If you want to edit or correct this summary, you can find the source LaTeX code here: https://github.com/alronz/tum-MA4800-fda-summary

- Transpose: $(A^T)_{ji} = A_{ij}$
- Symmetric Matrix: $A^T = A$
- Hermitian Matrix: $A^H = A$
- Matrix Multiplication:
  - $(Ax)_i = \sum_{j \in J} a_{ij} x_j$
  - $(AB)_{il} = \sum_{j \in J} A_{ij} B_{jl}$
- Kronecker Delta: $\delta_{ij}$ is 1 if $i = j$, 0 otherwise.
- Identity Matrix: Diagonal elements are 1, others are 0.

$$\text{range}(A) = \{Ax : x \in \mathbb{K}^J\} = \text{span}\{A_j : j \in J\}.$$

# Mathematical Concepts

## Range of a Matrix

- **Definition:** Span of columns of $A$.
- **Notation:** $\text{range}(A) = \text{span}\{A(j) \mid j \in J\}$.

## Euclidean Scalar Product

- **Scalar Product:** $\langle x, y \rangle = \sum_{i \in I} x_i y_i$.
- **For $K = \mathbb{R}$:** Ignore conjugate.

## Matrix-Vector and Matrix-Matrix Multiplications

- **Matrix-Vector:** $(Ax)_i = \langle A(i), x \rangle$.
- **Matrix-Matrix:** $(AB)_{i\ell} = \langle A(i), B(\ell) \rangle$.

## Orthogonality

- **Orthogonal Vectors:** $\langle x, y \rangle = 0$.
- **Orthogonal Sets:** $\langle x, y \rangle = 0$ for all $x \in X, y \in Y$.
- **Orthogonal Family:** $\langle x_\nu, x_{\nu'} \rangle = 0$ for $\nu \neq \nu'$.
- **Orthonormal Family:** $\langle x_\nu, x_\nu \rangle = 1$.

## Orthogonal and Unitary Matrices

- **Orthogonal Matrix:** $A^H A = I$.
- **Unitary Matrix:** $A^H A = A A^H = I, A^H = A^{-1}$.

## Diagonal Matrix

- **Diagonal Matrix:** $A_{ij} = 0$ for $i \neq j$.

- **Orthogonality of Subspaces:** $X \perp Y$ if every vector in $X$ is orthogonal to every vector in $Y$.
- **Checking with Bases:** $X \perp Y$ if and only if $\langle v_i, w_j \rangle = 0$ for all basis vectors $v_i$ of $X$ and $w_j$ of $Y$.
- **Square Matrix:** A matrix with the same number of rows and columns, $n \times n$.
- **Diagonal Elements:** Elements $a_{ii}$ where the row index equals the column index.
- **Types of Square Matrices:**
  - Identity Matrix: Diagonal elements are 1, off-diagonal elements are 0.
  - Diagonal Matrix: Only diagonal elements are non-zero.

– Symmetric Matrix: $A = A^T$.
– Skew-Symmetric Matrix: $A = -A^T$.

**Matrix Rank $r$:**

1. $r = \dim \text{range}(A)$ (dimension of column space)
2. $r = \dim \text{range}(A^H)$ (dimension of row space)
3. $r$ is the maximal number of linearly independent rows.
4. $r$ is the maximal number of linearly independent columns.
5. $r$ is minimal such that $A = \sum_{i=1}^{r} a_i b_i^H$.
6. $r$ is maximal such that there exists an invertible $r \times r$ submatrix.
7. $r$ is the number of positive singular values.

- **Invertible Submatrix:** A submatrix is invertible if its determinant is non-zero.
- **Rank:** The rank of a matrix is the size of the largest invertible submatrix.
- **Example:** For matrix $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$, the largest invertible submatrix is $2 \times 2$, so the rank is 2.

# Matrix Determinants

## 2x2 Matrix:

$$\det(A) = ad - bc$$

## 3x3 Matrix:

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg)$$

## Larger Matrices:

- **Cofactor Expansion:** Expand along a row or column.
- **Row Reduction:** Reduce to upper triangular form and multiply the diagonal elements.

## Example: Determinant of a 4x4 Matrix Using Row Reduction

Consider a $4 \times 4$ matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

This matrix is already in upper triangular form (all elements below the main diagonal are zero). The determinant is the product of the diagonal elements:

$$\det(A) = 1 \cdot 1 \cdot 1 \cdot 1 = 1$$

## Inner and Outer Products

| Feature | Inner Product | Outer Product |
|---|---|---|
| Definition | $\langle x, y \rangle = x^T y$ | $x \otimes y = xy^T$ |
| Result | Scalar | Matrix |
| Size | Single value | $m \times n$ matrix for $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ |
| Symmetry | Symmetric ($\langle x, y \rangle = \langle y, x \rangle$) | Generally asymmetric ($xy^T \neq yx^T$) |
| Applications | Measuring similarity, orthogonality, projections | Constructing rank-1 matrices, tensor product |
| Example Calculation | $\left\langle \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \right\rangle = 32$ | $\begin{pmatrix} 1 \\ 2 \end{pmatrix} \otimes \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 6 \\ 8 \end{pmatrix}$ |

Outer product is also called the tensor product.

# 1 Maximal Rank (Full-Rank Matrix):

- Definition: $r_{\max} = \min(m, n)$
- Full-Rank: Matrix with rank $r_{\max}$

# 2 Field Independence:

- Real and Complex: Rank of a real-valued matrix is the same over $\mathbb{R}$ and $\mathbb{C}$

# 3 Matrices of Bounded Rank $k$:

- Set $R_k : R_k = \{A \in K^{I \times J} : \text{rank}(A) \leq k\}$
- Non-Vector Space: Addition or scaling can result in higher rank matrices

# 4 Abstract Vector Space $V$:

- $V$ over a field $K$
- Example: Euclidean plane $\mathbb{R}^2$

# 5 Norm $\|\cdot\| : V \to [0, \infty)$:

- Properties:
  1. Non-negativity and Definiteness: $\|v\| = 0$ if and only if $v = 0$
  2. Homogeneity: $\|\lambda v\| = |\lambda| \|v\|$
  3. Triangle Inequality: $\|v + w\| \leq \|v\| + \|w\|$

# 6 Continuity:

- Inverse Triangle Inequality: $|\|v\| - \|w\|| \leq \|v - w\|$

# 7 Normed Vector Space $(V, \|\cdot\|)$:

- A vector space $V$ with a norm.
- Example: $\mathbb{R}^2$ with Euclidean norm.

## 8 Pre-Hilbert Space:

- A normed vector space $(V, \|\cdot\|)$ where the norm is derived from a scalar product.

## 9 Scalar Product Properties:

- **Positivity:** $\langle v, v\rangle > 0$ for $v \neq 0$
- **Symmetry:** $\langle v, w\rangle = \langle w, v\rangle$
- **Linearity in First Argument:** $\langle u + \lambda v, w\rangle = \langle u, w\rangle + \lambda\langle v, w\rangle$
- **Linearity in Second Argument:** $\langle w, u + \lambda v\rangle = \langle w, u\rangle + \lambda\langle w, v\rangle$

## 10 Norm from Scalar Product:

- $\|v\| = \sqrt{\langle v, v\rangle}$

## 11 Schwarz Inequality:

- $|\langle v, w\rangle| \leq \|v\|\|w\|$

## 12 Pre-Hilbert Space:

- Pair $(V, \langle\cdot,\cdot\rangle)$
- Equipped with a scalar product.

## 13 Scalar Product Properties:

- Positivity: $\langle v, v\rangle > 0$ for $v \neq 0$
- Symmetry: $\langle v, w\rangle = \langle w, v\rangle$
- Linearity:

$$\langle \alpha u + \beta v, w\rangle = \alpha\langle u, w\rangle + \beta\langle v, w\rangle$$

## 14 Euclidean Norm:

- Defined by:

$$\|v\|_2 = \sqrt{\sum_{i\in I} |v_i|^2}$$

## 15 Orthogonality:

- Orthogonal Vectors: $\langle v, w\rangle = 0$
- Orthogonal Set: $\langle v_i, v_j\rangle = 0$ for $i \neq j$
- Orthonormal Set: $\|v_i\| = 1$ and $\langle v_i, v_j\rangle = 0$ for $i \neq j$

## 16 Hilbert Space:

- A complete normed vector space with an inner product.

## 17 Pre-Hilbert Space:

- A vector space $(V, \langle \cdot, \cdot \rangle)$ with an inner product.

## 18 Completeness:

- Every Cauchy sequence in $V$ converges to a limit in $V$.

## 19 Cauchy Sequence:

- A sequence $\{v_n\}$ where $\|v_n - v_m\|$ becomes arbitrarily small for sufficiently large $n$ and $m$.

## 20 Finite Dimensional Spaces:

- Always complete, hence always Hilbert spaces.

## 21 Convex Set:

- $C$ is convex if $tv + (1-t)w \in C$ for all $v, w \in C$ and $t \in [0, 1]$.

## 22 Projection $P_C(v)$:

- $P_C(v) = \arg\min_{w \in C} \|v - w\|$
- Unique and well-posed.

## 23 Equivalent Definition:

- $\langle z - P_C(v), v - P_C(v) \rangle \leq 0$ for all $z \in C$.

## 24 Projection onto Subspace $W$:

- If $W$ has an orthonormal basis $\{w_\nu\}_{\nu \in F}$:

$$P_W(v) = \sum_{\nu \in F} \langle v, w_\nu \rangle w_\nu,$$

## 25 Pythagoras-Fourier Theorem:

- For $v \in V$:

$$\|P_W(v)\|^2 = \sum_{\nu \in F} |\langle v, w_\nu \rangle|^2.$$

## 26 Example Calculation:

- Given $v = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ and $W$ spanned by $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$:

$$P_W(v) = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$$

$$\|P_W(v)\|^2 = 5$$

$$\sum_{\nu \in \{1,2\}} |\langle v, w_\nu \rangle|^2 = 5.$$

## 27 Orthonormal Basis:

- $\{w_\nu\}_{\nu \in F}$ such that $\langle w_\nu, w_\mu \rangle = \delta_{\mu\nu}$ and $\|w_\nu\| = 1$.

## 28 Projection Operator $P_V$:

- $P_V(v) = v$ when $W = V$.

## 29 Orthonormal Expansion:

- $v = \sum_{\nu \in F} \langle v, w_\nu \rangle w_\nu$.

## 30 Pythagoras-Fourier Theorem:

- $\|v\|^2 = \sum_{\nu \in F} |\langle v, w_\nu \rangle|^2$.

## 31 Example in $\mathbb{R}^3$:

- Vector $v = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$.
- Orthonormal basis $\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \}$.
- Orthonormal expansion $v = 2 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 4 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.
- Norm equivalence $\|v\|^2 = 29$.

