ECE367 Cheatsheet

John Doe

October 1, 2024

Contents

List of Figures

List of Tables

1 Vectors, Norms, Inner Products (Ch. 2.1-2.2)

1.1 Linear transformation

Definition: $T: X \to Y$ that satisfies

1. **Additivity:** $T(x_1 + x_2) = T(x_1) + T(x_2)$
2. **Homogeneity:** $T(\alpha x) = \alpha T(x)$

- **Note:** Linear algebra is the study of linear transformations over vector spaces.

1.1.1 Matrix representation of a linear transformation

Definition: Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. Let $T: \mathcal{V} \to \mathcal{W}$ be a linear transformation. When $\mathcal{V} = \mathbb{R}^n$ (or $\mathbb{C}^n$) and $\mathcal{W} = \mathbb{R}^m$ (or $\mathbb{C}^m$), then T can be uniquely represented as a matrix $A \in \mathbb{R}^{m \times n}$ such that:

$$T(\mathbf{x}) = A\mathbf{x}$$

where

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

- **Key:** Any linear transformation is a matrix multiplication. Any matrix multiplication is a linear transformation.

1.2 Vectors

Definition: Ordered collection of numbers, where $x_i \in \mathbb{R}$ or $\mathbb{C}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}.$$

- n: Dimension of $\mathbf{x}$
- $\mathbf{x}$: Column vector
- $\mathbf{x}^T$: Transpose of x (row vector)
- T: Transpose
- x_i: i-th element of x.

1.3 Vector spaces

Definition: A vector space over a field $\mathbb{F}$ (e.g. $\mathbb{R}/\mathbb{C}$) consists of:

1. A set of vectors $\mathcal{V}$
2. A vector addition operator $+$: $\mathcal{V} \times \mathcal{V} \to \mathcal{V}$ s.t. $\forall x, y \in \mathcal{V} \to x + y \in \mathcal{V}$ (i.e. closed under VA)
3. A scalar multiplication operator $\cdot$: $\mathbb{F} \times \mathcal{V} \to \mathcal{V}$ s.t. $\forall \alpha \in \mathbb{F}$, $\forall x \in \mathcal{V} \to \alpha x \in \mathcal{V}$ (i.e. closed under SM)

- $\times$ is not scalar multiplication.

For $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{F}$. The following properties are satisfied:

- **Vector addition** satisfies (i.e., Abelian group):
  1. **Commutativity:** $\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$.
  2. **Associativity:** $\mathbf{x}+(\mathbf{y}+\mathbf{z})=(\mathbf{x}+\mathbf{y})+\mathbf{z}$.
  3. **Additive identity:** $\exists\, \mathbf{0} \in \mathcal{V}$ s.t. $\mathbf{x}+\mathbf{0}=\mathbf{0}+\mathbf{x}=\mathbf{x}$.
  4. **Additive inverse:** $\forall\, \mathbf{x},\ \exists\, \mathbf{y}$ s.t. $\mathbf{x}+\mathbf{y}=\mathbf{0}$ (i.e. $\mathbf{y}=-\mathbf{x}$).
- **Scalar multiplication** satisfies:
  1. **Associativity:** $\alpha\cdot(\beta\cdot\mathbf{x})=(\alpha\cdot\beta)\cdot\mathbf{x}$.
  2. **Multiplicative Identity:** $\exists\, 1 \in \mathbb{F}$ s.t. $1\cdot\mathbf{x}=\mathbf{x}$.
  3. **Right Distributivity:** $\alpha\cdot(\mathbf{x}+\mathbf{y})=\alpha\cdot\mathbf{x}+\alpha\cdot\mathbf{y}$.
  4. **Left Distributivity:** $(\alpha+\beta)\cdot\mathbf{x}=\alpha\cdot\mathbf{x}+\beta\cdot\mathbf{x}$.

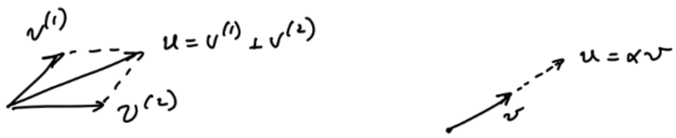

Figure 1: Vector addition and scalar multiplication.

1.3.1 How to prove or disprove a vector space?

Process:
Prove:
1. Prove that $\mathcal{V}$ is closed under VA and SM.
2. Prove all the properties under VA and SM.

Disprove:
1. Disprove one of the properties or that it isn't closed under VA and SM.

Warning: If standard addition and multiplication then, closed under VA and SM properties is enough to prove it's a vector space.

Example:
- Let $\mathcal{V}=\mathbb{R}^n$ and $\mathbb{F}=\mathbb{R}$: This represents vectors of dimension n where each element belongs to $\mathbb{R}$.

$$\mathcal{V}=\mathbb{R}^n=\left\{\begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix} : x_i \in \mathbb{R}\right\}$$

For $\mathbf{x}=\begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix} \in \mathbb{R}^n$ and $\mathbf{y}=\begin{bmatrix}y_1\\ \vdots \\ y_n\end{bmatrix} \in \mathbb{R}^n$:

$$\mathbf{x}+\mathbf{y}=\begin{bmatrix}x_1+y_1\\ \vdots \\ x_n+y_n\end{bmatrix}$$

For $\alpha, \beta \in \mathbb{R}$:

$$\alpha\cdot\mathbf{x}=\begin{bmatrix}\alpha x_1\\ \vdots \\ \alpha x_n\end{bmatrix}$$

- Let $\mathcal{V} = \mathbb{C}^n$ and $\mathbb{F} = \mathbb{C}$: This represents vectors of dimension n with complex components.

$$\mathcal{V} = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} : x_i \in \mathbb{C} \right\}$$

  $\mathcal{V}$ is a vector space over $\mathbb{C}$ under element-wise addition and scalar multiplication.
- Let $\mathcal{V} = \{\text{set of all continuous functions } f : \mathbb{R} \to \mathbb{R}^n\}$ and $\mathbb{F} = \mathbb{R}$:
  Let $f_1, f_2 \in \mathcal{V}$, and for $t \in \mathbb{R}$:

$$(f_1 + f_2)(t) = f_1(t) + f_2(t) \quad \Rightarrow \quad f_1 + f_2 \in \mathcal{V}$$

  For $\alpha \in \mathbb{R}$:

$$(\alpha f)(t) = \alpha f(t) \quad \Rightarrow \quad \alpha f \in \mathcal{V}$$

  - f is the vector, $\mathbb{R} \to \mathbb{R}^n$ is the input-output relationship. For 2D, $f(x) = [x_1, x_2]^T$, where x is the input, the vector is the output in 2D, and the vector is f.
- Let $\mathcal{V} = \mathcal{P}_n$, the set of all polynomials with real coefficients and degree $\leq n$:

$$\mathcal{V} = \mathcal{P}_n = \left\{ p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n : a_0, a_1, \ldots, a_n \in \mathbb{R} \right\}$$

  $\mathcal{V}$ is a vector space over $\mathbb{R}$ under standard addition and scalar multiplication.

1.4 Subspace

Definition: A **subspace** is a subset of a vector space $\mathcal{V}$ that is a vector space by itself.
- **Test:** To check whether a subset is a subspace, check that it is closed under VA & SM.

Example:
- Let $\mathcal{V} = \mathbb{R}^3$, and consider the set:

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} : x_1, x_2 \in \mathbb{R} \right\}$$

  This set S is a subspace of $\mathbb{R}^3$.
- Let $\mathcal{V} = \mathbb{R}^3$, and consider the set:

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} : x_1, x_2 \in \mathbb{R} \right\}$$

  This set S is **not** a subspace of $\mathbb{R}^3$ because adding two vectors will make the last component 2.
- Let $\mathcal{V} = \mathbb{R}^n$, and consider the set:

$$S = \{\mathbf{0}\}$$

  This set S is a subspace of $\mathbb{R}^n$.

1.5 Span

Definition: Given a finite set of vectors $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ in the same vector space $\mathcal{V}$ over some field $\mathbb{F}$ then,

$$\text{Span}(S) = \left\{ \sum_{i=1}^{m} \alpha_i \mathbf{v}_i \mid \alpha_i \in \mathbb{F} \right\}$$

- **Note:** Span(S) is always a subspace of V.

1.5.1 How to draw the span?

Process:

1. Identify the vectors.
2. Plot the vectors: Plot each vector on a coordinate plane starting at the origin.
3. Draw the span: Extend the vectors in both directions to show the line or plane formed by their span. If they span the entire plane, draw dashed lines extending their direction.

Example:

- Let $S = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right\}$:

$$\text{span}(S) = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} : x_1, x_2 \in \mathbb{R} \right\}$$

  This set span(S) forms a plane in $\mathbb{R}^3$. The vectors span the xy-plane with the z-coordinate fixed at zero.

- Let $S = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix} \right\}$:

$$\text{span}(S) = \left\{ x \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} : x \in \mathbb{R} \right\}$$

  In this case, span(S) is a line in $\mathbb{R}^3$ along the x-axis with y and z coordinates fixed at zero.

1.6 Linear independent (LI) set

Definition: A set of vectors $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is LI if no vector in S can be written as a LC of other vectors in S. In other words, the only α_i's that makes $\sum_{i=1}^{m} \alpha_i \mathbf{v}_i = 0$ is $\alpha_i = 0,\ \forall i$.

- If $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is LI, then $\forall$ $\mathbf{u} \in \text{span}(S)$, there is a **unique** set of α_i's s.t. $\mathbf{u} = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i$ (i.e. there is no redundancies in representation)
  - **Coordinates:** $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ of $\mathbf{u}$ w.r.t. S.
- If S is linearly dependent, then one of the vectors can be written as a LC of the other vectors. In this case, we can remove that vector and continue this process until the remaining set is LI.
  - **Note:** Such an irreducible linearly independent set is called a **basis** of span(S).

1.6.1 How to determine if a set is linearly independent

Process:

1. Write a linear combination with coefficients $\alpha_1, \ldots, \alpha_k$.
2. Set the linear combination equal to 0.
3. Solve for $\alpha_1, \ldots, \alpha_k$ by solving the set of equations (i.e. each component is one equation).
4. If $\alpha_1 = \ldots = \alpha_k = 0$, then it is linearly independent.
5. Else, linearly dependent by finding a counter example, where the linear combination is 0 for $\alpha_1, \ldots, \alpha_k$ not all equal to 0.

1.7 Basis

Definition: A set of vectors B is a basis of a vector space $\mathcal{V}$ if
- B is LI
- $Span(B) = \mathcal{V}$

Example: What is the standard basis for $\mathcal{V} = \mathbb{R}^n$?

$$\mathbf{e}^{(1)} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}^{(2)} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots \quad , \quad \mathbf{e}^{(n)} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad \text{for } \mathbb{R}^n$$

If $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, then:

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$$

1.7.1 Dimension

Definition: The dimension is the number of basis vectors.
- **Note:** Basis is not unique. But $dim(\mathcal{V})$ is well-defined.

Example:
- $\dim\left(\text{span}\left(\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}\right)\right) = 2$
- $\dim\left(\text{span}\left(\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}\right)\right) = 1$
- $\dim(\{\mathbf{0}\}) = 0$
- The dimension for $\mathcal{V} = \mathbb{R}^n$ of the standard basis is n

Process:
1. Given a set of vectors, S
2. If l.i. and $span(S) = \mathcal{V}$, then it's a basis.
3. If for each pair of vectors, the inner product is 0, then it's orthogonal basis.
4. If unit norm, then it's orthonormal basis.

1.8 Norms (Notion of distance)

Definition: Let $\mathcal{V}$ be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A norm is a function $\|\cdot\|\colon \mathcal{V} \to \mathbb{R}$ that satisfies
1. **Non-negativity:** $\|\mathbf{x}\| \geq 0,\ \forall \mathbf{x} \in \mathcal{V},$ and $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$
2. **Homogeneity:** $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|\ \forall \mathbf{x} \in \mathcal{V},\ \alpha \in \mathbb{F}$
3. **Triangle inequality:** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$ (triangular inequality)

Example: ℓ_p norms:

$$\|\mathbf{x}\|_p \equiv \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p}, \quad 1 \leq p < \infty.$$

- **Note:** For $p < 1$, triangular inequality doesn't hold.

1. **Sum-of-absolute-values length** $p = 1$: $\|\mathbf{x}\|_1 \equiv \sum_{k=1}^{n} |x_k|$

2. **Euclidean length** $p = 2$: $\|\mathbf{x}\|_2 \equiv \sqrt{\sum_{k=1}^{n} x_k^2}$
3. **Max absolute value norm** $p = \infty$: $\|\mathbf{x}\|_\infty \equiv \max_{k=1,\ldots,n} |x_k|$
   - Largest term will dominate as if we common factor out the largest term, each of the other terms will go to 0 as noted in the lp norm.
4. **Cardinality** $p = 0$: The number of non-zero vectors in x is

$$\|\mathbf{x}\|_0 = \text{card}(\mathbf{x}) \equiv \sum_{k=1}^{n} \mathbb{I}(x_k \neq 0), \quad \text{where} \quad \mathbb{I}(x_k \neq 0) \equiv \begin{cases} 1 & \text{if } x_k \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

   - **Key:** Not a norm since $\|\alpha \mathbf{x}\|_0 = \|\mathbf{x}\|_0 \neq |\alpha| \cdot \|\mathbf{x}\|_0$ (e.g. if $\alpha = 2$ then this would double the count of number of non-zero vectors for the RS)

1.8.1 Norm balls

Definition: The set of all vectors with ℓ_p norm less than or equal to one,

$$B_p = \{\mathbf{x} : \|\mathbf{x}\|_p \leq 1\} \tag{1}$$

Example: For 2D, the norm balls are as follows:

- $\ell_2 : \ B_2 = \left\{\mathbf{x} \;\middle|\; \sqrt{x_1^2 + x_2^2} \leq 1\right\}$
- $\ell_1 : \ B_1 = \{\mathbf{x} \mid |x_1| + |x_2| \leq 1\}$
- $\ell_\infty : \ B_\infty = \{\mathbf{x} \mid \max |x_i| \leq 1 \text{ or } |x_1| \leq 1, |x_2| \leq 1\}$
- $\ell_0 : \ B_0 = \{\mathbf{x} \mid \text{card}(\mathbf{x}) \leq 1\}$

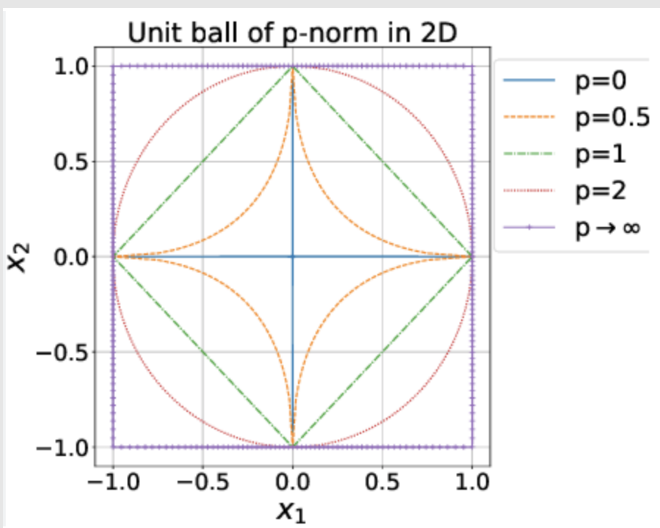

Figure 2: Norm balls of different p values.

