ECE367 Cheatsheet

John Doe

November 25, 2024

Contents

List of Figures

List of Tables

1 QA

Intuition: **Optional" problems:** Optional problems provide extra practice or introduce interesting connections or extensions.

- They need not be turned in. We will not grade them, but we will assume you have reviewed and understood the solutions to the optional problems when designing the exams.

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

2.1 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.

2.1.1 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.

2.2 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.

2.3 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:** $(\alpha + \beta) \cdot \mathbf{x} = \alpha \cdot \mathbf{x} + \beta \cdot \mathbf{x}$.

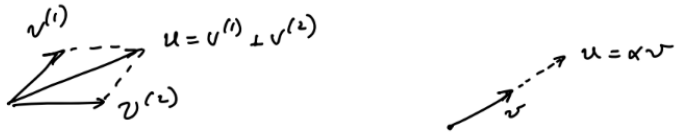

Figure 1: Vector addition and scalar multiplication.

2.3.1 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.

2.4 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 $\mathbb{R}^n$.

2.5 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:** $\text{Span}(S)$ is always a subspace of V.

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

2.6 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).

2.6.1 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.

2.7 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 $\mathcal{V} = \mathbb{R}^n$?

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

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

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

2.7.1 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.

2.8 Norms (Notion of distance)

Definition: Let $\mathcal{V}$ be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A norm is a function $\|\cdot\|: \mathcal{V} \to \mathbb{R}$ 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)

2.8.1 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:
- $\ell_2:\ B_2 = \left\{\mathbf{x} \;\middle|\; \sqrt{x_1^2 + x_2^2} \leq 1\right\}$
- $\ell_1:\ B_1 = \{\mathbf{x} \mid |x_1| + |x_2| \leq 1\}$
- $\ell_\infty:\ B_\infty = \{\mathbf{x} \mid \max|x_i| \leq 1 \text{ or } |x_1| \leq 1, |x_2| \leq 1\}$
- $\ell_0:\ B_0 = \{\mathbf{x} \mid \text{card}(\mathbf{x}) \leq 1\}$

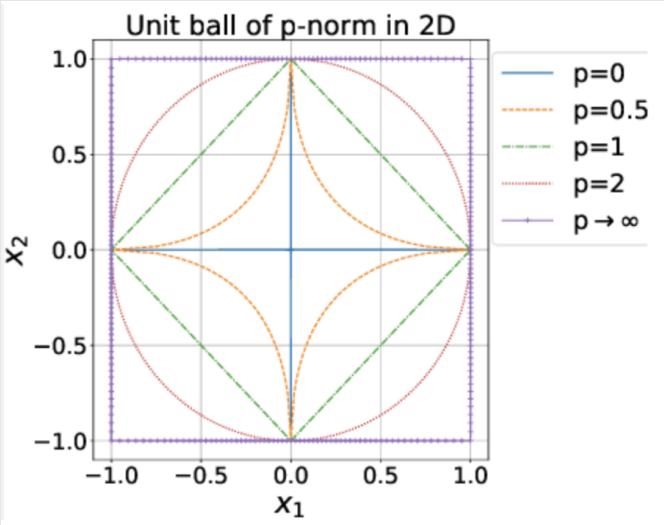

Figure 2: Norm balls of different p values.

2.8.2 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)

2.8.3 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

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

2.9 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:** $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$
   - $\langle \mathbf{x}, \mathbf{y} \rangle = \underline{\langle \mathbf{y}, \mathbf{x} \rangle}$ in $\mathbb{R}^n$
   - $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ 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?

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

2.9.1 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)

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

2.9.2 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.

2.9.3 Cauchy-Schwartz inequality and its generalization

Definition:

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

Hölder's Inequality (generalization):

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

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

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

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

2.9.4 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 $\mathbb{C}^n$

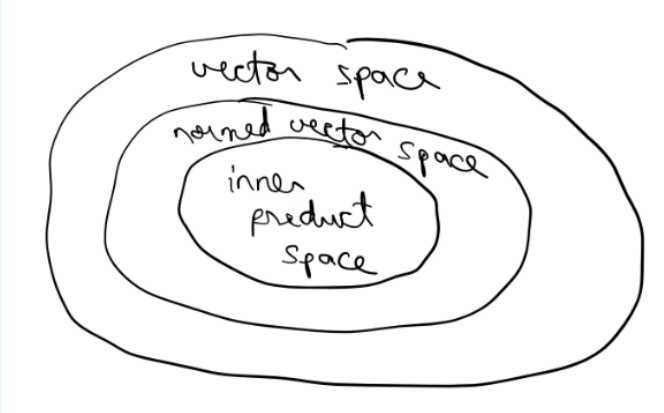

Figure 3: Ordering of the vector spaces.

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

2.10 Orthogonal decomposition

2.10.1 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.

2.10.2 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 $\mathbb{R}^n$

2.10.3 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: $\mathbf{x} \perp S$.

2.10.4 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 $\mathbf{x}_{S^\perp} \in S^\perp$

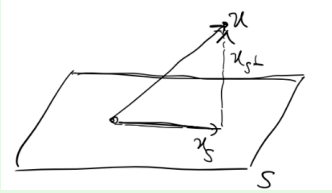

Figure 4: Drawing any x.

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

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

3.1 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.

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

3.1.1 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 $\|x - y\|$

3.1.2 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.

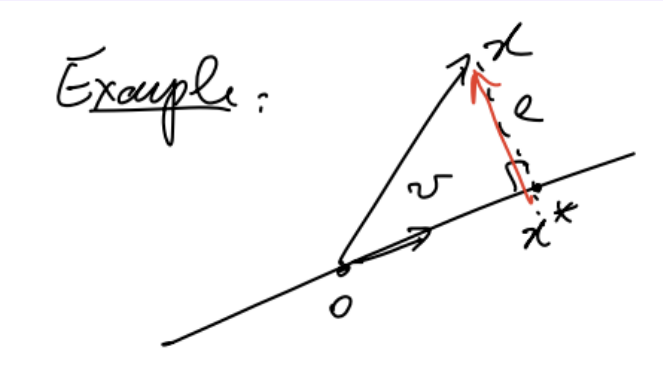

Figure 6: Visual representation of the projection problem.

3.1.3 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,

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

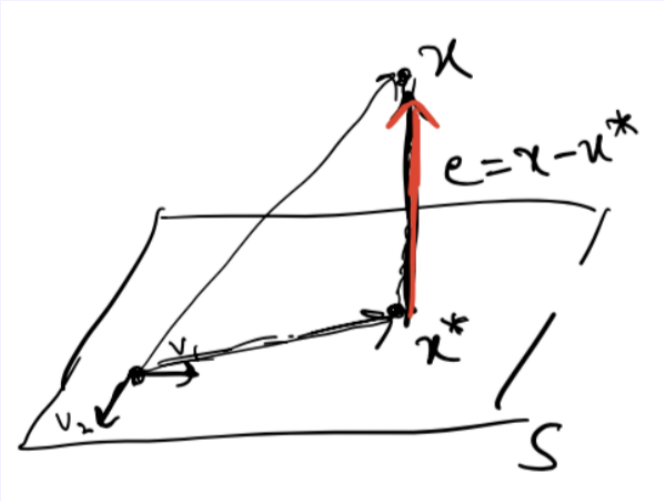

Figure 7: Generalization of projection.

3.1.4 Application of projections: Fourier series

Example: **Fourier series:**

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

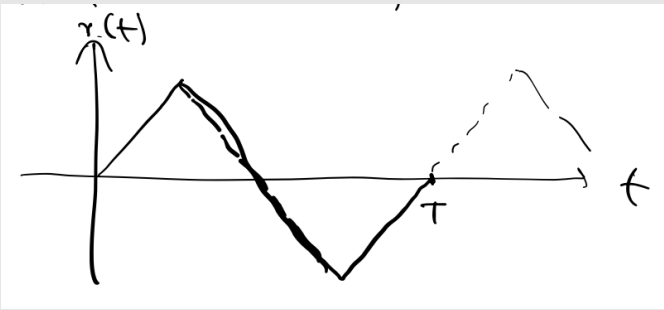

Figure 8: Periodic triangle function.

2. **Inner product for time domain (complex version):** $a_k = \langle x(t), y(t) \rangle = \frac{1}{T} \int_T x(t)\overline{y(t)}\, dt$
   - **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:** $a_k = \langle x(t), e^{jk\omega_0 t} \rangle = \frac{1}{T_0} \int_0^T x(t) e^{-jk\omega_0 t} dt$
   - **Fundamental frequency:** $\omega_0 = \frac{2\pi}{T_0}$.
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.

3.2 Gram-Schmidt and QR decomposition

3.2.1 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 $x^* = \sum_{i=1}^{d} \alpha_i u^{(i)}$

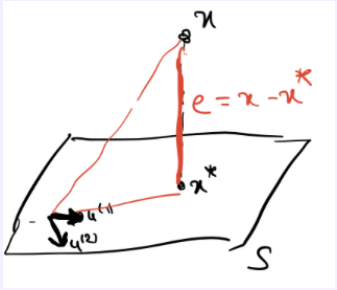

Figure 9: Not orthogonal, but similar to projection with orthonormal basis.

- **Note:** If $\{u^{(1)}, \ldots, u^{(d)}\}$ 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:

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

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 $\alpha_0, \alpha_1, \ldots, \alpha_d$, 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

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

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

3.2.2 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 $S = \text{span}\{u^{(1)}, \ldots, u^{(d)}\}$ is to first find an orthonormal basis of S, and then the projection problem becomes easier.

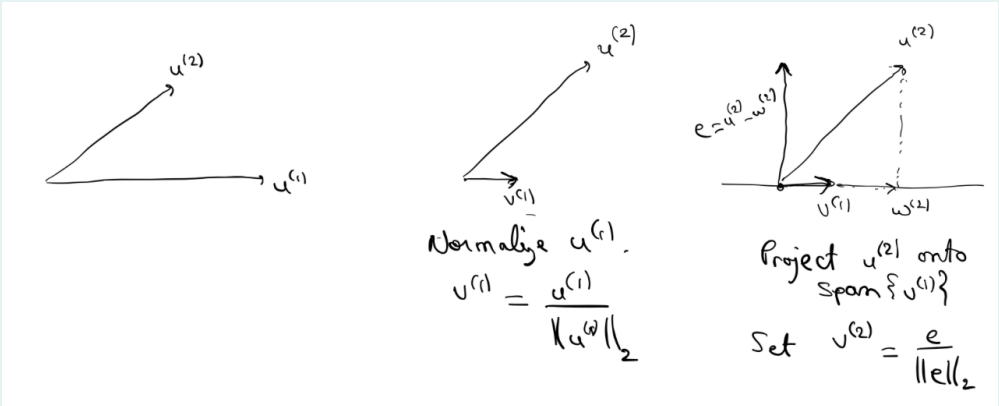

Figure 10: Gram-Schmidt Process for 2D.

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, $\{v^{(1)}, \ldots, v^{(d)}\}$ is an orthonormal basis for $\text{span}\{u^{(1)}, \ldots, u^{(d)}\}$.

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 $(1/\sqrt{2}, 1/\sqrt{2}), (1/\sqrt{2}, 1/\sqrt{2})$.

3.2.3 QR decomposition

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

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

$$\begin{bmatrix} u^{(1)} & \cdots & u^{(d)} \end{bmatrix} = QR$$

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

$$= \begin{bmatrix} r_{11}v^{(1)} & r_{12}v^{(1)} + r_{22}v^{(2)} & \cdots \end{bmatrix}$$

- Q: Orthonomral matrix (i.e., its columns are orthogonal to each other and have unit norm)
  - Q has zero columns when A is linearly dependent because it's impossible to construct orthogonal vectors for linearly dependent directions of A.
- 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:

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

3.3 Projection of a subspace defined by its orthogonal vectors

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

3.3.2 Projection

Derivation:

1. Projecting a vector x onto a subspace S spanned by the vectors $\{a^{(1)}, \ldots, a^{(m)}\}$. The projection x^* is given by:

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

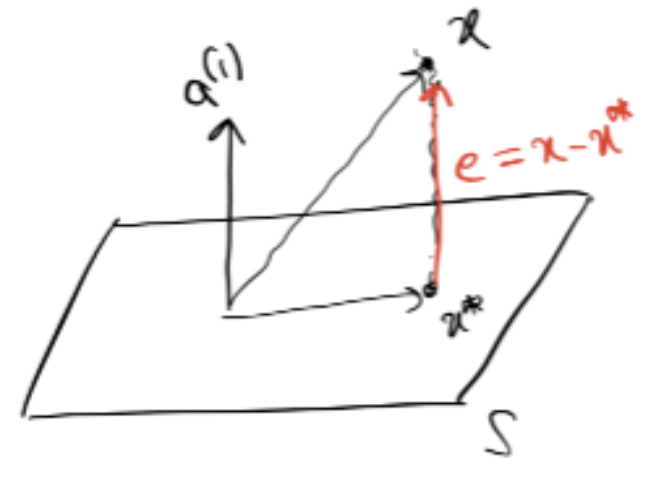

Figure 11: Projection onto a subspace defined by its orthogonal vectors

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:

$$e \in \text{span}\{a^{(1)}, \ldots, a^{(m)}\}$$

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 $\beta_1, \ldots, \beta_m$.

4. Since $x^* \in S$, we have the condition:

$$\langle x^*, a^{(j)} \rangle = 0 \quad \forall j = 1, \ldots, m$$

which leads to the following equation:

$$(a^{(j)})^T x^* = 0 \quad \forall j = 1, \ldots, m$$

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:

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

3.4 Projection onto affine sets

3.4.1 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 shifting vector and S is the original subspace. This represents a shifted version of the subspace S.

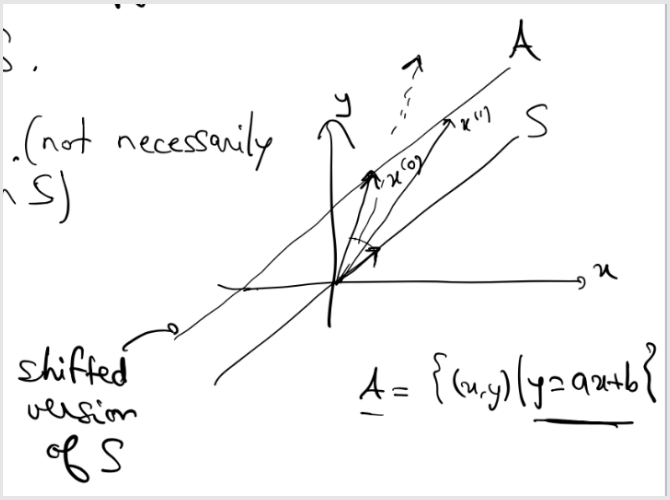

Figure 12: Affine space of a 2D space, where the $x^{(0)}$ is the constant (i.e. shifting origin to the Affine set) that we are adding to shift all the vectors to the affine space.

3.4.2 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).

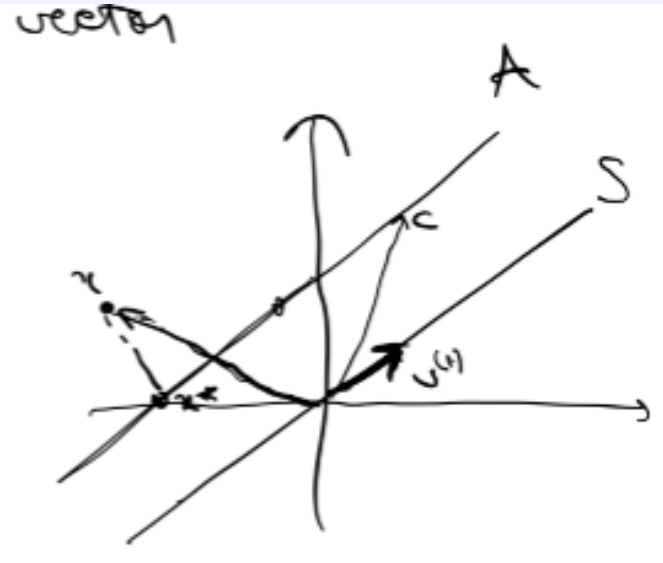

Figure 13: Projection problem visualization, where we are projecting the vector onto the Affine space.

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.

3.4.3 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.)