## 32 Trace Definition:

- For $A \in K^{I \times I}$:

$$\mathrm{tr}(A) = \sum_{i \in I} A_{ii}$$

### 33 Example Calculation:

- For $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$:

$$\mathrm{tr}(A) = 1 + 5 + 9 = 15$$

### 34 Properties:

- Linearity: $\mathrm{tr}(\alpha A + \beta B) = \alpha \mathrm{tr}(A) + \beta \mathrm{tr}(B)$
- Cyclic Property: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- Invariance under Similarity: $\mathrm{tr}(A) = \mathrm{tr}(PBP^{-1})$
- Sum of Eigenvalues: $\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i$

## 35 Matrix Decomposition:

- A matrix $A$ of rank $r$ can be decomposed as:

$$A = \sum_{i=1}^{r} a_i b_i^T$$

- Where $a_i \in \mathbb{R}^N$ and $b_i \in \mathbb{R}^M$.

## 36 Storage Requirements:

- Direct Storage: $N \times M$ entries.
- Decomposition Storage: $r(N + M)$ entries.

## 37 Efficiency for Low-Rank Matrices:

- When $r \ll \min(N, M)$, storing the decomposition is much more efficient than storing the entire matrix.

## Example:

Consider a matrix $A$ of size $1000 \times 500$ with rank $r = 10$:

- Direct Storage:
  - $1000 \times 500 = 500,000$ entries.
- Decomposition Storage:
  - $r(N + M) = 10(1000 + 500) = 15,000$ entries.

## Frobenius Norm $\|A\|_F$

- **Definition:**

$$\|A\|_F = \sqrt{\sum_{i \in I} \sum_{j \in J} |A_{ij}|^2}$$

- **Alternative Names:**
  - Schur norm
  - Hilbert-Schmidt norm
- **Scalar Product:**

$$\langle A, B \rangle_F = \sum_{i \in I} \sum_{j \in J} A_{ij} B_{ij}$$

$$\langle A, B \rangle_F = \operatorname{tr}(AB^H) = \operatorname{tr}(B^H A)$$

- **Property:**

$$\|A\|_F^2 = \operatorname{tr}(AA^H) = \operatorname{tr}(A^H A)$$

# Matrix Norm

## Definition:

$$\|A\| = \|A\|_{X\to Y} = \sup_{z\neq 0} \frac{\|Az\|_Y}{\|z\|_X}$$

$\|\cdot\|_X$ and $\|\cdot\|_Y$ are vector norms on spaces $X = \mathcal{K}^I$ and $Y = \mathcal{K}^J$.

## Spectral Norm ($\|A\|_2$ or $\|A\|$)

**Definition (for Euclidean norms):**

$$\|A\|_2 = \sup_{z\neq 0} \frac{\|Az\|_2}{\|z\|_2}$$

**Key Points:**

- The spectral norm is derived from Euclidean vector norms.
- It is crucial in many applications and often denoted simply by $\|A\|$.

## Summary of Terms

- Supremum (sup): The least upper bound of a set.
- Euclidean Norm ($\|\cdot\|_2$): The standard norm in Euclidean space, also known as the $L^2$ norm.

# Unitary Invariance

## Unitary Matrix:

A matrix $U$ is unitary if $U^H U = I$, where $U^H$ is the conjugate transpose of $U$.

## Frobenius Norm:

$$\|A\|_F = \|UAV^H\|_F$$

## Spectral Norm:

$$\|A\|_2 = \|UAV^H\|_2$$

# Submultiplicativity

## Frobenius Norm:

$$\|AB\|_F \leq \|A\|_F\|B\|_F \leq \|A\|_F\|B\|_F$$

## Spectral Norm:

$$\|AB\| \leq \|A\|\|B\|$$

## Components of SVD

- **Orthogonal Matrices $U$ and $V$**:
  - $U^H U = I$ and $V^H V = I$
  - Columns of $U$ and $V$ are orthonormal vectors.
- **Diagonal Matrix $\Sigma$**:
  - Contains singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$ (where $r$ is the rank of $A$) on the diagonal.
  - Singular values are non-negative and sorted in descending order.

## Properties and Uses

- **Orthogonality**: $U$ and $V$ preserve Euclidean norm.
- **Data Compression**: Truncate small singular values for matrix approximation.
- **Principal Component Analysis (PCA)**: Used to find principal components that capture the most variance in data.

## First Singular Vector $\mathbf{v_1}$

- **Definition:**

$$\mathbf{v_1} = \arg \max_{\|\mathbf{v}\|_1 = 1} \|A\mathbf{v}\|_2$$

  - $\mathbf{v_1}$ is the column vector representing the best-fit line through the origin.
  - The rows of $A$ are points in $d$-dimensional space.

## First Singular Value $\sigma_1(A)$

- **Definition:**

$$\sigma_1(A) = \|A\mathbf{v_1}\|_2$$

- **Interpretation:**
  - $\sigma_1(A)^2$ is the sum of the squares of the projections of the points onto the line determined by $\mathbf{v_1}$.

## Key Points

- **Best-Fit Line via SVD:**
  - The first singular vector $\mathbf{v_1}$ provides the direction of the best-fit line.
  - The first singular value $\sigma_1(A)$ measures how well the data fits along this direction.
- **Variance Maximization:**
  - $\mathbf{v_1}$ maximizes the variance of the projected data.
  - $\sigma_1(A)$ quantifies this maximum variance.

# Singular Value Decomposition (SVD) - Extended

### First Singular Vector $v_1$:

$$v_1 = \arg \max_{\|x\|_2=1} \|Av\|_2$$

### First Singular Value $\sigma_1(A)$:

$$\sigma_1(A) = \|Av_1\|_2$$

### Second Singular Vector $v_2$:

**Definition:**

$$v_2 = \arg \max_{\|x\|_2=1,\ \langle v_1, v_2 \rangle = 0} \|Av\|_2$$

### Second Singular Value $\sigma_2(A)$:

$$\sigma_2(A) = \|Av_2\|_2$$

### General $k$-th Singular Vector $v_k$:

**Definition:**

$$v_k = \arg \max_{\|x\|_2=1,\ \langle v_1, v_2, \ldots, v_{k-1} \rangle = 0} \|Av\|_2$$

### General $k$-th Singular Value $\sigma_k(A)$:

$$\sigma_k(A) = \|Av_k\|_2$$

### Stopping Criterion

### Termination Condition:

$$0 = \max_{\|v\|_2=1,\ \langle v_1, v_2, \ldots, v_k \rangle = 0} \|Av\|_2$$

### Rank $r$:

The process stops after finding $r$ singular vectors and singular values, where $r$ is the rank of $A$.

# Singular Value Decomposition (SVD) and Frobenius Norm

label=• **Row Space Spanning:**

- Singular vectors $v_1, v_2, \ldots, v_r$ span the row space of $A$.

lbbel=• **Orthogonality:**

- $\langle A^{(i)}, v \rangle = 0$ for any $v$ orthogonal to $v_1, v_2, \ldots, v_r$.

lcbel=• **Pythagoras' Theorem for Rows:**

$$\|A^{(i)}\|_2^2 = \sum_{k=1}^{r} |\langle A^{(i)}, v_k \rangle|^2$$

ldbel=• **Summing Over All Rows:**

$$\sum_{i\in I} \|A^{(i)}\|_2^2 = \sum_{k=1}^{r} \|A_v k\|_2^2 = \sum_{k=1}^{r} \sigma_k(A)^2$$

lebel=• **Connection to Frobenius Norm:**

$$\|A\|_F^2 = \sum_{i\in I}\sum_{j\in J} |A_{ij}|^2 = \sum_{k=1}^{r} \sigma_k(A)^2$$

lfbel=• **Final Result:**

$$\|A\|_F = \sqrt{\sum_{k=1}^{r} \sigma_k(A)^2}$$

# Spectral Norm

– **Definition:**

$$\|A\| = \sigma_1(A) = \max_{\|v\|_2=1} \|Av\|_2$$

# Schatten-$p$ Norms

– **Definition (for $1 \le p < \infty$):**

$$\|A\|_p = \left(\sum_{k=1}^{r} \sigma_k(A)^p\right)^{1/p}$$

– **Special Cases:**

∗ Frobenius Norm ($p = 2$):

$$\|A\|_F = \left(\sum_{k=1}^{r} \sigma_k(A)^2\right)^{1/2}$$

∗ Spectral Norm ($p = \infty$):

$$\|A\|_\infty = \max_{k=1,\dots,r} \sigma_k(A) = \sigma_1(A)$$

∗ Nuclear Norm ($p = 1$):

$$\|A\|_* = \|A\|_1 = \sum_{k=1}^{r} \sigma_k(A)$$

## Symmetric and Positive (Semi-)Definite Matrices

1. **Symmetric Matrix:**
   – A matrix $A \in \mathbb{R}^{m\times m}$ is symmetric if $A = A^T$.
2. **Inner Product:**
   – For $x \in \mathbb{R}^m$,
   $$\langle x, Ax\rangle = x^T(Ax).$$
3. **Positive Semi-Definite Matrix:**
   – A matrix $A$ is positive semi-definite if:
   $$\langle x, Ax\rangle \ge 0 \quad \forall x \in \mathbb{R}^m.$$

4. **Positive Definite Matrix:**
   - A matrix $A$ is positive definite if:

$$\langle x, Ax \rangle > 0 \quad \forall x \in \mathbb{R}^m,\, x \neq 0.$$

**Key Points**

- **Symmetric Matrices:** Have the property $A = A^T$.
- **Positive Semi-Definite:** The inner product $\langle x, Ax \rangle$ is non-negative for all $x$.
- **Positive Definite:** The inner product $\langle x, Ax \rangle$ is positive for all non-zero $x$.

# Infimum and Supremum

- **Infimum (Inf):**
  - Definition: Greatest lower bound of a set.
  - Not necessarily in the set.
  - Example: $\inf\{\frac{1}{x} : x \in (0, 1)\} = 1$.
- **Supremum (Sup):**
  - Definition: Least upper bound of a set.
  - Not necessarily in the set.
  - Example: $\sup\{-x^2 : x \in [-1, 2]\} = 0$.

# Positive Definite Matrices

- **Full Rank:** Always full rank $\text{rank}(A) = n$.
- **Eigenvalues:** All positive $\lambda_i > 0$.
- **Invertibility:** Always invertible (no zero eigenvalues).

# Positive Semi-Definite Matrices

- **Possibly Not Full Rank:** May not be full rank $\text{rank}(A) \leq n$.
- **Eigenvalues:** All non-negative $\lambda_i \geq 0$.
- **Singularity:** May be singular (zero eigenvalues possible).
- **Rank and Eigenvalues:** Rank equals the number of non-zero eigenvalues.

# Key Differences

- **Rank:**
  - Positive Definite: Full rank.
  - Positive Semi-Definite: Can be less than full rank.
- **Invertibility:**
  - Positive Definite: Always invertible.
  - Positive Semi-Definite: May be singular.