1.8.2 Motivation for Norms

Example: In optimization problems, different norms are used to achieve various goals. Suppose we are trying to solve an optimal control problem, where $x = (x_1, \ldots, x_n)$ are some action variables.

- $\min \|\mathbf{x}\|_2^2 = x_1^2 + \ldots + x_n^2$ (i.e. minimizing the total energy (power) in $\mathbf{x}$)
- $\min \|\mathbf{x}\|_\infty$ (i.e. minimizing the peak energy in $\mathbf{x}$).
- $\min \|\mathbf{x}\|_1$ (i.e. minimizing the sum of action variables).
- $\min \|\mathbf{x}\|_0$ (i.e. find sparse solution)

1.8.3 Distance metric

Definition: A norm induces a distance metric between two vectors x and y in $\mathbb{V}$ as

$$d(x, y) = \|x - y\|$$

- **Note:** The ℓ_2-norm induces the Euclidean distance

$$\|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

1.9 Inner product (Notion of angle)

Definition: An inner product on a vector space $\mathcal{V}$ is a function $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathcal{F}$ such that:

1. **Positive definiteness:** $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \ \forall \mathbf{x} \in \mathcal{V}$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ iff $\mathbf{x} = 0$
2. **Conjugate Symmetry:** $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$
   - $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ in $\mathbb{R}^n$
   - $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$ in $\mathbb{C}^n$.
3. **Linearity in first argument:** $\langle \alpha \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}, \ \alpha \in \mathbb{F}$

Example: How to use the properties of inner products?

$$\begin{aligned} \langle x, \alpha y + z \rangle &\overset{(2)}{=} \overline{\langle \alpha y + z, x \rangle} \\ &\overset{(3)}{=} \overline{\alpha}\, \overline{\langle y, x \rangle} + \overline{\langle z, x \rangle} \qquad \text{also by conjugate prop.} \\ &\overset{(2)}{=} \overline{\alpha} \langle x, y \rangle + \langle x, z \rangle \end{aligned}$$

1.9.1 Examples of inner products

Example:

- In $\mathbb{R}^n$ (Dot product): $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$
  - **Key:** $\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^{n} x_i^2 = \mathbf{x}^\top \mathbf{x} = \|\mathbf{x}\|_2^2$
- In $\mathbb{C}^n$: $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} \overline{x_i} y_i = \mathbf{y}^H \mathbf{x} = \overline{\mathbf{x}^H \mathbf{y}}$
  - $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{x}^H = \begin{bmatrix} \overline{x_1} & \cdots & \overline{x_n} \end{bmatrix}$
- $\mathcal{V} = \left\{ f : \mathbb{R} \to \mathbb{R} \,;\, \int_{-\infty}^{+\infty} f^2(t)\, dt < \infty \right\}$ (i.e. the set of square integrable functions)

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) g(t)\, dt$$

1.9.2 Connection of inner product to angle

In $\mathbb{R}^n$, the notion of inner product has a geometric interpretation, and is closely related to the notion of angle between vectors.

Definition:

$$\cos\theta = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \right\rangle \tag{2}$$

- $\langle \mathbf{x}, \mathbf{y}\rangle = 0 \Rightarrow \cos\theta = 0 \Rightarrow \theta = \frac{\pi}{2}$ (i.e. perpendicular)
- $\langle \mathbf{x}, \mathbf{y}\rangle = \|\mathbf{x}\|_2\|\mathbf{y}\|_2 \Rightarrow \cos\theta = 1 \Rightarrow \theta = 0$ (i.e. $\mathbf{x}$ and $\mathbf{y}$ are aligned)
- $\langle \mathbf{x}, \mathbf{y}\rangle = -\|\mathbf{x}\|_2\|\mathbf{y}\|_2 \Rightarrow \cos\theta = -1 \Rightarrow \theta = \pi$ (i.e $\mathbf{x}$ and $\mathbf{y}$ are in opposite directions)
- $\langle \mathbf{x}, \mathbf{y}\rangle > 0 \Rightarrow \cos\theta > 0 \Rightarrow$ angle is acute
- $\langle \mathbf{x}, \mathbf{y}\rangle < 0 \Rightarrow \cos\theta < 0 \Rightarrow$ angle is obtuse

Derivation: L3: Inner products and orthogonality.

1.9.3 Cauchy-Schwartz inequality and its generalization

Definition:

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2 \tag{3}$$

Hölder's Inequality (generalization):

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|_p\|\mathbf{y}\|_q \quad \text{where } 1 \leq p, q < \infty \text{ and } \frac{1}{p} + \frac{1}{q} = 1 \tag{4}$$

Example: For $p = 1$ and $q = \infty$, we have:

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|_1 \cdot \|\mathbf{y}\|_\infty$$

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \left(\sum_{i=1}^{n} |x_i|\right) \cdot \max_i |y_i|$$

1.9.4 Inner product induces a norm

Definition: Any inner product induces a norm, but not all norms are induced by an inner product.

- **Key:** If given an inner product, take the square root of the inner product to get the norm.
  - e.g. $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$, which holds for $\mathbb{R}^n$ and $\mathbb{C}^n$

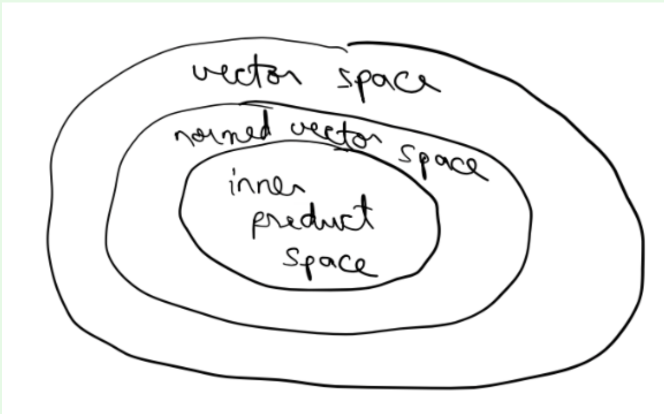

Figure 3: Ordering of the vector spaces.

Warning: A norm doesn't induce an inner product (e.g. l_1 or l_∞)

1.10 Orthogonal decomposition

1.10.1 Mutually orthogonal

Definition: A set of non-zero vectors $S = \left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(d)}\right\}$ is **mutually orthogonal** if $\langle \mathbf{v}^{(i)}, \mathbf{v}^{(j)} \rangle = 0 \; \forall \; i \neq j$.

- **Fact:** Orthogonal set of vectors $S = \left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(d)}\right\}$ is linearly independent.
  - **Proof:** In L3.

1.10.2 Orthonormal basis

Definition: Set of orthogonal basis vectors that have unit norm.

If $S = \left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(d)}\right\}$ is a set of mutually orthogonal vectors, then $\left\{\frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}, \ldots, \frac{\mathbf{v}_d}{\|\mathbf{v}_d\|}\right\}$ is an orthonormal basis for $span(S)$

Example: Standard basis is an orthonormal basis for $\mathbb{R}^n$

1.10.3 Orthogonal

Definition: Consider $\mathbf{x} \in \mathcal{V}$, and let S be a subspace of $\mathcal{V}$. We say $\mathbf{x}$ is orthogonal to S if:

$$\langle \mathbf{x}, \mathbf{v} \rangle = 0 \quad \forall \, \mathbf{v} \in S.$$

We write: $\mathbf{x} \perp S$.

1.10.4 Orthogonal complement

Definition: The **orthogonal complement** of S, denoted $S^\perp$, is the set of all orthogonal vectors to S:

$$S^\perp = \{\mathbf{x} \in \mathcal{V} \, : \, \mathbf{x} \perp S\}$$

- $S^\perp$ is a subspace. (Closed under addition and scalar multiplication)
- $S \cap S^\perp = \{\mathbf{0}\}$
- **Orthogonal decomposition:** Any $\mathbf{x} \in \mathcal{V}$ can be uniquely written as: $\mathbf{x} = \mathbf{x}_S + \mathbf{x}_{S^\perp}$ where $\mathbf{x}_S \in S$ and $\mathbf{x}_{S^\perp} \in S^\perp$

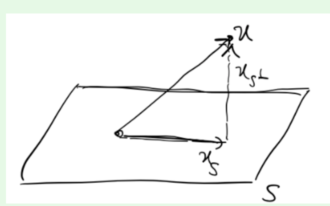

Figure 4: Drawing any x.

- $\mathcal{V} = S + S^\perp = \left\{\mathbf{u} + \mathbf{v} \, : \, \mathbf{u} \in S, \, \mathbf{v} \in S^\perp\right\}$

2 Orthogonal Decomposition, Projecting onto Subspaces, Gram-Schmidt, QR Decomposition, Projection onto Affine Sets, Hyperplanes and Half-Spaces (Ch. 2.2-2.3)

2.1 Projection onto subspaces

Definition:

$$x^* = \text{Proj}_S(x) = \arg\min_{y \in S} \|x - y\|_2 \tag{5}$$

If $\{v^{(1)}, \ldots, v^{(d)}\}$ is an orthonormal basis of S then

$$x^* = \sum_{i=1}^{d} \langle x, v^{(i)} \rangle v^{(i)} \tag{6}$$

- The error vector should be orthogonal to each vector in the subspace.

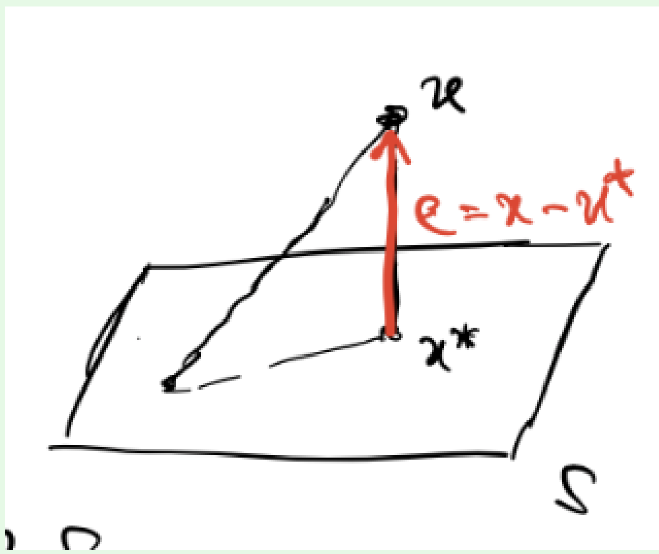

Figure 5: Error vector being perp. to S.

Example: For $v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$

The ith component can be extracted by doing the inner product with the ith standard basis:

$$\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = v_1$$

$$\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = v_2$$

$$\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = v_3$$

Therefore, analogous to x*, we can write them as the sum of the inner product times the standard basis.

$$v = v_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + v_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + v_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

2.1.1 Basic problem

Intuition: Given $x \in \mathcal{V}$ and a subspace S. Find the closest point (in norm) in S to x:

$$\text{Proj}_S(x) = \arg\min_{y \in S} \|y - x\| \tag{7}$$

- $\|y - x\|$: Some norm.
- **Subspace:** S doesn't have to be a subspace.
- $\arg\min$: Vector y that minimizes $\|x - y\|$

2.1.2 Projection onto a 1D subspace

Derivation: Projection onto a 1-dimensional subspace.
Let $S = \text{span}(\mathbf{v})$, and we denote the projection of $\mathbf{x}$ onto S as:

$$\text{Proj}_S(\mathbf{x}) = \mathbf{x}^*$$

Under the Euclidean norm (i.e. l2 norm), we have nice geometry: we should have

$$\langle \mathbf{x} - \mathbf{x}^*, \mathbf{v} \rangle = 0$$

Since $\mathbf{x}^* \in S$, $\mathbf{x}^* = \alpha \mathbf{v}$ for some scalar α.
We need to find α.
So,

$$\langle \mathbf{x} - \alpha \mathbf{v}, \mathbf{v} \rangle = 0$$

$$\Rightarrow \langle \mathbf{x}, \mathbf{v} \rangle - \alpha \langle \mathbf{v}, \mathbf{v} \rangle = 0$$

$$\Rightarrow \alpha = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}$$

Thus,

$$\mathbf{x}^* = \alpha \mathbf{v} = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v}$$

which simplifies to:

$$\mathbf{x}^* = \frac{\mathbf{x}^\top \mathbf{v}}{\|\mathbf{v}\|_2^2} \mathbf{v} = \left\langle \mathbf{x}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\rangle \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$$

- **Orthonormal Basis for S:** $\left\{ \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\}$ since $\left\| \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 = 1$
- **Projection Coefficient:** $\left\langle \mathbf{x}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\rangle$
- **Note:** x^* is the point we are looking for in the projection problem.

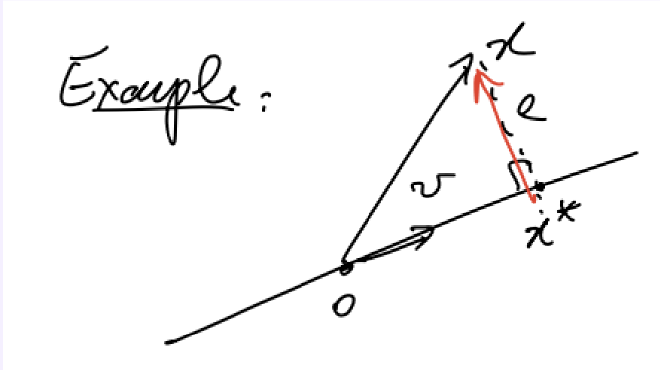

Figure 6: Visual representation of the projection problem.

2.1.3 Projection onto an n dimensional space

Derivation: Let S be a subspace of $\mathcal{V}$, and let $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ be an orthonormal basis of S.

1. **Problem setup**

$$\mathbf{x}^* = \sum_{i=1}^{d} \alpha_i \mathbf{v}_i$$

Goal: Find $\alpha_1, \ldots, \alpha_d$ so as to minimize the norm $\|\mathbf{x} - \mathbf{x}^*\|_2$.