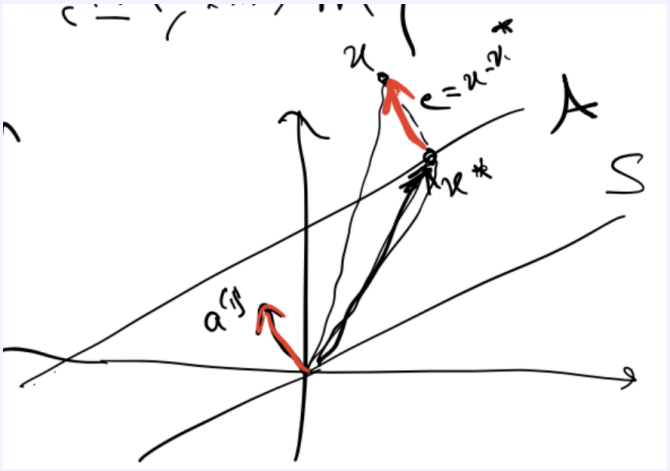

Figure 14: Projection problem visualization, where we are projecting the vector onto the affine space, which is in line with the orthogonal vectors.

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 $\mathcal{A}$.

3.4.4 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.

3.5 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.

3.6 Hyperplanes and half-spaces

Definition:

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

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

- **Half-space:**

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

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

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

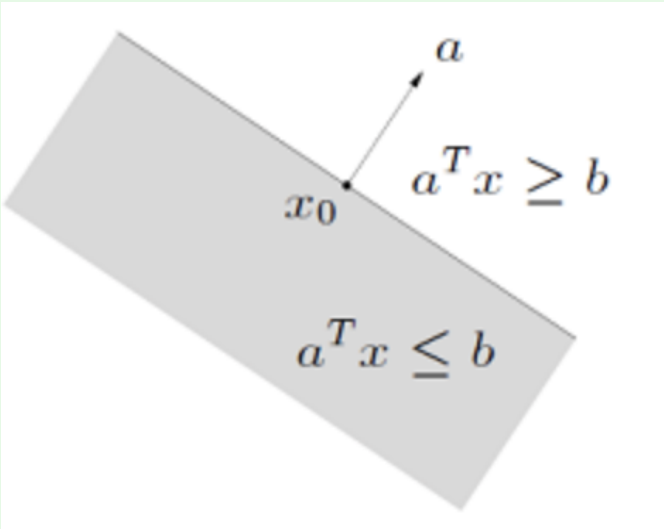

Figure 15: 2D Hyperplane which is the line then the half space, which is separating the line into the above and below b.

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.

4 Problem set 1

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

5.1 Non-Euclidean projection

Intuition:

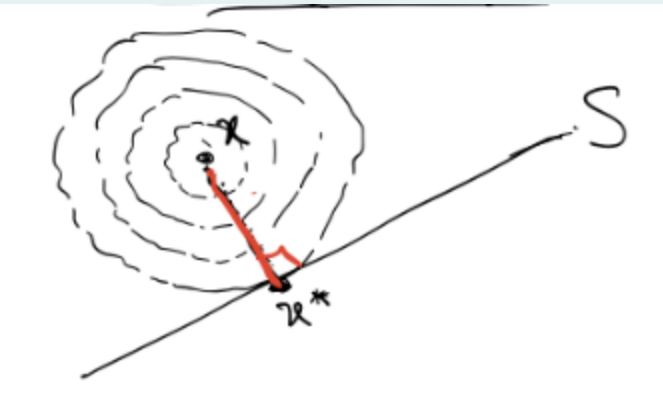

Figure 16: l2 norm

- **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.

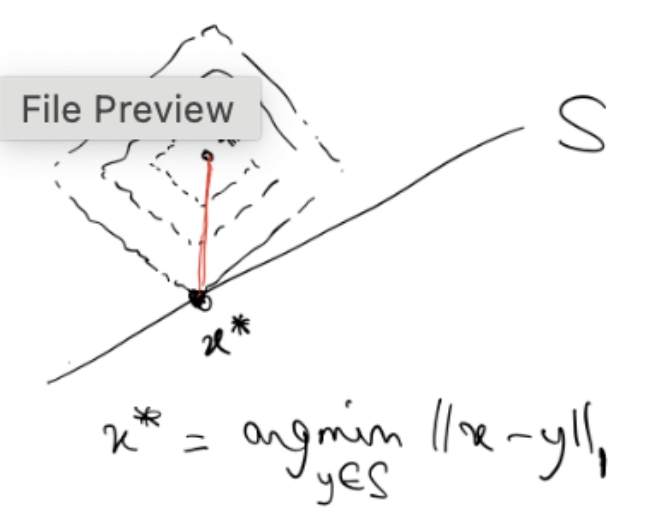

Figure 17: l1 norm

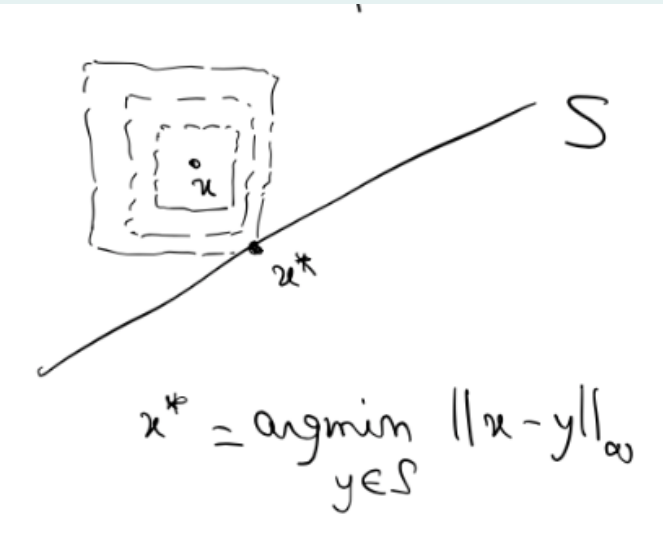

Figure 18: l-infinity norm

5.2 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.

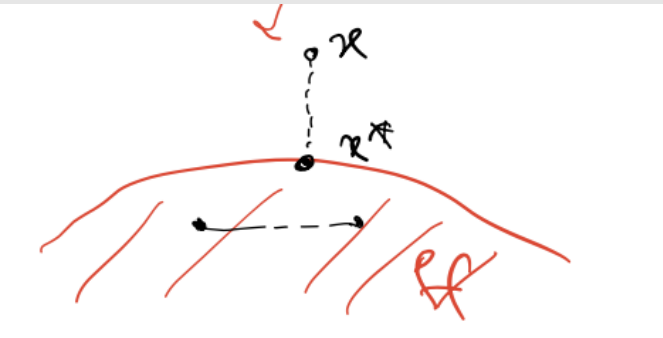

Figure 19: Function: Convex projection because we are projecting one input and getting one output

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

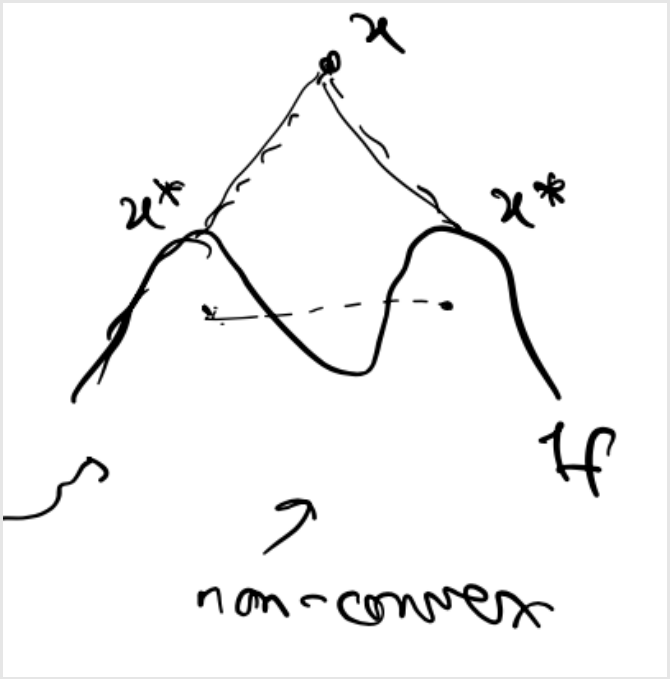

Figure 20: Not a function: Non Convex projection because we are projecting one input and getting two outputs

5.2.1 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. $f(x) = ||x||_p$

5.2.2 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:

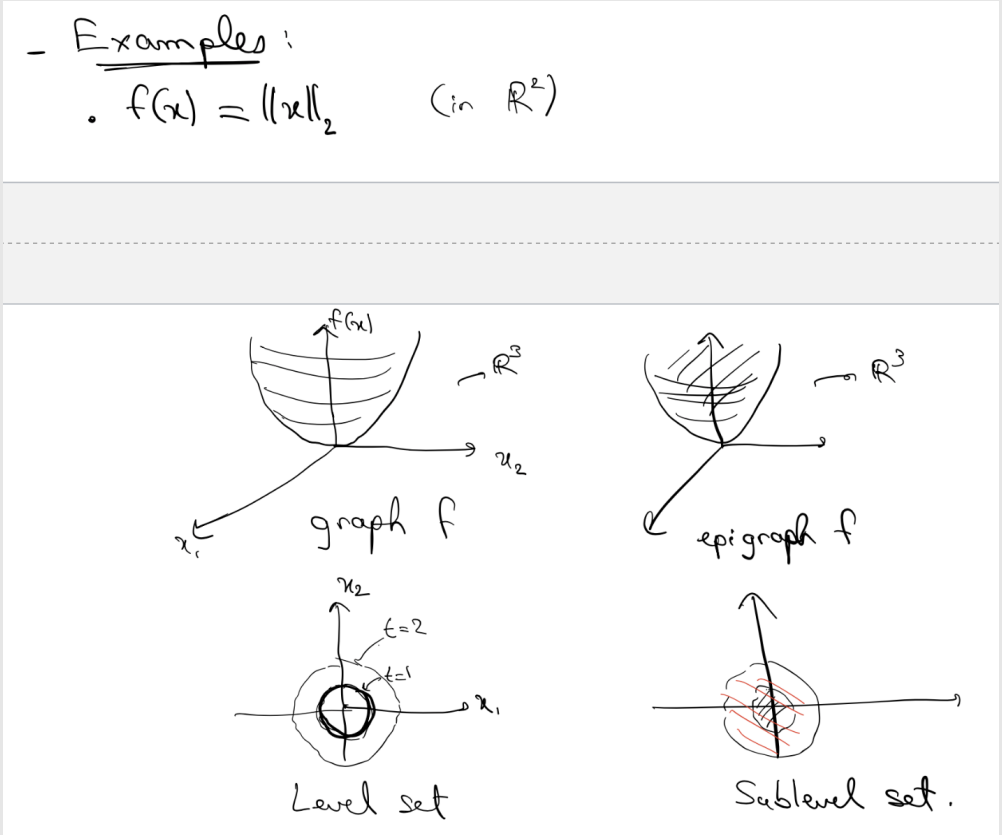

Figure 21: Example of the different common sets.

- 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:

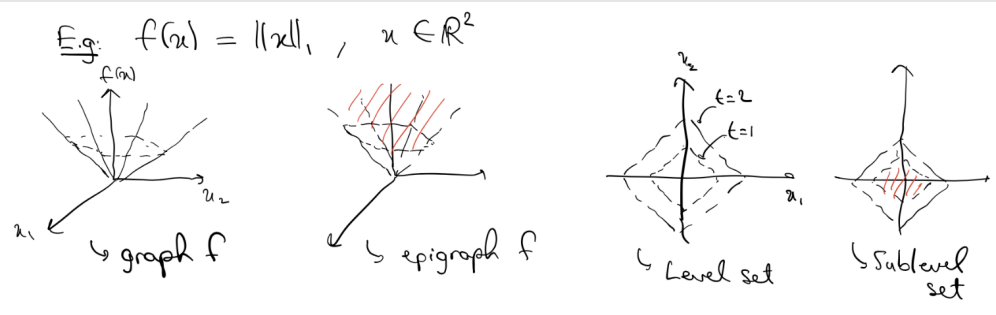

Figure 22: Example of the different common sets for the l1 norm.

Example:

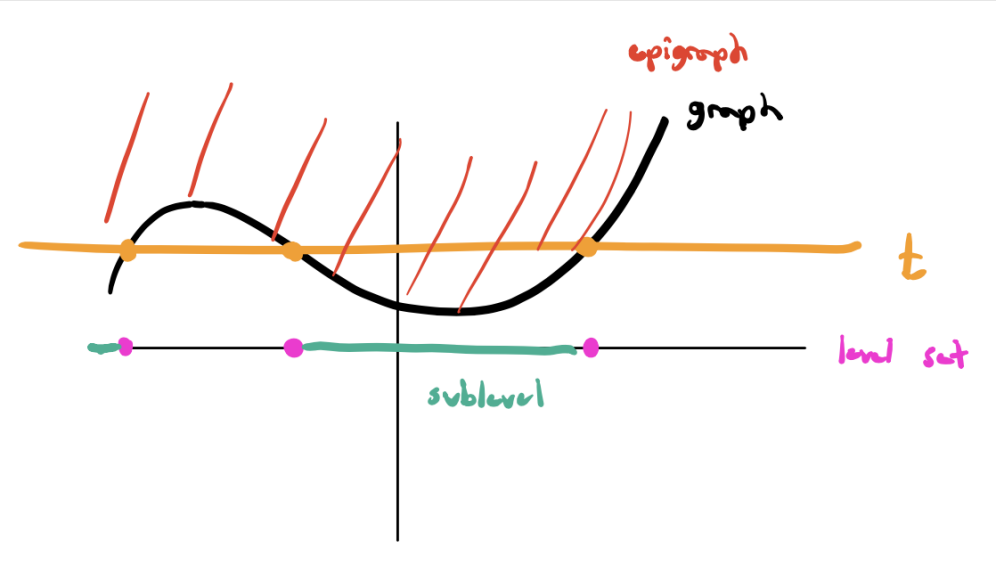

Figure 23: Example of a 2D function.

5.2.3 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:

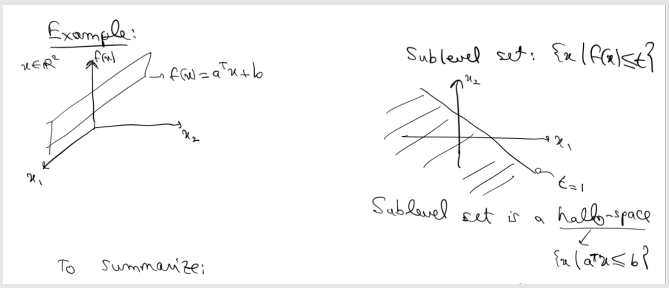

Figure 24: Example 2 of using the linear function, where the sublevel sets are the half-spaces.

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.

5.3 Function approximations

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

5.3.1 Gradients

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

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

5.3.2 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:

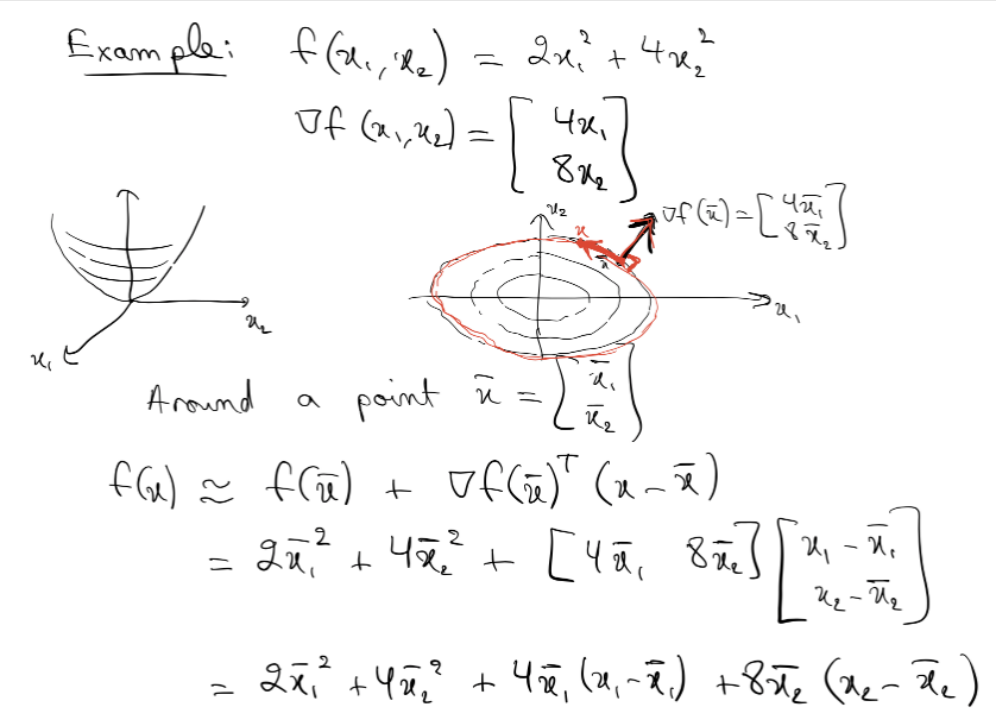

Figure 25: Example of using function approximation for first order. The arrow denotes $x - x^*$ along the level curve.

