BITS Pilani
Pilani | Dubai | Goa | Hyderabad | Mumbai
WORK INTEGRATED LEARNING PROGRAMMES

Birla Institute of Technology and Science, Pilani

Work Integrated Learning Programmes Division

Important Definitions and Concepts in * ZC416

Mathematical Foundations for Data Science

&

Mathematical Foundations for Machine Learning

G. Venkiteswaran

Contents

1 Matrix Algebra 3
2 Solutions of Linear Systems 4
3 Vector Spaces and Linear Transformations 5
4 Inner Product and Orthogonality 7
5 Eigenvalues and Eigenvectors 10
6 Matrix Decompositions 11
7 Calculus and Vector Calculus 12
8 Gradient Descent Methods 15
9 Principal Component Analysis 16
10 KKT Conditions and Strong Duality 18
11 Support Vector Machine 19

Important Definitions and Properties

1 Matrix Algebra

1. A **matrix** is a rectangular array of numbers or functions.
2. The **size of a matrix** is definied as $m \times n$ where m is the number of rows and n is the number of columns. A matrix with $m = n$ is called a square matrix. The element in the i^{th} row and j^{th} column is denoted by a_{ij}.
3. A **zero matrix** is a matrix in which all elements are zero and denoted by $\mathbf{0}$.
4. **Equality of two matrices** is established if and only if the sizes are the same and the corresponding entries are equal.
5. **Addition of two matrices** $\boldsymbol{A}(= a_{ij})$ and $\boldsymbol{B}(= b_{ij})$ of the same size is the matrix with elements $a_{ij} + b_{ij}$.
6. **Scalar multiplication** of a matrix $\boldsymbol{A}(= a_{ij})$ with a scalar c is a matrix whose elements are ca_{ij}.
7. **Matrix multiplication** of $\boldsymbol{A}_{m\times n}(= a_{ij})$ and $\boldsymbol{B}_{n\times p}(= b_{ij})$ yields a matrix $\boldsymbol{C}_{m\times p}$ whose ij^{th} element is $\sum_{k=1}^{n} a_{ik}b_{kj}$. The equality of the number of columns of $\boldsymbol{A}$ and the number of rows of $\boldsymbol{B}$ should be noted. Also, in general $\boldsymbol{AB} \neq \boldsymbol{BA}$.
8. **Properties of matrices** - assiming suitable sizes for matrices $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ and c, k being scalars.
   i) **Associativity under addition:** $(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$
   ii) **Commutativity under addition:** $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$
   iii) **Distributivity:** $c(\boldsymbol{A} + \boldsymbol{B}) = c\boldsymbol{A} + c\boldsymbol{B}$ and $c(k\boldsymbol{A}) = ck(\boldsymbol{A})$
   iv) **Additive identity:** $\exists$ a matrix $\mathbf{0}$ such that $\boldsymbol{A} + \mathbf{0} = \boldsymbol{A}$
   v) **Additive inverse:** $\exists$ a matrix $-\boldsymbol{A}$ such that $\boldsymbol{A} + (-\boldsymbol{A}) = \mathbf{0}$
9. **Transpose** of a matrix $\boldsymbol{A}(= a_{ij})$ is obtained by changing the rows to columns and columns to rows and denoted by $\boldsymbol{A}^T(= a_{ij}^T)$. By definition $a_{ij}^T = a_{ji}$. The important properties of transposes are summarized below.
   i) $(\boldsymbol{A}^T)^T = \boldsymbol{A}$
   ii) $(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$

iii) $(c\boldsymbol{A})^T = c\boldsymbol{A}^T$, for any scalar c

iv) $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$

10. **Types of matrices**

i) **Symmetric** if $\boldsymbol{A}^T = \boldsymbol{A}$

ii) **Skew-symmetric** if $\boldsymbol{A}^T = -\boldsymbol{A}$

iii) **Upper triangular** if $a_{ij} = 0$ for $i < j$

iv) **Lower triangular** if $a_{ij} = 0$ for $i > j$

v) **Diagonal** if $a_{ij} = 0$ for $i \neq j$

vi) **Identity** if $a_{ij} = 1$ for $i = j$ and 0 otherwise

vii) **Positive definite** if $\forall\ \boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{x}^T\boldsymbol{Ax} > 0$

viii) **Positive semi-definite** if $\forall\ \boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{x}^T\boldsymbol{Ax} \geq 0$

2 Solutions of Linear Systems

1. **Elementary row operations** consist of

i) Interchanging two rows

ii) Muliplying one row with a non-zero constant

iii) Adding a constant multiple of one row to another

2. A matrix is said to be in **Row Echelon Form (REF)** if the following conditions are satisfied after performing the necessary elementary row operations.

i) All elements below the leading non-zero entry in a row are zero

ii) The leading non-zero entry in a row occurs to the right of the leading non-zero entry in the row above it

iii) The zero rows occur below the non-zero rows.

It is said to be in **Reduced Row Echelon Form (RREF)** if apart from the above, the leading non-zero entries are scaled to 1 and all elements above the leading non-zero entries are zero. The leading non-zero entres are called the **pivots** and the columns containing them are called the **pivot columns.**

3. The number of non-zero rows in the REF / RREF is called the **rank** of the matrix.

4. A **system of linear equations** is written as $\boldsymbol{Ax} = \boldsymbol{b}$ where $\boldsymbol{A}$ is a matrix of size $m \times n$. It is said to be **non-homogeneous** when $b \neq \boldsymbol{0}$ and **homogeneous** otherwise.

5. REF applied to the augmented matrix $(\boldsymbol{A}|\boldsymbol{b})$ associated with a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is a square matrix is called the **forward elimination process** of **Gaussian elimination**. Solving the resulting system is called the **backward substitution**.

6. For the general case where $\boldsymbol{A}$ is of size $m \times n$ having rank r, we have the following scenarios

   i) if $\text{rank}(\boldsymbol{A}|\boldsymbol{b}) = r$, then the system has atleast one solution. Else no solution.

   ii) if $n > r$, then $n - r$ variables can be given arbitrary values and the values of r variables can be uniquely obtained.

