $\boldsymbol{x} \in V$ is a linear combination of vectors from $\mathcal{B}$, and every linear combination is unique. This means that if $\boldsymbol{x} = \sum_{i=1}^{i=k} \lambda_i \boldsymbol{b}_i = \sum_{i=1}^{i=k} \psi_i \boldsymbol{b}_i$, then $\lambda_i = \psi_i, 1 \leq i \leq k$.

Basis is not unique

- Consider the space $\mathbb{R}^3$. The canonical basis is $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$
- Another basis for $\mathbb{R}^3$ is $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\1\end{bmatrix}$
- Yet another basis for $\mathbb{R}^3$ is $\begin{bmatrix}0.5\\0.8\\0.4\end{bmatrix}, \begin{bmatrix}1.8\\0.3\\0.3\end{bmatrix}, \begin{bmatrix}-2.2\\-3.3\\1.5\end{bmatrix}$

Not a basis

- Is any linear independent set a basis?
- No, we can have a linearly independent set that has too few vectors to become a basis
- As an example, consider the following set of vectors in $\mathbb{R}^4$
$$\begin{bmatrix}1\\2\\3\\4\end{bmatrix}, \begin{bmatrix}2\\-1\\0\\2\end{bmatrix}, \begin{bmatrix}1\\1\\0\\4\end{bmatrix}$$
- Why is the above not a basis?

Some properties

- Vector spaces can be finite or infinite-dimensional.
- We consider only finite-dimensional vector spaces.
- In a finite-deimensional vector space V, the number of vectors in the basis is known as $dim(V)$.
- If $U \subseteq V$, then $dim(U) \leq dim(V)$.
- $dim(U) = dim(V)$ if and only if $U = V$.

Dimension of a vector space

- Is the size of the basis merely the number of components in the vector?
- So far, that has been the case.
- But consider the vector space $span(\begin{bmatrix} 0 \\ 1 \end{bmatrix})$ - the vector has two components, but the basis is just one vector. All vectors in this vector space are multiples of the vector $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Finding a basis

There are three steps to finding a basis for a vector space. A basis of a subspace $U = span(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \subseteq \mathbb{R}^n$ can be found by executing the following steps:

- Write the spanning vectors as columns of a matrix $\boldsymbol{A}$.
- Determine the row-echelon form of $\boldsymbol{A}$.
- The spanning vectors associated with the pivot columns are a basis of U.

Example

- Consider the vector subspace $U \subseteq \mathbb{R}^5$ which is spanned by the following vectors:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 2 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} 2 \\ -1 \\ 1 \\ 2 \\ -2 \end{bmatrix}, \boldsymbol{x}_3 = \begin{bmatrix} 3 \\ -4 \\ 3 \\ 5 \\ -3 \end{bmatrix}, \boldsymbol{x}_4 \begin{bmatrix} -1 \\ 8 \\ -5 \\ -6 \\ 1 \end{bmatrix}$$

- We would like to check if the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ constitute a basis for the subspace U.
- We therefore need to check whether the given vectors are linearly independent or not.

Lecture 3

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Analytic geometry

- We have studied vector spaces in the previous lecture.
- Now we would like to provide some geometric interpretation to these concepts.
- We shall take a close look at geometric vectors and the concepts of lengths of vectors and angles between vectors.
- But first we need to add the concept of an inner product to our vector space.

Norms

- A norm on a vector space is a function $\|.\| : V \to \mathbb{R}$, $\boldsymbol{x} \to \|\boldsymbol{x}\|$ which assigns to each vector $\boldsymbol{x}$ a length $\|\boldsymbol{x}\|$ such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following properties hold:
  - Absolutely homogeneous: $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$
  - Triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
  - Positive definite: $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \implies \boldsymbol{x} = 0$
- Manhattan norm : $\|\boldsymbol{x}\| = \sum_{i=1}^{i=n} |x_i|$
- Euclidean norm : $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{i=n} x_i^2}$.

Inner products

- Dot product in $\mathbb{R}^n$ is given by $\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{i=n} x_i y_i$
- A bilinear mapping Ω is a mapping with two arguments and is linear in both arguments: Let V be a vector space such that $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$, and let $\lambda, \psi \in \mathbb{R}$. Then we have $\Omega(\lambda\boldsymbol{x} + \psi\boldsymbol{y}, \boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{z} + \psi\Omega(\boldsymbol{y}, \boldsymbol{z})$, and $\Omega(\boldsymbol{x}, \lambda\boldsymbol{y} + \psi\boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{y}) + \psi\Omega(\boldsymbol{x}, \boldsymbol{z})$.
- Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors as arguments and returns a real number. Then Ω is called symmetric if $\Omega(\boldsymbol{x}, \boldsymbol{y}) = \Omega(\boldsymbol{y}, \boldsymbol{x})$. Also Ω is called positive-definite if $\forall x \in V \setminus \{0\}, \;\; \Omega(\boldsymbol{x}, \boldsymbol{x}) > 0$ and $\Omega(0, 0) = 0$.

Inner products

- A positive-definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an inner product. To denote an inner product on V we generally write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$.
- The pair $(V, \langle ., . \rangle)$ is called an inner product space.
- Next we introduce the concept of symmetric, positive-definite matrices and show we can express an inner product using such matrices.
- We call that in a vector space V any vector $\boldsymbol{x}$ can be written as linear combination of the basis vectors. We use this to express an inner product in terms of a matrix.

Symmetric, positive-definite matrices

Theorem: For a real-valued, finite-dimensional vector space V and an ordered basis B of V, it holds that $\langle ., . \rangle : V \times V \to \mathbb{R}$ is an inner product if and only if there exists a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$.

Proof: One direction $\rightarrow$: $\langle ., . \rangle$ is an inner product $\implies$ $\boldsymbol{A}$ is symmetric, positive-definite such that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$.

Other direction $\leftarrow$: $\boldsymbol{A}$ is symmetric, positive definite such that the operation $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$ $\implies$ the operation defined is an inner product.

Symmetric, positive-definite matrices

- We prove the $\rightarrow$ direction.
- Let $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ be the inner product between the vectors $\boldsymbol{x}, \boldsymbol{y}$ in V. We can write $\boldsymbol{x}$ in terms of say n basis vectors as $\boldsymbol{x} = \sum_{i=1}^{i=n} \psi_i \boldsymbol{b}_i$. Similarly $\boldsymbol{y} = \sum_{i=1}^{i=n} \lambda_i \boldsymbol{b}_i$.
- Since the inner product is bilinear we can write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \sum_{i=1}^{i=n} \psi_i \boldsymbol{b}_i, \sum_{i=1}^{i=n} \lambda_i \boldsymbol{b}_i \rangle = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} \psi_i \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle \lambda_j = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$ where $A_{ij} = \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle$.
- Here $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ are vectors which represent the coordinates of the original vectors $\boldsymbol{x}, \boldsymbol{y}$ with respect to the basis vectors.

Symmetric, positive-definite matrices

- This means that the inner product is entirely determined through the matrix $\boldsymbol{A}$. The symmetry of the inner product means that $A_{ij} = \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = A_{ji} = \langle \boldsymbol{b}_j, \boldsymbol{b}_i \rangle$. Thus $\boldsymbol{A}$ is symmetric.
- The positive-definiteness of the inner product means that $\forall \boldsymbol{x} \in V \backslash \{0\}, \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$.

Symmetric, positive-definite matrices

- Now let us consider an operation op such that $\boldsymbol{x} op \boldsymbol{y} = \hat{\boldsymbol{x}}^T A \hat{\boldsymbol{y}}$ where $\boldsymbol{A}$ is a symmetric, positive definite matrix.
- We shall show that "op" is an inner product by showing that it has all the properties of an inner product:
  - "op" has symmetry because $\boldsymbol{x} op \boldsymbol{y} = \hat{\boldsymbol{x}}^T A \hat{\boldsymbol{y}}$ and $\boldsymbol{y} op \boldsymbol{x} = \hat{\boldsymbol{y}}^T A \hat{\boldsymbol{x}} = \hat{\boldsymbol{y}}^T (A\hat{\boldsymbol{x}})$. By a property of the dot product we can write $\hat{\boldsymbol{y}}^T (A\hat{\boldsymbol{x}}) = (A\hat{\boldsymbol{x}})^T \hat{\boldsymbol{y}} = \hat{\boldsymbol{x}}^T \boldsymbol{A}^T \hat{\boldsymbol{y}} = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$ where the last equality in the chain is possible since $\boldsymbol{A}$ is symmetric.
  - "op" also has bilinearity since we see that for $r \in R, (r\boldsymbol{x}) op \boldsymbol{y} = (r\hat{\boldsymbol{x}})^T \boldsymbol{A} \hat{\boldsymbol{y}} = r\hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}} = r\boldsymbol{x} op \boldsymbol{y}$.
  - $(\boldsymbol{x} + \boldsymbol{y}) op \boldsymbol{z} = (\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}})^T \boldsymbol{A} \hat{\boldsymbol{z}} = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{z}} + \hat{\boldsymbol{y}}^T \boldsymbol{A} \hat{\boldsymbol{z}} = \boldsymbol{x} op \boldsymbol{z} + \boldsymbol{y} op \boldsymbol{z}$.
  - Finally if $\boldsymbol{x}$ is a non-zero vector then $\hat{\boldsymbol{x}}$ is also a non-zero vector, $\boldsymbol{x} op \boldsymbol{x} = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{x}} > 0$ since we are given that $\boldsymbol{A}$ is positive-definite.

Symmetric, positive definite matrices

- Can a symmetric, positive-definite matrix have less than full rank? We have $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} > 0$ for all non-zero $\boldsymbol{x}$. Thus $\boldsymbol{x} = \boldsymbol{0}$ is the only vector allowed in the nullspace. The nullspace is 0-dimensional so $\boldsymbol{A}$ has full rank.
- What can be said about the diagonal elements of a positive-definite matrix? From $(\boldsymbol{e}_i)^T \boldsymbol{A}\boldsymbol{e}_i > 0$ where $\boldsymbol{e}_i$ is the ith canonical basis vector, we see that $A_{ii} > 0$. Thus the diagonal entries are all strictly positive.

Lengths and distances

- Inner products and norms are closely related in the sense that any inner product induces a norm: $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$
- Not every norm is induced by an inner product, for example the Manhattan norm.
- For an inner product vector space $(V, \langle ., . \rangle)$, the induced norm $\|.\|$ satisfies the Cauchy-Schwarz inequality: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$. Why is this true?

Cauchy-Schwarz inequality

- Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two vectors and let us consider the length of the vector $\boldsymbol{u} - \alpha\boldsymbol{v}$ where α is a constant.
- The length of the vector $\boldsymbol{u} - \alpha\boldsymbol{v}$ is greater than or equal to zero. The length of the vector $\boldsymbol{u} - \alpha\boldsymbol{v}$ is $\|\boldsymbol{u} - \alpha\boldsymbol{v}\| = \langle \boldsymbol{u} - \alpha\boldsymbol{v}, \boldsymbol{u} - \alpha\boldsymbol{v} \rangle = (\boldsymbol{u} - \alpha\boldsymbol{v})^T(\boldsymbol{u} - \alpha\boldsymbol{v})$.
- We can expand the dot product $(\boldsymbol{u} - \alpha\boldsymbol{v})^T(\boldsymbol{u} - \alpha\boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{u} - \alpha\boldsymbol{u}^T\boldsymbol{v} - \alpha\boldsymbol{v}^T\boldsymbol{u} + \alpha^2\boldsymbol{v}^T\boldsymbol{v} \geq 0$
- Now set $\alpha = \frac{\boldsymbol{u}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}}$ to get $\boldsymbol{u}^T\boldsymbol{u} - \frac{(\boldsymbol{u}^T\boldsymbol{v})^2}{\boldsymbol{v}^T\boldsymbol{v}} \geq 0$ which leads us to $(\boldsymbol{u}^T\boldsymbol{u})(\boldsymbol{v}^T\boldsymbol{v}) \geq (\boldsymbol{u}^T\boldsymbol{v})^2$ which is Cauchy-Schwarz inequality.

Metric space

- Consider an inner product space $(V, \langle ., . \rangle)$. Define $d(\boldsymbol{x}, \boldsymbol{y})$ the distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to be $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$.
- If we use the dot product as the inner product, then the distance is called the Euclidean distance.
- The mapping $d : V \times V \to \mathbb{R}$ is called a metric.

Properties of a metric space

A metric d has the following properties:

- d is positive-definite which means $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in V$. $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \implies \boldsymbol{x} = \boldsymbol{y}$.
- d is symmetric which means $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in V$.
- d obeys the triangle inequality as follows: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}) \;\; \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$

Inner products and metrics seem to be very similar in terms of their properties - however there is one important difference. When $\boldsymbol{x}$ and $\boldsymbol{y}$ are close to each other the inner product is large but the distance metric is small. On the other hand when $\boldsymbol{x}$ and $\boldsymbol{y}$ are far apart, then the inner product is small but the distance metric is large.

Angles and orthogonality

- In addition to being able to capture the lengths of vectors and the distance between vectors, inner products can also capture the **angle** ω between two vectors and can thus capture the geometry of a vector space.
- The key to using the inner product to characterize the angle between two vectors is the Cauchy-Schwarz inequality.
- Assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are not the $\mathbf{0}$ vector. Then the Cauchy-Schwarz inequality tells us that

$$-1 \leq \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \leq 1 \tag{1}$$

Angles and orthogonality

- Since the Cauchy-Schwarz ratio lies between -1 and 1 we can set it equal to the cosine of a unique angle $\omega \in [0, \pi]$ such that

$$\cos(\omega) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \tag{2}$$

- The angle ω is the angle between two vectors. What does it capture?
- The notion of angle captures similarity of orientation between two vectors. When the dot product is close to zero, the vectors are more or less pointing in the same direction and $\omega \approx 0$.

Angles and orthogonality

- Food for thought: Suppose we choose vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ uniformly at random in high dimensions. What happens to the dot product between the vectors and hence the angle between them?
- To choose a vector uniformly at random over a sphere let every component in the vector be an independent Gaussian random variable of mean 0 and unit variance.
- Write a small program to see what happens ...

Angles and orthogonality

- A key feature of the inner product is that we can use it to characterize vectors that are orthogonal.
- Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal if and only if the inner product between them is 0. For an orthogonal pair of vectors $\boldsymbol{x}$, $\boldsymbol{y}$ we can write $\boldsymbol{x} \perp \boldsymbol{y}$.
- By the above definition the $\mathbf{0}$-vector is orthogonal to all vectors.
- Vectors which are othrogonal with respect to one inner product need not be orthogonal with respect to another inner product.

Example - angles and orthogonality

- Consider the vectors $\boldsymbol{x} = [1, 1]^T$ and $\boldsymbol{y} = [-1, 1]^T$
- With respect to the inner product defined as a dot product we see that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = 1 * -1 + 1 * 1 = 0$.
- With respect to the inner product $\boldsymbol{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y}$, the angle between the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ becomes

$$\cos(\omega) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

Example - angles and orthogonality

- Continuing with our example we have

$$\begin{aligned}\cos(\omega) &= \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\ \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}} \\ &= \frac{2x_1y_1 + x_2y_2}{(2x_1^2 + x_2^2)(2y_1^2 + y_2^2)} \\ &= \frac{-1}{3}\end{aligned}$$

where $\boldsymbol{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$.

- Thus, with respect to the new definition of inner product the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are no longer orthogonal.

Orthonormal matrix

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix if and only if its columns are orthonormal:

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^T$$
$$\boldsymbol{A}^T = \boldsymbol{A}^{-1}$$

- If the columns of a matrix are orthonormal, why are its rows orthonormal too? This follows from the fact that the left-inverse of a square matrix is the same as the right-inverse. Let $\boldsymbol{A}$ be a square matrix with $\boldsymbol{B}$ and $\boldsymbol{C}$ the left and right inverses of $\boldsymbol{A}$: $\boldsymbol{B}\boldsymbol{A} = I = \boldsymbol{A}\boldsymbol{C} \implies \boldsymbol{B} = \boldsymbol{C}$. Why is this true?

Orthonormal matrix

- Transformations by an orthonormal matrix preserve lengths. This can be seen as follows, using the dot product as the definition of the inner product:
  $\|\boldsymbol{Ax}\|^2 = \boldsymbol{Ax}^T\boldsymbol{Ax} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{x}^T I\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x}$.
- An example of an orthonormal matrix is the 2D-rotation matrix which can be expressed as $\begin{bmatrix} \cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$ where θ is the angle of rotation.

Orthonormal matrix

- Also the angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ does not change after transformation by an orthonormal matrix. This can be seen as follows:

$$
\begin{aligned}
\cos(\omega) &= \frac{\boldsymbol{A}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}}{\|\boldsymbol{A}\boldsymbol{x}\|\ \|\boldsymbol{A}\boldsymbol{y}\|} \\
&= \frac{\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{y}}{\|\boldsymbol{x}\|\ \|\boldsymbol{y}\|} \\
&= \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\ \|\boldsymbol{y}\|}
\end{aligned}
$$

Orthonormal basis

- We already looked at the concept of a basis of a vector space, and found that for the vector space $\mathbb{R}^n$ we need n basis vectors.
- Our basis vectors needed only to be linearly independent - we can ensure linear independence by ensuring that our basis vectors point in different directions, so that a linear combination of $n-1$ basis vectors cannot cancel out the nth basis vector.
- Now we will look at a special case of a basis where the vectors are all mutually orthogonal in the sense of the inner product, and each vector is of unit length. We call such a basis an orthonormal basis.

Orthonormal basis

- Question: Can you immediately think of an orthonormal basis for $\mathbb{R}^n$? Is an orthonormal basis for a vector space unique?
- Formal definition of an orthonormal basis: Consider an n-dimensional vector space V and n basis vectors $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_n\}$. If it is true that $\forall i, j = 1, \ldots n, i \neq j \ \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0$ and $\langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1$, then the basis is called an orthonormal basis.
- If the basis vectors are only mutually orthogonal but not of length unity, then we have an orthogonal basis.