5.3.3 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).

5.3.4 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:

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

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

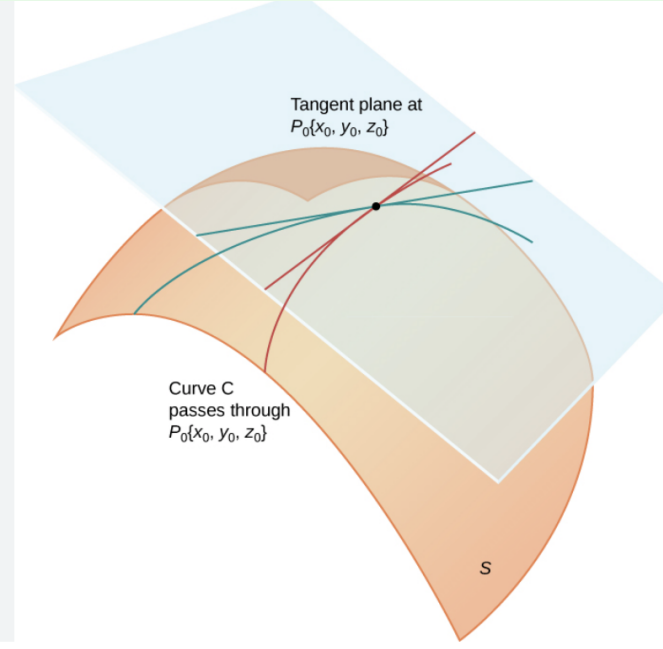

Figure 26: Tangent plane.

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:

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

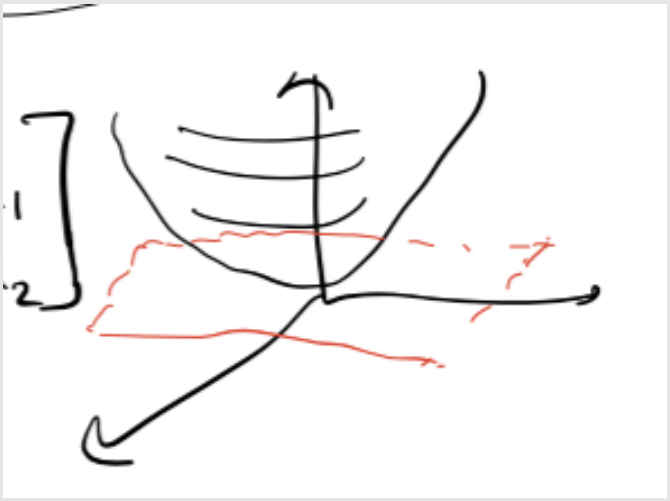

Figure 27: The tangent plane is defined as the xy-plane since the first order approximation is zero.

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:

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

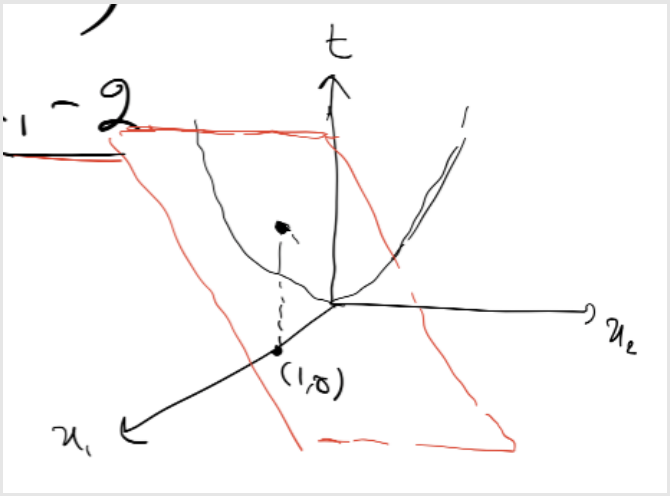

Figure 28: The tangent plane is defined as this plane.

5.3.5 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:

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

5.3.6 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:

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

5.3.7 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.

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

6.1 Matrices

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

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

6.1.1 Matrix Transpose

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

$$A^\top = \begin{bmatrix} (a^{(1)})^\top \\ \vdots \\ (a^{(n)})^\top \end{bmatrix} \in \mathbb{R}^{n \times m} \tag{19}$$

- **Note:** $(A^T A)^T = A^T A$

6.1.2 Matrix Multiplication

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

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

For regular matrix multiplication AB, we have:

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

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

6.1.3 Matrix vector multiplication

Definition:

$$Ax = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i a^{(i)} \in \mathbb{R}^{m \times 1}$$

6.1.4 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: $[a\ b] \begin{bmatrix} x \\ y \end{bmatrix} = ax + by$)

6.1.5 Types of Matrices

Definition:

- **Square matrix**: $m = n$
- **Diagonal matrix**: $a_{ij} = 0 \quad \forall i \neq j$
- **Identity matrix**: $a_{ij} = 0\ \forall i \neq j,\ a_{ii} = 1\ \forall i$
- **Triangular matrix**: $A = QR$ (R is upper triangular)
- **Orthogonal matrix**: $A^\top A = AA^\top = cI$
- **Orthonormal matrix**: $A^\top A = AA^\top = I$ (i.e. inverse of A is A^T)
- **Symmetric matrix**: $A = A^\top$

6.1.6 Matrices as Linear Maps

Definition: Matrices are used as linear maps, where:

$$y = Ax$$

$$\begin{bmatrix} y \end{bmatrix} \in \mathbb{R}^m \quad = \quad \begin{bmatrix} A \end{bmatrix} \in \mathbb{R}^{m \times n} \quad \begin{bmatrix} x \end{bmatrix} \in \mathbb{R}^n$$

This represents a **linear function**.

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

$$y = Ax + b$$

Intuition: Linear functions imply a linear mapping. Let $\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

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

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

   - Geometric Interpretation: A leaves the vector $\mathbf{v}$ unchanged (identity transformation).

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

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

   - Geometric Interpretation: A swaps the x and y components of the vector, flip it over the line $y = x$.

3. $A = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix}$

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

   - Geometric Interpretation: A scales the x component of $\mathbf{v}$ by a factor of $\frac{1}{\sqrt{2}}$ while leaving the y component unchanged.

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

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

- Geometric Interpretation: A rotates the vector $\mathbf{v}$ by $90°$ counterclockwise.

5. $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

$$A\mathbf{v} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta - \sin\theta \\ \sin\theta + \cos\theta \end{bmatrix}$$

- Geometric Interpretation: A rotates the vector $\mathbf{v}$ by an angle θ counterclockwise.

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

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

- Geometric Interpretation: A shears the vector $\mathbf{v}$, combining both components into the x-coordinate while leaving the y-coordinate unchanged.

6.2 Range Space

Definition:

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m \tag{23}$$

$$= \left\{ \sum_{i=1}^{n} x_i a^{(i)} \mid x \in \mathbb{R}^n \right\} \tag{24}$$

$$= \text{span}\{a^{(1)}, a^{(2)}, \ldots, a^{(n)}\} \tag{25}$$

- $\mathcal{R}(A)$ is a subspace (closed under addition and scalar multiplication).
- **Note:** A can consist of different column vectors that still span the space. E.g. In 2D, $(1, 0)$ and $(0, 1)$ can be used, but $(1, 1)$ and $(1, 0)$ can be used as well, which would have different A.
- **Note:** $R(A) = R(A^T)$ iff $A = A^T$ (i.e. symmetric).

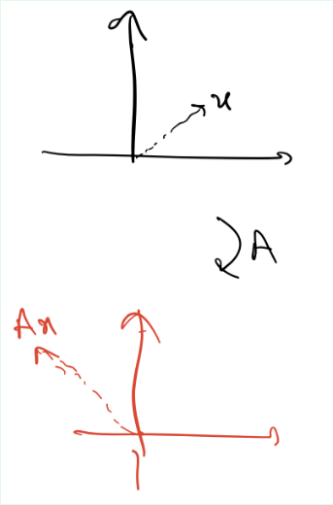

Figure 29: Range space.

6.3 Null Space

Definition: Set of all vectors that map to zero.

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

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

6.4 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)

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

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

Therefore, since $S \oplus S^\perp = V$, then 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 $\mathbb{R}^m$)

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

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

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

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

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,

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

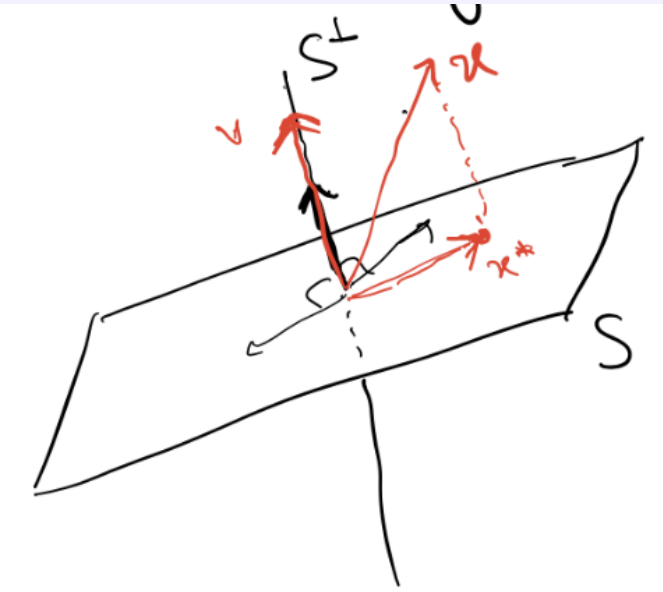

Figure 30: Orthogonal complement.

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:

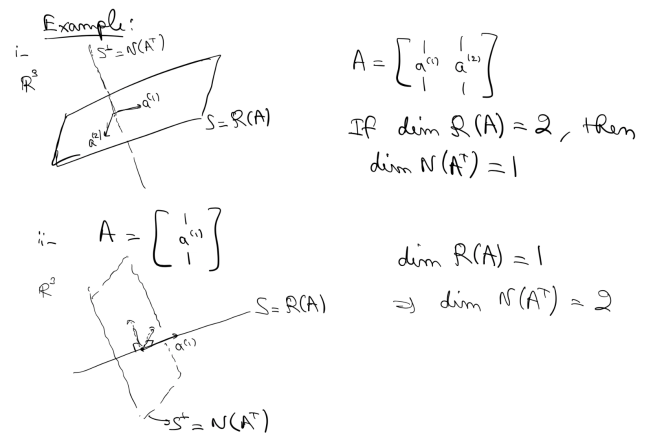

Figure 31: Example of fundamental theorem of linear algebra.

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

6.4.1 Rank of A

Definition:

$$\text{Rank}(A) = \dim(\mathcal{R}(A^T)) = \dim(\mathcal{R}(A))$$

- i.e. number of independent rows or independent columns.

6.5 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:

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

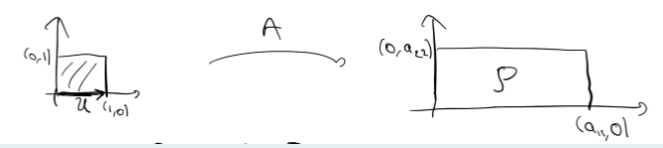

Figure 32: Diagonal

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:

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

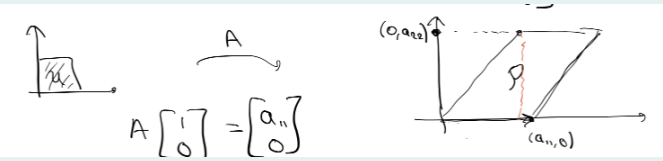

Figure 33: Upper triangular

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.

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

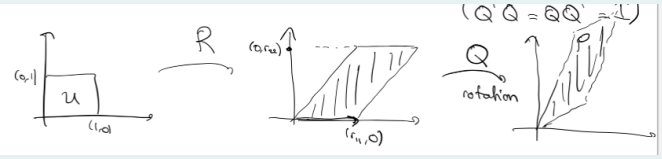

Figure 34: General matrix

6.6 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:

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

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

6.6.1 Frobenius Norm

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

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

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.

6.6.2 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 $\mathbb{R}^m$.

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

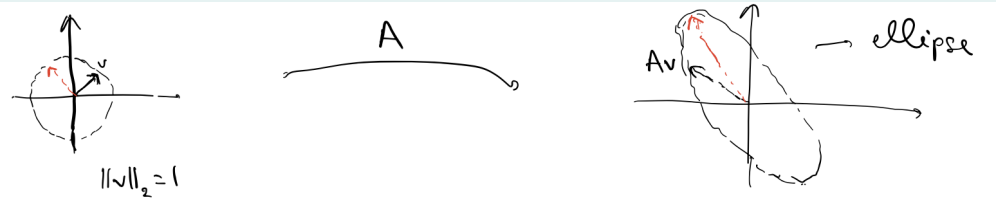

Figure 35: Matrix norm tht maximizes Av given v with unit norm.

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)

6.7 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 $(A - \lambda I)$.

6.7.1 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, so we want to find λ s.t. the determinant is 0.

Warning: n roots doesn't mean there are n distinct eigen values.
- e.g. For $(\lambda - 1)^2$, there are 2 roots, but there is only one eigenvalue.

6.7.2 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.
- i.e. If $p(x) = (x - \lambda)^m q(x)$, where λ is not a root of $q(x)$, then $AM(\lambda) = m$

Geometric multiplicity: $\text{GM}(\lambda) = \dim \mathcal{N}(A - \lambda I)$
- **Key:** If we know the rank of $A - \lambda I$, then we can use FToLA to find the $\text{GM}(\lambda) = \dim \mathcal{N}(A - \lambda I)$.

Theorem: In general, for any matrix $A \in \mathbb{R}^{n\times n}$ (square matrix),

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

6.7.3 Defective and non-defective

Definition:
Defective Matrix: If $GM(\lambda) < AM(\lambda)$ for some λ
- **Note:** Defective doesn't mean that it's not invertible.
- **Key:** Cannot construct n linearly independent vectors because $\exists$ a λ s.t. $\sum \text{AM} = n$ but $\sum \text{GM} < n$.

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).

6.7.4 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.

Warning: Holds for any defective or non-defective matrix.

- **Same eigenvalue (key subtle detail):** Eigenvectors corresponding to an eigenvalue can be linearly dependent or independent (**but not guaranteed to be l.i.**).
  - e.g. If $\dim((N)(A - \lambda I)) = 2$, you can find 2 basis vectors (i.e. eigenvectors), but **not** any two vectors.
  - Since we can choose two vectors to be linearly independent or dependent, you have to choose the linearly independent vectors.
    - i.e. Any vector in $(N)(A - \lambda I)$ is an eigenvector, but not every 2 eigenvectors in $(N)(A - \lambda I)$ is linearly independent.
    - e.g. If v is an eigenvector, then $2v$ is also an eigenvector.
  - **Construction:** So choose two linearly independent vectors to be the eigenvectors of that space.
- **Different eigenvalue:** For eigenvectors corresponding to different eigenvalues though, we are guaranteed for them to be linearly independent since they live in different null spaces.

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:

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

- 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 $\{u^{(2)}, \ldots, u^{(k)}\}$ is linearly dependent because $\lambda_1 - \lambda_i \neq 0$ since they are distinct, so there exists constants not all 0 to make the linear combination equal to 0

- i.e. not linearly independent as the only constants that would satisfy this is all constants being 0.