7. The particular solution can be uniquely found using the pivot columns and the right hand side and the non-pivotal columns can be used to find the solution of $\boldsymbol{Ax} = \boldsymbol{0}$. The general solution is then the combination of the particular solution and the general solution of $\boldsymbol{Ax} = \boldsymbol{0}$.

3 Vector Spaces and Linear Transformations

1. A **binary operator** $*$ on a non-empty set $\boldsymbol{X}$ gives us a rule to perform an operation on two given elements of the set. A binary operator is said to follow the closure property if $\forall\ a, b \subset \boldsymbol{X}, a * b \subset \boldsymbol{X}$.

2. Let G be a non-empty set with a binary operator $*$. $< G, * >$ is said to be a **Group** if

   G1. $*$ is closed

   G2. $*$ is associative, that is, $a * (b * c) = (a * b) * c\ \forall\ a, b, c \in G$

   G3. $*$ has an identity, that is, $\forall\ a \in G,\ \exists\ e \in G$ such that $a * e = a$

   G4. $*$ has an inverse, that is, $\forall\ a \in G,\ \exists\ b \in G$ such that $a * b = e$

3. A group $< G, * >$ is said to be **Abelian** if $a * b = b * a\ \forall\ a, b \in G$.

4. A **Field** is an Abelian group with respect to the usual addition and multiplication.

5. Let $\mathbf{V}$ be a non-empty set over a field $\mathbf{F}$. $\mathbf{V}$ is called a vector space if the following conditions are satisfied.

   A1. Associativity for addition: $\forall\ \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\omega} \in \mathbf{V}, \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{\omega}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{\omega}$

   A2. Commutativity for addition: $\forall\ \boldsymbol{u}, \boldsymbol{v} \in \mathbf{V}, \boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$

   A3. Additive identity: $\exists\ \boldsymbol{0} \in V$ such that $\forall\ \boldsymbol{v} \in V, \boldsymbol{v} + \boldsymbol{0} = \boldsymbol{v}$

   A4. Additive inverse: $\forall\ \boldsymbol{v} \in V, \exists\ -\boldsymbol{v} \in V$ such that $\boldsymbol{v} + (-\boldsymbol{v}) = \boldsymbol{0}$

M1. Left distributivity: $\forall\ \boldsymbol{u}, \boldsymbol{v}, \in \mathbf{V}$ and $c \in \mathbf{F}$ $c(\boldsymbol{u}+\boldsymbol{v}) = c\boldsymbol{u}+c\boldsymbol{v}$

M2. Right distributivity: $\forall\ \boldsymbol{v} \in \mathbf{V}$ and $c, d \in \mathbf{F}$ $(c+d)\boldsymbol{v} = c\boldsymbol{v}+d\boldsymbol{u}$

M3. Scalar multiplication: $\forall\ \boldsymbol{v} \in \mathbf{V}$ and $c, d \in \mathbf{F}$ $c(d\boldsymbol{v}) = (cd)\boldsymbol{v}$

M4. Multiplicative identity: $\exists\ 1 \in \mathbf{F}$ such that $\forall\ \boldsymbol{v} \in \mathbf{V}, 1.\boldsymbol{v} = \boldsymbol{v}$

6. The **linear span** of a set of vectors $S = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$ is the set

$$LS(S) = \left\{ \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i \text{ where } \alpha_i \in \mathbf{F}\ \forall\ i = 1, 2, \ldots, n \right\}.$$

7. A set of vectors $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$ is **linearly independent** if $\sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i = \boldsymbol{0}$ has $\alpha_i = 0$ as the only solution $\forall\ i = 1, 2, \ldots, n$. If at least one of the $\alpha_i \neq 0$, then it is said to be **linearly dependent**.

8. A set of vectors $S = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$ is called a **basis** of a vector space V over F if S is linearly independent and spans V. An equivalent definition is that S is the maximum number of linearly independent elements in V. A vector space can have multiple bases.

9. $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2$ are linearly dependent if there exists a $\lambda \in \mathbb{R}$ such that $\boldsymbol{v}_2 = \lambda \boldsymbol{v}_1$ or $\boldsymbol{v}_1 = \lambda \boldsymbol{v}_2$. In $\mathbb{R}^2$, if we cannot express one of the two vectors as a linear multiple of the other then they are linearly independent.

10. The number of elements in a basis is called the **dimension** of V

11. Let $\mathbf{V}$ be a vector space over F and $\boldsymbol{W} \subset \mathbf{V}$. $\boldsymbol{W}$ is a **subspace** over F if a) $\boldsymbol{0} \in \boldsymbol{W}$ and b) $\alpha\boldsymbol{\omega}_1 + \beta\boldsymbol{\omega}_2 \in \boldsymbol{W}\ \forall\ \boldsymbol{\omega}_1, \boldsymbol{\omega}_2 \in \boldsymbol{W}$ and $\alpha, \beta \in F$.

12. **Row space** of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the linear span of the row vectors of the matrix $\boldsymbol{A}$, denoted by row($\boldsymbol{A}$) and is a subspace of $\mathbb{R}^n$.

13. **Column space** of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the linear span of column vectors of the matrix $\boldsymbol{A}$, denoted by col($\boldsymbol{A}$) and is a subspace of $\mathbb{R}^m$.

14. **Equivalence of dimensions** dim row($\boldsymbol{A}$) = dim col($\boldsymbol{A}$) = rank($\boldsymbol{A}$).

15. The **null space** of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is defined as $\mathrm{N}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$ and is a subspace of $\mathbb{R}^n$.

16. **Rank Nullity Theorem:** for $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, rank($\boldsymbol{A}$) + dim N($\boldsymbol{A}$) = n.