2. **Derivation:** By geometry, we require that

$$\langle \mathbf{e}, \mathbf{v}_j \rangle = 0 \quad \forall\, j = 1, \ldots, d$$

which implies:

$$\langle \mathbf{x} - \mathbf{x}^*, \mathbf{v}_j \rangle = 0 \quad \forall\, j$$

$$\Rightarrow \langle \mathbf{x} - \sum_{i=1}^{d} \alpha_i \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \quad \forall\, j$$

Using linearity of the inner product:

$$\Rightarrow \langle \mathbf{x}, \mathbf{v}_j \rangle = \sum_{i=1}^{d} \alpha_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle$$

Since $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ if $i \neq j$ and 1 if $i = j$, this simplifies to:

$$\alpha_j = \langle \mathbf{x}, \mathbf{v}_j \rangle \quad \text{b/c only the i=j term survives}$$

Thus,

$$\mathbf{x}^* = \sum_{i=1}^{d} \alpha_i \mathbf{v}_i = \sum_{i=1}^{d} \langle \mathbf{x}, \mathbf{v}_i \rangle \mathbf{v}_i$$

3. **Solution:**

$$= \sum_{i=1}^{d} (\mathbf{x}^\top \mathbf{v}_i) \mathbf{v}_i$$

- $\mathbf{v}_i \in \mathbb{R}^n$.
- **Projection Coefficients:** $\mathbf{x}^\top \mathbf{v}_i$

4. **Example of Orthogonal Decomposition:**

$$\mathbf{e} = \mathbf{x} - \mathbf{x}^* \in S^\perp, \quad \mathbf{x}^* \in S$$

So,

$$\mathbf{x} = \mathbf{x}^* + \mathbf{e}, \quad \text{where} \quad \mathbf{x}^* \in S, \quad \mathbf{e} \in S^\perp$$

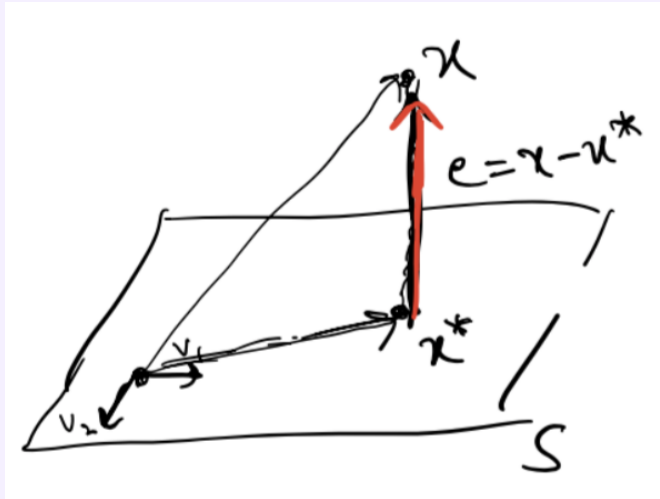

Figure 7: Generalization of projection.

2.1.4 Application of projections: Fourier series

Example: **Fourier series:**

1. Suppose we have a periodic function $x(t)$ with period T_0.

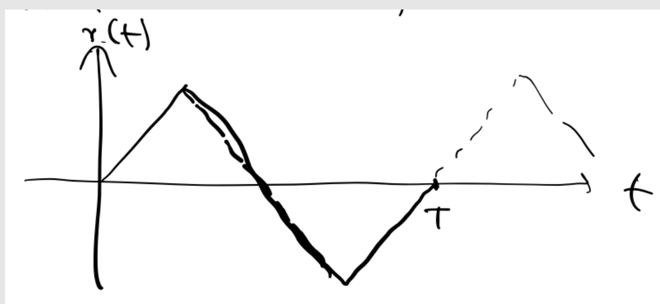

Figure 8: Periodic triangle function.

2. **Inner product for time domain (complex version):** $a_k = \langle x(t), y(t)\rangle = \dfrac{1}{T}\displaystyle\int_T x(t)\overline{y(t)}\,dt$
   - **Note:** Real version is without the conjugate.
3. **Projection (i.e. one component of the sum):** $\text{Proj}_{\underline{v}_i}(\underline{x}) = \langle \underline{x}, \underline{v}_i\rangle \underline{v}_i$
4. **Goal:** Express $x(t)$ (i.e. any periodic function) as a sum of complex exponentials:

$$x^*(t) = \sum_{k=-\infty}^{\infty} a_k e^{jk\omega_0 t}$$

   - **Projection:** $\text{Proj}_{e^{jk\omega_0 t}}(x(t)) = \langle x(t), \exp(jk\omega_0 t)\rangle\, e^{jk\omega_0 t} = a_k e^{jk\omega_0 t}$ for a certain value of k.
   - **Projection coefficient:** $a_k = \left\langle x(t), e^{jk\omega_0 t}\right\rangle = \dfrac{1}{T_0}\displaystyle\int_0^T x(t)e^{-jk\omega_0 t}dt$
   - **Fundamental frequency:** $\omega_0 = \dfrac{2\pi}{T_0}$.
5. **Prove orthonormal basis for the complex exponentials:** To prove it's a orthnormal basis, must prove it has unit norm 1 and each pair of vectors are orthogonal (i.e. inner product is 0).
   (a) **Magnitude of exp:** $\left|e^{j\theta}\right| = 1$. Therefore, it has unit norm.

(b) **Orthogonality:**

$$\langle e^{ji\omega_0 t}, e^{jl\omega_0 t}\rangle = \begin{cases} 1, & i = l \\ 0, & i \neq l \end{cases}$$

Therefore, for each pair of basis vectors, they are orthogonal.

- **Conjugate of exp:** $(e^{j\theta}) = e^{-j\theta}$

6. **Conclusion:** Fourier series is a projection of a function onto the set of othonormal basis functions $\exp(jk\omega_0 t)$, where k is an integer.
   - **Optimal:** This projection is optimal as it minimizes the approximation error $\|x(t) - x^*(t)\|$, i.e.

$$\frac{1}{T}\int_0^T (x(t) - x^*(t))^2\, dt$$

As the number of terms in the summation increases to infinity, the error goes to 0.

2.2 Gram-Schmidt and QR decomposition

2.2.1 What if the set of basis vectors is not orthonormal?

Derivation: Let $\{u^{(1)}, \ldots, u^{(d)}\}$ be a set of basis vectors for a subspace $S($ not necessarily orthonormal)

We can still use the orthogonality principle, i.e.,

$$e = x - x^* \perp S$$

Therefore,

$$\langle x - x^*, u^{(j)}\rangle = 0 \quad \forall j = 1, \ldots, d$$

Also, $x^* \in S$ so x^*can be written as a linear combination of basis vectors, so $x^* = \sum_{i=1}^{d} \alpha_i u^{(i)}$

Need to find $\alpha_1, \ldots, \alpha_d$ s.t.

$$\langle x - \sum_{i=1}^{d} \alpha_i u^{(i)}, u^{(j)}\rangle = 0 \quad \forall j = 1, \ldots, d$$

$$\Rightarrow \langle x, u^{(j)}\rangle = \sum_{i=1}^{d} \alpha_i \langle u^{(i)}, u^{(j)}\rangle \quad \forall j = 1, \ldots, d$$

$$\begin{bmatrix} \langle u^{(1)}, u^{(1)}\rangle & \langle u^{(2)}, u^{(1)}\rangle & \ldots & \langle u^{(d)}, u^{(1)}\rangle \\ \langle u^{(1)}, u^{(2)}\rangle & \langle u^{(2)}, u^{(2)}\rangle & \ldots & \langle u^{(d)}, u^{(2)}\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u^{(1)}, u^{(d)}\rangle & \langle u^{(2)}, u^{(d)}\rangle & \ldots & \langle u^{(d)}, u^{(d)}\rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_d \end{bmatrix} = \begin{bmatrix} \langle x, u^{(1)}\rangle \\ \langle x, u^{(2)}\rangle \\ \vdots \\ \langle x, u^{(d)}\rangle \end{bmatrix}$$

Solve for $\alpha_1, \ldots, \alpha_d$, Then, we get $x^* = \sum_{i=1}^{d} \alpha_i u^{(i)}$

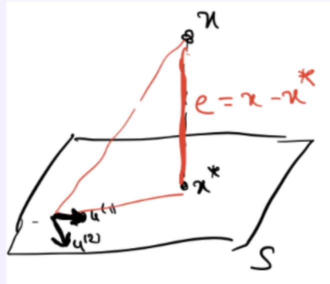

Figure 9: Not orthogonal, but similar to projection with orthonormal basis.

- **Note:** If $\{u^{(1)}, \dots, u^{(d)}\}$ is an orthonormal basis, then the matrix is the identity matrix, and we get $\alpha_j = \langle x, u^{(j)} \rangle$ as before.

Example: **Function approximation.** Let B be the set of basis functions that is not orthonormal:

$$\mathcal{B} = \{1, t, \dots, t^d\}$$

Let $x(t)$ be a function over $[0, 1]$.

- **1st Goal** Approximate $x(t)$ by $x^*(t) = \sum_{n=0}^{d} \alpha_n t^n$
- To find $\alpha_0, \alpha_1, \dots, \alpha_d$, need to solve the $Ax = b$.
- **2nd Goal:** Minimize the approximation error $\|x(t) - x^*(t)\|_2 = \left(\int_0^1 (x(t) - x^*(t))^2 dt \right)^{1/2}$

Recall: Taylor series expansion

$$x(t) \approx x(0) + x'(0)t + \frac{x''(0)}{2} t^2 + \dots$$

- Taylor series expansion is completely different from the projection method, and the reason is that Taylor series expansion is a local approximation.

2.2.2 Gram-Schmidt Procedure

Motivation: This is to get an orthonormal basis, so we can use the easier projection method.

Intuition: Another way to find the projection of x onto $S = \text{span}\{u^{(1)}, \dots, u^{(d)}\}$ is to first find an orthonormal basis of S, and then the projection problem becomes easier.

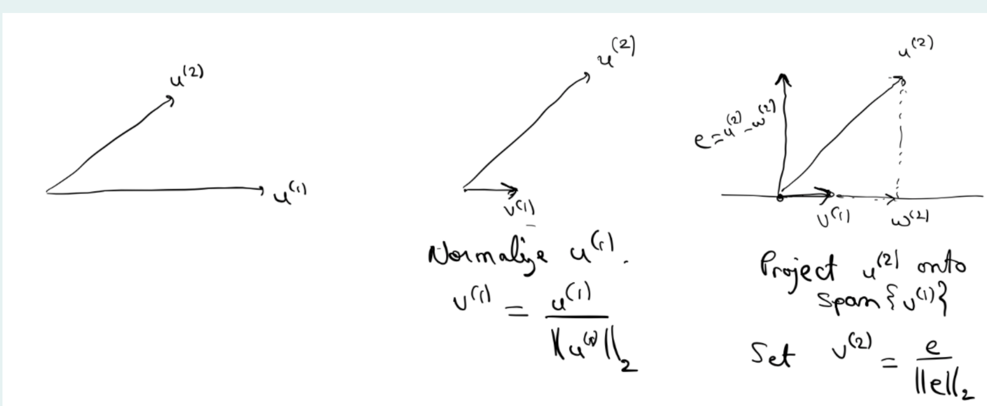

Figure 10: Gram-Schmidt Process for 2D.

1. Normalize $u^{(1)}$
2. Find the error vector by projecting $u^{(2)}$ onto the subspace $v^{(1)}$.
3. Normalize the error vector.
4. Now you have two vectors that form an orthonormal basis in 2D.

Definition: Turns any set of basis vectors of a subspace into an **orthonormal** set of basis vectors.

Process:
1. Normalize $u^{(1)}$ to get $v^{(1)}$:
$$v^{(1)} = \frac{u^{(1)}}{\|u^{(1)}\|_2}$$
2. (a) Project $u^{(2)}$ onto $S = \text{span}\{v^{(1)}\}$ to get:
$$w^{(2)} = \langle u^{(2)}, v^{(1)} \rangle v^{(1)}$$
   (b) Set:
$$v^{(2)} = \frac{u^{(2)} - w^{(2)}}{\|u^{(2)} - w^{(2)}\|_2}$$
3. Continue similarly:
   (a) Project $u^{(3)}$ onto $S = \text{span}\{v^{(1)}, v^{(2)}\}$ to get:
$$w^{(3)} = \langle u^{(3)}, v^{(1)} \rangle v^{(1)} + \langle u^{(3)}, v^{(2)} \rangle v^{(2)}$$
   (b) Set:
$$v^{(3)} = \frac{u^{(3)} - w^{(3)}}{\|u^{(3)} - w^{(3)}\|_2}$$
4. Continue this process for higher dimensions. Therefore, $\{v^{(1)}, \dots, v^{(d)}\}$ is an orthonormal basis for $\text{span}\{u^{(1)}, \dots, u^{(d)}\}$.

Intuition:
- Create the Gram matrix by making an orthonormal basis.
- Then this terms the matrix into the identity.

Warning: Not every set of vectors in $\mathbb{R}^n$ turns into the standard basis. It depends on how you apply the Gram-Schmidt.
- For example if you have the vectors $(1, 0)$ and $(1, 1)$.
- If you apply Gram-Schmidt starting with $(1, 0)$, then you will get the standard basis.
- But if you apply Gram-Schmidt starting with $(1, 1)$, then you will get $(1/\sqrt{2}, 1/\sqrt{2}), (1/\sqrt{2}, 1/\sqrt{2})$.

2.2.3 QR decomposition

Another way to see Gram-Schmidt procedure is through matrix multiplication.

Definition: Stack all $u^{(i)}$ vectors as columns of a matrix

$$\begin{aligned}
\begin{bmatrix} u^{(1)} & \cdots & u^{(d)} \end{bmatrix} &= QR \\
\begin{bmatrix} u^{(1)} & \cdots & u^{(d)} \end{bmatrix} &= \begin{bmatrix} v^{(1)} & \cdots & v^{(d)} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1d} \\ 0 & r_{22} & \cdots & r_{2d} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & r_{dd} \end{bmatrix} \\
&= \begin{bmatrix} r_{11}v^{(1)} & r_{12}v^{(1)} + r_{22}v^{(2)} & \cdots \end{bmatrix}
\end{aligned}$$

- Q: Orthonomral matrix (i.e., its columns are orthogonal to each other and have unit norm)
- R: Upper triangular.

Intuition: For $Ax = b$, therefore,

$$QRx = b$$

- Since Q has columns of orthonormal basis, then it has an inverse, which is $Q^{-1} = Q^T$.

Then,

$$Rx = Q^T b$$

Example:

$$\{1, t, t^2, \cdots, t^d\}$$

is *not an orthonormal basis*, which as an example is defined from $[0, 1]$

The L^2-norm for this example is given by

$$\|f\|_2 = \left(\int_0^1 f^2(t)\, dt\right)^{\frac{1}{2}}.$$

The inner product between two functions $f(t)$ and $g(t)$ is defined as:

$$\langle f, g\rangle = \int_0^1 f(t)g(t)\, dt.$$

1. Start with $u^{(1)} = 1$, which is equivalent to $v^{(1)}$ because it's unit norm.
2. For $u^{(2)}$, calculate the projection:

$$\omega^{(2)} = \text{Proj}_{\text{span}\{v^{(1)}\}} u^{(2)} = \langle u^{(2)}, v^{(1)}\rangle = \int_0^1 t \cdot 1\, dt = \frac{1}{2}..$$

   So, the projection of $u^{(2)}$ onto $u^{(1)}$ is:

$$\frac{1}{2} v^{(1)}.$$

3. Now subtract the projection from $u^{(2)}$ and normalize:

$$v^{(2)} = \frac{u^{(2)} - \omega^{(2)}}{\|u^{(2)} - \omega^{(2)}\|_2} = \frac{t - \frac{1}{2}}{\left(\int_0^1 \left(t - \frac{1}{2}\right)^2 dt\right)^{\frac{1}{2}}}.$$

2.3 Projection of a subspace defined by its orthogonal vectors

2.3.1 Subspace defined by its orthogonal vectors

Intuition:

1. So far, we have defined a subspace by its basis vectors:

$$S = \text{span}\{v^{(1)}, \ldots, v^{(d)}\}.$$

2. But, in many cases, we can define S in terms of the set of vectors that are orthogonal to it.

Definition:

$$S = \left\{x \mid \left(a^{(i)}\right)^T x = 0,\, i = 1, \ldots, m\right\}$$

then the vectors $a^{(1)}, \dots, a^{(m)}$ are orthogonal to all vectors in S (i.e. the inner products are 0 for all vectors x with $a^{(i)}$). Therefore, $S^\perp = \text{span}\{a^{(1)}, \dots, a^{(m)}\}$.