Continue with the same argument to show that $\{u^{(3)}, \ldots, u^{(k)}\}$ 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.

6.8 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 (i.e. non-defective by definition means that $AM(\lambda) = GM(\lambda)$) 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.

Warning:

- This is an extension of the previous theorem, since we are only considering non-defective matrices, which means $AM = GM$, so then you can find n linearly independent vectors.
  - **Defective:** We cannot find n linearly independent vectors because $GM < AM$.
- But, the same key subtlety pops up as for the same eigenvalue, we have to choose eigenvectors in that space that are linearly independent by **construction**.

Warning: You cannot orthnormalize these n linearly indpeendent vectors using Gram because they are not in the same space (i.e. if you Gram-Schmidt it, you will get non-eigenvectors).

- Therefore orthonormalizing vectors in the same space (i.e. same eigenvalue) retains the eigenvector property, but across spaces, will lose the eigenvector properties.

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.

6.8.1 What does this diagonalization mean?

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

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

- $\tilde{x} = U^{-1}x$: Coordinates of x in the new basis $\{u^{(1)}, \dots, u^{(n)}\}$.
  - Since x is a linear combination of eigenvectors, therefore, we are **transforming** to the basis of eigenvectors, where x can be written as $\sum_{i=1}^{n} \tilde{x}_i u^{(i)}$.
- $\Lambda\tilde{x}$ means that we are scaling $\tilde{x}_i$ by the corresponding eigenvalue λ_i.
- $y = Ax = U\Lambda\tilde{x}$ means that we are changing the coordinates back to the original basis.
  - Since the inverse of U^{-1} is U, therefore, after scaling by Λ, we are essentially transforming from the basis of eigenvectors to the original basis by applying the inverse transformation.

Warning:

- **Change of basis to eigenvector basis:** $U^{-1}x$ transforms x from the original basis to the eigenvector basis. The ith component of x is represented by $\tilde{x}_i$ (i.e. coordinate) in the new basis.
- **Scaling:** $\tilde{x}$ get scaled by their corresponding eigenvalues in Λ
- **Change of basis to original basis:** $\tilde{x}$ then gets transformed back to original basis by applying the inverse transformation of U^{-1}, which is U.

6.9 Application: Google's PageRank Algorithm

6.9.1 How to measure importance of a webpage

Intuition: The following two methods are insufficient as there are loop holes.

- **Web Search:** Mapping keywords to URLs.
- **Hyperlinks:** Hyperlinks that link to a webpage give an indication of how "important" the webpage is.

Therefore, the importance of a webpage should depend on the importance of the webpages linking to it.

6.9.2 Example: Random Walk across Webpages

Example: Bot performs a random walk across the webpages. Let $x_i^{(t)}$ be the probability that the bot is on page i at time t.

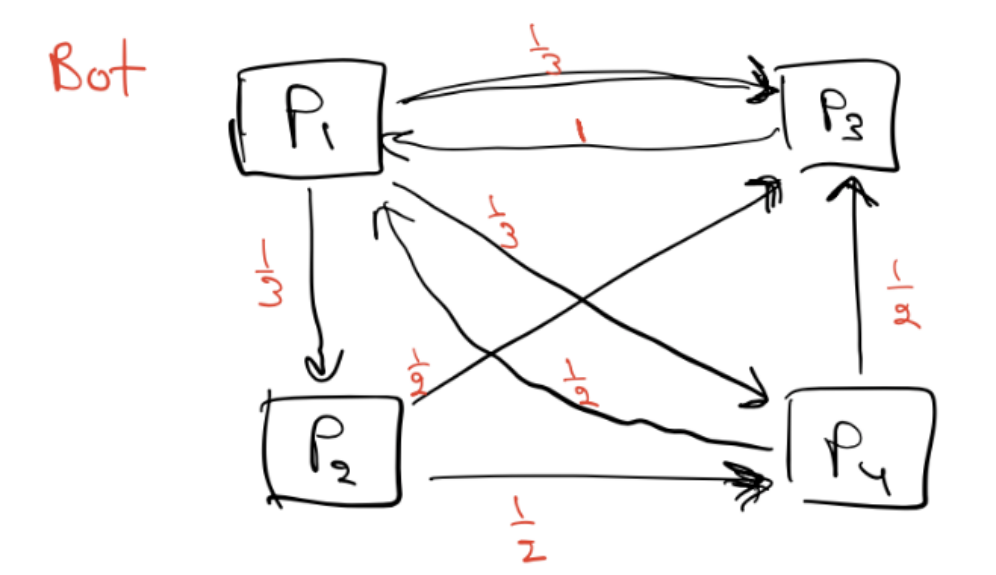

Figure 36: P are the pages, and the numbers in red are the probabilities of going to that page with the bot

$$x_1^{(t+1)} = 1 \cdot x_3^{(t)} + \frac{1}{2}x_4^{(t)}$$

$$x_2^{(t+1)} = \frac{1}{3}x_1^{(t)}$$

$$x_3^{(t+1)} = \frac{1}{3}x_1^{(t)} + \frac{1}{2}x_2^{(t)} + \frac{1}{2}x_4^{(t)}$$

$$x_4^{(t+1)} = \frac{1}{3}x_1^{(t)} + \frac{1}{2}x_2^{(t)}$$

This system can be written as:

$$\begin{bmatrix} x_1^{(t+1)} \\ x_2^{(t+1)} \\ x_3^{(t+1)} \\ x_4^{(t+1)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ x_3^{(t)} \\ x_4^{(t)} \end{bmatrix}$$

- P is the transition probability matrix.

6.9.3 Convergence to a Steady State Distribution

Definition: The evolution of the system is given by $x^{(t+1)} = Px^{(t)}$. As $t \to \infty$, $x^{(t)} \to x^{(\infty)}$ (under certain conditions) s.t. $Px^{(\infty)} = x^{(\infty)}$.

- i.e. The steady state $x^{(\infty)}$ is the eigenvector of P with eigenvalue 1. That is, $Px^{(\infty)} = x^{(\infty)}$.

6.9.4 Conditions for Convergence

Definition: When the Markov chain is both irreducible and aperiodic, then $x^{(t)} \to x^{(\infty)}$ as $t \to \infty$.

- An **aperiodic Markov chain** means the Markov chain does not exhibit periodic behavior, meaning it doesn't alternate between certain states but can visit them in an irregular pattern.
- An **irreducible Markov chain** means that there is a path from every state (webpage) to every other state, ensuring eventual convergence to a steady state.

Example: Examples of divergence

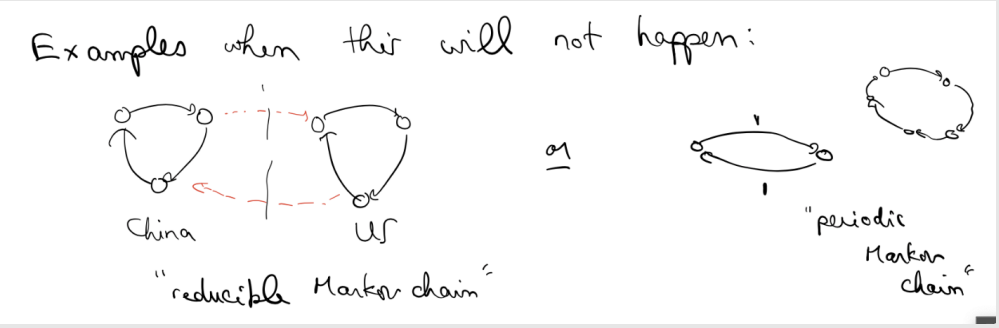

Figure 37: Non-convergent Markov chains

- Reducible Markov chains have disconnected chains, meaning that it will never converge since they are separated.
- Periodic Markov chains will always oscillate between webpages.
- **Key:** The convergence of the PageRank vector depends on the initial condition and the structure of the Markov chain.

6.9.5 Practical Modification

Intuition: In practice, the following adjustment ensures irreducibility and aperiodicity in web surfing:

- With probability 85%, the robot follows one of the hyperlinks from the current webpage.
- With probability 15%, the robot jumps to a totally random webpage on the Internet.

This ensures that the Markov chain is irreducible and aperiodic, i.e., $x^{(t)} \to x^{(\infty)}$.

6.9.6 Why is 1 an eigenvalue of P?

Derivation:

1. Fact: $\det(A) = \det\left(A^T\right)$.
2. Recall that the set of eigenvalues of a matrix is given by:

$$\{\lambda : \det(P - \lambda I) = 0\}$$

3. Using the fact that the determinant of a matrix equals the determinant of its transpose, we have:

$$\{\lambda : \det\left((P - \lambda I)^T\right) = 0\}$$

4. This implies:

$$\{\lambda : \det(P^T - \lambda I) = 0\}$$

5. Thus, the set of eigenvalues of P is the same as the set of eigenvalues of P^T.
6. Consider the following equation:

$$P^T \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 1 \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Since the sum of the columns equal to 1 for P, therefore, the sums of the rows equal to 1 for P^T.
- Therefore, if multiply P^T by the ones vector, then this will yield the sums of each row (i.e. 1)
- So, 1 is an eigenvalue of P.

6.9.7 What about other eigenvalues?

Definition: All other eigenvalues λ_i satisfy:

$$|\lambda_i| \leq 1$$

Derivation:

1. We know that $Px = \lambda x$. By applying P t times, we get:

$$P(P(\dots(Px))) = P^t x = \lambda^t x$$

2. Thus, $P^t x = \lambda^t x$, and since every entry of $P^t x \leq 1$ because it's a probability distribution. Therefore, we should have that every entry of $P^t x$ is bounded.
3. If $|\lambda_i| > 1$, then λ_i^t diverges to infinity as $t \to \infty$. But, for stability:
   - Since $P^t x = \lambda^t x$ and $P^t x \leq 1 \implies |\lambda_i| \leq 1 \quad \forall i$

6.9.8 Convergence via Power Iteration

Definition: We can express the eigenvalue decomposition of the matrix P as follows:

1. The eigenvalue decomposition of P is given by:

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

   where U is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues.
2. After t iterations of applying P, we have:

$$P^t x = (U \Lambda U^{-1})(U \Lambda U^{-1}) \dots (U \Lambda U^{-1}) x = U \Lambda^t U^{-1} x$$

3. Breaking this down, we get:

$$P^t x = U \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \lambda_2^t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n^t \end{bmatrix} U^{-1} x$$

4. As $t \to \infty$, if $|\lambda_i| < 1$ for $i = 2, 3, \dots, n$, then:

$$\lambda_i^t \to 0$$

   Hence, the system converges to:

$$a_1 u^{(1)}$$

   - $u^{(1)}$ is the eigenvector corresponding to the largest eigenvalue $\lambda_1 = 1$.
   - $U^{-1} x = a$
5. By normalizing the vector $x^{(t)}$ at each iteration, we obtain $u^{(1)}$, which corresponds to the eigenvalue $\lambda_1 = 1$. This process is known as the **power iteration method**.

7 Problem set 2

7.1 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:

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

7.2 Finding rank and dimension/basis for range and null space of A and A transpose

Process:

1. Find the REF of the matrix (i.e. each leading entry in a row is to the right of the leading entry in the previous row, with all entries below leading entries being zeros).
2. rank(A) is the number of pivots.
3. dim($N(A^T)$) = Total num. of rows − rank(A)
4. dim($N(A)$) = Total num. of cols − rank(A)
   - **Note:** This makes sense from the FTLA, since the total is basically the universal space.
5. $R(A)$: Span of column vectors where there is a pivot (can use REF or original matrix's vectors).
6. $R(A^T)$: Span of row vectors where there is a pivot (can use REF or original matrix's vectors).
7. $N(A)$: Solve $Ax = 0$ and find x, where there is probably a free variable. Check by subbing x into $Ax = 0$
   - You can use the REF of A for this.
8. $N(A^T)$: Solve $A^T x = 0$ and find x, where there is probably a free variable. Check by subbing x into $A^T x = 0$
   - **Key:** Have to find the REF of A^T, which is not equivalent to the REF of A.

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

8.1 QA

Summary:

1. Why for defective matrices, we cannot form n l.i. vectors?
   - B/c $GM < AM$, which means that the $\sum AM = n$, but $\sum GM < n$, so there won't exist n l.i. vectors.
2. Watch 3b1b for eigenvectors and eigenvalues.
3. Are we describing the shape in the new coordinate system for quadratic constraints?
   - No, we are just using this transformed coordiante to interpret A easier because then it becomes a scaling of Λ in this coordinate system. Afterwards, we apply the rotation by seeing where $1, 0$ and $0, 1$ vectors go.
4. Watch MyWhyU linear algebra 54-59 for REF.
5. Proofs will be one part of a problem in the midterm. But there will also be computational problems.
6. Do we have to normalize the eigenvectors for decomposition?
   - Always normalize the vectors. You just have to show they are l.i. for eigendecomposition, but for spectral decomposition, you normalzie the orthogonal vectors. But as a convention, **Always noramlzie the vectors**.
7. For a symmetric matrix, and using spectral decomposition, will we always get an orthonormal basis or do we have to construct this ourselves?
   - No, you have to normalize the vectors to get an orthogonal matrix.
8. Doesn't the rotation depend on how we set up the orthogonal matrix?
   - No, see this by applying $1, 0$ and $0, 1$ to the rotation matrix and seeing where they get transformed in the x space.
9. How to determine the rotations of a matrix?
   - Apply $1, 0$ and $0, 1$ in the tilde space to the rotation matrix to see where x lands after the transformation.
10. For eigendecomposiiton of a diagonalizable matrix, aren't they n linearly independent eigen vectors so why can't we apply gram Schmidt to make them orthogonal. I know for the theorem we proved for symmetric matrices that they have to orthogonal for different eigen values but still
    - Because we applying GS across different null spaces can results in vectors that do not span the same space.

8.2 Symmetric matrices

Definition: Set of real symmetric matrices is

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

- i.e. a matrix A is symmetric if $a_{ij} = a_{ji}\ \forall i, j$

Set of complex symmetric matrices is

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

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

8.2.1 Example of symmetric matrix: Hessian Matrix

Example: Recall: The Hessian matrix $\nabla^2 f$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ with continuous partial derivatives (i.e. must be continuous for symmetric matrix) is a symmetric matrix:

$$[\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} = [\nabla^2 f]_{ji}.$$

Example: Consider a quadratic function $q : \mathbb{R}^n \to \mathbb{R}$:

$$
\begin{aligned}
q(x) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j + \sum_{i=1}^{n} c_i x_i + d \\
&= \frac{1}{2}\begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} c_1 & \dots & c_n \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + d \\
&= \frac{1}{2}x^\top Ax + c^\top x + d.
\end{aligned}
$$

Gradient of the Quadratic Function
Let

$$q(x) = \frac{1}{2}x^\top Ax + c^\top x + d.$$

Then,

$$\nabla(c^\top x) = \begin{bmatrix} \frac{\partial}{\partial x_1}(c^\top x) \\ \vdots \\ \frac{\partial}{\partial x_n}(c^\top x) \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c.$$

Now let's solve $\nabla\left(\frac{1}{2}x^\top Ax\right)$, for the kth component:

$$
\begin{aligned}
\frac{\partial}{\partial x_k}\left(\frac{1}{2}x^\top Ax\right) &= \frac{\partial}{\partial x_k}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j\right) \\
&= \frac{\partial}{\partial x_k}\left(\frac{1}{2}a_{kk}x_k^2 + \frac{1}{2}\sum_{i=1,i\neq k}^{n} a_{ik}x_i x_k + \frac{1}{2}\sum_{j=1,j\neq k}^{n} a_{kj}x_k x_j\right) \\
&\quad \text{1st term: } i=j=k\text{, 2nd term: } j=k\text{, 3rd term: } i=k \\
&= a_{kk}x_k + \sum_{i=1,i\neq k}^{n} a_{ik}x_i \quad \text{since } a_{ij} = a_{ji} \\
&= \sum_{i=1}^{n} a_{ik}x_i \quad \text{bringing the term in} \\
&= \begin{bmatrix} a_{k1} & \dots & a_{kn} \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.
\end{aligned}
$$

Therefore, generalizing this to all the components, then

$$\nabla\left(\frac{1}{2}x^T Ax\right) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = Ax$$

Thus,

- $\nabla\left(\frac{1}{2}x^\top Ax\right) = Ax$
- $\boxed{\nabla q(x) = Ax + c}$ is the final gradient.