Gram-Schmidt process

- Given a set of basis vectors for a vector space, can we convert the given basis into an orthogonal basis? Yes, we shall use Gaussian elimination to construct such a basis.
- Let us start with an example: Consider $\mathbb{R}^2$ and two basis vectors $\boldsymbol{v}_1 = (3,1)^T$ and $\boldsymbol{v}_2 = (2,2)^T$. Put these vectors into columns of a matrix $\boldsymbol{A}$ such that $\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}$.
- The next step is to perform Gaussian elimination on the following augmented matrix: $[\boldsymbol{A}^T\boldsymbol{A}|\boldsymbol{A}^T] = \begin{bmatrix} 10 & 8| & 3 & 1 \\ 8 & 8| & 2 & 2 \end{bmatrix}$
- On performing Gaussian elimination of this augmented matrix we end up with $\begin{bmatrix} 1 & 0| & 0.3 & 0.1 \\ 0 & 1| & -0.25 & 0.75 \end{bmatrix}$

Gram-Schmidt process

- Note that after the completion of Gaussian elimination the two rows on the right hand side are orthogonal. They form a basis for $\mathbb{R}^2$. We can normalize the vectors to get an orthonormal basis.
- What is the justification for this technique?
- First we see that when the $m \times n$ matrix $\boldsymbol{A}$ has full column rank, then the matrix $\boldsymbol{A}^T\boldsymbol{A}$ is positive definite. To see this note that any solution $\boldsymbol{x}$ to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ is also a solution to $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ and vice-versa. Why is this the case?
- When $\boldsymbol{A}$ is full rank, there are no non-trivial solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. Thus the fact that there are no non-trivial solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ means that $\forall \boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > \boldsymbol{0}$.

Elementary transformations

- One of the steps of Gaussian elimination is the the subtraction of a multiple of a given row from a row below it. This step can be achieved by pre-multiplication of the given matrix by an elementary matrix. An elementary matrix is like an identity matrix except that one of the entries below the diagonal is allowed to be non-zero.
- To show how the process of elimination works using an elementary matrix consider the matrix $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ and assume that we want to subtract two times the first row from the second row.

Elementary transformations

This can be accomplished by the following elementary matrix

$\boldsymbol{E} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ so that the product

$$\begin{aligned} EA &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} - 2a_{11} & a_{22} - 2a_{12} & a_{23} - 2a_{13} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \end{aligned}$$

Product of elementary transformations

- A series of Gaussian elimination steps can be represented as a product of elementary transformations acting on $\boldsymbol{A}$: $\boldsymbol{E}_m \boldsymbol{E}_{m-1} \ldots \boldsymbol{E}_1 \boldsymbol{A}$.
- The product of lower triangular matrices can be seen to be lower triangular, and the inverse of a lower triangular matrix can also be seen as a lower triangular matrix.
- Thus the action of Gaussian elimination operations can be seen in the following terms $\boldsymbol{L}^{-1}\boldsymbol{A} = \boldsymbol{U}$ where the product of the elementary transformations is represented as the inverse of a lower triangular matrix for notational convenience, and the right hand side U is an upper triangular matrix. Thus we have $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$.

Final argument

- Returning to our problem we are performing Gaussian elimination on the matrix $\boldsymbol{A}^T\boldsymbol{A}$ where $\boldsymbol{A}$ contains the basis vectors as its columns. Upon Gaussian elimination on the augmented matrix we reduce $[\boldsymbol{A}^T\boldsymbol{A}|\boldsymbol{A}^T]$ to get $[\boldsymbol{U}|\boldsymbol{L}^{-1}\boldsymbol{A}^T$ where $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$.
- Now we shall show that $\boldsymbol{Q}^T = \boldsymbol{L}^{-1}\boldsymbol{A}^T$ is an orthogonal matrix whose rows are orthogonal.
- Consider $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{L}^{-1}\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{L}^{-1})^T = \boldsymbol{U}(\boldsymbol{L}^{-1})^T =$ some upper triangular matrix
- But $\boldsymbol{Q}^T\boldsymbol{Q}$ is a symmetric matrix and can only be upper triangular if it is diagonal. Therefore $\boldsymbol{Q}$ is an orthogonal matrix whose columns are orthogonal. They can be normalized to obtain an orthonormal basis.

Lecture 4

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Matrix decompositions

- We studied vectors and how to manipulate them in preceding lectures.
- Mappings and transformations of vectors can be conveniently described in terms of operations performed by matrices.
- In this lecture we shall study three aspects of matrices: how to summarize matrices, how matrices can be decomposed, and how the decompositions can be used for matrix approximations.

Determinant and trace

A determinant of order $n \times n$ is a scalar associated with a $n \times n$ matrix and is denoted as follows: $det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$

We have a cofactor formula to calculate a determinant of order n: for $n = 1$, we have $det(\boldsymbol{A}) = a_{11}$. For $n \geq 2$ we have

$$\begin{aligned} D = det(\boldsymbol{A}) &= a_{j1}C_{j1} + a_{j2}C_{j2} + \ldots + a_{jn}C_{jn} \\ &= a_{1k}C_{1k} + a_{2k}C_{2k} + \ldots + a_{nk}C_{nk} \end{aligned}$$

The first line above represents expansion along the jth row, while the second line represents expansion along the kth column.

Cofactor formula

- In the preceding slide, the $C_{jk} = (-1)^{j+k} M_{jk}$ where M_{jk} represents the $n-1$ order determinant of the submatrix of $\boldsymbol{A}$ obtained by removing the jth row and kth column.
- M_{jk} is called the minor of a_{jk} in D and C_{jk} is called the cofactor of a_{jk} in D.
- Our definition for the determinant in the previous slide shows that the $n \times n$ determinant is defined in terms of $(n-1) \times (n-1)$ determinants which in turn are defined in terms of $(n-2) \times (n-2)$ determinants recursively.
- Let us examine the computation for a simple 3×3 determinant.

Example

Let us compute $D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$

For the entries in the second row the minors are $M_{21} = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix}$, $M_{22} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$, $M_{23} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}$. The cofactors are
$C_{21} = (-1)^{1+2} M_{21} = -M_{21}$, $C_{22} = (-1)^{2+2} M_{22} = M_{22}$,
$C_{23} = (-1)^{2+3} M_{23} = -M_{23}$.
Expanding along the second row we can write
$D = a_{21} C_{21} + a_{22} C_{22} + a_{23} C_{23}$.

Behaviour of determinant

Theorem: We can state the following for a nth order determinant under elementary row operations:(a) interchanging two rows multiplies the value of the determinant by -1, and (b) adding a multiple of a row to another row does not change the value of the determinant.

Proof Sketch Let us look at how to prove (a). The proof is by induction. The statement holds for $n = 2$ since $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$ whereas $\begin{vmatrix} c & d \\ a & b \end{vmatrix} = bc - ad = -(ad - bc)$. We make the induction hypothesis that the statement is true for all determinants of order $(n-1)$. Let D represent the original determinant and E represent the determinant with rows interchanged.

Proof sketch continued

Let us expand D and E along a **row that is not interchanged**. We have

$$
\begin{aligned}
D &= \sum_{k=1}^{k=n} (-1)^{j+k} a_{jk} M_{jk} \\
E &= \sum_{k=1}^{k=n} a_{jk} (-1)^{j+k} N_{jk}
\end{aligned}
$$

where N_{jk} in E is obtained by exchanging ttwo rows of the minor M_{jk} in D. M_{jk} and N_{jk} are determinants of order $n-1$ where one of the determinants has a pair of rows interchanged as compared to the other determinant. Therefore $M_{jk} = -N_{jk}$, and $D = -E$.

Proof sketch continued

Let us now look at adding multiples of a row to another row.

- Add c times row i to row j. Then we get a new determinant D' whose entries in the jth row has $a_{jk} + ca_{ik}$. Expanding the determinant D' we have $D' = \sum_{k=1}^{k=n}(a_{jk} + ca_{ik})(-1)^{j+k}M_{jk} = \sum_{k=1}^{k=n}(-1)^{j+k}a_{jk}M_{jk} + \sum_{k=1}^{k=n}(-1)^{j+k}ca_{ik}M_{jk}$.
- The summation can be written as $D' = D_1 + cD_2$ where $D_1 = D$ and D_2 represents the determinant of a matrix similar to the one we started out with except that rows j and i both have coefficients a_{ik} in them. Thus two rows of D_2 are equal - this makes $D_2 = 0 \rightarrow$ why is this?

Proof sketch continued

- In part (a) we showed that interchanging two rows will negate the determinant. If we interchange rows i and j we will get the same determinant since the two rows are identical. But one of the determinants must be the negative of the other. This is only possible when both determinants are zero. Thus $D_2 = 0$ and $D' = D$.
- **Bottom line** $\rightarrow$ Adding a multiple of one row to another row does not change the determinant.
- This will lead us to our next result.

Another result

Theorem An $n \times n$ matrix A has rank n if and only if its determinant is not equal to zero.

Proof sketch: $\boldsymbol{A}$ has full rank $\implies$ $\det(\boldsymbol{A}) \neq 0$: Gaussian elimination reduces $\boldsymbol{A}$ to upper triangular matrix $\boldsymbol{U} = (U_{ij})$ whose determinant is the product of all the elements U_{ii}. But we know that $\det(\boldsymbol{A}) = (-1)^s \det(\boldsymbol{U})$ where s is the number of row interchanges performed during Gaussian elimination. Since $\boldsymbol{A}$ has full rank, the columns of $\boldsymbol{A}$ are linearly independent and the only solution to $\boldsymbol{Ax} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$. The system $\boldsymbol{Ax} = \boldsymbol{0}$ has the same set of solutions as $\boldsymbol{Ux} = \boldsymbol{0}$, so this means $\boldsymbol{U}$ has only pivot columns. The pivots are all the elements U_{ii}. The product of all pivots is non-zero, and hence $\det(\boldsymbol{U}) = \det(\boldsymbol{A}) \neq 0$.

Another result

$\det(\boldsymbol{A}) \neq 0 \implies \boldsymbol{A}$ has full rank: If the determinant of $\boldsymbol{A}$ is non-zero, $\det(\boldsymbol{A}) = (-1)^s \det(\boldsymbol{U}) = \prod_{i=1}^{i=n} U_{ii}$ means that all the U_{ii} are non-zero. Therefore all the columns of $\boldsymbol{U}$ are pivot columns with the pivots being the U_{ii}. The pivot columns are all linearly independent, so the only solution to $\boldsymbol{Ux} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$. This is also the only solution to $\boldsymbol{Ax} = \boldsymbol{0}$ which means $\boldsymbol{A}$ has full rank.

Trace

- The trace of a $n \times n$ square matrix $\boldsymbol{A}$ is defined as $tr(\boldsymbol{A}) = \sum_{i=1}^{i=n} a_{ii}$, i.e the sum of the diagonal elements of A.
- The trace has the following properties:
  - $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$, for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$
  - $tr(\alpha \boldsymbol{A}) = \alpha tr(\boldsymbol{A})$, $\alpha \in \mathbb{R}$
  - $tr(\boldsymbol{I}_n) = n$
  - $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$ for $\boldsymbol{A} \in \mathbb{R}^{n \times k}, \boldsymbol{B} \in \mathbb{R}^{k \times n}$.
- The proofs of these properties are not difficult.

Characteristic polynomial

- For $\lambda \in \mathbb{R}$ and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ we can define $p_{\boldsymbol{A}}(\lambda) = \det(A - \lambda \boldsymbol{I})$ and show that it can be written as $c_0 + c_1\lambda + \ldots c_{n-1}\lambda^n + (-1)^n\lambda^n$ where $c_0, c_1 \ldots c_{n-1} \in R$.
- We can show that $c_0 = \det(\boldsymbol{A})$ and $c_{n-1} = (-1)^{n-1} tr(\boldsymbol{A})$
- To see that $c_0 = \det(\boldsymbol{A})$, set $\lambda = 0$ in $\det(\boldsymbol{A} - \lambda \boldsymbol{I})$ to get $p_{\boldsymbol{A}}(0) = \det(\boldsymbol{A}) = c_0$
- The formula for c_{n-1} takes a little bit of work - let us expand a 3×3 determinant $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix}$

Characteristic polynomial

- Expanding the determinant along the first row we see that the $(a_{11} - \lambda)C_{11}$ term contains the product $\prod_{i=1}^{i=3}(a_{ii} - \lambda)$ which contains the powers λ^3 and λ^2. The other contributors to the determinant i.e $a_{12}C_{12}$ and $a_{12}C_{13}$ expand into terms where the maximum power of $\lambda = 1$.
- Carrying this analogy over to the general case of $n > 3$ we see that expanding along the first row the first contributor to the determinant will have the term $\prod_{i=1}^{i=n}(a_{ii} - \lambda)$ and subsequent contributors will have a maximum power of λ^{n-2} as the minors for each such contributor will kill off a term containing λ in a given row and column.

Characteristic polynomial

- Thus in the determinant expansion to obtain the characteristic polynomial we see that coefficient to λ^{n-1} can only come from the expansion of $\prod_{i=1}^{i=n}(a_{ii} - \lambda)$ and can be seen to be seen to be $(-1)^{n-1}\sum_{i=1}^{i=n} a_{ii} = (-1)^{n-1} tr(\boldsymbol{A})$.
- As a corollary to this argument we can see that the coefficient to λ^n in the characteristic polynomial is $(-1)^n$
- We will use the characteristic polynomial to compute eigenvalues and eigenvectors.

Eigenvalues and eigenvectors

- Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n \setminus 0$ is the corresponding eigenvector of λ if $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$. This equation is called the eigenvalue equation.
- The following statements are equivalent:
  - λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.
  - There exists an $\boldsymbol{x} \in \mathbb{R}^n \setminus 0$ with $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$, or equivalently, $(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$ can be solved non-trivially, i.e., $\boldsymbol{x} \neq \boldsymbol{0}$.
  - $\mathrm{rank}(\boldsymbol{A} - \lambda \boldsymbol{I}_n) < n$.
  - $\det(\boldsymbol{A} - \lambda \boldsymbol{I}_n) = 0$.
- If $\boldsymbol{x}$ is an eigenvector corresponding to a particular eigenvalue λ, $c\boldsymbol{x}, c \in \mathbb{R} \setminus 0$ is also an eigenvector.

Eigenvalues and eigenvectors - example

- Consider the matrix $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. The characteristic polynomial $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (1-\lambda)^2 - 1$ and setting it to zero gives us the roots of the characetristic polynomial: $(1-\lambda)^2 - 1 = 0$ has roots $\lambda = 2, 0$.
- What are the eigenvectors? For $\lambda = 0$ we solve for $\boldsymbol{Ax} = 0\boldsymbol{x}$, so we find the nullspace of the matrix $\boldsymbol{A}$. Using Gaussian elimination we convert $\boldsymbol{Ax} = \boldsymbol{0}$ to $\boldsymbol{Ux} = \boldsymbol{0}$ where $\boldsymbol{U} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$. Thus we discover the eigenvector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ for $\lambda = 0$.
- Similarly we discover the eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for $\lambda = 2$.

Eigenvalues and eigenvectors - example

- The general procedure to find eigenvalues and eigenvectors is to first find the roots of the characteristic polynomials and then find the nullspaces of the matrices $\boldsymbol{A} - \lambda \boldsymbol{I}$ for the different roots λ.
- Does every $n \times n$ matrix have a full set of eigenvectors, i.e n eigenvectors?
- Look at $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. What are its eigenvalues and eigenvectors?
- **Point to ponder** Looking at the equation $\boldsymbol{Ax} = \lambda \boldsymbol{x}$ it seems that the action of $\boldsymbol{A}$ on $\boldsymbol{x}$ is to preserve the direction of $\boldsymbol{x}$ but scale it up or down according to λ. Does this mean that a rotation matrix has no eigenvalues and eigenvectors?

Some additional properties

- λ is an eigenvalue of $\boldsymbol{A}$ if and only if λ is a root of the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ of $\boldsymbol{A}$. This can be easily seen as a consequence of the definition of $p_{\boldsymbol{A}}(\lambda)$.
- For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the set of eigenvectors corresponding to an eigenvalue λ spans a subspace of $\mathbb{R}^n$ called the Eigenspace of $\boldsymbol{A}$ with respect to λ and is denoted by E_λ.
- The set of all eigenvalues of $\boldsymbol{A}$ is called the spectrum of $\boldsymbol{A}$.
- Look at the eigenvalues and eigenspace of the $n \times n$ identity matrix $\boldsymbol{I}_n$. It has one eigenvalue $\lambda = 1$ and the eigenspace is $\mathbb{R}^n$. Every canonical vector is a basis vector for the eigenspace.

Some additional properties

- A matrix and its transpose have the same eigenvalues. To see this, first note that $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$. Then $\det(\boldsymbol{A} - \lambda I) = \det((\boldsymbol{A} - \lambda \boldsymbol{I})^T) = \det(\boldsymbol{A}^T - \lambda \boldsymbol{I}^T) = \det(\boldsymbol{A}^T - \lambda \boldsymbol{I})$. The last expression in the chain of equalities is the characteristic polynomial for $p_{\boldsymbol{A}^T}(\lambda)$. Thus we have $p_{\boldsymbol{A}}(\lambda) = p_{\boldsymbol{A}^T}(\lambda)$ which means the characteristic polynomials are equal and so the roots of the polynomials or the eigenvalues must be equal.
- The eigenspace E_λ is the nullspace of $\boldsymbol{A} - \lambda \boldsymbol{I}$.
- Symmetric, positive-definite matrices always have positive, real eigenvalues.

Some theorems

- The eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n$ of a $n \times n$ matrix $\boldsymbol{A}$ with n **distinct** eigenvalues are linearly independent $\rightarrow$ why?
- Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ we can show that $\boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a symmetric, positive-definite matrix when the rank of $\boldsymbol{A} = n$. Why is this true? Clearly $\boldsymbol{A}^T\boldsymbol{A}$ is a symmetric matrix and it is positive definite since $\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|^2 > 0 \;\; \forall \boldsymbol{x} \in \mathbb{R}^n \setminus 0$ since the nullspaces of $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}$ are the same, and $\boldsymbol{A}$ is a full column rank matrix.
- The matrix $\boldsymbol{A}^T\boldsymbol{A}$ is important in machine learning since it figures in the least-squares solution to a data matrix represented as $\boldsymbol{A}$ where n represents the number of features and m is the number of data vectors.