2.3.2 Projection

Derivation:

1. Projecting a vector x onto a subspace S spanned by the vectors $\{a^{(1)}, \dots, a^{(m)}\}$. The projection x^* is given by:
$$x^* = \text{Proj}_S(x) = \arg\min_{y \in S} \|x - y\|_2$$

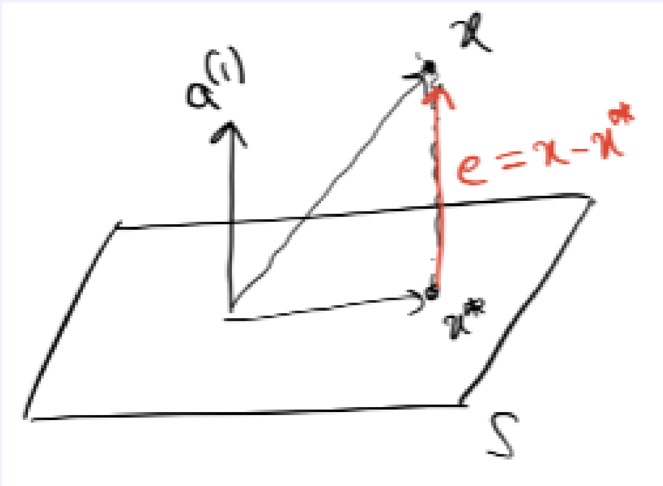

Figure 11: Projection onto a subspace defined by its orthogonal vectors

2. Using the orthogonality principle, the error $e = x - x^*$ must be orthogonal to the subspace S, i.e.,
$$e \perp S$$
This implies that:
$$e \in \text{span}\{a^{(1)}, \dots, a^{(m)}\}$$
3. The error can be written as a linear combination of the basis vectors:
$$e = x - x^* = \sum_{i=1}^{m} \beta_i a^{(i)}$$
We need to find the coefficients $\beta_1, \dots, \beta_m$.
4. Since $x^* \in S$, we have the condition:
$$\langle x^*, a^{(j)} \rangle = 0 \quad \forall j = 1, \dots, m$$
which leads to the following equation:
$$(a^{(j)})^T x^* = 0 \quad \forall j = 1, \dots, m$$
5. Substituting $x^* = x - \sum_{i=1}^{m} \beta_i a^{(i)}$ into the above equation, we get:
$$(a^{(j)})^T \left(x - \sum_{i=1}^{m} \beta_i a^{(i)} \right) = 0 \quad \forall j$$

6. Expanding the terms using linearity in the first argument for inner products:

$$(a^{(j)})^T x = \sum_{i=1}^{m} \beta_i (a^{(j)})^T a^{(i)}$$

This system of equations can be written in matrix form as:

$$\begin{bmatrix} (a^{(1)})^T a^{(1)} & \cdots & (a^{(1)})^T a^{(m)} \\ \vdots & \ddots & \vdots \\ (a^{(m)})^T a^{(1)} & \cdots & (a^{(m)})^T a^{(m)} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} (a^{(1)})^T x \\ \vdots \\ (a^{(m)})^T x \end{bmatrix}$$

We can solve this system of linear equations to obtain the values of $\beta_1, \ldots, \beta_m$.

7. Once we have the values of β_i, we can compute the projection as:

$$x^* = x - \sum_{i=1}^{m} \beta_i a^{(i)}$$

8. **Note:** If the set $\{a^{(i)}\}$ is orthonormal, the matrix on the left-hand side becomes the identity matrix I, and the coefficients simplify to:

$$\beta_j = (a^{(j)})^T x = \langle x, a^{(j)} \rangle$$

2.4 Projection onto affine sets

2.4.1 Affine spaces

Definition: An affine space (or affine set) is a translation (or shift) of a subspace S.

Example: Consider a vector $x^{(0)}$ (not necessarily in S). The affine space $\mathcal{A}$ is defined as:

$$\mathcal{A} = \{u + x^{(0)} \mid u \in S\}$$

where $x^{(0)}$ is the shifted vector and S is the original subspace. This represents a shifted version of the subspace S.

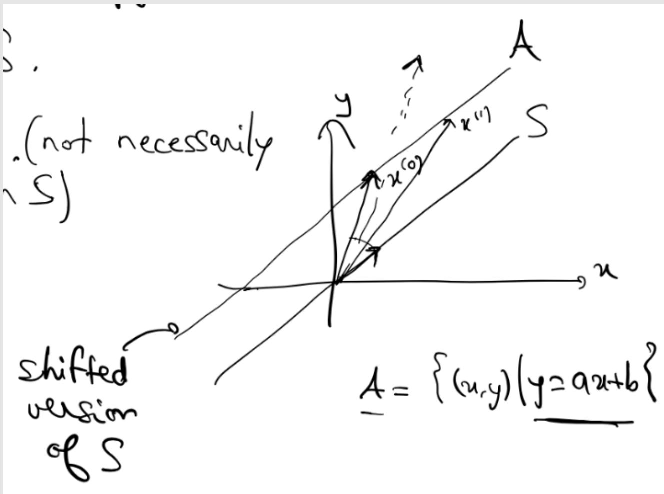

Figure 12: Affine space of a 2D space, where the $x^{(0)}$ is the constant (i.e. shifting origin to the Affine set) that we are adding to shift all the vectors to the affine space.

2.4.2 Projection of Affine space defined in terms of basis vectors of corresponding subspace

Derivation:

1. The affine space is described by:
$$\mathcal{A} = \left\{ x \mid x = \sum_{i=1}^{d} \alpha_i v^{(i)} + c \right\}$$
   - $\{v^{(1)}, \ldots, v^{(d)}\}$: Basis vectors of the subspace S
   - c: Vector (i.e. shift).

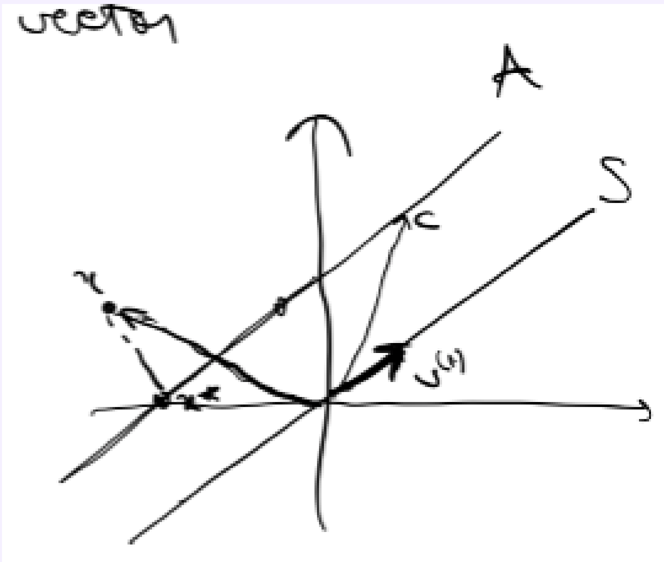

Figure 13: Projection problem visualization, where we are projecting the vector onto the Affine space.

2. Using the orthogonality principle, we must have:
$$\langle x - x^*, v^{(j)} \rangle = 0 \quad \forall j = 1, \ldots, d$$
where $x^* \in \mathcal{A}$. Therefore:
$$x^* = \sum_{i=1}^{d} \alpha_i v^{(i)} + c$$
3. This leads to the condition:
$$\left\langle x - \sum_{i=1}^{d} \alpha_i v^{(i)} - c, v^{(j)} \right\rangle = 0 \quad \forall j = 1, \ldots, d$$
4. Simplifying this expression using the linearity in first argument for inner product, we obtain:
$$\langle x - c, v^{(j)} \rangle = \sum_{i=1}^{d} \alpha_i \langle v^{(i)}, v^{(j)} \rangle \quad \forall j = 1, \ldots, d$$
5. To solve for $\alpha_1, \ldots, \alpha_d$, we set up the following system of linear equations in matrix form:
$$\begin{bmatrix} \langle v^{(1)}, v^{(1)} \rangle & \cdots & \langle v^{(1)}, v^{(d)} \rangle \\ \vdots & \ddots & \vdots \\ \langle v^{(d)}, v^{(1)} \rangle & \cdots & \langle v^{(d)}, v^{(d)} \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{bmatrix} = \begin{bmatrix} \langle x - c, v^{(1)} \rangle \\ \vdots \\ \langle x - c, v^{(d)} \rangle \end{bmatrix}$$
   - **Note:** We are projecting onto $x - c$, which we can see on the RS, which is the subspace.
6. Solving this system gives us the values for $\alpha_1, \ldots, \alpha_d$. Finally, the projection x^* onto the affine space $\mathcal{A}$ is:
$$x^* = \sum_{i=1}^{d} \alpha_i v^{(i)} + c$$
Note: We are projecting onto $x - c$ (i.e. subspace), then adding the shift into the end to be back on the Affine set.

2.4.3 Projection of Affine space defined in terms of orthogonal vectors to corresponding subspace

Derivation:

1. The affine set $\mathcal{A}$ is defined as:
$$\mathcal{A} = \left\{ x \mid \langle x, a^{(i)} \rangle = d_i,\ i = 1, \ldots, m \right\}$$
   - d_i: Scalars
   - $\{a^{(1)}, \ldots, a^{(m)}\}$: A set of vectors spanning the affine space. (Check why this is equivalent to the previous definition of an affine set.)

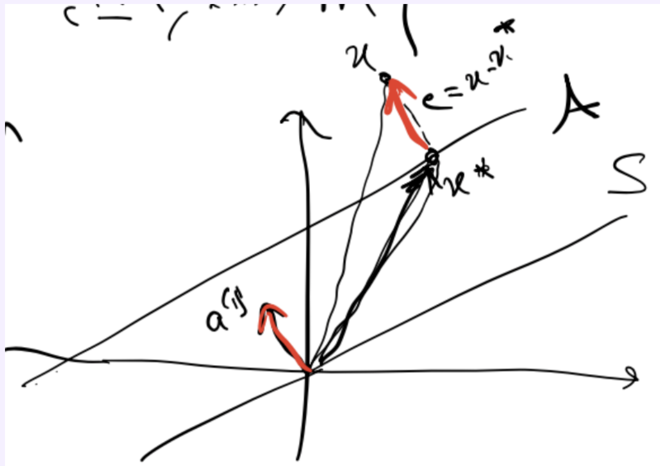

Figure 14: Projection problem visualization, where we are projecting the vector onto the affine space, which is in line with the orthogonal vectors.

2. Since $x - x^*$ lies in the span of $\{a^{(1)}, \ldots, a^{(m)}\}$:
$$x - x^* = \sum_{i=1}^{m} \beta_i a^{(i)}$$
where $\beta_1, \ldots, \beta_m$ are the coefficients to be determined.
3. Since $x^* \in \mathcal{A}$, we also have:
$$\langle x^*, a^{(j)} \rangle = d_j \quad \forall j = 1, \ldots, m$$
This implies the orthogonality condition for the projection:
$$\langle x - \sum_{i=1}^{m} \beta_i a^{(i)}, a^{(j)} \rangle = d_j \quad \forall j = 1, \ldots, m$$
4. Expanding the above expression using the linearity in first argument for inner product, we get:
$$\langle x, a^{(j)} \rangle - \sum_{i=1}^{m} \beta_i \langle a^{(i)}, a^{(j)} \rangle = d_j \quad \forall j = 1, \ldots, m$$
5. This leads to the system of linear equations:
$$\langle x, a^{(j)} \rangle - d_j = \sum_{i=1}^{m} \beta_i \langle a^{(i)}, a^{(j)} \rangle \quad \forall j$$
6. We now solve this system of linear equations for the coefficients $\beta_1, \ldots, \beta_m$. The system can be written in matrix form as:
$$\begin{bmatrix} \langle a^{(1)}, a^{(1)} \rangle & \cdots & \langle a^{(1)}, a^{(m)} \rangle \\ \vdots & \ddots & \vdots \\ \langle a^{(m)}, a^{(1)} \rangle & \cdots & \langle a^{(m)}, a^{(m)} \rangle \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} \langle x, a^{(1)} \rangle - d_1 \\ \vdots \\ \langle x, a^{(m)} \rangle - d_m \end{bmatrix}$$

7. Solving this system gives the values for $\beta_1, \ldots, \beta_m$. Once the β_i values are known, the projection x^* is given by:

$$x^* = x - \sum_{i=1}^{m} \beta_i a^{(i)}$$

- **Intuition:** This is subtracting the orthogonal components of x (i.e. removing the error vector) to get x in the subspace.

Example:

1. Consider the case where $m = 1$. The affine set $\mathcal{A}$ is defined as:

$$\mathcal{A} = \{x \mid a^T x = d\}$$

where a is a vector and d is a scalar.

2. To project x onto the affine subspace, we start by using the orthogonality condition:

$$\langle x, a\rangle - d = \beta\langle a, a\rangle$$

This ensures that the difference between x and its projection x^* lies in the direction of a.

3. Solving for β, we get:

$$\beta = \frac{\langle x, a\rangle - d}{\langle a, a\rangle} = \frac{a^T x - d}{\|a\|_2^2}$$

This provides the scalar β, which tells us how much of the vector a needs to be subtracted from x.

4. The projection x^* onto the affine subspace is then:

$$x^* = x - \beta a = x - \left(\frac{a^T x - d}{\|a\|_2^2}\right) a$$

This gives the final expression for the projection of x onto the affine set $\mathcal{A}$.

2.4.4 Show that the two affine sets are equal

Derivation: We define set A as follows:

$$A = \left\{ x \in \mathbb{R}^n \;\middle|\; x = \sum_{i=1}^{d} \alpha_i v^{(i)} + c \right\}$$

where $v^{(i)}$ are the basis vectors, α_i are scalar coefficients, and c is a fixed vector (the translation vector of the affine set).

Now, we define set B as:

$$B = \left\{ x \in \mathbb{R}^n \;\middle|\; a^{(i)T} x = d_i,\ i = 1, 2, \ldots, m \right\}$$

where $a^{(i)}$ are orthogonal vectors, and d_i are scalars defining the affine constraints.

Step 1: Show $A \subseteq B$.

Assume $x \in A$, then we can write:

$$x = \sum_{i=1}^{d} \alpha_i v^{(i)} + c$$

Substitute this into the condition for set B:

$$a^{(i)T} x = a^{(i)T} \left(\sum_{j=1}^{d} \alpha_j v^{(j)} + c \right)$$

Expanding the expression:

$$a^{(i)T}x = \sum_{j=1}^{d} \alpha_j a^{(i)T} v^{(j)} + a^{(i)T} c$$

Since the vectors $a^{(i)}$ are orthogonal to the vectors $v^{(j)}$, we have:

$$a^{(i)T} v^{(j)} = 0 \quad \text{for all } i, j$$

Therefore, the equation simplifies to:

$$a^{(i)T} x = a^{(i)T} c$$

We define $d_i = a^{(i)T} c$. Hence,

$$a^{(i)T} x = d_i \quad \text{for all } i$$

Thus, $x \in B$.
Step 2: Show $B \subseteq A$.
Assume $x \in B$, then we know:

$$a^{(i)T} x = d_i \quad \text{for all } i = 1, 2, \ldots, m$$

This implies that the vector x satisfies all the affine constraints defined by the vectors $a^{(i)}$ and scalars d_i. Now, consider the vector c such that:

$$a^{(i)T} c = d_i$$

Subtracting this from the affine constraint for x, we get:

$$a^{(i)T} (x - c) = 0 \quad \text{for all } i$$

This shows that $x - c$ lies in the null space of the vectors $a^{(i)}$. Therefore, $x - c$ must lie in the span of the vectors $v^{(i)}$, meaning:

$$x - c = \sum_{i=1}^{d} \alpha_i v^{(i)}$$

for some scalars $\alpha_1, \alpha_2, \ldots, \alpha_d$. Thus,

$$x = \sum_{i=1}^{d} \alpha_i v^{(i)} + c$$

Therefore, $x \in A$.
Conclusion:
Since we have shown that $A \subseteq B$ and $B \subseteq A$, we conclude that:

$$A = B$$

This proves that the definition of the affine set in terms of orthogonal vectors and the definition in terms of basis vectors are equivalent.