17. A mapping $T : \mathbf{V} \to \boldsymbol{W}$, where $\mathbf{V}$ and $\boldsymbol{W}$ are vector spaces over the same field F is called a **linear transformation** if

    i) $T(\boldsymbol{v}_1 + \boldsymbol{v}_2) = T(\boldsymbol{v}_1) + T(\boldsymbol{v}_2)\ \forall\ \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbf{V}$

ii) $T(c\boldsymbol{v}) = cT(\boldsymbol{v})\ \forall\ c \in F, \boldsymbol{v} \in \mathbf{V}$

18. The sets $\boldsymbol{R}(T) = \{\boldsymbol{u} \in \boldsymbol{W} \mid \boldsymbol{u} = T(\boldsymbol{v}) \text{ for some } \boldsymbol{v} \in \mathbf{V}\}$ and $\boldsymbol{N}(T) = \{\boldsymbol{v} \in \mathbf{V} \mid T(\boldsymbol{v}) = \mathbf{0}\}$ are subspaces of $\boldsymbol{W}$ and $\mathbf{V}$ and are called the and **null space** respectively.

19. The **Rank-Nullity theorem** for linear transformation states that

$$\dim(R(T)) + \dim(N(T)) = \dim(\mathbf{V}).$$

20. Associated with every linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ is an $n \times m$ matrix which is the **matrix representation of the linear transformation**.

21. A linear transformation of a vector in $\mathbb{R}^n$ is a combination of rotation and scaling.

4 Inner Product and Orthogonality

1. For $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^n$, $\boldsymbol{a}^T\boldsymbol{b}$ is called the **dot product** of $\boldsymbol{a}$ and $\boldsymbol{b}$ and is denoted by $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ or $\boldsymbol{a} \cdot \boldsymbol{b}$.

2. Properties of dot product

   i) $\langle k\boldsymbol{u} + l\boldsymbol{v}, \boldsymbol{w} \rangle = k\langle \boldsymbol{u}, \boldsymbol{w} \rangle + l\langle \boldsymbol{v}, \boldsymbol{w} \rangle,\ \forall k, l \in \mathbb{R},\ \forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ **(linearity)**

   ii) $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{u} \rangle\ \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ **(symmetry)**

   iii) $\langle \boldsymbol{u}, \boldsymbol{u} \rangle \geq 0,\ \forall \boldsymbol{u} \in \mathbb{R}^n\ \langle \boldsymbol{u}, \boldsymbol{u} \rangle = 0$ if and only if $\boldsymbol{u} = \mathbf{0}$ **(positive definite)**

3. The **norm** of a vector $\boldsymbol{a} \in \mathbb{R}^n$ is defined as $\|\boldsymbol{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle} \geq 0$.

4. Properties of norm

   i) $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$ **(Cauchy Schwarz inequality)**

   ii) $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$ **(Triangle inequality)**

5. For any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$

$$-1 \leq \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \leq 1$$

6. A bilinear mapping Ω is a mapping with two arguments and is linear in both arguments: Let $\mathbf{V}$ 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 $\Omega(\boldsymbol{x}, \lambda\boldsymbol{y} + \psi\boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{y}) + \psi\Omega(\boldsymbol{x}, \boldsymbol{z})$.

7. Let $\mathbf{V}$ be a vector space and $\Omega : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ 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 $\forall x \in \mathbf{V} \setminus \{0\}$, $\Omega(\boldsymbol{x}, \boldsymbol{x}) > 0$ and $\Omega(0, 0) = 0$.

8. A bilinear mapping Ω is a mapping with two arguments and is linear in both arguments: Let $\mathbf{V}$ 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 $\Omega(\boldsymbol{x}, \lambda\boldsymbol{y} + \psi\boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{y}) - \psi\Omega(\boldsymbol{x}, \boldsymbol{z})$.

9. Let $\mathbf{V}$ be a vector space and $\Omega : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ 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 $\forall x \in \mathbf{V} \setminus \{0\}$, $\Omega(\boldsymbol{x}, \boldsymbol{x}) > 0$ and $\Omega(0, 0) = 0$.

10. A positive-definite, symmetric bilinear mapping $\Omega : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ is called an inner product. To denote an inner product on $\mathbf{V}$ we generally write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. The pair $(\mathbf{V}, \langle ., . \rangle)$ is called an inner product space.

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

12. Inner products and norms are closely related in the sense that any inner product induces a norm: $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$

13. Not every norm is induced by an inner product, for example the Manhattan norm.

14. For an inner product vector space $(V, \langle ., . \rangle)$, the induced norm $\|.\|$ satisfies the Cauchy-Schwarz inequality: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$.

15. The **angle** between two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ is
$$\alpha = \cos^{-1}\left(\frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}\right).$$

16. Two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ are said to be **orthogonal** if $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = 0$.

17. A set of vectors $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ is called as an **orthogonal set** if $\boldsymbol{v}_i$ is orthogonal to $\boldsymbol{v}_j$, $\forall i \neq j$.

18. A set of vectors $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ is called as an **orthonormal set** if $\boldsymbol{v}_i$ is orthogonal to $\boldsymbol{v}_j$, $\forall i \neq j$ and each $\boldsymbol{v}_i$ is of unit norm.

19. The **projection** of $\boldsymbol{v}_2$ onto the vector $\boldsymbol{v}_1$ is $\boldsymbol{v} = \lambda \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}$ where $\lambda = \frac{\langle \boldsymbol{v}_2, \boldsymbol{v}_1 \rangle}{\|\boldsymbol{v}_1\|}$.

20. The canonical basis or the standard basis forms an orthonormal set.

21. A matrix $\boldsymbol{A}$ is said to be an orthonormal matrix if the column vectors of the matrix form a orthonormal set and $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} = \boldsymbol{A}^T\boldsymbol{A}$.