Hessian:

$$\begin{aligned}\left[\nabla^2 g(x)\right]_{ij} &= \frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} g(x)\right) \\ &= \frac{\partial}{\partial x_j}\left(\sum_{i=1}^{n} a_{ij} x_j\right) = a_{ji} = a_{ij}\end{aligned}$$

Therefore, generalizing this to the entire Hessian:

$$\nabla^2 g(x) = A$$

- **Note:** If A is not symmetric, then the Hessian becomes $B = \frac{A + A^T}{2}$

Taylor approximation: The second-order Taylor approximation of $g(x)$ around x_0:

$$\begin{aligned}q(x) &\approx q(x_0) + \nabla q(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 q(x_0)(x - x_0) \\ &= \left(\frac{1}{2} x_0^\top A x_0 + c^\top x_0 + d\right) + (Ax_0 + c)^T (x - x_0) + \frac{1}{2}(x - x_0)^T A (x - x_0) \\ &= \left(\frac{1}{2} x_0^\top A x_0 + c^\top x_0 + d\right) + (x_0^T A^T + c^T)(x - x_0) + \frac{1}{2}(x^T A x - x^T A x_0 - x_0^T A x + x_0^T A x_0) \\ &= \left(\frac{1}{2} x_0^\top A x_0 + c^\top x_0 + d\right) + x_0^T A x - x_0^T A x_0 + c^T x - c^T x_0 + \frac{1}{2}(x^T A x - 2 x_0^T A x + x_0^T A x_0) \\ &= \left(\frac{1}{2} x_0^\top A x_0 + c^\top x_0 + d\right) + x_0^T A x - x_0^T A x_0 + c^T x - c^T x_0 + \frac{1}{2} x^T A x - x_0^T A x + \frac{1}{2} x_0^T A x_0 \\ &= x_0^\top A x_0 + c^\top x_0 + d + x_0^T A x - x_0^T A x_0 + c^T x - c^T x_0 + \frac{1}{2} x^T A x - x_0^T A x \\ &= c^\top x + d + \frac{1}{2} x^T A x \qquad = q(x)\end{aligned}$$

- This is independent of x_0 (check why this holds).

Warning: If A is not symmetric, we can replace A by $\frac{A + A^\top}{2}$ without changing the evaluation of the function.

- e.g. $4x_1x_2 + 5x_2x_1 = 9x_1x_2$. This is non-symmetric since $a_{12} \neq a_{21}$, but we can symmetrize this by taking the average of the two points
  - $4.5x_1x_2 + 4.5x_2x_1 = 9x_1x_2$. This symmetry was found as $\frac{a_{12} + a_{21}}{2}$

Therefore, we can always assume that A is symmetric

- Since we can always change a non-symmetric matrix A into a symmetric matrix $\frac{A + A^\top}{2}$ since $\left(\frac{A + A^\top}{2}\right)^T = \frac{A + A^\top}{2}$

Therefore, in general, always replace with $\frac{A + A^\top}{2}$ because if it's symmetric already, then it becomes A. If not it becomes $B = \frac{A + A^\top}{2}$, which is a symmetric matrix by construction.

8.2.2 Theorem

Theorem: Let $A \in S^n$ be a real symmetric matrix. Let $\lambda_1, \ldots, \lambda_n$ be its eigenvalues (not necessarily distinct). Then:

1. $\lambda_1, \ldots, \lambda_n$ are real, and $\text{AM}(\lambda_i) = \text{GM}(\lambda_i)$ for all i.
2. Eigenvectors corresponding to different eigenvalues are orthogonal.

Warning: The implication from the theorem

1. This means all symmetric matrices are diagonalizable.
2. $U^T U = I$ (i.e. $U^{-1} = U^T$)

Derivation: **1)** Let λ be an eigenvalue with eigenvector u.

$$Au = \lambda u \tag{1}$$

Let's consider the conjugate transpose:

$$\begin{aligned} &(Au)^H = (\lambda u)^H \\ \implies &u^H A^H = \bar{\lambda} u^H \end{aligned} \tag{2}$$

Now set up two equivalent equations starting with $u^H Au$:

- $$u^H Au = \lambda u^H u \quad \text{Multiply equation (1) by } u^H$$
- $$\begin{aligned} u^H Au &= u^H A^H u \quad \text{since A is symmetric } A^H = A \\ &= \bar{\lambda} u^H u \quad \text{Multiply equation (2) by } u \end{aligned}$$

  – A is real and symmetric and $A = A^H$ since the Hermitian takes the transpose and the conjugage.
  - $A^H = A^T$ since the conjugate of a real matrix is the real matrix
  - $A = A^T$ because it's symmetric.
  - Therefore, $A^H = A$

Thus, we have:

$$(\lambda - \bar{\lambda}) u^H u = 0$$

So, $\lambda = \bar{\lambda}$, meaning that λ is real.

For the proof of $\text{AM}(\lambda) = \text{GM}(\lambda)$, check Theorem 4.1 in the OptM book.

Derivation: **2)** Let λ_i and λ_j be two different eigenvalues, and let $v^{(i)}$ and $v^{(j)}$ be the corresponding eigenvectors.

$$Av^{(i)} = \lambda_i v^{(i)} \tag{1}$$

$$Av^{(j)} = \lambda_j v^{(j)} \tag{2}$$

Now set up two equivalent equations starting with $(v^{(j)})^T Av^{(i)}$:

- $$(v^{(j)})^T Av^{(i)} = \lambda_i (v^{(j)})^T v^{(i)} \quad \text{Multiply (1) by } (v^{(j)})^T$$
- $$\begin{aligned} (v^{(j)})^T Av^{(i)} &= (v^{(j)})^T A^T v^{(i)} \quad \text{Since } A = A^T \\ &= (Av^{(j)})^T v^{(i)} \quad (AB)^T = B^T A^T \\ &= \lambda_j (v^{(j)})^T v^{(i)} \end{aligned}$$

Thus, we have:

$$(\lambda_i - \lambda_j)(v^{(j)})^T v^{(i)} = 0$$

Since $\lambda_i \neq \lambda_j$, we must have:

$$(v^{(j)})^T v^{(i)} = 0$$

Therefore, $v^{(i)}$ and $v^{(j)}$ are orthogonal.

Warning: For these two derivations: $u^H Au$ and $(v^{(j)})^T Av^{(i)}$ were used to use the properties of symmetric matrices.
- $(u^H Au)^H = u^H Au$

8.3 Spectral decomposition

8.3.1 Spectral theorem

Theorem: Let $A \in S^n$ be a real symmetric matrix. Let $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ be its eigenvalues.

Then, there exists a set of real eigenvectors $u^{(1)}, \dots, u^{(n)}$, which are orthonormal (i.e., $u^{(i)} \perp u^{(j)}$ for $i \neq j$, and $\|u^{(i)}\| = 1$ for all i).

Equivalently, the eigendecomposition of A becomes:

$$A = U\Lambda U^T$$

- $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$
- $U = \begin{bmatrix} u^{(1)} & \dots & u^{(n)} \end{bmatrix}$
- **Note:** This is called the "spectral decomposition" of A.
  - Previously, we had $A = U\Lambda U^{-1}$, but for a symmetric matrix $U^{-1} = U^T$.

Derivation: We already proved that $\lambda_1, \dots, \lambda_n$ are real.

Without loss of generality (WLOG), we can take the eigenvectors to be real. Why?

1. $A\mathbf{v} = \lambda\mathbf{v}$ implies
$$\text{Re}\{A\mathbf{v}\} = \text{Re}\{\lambda\mathbf{v}\}$$
2. Since λ is real, this becomes
$$A\,\text{Re}\{\mathbf{v}\} = \lambda\,\text{Re}\{\mathbf{v}\}$$
3. Therefore, $\text{Re}\{\mathbf{v}\}$ is an eigenvector with eigenvalue λ.

Now, let's prove that the set of real eigenvectors are orthonormal:

1. Now, since $\text{AM}(\lambda_i) = \text{GM}(\lambda_i) = m_i$ (since diagonalizable), we can take m_i eigenvectors $u^{(i,1)}, \dots, u^{(i,m_i)}$ which form an orthonormal basis of $\mathcal{N}(A - \lambda_i I)$
   - i.e. by construction, we can form an orthonormal basis for this space by using Gram-Schmidt for example
   - λ_1: $u^{(1)}, \dots, u^{(1,m_1)}$ (these vectors can be chosen to be orthogonal within this space.)
   - λ_2: $u^{(2)}, \dots, u^{(2,m_2)}$ (these vectors can be chosen to be orthogonal within this space.)
2. Previously, we showed that across different $\mathcal{N}(A - \lambda_i I)$'s, the $u^{(i)}$ are orthogonal.
3. Therefore, $\{u^{(1,1)}, \dots, u^{(1,m_1)}, u^{(2,1)}, \dots, u^{(2,m_2)}, \dots\}$ forms an orthonormal basis of n eigenvectors

Now we can assemble everything together:

1. So,
$$Au^{(i,j)} = \lambda_i u^{(i,j)}$$
2. Therefore, if we let $U = \begin{bmatrix} u^{(1,1)} & \dots & u^{(n,m_n)} \end{bmatrix}$. Then,
$$AU = U\Lambda$$

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

3. Since the columns of U form an orthonormal basis, we have

$$U^\top U = \begin{bmatrix} (u^{(1)})^\top \\ \vdots \\ (u^{(n)})^\top \end{bmatrix} \begin{bmatrix} u^{(1)} & \cdots & u^{(n)} \end{bmatrix} = I$$

4. Therefore, $U^{-1} = U^\top$, meaning U is an orthogonal matrix.
5. Multiply both sides of $AU = U\Lambda$ by $U^\top$:

$$A = U\Lambda U^\top$$

8.3.2 Difference between Eigendecomposition and Spectral Decomposition

Definition: Eigendecomposition of a diagonalizable matrix

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

- $U = \begin{bmatrix} u^{(1)} & \dots & u^{(n)} \end{bmatrix}$ eigenvectors form a basis for $\mathbb{R}^n$
- $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$

Eigendecomposition of a symmetric matrix A

$$A = U\Lambda U^T$$

- $U = \begin{bmatrix} u^{(1)} & \dots & u^{(n)} \end{bmatrix}$ eigenvectors form an orthonormal basis

8.4 Orthogonal matrices

Definition: U is an orthogonal matrix if

$$U = \begin{bmatrix} u^{(1)} & \cdots & u^{(n)} \end{bmatrix},$$

- $(u^{(i)})^\top u^{(j)} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$
- i.e. $U^{-1} = U^T$ or $UU^T = U^TU = I$
- **Note:** Columns of an orthogonal matrix are orthonormal.

8.4.1 What happens when we multiply a vector by an orthogonal matrix?

Derivation:

$$y = Ux$$

Answer: Ux is a rotation of x, plus a possible flip.

Why? Let's start with proving two facts:

- **Why? Fact 1**: If U is orthogonal, then

$$\|Ux\|_2 = \|x\|_2.$$

– **Proof**:

$$\begin{aligned}\|Ux\|_2^2 &= (Ux)^\top (Ux) \\ &= x^\top U^\top U x \\ &= x^\top I x \\ &= \|x\|_2^2.\end{aligned}$$

- **Fact 2**: Suppose x and y have an angle Θ between them. Then Ux and Uy have the same angle Θ between them.
  – **Proof (1-2)**: By previous lecture, we showed that

$$\cos\Theta = \frac{\langle x, y\rangle}{\|x\|_2\|y\|_2} = \frac{x^\top y}{\|x\|_2\|y\|_2}.$$

  – Now, prove 2-1,

$$\begin{aligned}\frac{\langle Ux, Uy\rangle}{\|Ux\|_2\|Uy\|_2} &= \frac{(Uy)^\top (Ux)}{\|Ux\|_2\|Uy\|_2} \\ &= \frac{y^\top U^\top U x}{\|x\|_2\|y\|_2} \quad \text{by fact 1} \\ &= \frac{y^\top x}{\|x\|_2\|y\|_2} = \cos\Theta \quad \text{by } U^T U = I\end{aligned}$$

Thus, the angle remains unchanged after applying U.

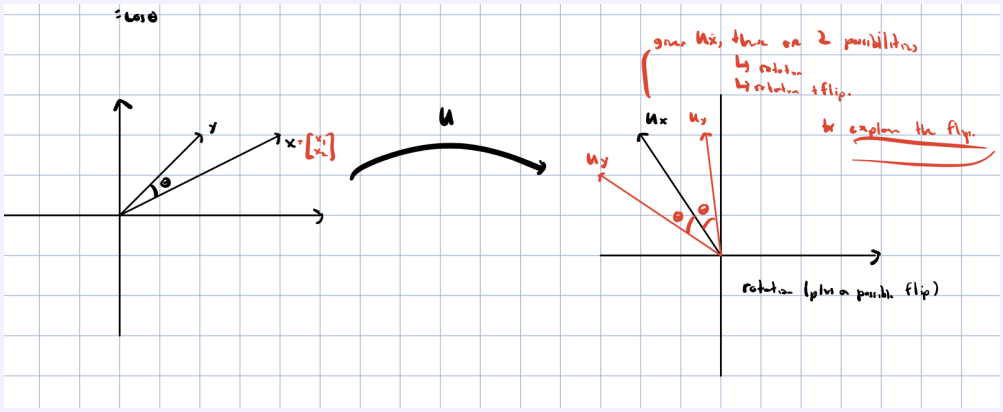

Figure 38: Orthogonal matrix performed on two vectors causes a rotation and possibly a flip.

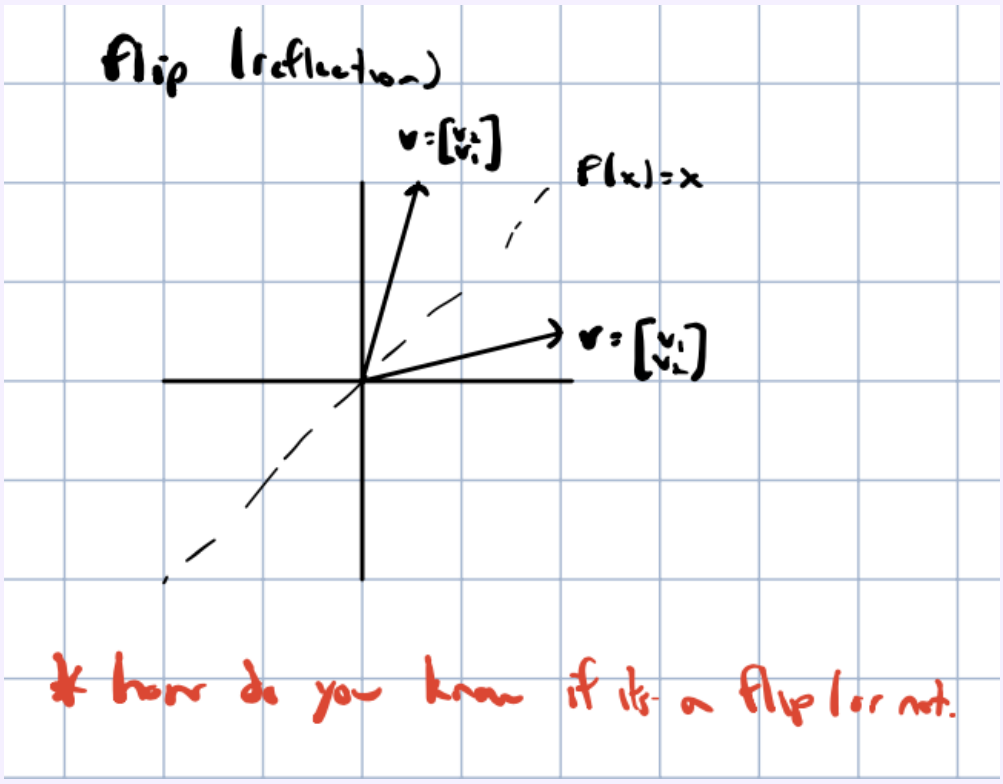

Figure 39: Flip.

Example: Examples of flipping in orthogonal matrices.