Spectral theorem

Theorem: If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric there exists an orthonormal basis of the corresponding vector space V consisting of the eigenvectors of $\boldsymbol{A}$, and each eigenvalue is real.
Proof: We will not attempt a full proof of this theorem but provide some intuitions about why it is true. The theorem relies on the following three statements, shown in the next slide.

Spectral theorem

- All roots of the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ are real.
- For each eigenvalue λ we can compute an orthonormal basis for its eigenspace. We can string together the orthonormal bases for the different eigenvalues of $\boldsymbol{A}$ to come up with the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \ldots$
- The dimension of the eigenspace E_λ, called its geometric multiplicity, is the same as the algebraic multiplicity of λ which is the number of times λ appears as a root of the characteristic polynomial.
- All the basis vectors from the different Eigenspaces combine to provide an orthonormal basis for $\mathbb{R}^n$.

Complex vectors

- In the old formulation with real vectors, length-squared according to the Euclidean norm was $x_1^2 + x_2^2 + \ldots x_n^2$. If the x_i are complex we should take length-squared to be $|x_1|^2 + |x_2|^2 + \ldots |x_n|^2$ where $|.|$ denotes modulus. For the complex number $a + bi$, the modulus is $\sqrt{(a+bi)(a-bi)} = \sqrt{a^2 + b^2}$
- For complex vectors we would like to preserve the idea as possible that $\|x\|^2 = \boldsymbol{x}^T \boldsymbol{x}$. If we keep the old definition of inner product for complex vectors we will not get a real number as length as shown in the next bullet.
- Let $\boldsymbol{x} = \begin{bmatrix} 1+i \\ 2+i \end{bmatrix}$. We have $\boldsymbol{x}^T \boldsymbol{x} = (1+i)^2 + (2+i)^2 = 1 + 2i + i^2 + 4 + 4i + i^2 = 6i + 3.$

Hermitian matrices

- We modify the inner product between two complex vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to $\boldsymbol{x}^H\boldsymbol{y}$, where $\boldsymbol{x}^H = \overline{\boldsymbol{x}}^T$.
- Now $\boldsymbol{x}^H\boldsymbol{x} = \overline{x_1}x_1 + \ldots \overline{x_n}x_n = \|x\|^2$ according to the new definition of length.
- A Hermitian matrix is a generalization of a symmetric matrix.
- Instead of requiring $\boldsymbol{A}^T = \boldsymbol{A}$, we say a matrix is Hermitian if it is equal to its conjugate-transpose, ie $\boldsymbol{A}$ is a Hermitian matrix if $\boldsymbol{A}^H = \boldsymbol{A}$ or $\overline{\boldsymbol{A}}^T = \boldsymbol{A}$
- As an example consider the matrix $\boldsymbol{A} = \begin{bmatrix} 1 & 3-i \\ 3+i & 4 \end{bmatrix}$. It is a Hermitian matrix since $\boldsymbol{A}^H = \begin{bmatrix} 1 & 3-i \\ 3+i & 4 \end{bmatrix} = \boldsymbol{A}$.

Spectral theorem

We shall now show that all eigenvalues for a symmetric matrix are real. Let $\boldsymbol{Ax} = \lambda \boldsymbol{x}$. Then premultiplying with $\boldsymbol{x}^H$ on both sides we have $\boldsymbol{x}^H \boldsymbol{Ax} = \lambda \boldsymbol{x}^H \boldsymbol{x}$

Now $\boldsymbol{x}^H \boldsymbol{Ax}$ is a 1×1 matrix. Taking the Hermitian of this matrix we have $(\boldsymbol{x}^H \boldsymbol{Ax})^H = \boldsymbol{x}^H \boldsymbol{A}^H \boldsymbol{x} = \boldsymbol{x}^H \boldsymbol{Ax}$, so the Hermitian of the matrix is itself which means that the matrix is real.

On the right hand side we note that $\boldsymbol{x}^H \boldsymbol{x}$ is real, so this means that λ must be real.

Spectral theorem

Let us show that eigenvectors belonging to different eigenvalues are orthogonal. Let $\boldsymbol{Ax} = \lambda \boldsymbol{x}$ and $\boldsymbol{Ay} = \mu \boldsymbol{y}$. Then we have

$$\begin{aligned} \boldsymbol{y}^H \boldsymbol{Ax} &= \lambda \boldsymbol{y}^H \boldsymbol{x} \\ \boldsymbol{x}^H \boldsymbol{Ay} &= \mu \boldsymbol{x}^H \boldsymbol{y} \end{aligned}$$

But $\boldsymbol{x}^H \boldsymbol{Ay} = (\boldsymbol{y}^H \boldsymbol{A}^H \boldsymbol{x})^H = (\boldsymbol{y}^H \boldsymbol{Ax})^H = \lambda \boldsymbol{x}^H \boldsymbol{y}$. We already know that $\boldsymbol{x}^H \boldsymbol{Ay} = \mu \boldsymbol{x}^H \boldsymbol{y}$. This means $\lambda \boldsymbol{x}^H \boldsymbol{y} = \mu \boldsymbol{x}^H \boldsymbol{y}$. Since $\lambda \neq \mu$, this must mean $\boldsymbol{x}^H \boldsymbol{y} = 0$.
This shows that eigenvectors corresponding to different eigenvalues are orthogonal.

Spectral theorem

- So we see that the eigenvalues of a symmetric matrix are real and eigenvectors belonging to different eigenvalues are orthogonal.
- This suggests that one can string together all the orthonormal bases for the different eigenvalues and get an orthonormal basis for $\mathbb{R}^n$.
- But who is to say that when we string together the basis vectors for all the eigenvalues, we will have enough vectors to describe $\mathbb{R}^n$? We need n basis vectors and might end up having fewer than n vectors.
- If the eigenvalues are all different, we can see that we will indeed have enough basis vectors. But what about when there are repeating eignevalues?

Spectral theorem

- We need one more piece to complete the puzzle and show that we will have enough eigenvectors to complete the orthonormal basis - this part we shall not prove!
- As a consequence of the spectral theorem we can write a real symmetric matrix $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^T$ where $\boldsymbol{Q}$ is an orthonormal basis (think orthonormal basis vectors for example), and Λ is a diagonal matrix consisting of non-zero entries only along the diagonal.
- The spectral theorem can be used in a machine learning context since we can take the data matrix $\boldsymbol{A}$ and create a symmetric matrix out of it - $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^T$ which are both used in Singular-Value Decomposition and PCA.

Trace and eigenvalues

- We can show that the sum of the eigenvalues of a matrix is equal to the trace of the matrix, i.e $\sum_{i=1}^{i=n} \lambda_i = \sum_{i=1}^{i=n} a_{ii}$. To see why this is true, note that the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ can be written as $\prod_{i=1}^{i=n}(\lambda_i - \lambda)$. The coefficient to λ^{n-1} in this expansion is $(-1)^{n-1}\sum_{i=1}^{i=n} \lambda_i$. Early on in this lecture we showed from a direct expansion of the determinant that the coefficient of λ^{n-1} is $(-1)^{n-1}\sum_{i=1}^{i=n} a_{ii}$. Thus we have our result.
- The product of all eigenvalues is the determinant of the matrix, i.e $\det(\boldsymbol{A}) = \prod_{i=1}^{i=n} \lambda_i$. To see why this is true, let us once again look at the factorisation of $p_{\boldsymbol{A}}(\lambda)$ as $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = p_{\boldsymbol{A}}(\lambda) = \prod_{i=1}^{i=n}(\lambda_i - \lambda)$. Setting $\lambda = 0$ in this equation gives the result.

Cholesky decomposition

Theorem A symmetric, positive-definite matrix A can be factorized into a product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is a lower-triangular matrix with positive elements.
For an example 3×3 matrix we can write

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

We can solve for the elements of the lower triangular matrix to get $l_{11} = \sqrt{a_{11}}$, $l_{22} = \sqrt{a_{22} - l_{21}^2}$, $l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)}$.
For the elements below the diagonal we have $l_{21} = \frac{a_{21}}{l_{11}}$, $l_{31} = \frac{a_{31}}{l_{11}}$ and $l_{32} = \frac{a_{32} - l_{31} l_{21}}{l_{22}}$.

An application of Cholesky decomposition

- In the Data Science / Machine Learning context, distributions on data are frequently multivariate Gaussian.
- Multivariate Gaussian distributions are governed by a covariance matrix which is symmetric, positive-definite.
- We may need to draw samples from such distributions which is where the Cholesky decomposition finds an important application.
- To generate a sample from a multivariate Gaussian distribution, we factor the covariance matrix into its Cholesky factor $\boldsymbol{L}$, generate a Gaussian random vector $\boldsymbol{x}$ on independent variables which is easy to do, and compute $\boldsymbol{Lx}$ which will be a sample according to the required distribution.

Lecture 5

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction

- In the previous lecture, we discussed eigenvalues and eigenvectors of matrices
- In this lecture, we will look at two related methods for factorizing matrices into canonical forms.
- The first one is known as Eigenvalue decompostion. It uses the concepts of eigenvalues and eigenvectors to generate the decomposition
- The second method known as singular value decomposition or SVD is applicable to all matrices

Diagonal Matrices

- A diagonal matrix is a matrix that has value zero on all off diagonal elements.

$$\mathbf{D} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$$

- For a diagonal matrix $\mathbf{D}$, the determinant is the product of its diagonal entries.
- A matrix power $\mathbf{D}^k$ is given by each diagonal element raised to the power k.
- Inverse of a diagonal matrix is obtained by taking inverse of non-zero diagonal entry.

Diagonalizable Matrices

- A matrix $A \in \mathbb{R}^{n \times n}$ is diagonalizable if there exists an invertible matrix $P \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$
- In the definition of diagonalization, it is required that $\mathbf{P}$ is an invertible matrix. Assume $p_1, p_2, \ldots.., p_n$ are the n columns of $\mathbf{P}$
- Rewriting we get $\mathbf{AP} = \mathbf{PD}$. By observing that $\mathbf{D}$ is a diagonal matrix, we can simplify as

$$\mathbf{A}p_i = \lambda_i p_i$$

where λ_i is the ith diagonal entry in $\mathbf{D}$.

Diagonalizable Matrix

- Consider a square matrix

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

- Consider the invertible matrix

$$\mathbf{P} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$$

- Now consider the product $\mathbf{P}^{-1}\mathbf{AP}$ as follows

$$\begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 5 \end{bmatrix}$$

Eigendecomposition of a matrix

- Recall the existence of eigenvalues and eigenvectors for square matrices
- Eigenvalues can be used to create a matrix decomposition known as Eigenvalue Decomposition
- A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be factored into

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$$

- where $\mathbf{P}$ is an invertible matrix of eigenvectors of A assuming we can find n eigenvectors that form a basis of $\mathbb{R}^n$
- and $\mathbf{D}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathbf{A}$

Example of Eigendecomposition

Let us compute the eigendecomposition of the matrix A

$$\mathbf{A} = \begin{bmatrix} 2.5 & -1 \\ -1 & 2.5 \end{bmatrix}$$

- Step 1: Find the eigenvalues and eigenvectors

$$\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} 2.5 - \lambda & -1 \\ -1 & 2.5 - \lambda \end{bmatrix}$$

- The characteristic equation is given by $\det(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$
- This leads to the equation $\lambda^2 - 5\lambda + \frac{21}{4} = 0$
- Solving the quadratic equation gives us $\lambda_1 = 3.5$ and $\lambda_2 = 1.5$

Example of Eigendecomposition

- The eigenvector corresponing to $\lambda_1 = 3.5$ is derived as $p_1 = \begin{bmatrix} -\frac{1}{\sqrt{2})} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$
- The eigenvector corresponing to $\lambda_1 = 1.5$ is derived as $p_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$
- Step 2 : Construct the matrix **P** to diagonalize **A**

$$\mathbf{P} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Example of Eigendecomposition

- The inverse of matrix P is given by

$$\mathbf{P}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

- The eigendecompostion of the matrix A is given by

$$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 3.5 & 0 \\ 0 & 1.5 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

- In summary we have obtained the required matrix factorization using eigenvalues and eigenvectors.

Symmetric Matrices and Diagonalizability

- Recall that a matrix A is called symmetric matrix if $\mathbf{A} = \mathbf{A}^{\mathbf{T}}$

$$\mathbf{A} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$$

- A Symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can always be diagonalized.
- This follows directly from the spectral theorem discussed in previous lecture
- Moreover the spectral theorem states that we can find an orthogonal matrix $\mathbf{P}$ of eigenvectors of $\mathbf{A}$.

Motivation for Singular Value Decomposition

- The singular value decomposition or (SVD) of a matrix is a central matrix decomposition method in linear algebra.
- The eigenvalue decomposition is applicable to square matrices only.
- The singular value decomposition exists for all rectangular matrices
- SVD involves writing a matrix as a product of three matrices $\mathbf{U}, \mathbf{\Sigma}$ and $\mathbf{V}^{\mathbf{T}}$.
- The three component matrices are derived by applying eigenvalue decomposition discussed previously

Singular Value Decomposition Theorem

- Let $\mathbf{A} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ be a rectangular matrix. Assume that $\mathbf{A}$ has rank r.
- The Singular value decomposition of $\mathbf{A}$ is defined as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$$

- $\mathbf{U} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix with column vectors u_i where $i = 1, ..., m$
- $\mathbf{V} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix with column vectors v_j where $j = 1, ..., n$
- $\mathbf{\Sigma}$ is an m $\times$ n matrix with $\mathbf{\Sigma_{ii}} = \sigma_{\mathbf{i}} > \mathbf{0}$
- The diagonal entries $\sigma_i, i = 1, ..., r$ of $\mathbf{\Sigma}$ are called the singular values.
- By convention, the singular values are ordered i.e $\mathbf{\Sigma}_{ii} > \mathbf{\Sigma}_{jj}$ where $i < j$.

Properties of Σ

- The singular value matrix $\mathbf{\Sigma}$ is unique.
- Observe that the $\mathbf{\Sigma} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$matrix is rectangular. In particular, $\mathbf{\Sigma}$ is of the same size as $\mathbf{A}$.
- This means that $\mathbf{\Sigma}$ has a diagonal submatrix that contains the singular values and needs additional zero padding.
- Specifically, if $m > n$, then the matrix $\mathbf{\Sigma}$ has diagonal structure up to row n and then consists of zero rows.
- If $m < n$, the matrix Σ has a diagonal structure up to column m and columns that consist of 0 from $m + 1$ to n.

Construction of V

- It can be observed that

$$\mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$$

- Since $A^T A$ has the following eigendecomposition

$$\mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{\mathbf{T}}$$

- Therefore, the eigenvectors of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ that compose $\mathbf{P}$ are the right-singular vectors $\mathbf{V}$ of $\mathbf{A}$.
- The eigenvalues of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ are the squared singular values of $\mathbf{\Sigma}$

Construction of U

- It can be observed that

$$\mathbf{A}\mathbf{A}^{\mathbf{T}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}\mathbf{V}\boldsymbol{\Sigma}^{\mathbf{T}}\mathbf{U}^{\mathbf{T}}$$

- Since $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ has the following eigendecomposition

$$\mathbf{A}\mathbf{A}^{\mathbf{T}} = \mathbf{S}\mathbf{D}\mathbf{S}^{\mathbf{T}}$$

- Therefore, the eigenvectors of $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ that compose $\mathbf{S}$ are the left-singular vectors $\mathbf{U}$ of $\mathbf{A}$

Construction of U continued

- $\mathbf{A} = \mathbf{U\Sigma V^T}$ can be rearranged to obtain a simple formulation for u_i
- By postmultiplying by $\mathbf{V}$ we get $\mathbf{AV} = \mathbf{U\Sigma V^T V}$
- By observing that $\mathbf{V}$ is orthogonal we obtain a simple form

$$\mathbf{AV} = \mathbf{U\Sigma}$$

- This is equivalent to the following

$$u_i = \frac{1}{\sigma_i}\mathbf{A}v_i \quad \forall i = 1, 2, ..., r$$

Computing Singular Value Decomposition 1

- We want to find SVD of the following rectangular matrix $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}$$

- Let us consider the matrix $\mathbf{A}^T\mathbf{A}$ derived from $\mathbf{A}$ given by

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

- It is a symmetric matrix

Computing Singular Value Decomposition 2

- Derive the eigendecomposition of $\mathbf{A}^T\mathbf{A}$ in the form PDP^T
- The matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

- The matrix $\mathbf{D}$ is given by

$$\mathbf{D} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Computing Singular Value Decomposition 3

Now we construct the singular value matrix $\mathbf{\Sigma}$

- The matrix $\mathbf{\Sigma}$ has the dimension same as $\mathbf{A}$. In this case $\mathbf{\Sigma}$ is hence a 2×3 matrix.
- The diagonal entries of submatrix is obtained by taking square root of 6 and 1 respectively
- Singular-value matrix $\mathbf{\Sigma}$ is given by:

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

- The last column is a column of zeros only

Computing Singular Value Decomposition 4

Left singular vectors as the normalized image of the right singular vectors. Recall that $u_i = \frac{1}{\sigma_i}\mathbf{A}v_i$

- The first vector

$$u_1 = \frac{1}{\sigma_1}\mathbf{A}v_1 = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{bmatrix}$$

- The second vector

$$u_2 = \frac{1}{\sigma_2}\mathbf{A}v_2 = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}$$

Final Step : Combining U, Σ and V

We compile all the three matrices together to generate the SVD

- $$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix}^T$$
- The matrix **U** is an 2×2 matrix satifying orthogonality property.
- The matrix **V** is an 3×3 matrix satifying orthogonality property.

Comparing SVD and EVD

- The left-singular vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}\mathbf{A}^{\mathbf{T}}$
- The right-singular vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$
- The non-zero singular values of $\mathbf{A}$ are the square roots of the nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$.
- The SVD always exists for any matrix in $\mathbb{R}^{m\times n}$
- The eigendecomposition is only defined for square matrices in $\mathbb{R}^{n\times n}$ and only exists if we can find a basis of eigenvectors of $\mathbb{R}^n$

Comparing SVD and EVD

- The vectors in the eigendecomposition matrix $\mathbf{P}$ are not necessarily orthogonal.
- On the other hand, the vectors in the matrices $\mathbf{U}$ and $\mathbf{V}$ in the SVD are orthonormal.
- Both the eigendecomposition and the SVD are compositions of three linear mappings:
- A key difference between the eigendecomposition and the SVD is that in the SVD, domain and codomain can be of different dimensions
- In the SVD, the left and right singular vector matrices $\mathbf{P}$ and $\mathbf{P}$ are generally not inverse of each other.