22. **Gram-Schmidt Process** If $\{\boldsymbol{u}_1, \cdots \boldsymbol{u}_m\}$ be a basis of a subspace U of $\mathbb{R}^n$, and

$$\begin{aligned}
\boldsymbol{v}_1 &= \boldsymbol{u}_1 \\
\boldsymbol{v}_2 &= \boldsymbol{u}_2 - \frac{\boldsymbol{u}_2 \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\boldsymbol{v}_1 \\
\boldsymbol{v}_3 &= \boldsymbol{u}_3 - \frac{\boldsymbol{u}_3 \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\boldsymbol{v}_1 - \frac{\boldsymbol{u}_3 \cdot \boldsymbol{v}_2}{\boldsymbol{v}_2 \cdot \boldsymbol{v}_2}\boldsymbol{v}_2 \\
&\vdots \\
\boldsymbol{v}_m &= \boldsymbol{u}_m - \frac{\boldsymbol{u}_m \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\boldsymbol{v}_1 - \frac{\boldsymbol{u}_m \cdot \boldsymbol{v}_2}{\boldsymbol{v}_2 \cdot \boldsymbol{v}_2}\boldsymbol{v}_2 - \cdots - \frac{\boldsymbol{u}_m \cdot \boldsymbol{v}_{m-1}}{\boldsymbol{v}_{m-1} \cdot \boldsymbol{v}_{m-1}}\boldsymbol{v}_{m-1}
\end{aligned}$$

Then, $\{\boldsymbol{v}_1, \cdots \boldsymbol{v}_m\}$ is an orthogonal basis for U and $\text{LS}\{\boldsymbol{v}_1, \cdots \boldsymbol{v}_k\} = \text{LS}\{\boldsymbol{u}_1, \cdots \boldsymbol{u}_k\}$ for $1 \leq k \leq m$.

23. 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 $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$.

24. If we use the dot product as the inner product, then the distance is called the Euclidean distance.

25. The mapping $d : V \times V \to \mathbb{R}$ is called a metric.

26. d is positive-definite which means $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in V$. $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \implies \boldsymbol{x} = \boldsymbol{y}$.

27. d is symmetric which means $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in V$.

28. d obeys the triangle inequality as follows: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}) \ \ \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$

29. 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 $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. $\boldsymbol{Q}^T = \boldsymbol{L}^{-1}\boldsymbol{A}^T$ is an orthogonal matrix whose rows are orthogonal. This is the alternatuive form of **Gram-Schmidt Orthogonalization**.

5 Eigenvalues and Eigenvectors

1. The **minor** of an element a_{ij}, denoted by m_{ij}, of a square matrix $\boldsymbol{A}(=a_{ij})$ of size $n \times n$ is the determinant of the submatrix of $\boldsymbol{A}$ obtained by deleting the i^{th} row and the j^{th} column from $\boldsymbol{A}$.

2. The **cofactor** of the element a_{ij} is defined as $(-1)^{i+j}$.

3. The **determinant** of a matrix is then defined as $\text{Det}(\boldsymbol{A}) = \sum_{k=1}^{n} a_{rk}c_{rk}$ where we have taken the product over the r^{th} row. The value is invariant for the product taken over columns. The following are some of the interesting properties.

   i) $\text{Det}(\boldsymbol{A}) = 0$ whenever there is linear dependence in the rows / columns.

   ii) Multiplying the r^{th} row by a scalar c is same as multiplying the determinant value with c.

   iii) $\text{Det}(c\boldsymbol{A}) = c^n\text{Det}(\boldsymbol{A})$

   iv) $\text{Det}(\boldsymbol{AB}) = \text{Det}(\boldsymbol{A})\text{det}(\boldsymbol{B})$

4. The **adjoint** of a matrix $\boldsymbol{A}$, denoted by $\text{adj}(\boldsymbol{A})$ is defined as the transpose of the cofactor matrix.

5. The **inverse** of a matrix $\boldsymbol{A}$ is defined as $A^{-1} = \dfrac{\text{adj}(\boldsymbol{A})}{\text{Det}(\boldsymbol{A})}$.

6. An alternative way to calculate the inverse of $\boldsymbol{A}$ is to start with the augmented matrix $[\boldsymbol{A} \mid \boldsymbol{I}]$ and use elementary row operations to convert this to $[\boldsymbol{I} \mid A^{-1}]$. This procededure is called the **Gauss Jordan** method.

7. The roots of the **characteristic equation** $\text{Det}(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ are called the **eigenvalues**. Any non-zero vector $\boldsymbol{x}$ which satisfies the equation $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ is called the **eigenvector** corresponding to the eigenvalue λ

8. There are n eigenvalues for a matrix $\boldsymbol{A}$ of size $n \times n$ and they can be real or complex.

9. For a square matrix the **rank** is the number of non-zero eigenvalues.

10. A symmetric positive definite matrix has full rank.

11. The eigenvalues of a positive definite matrix are always positive.

12. The sum of diagonal elements of a matrix, called the **trace** is equal to the sum of the eigenvalues and the product of the eigenvalues is equal to the determinant.

13. The eigenvalues of a symmetric matrix are all real whereas that of a skew-symmetric matrix are either purely complex or zero.

14. The eigenvectors corresponding to distinct eigenvalues are linearly independent.

15. The **Spectral theorem** states that for a symmetric matrix $\boldsymbol{A}$ of size $n \times n$, the eigenvalues are real and that the eigenvectors for an orthogonal basis of $\mathbb{R}^n$.

6 Matrix Decompositions

1. If $\boldsymbol{x_1, x_2, \ldots, x_n}$ are the linearly independent eigenvectors of a matrix correspoding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where $\boldsymbol{P}$ is the matrix with eigenvectors as columns. This is called the **eigenvalue decomposition**.

2. Under certain conditions a matrix $\boldsymbol{A}$ can be decomposed into a product of a lower triangular matrix ($\boldsymbol{L}$) and an upper triangular matrix ($\boldsymbol{U}$), that is $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. This method is computationally effective in solving systems of equations with the same matrix $\boldsymbol{A}$ and different right hand sides. This procedure is called **LU decomposition method**.

3. The steps in LU decomposition include, solving for $\boldsymbol{y}$ in $\boldsymbol{L}\boldsymbol{y} = \mathbf{b}$ and then $\boldsymbol{U}x = \boldsymbol{y}$. The following methods are popular.

   i) **Doolittle's method** if $\boldsymbol{L}_{ii} = 1 \; \forall \; i$

   ii) **Crout's method** if $\boldsymbol{U}_{ii} = 1 \; \forall \; i$

   iii) **Cholesky's method** if $\boldsymbol{A}$ is positive definite and $\boldsymbol{L} = \boldsymbol{U}^T$.