2.5 Summary

Definition:
Subspace:

$$S = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \sum_{i=1}^{d} \alpha_i \mathbf{v}^{(i)} \right\}$$

$$S = \left\{ \mathbf{x} \;\middle|\; (\mathbf{a}^{(i)})^\top \mathbf{x} = 0,\ i = 1, \ldots, m \right\}$$

Affine Space:

$$A = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \sum_{i=1}^{d} \alpha_i \mathbf{v}^{(i)} + \mathbf{c} \right\}$$

$$A = \left\{ \mathbf{x} \;\middle|\; (\mathbf{a}^{(i)})^\top \mathbf{x} = d_i,\ i = 1, \ldots, m \right\}$$

Process: You have three types of Gram Matrix that you can have, which can all be used to solve the projection problems.

- If you have a set of vectors that are linearly independent, then Gram matrix (i.e. the matrix in every projection problem) is invertible.
- If Gram matrix is orthogonal, then Gram Matrix is diagonal matrix (i.e. identity scaled by some factor)
- If Gram matrix is orthonormal, you have an identity matrix.

2.6 Hyperplanes and half-spaces

Definition:

- **Hyperplane:** an affine space for the special case $m = 1$.

$$\mathcal{H} = \left\{ \mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b \right\} \tag{8}$$

- **Half-space:**

$$\mathcal{H}_+ = \left\{ \mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \geq b \right\} \tag{9}$$

$$\mathcal{H}_- = \left\{ \mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \leq b \right\} \tag{10}$$

  - $\mathbf{x} \in \mathcal{H}_+ \implies$ angle between $\mathbf{a}$ and $\mathbf{x}$ is acute.
  - $\mathbf{x} \in \mathcal{H}_- \implies$ angle between $\mathbf{a}$ and $\mathbf{x}$ is obtuse.

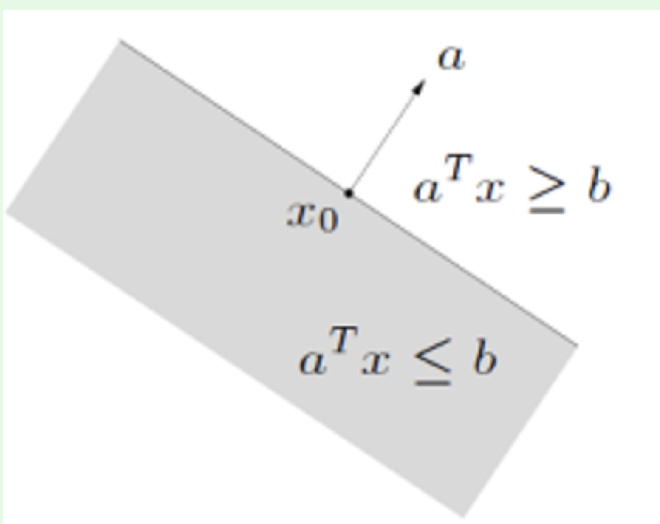

Figure 15: 2D Hyperplane which is the line then the half space, which is separating the line into the above and below b.

Intuition:

- In 2D: A hyperplane is a straight line dividing the plane into two regions.
- In 2D: A half-space is the region on one side of a line, which describes all points on one side of the line.
- In 3D: A hyperplane is a flat plane dividing the space into two regions.
- In 3D: A half-space is the region on one side of a plane, which includes all points on one side of the plane.

3 Problem set 1

4 Non-Euclidean Projection, Functions, Gradients and Hessians (Ch. 2.3-2.4)

4.1 Non-Euclidean projection

Intuition:

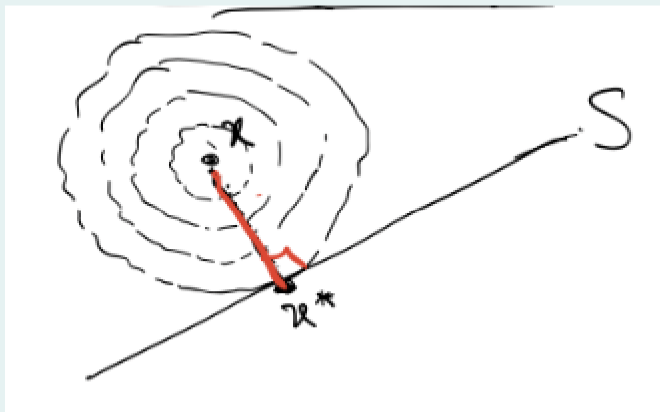

Figure 16: l2 norm

- **Key:** $\mathbf{x}^* = \arg\min_{\mathbf{y}\in\mathcal{S}} \|\mathbf{x}-\mathbf{y}\|_2$ has the orthogonality principle holds (i.e. $x - x^* \perp S$)

The projection problem is well-defined in the case of $l1$-norm or l_∞-norm, but these norms have **no associated notion of angle and orthogonality**. The following plots show that the vector x is not orthogonal to x^*.

- **Note:** To get these projections, this must be done through linear programming.

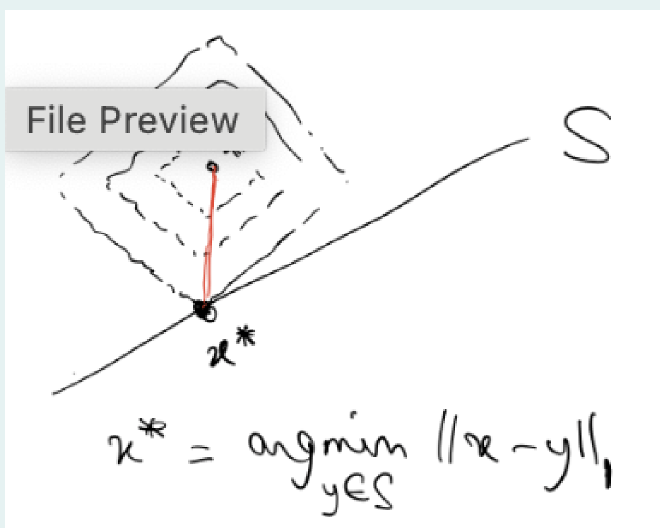

Figure 17: l1 norm

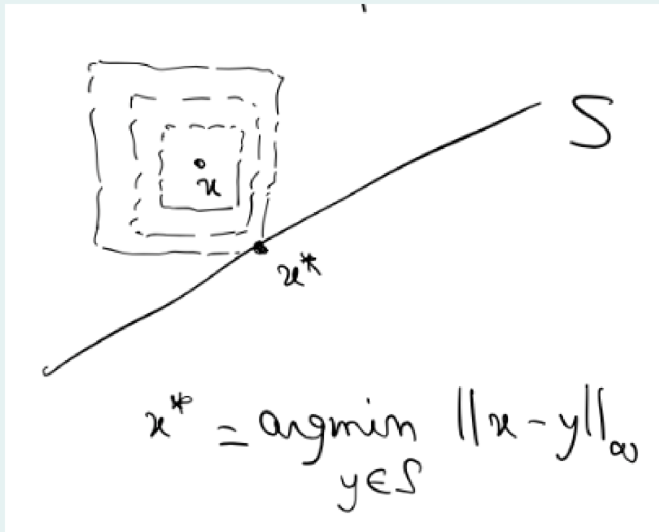

Figure 18: l-infinity norm

4.2 Functions

Definition: A function is a mapping from a set (domain) to another set (range).

Example: Projection function

$$x^* = \arg\min_{y \in \mathcal{H}} \|x - y\| = \text{Proj}_{\mathcal{H}}(x)$$

If $\mathcal{H}$ is linear (or convex), then x^* is unique.

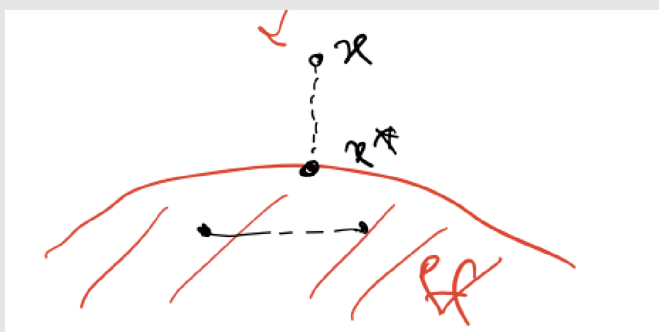

Figure 19: Function: Convex projection because we are projecting one input and getting one output

If $\mathcal{H}$ is non-convex, the projection might not be unique and the function is not well-defined (i.e. not a function).

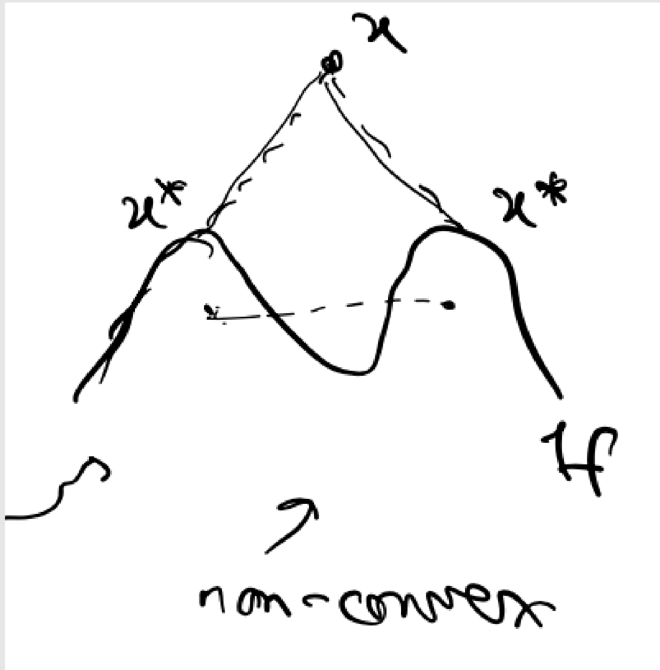

Figure 20: Not a function: Non Convex projection because we are projecting one input and getting two outputs

4.2.1 Terminology of functions

Definition:
- **Map**: $f : \mathbb{R}^n \to \mathbb{R}^m$, e.g. $f(x) = Ax$
- **Operator**: $f : \mathbb{R}^n \to \mathbb{R}^n$, e.g. $f(x) = e^x$
- **Functional**: $f : \mathbb{R}^n \to \mathbb{R}$, e.g. $f(x) = ||x||_p$

4.2.2 Common Sets

Definition: Given a functional $f : \mathbb{R}^n \to \mathbb{R}$:
- **Graph**: $\{(x, f(x)) \in \mathbb{R}^{n+1} | x \in \mathbb{R}^n\} \subseteq \mathbb{R}^{n+1}$
- **Epigraph (on or above)**: $\{(x, t) \in \mathbb{R}^{n+1} | x \in \mathbb{R}^n, f(x) \leq t\} \subseteq \mathbb{R}^{n+1}$
- **Level set**: $L_t = \{x \in \mathbb{R}^n | f(x) = t\} \subseteq \mathbb{R}^n$
  - Parametrized by t (i.e. scalar).
- **Sublevel set**: $\{x \in \mathbb{R}^n | f(x) \leq t\} \subseteq \mathbb{R}^n$

Example:

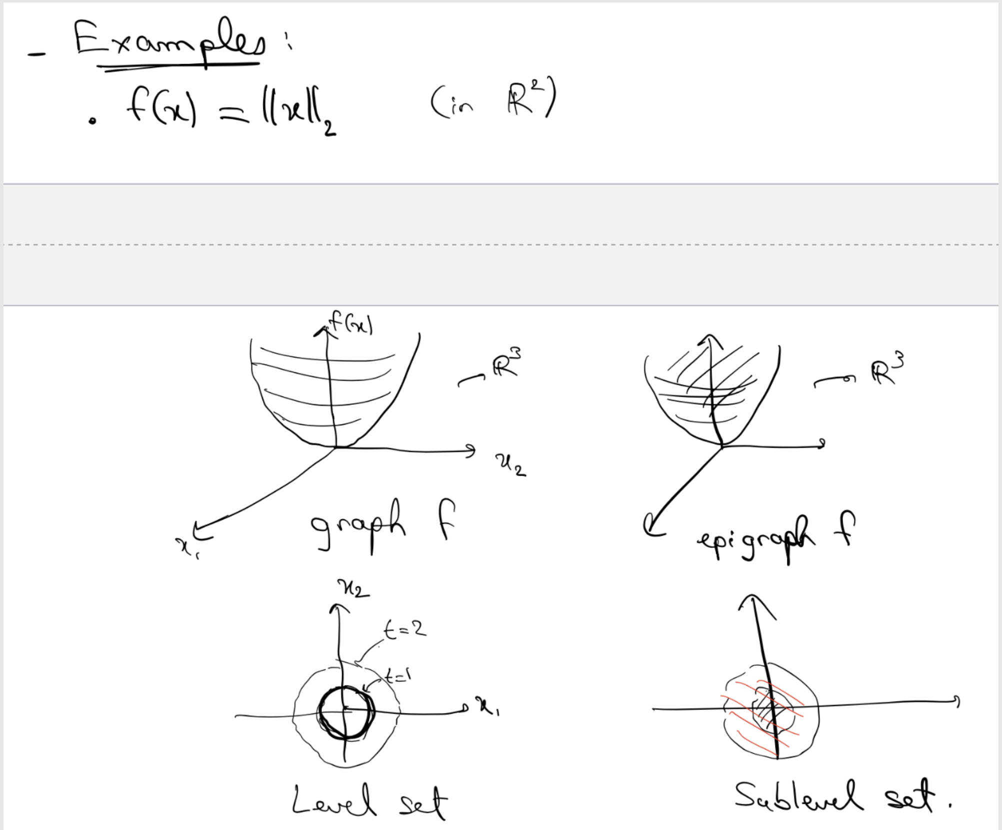

Figure 21: Example of the different common sets.

- The graph is the cup.
- The epigraph is the cup and every inside the cup (i.e. volume)
- The level set is the cross section which is projected onto the x1 and x2 axis. (i.e. where $f(x) = t$)
- The sublevel set is the level set and everything inside of it.

Example:

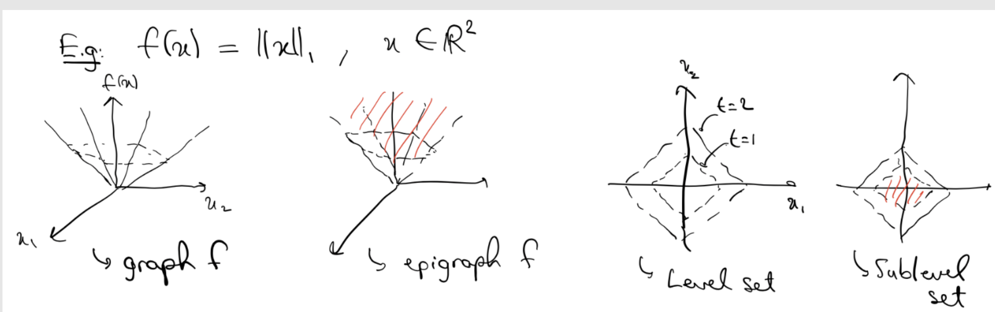

Figure 22: Example of the different common sets for the l1 norm.