## Pseudo-Inverse Calculation:

1. SVD of A:
$$A = U\Sigma V^H$$
2. Diagonal Matrix $\Sigma^{-1}$:
   - Invert nonzero singular values $\sigma_k$ to get $\sigma_k^{-1}$.
   - Maintain zeros for zero singular values.
3. Construct $A^+$:
$$A^+ = V\Sigma^{-1}U^H$$

**Formula:**
$$A^+ = \sum_{k=1}^{r} \sigma_k^{-1} v_k u_k^H$$

## Pseudo-Inverse Matrices:

1. **Condition:** $A^H A \in \mathbb{K}^{J\times J}$ is invertible.
   - Implies $A$ has full column rank: $\text{rank}(A) = J$.
2. **Pseudo-Inverse Formula:**
$$A^+ = (A^H A)^{-1} A^H$$
3. **Verification Using SVD:**
   - SVD of $A$ : $A = U\Sigma V^H$
   - Compute $A^H A$ : $A^H A = V\Sigma^2 V^H$
   - Invert $A^H A$ : $(A^H A)^{-1} = V\Sigma^{-2} V^H$

- Multiply by $A^H$ : $(A^H A)^{-1} A^H = V\Sigma^{-1}U^H = A^+$

4. **Fundamental Properties:**
   - Left Inverse: $I = (A^H A)^{-1} A^H A = A^+ A$
   - Orthogonal Projection: $P_{\text{range}(A)} = AA^+$

## Pseudo-Inverse Matrices (Right Inverse Condition):

1. Condition: $AA^H \in \mathbb{K}^{I\times I}$ is invertible.
   - Implies $I \leq J$ (more columns than rows).
   - Rank $r = I$, and rows of $A$ are linearly independent.
2. Pseudo-Inverse Formula:
$$A^+ = A^H (AA^H)^{-1}$$
3. Verification Using SVD:
   - SVD of $A$: $A = U\Sigma V^H$
   - Compute $AA^H$: $AA^H = U\Sigma\Sigma^H U^H$
   - Invert $AA^H$: $(AA^H)^{-1} = U(\Sigma\Sigma^H)^{-1}U^H$
   - Multiply by $A^H$: $A^H(AA^H)^{-1} = V\Sigma^{-1}U^H = A^+$
4. Fundamental Properties:
   - Right Inverse: $I = A^H(AA^H)^{-1}A = AA^+$
   - Orthogonal Projection: $P_{\text{range}(A^H)} = A^+A$

## Least Squares Problems:

### 1. Overdetermined Systems ($m > n$):

- **Problem:** More equations than unknowns; no exact solution.
- **Objective:** Minimize the mismatch $\|Ax - y\|_2$.
- **Solution:**
  - Normal Equations: $A^T Ax = A^T y$
  - Least Squares Solution: $x = (A^T A)^{-1} A^T y$
  - Using Pseudo-Inverse: $\hat{x} = A^+ y$

### 2. Underdetermined Systems ($m < n$):

- **Problem:** Fewer equations than unknowns; infinite solutions.
- **Objective:** Find the solution with minimal Euclidean norm $\|x\|_2$.
- **Solution:**
  - Minimal Norm Solution: $x = A^+ y = A^H (AA^H)^{-1} y$

### Key Concepts:

- **Overdetermined:** Approximate solution minimizing $\|Ax - y\|_2$.
- **Underdetermined:** Minimal norm solution $\|x\|_2$.
- **Pseudo-Inverse:** A common tool for both cases, providing a systematic way to solve these problems.

# 38 Weyl's Bounds

## 38.1 Context

Perturbation of eigenvalues of Hermitian matrices.

## 38.2 Theorem

$$\lambda_k(A) + \lambda_n(E) \leq \lambda_k(A + E) \leq \lambda_k(A) + \lambda_1(E)$$

- Applies to Hermitian matrices $A$ and $E$.
- The $k$-th largest eigenvalue of $A + E$ is bounded by eigenvalues of $A$ and $E$.

## 38.3 Corollary

$$|\sigma_k(A + E) - \sigma_k(A)| \leq \|E\|$$

- Applies to any matrices $A$ and $E$.
- Bounds the change in singular values by the spectral norm of $E$.

# 39 Mirsky's Bounds

## 39.1 Context

Perturbation of singular values of arbitrary matrices.

## 39.2 Theorem

$$\sum_{k=1}^{n} |\sigma_k(A+E) - \sigma_k(A)|^2 \leq \|E\|_F^2$$

- Applies to any matrices $A$ and $E$.
- Bounds the sum of squared differences in singular values by the Frobenius norm of $E$.

## 39.3 Generalization

- Holds for any unitarily invariant norm.
- Includes Weyl's theorem as a special case.

# Probability Density Function (PDF)

- **Definition:** A random variable $X$ has a PDF $\varphi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ if

$$P(a < X \leq b) = \int_a^b \varphi(t)\, dt \quad \text{for all } a < b \in \mathbb{R}$$

- **Relationship with Distribution Function:**

$$\varphi(t) = \frac{d}{dt} F(t)$$

# Expectation (Mean)

- **Definition:**

$$E[X] = \int_\Omega X(\omega)\, dP(\omega)$$

- **With PDF:**
  - For a function $g : \mathbb{R} \to \mathbb{R}$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(t)\varphi(t)\, dt$$

- **Special Case (mean of $X$):**

$$E[X] = \int_{-\infty}^{\infty} t\varphi(t)\, dt$$

# Moments and Absolute Moments

- Moments:

$$E[X^p] \quad \text{for } p > 0$$

- Absolute Moments:

$$E[|X|^p] \quad \text{for } p > 0$$

# Variance

- Definition:
$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

# Lp Norm

- Definition:
$$\|X\|_p = (E[|X|^p])^{1/p} \quad \text{for } 1 \leq p < \infty$$

# Triangle Inequality

- Inequality:
$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$$
- Holds for all $p$-integrable random variables $X, Y$ on $(\Omega, \Sigma, \mathcal{P})$.

# Hölder's Inequality

- Inequality:
$$|E[XY]| \leq \|X\|_p \|Y\|_q$$
- For $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$.

# Convergence of Random Variables

## Definition

A sequence of random variables $\{X_n\}$ converges to $X$ if:
$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \text{for all } \omega \in \Omega$$

## Lebesgue's Dominated Convergence Theorem (LDCT)

### Statement

If there exists a random variable $Y$ such that $E[|Y|] < \infty$ and $|X_n| \leq |Y|$ a.s., then:
$$\lim_{n\to\infty} E[X_n] = E\left[\lim_{n\to\infty} X_n\right] = E[X]$$

### Conditions

- $X_n \to X$ a.s.
- There exists an integrable random variable $Y$ such that $|X_n| \leq |Y|$ a.s.

# Formulation for Integrals of Sequences of Functions

## Statement

If $\{f_n\}$ is a sequence of measurable functions converging pointwise to a function $f$, and there exists an integrable function $g$ such that $|f_n| \leq g$ almost everywhere, then:
$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu$$

# Fubini's Theorem

## Setup

Let $f : A \times B \to \mathbb{C}$ be a measurable function.

- $(A, \nu)$ and $(B, \mu)$ are measurable spaces with measures $\nu$ and $\mu$.

## Condition

The integral of the absolute value of $f$ over $A \times B$ is finite:

$$\int_{A\times B} |f(x,y)| d(\nu \times \mu)(x,y) < \infty$$

## Conclusion

The iterated integrals are equal:

$$\int_A \left( \int_B f(x,y) d\mu(y) \right) d\nu(x) = \int_B \left( \int_A f(x,y) d\nu(x) \right) d\mu(y)$$

# Absolute Moments

- **Definition:** The absolute $p$-th moment of a random variable $X$ is:

$$E[|X|^p] \quad \text{for } p > 0$$

- **Proposition:**

$$E[|X|^p] = p \int_0^\infty P(|X| \geq t) t^{p-1} dt$$

- **Usage:** Use this formula to compute the $p$-th absolute moment by integrating the tail probabilities weighted by $t^{p-1}$.

# Cavalieri's Formula for Expectation

- **Corollary:**

$$E[X] = \int_0^\infty P(X \geq t) dt - \int_0^\infty P(X \leq -t) dt$$

- **Usage:** Use this formula to compute the expectation by integrating the tail probabilities of $X$ being above $t$ and below $-t$.

# Tail of a Random Variable

- **Definition:** The tail function of a random variable $X$ is:

$$t \mapsto P(|X| \geq t)$$

- Describes the probability that the absolute value of $X$ exceeds $t$.

# Markov's Inequality

- **Statement:** For any non-negative random variable $X$ and any $t > 0$,

$$P(|X| \geq t) \leq \frac{E[|X|]}{t}$$

- Provides an upper bound on the tail probability using the expectation of $|X|$.
- **Usefulness:**
  - Simple and general tool for bounding probabilities.
  - Useful for providing rough estimates when only the mean of the distribution is known.

# Inequalities and Transformations in Probability

## Generalized Markov Inequality

- **Statement:** For any random variable $X$ and $p > 0$,

$$P(|X| \geq t) = P(|X|^p \geq t^p) \leq \frac{E[|X|^p]}{t^p} \quad \text{for all } t > 0$$

- **Usefulness:** Provides a bound on the tail probability using the $p$-th moment of $X$.

## Chebyshev Inequality

- **Statement:** For any random variable $X$ with finite variance,

$$P(|X - E[X]| \geq t) \leq \frac{Var(X)}{t^2} \quad \text{for all } t > 0$$

- **Usefulness:** Provides a bound on the probability that $X$ deviates from its mean by at least $t$ standard deviations.

## Exponential Markov Inequality

- **Statement:** For any random variable $X$ and $\theta > 0$,

$$P(X \geq t) = P(e^{\theta X} \geq e^{\theta t}) \leq e^{-\theta t} E[e^{\theta X}] \quad \text{for all } t \in \mathbb{R}$$

- **Usefulness:** Provides a bound on the tail probability using the moment generating function (Laplace transform) of $X$.

## Laplace Transform (Moment Generating Function)

- **Definition:** The Laplace transform or moment generating function of a random variable $X$ is:

$$M_X(\theta) = E[e^{\theta X}]$$

- **Properties:** Helps in deriving bounds for tail probabilities and is useful in studying the distribution of $X$.

## Normal Distribution (Gaussian Distribution)

- **PDF:**
$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$
- $\mu$: Mean
- $\sigma^2$: Variance
- **Properties:**
  - Mean: $E[X] = \mu$
  - Variance: $\mathrm{Var}(X) = \sigma^2$
- **Notation:**
$$X \sim N(\mu, \sigma^2)$$

## Standard Normal Distribution (Standard Gaussian)

- **PDF:**
$$\phi(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$
- **Properties:**
  - Mean: $E[g] = 0$
  - Variance: $\mathrm{Var}(g) = 1$
- **Notation:**
$$g \sim N(0,1) \quad \text{or} \quad Z \sim N(0,1)$$

## Definition of Independence

- **Independence:** Random variables $X_1, \ldots, X_n$ are independent if:
$$P(X_1 \leq t_1, \ldots, X_n \leq t_n) = \prod_{\ell=1}^{n} P(X_\ell \leq t_\ell)$$

## Expectation of Product

- **Property:** For independent random variables $X_1, \ldots, X_n$,
$$E\left[\prod_{\ell=1}^{n} X_\ell\right] = \prod_{\ell=1}^{n} E[X_\ell]$$

## Joint Probability Density Function

- **Factorization:** If $X_1, \ldots, X_n$ have a joint PDF $\varphi$, then:
$$\varphi(t_1, \ldots, t_n) = \varphi_1(t_1) \cdot \varphi_2(t_2) \cdots \varphi_n(t_n)$$
- $\varphi_\ell(t_\ell)$ are the individual PDFs of $X_\ell$.