4. Any matrix $\boldsymbol{A}_{m\times n}$ can be written as $\boldsymbol{A} = \boldsymbol{U}\Sigma\mathbf{V}^T$ where $\boldsymbol{U}_{m\times m}, \mathbf{V}_{n\times n}$ are orthonormal matrices and $\Sigma_{m\times n}$ is the matrix of singular values. This is called the **Singular Value Decomposition** of $\boldsymbol{A}$.

5. The matrix $\boldsymbol{U}$ consists of eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ written as columns and $\mathbf{V}$ has the eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$ written as columns.

6. The elements of Σ are all non-negative and are arranged in decreasing order along the diagonal.

7. If $\boldsymbol{u_1, v_2, \ldots, u_m}$ and $\boldsymbol{v_1, v_2, \ldots, v_n}$ are the columns of $\boldsymbol{U}$ and $\mathbf{V}$ and σ_{ii}, $i = 1, 2, \ldots, r$ are the non-zero diagonal elements of Σ, of a matrix $\boldsymbol{A}$ of rank r, then a rank k approximation of $\boldsymbol{A}$ (where $k \leq r$) is given by $\sum_{i=1}^{k} \boldsymbol{u}_k \sigma_{ii} \boldsymbol{v}_k^T$.

8. The first diagonal element of Σ is called the 2 norm of $\boldsymbol{A}$.

9. If $\boldsymbol{A}$ is an $m \times n$ matrix with $m \geq n$ and all the columns linearly independent, then $\boldsymbol{A}$ has a decomposition of the form $\boldsymbol{QR}$ where $\boldsymbol{Q}$ is a matrix whose columns are orthogonal and $\boldsymbol{R}$ is an upper triangular matrix. This is called the **QR decomposition** of $\boldsymbol{A}$. The basic idea is to use Gram-Schmidt Orthogonolization process to get $\boldsymbol{Q}$ and get $\boldsymbol{R}$ via $\boldsymbol{Q}^T \boldsymbol{A}$.

7 Calculus and Vector Calculus

1. The **derivative** of f at x is defined as the limit

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

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

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

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

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

4. Rules for differentiation. We denote the derivative of f by f'

   i) **Product Rule:** $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$

   ii) **Sum Rule:** $(f(X) + g(X))' = f'(X) + g'(x)$

   iii) **Quotient Rule:** $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$

   iv) **Chain Rule:** $(g(f(x))' = (g \circ f)'(x) = g'(f(x))f'(x)$

5. 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**

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

6. The rules for partial differentiation are
   i) **Product rule**: $\frac{\partial}{\partial x}(f(x)g(x)) = \frac{\partial f}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x}$
   ii) **Sum rule**: $\frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$
   iii) **Chain rule**: $\frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x}(g(f(x)) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$

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

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

where d denotes the gradient and ∂ partial derivatives.

8. 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

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

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

9. **Taylor's Theorem:** Suppose $f : (a, b) \to R$ is a function on (a, b), where a, b in R with $a < b$. Assume that f is n-times differentiable in the open interval (a, b) and $f, f', f'', \ldots f^{n-1}$ all extend continuously to the closed interval $[a, b]$, such that the extended functions are still called $f, f'' \ldots f^{n-1}$. Then there exists $c \in (a, b)$ such that

$$f(b) = \sum_{k=0}^{k=n-1} \frac{f^k(a)}{k!}(b - a)^k + \frac{f^n(c)}{n!}(b - a)^n$$

10. For $n = 1$, the statement of Taylor's theorem boils down to the **Mean-Value Theorem** which is that if a function f is continuous on $[a, b]$ and differentiable on the interval (a, b), then there exists a value $c \in (a, b)$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$

11. 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 is symmetric. The Hessian is denoted as

$$\nabla^2_{x,y} f(x, y) = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}$$

12. For a function $f(x, y)$ which is twice differentiable in a neighbourhood of the point (a, b) and $f_x(a, b) = f_y(a, b) = 0$, the expression for $f(a + h, b + k)$ simplifies to $f(a + h, b + k) - f(a, b) = \frac{1}{2}(h^2 f_{xx} + 2hk f_{xy} + k^2 f_{yy})|_{a+ch,b+ck} =: Q(c)$. If $Q(0) \neq 0$, the sign of $Q(c)$ for small c will be the same as the sign of $Q(0)$ for sufficiently small values of h and k.

13. Since $f_{xx}Q(0) = (hf_{xx} + kf_{xy})^2 + (f_{xx}f_{yy} - f_{xy}^2)k^2$.

    i) If $f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b) then $Q(0) < 0$ for all sufficiently small non-zero values of h and k, then f has a local maximum value at (a, b).

    ii) If $f_{xx} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b) then $Q(0) > 0$ for all sufficiently small non-zero values of h and k, then f has a local minimum value at (a, b).

    iii) If $f_{xx}f_{yy} - f_{xy}^2 < 0$ at (a, b) there are combinations of small values for h and k for which $Q(0) > 0$ and other combinations of h and k for which $Q(0) < 0$. This means that f has a saddle point at (a, b).

    iv) If $f_{xx}f_{yy} - f_{xy}^2 = 0$ at (a, b) another test is needed.