Comparing SVD and EVD 3

- In the eigendecomposition, the matrices in decomposition are inverse of each other
- In the SVD, the entries in the diagonal matrix $\mathbf{\Sigma}$ are all real and nonnegative,
- In eigendecomposition diagonal matrix entries need not be real always.
- The leftsingular vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}\mathbf{A}^T$
- The rightsingular vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$.

Matrix Approximation

- We considered the SVD as a way to factorize $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$ into the product of three matrices, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\boldsymbol{\Sigma}$ contains the singular values on its main diagonal.
- Instead of doing the full SVD factorization, we will now investigate how the SVD allows us to represent a matrix $\mathbf{A}$ as a sum of simpler matrices $\mathbf{A}_i$
- This representation which lends itself to a matrix approximation scheme that is cheaper to compute than the full SVD.

Matrix Approximation continued

- A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank r can be written as a sum of rank-1 matrices so that $\mathbf{A} = \sum_{i=1}^{r} \sigma_i u_i v_i^T$
- The diagonal structure of the singular value matrix $\mathbf{\Sigma}$ multiplies only matching left and right singular vectors $u_i v_i^T$ and scales them by the corresponding singular value σ_i.
- All terms $\sigma_i u_i v_i^T$ vanish for $i \neq j$ because $\mathbf{\Sigma}$ is a diagonal matrix.
- Any term for $i > r$ would vanish because the corresponding singular value is 0.

Rank k Approximation

- We summed up the r individual rank-1 matrices to obtain a rank r matrix $\mathbf{A}$.
- If the sum does not run over all matrices $A_i \;\; i = 1, ..., r$ but only up to an intermediate value k we obtain a rank-k approximation
- The approximation represented by $\hat{\mathbf{A}}(k)$ is defined as follows

$$\hat{\mathbf{A}}(k) = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

- To measure the difference between $\mathbf{A}$ and its rank-k approximation we need the notion of a norm which is introduced next

Spectral Norm of a matrix

- We introduce the notation of a subscript in the matrix norm
- Spectral Norm of a Matrix. For $x \in \mathbb{R}^n$, $x \neq \mathbf{0}$, the spectral norm norm of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\|\mathbf{A}\|_2 = \max_x \frac{\|\mathbf{A}x\|_2}{\|x\|_2}$$

  where $\|y\|_2$ is the euclidean norm of y
- Theorem : The spectral norm of a matrix $\mathbf{A}$ is its largest singular value

Example : Spectral Norm of a matrix

- Example : Consider the following matrix **A**

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

- Singular value decomposition of this matrix will provide the matrix Σ as follows

$$\mathbf{\Sigma} = \begin{bmatrix} 5.465 & 0 \\ 0 & 0.366 \end{bmatrix}$$

- The 2 singular values are 5.4650 and 0.366.
- By definition the spectral norm is the largest singular value.
- Hence, the spectral norm is 5.4650

Lecture 6

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction

Many algorithms in machine learning optimize an objective function with respect to a set of desired model parameters that control how well a model explains the data: Finding good parameters can be phrased as an optimization problem.

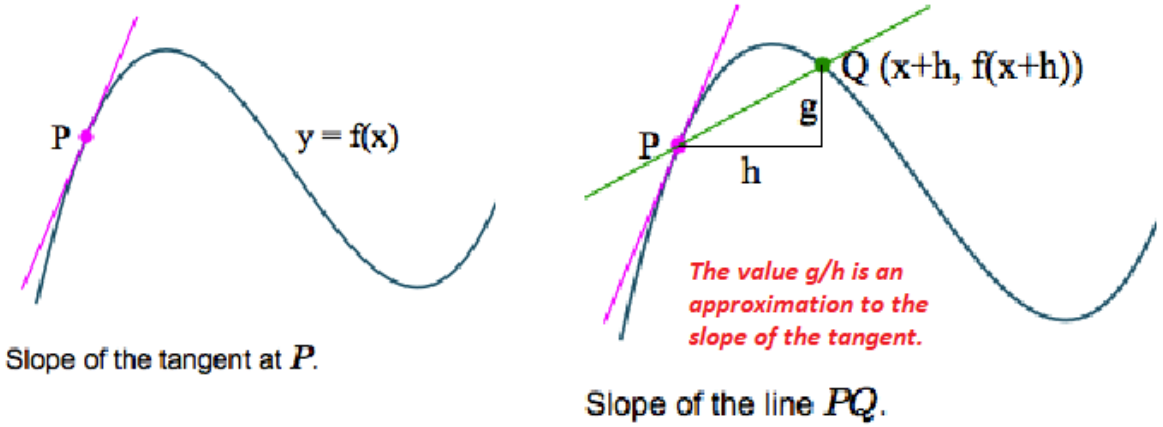
Examples include: linear regression, where we look at curve-fitting problems and optimize linear weight parameters to maximize the likelihood; neural-network auto-encoders for dimensionality reduction and data compression.

Differentiation of Univariate Functions

For $h > 0$, the derivative of f at x is defined as the limit

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$

The derivative of f points in the direction of steepest ascent of f.



Derivative of a Polynomial

To compute the derivative of $f(x) = x^n \; n \in N$ using the definition

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\
&= \lim_{h \to 0} \frac{(x+h)^n - x^n}{h} \\
&= \lim_{h \to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h} \\
&= \lim_{h \to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^i}{h}
\end{aligned} \tag{2}$$

Derivative of a Polynomial

$$
\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^{i-1} \\
&= \lim_{h\to 0} \binom{n}{1} x^{n-1} + \lim_{h\to 0} \sum_{i=2}^{n} \binom{n}{i} x^{n-i} h^{i-1} \\
&= nx^{n-1}
\end{aligned}
\tag{3}
$$

Taylor polynomial

The Taylor polynomial is a representation of a function f as an finite sum of terms. These terms are determined using derivatives of f evaluated at x_0.

Definition: The Taylor polynomial of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{4}$$

where $f^{(k)}(x_0)$ is the kth derivative of f at x_0 which we assume exists.

Taylor series

Definition: The Taylor series of smooth (continuously differentiable infinite many times) function $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5}$$

For $x_0 = 0$, we obtain the Maclaurin series as a a special instance of the Taylor series.

Remark: In general, a Taylor polynomial of degree n is an approximation of a function, which does not need to be a polynomial. The Taylor polynomial is similar to f in a neighborhood around x_0. However, a Taylor polynomial of degree n is an exact representation of a polynomial f of degree $k \leq n$ since all derivatives $f^{(i)} = 0$, for $i > k$.

Taylor Polynomial example

Consider the polynomial $f(x) = x^4$. Find the Taylor polynomial T_6 evaluated at $x_0 = 1$.
We compute $f^{(k)}(1)$ for $k = 0, 1, 2..., 6$
$f(1) = 1$, $f^{'}(1) = 4$, $f^{''}(1) = 12$, $f^{(3)}(1) = 24$, $f^{(4)}(1) = 24$, $f^{(5)}(1) = 0$, $f^{(6)}(1) = 0$. The desired Taylor polynomial is

$$\begin{aligned} T_6(x) &= \sum_{k=0}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \\ &= 1 + 4(x-1) + 12(x-1)^2 + 24(x-1)^3 + 24(x-1)^4 \\ &= x^4 = f(x) \end{aligned} \tag{6}$$

we obtain an exact representation of the original function.

Taylor Series example

Consider the smooth function $f(x) = sin(x) + cos(x)$. We compute Taylor series expansion of f at $x_0 = 0$, which is the Maclaurin series expansion of f. We obtain the following derivatives:

$f(0) = sin(0) + cos(0) = 1$

$f^{'}(0) = cos(0) - sin(0) = 1$

$f^{''}(0) = -sin(0) - cos(0) = -1$

$f^{(3)}(0) = -cos(0) + sin(0) = -1$

$f^{(4)}(0) = sin(0) + cos(0) = f(0) = 1$

The coefficients in our Taylor series are only ± 1 (since $sin(0) = 0$), each of which occurs twice before switching to the other one. Furthermore, $f^{(k+4)}(0) = f^{k}(0)$

Taylor Series example

Therefore, the full Taylor series expansion of f at $x_0 = 0$ is given by

$$
\begin{aligned}
T_\infty(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \\
&= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \dots \\
&= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \dots x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \dots \\
&= \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k)!}x^{2k} + \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k+1)!}x^{2k+1} \\
&= cos(x) + sin(x)
\end{aligned}
\tag{7}
$$

Differentiation Rules

We denote the derivative of f by $f^{'}$

- Product Rule: $(f(x)g(x))^{'} = f^{'}(x)g(x) + f(x)g^{'}(x)$
- Sum Rule: $(f(x) + g(x))^{'} = f^{'}(x) + g^{'}(x)$
- Quotient Rule: $(\frac{f(x)}{g(x)})^{'} = \frac{f^{'}(x)g(x) - f(x)g^{'}(x)}{(g(x))^2}$
- Chain Rule: $(g(f(x))^{'} = (g \circ f)^{'}(x) = g^{'}(f(x))f^{'}(x)$

Example: Chain Rule

Compute the derivative of function $h(x) = (2x+1)^4$

$h(x) = (2x+1)^4 = g(f(x))$

$f(x) = 2x+1,$

$g(f) = f^4$

Derivatives of f and g are

$f^{'}(x) = 2$

$g^{'}(f) = 4f^3$

$h^{'}(x) = g^{'}(f)f^{'}(x) = (4f^3).2 = 8(2x+1)^3$

Partial Differentiation and Gradients

Differentiation applies to functions f of a scalar variable $x \in R$. In the following, we consider the general case where the function f depends on one or more variables $x \in R^n$, e.g.,$f(x) = f(x_1, x_2)$. The generalization of the derivative to functions of several variables is the gradient. We find the gradient of the function f with respect to x by varying one variable at a time and keeping the others constant. The gradient is then the collection of these partial derivatives.

Partial derivatives and Gradients

Definition: For a function $f : \mathbb{R}^n \to \mathbb{R}$, $x \to f(x)$, $x \in R^n$ of n variables $x_1, \ldots, x_n$ we define the *partial derivatives* as

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$

$$\frac{\partial f}{\partial x_2} = \lim_{h \to 0} \frac{f(x_1, x_2 + h, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$

$$\vdots$$

$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_n + h) - f(x_1, x_2, \ldots, x_n)}{h}$$

We collect them in the row vector called the gradient of f or Jacobian

$$\Delta_x f = gradf = \frac{df}{dx} = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}\right] \tag{8}$$

Example 1: Find the partial derivatives of $f(x, y) = (x + 2y^3)^2$

$$\frac{\partial f(x, y)}{\partial x} = 2(x + 2y^3)\frac{\partial(x + 2y^3)}{\partial x} = 2(x + 2y^3) \tag{9}$$

$$\frac{\partial f(x, y)}{\partial y} = 2(x + 2y^3)\frac{\partial(x + 2y^3)}{\partial y} = 12y^2(x + 2y^3) \tag{10}$$

here we used the chain rule to compute the partial derivatives.

Example 2

Find the partial derivatives of $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3 \tag{11}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2 \tag{12}$$

So the gradient is then

$$\frac{df}{dx} = [\frac{\partial f(x_1, x_2)}{\partial x_1}, \frac{\partial f(x_1, x_2)}{\partial x_2}] = [2x_1 x_2 + x_2^3, x_1^2 + 3x_1 x_2^2] \in \mathbb{R}^{1\times 2} \tag{13}$$

Basic rules of partial differentiation

When we compute derivatives with respect to vectors $x \in \mathbb{R}^n$ we need to pay attention: Our gradients now involve vectors and matrices, and matrix multiplication is not commutative i.e., the order matters.

$$\text{Product rule: } \frac{\partial}{\partial x}(f(x)g(x)) = \frac{\partial f}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x} \tag{14}$$

$$\text{Sum rule: } \frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} \tag{15}$$

$$\text{chain rule: } \frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x}(g(f(x)) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x} \tag{16}$$

Chain Rule

Consider a function $f : \mathbb{R} \to \mathbb{R}$ of two variables x_1, x_2. Furthermore, $x_1(t)$ and $x_2(t)$ are themselves functions of t.

To compute the gradient of f with respect to t, we need to apply the chain rule for multivariate functions as

$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{17}$$

where d denotes the gradient and ∂ partial derivatives.

Example

Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$ then

$$\begin{aligned}\frac{df}{dt} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \\ &= 2\sin t\frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t} \\ &= 2\sin t\cos t - 2\sin t = 2\sin t(\cos t - 1)\end{aligned}$$

is the corresponding derivative of f with respect to t.

If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivatives:

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s} \tag{18}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{19}$$

and the gradient is obtained by the matrix multiplication

$$\begin{aligned}\frac{df}{d(s,t)} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial(s,t)} \\ &= \begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2}\end{bmatrix}\begin{bmatrix}\frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t}\end{bmatrix}\end{aligned}$$

Gradients of Vector-Valued Functions

We have discussed partial derivatives and gradients of functions $f : \mathbb{R}^n \to \mathbb{R}$ mapping to the real numbers. Now we will generalize the concept of the gradient to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.
For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $x = [x_1, \ldots, x_n]^T$ corresponding vector of function values is given as

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m \tag{20}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$

Gradients of Vector-Valued Functions

Therefore, the partial derivative of a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ w.r.t. $x_i \in R$, $i = 1, \ldots n$ is given as the vector

$$\frac{\partial f}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix}$$

$$= \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_1(x)}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_m(x)}{h} \end{bmatrix} \in \mathbb{R}^m$$

Gradients of Vector-Valued Functions

We know that the gradient of f with respect to a vector is the row vector of the partial derivatives. Every partial derivative $\frac{\partial f}{\partial x_i}$ is itself a column vector. Therefore, we obtain the gradient of $f : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\begin{aligned}\frac{df(x)}{dx} &= \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ & \vdots & \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}\end{aligned}$$

Example 1: Gradients of Vector-Valued Functions

Given $f(x) = Ax, \quad f(x) \in \mathbb{R}^M, \quad A \in \mathbb{R}^{M \times N}, \quad x \in \mathbb{R}^N$
Since $f : \mathbb{R}^N \to \mathbb{R}^M$, it follows that $df/dx \in \mathbb{R}^{M \times N}$. To compute the gradient we determine the partial derivatives of f w.r.t x_j:

$$f_i(x) = \sum_{i=1}^{N} A_{ij} x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij} \tag{21}$$

We obtain the gradient using Jacobian

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ & \vdots & \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} \ldots A_{1N} \\ \vdots \\ A_{M1} \ldots A_{MN} \end{bmatrix} = A \in \mathbb{R}^{M \times N} \tag{22}$$

Example 2: Gradients of Vector-Valued Functions

Consider the function $h : \mathbb{R} \to \mathbb{R}, \quad h(t) = (f \circ g)(t)$ with $f(x) = exp(x_1 x_2^2)$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix} \tag{23}$$

and compute the gradient of h w.r.t. t. Since $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}^2$ we note that

$$\frac{\partial f}{\partial x} \in \mathbb{R}^{1\times 2} \text{ and } \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1} \tag{24}$$

The desired gradient is computed by applying the chain rule:

$$\frac{dh}{dt} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} exp(x_1 x_2^2)x_2^2 & 2exp(x_1 x_2^2)x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix}$$

$$= exp(x_1 x_2^2)(x_2^2(\cos t - t\sin t) + 2x_1 x_2(\sin t + t\cos t))$$

where $x_1 = t\cos t$ and $x_2 = t\sin t$;

Lecture 7

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction

- In last lecture, we discussed about differentiation of univariate functions, partial differentiation, gradients and gradients of vector valued functions.
- Now we will look into gradients of matrices and some useful identities for computing gradients.
- Finally, we will discuss back propagation and automatic differentiation.