## Sum of Independent Random Variables

- **Definition:** If $X$ and $Y$ are independent random variables with PDFs $\varphi_X$ and $\varphi_Y$, then the PDF of $X + Y$ is given by the convolution of $\varphi_X$ and $\varphi_Y$.

## Convolution

- **Formula:**
$$\varphi_{X+Y}(t) = (\varphi_X * \varphi_Y)(t) = \int_{-\infty}^{\infty} \varphi_X(u)\varphi_Y(t-u)\,du$$
- $\varphi_X(u)$: PDF of $X$
- $\varphi_Y(t-u)$: PDF of $Y$ shifted by $t$.

## Fubini's Theorem for Expectations

- **Setup:**
  - $X, Y \in \mathbb{R}^n$ are independent random vectors (or variables).
  - $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a measurable function with:
$$E[|f(X,Y)|] < \infty$$
- **Functions:**
  - $f_1(x) = E[f(x,Y)]$
  - $f_2(y) = E[f(X,y)]$
- **Measurability and Integrability:**
  - $f_1$ and $f_2$ are measurable.
  - $E[|f_1(X)|] < \infty$
  - $E[|f_2(Y)|] < \infty$
- **Theorem:**
$$E[|f_1(X)|] = E[|f_2(Y)|] = E[|f(X,Y)|]$$
- **Conditional Expectation:**
  - $f_1(X) = E_Y[f(X,Y)]$

## Standard Gaussian Vector

- **Definition:** A vector $g \in \mathbb{R}^n$ with independent standard normal components:
$$g_i \sim N(0,1) \quad \text{independently for } i = 1, \ldots, n$$

## Multivariate Normal Distribution (Gaussian Vector)

- **Definition:** A vector $X \in \mathbb{R}^n$ is Gaussian if:

$$X = Ag + \mu$$

  where

  * $g \in \mathbb{R}^k$ is a standard Gaussian vector.
  * $A \in \mathbb{R}^{n \times k}$ is a matrix.
  * $\mu \in \mathbb{R}^n$ is the mean vector.

- **Covariance Matrix:**

$$\Sigma = AA^T = E[(X - \mu)(X - \mu)^T]$$

## Probability Density Function (PDF)

- **Non-degenerate Case ($\Sigma$ is positive definite):**

$$\psi(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

- **Degenerate Case ($\Sigma$ is not invertible):**

  * $X$ does not have a density.

## Rotation Invariance

- **Property:** If $U$ is an orthogonal matrix, $U^T g$ (where $g$ is a standard Gaussian vector) has the same distribution as $g$.

## Jensen's Inequality

- **For Convex Functions:**

$$f(E[X]) \leq E[f(X)]$$

- **Convex Function:** $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$.
- **For Concave Functions:**

$$E[f(X)] \leq f(E[X])$$

- **Concave Function:** $f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$ for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$.

## 1. Moment Generating Function (MGF) $M_X(\theta)$:

$$M_X(\theta) = E[\exp(\theta X)]$$

- Use this function to understand the distribution of the random variable $X$.

## 2. Cumulant Generating Function (CGF) $K_X(\theta)$:

$$K_X(\theta) = \ln(E[\exp(\theta X)])$$

- The CGF is the logarithm of the MGF.

## 3. Cumulant Generating Function (CGF) $C_X(\theta)$:

$$C_X(\theta) = \ln(E[\exp(\theta X)])$$

## 2. Probability Bound:

– For independent random variables $X_1, X_2, \ldots, X_M$ and a threshold $t > 0$:

$$P\left(\sum_{l=1}^{M} X_\ell \geq t\right) \leq \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{l=1}^{M} C_X(\theta)\right\}\right)$$

## Key Definitions and Theorem

1. Conditions:
   - Independent random variables $X_1, X_2, \ldots, X_M$.
   - Each $X_\ell$ has $\mathbb{E}[X_\ell] = 0$.
   - Each $X_\ell$ is bounded by $B_\ell$ (i.e., $|X_\ell| \leq B_\ell$ almost surely).
2. One-Sided Bound:
$$\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{\ell=1}^{M} B_\ell^2}\right)$$
3. Two-Sided Bound:
$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{\ell=1}^{M} B_\ell^2}\right)$$

Cheat Sheet for Bernstein's Inequality

## Key Definitions and Theorem

### 1. Conditions:

– Independent random variables $X_1, X_2, \ldots, X_M$.

– Each $X_\ell$ has $\mathbb{E}[X_\ell] = 0$.

– For all integers $n \geq 2$:

$$\mathbb{E}[|X_\ell|^n] \leq \frac{n!}{2} R^{n-2} \sigma_\ell^2$$

### 2. Probability Bound:

– For all $t > 0$:

$$P\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + Rt)}\right)$$

– Where $\sigma^2 = \sum_{t=1}^{M} \sigma_t^2$.

# Cheat Sheet for Bernstein's Inequality for Bounded Random Variables

## Key Definitions and Corollary

label=0. Conditions:

- Independent random variables $X_1, X_2, \ldots, X_M$.
- Each $X_k$ has $E[X_k] = 0$.
- Each $X_k$ is almost surely bounded by $K$ (i.e., $|X_k| < K$).
- The second moment is bounded by $\sigma_k^2$ (i.e., $E[|X_k|^2] \leq \sigma_k^2$).

lbbel=0. Probability Bound:

- For all $t > 0$:
$$P\left(\left|\sum_{k=1}^{M} X_k\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + \frac{1}{3}Kt)}\right)$$
- Where $\sigma^2 = \sum_{k=1}^{M} \sigma_k^2$.

# Cheat Sheet for Bernstein's Inequality for Subexponential Random Variables

## Key Definitions and Corollary

label=0. Conditions:

- Independent mean zero subexponential random variables $X_1, X_2, \ldots, X_M$.
- Each $X_k$ satisfies $P\left(|X_k| \geq t\right) \leq \beta e^{-\kappa t}$ for all $t > 0$.

lbbel=0. Probability Bound:

- For all $t > 0$:
$$P\left(\left|\sum_{k=1}^{M} X_k\right| \geq t\right) \leq 2\exp\left(-\frac{(\kappa t)^2}{2(2\beta M + \kappa t)}\right)$$

Cheat Sheet for Johnson-Lindenstrauss Lemma

# Key Definitions and Theorem

## 1. Parameters:

- $\epsilon$: A small positive constant ($0 < \epsilon < 1$).
- $k$: Target lower dimension, satisfying:
$$k \geq \frac{2\beta}{\epsilon^2/2 - \epsilon^3/3} \ln n$$
for some $\beta \geq 2$.

## 2. Mapping $f : \mathbb{R}^d \to \mathbb{R}^k$:

- For any set $P$ of $n$ points in $\mathbb{R}^d$:
$$(1-\epsilon)\|v-w\|_2^2 \leq \|f(v)-f(w)\|_2^2 \leq (1+\epsilon)\|v-w\|_2^2 \quad \text{for all } v, w \in P$$

### 3. Probability:

- The map $f$ can be generated at random with high probability:

$$1 - \left(n^{2-\beta} - n^{1-\beta}\right)$$

- For $\beta$ slightly larger than 2, the probability is very close to 1 for large $n$.

## Cheat Sheet for Johnson-Lindenstrauss Lemma Using Gaussian Matrices

### Key Definitions and Theorem

1. **Parameters**:
   - $\epsilon$: A small positive constant ($0 < \epsilon < \frac{1}{2}$).
   - $k$: Target lower dimension, satisfying:

$$k \geq \beta \epsilon^{-2} \ln n$$

     for some constant $\beta$.
2. **Mapping** $f : \mathbb{R}^d \to \mathbb{R}^k$:
   - For any set $P$ of $n$ points in $\mathbb{R}^d$:

$$(1-\epsilon)\|v-w\|_2^2 \leq \|f(v)-f(w)\|_2^2 \leq (1+\epsilon)\|v-w\|_2^2 \quad \text{for all } v, w \in P$$

3. **Probability**:
   - The map $f$ can be generated at random with high probability:

$$1 - \left(n^{2-\beta(1-\epsilon)} - n^{1-\beta(1-\epsilon)}\right)$$

   - For $\beta$ sufficiently large, the probability is very close to 1 for large $n$.

### PARALLELOGRAM LAW:

$$\|x+y\|_2^2 = 2\|x\|_2^2 + 2\|y\|_2^2 - \|x-y\|_2^2$$

$$(x-y)^2 = 2x^2 + 2y^2 - (x+y)^2$$

$$P_W(v) := \arg\min_{w \in W} \|v-w\| = \sum_{i=1}^{n} \langle v, w_i \rangle w_i$$

$$A^T A v_k = \sigma_k^2 v_k \tag{1}$$

## Convex Sets:

- **Definition:** A set $K \subseteq \mathbb{R}^N$ is convex if for all $x, z \in K$ and $t \in [0,1]$ : $\quad t \cdot x + (1-t) \cdot z \in K$
- **Alternative Definition (Convex Combination):** A set $K \subseteq \mathbb{R}^N$ is convex if for any $x_1, \ldots, x_n \in K$ and $t_1, \ldots, t_n \geq 0$ with $\sum_{j=1}^n t_j = 1$ : $\quad \sum_{j=1}^n t_j x_j \in K$

## Convex Hull:

– **Definition:** For $T \subseteq \mathbb{R}^N$, the convex hull $\text{conv}(T)$ is the smallest convex set containing $T$.

– **Characterization:** The convex hull consists of all convex combinations of points in $T$:

$$\text{conv}(T) = \left\{ \sum_i t_j x_j : t_j \geq 0, \sum_j t_j = 1, x_j \in T \right\}$$

## Examples of Convex Sets:

– Subspaces

– Affine spaces

– Half spaces

– Convex polygons

– Norm balls (e.g., Euclidean balls)

## Intersection Property:

– The intersection of convex sets is convex.

## Cones:

– **Definition:** A set $K \subseteq \mathbb{R}^N$ is a cone if for all $x \in K$ and all $t \geq 0$, $t \cdot x \in K$.

## Convex Cones:

– **Definition:** A set $K$ is a convex cone if it is a cone and convex.

– **Characterization:** For all $x, z \in K$ and $s, t \geq 0$, $s \cdot x + t \cdot z \in K$.

## Dual Cones:

– **Definition:** For a cone $K \subseteq \mathbb{R}^N$, the dual cone $K^*$ is:

$$K^* = \{z \in \mathbb{R}^N : \langle x, z \rangle \geq 0 \text{ for all } x \in K\}$$

## Properties:

– **Closed and Convex:** $K^*$ is always closed and convex.