Example:

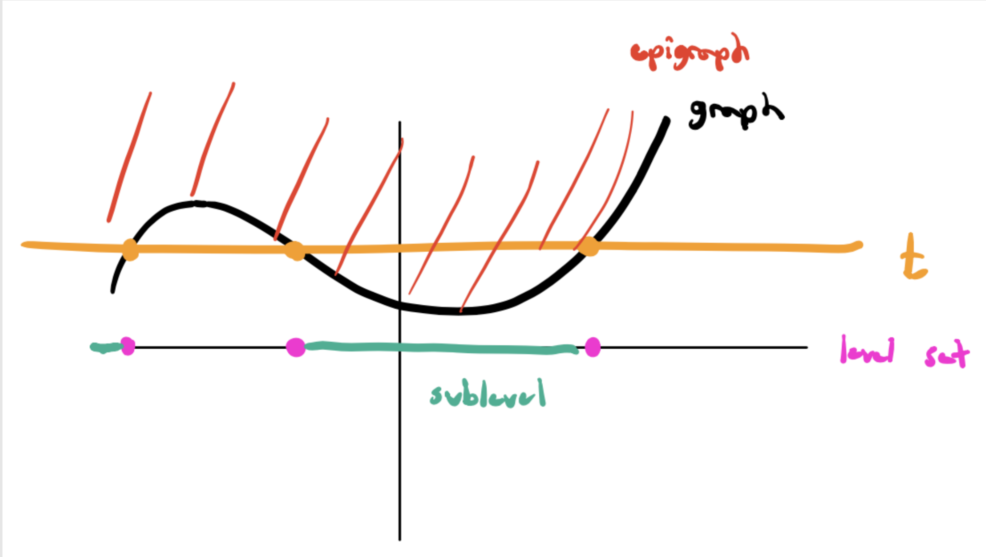

Figure 23: Example of a 2D function.

4.2.3 Linear functions

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear if:

- **Homogeneity**: $f(\alpha x) = \alpha f(x) \quad \forall x \in \mathbb{R}^n, \alpha \in \mathbb{R}$.
- **Additivity**: $f(x^{(1)} + x^{(2)}) = f(x^{(1)}) + f(x^{(2)}) \quad \forall x^{(1)}, x^{(2)} \in \mathbb{R}^n$.

Together, these two properties imply that

$$f\left(\sum_{i=1}^{d} \alpha_i x^{(i)}\right) = \sum_{i=1}^{d} \alpha_i f(x^{(i)})$$

- **Subspace**: A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if there exists a matrix $A \in \mathbb{R}^{m\times n}$ such that $f(x) = Ax$.
- **Affine translation:** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **affine** if and only if there exist a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^m$ such that: $f(x) = Ax + b$
- **Special Case** For $m = 1$,

$$f(x) = a^\top x + b \quad \to \text{affine function}$$

$$f(x) = a^\top x \quad \to \text{linear function}$$

Example:

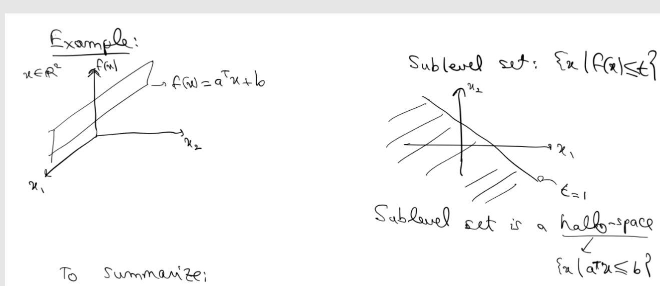

Figure 24: Example 2 of using the linear function, where the sublevel sets are the half-spaces.

Definition:
- If $f : \mathbb{R}^n \to \mathbb{R}$ is **linear**, then the graph of f is a subspace.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is **affine**, then the graph of f is a hyperplane.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is linear or affine, then the sublevel sets are half-spaces.

4.3 Function approximations

Motivation: 1st and 2nd order functions are useful to prove convex and non-convex functions.

4.3.1 Gradients

Definition: The gradient ∇f of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n \tag{11}$$

4.3.2 First-Order Approximation

Definition: For $f : \mathbb{R}^n \to \mathbb{R}$, the first-order approximation around $\bar{x}$ is:

$$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x}) \tag{12}$$

- **Recall**: The first-order approximation of $f : \mathbb{R} \to \mathbb{R}$ using the Taylor series at the point $\bar{x}$ is

$$f(x) \approx f(\bar{x}) + \left.\frac{\partial f}{\partial x}\right|_{x=\bar{x}} \cdot (x - \bar{x})$$

- **Notation:** $\nabla f(\bar{x}) = \nabla f(x)\Big|_{x=\bar{x}}$

Example:

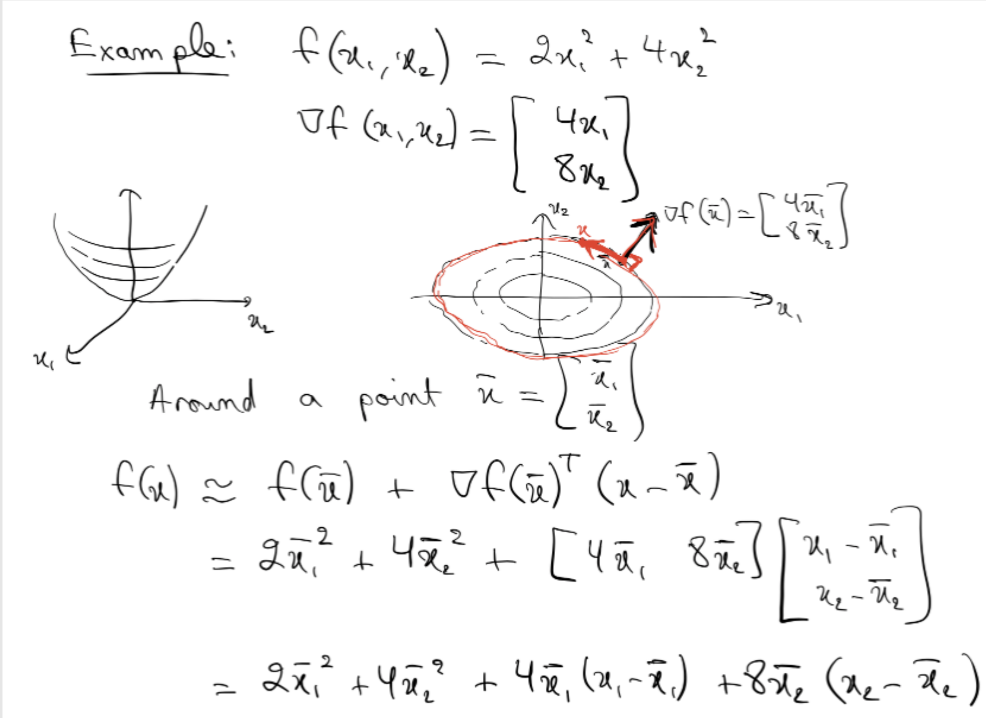

Figure 25: Example of using function approximation for first order. The arrow denotes $x - x^*$ along the level curve.

4.3.3 Gradient properties

Intuition:

- The gradient points in the direction where the function is increasing.
- **Why is the gradient orthogonal to the level sets?**
  - **Explanation:** Level sets are defined as: $\{x \mid f(x) = t\}$
    But by first order approximation:

    $$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x})$$

    So we want to find which direction of the gradient makes $f(x)$ constant (i.e. be on the level set), so we want to go in the direction such that $f(x) = f(\bar{x})$, therefore,

    $$\nabla f(\bar{x})^\top (x - \bar{x}) = 0$$

    Since we want the first order approximation to be $f(x) = f(\bar{x})$. This means that the gradient is orthogonal to $x - x^*$, which is on the level set. This means that the gradient is orthogonal to the level sets.
- **What about the norm of $\nabla f(x)$?** The norm of $\nabla f(x)$ gives an indication of the steepness of the graph.
  - **Explanation?**

    $$\begin{aligned} f(x) - f(\bar{x}) &\approx (\nabla f(\bar{x}))^\top (x - \bar{x}) \\ &= \|\nabla f(\bar{x})\|_2 \|x - \bar{x}\|_2 \left\langle \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}, \frac{x - \bar{x}}{\|x - \bar{x}\|_2} \right\rangle \\ &= \|\nabla f(\bar{x})\|_2 \|x - \bar{x}\|_2 \cos\theta \end{aligned}$$

    * θ is the angle between $\nabla f(x)$ and $(x - \bar{x})$.
    * **Key:** The larger $\|\nabla f(\bar{x})\|_2$, the larger the change in $f(x)$ (i.e. $f(x) - f(\bar{x})$)
    * **Minimizing/Maximizing:**
      · $\cos(\theta) = -1$, which means that this is minimizing $f(x) - f(\bar{x})$ (i.e. going in the opposite direction of the gradient) because the change (i.e. $f(x) - f(\bar{x})$) is negative.

- $\cos(\theta) = 1$, which means that this is maximizing $f(x) - f(\bar{x})$ (i.e. going in the direction of the gradient) because the change (i.e. $f(x) - f(\bar{x})$) is positive.
- $\cos(\theta) = 0$, which means orthogonal (perpendicular) to the gradient, meaning there is no change in the function value in this direction. The function remains constant in this direction (i.e. level set).

4.3.4 Tangent plane

Definition: In general, the tangent plane is defined by:

$$\left\{ \begin{bmatrix} x \\ t \end{bmatrix} \in \mathbb{R}^{n+1} \;\middle|\; t = f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x}) \right\} \tag{13}$$

- The tangent plane at $\bar{x}$ is defined by the first-order approximation.

This can also be written as:

$$\left\{ \begin{bmatrix} x \\ t \end{bmatrix} \in \mathbb{R}^{n+1} \;\middle|\; \begin{bmatrix} \nabla f(\bar{x})^\top & -1 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} = -f(\bar{x}) + \nabla f(\bar{x})^\top \bar{x} \right\} \tag{14}$$

In compact form:

$$a^\top \begin{bmatrix} x \\ t \end{bmatrix} = b \tag{15}$$

- $a^T = \begin{bmatrix} \nabla f(\bar{x})^\top & -1 \end{bmatrix}$
- $b = -f(\bar{x}) + \nabla f(\bar{x})^\top \bar{x}$

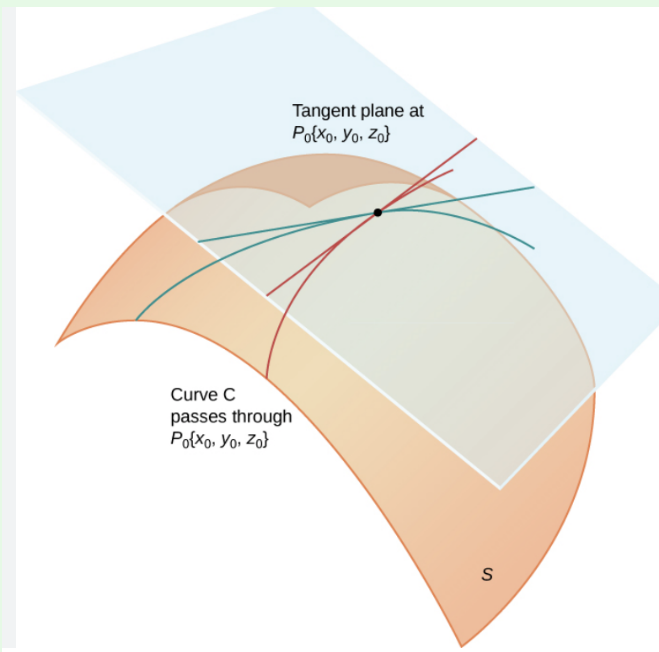

Figure 26: Tangent plane.

Example: Given $f(x_1, x_2) = 2x_1^2 + x_2^2$, the gradient of $f(x)$ is:

$$\nabla f(x) = \begin{pmatrix} 4x_1 \\ 2x_2 \end{pmatrix}$$

For the point $\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, we have:

$$\nabla f(\bar{x}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Thus, the first-order approximation is:

$$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x}) = 0$$

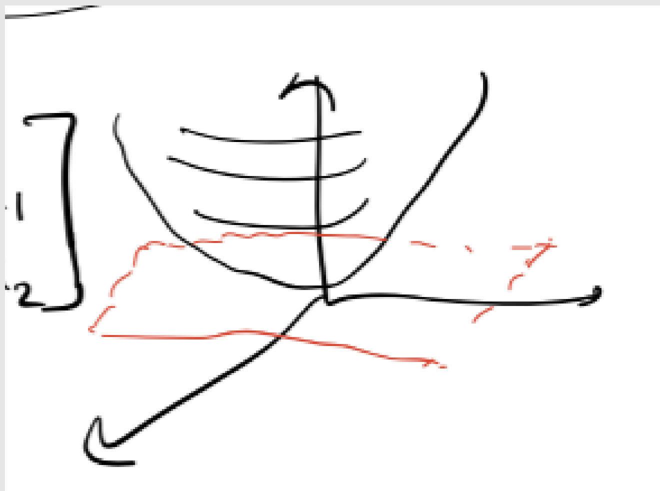

Figure 27: The tangent plane is defined as the xy-plane since the first order approximation is zero.

Now, for the point $\bar{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, we have:

$$\nabla f(\bar{x}) = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \quad f(\bar{x}) = 2$$

The first-order approximation is:

$$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x}) = 2 + \begin{pmatrix} 4 \\ 0 \end{pmatrix}^\top \begin{pmatrix} x_1 - 1 \\ x_2 \end{pmatrix}$$

$$= 2 + 4(x_1 - 1) = 2 + 4x_1 - 4$$

$$= 4x_1 - 2$$

Thus, the tangent plane is defined by:

$$\{(x_1, x_2, t) \mid t = 4x_1 - 2\}$$

Using the general form, it can be expressed as In the previous example, the tangent plane was defined by:

$$\left\{ \begin{bmatrix} x_1 \\ x_2 \\ t \end{bmatrix} \;\middle|\; \begin{bmatrix} 4 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ t \end{bmatrix} = 2 \right\}$$

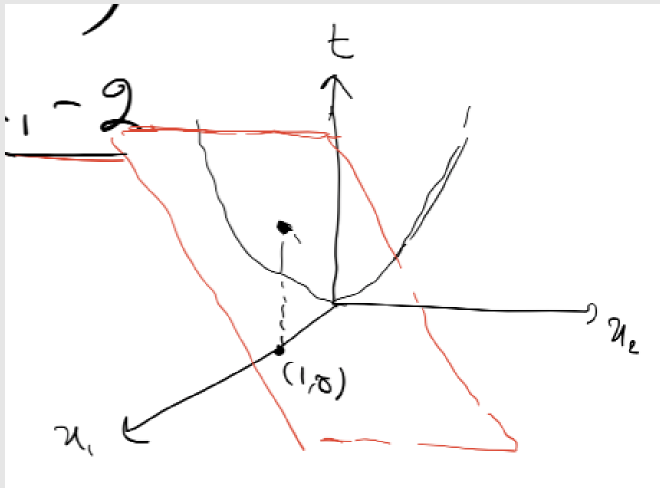

Figure 28: The tangent plane is defined as this plane.