14. The gradient of an $m \times n$ matrix $\boldsymbol{A}$ with respect to a $p \times q$ matrix $\boldsymbol{B}$, 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

$$J_{ijkl} = \frac{\partial \boldsymbol{A}_{ij}}{\partial \boldsymbol{B}_{kl}}$$

15. Some useful **gradient identities**

    i) $\frac{\partial}{\partial X} f(X)^T = \left(\frac{\partial f(X)}{\partial X}\right)^T$

ii) $\frac{\partial}{\partial X}\text{tr}(f(X)) = \text{tr}\left(\frac{\partial f(X)}{\partial X}\right)$

iii) $\frac{\partial}{\partial X}\det(f(X)) = \det(f(x))\text{tr}\left(f(X)^{-1}\frac{\partial f(X)}{\partial X}\right)$

iv) $\frac{\partial}{\partial X}f(X)^{-1} = -f(X)^{-1}\frac{\partial f(X)}{\partial X}f(X)^{-1}$

v) $\frac{\partial a^T X^{-1} b}{\partial X} = -(X^{-1})^T a b^T (X^{-1})^T$

vi) $\frac{\partial x^T a}{\partial x} = a^T$

vii) $\frac{\partial a^T x}{\partial x} = a^T$

viii) $\frac{\partial a^T X b}{\partial X} = ab^T$

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

x) $\frac{\partial}{\partial s}(x - As)^T W(x - As) = -2(x - As)^T WA$, for symmetric W.

8 Gradient Descent Methods

1. Gradient descent method: find the optimum of J, say at $J(\boldsymbol{x}_*)$, we can start at some initial point $\boldsymbol{x}_0$ and then iterate according to $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \alpha_i((\nabla J)(\boldsymbol{x}_i))^T$, where α is the **learning rate**.
2. The standard gradient descent procedure is a batch optimization method in that the update step considers the gradient of the entire loss function $L(\theta)$, i,e $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha_i \nabla L(\boldsymbol{\theta}_i)^T = \boldsymbol{\theta}_i - \alpha_i \sum_{n=1}^{n=N} \nabla L_n(\boldsymbol{\theta}_i)^T$.
3. With a learning rate dependent on time, the update step becomes $\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \alpha_t \nabla J$.
4. In line search, the step-size α_t is computed as $\alpha_t = \min_{\alpha} J(\boldsymbol{w}_t + \alpha \boldsymbol{g}_t)$.
5. The first step in optimization is to identify a range $[a, b] = [0, \alpha_{\max}]$ in which to perform the search for the optimum α.
6. It is then possible to narrow the search interval by using **binaryn search** or **golden section search** methods.
   i) In binary search, if the objective function is found to be increasing at $\frac{a+b}{2}$, we narrow the interval to $[a, \frac{a+b+\epsilon}{2}]$ and continue the search. Otherwise we narrow the interval to $[\frac{a+b}{2}, b]$ and continue the search. ϵ is usually taken as 10^{-8}.

ii) In golden section search, When $\alpha = a$ yields the minimum for the objective function, i.e $H(\alpha)$, we can drop the interval $(m_1, b]$. Similarly when $\alpha = b$ yields the minimum for $H(\alpha)$ we can drop the interval $[a, m_2)$. When $\alpha = m_1$ is the value at which the minimum is achieved we can drop $(m_2, b]$. When $\alpha = m_2$ is the value at which the minimum is achieved we can drop $[a, m_1)$.

7. In case of **mean centering** a vector of column-wise means is subtracted from each data point.

8. In case of **feature normalization**, each feature value is divided by its standard deviation.

9. In case of **min-max normalization** we scale the jth feature of the ith datapoint as follows: $x_{ij} = \frac{x_{ij} - \min_j}{\max_j - \min_j}$.

10. Modified Gradient Descent Methods

    i) The set S of data points can be treated as a sample and a sample-centric objective function can be constructed as follows: $J(S) = \sum_{i \in S} (\boldsymbol{w}^T \boldsymbol{X}_i - y_i)^2$. This is called as **mini batch gradient descent**.

    ii) In the extreme case S can contain only one index chosen at random, the approach is called as **stochastic gradient descent**.

    iii) The normal update procedure for gradient descent can be written as $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{v}$ where $v \leftarrow -\alpha \frac{\partial J}{\partial \boldsymbol{w}}$.

    iv) The gradient descent with **momentum based**, for $\beta \in (0, 1)$, the update can be written as $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{v}$ where $\boldsymbol{v} \leftarrow \beta \boldsymbol{v} - \alpha \frac{\partial J}{\partial \boldsymbol{w}}$.

    v) **AdaGrad Method:** The update step becomes $w_i \leftarrow w_i - \frac{\alpha}{\sqrt{A_i}} \frac{\partial J}{\partial w_i}$, $\forall i$ where $A_i \leftarrow A_i + \left(\frac{\partial J}{\partial w_i}\right)^2$, $\forall i$.

    vi) **RMS Prop Method:** The update step is $w_i \leftarrow w_i - \frac{\alpha}{\sqrt{A_i}} \frac{\partial J}{\partial w_i}$, $\forall i$ where $A_i \leftarrow \rho A_i + (1 - \rho) \left(\frac{\partial J}{\partial w_i}\right)^2$ with $\rho \in (0, 1)$.

    vii) **Adams Method:** The following update is used at the tth iteration: $w_i \leftarrow w_i - \frac{\alpha_t F_i}{\sqrt{A_i}}$ where $\alpha_t = \alpha \frac{\sqrt{1-\rho^t}}{1-\rho_f^t}$. And $A_i \leftarrow \rho A_i + (1 - \rho) \left(\frac{\partial J}{\partial w_i}\right)^2$ with $\rho \in (0, 1)$, $F_i \leftarrow \rho_f F_i + (1 - \rho_f) \frac{\partial J}{\partial w_i}$ with $\rho_f \in (0, 1)$.

9 Principal Component Analysis

1. Consider an iid dataset $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}, \boldsymbol{x}_n \in \mathbb{R}^D$ with mean $\boldsymbol{0}$ which possesses the **data covariance matrix** $\boldsymbol{S} = \frac{1}{N} \sum_{n=1}^{n=N} \boldsymbol{x_n} \boldsymbol{x_n^T}$.

2. We assume there exists a lower-dimensional compressed representation $\boldsymbol{z}_n$ of $\boldsymbol{x}_n$ such that $\boldsymbol{z}_n = \boldsymbol{B}^T\boldsymbol{x}_n$ where the **projection matrix** $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m] \in \mathbb{R}^{D\times M}$.