– **Conic Property:** $K^*$ is itself a cone.

– **Self-Dual:** A cone is self-dual if $K = K^*$.

## Bidual Cone:

– If $K$ is a closed cone, then $(K^*)^* = K$.

## Inclusion Relationship:

- If $H \subseteq K$, then $K^* \subseteq H^*$.

# Polar Cones

### Definition

For a cone $K \subseteq \mathbb{R}^N$, the polar cone $K^\circ$ is:

$$K^\circ = \{z \in \mathbb{R}^N : \langle x, z \rangle \leq 0 \text{ for all } x \in K\}$$

### Relation to Dual Cone

$K^\circ = -K^*$, where $K^*$ is the dual cone.

# Conic Hull

### Definition

For a set $T \subseteq \mathbb{R}^N$, the conic hull cone$(T)$ is the smallest convex cone containing $T$.

### Characterization

$$\text{cone}(T) = \left\{ \sum_j t_j x_j : t_j \geq 0, x_j \in T \right\}$$

# Geometrical Hahn-Banach Theorem

### Statement

For convex sets $K_1, K_2 \subseteq \mathbb{R}^N$ with empty interior intersection, there exists a vector $w \in \mathbb{R}^N$ and a scalar $\lambda$ such that:

$$K_1 \subseteq \{x \in \mathbb{R}^N : \langle x, w \rangle \leq \lambda\}$$
$$K_2 \subseteq \{x \in \mathbb{R}^N : \langle x, w \rangle \geq \lambda\}$$

# Key Points

### Hyperplane Separation

The theorem guarantees that two non-intersecting convex sets can be separated by a hyperplane.

### Hyperplane Definition

A hyperplane is defined by a normal vector $w \in \mathbb{R}^N$ and a scalar $\lambda$.

### Convex Set Position

Each convex set lies entirely on one side of the hyperplane.

# 1. Compact Convex Set

## Definition

A set is compact if it is closed and bounded.

## Property

For a compact convex set $\mathcal{K}$, every point in $\mathcal{K}$ can be written as a convex combination of its extreme points.

# 2. Convex Hull

## Definition

The convex hull of a set of points is the smallest convex set that contains all those points.

## Theorem

The theorem states that if you take all the extreme points of a compact convex set $\mathcal{K}$, the convex hull of these extreme points will be $\mathcal{K}$.

# Extended Overview Summary

## Extended-Valued Functions:

- Functions $F : \mathbb{R}^N \to (-\infty, \infty]$.
- Can take real values, $\infty$, or $-\infty$.

## Operations and Inequalities:

$$\begin{aligned} x + \infty &= \infty \quad \text{for all } x \in \mathbb{R} \\ x \cdot \infty &= \infty \quad \text{for } x > 0 \\ x \cdot (-\infty) &= -\infty \quad \text{for } x > 0 \\ x &\leq \infty \quad \text{for all } x \in \mathbb{R} \\ -\infty &\leq x \quad \text{for all } x \in \mathbb{R} \end{aligned}$$

## Domain of Extended-Valued Functions:

- Defined as $\text{dom}(F) = \{x \in \mathbb{R}^N : F(x) \neq \infty\}$.

## Proper Function:

- A function is proper if $\text{dom}(F) \neq \emptyset$.

## Canonical Extension:

- For $F : K \to \mathbb{R}$, extend by $F(x) = \infty$ for $x \notin K$.
- The domain of the extended function is $\text{dom}(F) = K$.

### Key Points:

- Proper functions have non-empty domains.
- Canonical extension ensures the function is defined on all of $\mathbb{R}^N$ by assigning $\infty$ outside the original domain.

## Convex Functions

**Definition:** $F : \mathbb{R}^N \to (-\infty, \infty)$ is convex if for all $x, z \in \mathbb{R}^N$ and $t \in [0, 1]$:

$$F(tx + (1-t)z) \leq tF(x) + (1-t)F(z)$$

### Strictly Convex Functions:

**Definition:** $F$ is strictly convex if for all $x \neq z$ and $t \in (0, 1)$:

$$F(tx + (1-t)z) < tF(x) + (1-t)F(z)$$

### Strongly Convex Functions:

**Definition:** $F$ is strongly convex with parameter $\gamma > 0$ if for all $x, z \in \mathbb{R}^N$ and $t \in [0, 1]$:

$$F(tx + (1-t)z) \leq tF(x) + (1-t)F(z) - \frac{\gamma}{2}(1-t)\|x - z\|_2^2$$

### Concave Functions:

**Definition:** $F$ is concave if $-F$ is convex.

**Strictly Concave:** $F$ is strictly concave if $-F$ is strictly convex.

**Strongly Concave:** $F$ is strongly concave with parameter $\gamma > 0$ if $-F$ is strongly convex with the same parameter.

### Key Points:

- Convex functions have a "bowl-shaped" graph.
- Strict convexity implies a stronger condition where the function value between two points is strictly less than the weighted average.
- Strong convexity introduces a quadratic term, enforcing even stronger curvature.

## Convex Functions

- **Strongly Convex Implies Strictly Convex:**
  - If $F$ is strongly convex, then $F$ is also strictly convex.
- **Convex Domain:**
  - The domain of a convex function $F$ is a convex set.
  - If $x, y$ are in the domain of $F$ and $t \in [0, 1]$, then $tx + (1-t)y$ is also in the domain.
- **Convex Function on a Subset:**
  - For $F : K \to \mathbb{R}$ on a convex subset $K \subseteq \mathbb{R}^N$, $F$ is convex if its canonical extension to $\mathbb{R}^N$ is convex.
  - Canonical extension: $F(x) = \infty$ for $x \notin K$.

# Epigraph

– **Definition:**

  * The epigraph of $F$ is $\text{epi}(F) = \{(x, r) : r \geq F(x)\}$.

– **Convexity and Epigraph:**

  * $F$ is convex if and only if $\text{epi}(F)$ is a convex set.
  * For $(x_1, r_1), (x_2, r_2) \in \text{epi}(F)$ and $t \in [0, 1]$:

$$(tx_1 + (1-t)x_2, tr_1 + (1-t)r_2) \in \text{epi}(F)$$

# Convexity and Gradient:

– **Condition:** A differentiable function $F : \mathbb{R}^N \to \mathbb{R}$ is convex if and only if:

$$F(x) \geq F(y) + \nabla F(y)^T (x - y)$$

– $\nabla F(y) \equiv (\partial_1 F(y), \ldots, \partial_n F(y))^T$ is the gradient at $y$.

# Strong Convexity:

– **Condition:** $F$ is strongly convex with parameter $\gamma > 0$ if and only if:

$$F(x) \geq F(y) + \nabla F(y)^T (x - y) + \frac{\gamma}{2}\|x - y\|^2$$

# Twice Differentiable Functions:

– **Condition:** A twice differentiable function $F$ is convex if and only if its Hessian $\nabla^2 F(x)$ is positive semi-definite:

$$\nabla^2 F(x) \succeq 0$$

– The Hessian matrix $\nabla^2 F(x)$ contains all second-order partial derivatives of $F$.

# Key Points:

– **Gradient Condition:** Convexity can be checked using the gradient. If the tangent line (or hyperplane) at any point $y$ lies below the function, $F$ is convex.

– **Strong Convexity:** Strong convexity includes an additional quadratic term that provides a lower bound on the curvature of $F$.

– **Hessian Condition:** For twice differentiable functions, convexity can be checked by ensuring the Hessian matrix is positive semi-definite at every point.

1. Let $F, G$ be convex functions on $\mathbb{R}^N$. Then, for $\alpha, \beta \geq 0$ the function $\alpha F + \beta G$ is convex.
2. Let $F : \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing, and $G : \mathbb{R}^N \to \mathbb{R}$ be convex. Then $H(x) = F(G(x))$ is convex.

Continuity and Minimization of Convex Functions

## Continuity of Convex Functions:

- **Proposition:** If $F : \mathbb{R}^N \to \mathbb{R}$ is a convex function, then $F$ is continuous on $\mathbb{R}^N$.
- **Implication:** Convex functions do not have discontinuities within their domain.

## Lower Semicontinuity:

- **Definition:** A function $F : \mathbb{R}^N \to (-\infty, \infty]$ is lower semicontinuous if for all $x \in \mathbb{R}^N$ and every sequence $(x_j)_{j\in\mathbb{N}} \subset \mathbb{R}^N$ converging to $x$:

$$\liminf_{j\to\infty} F(x_j) \geq F(x)$$

- **Implication:** Lower semicontinuity ensures the function does not have upward jumps. The function value at $x$ is a lower bound for the limit inferior of the function values at a converging sequence.

## Minimization in Convex Functions:

- **Global Minimum:** A point $x \in \mathbb{R}^N$ is a global minimum of $F$ if:

$$F(x) \leq F(y) \quad \text{for all } y \in \mathbb{R}^N$$

- **Local Minimum:** A point $x \in \mathbb{R}^N$ is a local minimum of $F$ if there exists $\epsilon > 0$ such that:

$$F(x) \leq F(y) \quad \text{for all } y \text{ satisfying } \|x - y\|_2 \leq \epsilon$$

## Proposition:

1. **Local Minimum is Global Minimum:**
   - For a convex function $F$, any local minimum is also a global minimum.
2. **Convex Set of Minima:**
   - The set of minima of a convex function $F$ is convex.
3. **Unique Minimum for Strictly Convex Functions:**
   - If $F$ is strictly convex, the minimum is unique.

## Jointly Convex Functions

### Definition:

A function $f(x, y)$ is jointly convex if it is convex in the combined variable $z = (x, y)$.

### Mathematical Form:

For any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^n \times \mathbb{R}^m$ and $t \in [0, 1]$:

$$f(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \leq tf(x_1, y_1) + (1-t)f(x_2, y_2)$$

### Theorem: Partial Minimization:

**Statement:** Let $f : \mathbb{R}^n \times \mathbb{R}^m \to (-\infty, \infty]$ be a jointly convex function. Then the function

$$g(x) = \inf_{y \in \mathbb{R}^m} f(x, y)$$

is convex.

**Implication:** The convexity of $f$ in both variables ensures that minimizing over one variable retains the convexity in the other variable.

## Theorem:

### Statement:

Let $K \subseteq \mathbb{R}^N$ be a compact convex set, and $F : K \to \mathbb{R}$ be a convex function. Then $F$ attains its maximum at an extreme point of $K$.

## Key Concepts:

1. **Compact Convex Set:**
   - Compact: Closed and bounded.
   - Convex: Any line segment between two points in the set lies entirely within the set.
2. **Extreme Points:**
   - A point $x \in K$ is extreme if it cannot be written as a convex combination of two distinct points in $K$.
   - For extreme point $x$, $x = ty + (1-t)z$ with $t \in (0, 1)$ implies $x = y = z$.
3. **Convex Function:**
   - A function $F$ is convex if:

     $$F(tx + (1-t)y) \leq tF(x) + (1-t)F(y)$$

     for all $x, y \in K$ and $t \in [0, 1]$.
4. **Maximum Attainment:**
   - A convex function $F$ on a compact convex set $K$ attains its maximum value at an extreme point of $K$.