4.3.5 Second order approximations

Definition: Given a function $f : \mathbb{R}^n \to \mathbb{R}$, the second-order approximation of $f(x)$ around $\bar{x}$ is:

$$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^\top \nabla^2 f(\bar{x})(x - \bar{x}) \tag{16}$$

- **Recall**: The Taylor series expansion of $f : \mathbb{R} \to \mathbb{R}$ is:

$$f(x) \approx f(\bar{x}) + \left.\frac{\partial f}{\partial x}\right|_{x=\bar{x}} (x - \bar{x}) + \frac{1}{2}\left.\frac{\partial^2 f}{\partial x^2}\right|_{x=\bar{x}} (x - \bar{x})^2$$

4.3.6 Hessians

Definition: **Hessian matrix** $\nabla^2 f(x)$ is given by:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{17}$$

- This is an $n \times n$ matrix.
- **Note**: The Hessian matrix is symmetric when f has continuous partial derivatives, because in that case:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

4.3.7 Approximating

Intuition: If the function you are approximating is 1st order, then 1st and 2nd order approximations will do fit it perfectly. If the function is 2nd order, then only 2nd order will approximate it perfectly. You can prove this by definition of Taylor series to show its equal to the function.

5 Matrices, Range, Null Space, Eigenvalues, Eigenvectors, Matrix Diagonalization (Ch. 3.1-3.5)

5.1 Matrices

Definition: Matrices are two-dimensional arrays of numbers. A general matrix A is denoted as:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} a^{(1)} & \cdots & a^{(n)} \end{bmatrix} \in \mathbb{R}^{m\times n} \quad (\text{or } \mathbb{C}^{m\times n} \text{ for complex numbers}) \tag{18}$$

5.1.1 Matrix Transpose

Definition: Given a matrix A with columns $a^{(i)}$, its transpose $A^\top$ is:

$$A^\top = \begin{bmatrix} (a^{(1)})^\top \\ \vdots \\ (a^{(n)})^\top \end{bmatrix} \tag{19}$$

5.1.2 Matrix Multiplication

Definition: The multiplication of $A^\top$ and a matrix B is given by:

$$A^\top B = \begin{bmatrix} (a^{(1)})^\top \\ \vdots \\ (a^{(n)})^\top \end{bmatrix} \begin{bmatrix} b^{(1)} & \cdots & b^{(n)} \end{bmatrix} = \begin{bmatrix} (a^{(1)})^\top b^{(1)} & \cdots & (a^{(1)})^\top b^{(n)} \\ \vdots & \ddots & \vdots \\ (a^{(n)})^\top b^{(1)} & \cdots & (a^{(n)})^\top b^{(n)} \end{bmatrix} \tag{20}$$

For regular matrix multiplication AB, we have:

$$AB = \begin{bmatrix} a^{(1)} & \cdots & a^{(n)} \end{bmatrix} \begin{bmatrix} (b^{(1)})^\top \\ \vdots \\ (b^{(n)})^\top \end{bmatrix} = \sum_{i=1}^{n} a^{(i)} (b^{(i)})^\top \tag{21}$$

(Note: Try to think why this is the same as before.)

5.1.3 Block Matrix Product

Definition: For a block matrix product:

$$\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = AX + BY \tag{22}$$

(Analogous to vectors: $[a \ b] \begin{bmatrix} x \\ y \end{bmatrix} = ax + by$)

5.1.4 Types of Matrices

Definition:

- **Square matrix**: $m = n$
- **Diagonal matrix**: $a_{ij} = 0 \quad \forall i \neq j$
- **Identity matrix**
- **Triangular matrix**

- **Orthogonal matrix**: $(A^\top A = AA^\top = I)$ (i.e. inverse of A is A^T)
- **Symmetric matrix**: $A = A^\top$

5.1.5 Matrices as Linear Maps

Definition: Matrices are used as linear maps, where:

$$y = Ax$$

$$[y] \in \mathbb{R}^m \quad = \quad [A] \in \mathbb{R}^{m\times n} \quad [x] \in \mathbb{R}^n$$

This represents a **linear function**.

For an **affine function**, the equation becomes:

$$y = Ax + b$$

5.2 Range Space

Definition:

$$\begin{aligned}
\mathcal{R}(A) &= \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m & (23)\\
&= \left\{\sum_{i=1}^{n} x_i a^{(i)} \mid x \in \mathbb{R}^n\right\} & (24)\\
&= \operatorname{span}\{a^{(1)}, a^{(2)}, \ldots, a^{(n)}\} & (25)
\end{aligned}$$

- $\mathcal{R}(A)$ is a subspace (closed under addition and scalar multiplication).

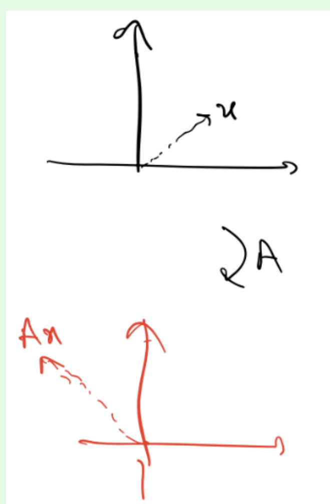

Figure 29: Range space.

5.3 Null Space

Definition: Set of all vectors that map to zero.

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \subseteq \mathbb{R}^n \quad (26)$$

- $\mathcal{N}(A)$ is also a subspace (closed under addition and scalar multiplication).

5.4 Fundamental theorem of algebra

Definition: For any matrix $A \in \mathbb{R}^{m\times n}$, the subspaces are orthogonal to each other (i.e. orthogonal complement)

$$\mathcal{R}(A) \perp \mathcal{N}(A^T) \tag{27}$$

$$\mathcal{R}(A^T) \perp \mathcal{N}(A) \tag{28}$$

Furthermore, the direct sum of the two subspaces (i.e. if you take one vector from one subspace, and another vector from the other subspace, the sum of these vectors would be in $\mathbb{R}^n$ or $\mathbb{R}^m$)

$$\mathcal{R}(A) \oplus \mathcal{N}(A^T) = \mathbb{R}^m \tag{29}$$

- $\dim(\mathcal{R}(A)) + \dim(\mathcal{N}(A^T)) = m$
  - Rank of A: $\dim(\mathcal{R}(A))$

$$\mathcal{R}(A^T) \oplus \mathcal{N}(A) = \mathbb{R}^n \tag{30}$$

- $\dim(\mathcal{R}(A^T)) + \dim(\mathcal{N}(A)) = n$
  - Rank of A: $\dim(\mathcal{R}(A^T))$

Derivation: Why Does This Hold? Recall the definition of orthogonal complement:
- $\forall x \in S$, $y \in S^{\perp}$, we have that $\langle x, y\rangle = 0$
- $x \in V$ can be decomposed as $x = x^* + v$, where $x^* \in S$ and $v \in S^{\perp}$

Therefore,

$$\dim V = \dim S + \dim S^{\perp}$$

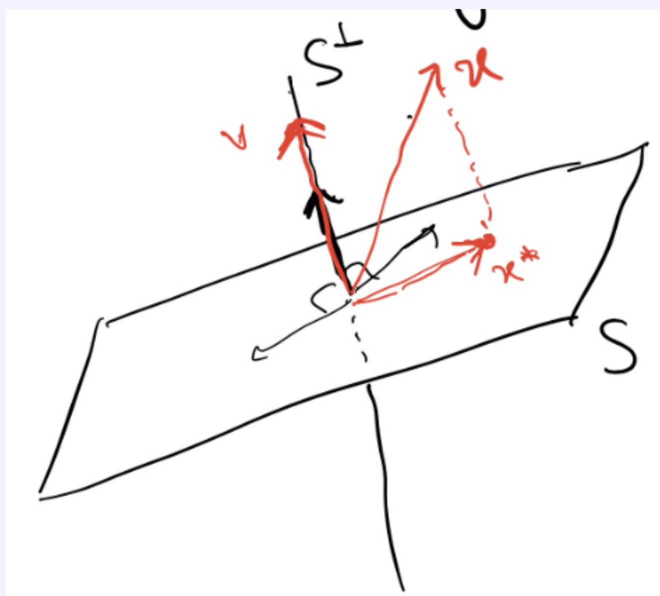

Figure 30: Orthogonal complement.

Now using this definition of orthogonal complement, let's apply it to the range space and null space of A: Consider

$$\mathcal{R}(A) = \left\{ y \mid y = \sum_{i=1}^{n} \alpha_i a^{(i)} \right\}$$

$$\mathcal{N}(A^{\top}) = \left\{ x \mid \left(a^{(i)}\right)^{\top} x = 0,\ i = 1, \ldots, n \right\}$$

$$A^\top = \begin{bmatrix} (a^{(1)})^\top \\ \vdots \\ (a^{(n)})^\top \end{bmatrix}$$

Consider $y \in \mathcal{R}(A)$ so $y = \sum_{i=1}^{n} \alpha_i a^{(i)}$, and $x \in \mathcal{N}(A^\top)$,

$$\langle y, x \rangle = \left\langle \sum_{i=1}^{n} \alpha_i a^{(i)}, x \right\rangle = \sum_{i=1}^{n} \alpha_i \langle a^{(i)}, x \rangle$$

Since $x \in \mathcal{N}(A^\top)$, this implies that $\langle a^{(i)}, x \rangle = 0$, so:

$$\langle y, x \rangle = 0$$

Thus, any $y \in \mathcal{R}(A)$ is orthogonal to any $x \in \mathcal{N}(A^\top)$.

Therefore, $\mathcal{R}(A) \oplus \mathcal{N}(A^\top) = \mathbb{R}^m$, and $\dim(\mathcal{R}(A)) + \dim(\mathcal{N}(A^\top)) = m$.

Example:

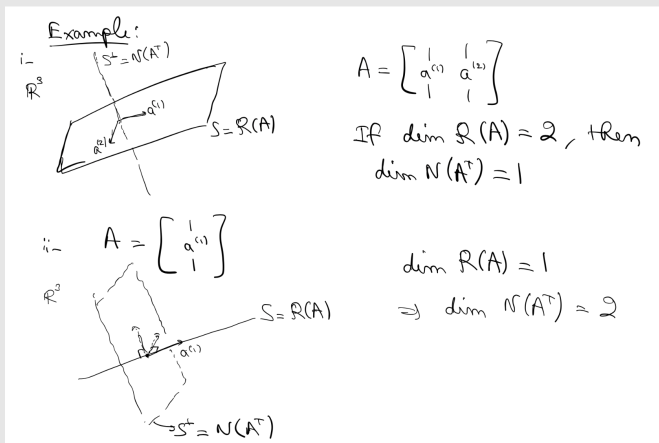

Figure 31: Example of fundamental theorem of linear algebra.

- Given the range of A, we can find the null space of A by finding what is orthogonal to it.

5.5 Determinant

Definition:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{31}$$

$$\det(A) = a_{11}a_{22} - a_{12}a_{21} \tag{32}$$

The volume of the region transformed by matrix A is given by:

$$\det(A) \tag{33}$$

Intuition: Graphically, the determinant represents the volume of the parallelogram spanned by the column vectors of A.

- Starting from a unit square, the matrix A transforms this square into a parallelogram.
- The area of the transformed parallelogram, denoted as P, equals the determinant:

$$\text{vol}(P) = \det(A)$$

Why? It is easy to see this for a diagonal matrix, where:

$$A = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}$$

The matrix A scales the unit square by a_{11} along one axis and by a_{22} along the other axis. This results in a rectangle with area:

$$\text{vol}(P) = a_{11}a_{22} = \det(A)$$

For instance, applying A to the basis vectors:

$$A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} a_{11} \\ 0 \end{bmatrix}, \quad A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ a_{22} \end{bmatrix}$$

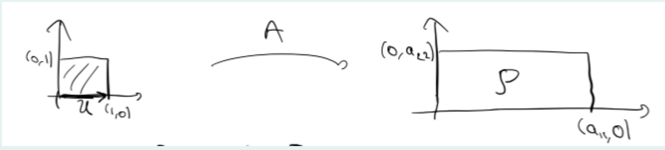

Figure 32: Diagonal

What if the matrix is not diagonal? Consider an upper triangular matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix}$$

The matrix A still transforms the unit square, but now the parallelogram P is slanted due to the non-zero a_{12}. The area of the parallelogram is still given by the determinant:

$$\text{vol}(P) = a_{11}a_{22} = \det(A)$$

Again, applying A to the basis vectors:

$$A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} a_{11} \\ 0 \end{bmatrix}, \quad A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}$$

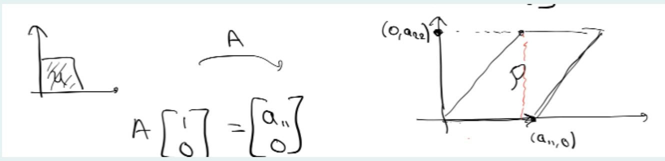

Figure 33: Upper triangular

For a general matrix A, we can use the QR factorization, where:

$$A = QR$$

R is an upper triangular matrix and Q is an orthogonal matrix (i.e., $Q^\top Q = I$).

- The orthogonal matrix Q represents a rotation, which preserves volume (the area remains unchanged).
- Therefore, the volume is determined by the determinant of R.

$$\text{vol}(P) = \det(R) = r_{11}r_{22} = \det(A)$$

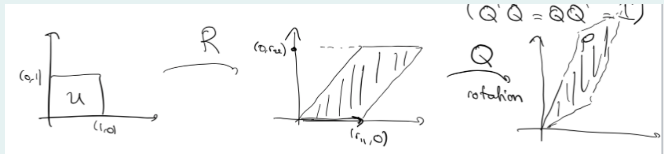

Figure 34: General matrix

5.6 Matrix inner product and norm

Definition: Set of matrices $A \in \mathbb{R}^{m \times n}$ is a vector space. We can define an inner product over this space:

$$\langle A, B \rangle = \operatorname{tr}(B^T A) = \operatorname{tr}(A^T B) \tag{34}$$

- tr : Trace is the sum of diagonal elements of a square matrix.

5.6.1 Frobenius Norm

Definition: For $A, B \in \mathbb{R}^{m \times n}$, then

$$\|A\|_F = \sqrt{\langle A, A \rangle} = \sqrt{\operatorname{tr}(A^T A)} = \sqrt{\sum_{i,j} a_{ij}^2} \tag{35}$$

Intuition: Let

$$A = \begin{bmatrix} a^{(1)} & \cdots & a^{(n)} \end{bmatrix}$$

then

$$A^T A = \begin{bmatrix} (a^{(1)})^T \\ \vdots \\ (a^{(n)})^T \end{bmatrix} \begin{bmatrix} a^{(1)} & \cdots & a^{(n)} \end{bmatrix} = \begin{bmatrix} (a^{(1)})^T a^{(1)} & \cdots & (a^{(1)})^T a^{(n)} \\ \vdots & \ddots & \vdots \\ (a^{(n)})^T a^{(1)} & \cdots & (a^{(n)})^T a^{(n)} \end{bmatrix}$$

- **Note:** The matrix multiplication shows us why the trace can be written as $\sum_{i,j} a_{ij}^2$.
  - e.g. $a^{(1)})^T a^{(1)} = a_{11}^2 + a_{21}^2 + \ldots + a_{n1}^2$

If we think of the matrix as a long column vector of the column vectors of A, then

$$\begin{bmatrix} a^{(1)} \\ \vdots \\ a^{(n)} \end{bmatrix}$$

so we can think of the Frobenius norm as the l_2 norm exactly since it's $\sqrt{\sum_{i,j} a_{ij}^2}$, which is identical to the expression for l2-norm.