3. The columns of B are orthonormal which means $\boldsymbol{b}_i^T\boldsymbol{b}_j = 0$ when $i \neq j$ and $\boldsymbol{b}_i^T\boldsymbol{b}_i = 1$.

4. There exists a linear relationship between the original data $\boldsymbol{x}$, its low-dimensional code $\boldsymbol{z}$ and the compressed data $\tilde{\boldsymbol{x}}$: $\boldsymbol{z} = \boldsymbol{B}^T\boldsymbol{x}$, and $\tilde{\boldsymbol{x}} = \boldsymbol{B}\boldsymbol{z}$ for a suitable matrix $\boldsymbol{B}$.

5. PCA can then be viewed as a dimensionality reduction algorithm that maximizes the variance in the low-dimensional representation of the data to retain as much information as possible.

6. For the data covariance matrix $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{n=N} \boldsymbol{x_n}\boldsymbol{x_n}^T$ we assume centred data, and can make this assumption without loss of generality.

7. Finding the direction $\boldsymbol{b}_1$ that maximizes variance can be set up as a constrained optimization problem

$$\max \boldsymbol{b}_1^T\boldsymbol{S}\boldsymbol{b}_1 \text{ subject to}$$
$$\|\boldsymbol{b}_1\| = 1$$

8. Our objective function boils to maximizing λ which means we are looking for the eigenvector of $\boldsymbol{S}$ that corresponds to its largest eigenvalue.

9. The mth **principal component** can be found by subtracting from the data the contribution of the first $m-1$ components $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_{m-1}$. Essentially, we are trying to find principal components that compress the remainder of the information.

10. We then arrive at a new data matrix $\hat{\boldsymbol{X}} = \boldsymbol{X} - \sum_{i=1}^{m-1} \boldsymbol{b}_i\boldsymbol{b}_i^T\boldsymbol{X}$ where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$ contains the data points as column vectors and $\boldsymbol{B}_{m-1} = \sum_{i=1}^{m-1} \boldsymbol{b}_i\boldsymbol{b}_i^T$ is a projection matrix that projects $\boldsymbol{X}$ onto the subspace spanned by $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_{m-1}$. $\hat{\boldsymbol{S}}$ is the covariance matrix of the data matrix $\hat{\boldsymbol{X}}$.

11. The sets of eigenvectors for $\hat{\boldsymbol{S}}$ and $\boldsymbol{S}$ are the same.

12. Assume the SVD of X as $X = U\Sigma V^T$. Then

$$S = \frac{1}{N}XX^T = \frac{1}{N}U\Sigma\Sigma^T U^T$$

13. The eigenvalues λ_d of S are related to the singular values of X via

$$\lambda_d = \frac{\sigma_d^2}{N} \tag{1}$$

14. Consider the best rank-M approximation of X defined as $\tilde{X}_M$

$$\tilde{X}_M = \text{argmin}_{rank(A)<=M} \|X - A\|_2 \tag{2}$$

15. The eigenvectors of XX^T can be computed from the eigenvectors of X^TX using the equation

$$\frac{1}{N} XX^T Xc_m = \lambda_m Xc_m \tag{3}$$

16. Steps for the computation of the PCA:

    i) We need to **standardize** x_* using the mean and **standard deviation** of the training data in the d th dimension

$$x_*^{(d)} = \frac{x_*^{(d)} - \mu_d}{\sigma_d}, \quad d = 1, ..., D \tag{4}$$

    where $x_*^{(d)}$ is the d th component of x_*.

    ii) We obtain the projection as

$$\tilde{x} = BB^T x_* \tag{5}$$

    iii) The coordinates are

$$z_* = B^T x_* \tag{6}$$

    with respect to the basis of the principal subspace.

10 KKT Conditions and Strong Duality

1. The **primal problem** is $\min f(\boldsymbol{x})$ subject to $g_i(\boldsymbol{x}) \leq 0, 1 \leq i \leq m$. Optimization is performed over the primal variables $\boldsymbol{x}$.
2. We create the **Lagrangian** of the given constrained optimization problem as follows: $\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{i=m} \lambda_i g_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{x})$, where $\lambda_i \geq 0$ for all i.
3. The associated Lagrangian **dual** problem is $\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ subject to $\boldsymbol{\lambda} \geq 0$ where $\boldsymbol{\lambda}$ are dual variables.
4. **Minmax inequality:** $\boldsymbol{x}, \boldsymbol{y}$: $\max_{\boldsymbol{y}} \min_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{y}) \leq \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \phi(\boldsymbol{x}, \boldsymbol{y})$.
5. A set C is called a **convex set** C if for any $x, y \in C$, $\theta x + (1-\theta)y \in C$, for $0 \leq \theta \leq 1$.
6. The function is a **convex function** if for any $\boldsymbol{x}, \boldsymbol{y} \in C$, $f(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1 - \theta) f(\boldsymbol{y})$

7. Another way of looking at a convex function is to use the gradient: for any two points $\boldsymbol{x}$ and $\boldsymbol{y}$, we have $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$.

8. For a primal optimization problem $\min f(\boldsymbol{x})$ subject to $g_i(\boldsymbol{x}) \leq 0$ for $i = 1, 2, \ldots, m$ and $h_j(\boldsymbol{x}) = 0$ for $j = 1, 2, \ldots, p$, we say that it obeys **Slater's condition** if the objective function f is convex, the constraints g_i are all convex and the contraint functions h_i are all linear and there exists a point $\bar{\boldsymbol{x}}$ in the interior of the region, i.e $g_i(\bar{\boldsymbol{x}}) < 0$ for all $i \in [m]$ and $h_j(\bar{\boldsymbol{x}}) = 0$ for all $j \in [p]$.