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \\ x_3 \end{bmatrix}$$

This is flipping across the x_1, x_2 hyperplane.

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix}$$

This is flipping across the x_3, x_1 hyperplane.

8.4.2 What kind of transformation corresponds to multiplying by a symmetric matrix?

Derivation:

$$y = Ax = U\Lambda U^\top x$$
$$U^\top y = \Lambda U^\top x$$
$$\tilde{y} = \Lambda \tilde{x}$$

- $\tilde{y}$: $U^T y$
- $\tilde{x}$: $U^T x$

So, an interpretation of $y = Ax$ is:
1. Rotation (or flip) of x to get $\tilde{x}$, where $\tilde{x} = U^\top x$.
2. The i-th component of $\tilde{x}$ is scaled by λ_i, where $\tilde{y} = \Lambda \tilde{x}$.
3. Rotation (or flip) $\tilde{y}$ back to $\mathbf{y}$, where $y = U\tilde{y}$.

Now, let's look at $\tilde{x} = U^\top x$:

$$\tilde{x} = U^\top x = \begin{bmatrix} (u^{(1)})^\top \\ \vdots \\ (u^{(n)})^\top \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \langle u^{(1)}, x \rangle \\ \vdots \\ \langle u^{(n)}, x \rangle \end{bmatrix}$$

Recall: The projection of x onto a subspace spanned by $u^{(i)}$ is just $\langle u^{(i)}, x \rangle u^{(i)}$

- So, we can think of $U^\top x$ as the coordinates of $\mathbf{x}$ in the orthonormal basis $\{u^{(1)}, \ldots, u^{(n)}\}$.

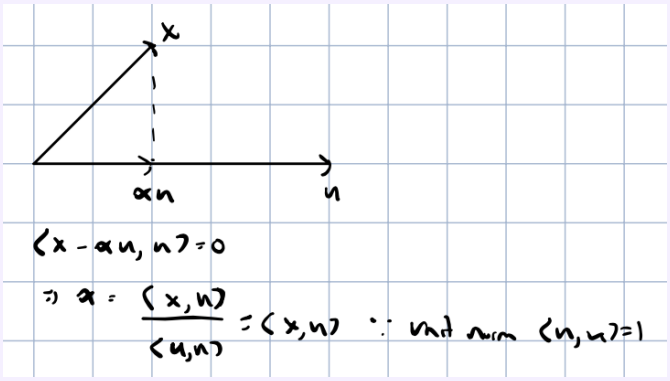

Figure 40: Projection

Warning: The interpretation is the same as the eigendecomposition for non-defective (but not necessarily symmetric matrices), but the change of basis is represented by rotations and flips since we have orthonormal matrices.

8.5 Positive semidefinite matrices and positive definite matrices

Definition: A symmetric matrix $A \in S^n$ is Positive Semi-Definite (PSD) if:

$$x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n$$

A symmetric matrix $A \in S^n$ is Positive Definite (PD) if:

$$x^T A x > 0 \quad \forall x \neq 0$$

- **Denote:** The set of PSD matrices as S^n_+, and the set of PD matrices as S^n_{++}.
- **Notation**

$$A \succeq 0 \quad \Leftrightarrow \quad A \in S^n_+$$
$$A \succ 0 \quad \Leftrightarrow \quad A \in S^n_{++}$$

Warning: $x \neq 0$ because if $x = 0$, then the product is 0 and A can be anything.

8.5.1 Theorem on PSD,PD for Eigenvalues

Theorem: A symmetric matrix $A \in S^n$ is:
- PSD if and only if $\lambda_i(A) \geq 0 \quad \forall i$
- PD if and only if $\lambda_i(A) > 0 \quad \forall i$

Derivation: We can write $A = U \Lambda U^T$, where Λ contains the eigenvalues of A since it's symmetric. Then:

$$x^T A x = x^T U \Lambda U^T x = \tilde{x}^T \Lambda \tilde{x}$$

Where $\tilde{x} = U^T x$. Expanding the right-hand side:

$$\tilde{x}^T \Lambda \tilde{x} = \begin{bmatrix} \tilde{x}_1 & \cdots & \tilde{x}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{bmatrix}$$
$$= \sum_{i=1}^{n} \lambda_i \tilde{x}_i^2$$

- **Right to left:** If $\lambda_i \geq 0$ for all i, then $x^T A x \geq 0$ since $\lambda_i \tilde{x}_i^2 > 0$ as $\tilde{x}_i^2 > 0$, proving that A is PSD.
- **Left to right:** Conversely, if $x^T A x \geq 0$ for all x, then we must have $\lambda_i \geq 0$ for all i.
  - **Contradiction:** Since otherwise, we could find an x such that $x^T A x < 0$, which contradicts the definition of PSD (i.e. $\tilde{x} = 1$ corresponding to the eigenvalue that is $\lambda_i < 0$) to make the $x^T A x < 0$.
- Similar proof for PD.

8.6 Ellipsoids

Definition: PSD and PD matrices are used to define ellipsoids. For a PD matrix,

$$\mathcal{E} = \left\{ x \in \mathbb{R}^n \mid (x - x^{(0)})^T P^{-1} (x - x^{(0)}) \leq 1 \right\}$$
$$\mathcal{E} = \left\{ x \in \mathbb{R}^n \mid x^T P^{-1} x - 2(x^{(0)})^T P^{-1} x + (x^{(0)})^T P^{-1} x^{(0)} \leq 1 \right\}$$

- P: PD matrix (i.e. defines the shape)
- $x^{(0)}$: Centered at $x^{(0)}$
- **Note:** $(x^{(0)})^T P^{-1} x = x^T P^{-1} x^{(0)}$
- **Note:** This is a quadratic function in x.

Derivation: How to visualize the ellipsoid? $P = U \Lambda U^T$ U is orthogonal, Λ has eigenvalues with $\lambda_i > 0$

$$\begin{aligned}
(x - x^{(0)})^T P^{-1} (x - x^{(0)}) &= \bar{x}^T P^{-1} \bar{x} \quad \text{where } \bar{x} = x - x^{(0)} \\
&= \bar{x}^T \left(U \Lambda U^T \right)^{-1} \bar{x} \quad \text{since } P \text{ is symmetric} \\
&= \bar{x}^T (U^T)^{-1} \Lambda^{-1} U^{-1} \bar{x} \quad \text{since for } X \text{ and } Y \text{ invertible } (XY)^{-1} = Y^{-1} X^{-1} \\
&= \bar{x}^T U \Lambda^{-1} U^T \bar{x} \quad \text{since } U^T = U^{-1} \\
&= \tilde{x}^T \Lambda^{-1} \tilde{x} \quad \text{where } \tilde{x} = U^T \bar{x} \\
&= \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix}^T \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_n} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix} \\
&= \sum_{i=1}^{n} \frac{\tilde{x}_i^2}{\lambda_i} \\
&= \sum_{i=1}^{n} \left(\frac{\tilde{x}_i}{\sqrt{\lambda_i}} \right)^2
\end{aligned}$$

Process:

1. Plot in the $\tilde{x}$-axis: $\mathcal{E} = \left\{ \tilde{x} \in \mathbb{R}^n \;\middle|\; \left(\frac{\tilde{x}_1}{\sqrt{\lambda_1}} \right)^2 + \cdots + \left(\frac{\tilde{x}_n}{\sqrt{\lambda_n}} \right)^2 \leq 1 \right\}$

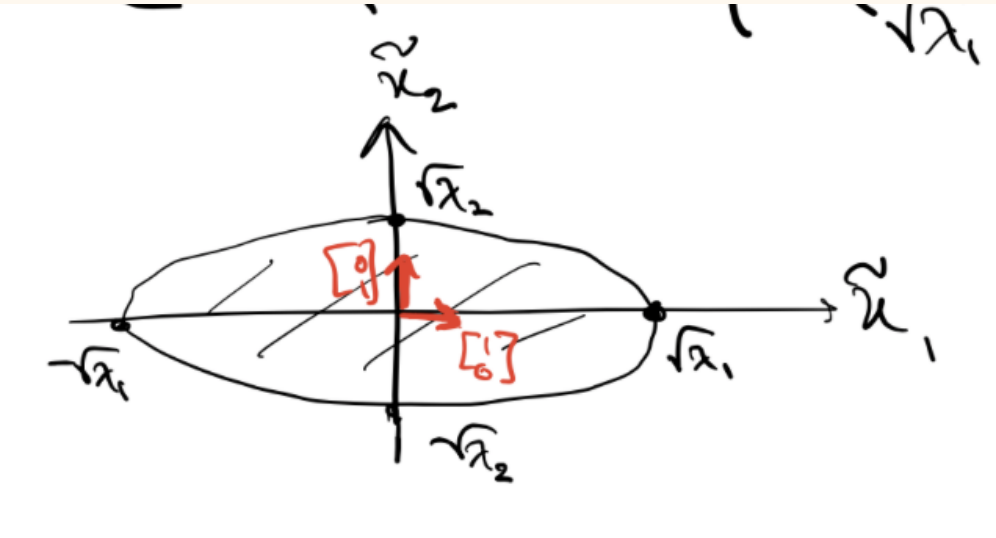

Figure 41: Tilde axis.

2. Plot in the $\bar{x}$-axis with rotation: $\tilde{x} = U^T \bar{x} \to \bar{x} = U\tilde{x}$

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} \xrightarrow{U} \bar{x} = \begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}$$

(a) For $\tilde{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$:

$$\bar{x} = \begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = u^{(1)}$$

(b) For $\tilde{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$:

$$\bar{x} = \begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = u^{(2)}$$

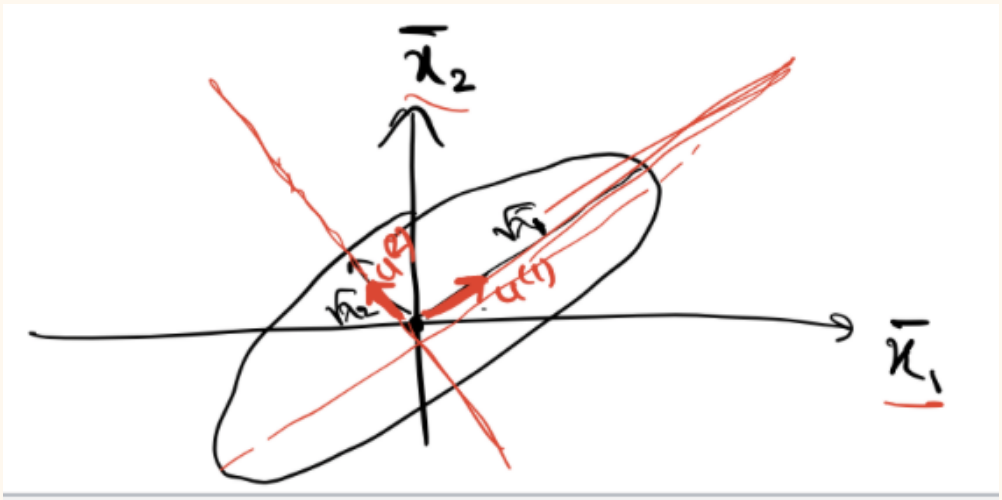

Figure 42: Bar axes.

3. Plot in the x-axis with translation: $\bar{x} = x - x^{(0)} \implies x = \bar{x} + x^{(0)}$

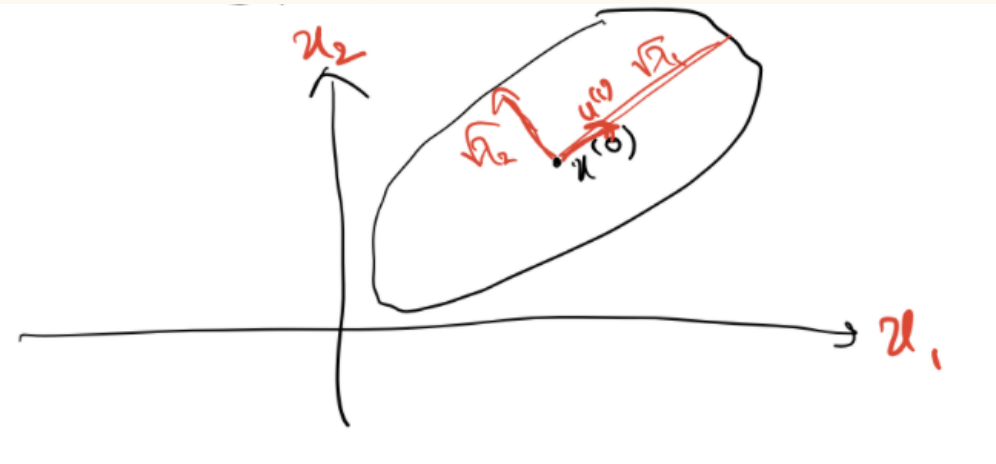

Figure 43: X axes.

8.7 Multivariate Gaussian Varibles

Example: Another reason why PD matrices are important is that they are used to define multivariate Gaussian variables.

Single-variable Gaussian:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

- σ: standard deviation
- σ^2: variance
- m: mean

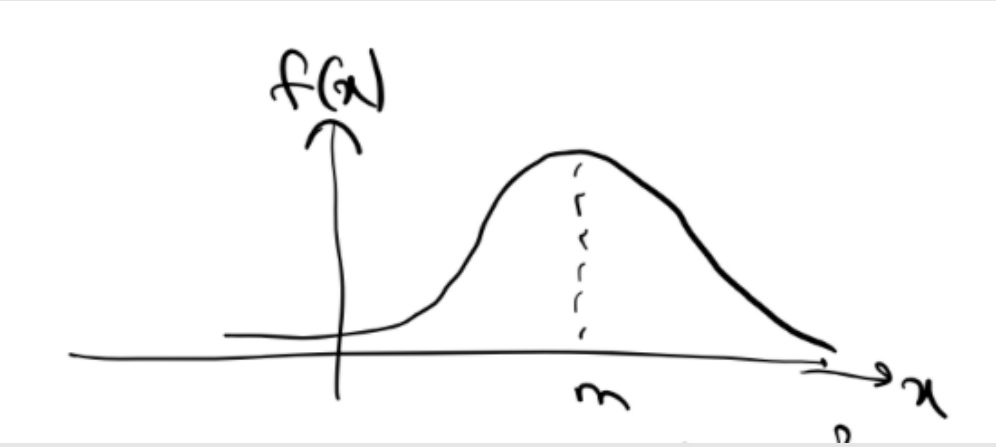

Figure 44: Single-variable Gaussian distribution

Multivariate Gaussian:

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)}$$

- $x \in \mathbb{R}^n$
- $m \in \mathbb{R}^n$: mean vector
- $\Sigma = \begin{bmatrix} \mathbb{E}[(x_1-m_1)^2] & \cdots & \mathbb{E}[(x_1-m_1)(x_n-m_n)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(x_n-m_n)(x_1-m_1)] & \cdots & \mathbb{E}[(x_n-m_n)^2] \end{bmatrix}$: covariance matrix (**PD matrix**)

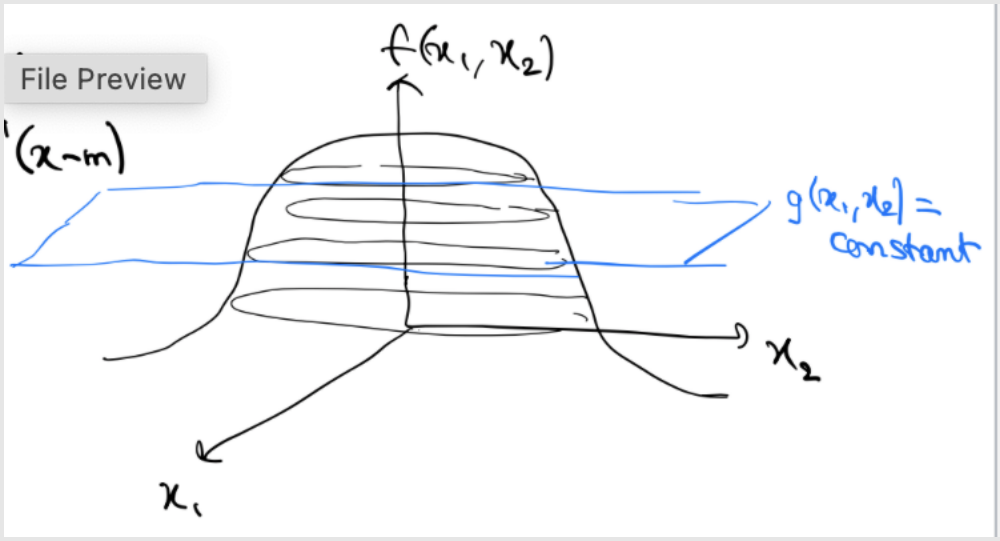

Figure 45: 3D-variable Gaussian distribution

How do we visualize the Gaussian distribution? See the intersection of f with g, which happens when f is the constant. f is a constant when the exponent term is equal to a constant, i.e.,

$$(x-m)^T\Sigma^{-1}(x-m) = \text{constant}$$

We take the cross-section of the Gaussian PDF with the horizontal hyperplane. This cross-section is defined by:

$$(x-m)^T\Sigma^{-1}(x-m) \leq \text{constant}$$

- **Note:** This is the form of the ellipse.