5.6.2 Operator norm (Motivation for eigenvalues and eigenvectors):

Intuition: **Let $A \in \mathbb{R}^{m \times n}$ be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$.**

- Think of the matrix as a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$.

$$v \quad \longrightarrow \quad Av \quad \text{(ellipse)}$$

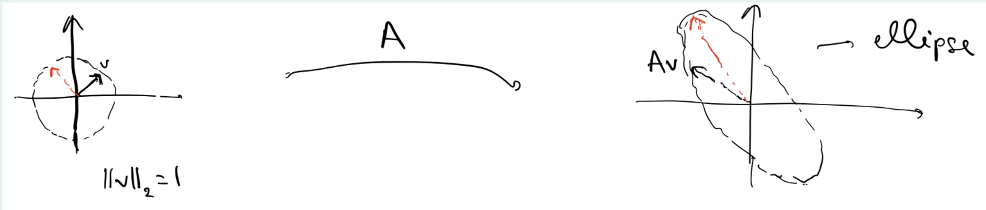

Figure 35: Matrix norm tht maximizes Av given v with unit norm.

Question: Which direction of v on the unit circle results in Av with the largest magnitude? **Operator norm**, which can defined for all p valiues.

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

$$\|A\|_1 = \max_{\|v\|_1=1} \|Av\|_1$$

$$\|A\|_\infty = \max_{\|v\|_\infty=1} \|Av\|_\infty$$

"Interesting" directions (i.e. directions of interest):

1. **Max/Min amplification direction:**

$$\max_{\|v\|_2=1} \|Av\|_2 = \|A\|_2, \quad \min_{\|v\|_2=1} \|Av\|_2$$

   - Later, we will see that this is related to the **singular values** of the matrix.

2. **Invariant direction:** (square matrix $A \in \mathbb{R}^{n\times n}$)

$$Av = \lambda v$$

   - v is the eigenvector and λ is the eigenvalue.
   - **Note:** In 2D, there are 2 eigenvectors (i.e. parallel with same and opposite direction)

5.7 Eigenvalues and Eigenvectors

Definition: For $A \in \mathbb{R}^{n\times n}$, a non-zero vector v is an eigenvector of A with associated eigenvalue λ if

$$Av = \lambda v$$

- **Consequence:** $(A - \lambda I)v = 0$, therefore, $v \in \mathcal{N}(A - \lambda I)$, where $\mathcal{N}(A - \lambda I)$ is the null space of $(A - \lambda I)$.

5.7.1 Characteristic polynomial

Definition: The above leads to a polynomial equation in λ of degree n. The roots are the eigenvalues of A.

$$\det(A - \lambda I) = 0$$

- **Roots can be real or complex.** If complex, they occur in conjugate pairs.
- **Note:** The determinant is equal to 0 because $Av = \lambda v$ (i.e. linearly dependent), and LD vectors have determinant of 0 by definition.

5.7.2 Geometric and algebraic multiplicity

Definition:
Algebraic multiplicity: $\text{AM}(\lambda)$ is the number of times a given eigenvalue appears as a root of the characteristic equation $\det(A - \lambda I) = 0$.

- For example, $(\lambda - 1)^2$ has root 1 with multiplicity 2.

Geometric multiplicity: $\text{GM}(\lambda) = \dim \mathcal{N}(A - \lambda I)$

Theorem: In general, for any matrix $A \in \mathbb{R}^{n\times n}$ (i.e. MUST BE SQUARE),

$$GM(\lambda) \leq AM(\lambda).$$

5.7.3 Defective and non-defective

Definition:
Defective Matrix: If $GM(\lambda) < AM(\lambda)$ for some λ

Non-defective Matrix: If $GM(\lambda) = AM(\lambda)$, then the matrix is **diagonalizable** (i.e. non-defective).

- **"Most" matrices are non-defective.** In other words, the set of non-defective matrices is dense (a random matrix is almost surely non-defective).

5.7.4 Eigenvectors corresponding to different eigenvalues are l.i.

Theorem: **Lemma:** If $u^{(1)}, \ldots, u^{(k)}$ are eigenvectors then,

$$u^{(i)} \notin \phi_j = \mathcal{N}(A - \lambda I) \text{ for } i \neq j \tag{36}$$

- i.e. An eigen vector of i of one null space cannot be in another eigenvector's null space.

Derivation: *Why?* Prove by contradiction, assume $u^{(i)}$ is in the null space of ϕ_j.

- If $Au^{(i)} = \lambda_i u^{(i)}$ and $u^{(i)} \in \mathcal{N}_j$, then $Au^{(i)} = \lambda_j u^{(i)}$.
- Therefore, $(\lambda_i - \lambda_j)u^{(i)} = 0$, which is impossible since $\lambda_i \neq \lambda_j$ and $u^{(i)}$ is a non-zero vector by definition of eigenvectors.

Theorem: Eigenvectors corresponding to different eigenvalues are linearly independent.

Derivation: Let $\lambda_1, \ldots, \lambda_k$ be distinct eigenvalues, $u^{(1)}, \ldots, u^{(k)}$ be the corresponding eigenvectors, and $\phi_i = \mathcal{N}(A - \lambda_i I)$, the null space corresponding to eigenvalue λ_i.

Proof by contradiction, assume that $u^{(1)}, \ldots, u^{(k)}$ are linearly dependent, i.e.,

$$u^{(1)} = \sum_{i=2}^{k} \alpha_i u^{(i)}.$$

- i.e. The first eigenvector can be written as a linear combination of the other eigenvectors.

Part 1:

$$Au^{(1)} = \sum_{i=2}^{k} \alpha_i A u^{(i)} = \sum_{i=2}^{k} \alpha_i \lambda_i u^{(i)}.$$

- 2nd part: By definition of $u^{(1)}$
- 3rd part: Since $u^{(i)}$ is an eigenvector with corresponding λ_i

Part 2:

$$\lambda_1 u^{(1)} = \sum_{i=2}^{k} \alpha_i \lambda_1 u^{(i)}.$$

- **Key:** λ_1 is inside the summation, while for part 1, it was λ_i

Combining both part 1 and part 2: But $Au^{(1)} = \lambda_1 u^{(1)}$, so

$$Au^{(1)} - \lambda_1 u^{(1)} = \sum_{i=2}^{k} \alpha_i (\lambda_1 - \lambda_i) u^{(i)} = 0,$$

which implies that $\{u^{(2)}, \dots, u^{(k)}\}$ is linearly dependent because $\lambda_1 - \lambda_i \neq 0$ since they are distinct, so the only α's to make the linear combination of FINISH LATER

Continue with the same argument to show that $\{u^{(3)}, \dots, u^{(k)}\}$ is linearly dependent, and so on.

Eventually, $\{u^{(k-1)}, u^{(k)}\}$ are linearly dependent, so:

$$u^{(k)} = \alpha u^{(k-1)}.$$

But this is impossible by the previous lemma since an eigenvector of one null space cannot be in another eigenvector's null space (i.e. cannot write on as a l.c. of the other). Therefore, $\{u^{(1)}, \dots, u^{(k)}\}$ is linearly independent.

5.8 Matrix Diagonalization

Intuition: The previous theorem leads us to the diagonalization of a non-defective square matrix $A \in \mathbb{R}^{n \times n}$.

The idea is to look at the null space $\mathcal{N}(A - \lambda_i I)$.

- When $\lambda_1, \dots, \lambda_n$ are distinct, then $u^{(1)}, \dots, u^{(n)}$ are linearly independent, so the set $\{u^{(1)}, \dots, u^{(n)}\}$ forms a basis for $\mathbb{R}^n$.
- If some λ_i is a repeated eigenvalue. For example, if $AM(\lambda_i) = 2$, then $GM(\lambda_i) = 2$ because it's non-defective matrix so $\dim \mathcal{N}(A - \lambda_i I) = 2$.
  - This means we can find $\{u^{(i,1)}, u^{(i,2)}\}$ that span this null space.
  - So, if we assemble all these $u^{(i,j)}$'s, then we get a basis for $\mathbb{R}^n$.

Theorem: Let λ_i for $i = 1, \dots, k$, be distinct eigenvalues of a non-defective matrix $A \in \mathbb{R}^{n \times n}$. Let μ_i be the multiplicity of λ_i (i.e., $AM(\lambda_i) = GM(\lambda_i) = \mu_i$).

Let

$$U^{(i)} = \begin{bmatrix} u^{(i,1)} & \dots & u^{(i,\mu_i)} \end{bmatrix}$$

be a matrix of eigenvectors that form a basis for $\mathcal{N}(A - \lambda_i I)$.

Then

$$U = \begin{bmatrix} U^{(1)} & \dots & U^{(k)} \end{bmatrix}$$

is a complete set of basis vectors for $\mathbb{R}^n$.

Then the **Eigendecomposition** is

$$A = U \Lambda U^{-1}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

is a diagonal matrix with the eigenvalues of A on the diagonal.

- Where λ_1 appears μ_1 times on the diagonal (indicating multiplicity), followed by λ_2 appears μ_2 and so on.

Derivation: $Au^{(i,j)} = \lambda_i u^{(i,j)}$, where $u^{(i,j)}$ is an eigenvector with eigenvalue λ_i, then this implies that

$$A \begin{bmatrix} u^{(1,1)} & \dots & u^{(1,\mu_1)} & u^{(2,1)} & \dots & u^{(2,\mu_2)} & \dots & u^{(n,\mu_n)} \end{bmatrix} = U \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

- i.e. Generalizing to all the eigenvectors with their corresponding eigenvalues.

Thus, we have:

$$AU = U\Lambda$$

which implies:

$$A = U\Lambda U^{-1}$$

where U is invertible because its columns (the eigenvectors) are linearly independent, and U is a square matrix.

5.8.1 What does this diagonalization mean?

Intuition: We start with the eigendecomposition of a matrix A:

$$A = U\Lambda U^{-1}$$

- $U = \begin{bmatrix} u^{(1)} & \dots & u^{(n)} \end{bmatrix}$.

Let $x \in \mathbb{R}^n$ be any vector. Then:

$$Ax = U\Lambda U^{-1}x.$$

Let $\tilde{x} = U^{-1}x$, so that:

$$Ax = U\Lambda\tilde{x} \quad \text{and} \quad x = U\tilde{x} = \begin{bmatrix} u^{(1)} & \dots & u^{(n)} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix} = \sum_{i=1}^{n} \tilde{x}_i u^{(i)}.$$

- The coordinates of $\tilde{x}$ are the components (i.e. coefficients) of x in the new basis $\{u^{(1)}, \dots, u^{(n)}\}$.
- $\Lambda\tilde{x}$ means that we are scaling $\tilde{x}_i$ by the corresponding eigenvalue λ_i.
- $U\Lambda\tilde{x}$ means that we are changing the coordinates back to the original basis.

This applies to any non-defective (or diagonalizable) matrix $A \in \mathbb{R}^{n\times n}$.

6 Problem set 2

6.1 Calculating inverse of a matrix

Process: Given an $n \times n$ matrix A, the inverse matrix A^{-1} can be found through the following steps:

1. **Check if the matrix is invertible**:
   - To be invertible, the matrix must have a non-zero determinant: $\det(A) \neq 0$.
   - If $\det(A) = 0$, the matrix is singular and does not have an inverse.
2. **Find the determinant of the matrix**:
   - The determinant of an $n \times n$ matrix A is a scalar value that can be computed using cofactor expansion (also known as Laplace's expansion) along any row or column of the matrix.
   - **e.g.** For the first row, the determinant of a matrix $A = [a_{ij}]$ is given by:

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j})$$

   - Here, A_{1j} is the $(n-1) \times (n-1)$ matrix obtained by removing the first row and j-th column of A.
3. **Find the cofactors of the matrix**:
   - The cofactor of an element a_{ij} in the matrix A is denoted by C_{ij}.
   - The cofactor is given by:

$$C_{ij} = (-1)^{i+j} \det(A_{ij})$$

   - Here, A_{ij} is the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and the j-th column from A.
   - You will compute the cofactor for each element in the matrix A, creating a cofactor matrix.
4. **Form the cofactor matrix (matrix of cofactors)**:
   - Construct the cofactor matrix by placing each cofactor C_{ij} at the corresponding position in a new matrix:

$$\text{Cofactor Matrix} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{pmatrix}$$

5. **Form the adjugate (or adjoint) matrix**:
   - The adjugate matrix, denoted as $\text{adj}(A)$, is the transpose of the cofactor matrix.
   - In other words, swap the rows and columns of the cofactor matrix:

$$\text{adj}(A) = \begin{pmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{pmatrix}$$

6. **Calculate the inverse matrix**:
   - The inverse of the matrix A is given by the formula:

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

   - Here, $\det(A)$ is the determinant of the matrix A, and $\text{adj}(A)$ is the adjugate matrix.
   - Multiply each element of the adjugate matrix by $\frac{1}{\det(A)}$.
7. **Verify the inverse (optional)**:
   - Multiply the matrix A by its inverse A^{-1}. You should get the identity matrix I_n, where:

$$A \times A^{-1} = A^{-1} \times A = I_n$$

   - The identity matrix I_n has 1's along the diagonal and 0's elsewhere:

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

7 Symmetric Matrices, Orthogonal Matrices, Spectral Decomposition, Positive Semidefinite Matrices, Ellipsoids (Ch. 4.1-4.4)

7.1 Symmetric matrices

Definition:

$$S^n = \left\{ A \in \mathbb{R}^{n \times n} \mid A = A^T \right\}$$

- In other words, a matrix A is symmetric if: $a_{ij} = a_{ji}$

$$S^n = \left\{ A \in \mathbb{C}^{n \times n} \mid A = A^H \right\}$$

- A^H represents the Hermitian transpose (conjugate transpose) of A.

7.2 Orthogonal matrices

7.3 Spectral decomposition

7.4 Positive semidefinite matrices

7.5 Ellipsoids

8 Singular Value Decomposition, Principal Component Analysis (Ch. 5.1, 5.3.2)

8.1 Singular value decomposition

8.2 Principle component analysis

9 Interpretations of SVD, Low-Rank Approximation (Ch. 5.2-5.3.1)

9.1 Interpretation of SVD

9.2 Low-rank approximation

10 Least Squares, Overdetermined and Underdetermined Linear Equations (Ch. 6.1-6.4)

10.1 Least squares

10.2 Overdetermined linear equation

10.3 Underdetermined linear equation

11 Regularized Least-Squares, Convex Sets and Convex Functions (Ch. 6.7.3, 8.1-8.4)

11.1 Regularized least-squares

11.2 Convex sets and convex functions

12 Lagrangian Method for Constrained Optimization, Linear Programming and Quadratic Programming (Ch. 8.5, 9.1-9.6)

12.1 Lagrangian method for constrained optimization

12.2 Linear programming and quadratic programming

13 Numerical Algorithms for Unconstrained and Constrained Optimization (Ch. 12.1-12.3)

13.1 Numerical algorithms for unconstrained optimization

13.2 Numerical algorithms for constrained optimization