## Convex Conjugate

- **Definition:** For $F : \mathbb{R}^N \to (-\infty, \infty]$, the convex conjugate $F^* : \mathbb{R}^N \to (-\infty, \infty]$ is defined by:

  $$F^*(y) := \sup_{x \in \mathbb{R}^N} \{\langle x, y \rangle - F(x)\}$$

- **Convexity:** $F^*$ is always a convex function, even if $F$ is not.

## Fenchel-Young Inequality

- **Statement:** For all $x, y \in \mathbb{R}^N$:

  $$\langle x, y \rangle \leq F(x) + F^*(y)$$

## Key Points

- The convex conjugate transforms a function $F$ into a new function $F^*$ using a supremum involving a linear term.
- $F^*$ is always convex, providing useful properties for analysis and optimization.
- The Fenchel-Young inequality provides a fundamental bound relating $\langle x, y \rangle$, $F(x)$, and $F^*(y)$.

## Properties of Convex Conjugate Functions:

1. **Lower Semicontinuity:**
   - $F^*$ is lower semicontinuous.
2. **Biconjugate:**
   - $F^{**}$ is the largest lower semicontinuous convex function satisfying $F^{**}(x) \leq F(x)$.
   - If $F$ is convex and lower semicontinuous, then $F = F^{**}$.
3. **Scaling Argument:**
   - For $\tau \neq 0$:
   $$(F_\tau)^*(y) = F^*\left(\frac{y}{\tau}\right)$$
4. **Scaling Function:**
   - For $\tau > 0$:
   $$(\tau F)^*(y) = \tau F^*\left(\frac{y}{\tau}\right)$$
5. **Translation:**
   - For $z \in \mathbb{R}^N$:
   $$(F_z)^*(y) = \langle z, y \rangle + F^*(y)$$

## Subdifferential and Subgradients

- **Definition:** For a convex function $F : \mathbb{R}^N \to \mathbb{R}$ at $x \in \mathbb{R}^N$,
$$\partial F(x) = \{v \in \mathbb{R}^N : F(z) - F(x) \geq \langle v, z - x \rangle \text{ for all } z \in \mathbb{R}^N\}$$
- **Subgradient:** $v$ is a subgradient at $x$ if it satisfies:
$$F(z) - F(x) \geq \langle v, z - x \rangle$$
- **Non-emptiness:** The subdifferential $\partial F(x)$ is always non-empty for a convex function.
- **Differentiable Case:** If $F$ is differentiable at $x$,
$$\partial F(x) = \{\nabla F(x)\}$$

Characterization of Minimizers

# Characterization of Minimizers

- **Theorem:** A vector $x$ is a minimum of a convex function $F$ if and only if:

$$0 \in \partial F(x)$$

- **Minimizer:** A point $x$ such that $F(x) \leq F(y)$ for all $y$ in the domain of $F$.
- **Subdifferential at $x$:** The set of all subgradients at $x$, denoted $\partial F(x)$.
- **Implication:**
  - If $x$ is a minimizer, then $0 \in \partial F(x)$.
  - If $0 \in \partial F(x)$, then $x$ is a minimizer.

# Subdifferential and Conjugation

- **Theorem:** Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be a convex function, and $x, y \in \mathbb{R}^N$. The following are equivalent:

  $y \in \partial F(x)$

  $F(x) + F^*(y) = \langle x, y \rangle$

  If $F$ is lower semicontinuous, $x \in \partial F^*(y)$

- **Convex Conjugate:**

$$F^*(y) = \sup_{x \in \mathbb{R}^N} (\langle x, y \rangle - F(x))$$

- **Lower Semicontinuous:** A function $F$ is lower semicontinuous if:

$$\liminf_{n \to \infty} F(x_n) \geq F(x) \text{ for any sequence } x_n \to x$$

- **Consequence:**
  - For a convex lower semicontinuous function $F$:

$$x \in \partial F^*(y) \longleftrightarrow y \in \partial F(x)$$

# Proximal Mapping

## Definition

For a convex function $F : \mathbb{R}^N \to (-\infty, \infty]$ and a point $z \in \mathbb{R}^N$,

$$\text{prox}_F(z) := \arg\min_{x \in \mathbb{R}^N} \left\{ F(x) + \frac{1}{2}\|x - z\|_2^2 \right\}$$

## Strict Convexity

The function $x \mapsto F(x) + \frac{1}{2}\|x - z\|_2^2$ is strictly convex, ensuring a unique minimizer.

## Special Case: Characteristic Function

**Characteristic Function $\chi_K$:**

$$\chi_K(x) = \begin{cases} 0 & \text{if } x \in K \\ \infty & \text{if } x \notin K \end{cases}$$

**Orthogonal Projection:** For the characteristic function $\chi_K$ of a convex set $K$,

$$P_K(z) := \arg\min_{x \in K} \|x - z\|_2$$

**Subspace:** If $K$ is a subspace of $\mathbb{R}^N$, then $P_K$ is the usual linear orthogonal projection.

### Proximal Mapping

**Proposition:** For a convex function $F : \mathbb{R}^N \to (-\infty, \infty]$,

$$x = P_F(z) \text{ if and only if } z \in x + \partial F(x)$$

**Proximal Mapping as Inverse:**

$$P_F = (I + \partial F)^{-1}$$

### Moreau's Identity

**Theorem (Moreau's Identity):** For a lower semicontinuous convex function $F : \mathbb{R}^N \to (-\infty, \infty]$ and all $z \in \mathbb{R}^N$,

$$P_F(z) + P_{F^*}(z) = z$$

**Implication:** If $P_F$ is easy to compute, then $P_{F^*}(z) = z - P_F(z)$ is also easy to compute.

### Moreau's Identity for Scaled Functions

**Scaling:** For $\tau > 0$,

$$P_{F,\tau}(z) + \tau P_{T^{-1}F^*}\left(\frac{z}{\tau}\right) = z$$

Nonexpansiveness of Proximal Mappings

## Theorem

For a convex function $F : \mathbb{R}^N \to (-\infty, \infty]$, the proximal mapping $P_F$ satisfies:

$$\|P_F(z) - P_F(z')\|_2 \leq \|z - z'\|_2 \quad \text{for all } z, z' \in \mathbb{R}^N.$$

**Implication:** The proximal mapping $P_F$ is a contraction, meaning it does not increase the distance between points in the Euclidean norm.

## Optimization Problem Form

$$\begin{aligned} \min_{x \in \mathbb{R}^N} \quad & F_0(x) \\ \text{s.t.} \quad & Ax = y \\ & F_j(x) \leq b_j, \quad j \in \{1, \ldots, M\} \end{aligned}$$

## Components

– **Objective Function** $F_0(x)$: Function to be minimized.

– **Constraint Functions** $F_j(x)$: Functions defining additional constraints.

– **Equality Constraint:** $Ax = y$, where $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$.

– **Inequality Constraints:** $F_j(x) \leq b_j$ for $j \in \{1, \ldots, M\}$.

## Key Terms

– **Feasible Point:** $x \in \mathbb{R}^N$ that satisfies all constraints.

– **Minimizer/Optimal Point:** Feasible point $x^*$ that minimizes $F_0(x)$.

## Optimal Value

$$F_0(x^*)$$

Reformulating Optimization Problems

– **Original Problem:**

$$\min_{x \in \mathbb{R}^N} F_0(x)$$

subject to

$$Ax = y, \quad F_j(x) \leq b_j, \quad j \in \{1, \ldots, M\}.$$

– **Equivalent Formulation Over Set $\mathcal{K}$:**

$$\min_{x \in \mathcal{K}} F_0(x),$$

where

$$\mathcal{K} = \{x \in \mathbb{R}^N : Ax = y \text{ and } F_j(x) \leq b_j, \forall j \in \{1, \ldots, M\}\}.$$

– **Characteristic Function $\chi_{\mathcal{K}}(x)$:**

$$\chi_{\mathcal{K}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{K} \\ \infty & \text{if } x \notin \mathcal{K} \end{cases}$$

– **Equivalent Unconstrained Problem:**

$$\min_{x \in \mathbb{R}^N} \left(F_0(x) + \chi_{\mathcal{K}}(x)\right).$$

## Lagrange Function

For an optimization problem:

$$\min_{x \in \mathbb{R}^N} F_0(x)$$

subject to

$$Ax = y,$$

$$F_j(x) \leq b_j, \quad j \in \{1, \ldots, M\},$$

the Lagrange function $L(x, \xi, \nu)$ is defined as:

$$L(x, \xi, \nu) = F_0(x) + \xi^T (Ax - y) + \sum_{l=1}^{M} \nu_l (F_l(x) - b_l).$$

– $x \in \mathbb{R}^N$: Decision variables.

– $\xi \in \mathbb{R}^m$: Lagrange multipliers for equality constraints. It can take any real values, positive or negative

– $\nu \in \mathbb{R}^M$: Lagrange multipliers for inequality constraints ($\nu_l \geq 0$).

### Special Case: No Inequality Constraints

$$L(x, \xi) = F_0(x) + \xi^T (Ax - y).$$

## Usage

- Convert constrained problems into unconstrained problems.
- Find critical points to identify potential optimal solutions.

## Lagrange Dual Function

For an optimization problem:

$$\min_{x \in \mathbb{R}^N} F_0(x)$$

subject to

$$Ax = y, \quad F_j(x) \leq b_j, \quad j \in \{1, \ldots, M\},$$

the Lagrange dual function $H(\xi, \nu)$ is defined as:

$$H(\xi, \nu) = \inf_{x \in \mathbb{R}^N} L(x, \xi, \nu),$$

where:

- $\xi \in \mathbb{R}^m$ (equality constraint multipliers).
- $\nu \in \mathbb{R}^M$ (inequality constraint multipliers, with $\nu \geq 0$).

### Lagrange Function

$$L(x, \xi, \nu) = F_0(x) + \xi^T (Ax - y) + \sum_{l=1}^{M} \nu_l (F_l(x) - b_l).$$

### Special Case: No Inequality Constraints

$$H(\xi) = \inf_{x \in \mathbb{R}^N} L(x, \xi) = \inf_{x \in \mathbb{R}^N} \{F_0(x) + \xi^T (Ax - y)\}.$$

### Properties

- **Concavity:** The dual function $H(\xi, \nu)$ is always concave because it is the pointwise infimum of a family of affine functions, regardless of whether the original problem is convex or not.

## Dual Problem

For the primal problem:

$$\min_{x \in \mathbb{R}^N} F_0(x)$$

subject to

$$Ax = y,$$

$$F_j(x) \leq b_j, \quad j \in \{1, \ldots, M\},$$

the dual problem is defined as:

$$\max H(\xi, \nu)$$

subject to

$$\nu \geq 0.$$

### Lagrange Dual Function $H(\xi, \nu)$

$$H(\xi, \nu) = \inf_{x \in \mathbb{R}^N} L(x, \xi, \nu),$$

where

$$L(x, \xi, \nu) = F_0(x) + \xi^T(Ax - y) + \sum_{j=1}^{M} \nu_j (F_j(x) - b_j).$$

### Properties

- **Concavity:** The dual function $H(\xi, \nu)$ is concave.
- **Dual Feasible:** $(\xi, \nu)$ is dual feasible if $\xi \in \mathbb{R}^m$ and $\nu \geq 0$.
- **Dual Optimal:** $(\xi^*, \nu^*)$ maximizes $H(\xi, \nu)$.
- **Primal-Dual Optimal:** $(x^*, \xi^*, \nu^*)$ where $x^*$ is optimal for the primal problem and $(\xi^*, \nu^*)$ are optimal for the dual problem.