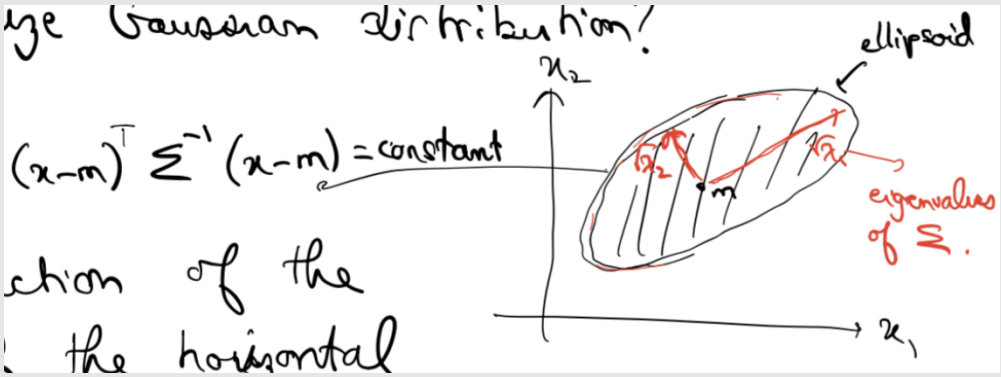

Figure 46: Cross-section of GD (i.e. Intersection of Gaussian distribution with horizontal hyperplane).

8.8 Sqaure roots of PSD

Definition: PSD matrices always has square roots (i.e. if $B = C^2$, then C is the square root of B)
If $A \in S_+^n$, then:

$$A = U\Lambda U^T$$

- Since $\lambda_i > 0$, therefore, we can always define the square root of Λ: $\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}$

Then:

$$A = U\Lambda^{1/2}\Lambda^{1/2}U^T = (U\Lambda^{1/2}U^T)(U\Lambda^{1/2}U^T) = B \cdot B$$

Therefore, $B = U\Lambda^{1/2}U^T$. B is the square root of A, denoted by $A^{1/2}$.

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

9.1 2nd Optimization Problem (Motivation for SVD)

Definition: Let $A \in S^n$ (not necessarily PSD). We are interested in solving the optimization problem:

$$\max_{s.t. ||x||_2=1} \|Ax\|_2$$

- s.t. is subject to

Intuition:

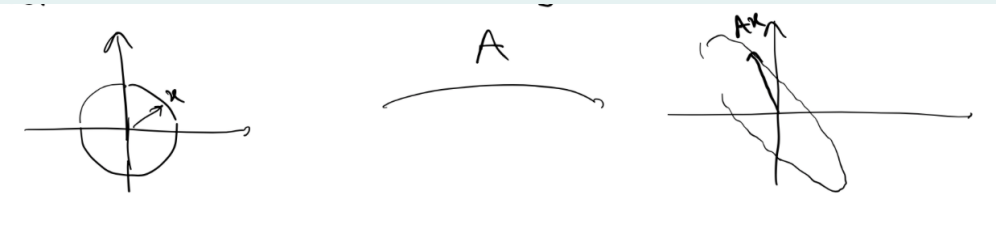

Figure 47: Optimization problem

- We are trying to find the maximum x after being transformed by A, subject to being contained within the unit circle (in 2D).

Derivation: Let $y = Ax$. Then, using the spectral decomposition $A = U\Lambda U^T$, we can express:

$$\begin{aligned} y &= Ax = U\Lambda U^T x \\ U^T y &= \Lambda U^T x \quad \text{multiply by } U^T \\ \tilde{y} &= \Lambda \tilde{x} \quad \text{where} \quad \tilde{x} = U^T x \text{ and } \tilde{y} = U^T y \end{aligned}$$

Thus, we have:

$$\begin{aligned} \|\tilde{y}\|_2 = \|U^T y\|_2 = \|y\|_2 \\ \|\tilde{x}\|_2 = \|U^T x\|_2 = \|x\|_2 \end{aligned}$$

- **Note:** This comes from a previous lecture saying how an orthogonal matrix applied to a vector doesn't change the l2-norm of it.

So the optimization problem becomes:

$$\max_{s.t. \|\tilde{x}\|_2=1} \|\Lambda \tilde{x}\|_2$$

Since $\Lambda\tilde{x}$ is diagonal, we can write:

$$\Lambda \tilde{x} = \begin{bmatrix} \lambda_1 \tilde{x}_1 \\ \lambda_2 \tilde{x}_2 \\ \vdots \\ \lambda_n \tilde{x}_n \end{bmatrix}$$

Solution: If $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, then the optimal solution is:

$$\tilde{x}^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- Putting all of our budget into the largest eigenvalue to maximize the product while staying within the constraint.

Therefore, the optimal $x^* = U\tilde{x}$ is:

$$x^* = U\tilde{x}^* = U \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = u^{(1)}$$

The maximum value is:

$$\|Ax^*\|_2 = \|y\|_2 = \|\tilde{y}\|_2 = |\lambda_1|$$

Conclusion: The optimal direction of x is the direction of the eigenvector corresponding to the largest eigenvalue. This is the solution where A is symmetric.

9.2 3rd Optimization Problem

Derivation: **Extending to rectangular matrices:** What if we need to maximize $\|Mx\|_2$ for some $M \in \mathbb{R}^{m\times n}$ (not necessarily square)? This leads us to the Singular Value Decomposition (SVD).

1. **Generalization for Non-Symmetric or Non-Square Matrices**:
   - For a matrix $M \in \mathbb{R}^{m\times n}$, the optimization problem becomes:
   $$\max_{\text{s.t. } \|x\|_2=1} \|Mx\|_2.$$
   - This can be rewritten as:
   $$\max_{\text{s.t. } \|x\|_2=1} \|Mx\|_2^2 = \max_{\text{s.t. } \|x\|_2=1} x^T(M^TM)x,$$
     - $A = M^TM$ is symmetric and positive semi-definite (PSD).
       - $A^T = (M^TM)^T = M^TM = A$
       - **Note:** A is in fact a PSD matrix because for any $x \in \mathbb{R}^n$, $x^TAx = x^TM^TMx = ||Mx||_2^2 \geq 0$
2. **Rayleigh Quotient:**
   - We aim to solve:
   $$\max_{\text{s.t. } \|x\|_2^2=x^Tx=1} x^TAx.$$
   - This is equivalent to maximizing the Rayleigh quotient:
   $$\max_{\text{s.t. } x\neq 0} \frac{x^TAx}{x^Tx}.$$
     - **Note:** This is equivalent bc if x is not unit norm, the division by x^Tx will normalize the function, therefore, any scaling factor will be canceled out. As a result, the only condition needed is that x is not zero, so we don't divide by 0.
3. **Upper and Lower Bounds**

- **Upper bound** can be achieved when $\tilde{x}$ is given by:

$$\tilde{x} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

  which corresponds to $x = u^{(1)}$, the eigenvector associated with the largest eigenvalue $\lambda_{\max}(A)$.
- **Lower bound** can be achieved when $\tilde{x}$ is given by:

$$\tilde{x} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

  which corresponds to $x = u^{(n)}$, the eigenvector associated with the smallest eigenvalue $\lambda_{\min}(A)$.

4. **Optimization Solution**
   Thus, this gives us a solution to the optimization problem:

$$\max_{\text{s.t. } \|x\|_2=1} \|Mx\|_2.$$

   The solution is the eigenvector of $M^T M$ corresponding to the largest eigenvalue, and the maximum is:

$$\sqrt{\lambda_{\max}(M^T M)}.$$

9.2.1 Theorem

Theorem: Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$, we have

$$\lambda_{\min}(A) \leq \frac{x^T A x}{x^T x} \leq \lambda_{\max}(A), \quad \forall x \in \mathbb{R}^n.$$

- $\lambda_{\max}(A) = \max_{x \neq 0} \frac{x^T A x}{x^T x}$
- $\lambda_{\min}(A) = \min_{x \neq 0} \frac{x^T A x}{x^T x}$

Derivation:

1. Given a symmetric matrix A, we can write its spectral decomposition as:

$$A = U \Lambda U^T,$$

   where U is orthogonal and Λ is diagonal.
2. For any vector x, we have:

$$\begin{aligned} x^T A x &= x^T U \Lambda U^T x \\ &= \tilde{x}^T \Lambda \tilde{x}, \end{aligned}$$

   where $\tilde{x} = U^T x$.
3. Expanding this expression, we get:

$$x^T A x = \sum_{i=1}^{n} \lambda_i \tilde{x}_i^2.$$

4. Since $x^T x = \|x\|_2^2 = \|\tilde{x}\|_2^2$, we have:

$$\|x\|_2^2 = \sum_{i=1}^{n} \tilde{x}_i^2$$

5. The Rayleigh quotient becomes:

$$\frac{x^T A x}{x^T x} = \frac{\sum_{i=1}^{n} \lambda_i \tilde{x}_i^2}{\sum_{i=1}^{n} \tilde{x}_i^2}.$$

- **Upper Bound**: Using the properties of eigenvalues, we get:

$$\frac{x^T A x}{x^T x} \leq \frac{\sum_{i=1}^{n} \lambda_{\max} \tilde{x}_i^2}{\sum_{i=1}^{n} \tilde{x}_i^2} = \lambda_{\max}(A).$$

  - **Note:** $\lambda_{\max}$ doesn't rely on index, so we can bring it out.
- **Lower Bound**: Similarly, we have:

$$\frac{x^T A x}{x^T x} \geq \frac{\sum_{i=1}^{n} \lambda_{\min} \tilde{x}_i^2}{\sum_{i=1}^{n} \tilde{x}_i^2} = \lambda_{\min}(A).$$

  - **Note:** $\lambda_{\min}$ doesn't rely on index, so we can bring it out.

6. Therefore,

$$\lambda_{\min}(A) \leq \frac{x^T A x}{x^T x} \leq \lambda_{\max}(A).$$

7. **Note:** If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, the maximum value is achieved when $\tilde{x} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$, giving:

$$x^T A x = \sum_{i=1}^{n} \lambda_i \tilde{x}_i^2 = \lambda_{\max} \|x\|_2^2$$

9.3 Singular value decomposition

Intuition: For any matrix $A \in \mathbb{R}^{m \times n}$, we can express it using Singular Value Decomposition (SVD) as:

$$A = U \Sigma V^T,$$

where:
- U is an orthogonal matrix of size $m \times m$.
- V is an orthogonal matrix of size $n \times n$.
- Σ is a "nearly diagonal" matrix of size $m \times n$ containing the singular values of A along its main diagonal.

The matrix Σ has the form:

$$\Sigma = \left[\begin{array}{cccc|ccc} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right]$$

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values of A, and $r = \text{rank}(A)$. The remaining entries in Σ are zero.

Definition: Any matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed into

$$A = U \Sigma V^T,$$

- $U \in \mathbb{R}^{m\times m}$: Orthogonal matrix.
- $V^T \in \mathbb{R}^{n\times n}$: Orthogonal matrix.
  - U and V^T live in different spaces.
- $\Sigma \in \mathbb{R}^{m\times n}$: Its first r diagonal entries $\sigma_1, \ldots, \sigma_r$ as positive, where $r = \text{rank}(A)$, and the rest is zero.

$$[A]_{m\times n} = [U]_{m\times m}[\Sigma]_{m\times n}[V^T]_{n\times n}$$

where

$$\Sigma = \left[\begin{array}{cccc|ccc} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right]$$

9.3.1 Intuition on SVD:

Intuition: The decomposition $y = Ax$ can be written as

$$\begin{aligned} y &= Ax \\ &= U\Sigma V^T x, \end{aligned}$$

and

$$\begin{aligned} U^T y &= \Sigma V^T x \\ \tilde{y} &= \Sigma \tilde{x}. \end{aligned}$$

The steps in the transformation are as follows:
1. Rotate/Flip using V^T,
2. Scale using Σ,
3. Rotate/Flip using U.

- **Note:** Since a linear map can be represented as $y = Ax$, therefore, SVD states that all linear maps are essentially scaling each component if
  - We allow some pre-processing of x by a rotation/flip into $\tilde{x}$.
  - And then some post-processing of $\tilde{y}$ by a rotation/flip to get y.

Warning: Whenever you get the matrix, do the SVD.

9.4 Proof of SVD

9.4.1 AA^T and A^TA

Derivation: Since we want $A = U\Sigma V^T$, let's consider:
1. For $AA^T \in m \times m$:

$$\begin{aligned} AA^T &= U\Sigma V^T (U\Sigma V^T)^T \\ &= U\Sigma V^T V \Sigma^T U^T \\ &= U\Sigma\Sigma^T U^T, \end{aligned}$$

- AA^T: Positive semi-definite (PSD) matrix (from previous lecture)
- Hence, $\sigma_i = \sqrt{\lambda_i(AA^T)}$ because we are dealing with a spectral decomposition and $\Sigma\Sigma^T$ acts as the Λ in

this case. Same goes for A^TA

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_r^2 & 0 & 0 \\ 0 & 0 & \cdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in m \times m$$

2. For $A^TA \in n \times n$:

$$\begin{aligned} A^TA &= (U\Sigma V^T)^T U\Sigma V^T \\ &= V\Sigma^T U^T U\Sigma V^T \\ &= V\Sigma^T\Sigma V^T, \end{aligned}$$

- Hence, $\sigma_i = \sqrt{\lambda_i(A^TA)}$

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_r^2 & 0 & 0 \\ 0 & 0 & \cdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in n \times n$$

- **Same eigenvalues:** AA^T and A^TA share the same set of non-zero eigenvalues, which will be proven as a fact, however, the total number of eigenvalues may differ.

9.4.2 Fact

Definition: For any matrices A and B, AB and BA share the same set of non-zero eigenvalues.

Derivation:

$$\begin{aligned} ABv &= \lambda v, \\ BABv &= \lambda Bv, \end{aligned}$$

which implies that λ is an eigenvalue of BA with eigenvector Bv.

9.4.3 How are u and v related?

Derivation: We don't need to do two spectral decompositions because $u^{(i)}$'s and $v^{(i)}$'s are related.

1. If $v^{(i)}$ is an eigenvector of A^TA, then

$$\begin{aligned} A^TAv^{(i)} &= \lambda_i v^{(i)}, \\ AA^T(Av^{(i)}) &= \lambda_i(Av^{(i)}), \end{aligned}$$

which implies that $Av^{(i)}$ is an eigenvector of AA^T.

2. Thus, we can find $u^{(i)}$ by normalizing $Av^{(i)}$ by the corresponding singular value σ_i since u lives in AA^T space.

$$
\begin{aligned}
u^{(i)} &= \frac{Av^{(i)}}{\|Av^{(i)}\|_2} \\
&= \frac{Av^{(i)}}{\sqrt{(Av^{(i)})^T(Av^{(i)})}} \\
&= \frac{Av^{(i)}}{\sqrt{v^{(i)T}A^TAv^{(i)}}} \\
&= \frac{Av^{(i)}}{\sqrt{v^{(i)T}\lambda_i v^{(i)}}} \quad \text{since } v^{(i)} \text{ is an eigenvector of } A^TA \\
&= \frac{Av^{(i)}}{\sqrt{\lambda_i\, v^{(i)T}v^{(i)}}} \\
&= \frac{Av^{(i)}}{\sqrt{\lambda_i}} \quad (\text{since } v^{(i)} \text{ is orthonormal vector}) \\
&= \frac{Av^{(i)}}{\sigma_i}.
\end{aligned}
$$

9.4.4 Summarization of Steps for SVD Construction

Derivation:

1. Do a spectral decomposition of A^TA to get $\{v^{(1)}, \ldots, v^{(n)}\}$ and $\lambda_1, \ldots, \lambda_r$, where $r = \text{rank}(A^TA)$.
2. Set $\sigma_i = \sqrt{\lambda_i}$, for $i = 1, \ldots, r$.
3. Set $u^{(i)} = \dfrac{Av^{(i)}}{\sigma_i}$, for $i = 1, \ldots, r$.