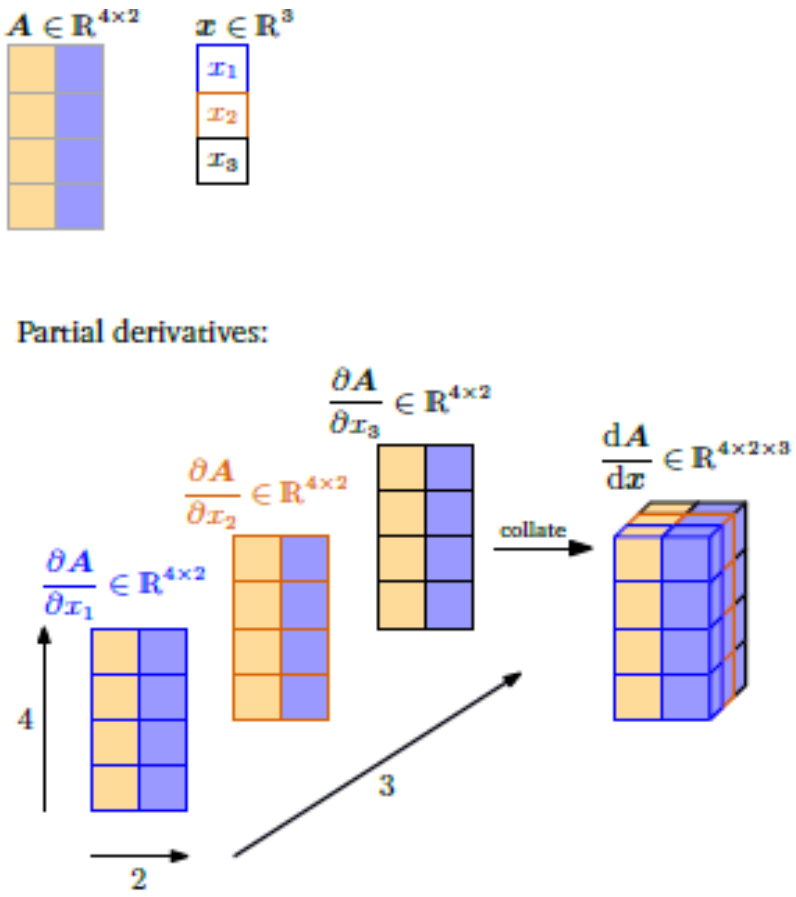
Gradients of Matrices

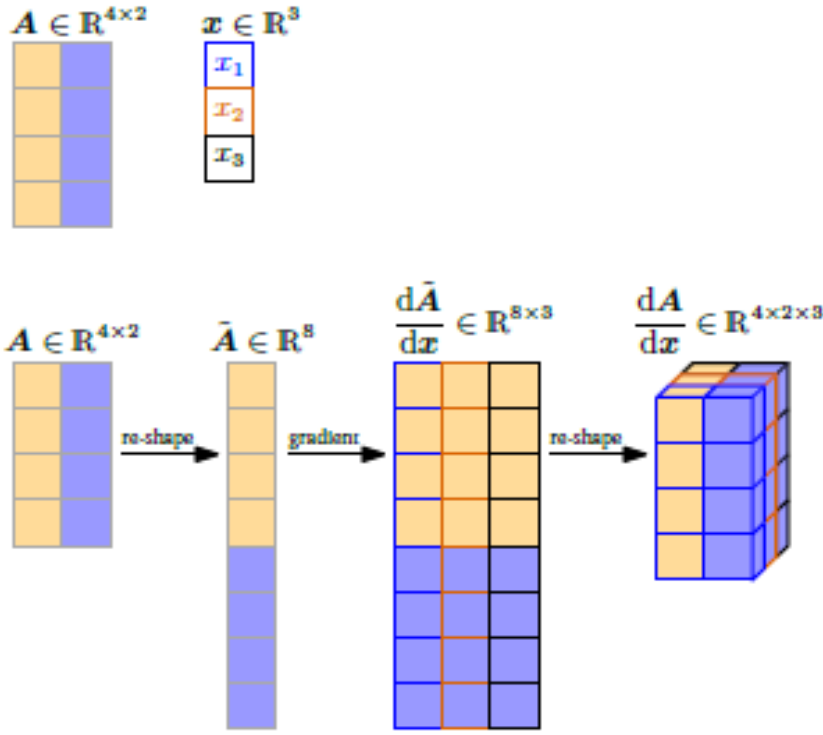
The gradient of an $m \times n$ matrix A with respect to a $p \times q$ matrix B, the resulting Jacobian would be an $(m \times n) \times (p \times q)$, i.e., a four-dimensional tensor J, whose entries are given as

$$J_{ijkl} = \frac{\partial A_{ij}}{\partial B_{kl}}$$

Since, we can consider $\mathbb{R}^{m \times n}$ as $\mathbb{R}^{mn}$, we can shape our matrix into vectors of length mn and pq respectively. The gradient using mn vectors results in a Jacobian of size $mn \times pq$

Gradients of Matrices





Gradients of Matrices

Let $f = Ax$ where $A \in \mathbb{R}^{m \times n}$, and $x \in \mathbb{R}^n$, then

$$\frac{\partial f}{\partial A} \in \mathbb{R}^{m \times (m \times n)}$$

By definition

$$\frac{\partial f}{\partial A} = \begin{bmatrix} \frac{\partial f_1}{\partial A} \\ \vdots \\ \frac{\partial f_m}{\partial A} \end{bmatrix}, \frac{\partial f_i}{\partial A} \in \mathbb{R}^{1 \times (m \times n)}$$

Gradients of Matrices

Now, we have

$$f_i = \sum_{j=1}^{n} A_{ij} x_j, i = 1, \cdots, m.$$

Therefore, by taking partial derivatives with respect to A_{iq}

$$\frac{\partial f_i}{\partial A_{iq}} = x_q.$$

Hence, i^{th} row becomes

$$\frac{\partial f_i}{\partial A_{i,:}} = x^T \in \mathbb{R}^{1 \times 1 \times n}$$

$$\frac{\partial f_i}{\partial A_{k,:}} = 0^T \in \mathbb{R}^{1\times 1\times n}, \text{for} k \neq i$$

Hence, by stacking the partial derivatives, we get

$$\frac{\partial f_i}{\partial A_{k,:}} = \begin{bmatrix} 0^T \\ \vdots \\ x^T \\ \vdots \\ 0^T \end{bmatrix} \in \mathbb{R}^{1\times m\times n}$$

Gradients of Matrices with respect to Matrices

Let $B \in \mathbb{R}^{m \times n}$ and $f : \mathbb{R}^{m \times n} \to \mathbb{R}^{n \times n}$ with

$$f(B) = B^T B =: K \in \mathbb{R}^{n \times n}$$

Then, we have

$$\frac{\partial K}{\partial B} \in \mathbb{R}^{(n \times n) \times (m \times n)}.$$

Moreover

$$\frac{\partial K_{pq}}{\partial B} \in \mathbb{R}^{1 \times (m \times n)}, \text{for} p, q = 1, \cdots n$$

where K_{pq} is the $(p, q)^{th}$ entry of $K = f(B)$

Gradients of Matrices with respect to Matrices

Let i^{th} column of B be b_i, then

$$K_{pq} = r_p^T r_q = \sum_{l=1}^{m} B_{lp} B_{lq}$$

Computing the partial derivative, we get

$$\frac{\partial K_{pq}}{\partial B_{ij}} = \sum_{l=1}^{m} \frac{\partial}{\partial B_{ij}} B_{lp} B_{lq} = \partial_{pqij}$$

Gradients of Matrices with respect to Matrices

Clearly, we have

$$\begin{aligned}
\partial_{pqij} &= B_{iq} \quad \text{if } j = p, p \neq q \\
\partial_{pqij} &= B_{ip} \quad \text{if } j = q, p \neq q \\
\partial_{pqij} &= 2B_{iq} \quad \text{if } j = p, p = q \\
\partial_{pqij} &= 0 \quad \text{otherwise}
\end{aligned}$$

where $p, q, j = 1, \cdots, n \; i = 1, \cdots, m$

Useful Identities for Computing Gradients

- $\frac{\partial}{\partial X} f(X)^T = (\frac{\partial f(X)}{\partial X})^T$
- $\frac{\partial}{\partial X} tr(f(X)) = tr(\frac{\partial f(X)}{\partial X})$
- $\frac{\partial}{\partial X} det(f(X)) = det(f(X)) tr(f(X)^{-1} \frac{\partial f(X)}{\partial X})$
- $\frac{\partial}{\partial X} f(X)^{-1} = -f(X)^{-1} \frac{\partial f(X)}{\partial X} f(X)^{-1}$

Useful Identities for Computing Gradients

- $\frac{\partial a^T X^{-1} b}{\partial X} = -(X^{-1})^T a b^T (X^{-1})^T$
- $\frac{\partial x^T a}{\partial x} = a^T$
- $\frac{\partial a^T x}{\partial x} = a^T$
- $\frac{\partial a^T X b}{\partial X} = a b^T$
- $\frac{\partial x^T B}{\partial x} = x^T (B + B^T)$
- $\frac{\partial}{\partial s}(x - As)^T W (x - As) = -2(x - As)^T WA$

for symmetric W.

Backpropagation and Automatic Differentiation

Consider the function

$$f(x) = \sqrt{x^2 + exp(x^2)} + cos(x^2 + exp(x^2))$$

Taking dervatives

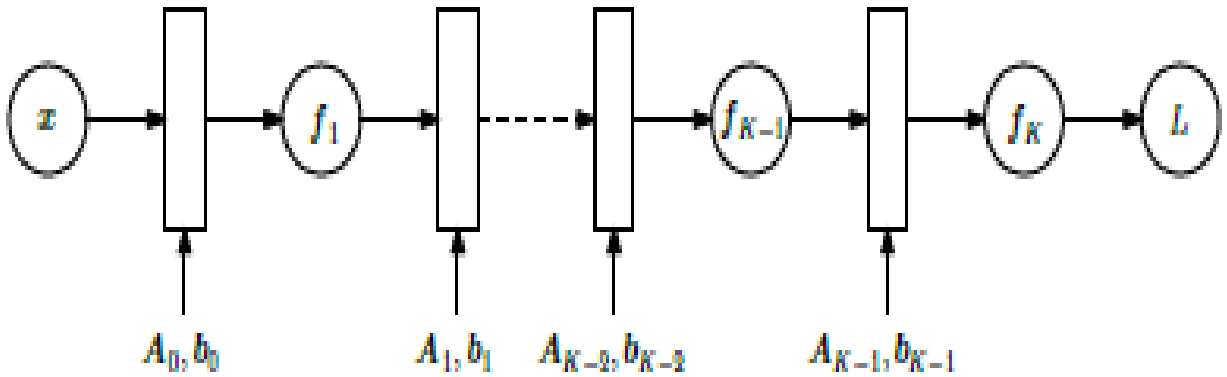
$$\begin{aligned} \frac{df}{dx} &= \frac{2x + 2xexp(x^2)}{2\sqrt{x^2 + exp(x^2)}} - sin(x^2 + exp(x^2))(2x + 2xexp(x^2)) \\ &= 2x(\frac{1}{2\sqrt{x^2 + exp(x^2)}} - sin(x^2 + exp(x^2)))(1 + exp(x^2)) \end{aligned}$$

Motivation

- The implementation of the gradient could be significantly more expensive than computing the function, which imposes unnecessary overhead where we get such lengthy expressions.
- We need an efficient way to compute the gradient of an error function with respect to the parameters of the model.
- For training deep neural network models, the backpropagation algorithm is one such method.

Backpropagation and Automatic Differentiation

In neural networks with multiple layers



$$f_i(x_{i-1}) = \sigma(A_{i-1}x_{i-1} + b_{i-1})$$

where x_{i-1} is the output of layer $i-1$ and σ is an activation function.

Backpropagation

To train these model, the gradient of the loss function L with respect to all model parameters $\theta_j = \{A_j, b_j\}, j = 1, \cdots, K$ and inputs of each layer needs to be computed. Consider,

$$\begin{aligned} f_0 &:= x \\ f_i &:= \sigma_i(A_{i-1} f_{i-1} + b_{i-1}), i = 1, \cdots, K. \end{aligned}$$

We have to find $\theta_j = \{A_j, b_j\}, j = 1, \cdots, K-1$ such that

$$L(\theta) = ||y - f_K(\theta, x)||^2$$

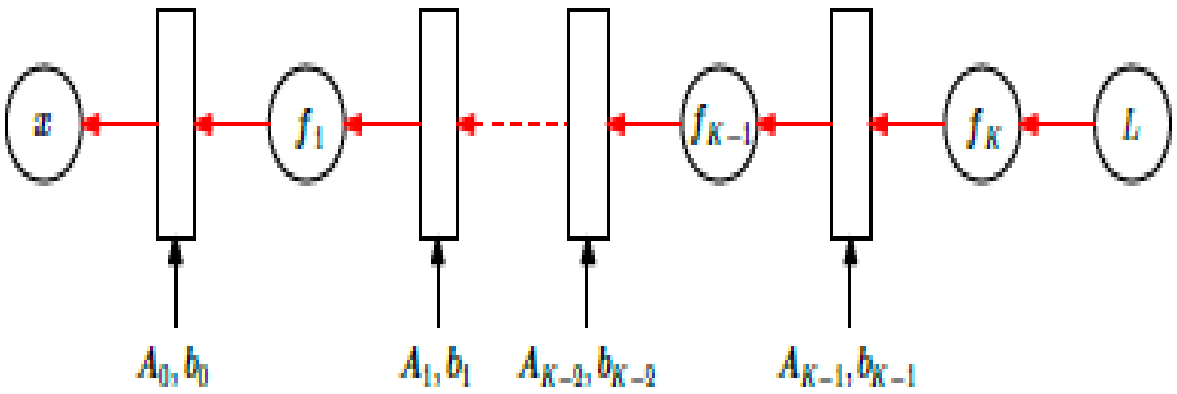
is minimum where $\theta = \{A_0, b_0, \cdots, A_{K-1}, b_{K-1}\}$

Backpropagation

Using the chain rule, we get

$$
\begin{aligned}
\frac{\partial L}{\partial \theta_{K-1}} &= \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial \theta_{K-1}} \\
\frac{\partial L}{\partial \theta_{K-2}} &= \frac{\partial L}{\partial f_K}\frac{f_K}{f_{K-1}}\frac{\partial f_{K-1}}{\partial \theta_{K-2}} \\
\frac{\partial L}{\partial \theta_{K-3}} &= \frac{\partial L}{\partial f_K}\frac{f_K}{f_{K-1}}\frac{\partial f_{K-1}}{\partial f_{K-2}}\frac{\partial f_{K-2}}{\partial \theta_{K-3}} \\
\frac{\partial L}{\partial \theta_i} &= \frac{\partial L}{\partial f_K}\frac{f_K}{f_{K-1}}\cdots\frac{\partial f_{i+2}}{\partial f_{i+1}}\frac{\partial f_{i+1}}{\partial \theta_i}
\end{aligned}
$$

Backpropagation



If the partial derivatives $\frac{\partial L}{\partial \theta_{i+1}}$ are computed, then the computation can be reused to compute $\frac{\partial L}{\partial \theta_i}$.

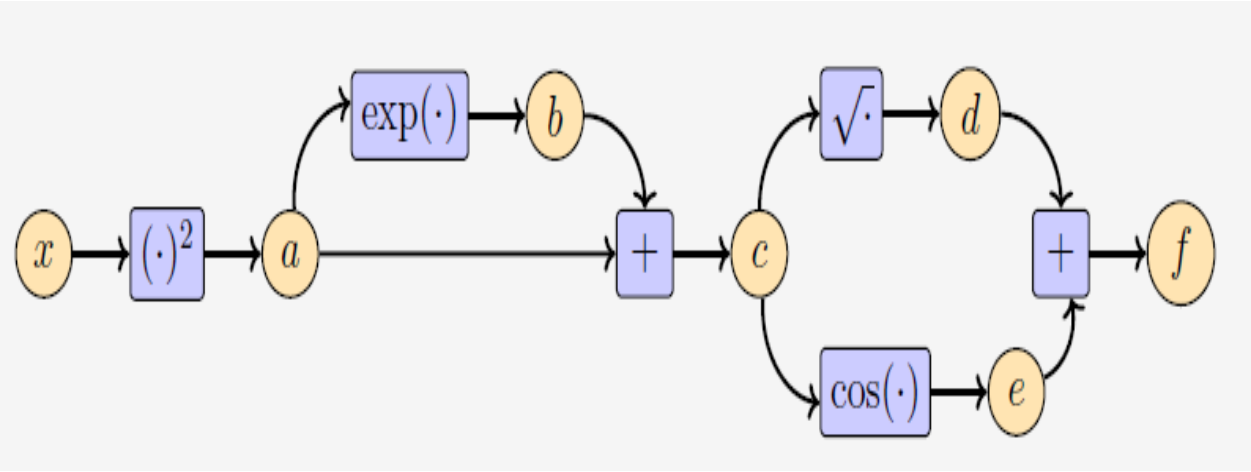
Example

Consider the function

$$f(x) = \sqrt{x^2 + exp(x^2)} + cos(x^2 + exp(x^2))$$

Let

$a = x^2, b = exp(a), c = a + b, d = \sqrt{c}, e = cos(c) \Rightarrow f = d + e$



Example

$$
\begin{aligned}
\Rightarrow \frac{\partial a}{\partial x} &= 2x \\
\frac{\partial b}{\partial a} &= exp(a) \\
\frac{\partial c}{\partial a} &= 1 = \frac{\partial c}{\partial b} \\
\frac{\partial d}{\partial c} &= \frac{1}{2\sqrt{c}} \\
\frac{\partial e}{\partial c} &= -sin(c) \\
\frac{\partial f}{\partial d} &= 1 = \frac{\partial f}{\partial e}
\end{aligned}
$$

Example

Thus, we have

$$
\begin{aligned}
\frac{\partial f}{\partial c} &= \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial c} \\
\frac{\partial f}{\partial b} &= \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} \\
\frac{\partial f}{\partial a} &= \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a} \\
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial a}\frac{\partial a}{\partial x}
\end{aligned}
$$

Example

Substituting the results, we get

$$
\begin{aligned}
\frac{\partial f}{\partial c} &= 1.(\frac{1}{2\sqrt{c}} + 1).(-sin(c)) \\
\frac{\partial f}{\partial b} &= \frac{\partial f}{\partial c}.1 \\
\frac{\partial f}{\partial a} &= \frac{\partial f}{\partial b} exp(a) + \frac{\partial f}{\partial c}.1 \\
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial a} 2x
\end{aligned}
$$

Thus, the computation for calculating the derivative is of similar complexity as the computation of the function itself.

Formalization of Automatic Differentiation

Let $x_1, \cdots, x_d$: input variables.
$x_{d+1}, \cdots, x_{D-1}$: intermediate variables.
x_D : output variable, then we have,

$$x_i = g_i(x_{Pa(x_i)})$$

Note that g_is are elementary functions and are also called as forward propagation function and $x_{Pa(x_i)}$ is the set of parent nodes of variable x_i in the graph.

Formalization of Automatic Differentiation

Now,

$$f = x_D \Rightarrow \frac{\partial f}{\partial_D} = 1$$

For other variables, using chain rule, we get

$$\frac{\partial f}{\partial x_i} = \sum_{x_j : x_i \in Pa(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial x_j}{\partial x_i} = \sum_{x_j : x_i \in Pa(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial g_i}{\partial x_i}$$

The last equation is the back propagation of the gradient through the computation graph. For neural network training, we back propagate the error of the prediction with respect to the label.

Lecture 8

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction

- Till now we have discussed about Taylor/Maclaurian series, Partial Derivatives and Gradients.
- Now we are interested in Higher order Derivatives.
- Multivariate Taylor Series and its uses in the expansion of a function with multivariables.

Higher-Order Derivatives

Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$

Notations for Higher-Order Partial Derivatives:

$\frac{\partial^2 f}{\partial x^2}$: Second Partial Derivative of x w.r.t. x

$\frac{\partial^n f}{\partial x^n}$: n^{th} Partial Derivative of x w.r.t. x

$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(\frac{\partial f}{\partial x})$: is the partial derivative obtained by first partial differentiating with respect to x and then with respect to y.

$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(\frac{\partial f}{\partial y})$: is the partial derivative obtained by first partial differentiating by y and then x.