## Weak Duality

- **Definition:** The value of the dual function at any feasible dual solution provides a lower bound on the value of the primal objective function at any feasible primal solution.
- **Formula:** $H(\xi^*, \nu^*) \leq F_0(x^*)$

## Strong Duality

- **Definition:** The optimal value of the dual problem is equal to the optimal value of the primal problem.
- **Formula:** $H(\xi^*, \nu^*) = F_0(x^*)$

## Slater's Constraint Qualification Theorem

- **Assumption:** $F_0, F_1, \ldots, F_M$ are convex functions with $\text{dom}(F_0) = \mathbb{R}^N$.
- **Condition:** There exists $x \in \mathbb{R}^N$ such that:

$$Ax = y \quad F_i(x) < b_i, \quad \forall i \in \{1, \ldots, M\}$$

- **Conclusion:** If the above conditions hold, then strong duality holds for the optimization problem.

## Saddle-Point Interpretation

For the primal problem:

$$\min_{x \in \mathbb{R}^N} F_0(x) \quad \text{subject to} \quad Ax = y,$$

the Lagrange function is:

$$L(x, \xi) = F_0(x) + \xi^T(Ax - y).$$

- **Supremum:**

$$\sup_{\xi \in \mathbb{R}^m} L(x, \xi) = \begin{cases} F_0(x) & \text{if } Ax = y \\ \infty & \text{otherwise} \end{cases}$$

– **Optimal Value:**

$$F_0(x^*) = \inf_{x \in \mathbb{R}^N} \sup_{\xi \in \mathbb{R}^m} L(x, \xi).$$

This saddle-point interpretation shows how the optimal value of the primal problem can be viewed as the balance point (saddle point) between the infimum over the primal variables and the supremum over the dual variables.

# Saddle-Point Interpretation

For the primal problem:

$$\min_{x \in \mathbb{R}^N} F_0(x)$$

subject to

$$Ax = y,$$

the Lagrange function is:

$$L(x, \xi) = F_0(x) + \xi^T (Ax - y).$$

– **Supremum:**

$$\sup_{\xi \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^N} L(x, \xi) \leq \inf_{x \in \mathbb{R}^N} \sup_{\xi \in \mathbb{R}^m} L(x, \xi)$$

– **Strong Duality:**

$$\sup_{\xi \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^N} L(x, \xi) = \inf_{x \in \mathbb{R}^N} \sup_{\xi \in \mathbb{R}^m} L(x, \xi)$$

## Saddle Point Property

For a primal-dual optimal pair $(x^*, \xi^*)$:

$$L(x^*, \xi) \leq L(x^*, \xi^*) \leq L(x, \xi^*)$$

for all $x \in \mathbb{R}^N$ and $\xi \in \mathbb{R}^m$.

This saddle-point interpretation shows that the optimal value of the primal problem can be viewed as the balance point (saddle point) between the infimum over the primal variables and the supremum over the dual variables.

# Gaussian Vectors

## Independent Gaussian Variables:

1. **Definition:**
   - $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, m$
   - $X_1, \ldots, X_m$ are independent
2. **Expectations:**
   - $E[X] = \mu = (\mu_1, \ldots, \mu_m)$
3. **Covariance Matrix:**

$$\mathrm{Cov}(X) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m^2 \end{pmatrix}$$

### General Gaussian Vectors:

1. **Definition:**
   - $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ Gaussian vector
   - $E[X] = \mu = (\mu_1, \ldots, \mu_m)$
   - $\text{Cov}(X) = \Sigma \in \mathbb{R}^{m \times m}$
2. **Component Distributions:**
   - $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$
3. **Linear Combinations:**
$$\langle \mathbf{v}, X \rangle \sim \mathcal{N}(\langle \mathbf{v}, \mu \rangle, \mathbf{v}^T \Sigma \mathbf{v})$$

# Taylor Expansions:

## 1. First-Order Taylor Expansion:

For a function $f : \mathbb{R} \to \mathbb{R}$ with continuous $f$ and $f'$:

$$f(x_0 + t) = f(x_0) + f'(x_0) \cdot t + o(t), \quad \text{as } t \to 0$$

- $f(x_0)$: Function value at $x_0$.
- $f'(x_0) \cdot t$: First-order term (derivative at $x_0$ times $t$).
- $o(t)$: Error term of smaller order than $t$ as $t \to 0$.

## 2. Little-o Notation ($o(t)$):

Describes a function that grows slower than $t$:

$$\lim_{t \to 0} \frac{o(t)}{t} = 0$$

Indicates that the error term $o(t)$ becomes negligible as $t$ approaches 0.

## 3. Example:

$$h_3(t) = t^2 : \quad t^2 = o(t) \text{ as } t \to 0$$

This summary covers the basic concept of Taylor expansions, the role of the first-order term, and the meaning of the error term $o(t)$, which is crucial for understanding and applying Taylor series approximations.

# Second-Order Taylor Expansion:

1. **Second-Order Expansion:**
   - For a function $f : \mathbb{R} \to \mathbb{R}$ with continuous $f$, $f'$, and $f''$:
$$f(x_0 + t) = f(x_0) + f'(x_0) \cdot t + \frac{1}{2} f''(x_0) \cdot t^2 + o(t^2), \quad \text{as } t \to 0$$
   - $f(x_0)$: Function value at $x_0$.
   - $f'(x_0) \cdot t$: First-order term.
   - $\frac{1}{2} f''(x_0) \cdot t^2$: Second-order term.

- $o(t^2)$: Error term of smaller order than $t^2$.

2. **Error Term $o(t^2)$:**
   - Describes a function that grows slower than $t^2$:
   $$\lim_{t \to 0} \frac{o(t^2)}{t^2} = 0$$

3. **Example:**
   - For $h_2(t) = t^3$:
   $$t^3 = o(t^2) \quad \text{as } t \to 0$$

Higher-Order Taylor Expansions

# Higher-Order Taylor Expansions:

1. **First-Order Taylor Expansion with Mean Value Theorem:**
$$f(x_0 + t) = f(x_0) + f'(z) \cdot t, \quad z \in (x_0, x_0 + t)$$

2. **Second-Order Taylor Expansion with Mean Value Theorem:**
$$f(x_0 + t) = f(x_0) + f'(x_0) \cdot t + \frac{1}{2} f''(z) \cdot t^2, \quad z \in (x_0, x_0 + t)$$

3. **General kth-Order Taylor Expansion:**
$$f(x_0 + t) = f(x_0) + \sum_{i=1}^{k} \frac{f^{(i)}(x_0)}{i!} \cdot t^i + \frac{f^{(k+1)}(z)}{(k+1)!} \cdot t^{k+1}, \quad z \in (x_0, x_0 + t)$$

# Multidimensional Case:

1. **First-Order Expansion in $\mathbb{R}^m$:**
$$f(x_0 + ty) = f(x_0) + \nabla f(x_0) \cdot y + o(t), \quad \text{as } t \to 0$$
   - $\nabla f(x_0)$: Gradient (vector of partial derivatives).
   - $y \in \mathbb{R}^m$: Direction of displacement.
   - $t \in \mathbb{R}$: Amount of movement.

2. **Alternative Notation:**
$$f(x_0 + y) = f(x_0) + \nabla f(x_0) \cdot y + o(\|y\|), \quad \text{as } \|y\| \to 0$$

3. **First-Order Mean Value Theorem:**
$$f(x_0 + y) = f(x_0) + \nabla f(z) \cdot y, \quad z \text{ is on the line segment } [x_0, x_0 + y]$$

4. **Case with Scalar $t$:**
$$f(x_0 + ty) = f(x_0) + \nabla f(z) \cdot y, \quad \text{for } z \in [x_0, x_0 + ty]$$

## Hessian Matrix:

1. Definition:

$$\nabla^2 f(x_0) = D^2 f(x_0) = \begin{pmatrix} \frac{\partial^2 f(x_0)}{\partial x_1^2} & \frac{\partial^2 f(x_0)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x_0)}{\partial x_1 \partial x_m} \\ \frac{\partial^2 f(x_0)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_0)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x_0)}{\partial x_2 \partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x_0)}{\partial x_m \partial x_1} & \frac{\partial^2 f(x_0)}{\partial x_m \partial x_2} & \cdots & \frac{\partial^2 f(x_0)}{\partial x_m^2} \end{pmatrix}$$

   - The Hessian is symmetric.

## Second-Order Taylor Expansion in $\mathbb{R}^m$:

1. Expansion:

$$f(x_0 + ty) = f(x_0) + \nabla f(x_0) \cdot y + \frac{1}{2} t^2 y^T \nabla^2 f(x_0) y + o(t^2), \quad \text{as } t \to 0$$

2. Correct Notation for Quadratic Form:
   - $y^T \nabla^2 f(x_0) y$: Correct form.

## Derivatives Along Curves:

1. First Derivative:
   - For $g(t) = f(\gamma(t))$:

$$g'(t) = \frac{d}{dt} g(t) = \nabla f(\gamma(t)) \cdot \gamma'(t)$$

2. Second Derivative:
   - For $g(t) = f(\gamma(t))$:

$$g''(t) = \frac{d^2}{dt^2} g(t) = \gamma'(t)^T \nabla^2 f(\gamma(t)) \gamma'(t) + \nabla f(\gamma(t)) \cdot \gamma''(t)$$

## Lower Semi-Continuity:

1. **Definition:**
   - A function $f : \mathbb{R}^m \to \mathbb{R}$ is lower semi-continuous if its epigraph is a closed set.
   - The epigraph of $f$ is:

$$\text{epi}(f) = \{(x, y) \in \mathbb{R}^m \times \mathbb{R} : y \geq f(x)\}$$

2. **Geometric Interpretation:**
   - The epigraph is the set of points lying on or above the graph of $f$.
3. **Properties:**
   - Any continuous function is lower semi-continuous.
   - A function can be lower semi-continuous but not continuous if the epigraph is closed even with discontinuities.
   - A function is not lower semi-continuous if there are gaps in the epigraph that cannot be filled.

Gradient Descent Cheat Sheet

## 40 1. Objective:

Minimize $f(x)$

- Find $x^* = \arg\min_x f(x)$.

## 41 2. Gradient:

- $\nabla f(x)$: Direction and rate of fastest increase.

## 42 3. Update Rule:

- $\hat{x} = x - \alpha \nabla f(x)$
- $\alpha$: Learning rate (step size).