9. Suppose Slater's condition holds and the region has a non-empty interior. Then we have **strong duality**.

10. Given a primal optimization problem, we say that $\boldsymbol{x}^*$ and $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ respect the **Karash-Kuhn-Tucker conditions** if:

    i) $g_i(\boldsymbol{x}^*) \leq 0 \; \forall i \in [m]$.

    ii) $h_i(\boldsymbol{x}^*) = 0 \; \forall i \in [p]$.

    iii) $\lambda_i^* \geq 0 \; \forall i \in [m]$.

    iv) $\lambda_i^* g_i(\boldsymbol{x}^*) = 0 \; \forall i \in [m]$.

    v) $\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{i=m} \lambda_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{i=1}^{i=p} \nu_i^* \nabla h_i(\boldsymbol{x}^*) = 0$.

11. For any optimization problem, if strong duality holds then any primal optimal solution $\boldsymbol{x}^*$ and dual optimal solution $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ respect the KKT conditions. Conversely if f and g_i are convex for all $i \in [m]$ and h_i are affine for all $i \in [p]$ then the KKT conditions are sufficient for strong duality.

11 Support Vector Machine

1. $\boldsymbol{\omega}$ is a normal vector to the hyperplane $\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle + b = 0$

2. **Hard Margin SVM**

$$\min_{\boldsymbol{\omega}, b} \frac{1}{2} \|w\|^2$$
$$\text{subject to } y_i(\langle w, x_i \rangle + b) \geq 1, \; \forall i = 1, \cdots N$$

3. Using the **Lagrangian** formulation and setting the partial derivatives to zero, $\boldsymbol{\omega} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$.

4. **Classification** of $\boldsymbol{x}$ is based on $\text{sign}(\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle + b)$

5. **Support vectors** are those which are on the left and right margins.

6. Using KKT conditions, $\alpha_i = 0$ for $\boldsymbol{x}_i$ that are not support vectors.

7. The **dual** of the Lagrangian is

$$D(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha} \geq \mathbf{0}} \frac{1}{2} \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

8. **Soft Margin SVM**

$$\begin{aligned} &\min_{\boldsymbol{\omega}, b} \frac{1}{2} \|w\| + C \sum_{n=1}^{N} \xi_n \\ &\text{subject to } y_n(\langle w, x_n \rangle + b) \geq 1 - \xi_n \\ &\xi_n \geq 0, \ n = 1, \cdots, N \end{aligned}$$

9. **Hinge's Loss Function**

$$\min_{\boldsymbol{\omega}, b} \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N} \max\{0, 1 - y_n(\langle w, x_n \rangle + b)\}$$

   where C is the parameter that controls the width of the margin. Larger C implies lesser width and smaller C signifies a larger width.

10. A **kernel function** is a function that corresponds to an **inner product** in some expanded feature space. For a function $\phi : \boldsymbol{x} \rightarrow \phi(\boldsymbol{x})$, the dot product can be replaced by $K(\boldsymbol{x}_i . \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. A widely used ϕ is $\phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$.

Index

Associativity under addition, 3
backward substitution, 5
eigenvector, 10
linearly dependent, 6
orthonormal set, 8

Abelian, 5
AdaGrad Method, 16
Adams Method, 16
Addition of two matrices, 3
Additive identity, 3
Additive inverse, 3
adjoint, 10
angle, 8

basis, 6
binary operator, 5
binary search, 15

Cauchy Schwarz inequality, 7
Chain rule, 12, 13
characteristic equation, 10
Cholesky's method, 11
Classification, 19
cofactor, 10
Column space, 6
Commutativity under addition, 3
convex function, 18
convex set, 18
Crout's method, 11

data covariance matrix, 16
derivative, 12
determinant, 10
Diagonal, 4
dimension, 6
Distributivity, 3
Doolittle's method, 11
dot product, 7
dual, 18, 20

eigenvalue decomposition, 11
eigenvalues, 10
Elementary row operations, 4
Equality of two matrices, 3
Equivalence of dimensions, 6

feature normalization, 16
Field, 5
forward elimination process, 5

Gauss Jordan, 10
Gaussian elimination, 5
golden section search, 15
grad, 13
gradient identities, 14
Gram-Schmidt Process, 9
Group, 5

Hard Margin SVM, 19
Hessian, 14
Hinge's Loss Function, 20
homogeneous, 4

Identity, 4
inner product, 20
inverse, 10

Jacobian, 13

Karash-Kuhn-Tucker conditions, 19
kernel function, 20

Lagrangian, 18, 19
learning rate, 15
linear span, 6
linear transformation, 6
linearity, 7
linearly independent, 6
Lower triangular, 4
LU decomposition method, 11

matrix, 3
Matrix multiplication, 3

matrix representation of the linear transformation, 7
mean centering, 16
Mean-Value Theorem, 14
min-max normalization, 16
mini batch gradient descent, 16
Minmax inequality, 18
minor, 10
momentum based learning, 16

non-homogeneous, 4
norm, 7
null space, 6

orthogonal, 8
orthogonal set, 8

partial derivatives, 13
pivot columns, 4
pivots, 4
Positive definite, 4
positive definite, 7
Positive semi-definite, 4
primal problem, 18
principal component, 17
Principal Component Analysis, 16
Product rule, 12, 13
projection, 8
projection matrix, 17
Properties of matrices, 3

QR decomposition, 12
Quotient rule, 12

range space, 7
rank, 4, 10
Rank Nullity Theorem, 6
Rank-Nullity theorem, 7
Reduced Row Echelon Form (RREF), 4
RMS Prop Method, 16
Row Echelon Form (REF), 4
Row space, 6

Scalar multiplication, 3
Singular Value Decomposition, 11
size of a matrix, 3
Skew-symmetric, 4
Slater's condition, 19
Soft Margin SVM, 20
Spectral theorem, 11
standard deviation, 18
standardize, 18
stochastic gradient descent, 16
strong duality, 19
subspace, 6
Sum rule, 12, 13
Support vectors, 19
Symmetric, 4
symmetry, 7
system of linear equations, 4

Taylor polynomial, 12
Taylor series, 12
Taylor's Theorem, 13
tensor, 14
trace, 11
Transpose, 3
Triangle inequality, 7

Upper triangular, 4

zero matrix, 3