Hessian Matrix

The Hessian is the collection of all second-order partial derivatives. If f(x, y) is a twice (continuously) differentiable function, then $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$ i.e., the order of differentiation does not matter, and the corresponding Hessian matrix

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

is symmetric. The Hessian is denoted as $\nabla^2_{x,y} f(x, y)$

Linearization and Multivariate Taylor Series

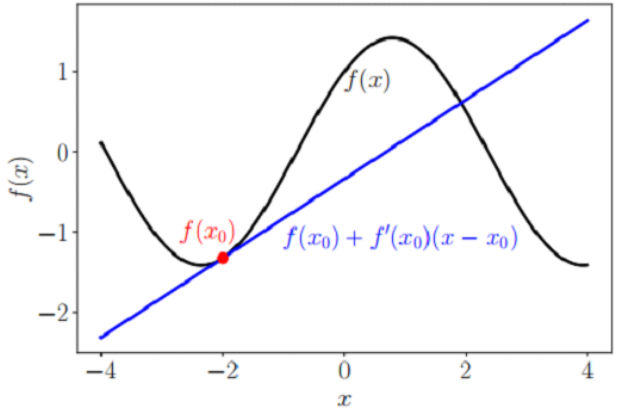
The gradient ∇f of a function f is often used for a locally linear approximation of f around x_0 :

$$f(x) \approx f(x_0) + (\nabla_x f)(x_0)(x - x_0) \tag{1}$$

Here $(\nabla_x f)(x_0)$ is the gradient of f with respect to x, evaluated atx_0. Figure illustrates the linear approximation of a function f at an input x_0. The original function is approximated by a straight line.

Linearization and Multivariate Taylor Series...

This approximation is locally accurate, but the farther we move away from x_0 the worse the approximation gets. Equation (1) is a special case of a multivariate Taylor series expansion of f at x_0, where we consider only the first two terms. We discuss the more general case in the following, which will allow for better approximations.



Linear approximation of a function. The original function f is linearized at x_0 = - 2 using a first-order Taylor series expansion.

Multivariate Taylor Series

Consider a function $f : \mathbb{R}^D \to \mathbb{R}$, $x \to f(x)$,

$$x \in \mathbb{R}^D$$

that is smooth at x_0. When we define the difference vector $\delta := x - x_0$ the multivariate Taylor series of f at (x_0) is defined as multivariate Taylor series

$$f(x) = \sum_{k=0}^{\infty} \frac{D_x^k f(x_0)}{k!} \delta^k \tag{2}$$

where $D_x^k f(x_0)$ is the kth (total) derivative of f with respect to x, evaluated at x_0.

Taylor Polynomial

The Taylor polynomial of degree n of Taylor polynomial f at x_0 contains the first n + 1 components of the series in (2) and is defined as

$$T_n(x) = \sum_{k=0}^{n} \frac{D_x^k f(x_0)}{k!} \delta^k \tag{3}$$

In (2) and (3), we used the slightly sloppy notation of δ^k, which is not defined for vectors

$$x \in \mathbb{R}^D,$$

$D > 1$, and $k > 1$. Note that both $D_x^k f$ and δ^k are k^{th} order tensors, i.e., k-dimensional arrays.

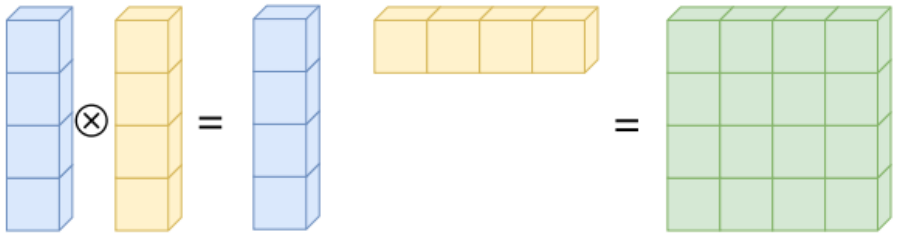
Taylor Polynomial...

kth-order tensor $\boldsymbol{\delta}^k \in \mathbb{R}^{\overbrace{D \times D \times \ldots \times D}^{k \text{ times}}}$ is obtained as a k-fold outer product, denoted by $\otimes$, of the vector $\boldsymbol{\delta} \in \mathbb{R}^D$. For example,

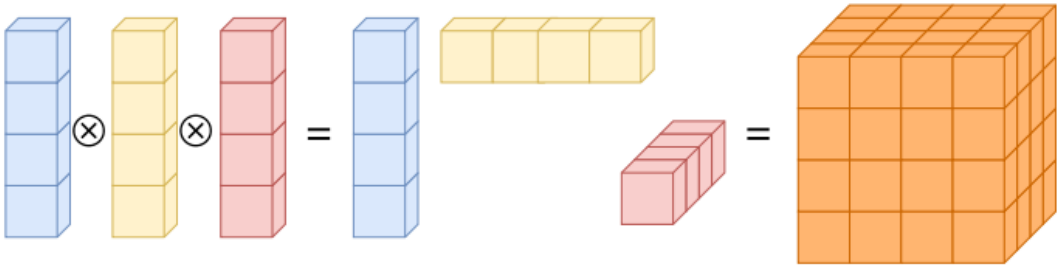
$$\boldsymbol{\delta}^2 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\delta}^\top, \quad \boldsymbol{\delta}^2[i,j] = \delta[i]\delta[j]$$

$$\boldsymbol{\delta}^3 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\delta}, \quad \boldsymbol{\delta}^3[i,j,k] = \delta[i]\delta[j]\delta[k].$$

Taylor Polynomial...



(a) Given a vector $\boldsymbol{\delta} \in \mathbb{R}^4$, we obtain the outer product $\boldsymbol{\delta}^2 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\delta}^\top \in \mathbb{R}^{4\times 4}$ as a matrix.



(b) An outer product $\boldsymbol{\delta}^3 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\delta} \in \mathbb{R}^{4\times 4\times 4}$ results in a third-order tensor ("three-dimensional matrix"), i.e., an array with three indexes.

$$D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)\boldsymbol{\delta}^k = \sum_{i_1=1}^{D} \cdots \sum_{i_k=1}^{D} D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)[i_1, \ldots, i_k]\delta[i_1] \cdots \delta[i_k]$$

Taylor Polynomial...

In general, we obtain the terms in the Taylor series, where $D_x^k f(x_0)\delta^k$ contains k^{th} order polynomials.Now that we defined the Taylor series for vector fields, let us explicitly write down the first terms $D_x^k f(x_0)\delta^k$ of the Taylor series expansion for

$$k = 0, \ldots, 3 \text{ and } \boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}_0:$$

$$k = 0 : D_{\boldsymbol{x}}^0 f(\boldsymbol{x}_0)\boldsymbol{\delta}^0 = f(\boldsymbol{x}_0) \in \mathbb{R}$$

$$k = 1 : D_{\boldsymbol{x}}^1 f(\boldsymbol{x}_0)\boldsymbol{\delta}^1 = \underbrace{\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)}_{1\times D}\underbrace{\boldsymbol{\delta}}_{D\times 1} = \sum_{i=1}^{D} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)[i]\delta[i] \in \mathbb{R}$$

$$k = 2 : D_{\boldsymbol{x}}^2 f(\boldsymbol{x}_0)\boldsymbol{\delta}^2 = \operatorname{tr}\big(\underbrace{\boldsymbol{H}(\boldsymbol{x}_0)}_{D\times D}\underbrace{\boldsymbol{\delta}}_{D\times 1}\underbrace{\boldsymbol{\delta}^\top}_{1\times D}\big) = \boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{\delta}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} H[i,j]\delta[i]\delta[j] \in \mathbb{R}$$

$$k = 3 : D_{\boldsymbol{x}}^3 f(\boldsymbol{x}_0)\boldsymbol{\delta}^3 = \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} D_x^3 f(\boldsymbol{x}_0)[i,j,k]\delta[i]\delta[j]\delta[k] \in \mathbb{R}$$

Here, $\boldsymbol{H}(\boldsymbol{x}_0)$ is the Hessian of f evaluated at $\boldsymbol{x}_0$.

Taylor Series Expansion of a Function with Two Variables

Consider the function $f(x, y) = x^2 + 2xy + y^3$.
We want to compute the Taylor series expansion of f at $(x_0, y_0) = (1, 2)$.
Before we start, let us discuss what to expect: The function in $f(x, y)$ is a polynomial of degree 3. We are looking for a Taylor series expansion,which itself is a linear combination of polynomials. Therefore, we do not expect the Taylor series expansion to contain terms of fourth or higher order to express a third-order polynomial. This means that it should be sufficient to determine the first four terms of $f(x) = \sum_{k=0}^{\infty} \frac{D_x^k f(x_0)}{k!} \delta^k$ for an exact alternative representation of $f(x, y)$. To determine the Taylor series expansion, we start with the constant term and the first-order derivatives, which are given by $f(1, 2) = 13$

Taylor Series Expansion of a Function with Two Variables...

$$\frac{\partial f}{\partial x} = 2x + 2y \implies \frac{\partial f}{\partial x}(1,2) = 6$$
$$\frac{\partial f}{\partial y} = 2x + 3y^2 \implies \frac{\partial f}{\partial y}(1,2) = 14\,.$$

Therefore, we obtain

$$D^1_{x,y} f(1,2) = \nabla_{x,y} f(1,2) = \begin{bmatrix} \frac{\partial f}{\partial x}(1,2) & \frac{\partial f}{\partial y}(1,2) \end{bmatrix} = \begin{bmatrix} 6 & 14 \end{bmatrix} \in \mathbb{R}^{1\times 2}$$

such that

$$\frac{D^1_{x,y} f(1,2)}{1!}\boldsymbol{\delta} = \begin{bmatrix} 6 & 14 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix} = 6(x-1) + 14(y-2)\,.$$

Therefore, we obtain

$$D^1_{x,y} f(1,2) = \nabla_{x,y} f(1,2) = \begin{bmatrix} \frac{\partial f}{\partial x}(1,2) & \frac{\partial f}{\partial y}(1,2) \end{bmatrix} = \begin{bmatrix} 6 & 14 \end{bmatrix} \in \mathbb{R}^{1\times 2}$$

Taylor Series Expansion of a Function with Two Variables...

When we collect the second-order partial derivatives, we obtain the Hessian

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 6y \end{bmatrix},$$

such that

$$\boldsymbol{H}(1,2) = \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Therefore, the next term of the Taylor-series expansion is given

$$\begin{aligned} \frac{D^2_{x,y} f(1,2)}{2!}\boldsymbol{\delta}^2 &= \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{H}(1,2)\boldsymbol{\delta} \\ &= \frac{1}{2}\begin{bmatrix} x-1 & y-2 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix} \\ &= (x-1)^2 + 2(x-1)(y-2) + 6(y-2)^2. \end{aligned}$$

Taylor Series Expansion of a Function with Two Variables...

The third-order derivatives are obtained as

$$D^3_{x,y} f = \begin{bmatrix} \frac{\partial \boldsymbol{H}}{\partial x} & \frac{\partial \boldsymbol{H}}{\partial y} \end{bmatrix} \in \mathbb{R}^{2\times 2\times 2},$$

$$D^3_{x,y} f[:,:,1] = \frac{\partial \boldsymbol{H}}{\partial x} = \begin{bmatrix} \frac{\partial^3 f}{\partial x^3} & \frac{\partial^3 f}{\partial x^2 \partial y} \\ \frac{\partial^3 f}{\partial x \partial y \partial x} & \frac{\partial^3 f}{\partial x \partial y^2} \end{bmatrix},$$

$$D^3_{x,y} f[:,:,2] = \frac{\partial \boldsymbol{H}}{\partial y} = \begin{bmatrix} \frac{\partial^3 f}{\partial y \partial x^2} & \frac{\partial^3 f}{\partial y \partial x \partial y} \\ \frac{\partial^3 f}{\partial y^2 \partial x} & \frac{\partial^3 f}{\partial y^3} \end{bmatrix}.$$

Taylor Series Expansion of a Function with Two Variables...

Since most second-order partial derivatives in the Hessian, are constant, the only nonzero third-order partial derivative is $\frac{\partial^3 f}{\partial y^3} = 6 \implies \frac{\partial^3 f}{\partial y^3}(1,2) = 6$ Higher-order derivatives and the mixed derivatives of degree 3 (e.g., $\frac{\partial^3 f}{\partial x^2 \partial y}$) vanish, such that

$$D^3_{x,y} f[:,:,1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad D^3_{x,y} f[:,:,2] = \begin{bmatrix} 0 & 0 \\ 0 & 6 \end{bmatrix}$$

and

$$\frac{D^3_{x,y} f(1,2)}{3!} \boldsymbol{\delta}^3 = (y-2)^3,$$

Taylor Series Expansion of a Function with Two Variables...

which collects all cubic terms of the Taylor series. Overall, the (exact) Taylor series expansion of f at $(x_0, y_0) = (1, 2)$ is

$$
\begin{aligned}
f(x) &= f(1,2) + D^1_{x,y} f(1,2)\boldsymbol{\delta} + \frac{D^2_{x,y} f(1,2)}{2!}\boldsymbol{\delta}^2 + \frac{D^3_{x,y} f(1,2)}{3!}\boldsymbol{\delta}^3 \\
&= f(1,2) + \frac{\partial f(1,2)}{\partial x}(x-1) + \frac{\partial f(1,2)}{\partial y}(y-2) \\
&\quad + \frac{1}{2!}\left(\frac{\partial^2 f(1,2)}{\partial x^2}(x-1)^2 + \frac{\partial^2 f(1,2)}{\partial y^2}(y-2)^2 \right. \\
&\quad \left. + 2\frac{\partial^2 f(1,2)}{\partial x \partial y}(x-1)(y-2)\right) + \frac{1}{6}\frac{\partial^3 f(1,2)}{\partial y^3}(y-2)^3 \\
&= 13 + 6(x-1) + 14(y-2) \\
&\quad + (x-1)^2 + 6(y-2)^2 + 2(x-1)(y-2) + (y-2)^3 .
\end{aligned}
$$

Mathematical Foundations for Machine Learning

BITS Pilani
Pilani Campus

MFML Team

innovate achieve lead

BITS Pilani
Pilani Campus

S2 -22 AIMLCZC416, MFML

Webinar#1

Agenda

Discussion on Solutions of

- Homework Problems
- Proof

Homework Problem

Q1 Let $A_{m\times n}$ be a given matrix with $m > n$. If the time taken to compute the determinant of a square matrix of size j is j^3, find upper bound on the

a) **total time taken to find the rank of A using determinants**

b) **number of additions and multiplications required to determine the rank using the elimination procedure.**

Homework Problem

a) total time taken to find the rank of A using determinants

Let $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{k1} & a_{k2} & \cdots & a_{kn} \\ & & \vdots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$

Since m> n the maximum possible rank of A is n.

Thus, we need to compute all determinates of order k, 1<k<n.

Therefore, Number of Determinants = $\sum_{i=1}^{i=\min(m,n)} {}^{m}C_i * {}^{n}C_i$,

Where, $min\ (m,n)$ is maximum size of square matrix in A

${}^{m}C_i$ is the combination of 'i' columns out of m

${}^{n}C_i$ is the combination of 'i' rows out of n

Given that time taken to compute the determinant of a square matrix of size i is i^3

Hence total time taken to compute determinant of a square matrix is $= \sum_{i=1}^{i=\min(m,n)} {}^{m}C_i \, {}^{n}C_i \, i^3$

Homework Problem

b) number of additions and multiplications required to determine the rank using the elimination procedure.

Solution:

Let $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{k1} & a_{k2} & \cdots & a_{kn} \\ & & \vdots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$

The number of multiplications required to make all the entries below a_{11} to zeros is $n(m-1)$

The number of multiplications required to make all the entries below a_{22} to zeros is $(n-1)(m-2)$

.

The number of multiplications required to make all the entries below a_{kk} to zeros is $(n-k+1)(m-k)$

Therefore, total number of multiplications required is $= n(m-1) + (n-1)(m-2) + \ldots$

$$= \sum_{K=1}^{n} (m-k)(n-k+1)$$

Similarly, total number of additions required is $= \sum_{K=1}^{n} (m-k)(n-k+1)$

Hence the total number additions and multiplications required $= 2\sum_{K=1}^{n} (m-k)(n-k+1)$

Homework Problem

Q2

Let A n×n be a given square matrix. Compute the number of multiplications and additions required to evaluate A^{28} using

a) the naive method, A^{28}= (A·A···A)28 times
b) A^2, A^4=A^2·A^2, etc.

Solution:

Let $$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}_{n\times n},$$

Now consider $A^2 = A * A$

$$= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}_{n\times n} * \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}_{n\times n}$$

Homework Problem

For multiplying the above matrices, we have

	Number of operations required for multiplication of R_i with all columns		Number of operations required for addition of R_i with all columns	
	Multiplications required	**Number of operations required for multiplication of**	**Additions required**	**Number of operations required for addition of**
$R_1C_1 = a_{11}a_{11} + a_{12}a_{21} + \ldots + a_{1n}a_{n1}$ $R_1C_2 = a_{11}a_{12} + a_{21}a_{22} + \ldots + a_{1n}a_{n2}$ $\vdots$ $R_1C_n = a_{11}a_{1n} + a_{12}a_{2n} + \ldots + a_{1n}a_{nn}$	n n $\vdots$ n	1st Row with all columns is n^2	$(n-1)$ $(n-1)$ $\vdots$ $(n-1)$	1st Row with all columns is $n(n-1)$
$R_2C_1 = a_{21}a_{11} + a_{22}a_{21} + \ldots + a_{2n}a_{n1}$ $R_2C_2 = a_{21}a_{12} + a_{22}a_{22} + \ldots + a_{2n}a_{n2}$ $\vdots$ $R_2C_n = a_{21}a_{1n} + a_{22}a_{2n} + \ldots + a_{2n}a_{nn}$	n n $\vdots$ n	2nd Row with all columns is n^2	$(n-1)$ $(n-1)$ $\vdots$ $(n-1)$	2nd Row with all columns is $n(n-1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$R_nC_1 = a_{n1}a_{11} + a_{n2}a_{21} + \ldots + a_{nn}a_{n1}$ $R_nC_2 = a_{n1}a_{12} + a_{n2}a_{22} + \ldots + a_{nn}a_{n2}$ $\vdots$ $R_nC_n = a_{n1}a_{1n} + a_{n2}a_{2n} + \ldots + a_{nn}a_{nn}$	n n $\vdots$ n	nth Row with all columns is n^2	$(n-1)$ $(n-1)$ $\vdots$ $(n-1)$	nth Row with all columns is $n(n-1)$
Total	$n\,n^2 = n^3$		$n\,n\,(n-1) = n^2(n-1)$	

Thus, the number of multiplications required to multiply the matrix A^2 is $= n^3$

Thus, the number of additions required to multiply the matrix A^2 is $= n^2(n-1)$

Homework Problem

a) For multiplication of the form the naive method, A^{28}= (A·A···A) 28 times is as fallows

The required number of additions and multiplications to get A^2 is $= n^3 + n^2(n-1)$
The required number of additions and multiplications to get A^3 is $= 2\left[n^3 + n^2(n-1)\right]$
And so on…

Similarly, the required number of additions and multiplications to get A^{28} is $= 27\left[n^3 + n^2(n-1)\right]$

b) For multiplication of the form A^2, A^4=$A^2 \cdot A^2$, etc.