## 43 4. Time-Dependent Gradient Flow:

- Differential equation: $\dot{x}(t) = -\nabla f(x(t))$
- Initial condition: $x(0) = x^{(0)}$

## 44 5. Lipschitz Continuity:

- Gradient $\nabla f(x)$ is Lipschitz continuous with constant $L$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- If $f$ is twice differentiable:

$$\lambda_{\max}(\nabla^2 f(x)) \leq L$$

## Gradient Descent Iteration

- **Objective:** Minimize a function $f(x)$ by iteratively moving towards the minimum.
- **Update Rule:**

$$x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f(x^{(n)})$$

- **Components:**
  - $x^{(n)}$: Current value.
  - $x^{(n+1)}$: Next value.
  - $\alpha^{(n)}$: Step size (learning rate).
  - $\nabla f(x^{(n)})$: Gradient of $f$ at $x^{(n)}$.
- **Step Size ($\alpha^{(n)}$):**
  - **Adaptive:** Changes at each iteration.
  - **Constant:** Fixed value, must be small enough to ensure convergence.
- **Convergence:** Sequence $x^{(n)}$ should approach the minimum of $f$.

# Convexity and Lipschitz Continuity

## Convexity:

- A convex function $f$ satisfies:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

- For all $x, y \in \mathbb{R}^d$.

## Lipschitz Continuity:

- If $\nabla f(x)$ is Lipschitz continuous with constant $L$, then:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$$

## Theorem: Convergence of Gradient Descent

**Assumptions:**

1. $f$ is convex.
2. $f$ is finite for all $x$.
3. A finite solution $x^*$ exists.
4. $\nabla f(x)$ is Lipschitz continuous with constant $L$.

**Condition:**

$$\alpha = \alpha^{(n)} \leq \frac{1}{L}$$

**Conclusion:**

The iteration

$$x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f(x^{(n)})$$

converges to $x^*$.

# Strong Convexity

- **Definition:** A function $f$ is strongly convex with coefficient $\gamma > 0$ if it satisfies certain conditions that make it "more curved" than a simple convex function.

1. **Gradient Condition:**

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2} \|y - x\|_2^2$$

2. **Gradient Difference Condition:**

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \gamma \|x - y\|_2^2$$

3. **Hessian Condition:**

$$w^T \nabla^2 f(x) w \geq \gamma \|w\|_2^2 \quad \forall x, w \in \mathbb{R}^n$$

   - Alternatively:

$$\nabla^2 f(x) \succeq \gamma I$$

4. **Jensen's Inequality for Strongly Convex Functions:**

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y) - \frac{\gamma}{2} \lambda (1 - \lambda) \|x - y\|_2^2$$

# 45 Orthogonal Projections

## 45.1 Definition of Projection $P_V(x)$:

- $P_V(x)$ is the projection of $x \in \mathbb{R}^d$ onto the subspace $V \subset \mathbb{R}^d$.
- If $V = \{v_1, v_2, \ldots, v_n\}$ (orthonormal basis), then:

$$P_V(x) = \sum_{i=1}^{n} \langle v_i, x \rangle v_i$$

## 45.2 Matrix Form of Projection $P_V$:

- $P_V$ is the operator that performs projections, given by:

$$P_V = VV^T \in \mathbb{R}^{d \times d}$$

- Here, $V$ is a $d \times n$ orthogonal matrix whose columns span the subspace $V$.

## 45.3 Relationship Between $P_V$ and $P_V(x)$:

- The matrix $P_V$ and the projection $P_V(x)$ are intimately related:

$$P_V \cdot x = P_V(x)$$

The convolution of two functions $f$ and $g$, denoted by $(f * g)$, is defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(u)g(t-u)\,du$$

For probability density functions (pdfs) $\phi_X$ and $\phi_Y$ of two independent random variables $X$ and $Y$, the pdf of the sum $Z = X + Y$ is given by the convolution of $\phi_X$ and $\phi_Y$:

$$\phi_{X+Y}(t) = (\phi_X * \phi_Y)(t) = \int_{-\infty}^{\infty} \phi_X(u)\phi_Y(t-u)\,du$$

- Given $a = b + c$
- If $c \geq 0$, then $a \geq b$

$$\sup_{i \in I}[g_i(x) + h_i(y)] \leq \sup_{i \in I} g_i(x) + \sup_{i \in I} h_i(y).$$

- Convexity Rule:

For a function $f : \mathbb{R}^n \to (-\infty, \infty)$, $f$ is convex if:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad \text{for all } x, y \in \mathbb{R}^n \text{ and } \lambda \in [0,1].$$

Excercises chapter convex analysis - equations needed

1. **Definition of Convex Function:**

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y \in C, \lambda \in [0,1]$$

2. **Gradient Inequality (First-Order Condition for Convexity):**

$$f(x+v) \geq f(x) + \nabla f(x) \cdot v \quad \forall x, v$$

3. **Positive Semi-Definiteness of the Hessian (Second-Order Condition for Convexity):**

$$\nabla^2 f(x) \geq 0 \quad \forall x \in C$$

This means that for any vector $v$,

$$v^T \nabla^2 f(x) v \geq 0$$

4. **Taylor Expansion:** For a twice continuously differentiable function $f$ around a point $x$,

$$f(x + tv) = f(x) + \nabla f(x) \cdot v + \frac{t^2}{2} v^T \nabla^2 f(x) v + o(t^2)$$

where $t$ is a small scalar, and $v$ is a vector.

5. **Limit Process:** When deriving inequalities from Taylor expansions, you often need to divide by $t^2$ and take the limit as $t$ approaches zero:

$$\lim_{t \to 0} \frac{o(t^2)}{t^2} = 0$$

## Gaussian Distribution (N($\mu$, $\sigma^2$)):

- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- MGF: $M_X(\theta) = \exp\left(\mu\theta + \frac{1}{2}\sigma^2\theta^2\right)$

## Standard Gaussian Distribution (N(0, 1)):

- PDF: $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
- MGF: $M_X(\theta) = \exp\left(\frac{1}{2}\theta^2\right)$

## Lipschitz Continuity:

- $|f(x) - f(y)| \leq L\|x - y\|$
- Lipschitz Continuity and Derivative: $\|Df(x)\| \leq L$
- Lipschitz Gradient: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

## Convex Functions with Lipschitz Gradient:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2$$

## Strong Convexity:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2$$

### Theorem:

If $f$ is twice continuously differentiable and the Hessian $\nabla^2 f(x)$ is bounded, then $\nabla f$ is Lipschitz continuous. Specifically,

$$\|\nabla^2 f(x)\| \leq L \quad \forall x \in D \subset \mathbb{R}^n$$

implies that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L$.

## Hessian Norm:

– Use the spectral norm (largest singular value) to estimate the Lipschitz constant.

$$\|\nabla^2 f(x)\| = \text{Largest Singular Value}$$

Meaning only eigenvalues is also enough without doing the full SVD.

## Quadratic Function:

$$f(x) = x^T Ax + \langle b, x \rangle + c$$

## Gradient:

$$\nabla f(x) = Ax + A^T x + b$$

(For symmetric $A$: $\nabla f(x) = 2Ax + b$)

## Hessian:

$$\nabla^2 f(x) = A + A^T$$

(For symmetric $A$: $\nabla^2 f(x) = 2A$)

## Summary

– **Valid Inner Product:**

$$\langle u, v \rangle \text{ where } u, v \in \mathbb{R}^n.$$

– **Matrix-Vector Multiplication:**

$$Ax \text{ where } A \in \mathbb{R}^{m \times n} \text{ and } x \in \mathbb{R}^n.$$

– **Quadratic Form:**

$$x^T Ax \text{ where } A \in \mathbb{R}^{n \times n} \text{ and } x \in \mathbb{R}^n.$$

## Example

**Inner Product of Vectors:**

$$\langle u, v \rangle = u^T v.$$

**Matrix-Vector Multiplication:**

$$y = Ax.$$

**Quadratic Form:**

$$x^T Ax = \langle Ax, x \rangle.$$

## Norms and Lipschitz Constant

– $\|\cdot\|_1$ Norm:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

– $\|\cdot\|_2$ Norm:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

– Lipschitz Constant Formula:

$$L = \sup_{x \in \mathbb{R}^n} \|\nabla f(x)\|_2$$

– Steps to Compute:

1. Compute the gradient $\nabla f(x)$.
2. Evaluate the Euclidean norm $\|\nabla f(x)\|_2$.
3. Find the supremum over the domain.

## Gradient of Norm Functions

### 1. Gradient of the Euclidean Norm ($L_2$ Norm):

For the Euclidean norm $\|x\|_2 = \sqrt{x^T x}$, the gradient is:

$$\nabla \|x\|_2 = \frac{x}{\|x\|_2} \text{ for } x \neq 0.$$

### 2. Gradient of the Squared Euclidean Norm:

For the squared Euclidean norm $\|x\|_2^2 = x^T x$, the gradient is:

$$\nabla \|x\|_2^2 = 2x.$$

### 3. Gradient of the Manhattan Norm ($L_1$ Norm):

The Manhattan norm $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ is not differentiable everywhere, but the subgradient can be given as:

$$\partial \|x\|_1 = \begin{cases} 1, & \text{if } x_i > 0, \\ -1, & \text{if } x_i < 0, \\ [-1, 1], & \text{if } x_i = 0. \end{cases}$$

**Convergence of Gradient Descent to Minimize:**
If the above rule $\gamma \leq \frac{1}{L}$ where $L$ is the Lipschitz constant of $f$,
**IMPORTANT:** In the first part of the exercise, we worked with Lipschitz constant of $f$, not of $\nabla f$.

**ALWAYS TRUE:**
$(H \cdot N)^{-1} = N^{-1} \cdot H^{-1}$

**ORTHOGONAL MATRICES:**
$O^{-1} = O^T$

$$UU^T = I \quad \text{for Orthonomal Matrices}$$

1. Orthogonal Matrix:

$$U^T U = I \quad \text{and} \quad UU^T = I \quad \text{and} \quad U^{-1} = U^T$$

2. Inverse and Identity:

$$AA^{-1} = I \quad \text{and} \quad A^{-1}A = I$$

3. Pseudo-inverse:

$$A^+ = (A^T A)^{-1} A^T$$
$$AA^+ = A \quad \text{and} \quad A^+ A = I$$
$$(AA^+)^T = AA^+ \quad \text{and} \quad (A^A)^T = A^+ A$$

4. Symmetric Matrix:

$$B = B^T \quad \text{and} \quad (B^{-1})^T = B^{-1}$$

5. SVD and Inverse:

$$B = U\Sigma V^T$$
$$B^{-1} = V\Sigma^{-1} U^T$$
$$\Sigma^{-1} = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n}\right)$$

1. Original Definition for Linearly Independent Columns:

$$A^+ = (A^T A)^{-1} A^T \quad (\text{sum form: } A^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T)$$

2. New Definition for Linearly Independent Rows:

$$A^+ = A^T (AA^T)^{-1} \quad (\text{sum form: } A^+ = \sum_{i=1}^{m} \frac{1}{\sigma_i} v_i u_i^T)$$