We know that
The number of additions and multiplications to get A^2 is $= n^3 + n^2(n-1)$
The required number of additions and multiplications to get $A^4 = A^2 * A^2$ is $= n^3 + n^2(n-1)$
The required number of additions and multiplications to get $A^8 = A^4 * A^4$ is $= n^3 + n^2(n-1)$
The required number of additions and multiplications to get $A^{16} = A^8 * A^8$ is $= n^3 + n^2(n-1)$
The required number of additions and multiplications to get $A^{28} = A^{16} * A^8 * A^4$ is $= 2\left[n^3 + n^2(n-1)\right]$

Thus, the total no of operations required to get A^{28} is $= 6\left[n^3 + n^2(n-1)\right]$.

Homework Problem

Q3: Let A_{nxk} and B_{kxk} be two given matrices with rank(B) =k. Estimate the rank of AB in terms of the ranks of A and B. Note that the rank of a matrix is the dimension of the vector space spanned by its rows/ columns.

Consider the spaces

C_1 ={A*x* : *x* is a vector of length k}

C_2 ={AB*y* : *y* is a vector of length k}

C_1 and C_2 are vector spaces. Let y be any vector of length k and x = By. Since AB*y* = A*x* belongs to C_1

C_2 is a subspace of C_1

dim(C2) ≤ dim(C1)

rank (AB) ≤ rank(A)

Homework Problem

Applying to the transpose

$\text{rank}(AB)^t = \text{rank}(B^tA^t) \leq \text{rank}(B^t) = \text{rank}(B)$

Rank $(AB) \leq \text{rank}(B)$

Thus

$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$

Homework Problem

Q4: Compute the total number of divisions, multiplications and additions required to perform the forward elimination and back substitution in solving a system of linear equations $A_{nxn}\, x = b$ using the Gauss elimination method.

Homework Problem

Number of divisions required in Gauss elimination $= \sum_{k=1}^{n-1} (n-k)$

Number of multiplications and subtractions in Gauss elimination $= 2\sum_{k=1}^{n-1} (n-k)(n-k+1)$

Forward elimination

Number of operations required back substitution $= 2\sum_{k=1}^{n} (n-k) + n$

Back substitution.

Now Let $f(n)$ be forward elimination $= \sum_{k=1}^{k=n-1} (n-k) + 2\sum_{k=1}^{k=(n-1)} (n-k)(n-k+1)$

$\sum_{k=1}^{n} k = 1+2+3+\cdots n = \frac{n(n+1)}{2}$

$\sum_{k=1}^{n} k^2 = \frac{k(k+1)(2k+1)}{6}$

Now $n-k = s$, $n-s = k$

$k=1$	$k=n-1$
$n-s=1$	$n-s=n-1$
$s=n-1$	$s=1$

Homework Problem

$$= \sum_{s=1}^{s=n-1} s + 2\sum_{s=1}^{s=n-1} s(s+1)$$

$$= \frac{(n-1)(n)}{2} + 2\left[\frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2}\right]$$

$$= \frac{n^2-n}{2} + \cancel{2}\left[\frac{(n^2-n)(2n-1) + 3n^2-3n}{\cancel{6}\,3}\right]$$

$$= \frac{3n^2-3n + 2(2n^3-2n^2-n^2+n+3n^2-3n)}{6}$$

$$= \frac{3n^2-3n + 2(2n^3-\cancel{3n^2}+n+\cancel{3n^2}-3n)}{6}$$

$$= \frac{3n^2-3n+4n^3+2n-6n}{6}$$

$$= \frac{4n^3}{6} + \frac{3n^2}{6} - \frac{7n}{6}$$

$$\boxed{f(n) = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{7}{6}n} \Rightarrow \boxed{O(n^3) = f(n)}$$

Homework Problem

Now

$$b(n) = 2\sum_{k=1}^{n}(n-k) + n$$

$$= 2\left(n \times \sum_{k=1}^{n} 1 - \sum_{k=1}^{n} k\right) + n$$

$$= 2\left(n \cdot n - \frac{n(n+1)}{2}\right) + n$$

$$= 2\left(n^2 - \frac{n^2}{2} - \frac{n}{2}\right) + n$$

$$= 2\left(\frac{n^2}{2} - \frac{n}{2}\right) + n$$

$$= n^2 - \cancel{n} + \cancel{n}$$

$$\boxed{b(n) = n^2}$$

$$\boxed{O(n^2) = b(n)}$$

Homework Problem

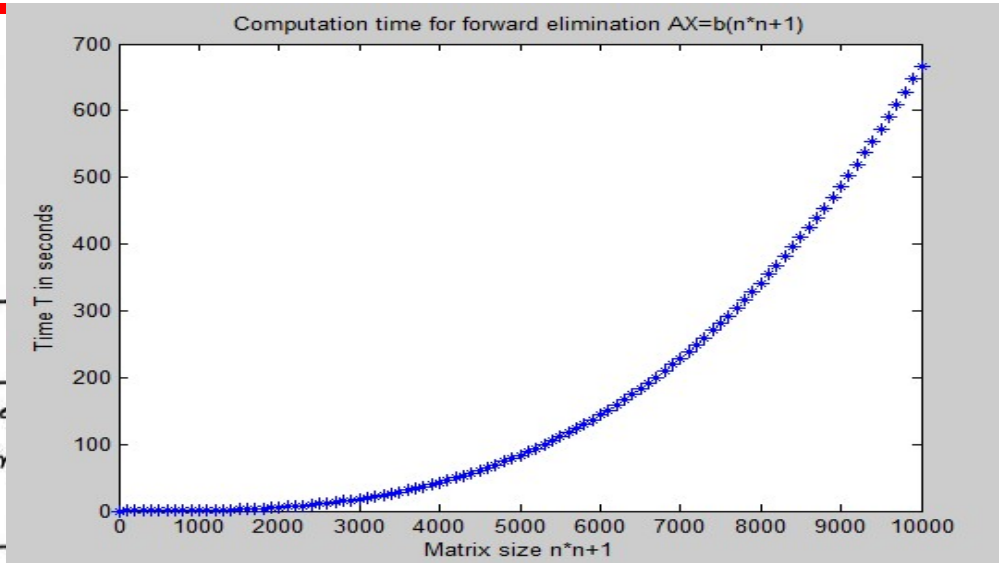
For solving $Ax = B$, $[A:B]_{(n \times n+1)}$

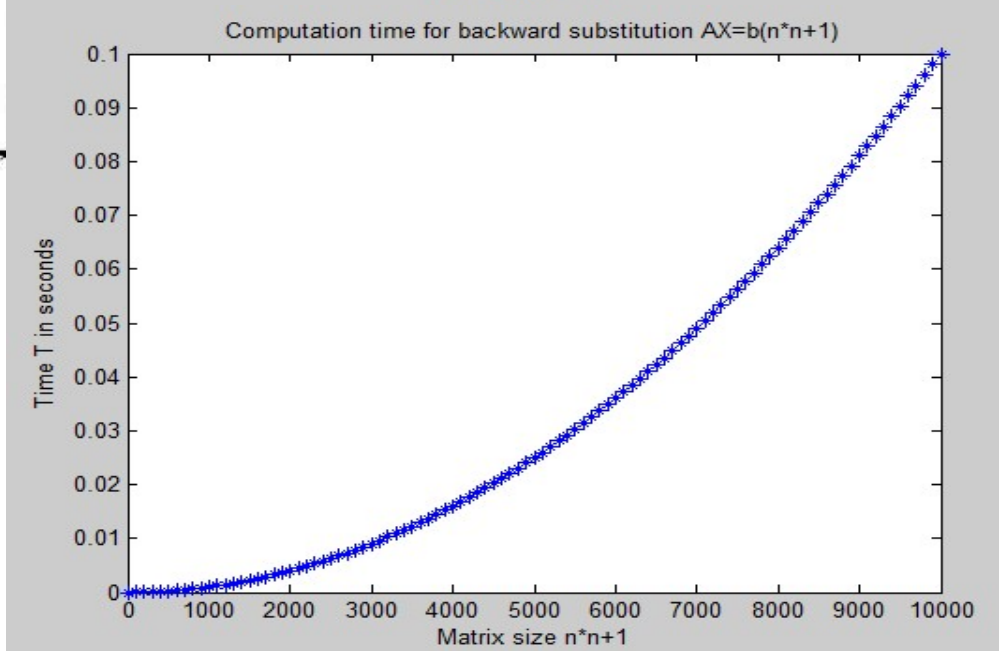
The number of operations required is

If for 1 count $- 10^{-9}$ sec

Algorithm	$n = 1000$	$n = 10{,}000$
Elimination $f(n) = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{7}{6}n$	$0.6671655 \simeq 0.7$ Sec	666.7166 Sec $\simeq 11.1119$ min
Back Substitution $b(n) = n^2$	0.001 Sec	0.1 Sec

```
N=0:100:10000;
f=(2/3)*(N.^3)+(1/2)*(N.^2)-(7/6)*(N);
%f=(N.^2); for backward sub
T=f*10^(-9);
plot(N,T,'*')
```





Theorem: Let V be a finite dimensional vector space. Then any two bases have the same cardinality.

Proof: Suppose that $B = \{v_1, v_2, \ldots, v_m\}$ **and** $C = \{w_1, w_2, \ldots. w_n\}$ **Are two bases of V.**

We want to show that m=n.

We may assume that m, n > 0.

Consider the set $B \cup \{w_n\}$ **.** $w_n \in V$ **Belongs to the span of** $v_1, v_2, \ldots, v_m$.

Therefore, we may find $r_1, r_2, \ldots, r_m$ **such that**

$$w_n = r_1 v_1 + r_2 v_2 + \ldots + r_m v_m.$$

Theorem: Let V be a finite dimensional vector space. Then any two bases have the same cardinality.

$w_n \neq 0$ and so at least one of $r_1, r_2, \ldots, r_m$ is non-zero.
We may assume that $r_m \neq 0$.
In this case

$$v_m = w_n - \left(r_1 v_1 + r_2 v_2 + \ldots + r_{m-1} v_{m-1}\right) / r_m$$

and so v_m is a linear combination of $v_1, v_2, \ldots, v_m$ and w_n.
We know that If v_n is a lnear combination of $v_1, v_2, \ldots, v_{n-1}$ then span$\{v_1, v_2, \ldots, v_{n-1}\}$=span$\{v_1, v_2, \ldots, v_n\}$
Therefore, $B - \{v_m\} \cup \{w_n\}$ also spans V.
So, now consider $v_1, v_2, \ldots, v_{m-1}, w_{n-1}$ and w_n.
As $v_1, v_2, \ldots, v_{m-1}$ and w_n span V it follows that w_{n-1} is a

Theorem: Let V be a finite dimensional vector space. Then any two bases have the same cardinality.

linear combination of $v_1, v_2, \ldots, v_{m-1}$ and w_n.

$$w_{n-1} = r_1 v_1 + r_2 v_2 + \ldots + r_{m-1} v_{m-1} + s_n w_n$$

Suppose that every $r_i = 0$. Then w_{n-1} and w_n are dependent, which contradicts the fact that C is a basis. Thus $r_i \neq 0$ some i. Relabelling we may suppose that $r_{m-1} \neq 0$. As before this implies that v_{m-1} is a linear combination of $v_1, v_2, \ldots, v_{m-2}, w_{n-1}$ and w_n. But from the known results $v_1, v_2, \ldots, v_{m-1}, w_{n-1}$ and w_n span V.

We can repeat this process for every vector in C. It follows that $m \leq m$. By symmetry $n \leq m$. But then $m=n$.

Product of any two lower triangular matrices is a lower triangular matrix

$A=[a_n]$, $B=[b_n]$ be lower triangular matrices of order n and let C=AB

(i) The diagonal elements of C are given by

$$\forall j \in [1, 2, \ldots, n] : c_{jj} = a_{jj} b_{jj}$$

(ii) The matrix C is itself lower triangular matrix

Proof: From the definition of matrix product, we have

$$\forall i, j \in [1, 2, \ldots n],\ c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.\ When\ i = j,\ c_{jj} == \sum_{k=1}^{n} a_{jk} b_{kj}$$

Now both A and B are lower triangular matrices

Product of any two lower triangular matrices is a lower triangular matrix

Thus:

$$\left.\begin{array}{l} if\ k < j,\ b_{kj} = 0\ and\ a_{jk}b_{kj} = 0 \\ if\ k > j, a_{jk} = 0\ and\ a_{jk}b_{kj} = 0 \\ and\ a_{jk}b_{kj} \neq 0\ when\ j = k \end{array}\right\} \Rightarrow c_{jj} = \sum_{k=1}^{n} a_{jk}b_{kj} = a_{jj}b_{jj}$$

Now if $i < j$, it follows that either a_{ik} or b_{kj} is zero for all k, and thus $c_{ij} = 0$.

Therefore, C is also a lower triangular matrix.

Thank You

BITS Pilani
Pilani Campus

Mathematical Foundations for Machine Learning

MFML Team

innovate achieve lead

BITS Pilani

Pilani Campus

S2 -22_AIMLCZC416, MFML

Webinar 2

Agenda

Discussion on

- Solutions of Homework Problems
- Gram-Schmidt process with example
- Example of SVD

Homework Problem

Qus-1: Let $B = (b_1, b_2, \ldots, b_{r-1}, b_r, b_{r+1}, \ldots, b_n)$ be a non-singular matrix. If column b_r is replaced by a and that the resulting matrix is called B_a along with $a = \sum_{i=1}^{n} y_i b_i$, then state the necessary and sufficient conditions for B_a to be non-singular.

Sol: Given

$$
\begin{array}{cc}
 & \begin{array}{cccccccc} b_1 & b_2 & \ldots & b_{r-1} & b_r & b_{r+1} & \ldots & b_n \end{array} \\
B = & \begin{bmatrix} * & * & \ldots & * & * & * & \ldots & * \\ * & * & \ldots & * & * & * & \ldots & * \\ * & * & \ldots & * & * & * & \ldots & * \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \end{bmatrix}
\end{array}
$$

Linearly independent columns

is non-singular.

Homework Problem

$$\begin{array}{c} \begin{array}{cccccccc} b_1 & b_2 & \dots & b_{r-1} & b_r & b_{r+1} & \dots & b_n \end{array} \\ B = \begin{bmatrix} * & * & \dots & * & * & * & \dots & * \\ * & * & \dots & * & * & * & \dots & * \\ * & * & \dots & * & * & * & \dots & * \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \end{bmatrix} \end{array}$$

Linearly independent columns

This means no column of B can be written as a linear combination of other columns.

Replacing column b_r by a, we get

$$\begin{array}{c} \begin{array}{cccccccc} b_1 & b_2 & \dots & b_{r-1} & a & b_{r+1} & \dots & b_n \end{array} \\ B_a = \begin{bmatrix} * & * & \dots & * & * & * & \dots & * \\ * & * & \dots & * & * & * & \dots & * \\ * & * & \dots & * & * & * & \dots & * \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \end{bmatrix} \end{array}$$

Homework Problem

Looking at, $a = \sum_{i=1}^{n} y_i b_i$, we see that

$$\begin{array}{c} \begin{array}{cccccccc} b_1 & b_2 & \dots & b_{r-1} & a & b_{r+1} & \dots & b_n \end{array} \\ \begin{bmatrix} * & * & \dots & * & * & * & \dots & * \\ * & * & \dots & * & * & * & \dots & * \\ * & * & \dots & * & * & * & \dots & * \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \end{bmatrix} \\ B_a \end{array}$$

- a is a linear combination of $b_1, b_2, \dots, b_n$.
- Suppose $y_r = 0$, i.e.

$$a = y_1 b_1 + y_2 b_2 + \dots + y_{r-1} b_{r-1} + y_{r+1} b_{r+1} + \dots + y_n b_n \,.$$

This means a is a linear combination of other columns of B_a. Hence B_a is singular.

Homework Problem

- Suppose $y_r \neq 0$, then

$$a = y_r b_r + y_1 b_1 + y_2 b_2 + \ldots + y_{r-1} b_{r-1} + y_{r+1} b_{r+1} + \ldots + y_n b_n \,.$$

Not a linear combination of columns $b_1, \ldots, b_{r-1}, b_{r+1}, \ldots, b_n$ of B_a

$$\begin{array}{c} \begin{array}{cccccccc} b_1 & b_2 & \ldots & b_{r-1} & a & b_{r+1} & \ldots & b_n \end{array} \\ \begin{bmatrix} * & * & \ldots & * & * & * & \ldots & * \\ * & * & \ldots & * & * & * & \ldots & * \\ * & * & \ldots & * & * & * & \ldots & * \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \end{bmatrix} \end{array}$$

B_a

- Thus, a is not a linear combination of other columns of B_a.
- Hence, B_a is non-singular, iff $y_r \neq 0$.

Homework Problem

Qus-2: Let V be a finite dimensional vector space over $\mathbb{R}$. If S is a set of elements in V such that $\text{span}(S) = V$, what is the relationship between S and the basis of V?

Sol: Given

$$\text{span}(S) = V$$

- Every vector in V can be expressed as a linear combination of elements of S.
- S is a generating set of V.
- Since basis is a minimal generating set, we have

$$\text{Number of elements in a basis of } V \leq \text{Number of elements in } S$$

Homework Problem

Qus-3:

a) Let P be a real square matrix satisfying $P = P^T$and $P^2 = P$.

1. Can the matrix P have complex eigenvalues? If so, construct an example, else, justify your answer.
2. What are the eigenvalues of P?

Sol: 1. $P = P^T \implies P$ is symmetric.

Spectral theorem: If $A \in \mathbb{R}^{n \times n}$ is symmetric there exists an orthonormal basis of the corresponding vector space V consisting of the eigenvectors of A, and each eigenvalue is real.

→ Therefore, by the Spectral theorem, P cannot have complex eigenvalues.

Homework Problem

Sol: 2. Suppose λ is an eigenvalue of P and let v be the corresponding eigenvector. Then

$$
\begin{aligned}
& Pv = \lambda v \\
\implies & P^2 v = \lambda P v = \lambda^2 v \\
\implies & Pv = \lambda^2 v \\
\implies & \lambda v = \lambda^2 v && (\because P^2 = P) \\
\implies & (\lambda^2 - \lambda) v = \mathbf{0}
\end{aligned}
$$

Since $v \neq \mathbf{0}$, we have

$$\lambda^2 - \lambda = 0 \implies \lambda(\lambda - 1) = 0 \qquad \implies \lambda = 0 \text{ or } \lambda = 1$$

Hence, the eigenvalues of P are 0 and 1.

Homework Problem

b) Given the following matrix

$$A = \begin{bmatrix} 1 & 2 & r \\ c & 1 & 7 \\ c & 1 & 7 \end{bmatrix}$$

where c and r are arbitrary real numbers and $5.5 < r \leq 6.5$, and the fact that λ_1= 3 is one of the eigenvalues, is it possible to determine the other two eigenvalues? If so, compute them and give reasons for your answer.

Sol: (ii) Since A has two identical rows

$$\det(A) = 0$$

i.e. 0 is an eigenvalue of A.

Homework Problem

Next, we recall that

Trace of a matrix = sum of its eigenvalues

$$A = \begin{bmatrix} 1 & 2 & r \\ c & 1 & 7 \\ c & 1 & 7 \end{bmatrix}$$

- Let λ_3 be the third eigenvalue of A.
- Since trace$(A) = 9$, we have

$$3 + 0 + \lambda_3 = 9 \implies \lambda_3 = 6.$$

- Hence, the other two eigenvalues of A are 0 and 6.

Homework Problem

Qus-4: The Fibonacci sequence is defined by

$$V_n = V_{n-1} + V_{n-2}, \qquad \text{for } n \geq 2$$

with starting values $V_0 = 0$ and $V_1 = 1$. Observe that the calculation of V_k requires the calculation of $V_2, V_3, \ldots, V_{k-1}$. To avoid this, could this problem be written as an eigenvalue problem and solved for V_n directly? If so, find the explicit formula for V_n.

Sol: We can write

$$V_{n+1} = V_n + V_{n-1}$$

$$V_n = 1.V_n$$

Or

$$\begin{bmatrix} V_{n+1} \\ V_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} V_n \\ V_{n-1} \end{bmatrix}$$

Homework Problem

Define

$$F_n = \begin{bmatrix} V_{n+1} \\ V_n \end{bmatrix} \qquad \text{and} \qquad A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

Then, we have

$$F_n = A \begin{bmatrix} V_n \\ V_{n-1} \end{bmatrix} = AF_{n-1}$$

- $F_1 = AF_0$
- $F_2 = AF_1 = A^2F_0$
- $F_3 = AF_2 = A^3F_0, \ldots$

Homework Problem

Continuing like this, we get

$$F_n = A^n F_0$$

where $F_0 = \begin{bmatrix} V_1 \\ V_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

$$F_n = \begin{bmatrix} V_{n+1} \\ V_n \end{bmatrix}$$

- If we find A^n, then from F_n, we can find the value of V_n.
- Since A is a symmetric matrix, it is diagonalisable, i.e.

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

$$A = PDP^{-1}$$

where columns of P consists of eigenvectors of A and D is the diagonal matrix with eigenvalues of A as its diagonal entries.

Homework Problem

- Note that

$$
\begin{aligned}
A^n &= (PDP^{-1})(PDP^{-1})\ldots(PDP^{-1}) && (n \text{ times}) \\
&= (PD^2P^{-1})(PDP^{-1})\ldots(PDP^{-1}) && (n-1 \text{ times}) \\
&\quad\vdots \\
&= PD^nP^{-1}
\end{aligned}
$$

- Therefore, to find A^n, it is enough to find P and D.

- To find P and D, we need to find the eigenvalues and eigenvectors of A.

Homework Problem

Suppose λ is an eigenvalue of A. Then

$$\det(A - \lambda I) = 0 \qquad \Longrightarrow \begin{vmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{vmatrix} = 0$$

$$\Longrightarrow -\lambda(1-\lambda) - 1 = 0$$

$$\Longrightarrow \lambda^2 - \lambda - 1 = 0$$

$$\Longrightarrow \lambda = \frac{1 \pm \sqrt{5}}{2}$$

Thus, the eigenvalues of A are $\dfrac{1 \pm \sqrt{5}}{2}$ and hence $D = \begin{bmatrix} \frac{1+\sqrt{5}}{2} & 0 \\ 0 & \frac{1-\sqrt{5}}{2} \end{bmatrix}$.

Homework Problem

Suppose $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be an eigenvector of A corresponding to the eigenvalue $\dfrac{1+\sqrt{5}}{2}$.

Then

$$Av = \frac{1+\sqrt{5}}{2} v \qquad \Longrightarrow \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \frac{1+\sqrt{5}}{2} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$\Longrightarrow v_1 + v_2 = \frac{1+\sqrt{5}}{2} v_1 \quad \text{and } v_1 = \frac{1+\sqrt{5}}{2} v_2$$

- Solving these two equations, we see that v_2 is an independent variable.
- Choosing $v_2 = 1$, we get $v = \begin{bmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{bmatrix}$

Homework Problem

- Similarly, we can see $\begin{bmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{bmatrix}$ is an eigenvector of A corresponding to the eigenvalue $\frac{1-\sqrt{5}}{2}$.

- Hence, $P = \begin{bmatrix} \frac{1+\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \\ 1 & 1 \end{bmatrix}$.

- It is easy to compute that, $P^{-1} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & \frac{-1+\sqrt{5}}{2} \\ -1 & \frac{1+\sqrt{5}}{2} \end{bmatrix}$.

Homework Problem

- Using $A^n = PD^nP^{-1}$, we compute

$$A^n = \frac{1}{\sqrt{5}} \begin{bmatrix} \frac{1+\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^n & 0 \\ 0 & \left(\frac{1-\sqrt{5}}{2}\right)^n \end{bmatrix} \begin{bmatrix} 1 & \frac{-1+\sqrt{5}}{2} \\ -1 & \frac{1+\sqrt{5}}{2} \end{bmatrix}$$

$$= \frac{1}{\sqrt{5}} \begin{bmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^{n+1} & \left(\frac{1-\sqrt{5}}{2}\right)^{n+1} \\ * & * \end{bmatrix} \begin{bmatrix} 1 & \frac{-1+\sqrt{5}}{2} \\ -1 & \frac{1+\sqrt{5}}{2} \end{bmatrix}$$

$$= \frac{1}{\sqrt{5}} \begin{bmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1} & * \\ * & * \end{bmatrix}.$$

Homework Problem

- Using

$$F_n = A^n F_0, \qquad F_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad F_n = \begin{bmatrix} V_{n+1} \\ V_n \end{bmatrix}$$

we get

$$V_{n+1} = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right]$$

or

$$V_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^{n} - \left(\frac{1-\sqrt{5}}{2} \right)^{n} \right].$$

Homework Problem

Qus-5: Prove that if A is a square diagonalisable matrix of size $n \times n$, then $A^k \to 0$ as $k \to \infty$ if and only if $|\lambda_i| < 1, \forall i$.

Sol:

- Since A is diagonalisable, we can write

$$A = PDP^{-1}$$

where D is the diagonal matrix with eigenvalues of A as its diagonal entries.

- Multiplying A by P^{-1} on the left and by P on the right, we get

$$P^{-1}AP = P^{-1}PDP^{-1}P$$
$$D = P^{-1}AP$$

$$\implies D^k = P^{-1}A^kP$$

Homework Problem

- Suppose $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. Then

$$D = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \implies D^k = \begin{bmatrix} \lambda_1^k & 0 & \ldots & 0 \\ 0 & \lambda_2^k & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \lambda_n^k \end{bmatrix}$$

- Since $D^k = P^{-1}A^kP$,

$$\begin{aligned} & A^k \to 0 \text{ as } k \to \infty \\ \iff & D^k \to 0 \text{ as } k \to \infty \\ \iff & (\lambda_i)^k \to 0 \text{ as } k \to \infty, \ \forall i \\ \iff & |\lambda_i| < 1, \forall i \end{aligned}$$

Homework Problem

Qus-6: Construct examples of matrices for which the defect is positive, negative and zero wherever possible.

Sol: Suppose λ is an eigenvalue of a matrix A. Then

Algebraic multiplicity of λ $\longrightarrow$ Multiplicity of λ as a root of the characteristic polynomial of A

Geometric multiplicity of λ $\longrightarrow$ the maximum number of linearly independent eigenvectors associated with λ

$\longrightarrow$ **<u>Defect of λ</u>** = Algebraic multiplicity − Geometric multiplicity

Homework Problem

- Since

$$\text{Algebraic multiplicity} \geq \text{Geometric multiplicity}$$

The defect cannot be negative.

- **<u>Example of matrix with zero defect.</u>**

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

A is a diagonal matrix with eigenvalues 2 and 3. Since all the eigenvalues are distinct, the defect of each eigenvalue is 0.

It is also easy to compute the eigenvectors corresponding to each eigenvalue and check algebraic and geometric multiplicity is same for each eigenvalue.

Homework Problem

Example of matrix with positive defect.

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

The characteristic polynomial of A is

$$\lambda^2 - 2\lambda + 1 = 0$$

Therefore, 1 is the only eigenvalue of A and has multiplicity 2.

$\Longrightarrow$ Algebraic multiplicity of $1 \ = 2$.

Homework Problem

Suppose $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be an eigenvector of A corresponding to the eigenvalue 1. Then

$$Av = v \qquad \Longrightarrow \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$\Longrightarrow v_1 + v_2 = v_1$$

$$\Longrightarrow v_2 = 0 \quad \text{and } v_1 \text{ is a free variable}$$

$$\Longrightarrow v = v_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$\Longrightarrow$ Geometric multiplicity of $1 \ = 1.$

Hence, Defect of eigenvalue 1 $= 2 - 1 = 1 > 0.$

Gram Schmidt process

- The Gram-Schmidt process is a method for orthonormalizing a set of vectors in an inner product space

- The Gram-Schmidt process takes a finite, linearly independent set of vectors

$$S = \{v_1, \ldots, v_k\} \quad \text{for } k \leq n$$

and generates an orthogonal set

$$S' = \{u_1, \ldots, u_k\}$$

- The Gram-Schmidt process works as follows:

$$u_1 = v_1$$

$$u_2 = v_2 - \frac{\langle v_2, u_1 \rangle}{\langle u_1, u_1 \rangle} u_1$$

$$u_3 = v_3 - \frac{\langle v_3, u_1 \rangle}{\langle u_1, u_1 \rangle} u_1 - \frac{\langle v_3, u_2 \rangle}{\langle u_2, u_2 \rangle} u_2$$

Gram Schmidt process (contd.)

$$u_4 = v_4 - \frac{\langle v_4, u_1 \rangle}{\langle u_1, u_1 \rangle} u_1 - \frac{\langle v_4, u_2 \rangle}{\langle u_2, u_2 \rangle} u_2 - \frac{\langle v_4, u_3 \rangle}{\langle u_3, u_3 \rangle} u_3$$

$$\vdots$$

$$\vdots$$

$$u_k = v_k - \sum_{j=1}^{k-1} \frac{\langle v_k, u_j \rangle}{\langle u_j, u_j \rangle} u_j$$

- The sequence $u_1, \ldots, u_k$ is the required system of orthogonal vectors.
- The normalized vectors

$$e_1 = \frac{u_1}{||u_1||}, e_2 = \frac{u_2}{||u_2||} \ldots, e_k = \frac{u_k}{||u_k||}$$

form an orthonormal set.

Example of Gram Schmidt process

Qus: Obtain an orthogonal basis for the subspace of $\mathbb{R}^4$ spanned by

$$x_1 = (1,0,1,0), x_2 = (1,1,1,1), x_3 = (-1,2,0,1)$$

Sol: Following the Gram-Schmidt process, we set

$$v_1 = x_1 = (1,0,1,0)\,.$$

Next, we have

$$v_2 = x_2 - \frac{\langle x_2, v_1 \rangle}{||v_1||^2} v_1 = (1,1,1,1) - \frac{2}{2}(1,0,1,0) = (0,1,0,1)$$

and

Example of Gram Schmidt process (contd.)

$$v_3 = x_3 - \frac{\langle x_3, v_1 \rangle}{||v_1||^2} v_1 - \frac{\langle x_3, v_2 \rangle}{||v_2||^2} v_2$$

$$= (-1,2,0,1) + \frac{1}{2}(1,0,1,0) - \frac{3}{2}(0,1,0,1)$$

$$= (-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}).$$

The orthogonal basis so obtained is

$$\{(1,0,1,0), (0,1,0,1), (-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2})\}$$

We can normalise the vectors to get

$$\{\frac{1}{\sqrt{2}}(1,0,1,0), \frac{1}{\sqrt{2}}(0,1,0,1), (-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2})\}$$

Example of SVD

Qus: Find Singular value decomposition of the matrix

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

Sol: **Step 1**: Compute $A^T A$

$$A^T A = \begin{bmatrix} -3 & 6 & 6 \\ 1 & -2 & -2 \end{bmatrix} \begin{bmatrix} -3 & 1 \\ 6 & -2 \\ 6 & -2 \end{bmatrix} = \begin{bmatrix} 81 & -27 \\ -27 & 9 \end{bmatrix}$$

Step 2: Find the eigenvalues of $A^T A$.

The characteristic polynomial of $A^T A$ is:

$$\lambda^2 - trace(A^T A)\lambda + det(A^T A) = 0$$

$$\implies \lambda^2 - 90\lambda = 0 \qquad (\because det(A^T A) = 0)$$

Example of SVD (contd.)

$$\implies \lambda(\lambda - 90) = 0$$

$\implies \lambda_1 = 90$ and $\lambda_2 = 0$ are the two eigenvalues of $A^T A$.

Step 3: Find the singular values of A.

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{90} \text{ and } \sigma_2 = \sqrt{\lambda_2} = 0.$$

Step 4: Construct the diagonal matrix of same size as A with singular values as its diagonal entries in decreasing order.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{90} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Example of SVD (contd.)

Step 5: Find the eigenvectors (called right singular vectors) corresponding to each eigenvalue of $A^T A$.

Suppose $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ be an eigenvector of A corresponding to the eigenvalue λ_2. Then

$$A^T A z = \lambda_2 z \implies \begin{bmatrix} 81 & -27 \\ -27 & 9 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\implies 81z_1 - 27z_2 = 0 \text{ and } -27z_1 + 9z_2 = 0$$

$$\implies z_2 = 3z_1$$

$$\implies z = z_1 \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Choose $v_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ as an eigenvector corresponding to λ_2.

Example of SVD (contd.)

Similarly, we can find $v_1 = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$ as an eigenvector corresponding to λ_1.

Step 6: Write the orthogonal matrix consisting of the normalised eigenvectors of $A^T A$.

$$V = \frac{1}{\sqrt{10}} \begin{bmatrix} -3 & 1 \\ 1 & 3 \end{bmatrix}.$$

Step 7: Find the left eigenvectors (called left singular vectors) $\hat{u}_i$ using

$$\hat{u}_i = \frac{Av_i}{\sigma_i}.$$

For $i = 1$, we have.

$$\hat{u}_1 = \frac{Av_1}{\sigma_1} = \frac{1}{\sqrt{90}} \begin{bmatrix} -3 & 1 \\ 6 & -2 \\ 6 & -2 \end{bmatrix} \begin{bmatrix} -\frac{3}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} \end{bmatrix} = \frac{1}{\sqrt{900}} \begin{bmatrix} 10 \\ -20 \\ -20 \end{bmatrix}.$$

Example of SVD (contd.)

Since $\sigma_2 = 0$, to find the other vector we need to use the orthogonality condition

Let $w = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ be the next vector. Then

$$w \cdot u_1 = 0 \implies 10x - 20y - 20z = 0 \implies x = 2y + 2z$$

$$\implies w = y \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + z \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

Let

$$w_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \qquad \text{And} \qquad w_2 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

Example of SVD (contd.)

Hence the vectors orthogonal to u_1 are

$$u_2 = w_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \implies \hat{u}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and}$$

$$u_3 = w_2 - \frac{\langle w_2, u_2 \rangle}{||u_2||^2} u_2 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} - \frac{4}{5} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix} \implies \hat{u}_3 = \frac{1}{\sqrt{45}} \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}$$

Hence, $U = \begin{bmatrix} \frac{10}{\sqrt{900}} & \frac{2}{\sqrt{5}} & \frac{2}{\sqrt{45}} \\ \frac{-20}{\sqrt{900}} & \frac{1}{\sqrt{5}} & -\frac{4}{\sqrt{45}} \\ \frac{-20}{\sqrt{900}} & 0 & \frac{5}{\sqrt{45}} \end{bmatrix}.$

Example of SVD (contd.)

Step 8: The singular value decomposition of A is

$$A = U\Sigma V^T = \begin{bmatrix} \frac{10}{\sqrt{900}} & \frac{2}{\sqrt{5}} & \frac{2}{\sqrt{45}} \\ \frac{-20}{\sqrt{900}} & \frac{1}{\sqrt{5}} & -\frac{4}{\sqrt{45}} \\ \frac{-20}{\sqrt{900}} & 0 & \frac{5}{\sqrt{45}} \end{bmatrix} \begin{bmatrix} \sqrt{90} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \end{bmatrix}$$

Thank You