- **Note:** We want to extend this to complete the orthonormal basis of $\mathbb{R}^m$.

9.4.5 Orthogonality of $u^{(i)}$'s

Derivation: $u^{(i)}$'s are orthonormal because

$$
\begin{aligned}
u^{(i)T}u^{(j)} &= \left(\frac{Av^{(i)}}{\sigma_i}\right)^T \left(\frac{Av^{(j)}}{\sigma_j}\right) \\
&= \frac{(v^{(i)})^TA^TAv^{(j)}}{\sigma_i\sigma_j} \\
&= \frac{(v^{(i)})^T(\lambda_j v^{(j)})}{\sigma_i\sigma_j} \quad \text{since } v^{(i)} \text{ is an eigenvector of } A^TA \\
&= \frac{\lambda_j\,(v^{(i)})^Tv^{(j)}}{\sigma_i\sigma_j} \\
&= \frac{\sigma_j^2\,(v^{(i)})^Tv^{(j)}}{\sigma_i\sigma_j} \\
&= \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}
\end{aligned}
$$

9.4.6 Why does this correspond to SVD?

Derivation:

$$
u^{(i)T}Av^{(j)} = u^{(i)T}u^{(j)} \cdot \sigma_j = \begin{cases} \sigma_j & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}
$$

- 1st to 2nd line comes from $u^{(i)} = \dfrac{Av^{(i)}}{\sigma_i}$.

We can represent this in matrix form as follows:

$$\begin{bmatrix} u^{(1)T} \\ \vdots \\ u^{(r)T} \end{bmatrix} A \begin{bmatrix} v^{(1)} & \dots & v^{(r)} \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}.$$

9.4.7 Completing SVD:

Derivation: If $r = \text{rank}(A) < m$, i.e., $\{u^{(1)}, \dots, u^{(r)}\}$ does not span $\mathbb{R}^m$, we can find $u^{(r+1)}, \dots, u^{(m)}$ to complete the orthonormal basis of $\mathbb{R}^m$.

$$\begin{bmatrix} u^{(1)T} \\ \vdots \\ u^{(r)T} \\ u^{(r+1)T} \\ \vdots \\ u^{(m)T} \end{bmatrix} A \begin{bmatrix} v^{(1)} & \dots & v^{(r)} & v^{(r+1)} & \dots & v^{(n)} \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}.$$

-

This matrix is denoted by Σ, where

$$U^T AV = \Sigma \Rightarrow A = U\Sigma V^T.$$

9.4.8 Fact:

Definition:

$$\mathcal{N}(A^T A) = \mathcal{N}(A)$$

Derivation:

1. If $x \in \mathcal{N}(A)$, then:
$$Ax = 0 \Rightarrow A^T Ax = 0 \Rightarrow x \in \mathcal{N}(A^T A).$$
2. Conversely, if $x \in \mathcal{N}(A^T A)$, then:
$$\begin{aligned} A^T Ax &= 0, \\ \Rightarrow x^T A^T Ax &= 0, \\ \Rightarrow \|Ax\|_2^2 &= 0, \\ \Rightarrow Ax &= 0, \\ \Rightarrow x &\in \mathcal{N}(A). \end{aligned}$$

Therefore, $\mathcal{N}(A^T A) = \mathcal{N}(A)$.

9.5 SVD in Two Forms

Definition: We can write SVD in two forms:

1. Full SVD:
$$A = U\Sigma V^T$$
where
   - A is $m \times n$,

- U is $m \times m$ (orthogonal matrix),
- Σ is $m \times n$ (diagonal matrix with singular values),
- V^T is $n \times n$ (orthogonal matrix).

2. Compact or Reduced SVD:

$$[A] = \begin{bmatrix} u^{(1)} & \cdots & u^{(r)} & | & u^{(r+1)} & \cdots & u^{(m)} \end{bmatrix} \left[\begin{array}{ccc|ccc} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right] \left[\begin{array}{c} u^{(1)T} \\ \vdots \\ u^{(r)T} \\ \hline u^{(r+1)T} \\ \vdots \\ u^{(n)T} \end{array}\right]$$

where
- $\tilde{U}$ is $m \times r$, the first r columns of U (not orthogonal matrix)
- $\tilde{\Sigma}$ is $r \times r$, containing the non-zero singular values $\sigma_1, \ldots, \sigma_r$,
- $\tilde{V}^T$ is $r \times n$, the first r columns of V^T (not orthogonal matrix)

Thus,

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^T.$$

3. **Summation:**

$$A = \sum_{i=1}^{r} \sigma_i u^{(i)} v^{(i)T}.$$

- $u^{(i)}$: $m \times 1$
- $v^{(i)T}$: $1 \times n$
- $\sigma_i u^{(i)} v^{(i)T}$: $m \times n$

9.6 Maximization Problem

Derivation: The goal is to maximize $\|Ax\|_2$ subject to $\|x\|_2 = 1$.

1. Formulation

$$\begin{aligned} y &= Ax \\ &= U\Sigma V^T x, \end{aligned}$$

where $U^T y = \Sigma V^T x$, leading to

$$\tilde{y} = \Sigma \tilde{x}$$

2. Optimization Reformulation:
   (a) Rewrite $\|y\|_2 = \|\tilde{y}\|_2$ and $\|\tilde{x}\|_2 = \|x\|_2$ since orthogonal matrix preserves the norms.
   (b) The optimization problem becomes:

$$\max_{\text{s.t. } \|\tilde{x}\|_2 = 1.} \|\Sigma \tilde{x}\|_2$$

3. Solution:

$$\Sigma\tilde{x} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_r \\ \tilde{x}_{r+1} \\ \vdots \\ \tilde{x}_n \end{bmatrix} = \begin{bmatrix} \sigma_1\tilde{x}_1 \\ \sigma_2\tilde{x}_2 \\ \vdots \\ \sigma_r\tilde{x}_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

4. Notation:
   - $\sigma_1, \ldots, \sigma_r$: Singular values.

- $u^{(1)}, \ldots, u^{(m)}$: Left singular vectors.
- $v^{(1)}, \ldots, v^{(n)}$: Right singular vectors.

5. Result: If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, then the solution is

$$\tilde{x}^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Thus, $x^* = V\tilde{x}^* = v^{(1)}$, and we have:

$$\|y^*\|_2 = \|Ax^*\|_2 = \|\tilde{y}^*\|_2 = \|\Sigma\tilde{x}^*\|_2 = \sigma_1(A).$$

9.7 SVD Relation to Range space and Null space of A

Definition:

$$\mathcal{R}(A) = \text{span}\{u^{(1)}, \ldots, u^{(r)}\}$$
$$\mathcal{N}(A) = \text{span}\{v^{(r+1)}, \ldots, v^{(n)}\}$$

Derivation: Given the decomposition:

$$Ax = U\Sigma V^T x$$

we have:

$$Ax = \begin{bmatrix} u^{(1)} & \cdots & u^{(r)} & | & u^{(r+1)} & \cdots & u^{(m)} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \left[\begin{array}{c} v^{(1)T} \\ \vdots \\ v^{(r)T} \\ \hline v^{(r+1)T} \\ \vdots \\ v^{(n)T} \end{array}\right] x$$

Range of A The range of A, $\mathcal{R}(A)$, is defined as the span of the columns of A:

$$\mathcal{R}(A) = \text{span of columns of } A = \{Ax \mid x \in \mathbb{R}^n\}.$$

We can write this as:

$$Ax = \begin{bmatrix} u^{(1)} & \cdots & u^{(r)} & u^{(r+1)} & \cdots & u^{(m)} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- **Note:** 0 for $r+1$ to n (i.e. $n-r$ entries) because they get multiplied by 0 in Σ.
- **Note:** x_i for 1 to r (i.e. r entries) because they get multiplied by σ_i in Σ.

Therefore,

$$\mathcal{R}(A) = \text{span}\{u^{(1)}, \ldots, u^{(r)}\}.$$

Null Space of A The null space of A, $\mathcal{N}(A)$, is defined as:

$$\mathcal{N}(A) = \{x \mid Ax = 0\}.$$

If $x = \sum_{i=r+1}^{n} \beta_i v^{(i)}$, then

$$Ax = U\Sigma \begin{bmatrix} v^{(1)T} \\ \vdots \\ v^{(r)T} \\ v^{(r+1)T} \\ \vdots \\ v^{(n)T} \end{bmatrix} x = U\Sigma \begin{bmatrix} 0 \\ \vdots 0 \\ x_{r+1} \\ \vdots \\ x_n \end{bmatrix}.$$

- **Note:** First r entries are 0 because of inner product with $v^{(i)}$ for $i = r+1$ to n (i.e. how x is defined)
- **Note:** Last $n - r$ entries are x_{r+1} to x_n because of inner product with $v^{(i)}$ for $i = r+1$ to n (i.e. how x is defined)

Thus,

$$\mathcal{N}(A) = \text{span}\{v^{(r+1)}, \ldots, v^{(n)}\}.$$

Derivation:

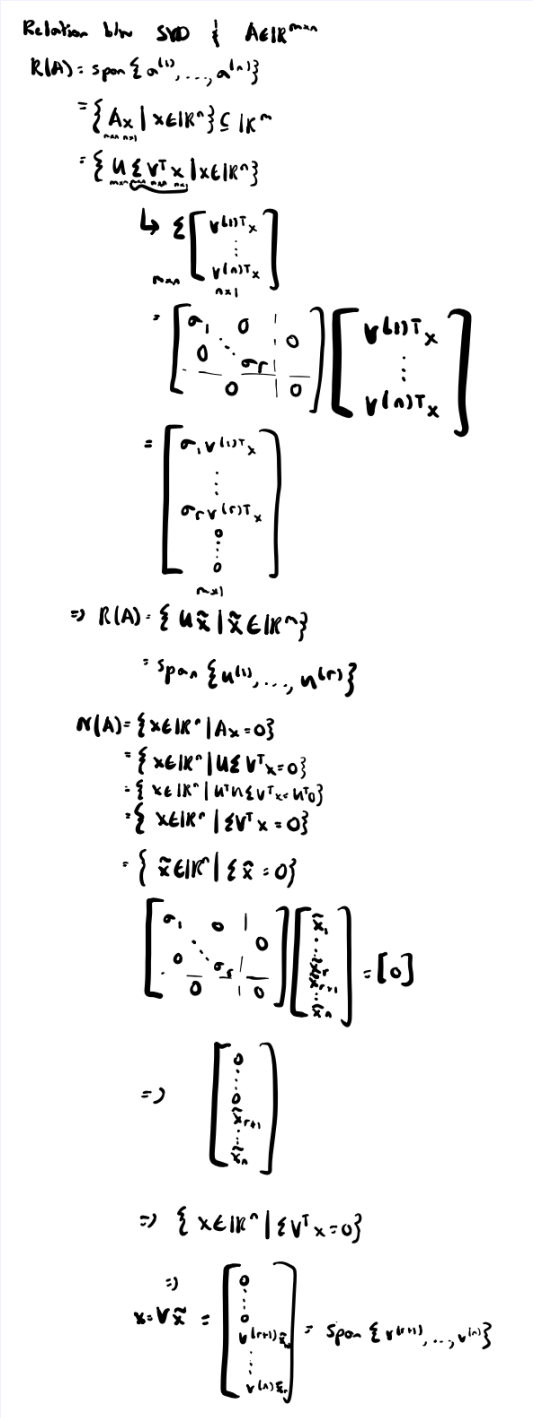

Figure 48: Another Derivation

9.8 What about A transpose?

Definition:

$$A = U\Sigma V^T$$

$$A^T = (U\Sigma V^T)^T = V\Sigma^T U^T$$

- V: $n \times n$
- Σ^T: $n \times m$

- U^T: $m \times m$

$$\begin{bmatrix} v^{(1)} & \dots & v^{(r)} & | & v^{(r+1)} & \dots & v^{(n)} \end{bmatrix} \Sigma^T \begin{bmatrix} u^{(1)T} \\ \vdots \\ u^{(r)T} \\ u^{(r+1)T} \\ \vdots \\ u^{(m)T} \end{bmatrix}$$

9.8.1 Range and Null Space

Definition:

$$\mathcal{R}(A^T) = \text{span}\{v^{(1)}, \dots, v^{(r)}\}$$
$$\mathcal{N}(A^T) = \text{span}\{u^{(r+1)}, \dots, u^{(m)}\}$$

FToLA (Use as a check):

$$\mathcal{N}(A) \oplus \mathcal{R}(A^T) = \mathbb{R}^n, \quad \mathcal{R}(A) \oplus \mathcal{N}(A^T) = \mathbb{R}^m$$

9.9 Matrix Inverse

Definition: If A is a square matrix and invertible, then the SVD of A^{-1}

$$A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T$$

- **Note:** $U^{-1} = U^T$ and $V^{-1} = V^T$ because they are orthogonal matrices.

$$A^{-1} = V \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_r} \end{bmatrix} U^T$$

- **Note on A^{-1}:** Rotate/Flip using U^T, Scale using Σ^{-1}, Rotate/Flip using V.
- **Note on A:** Rotate/Flip using V^T, Scale using Σ, Rotate/Flip using U.
- **Conclusion:** The inverse of A does the opposite transformation of A in the SVD.

9.10 Pseudo-Inverse

Definition: For any matrix $A = U\Sigma V^T$, the pseudo-inverse is given by:

$$A^\dagger = V \left[\begin{array}{ccc|ccc} 1/\sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_r & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right] U^T$$

- **Note:** It turns out this can be generalized to any matrix (not necessarily invertible) to the Moore-Penrose pseudo-inverse.

9.10.1 Least Squares Solution

Definition:

$$A = [\text{tall matrix with linearly independent columns}] \Rightarrow A^\dagger = (A^TA)^{-1}A^T, \quad A^\dagger A = I$$
$$A = [\text{fat matrix with linearly independent rows}] \Rightarrow A^\dagger = A^T(AA^T)^{-1}, \quad AA^\dagger = I$$

9.11 Matrix Norms

9.11.1 Frobenius Norm

Definition:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\text{tr}(A^TA)} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$$

Intuition:

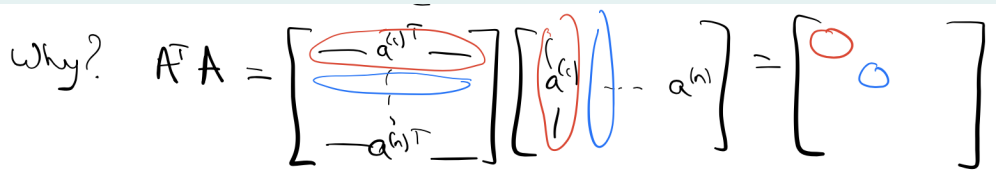

Figure 49: Frobenius Norm

- $\text{tr}(A^TA)$ makes sense because we are taking first row of A^T and multiplying it with first column of A and summing it up. Repeating this for all rows and columns gives us the sum of squares of all elements.

Derivation:

$$A = U\Sigma V^T$$
$$A^TA = (V\Sigma^TU^T)(U\Sigma V^T) = V\Sigma^T\Sigma V^T$$
$$= V \left[\begin{array}{ccc|ccc} \sigma_1^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^2 & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right] V^T$$

$$\begin{aligned}
\operatorname{tr}(A^TA) &= \operatorname{tr}\left(V \left[\begin{array}{ccc|ccc} \sigma_1^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^2 & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right] V^T\right) \\
&= \operatorname{tr}\left(VV^T \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^2 \end{bmatrix}\right) \quad \text{since } \operatorname{tr}(AB) = \operatorname{tr}(BA) \\
&= \operatorname{tr}\left(\begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^2 \end{bmatrix}\right) \\
&= \sum_{i=1}^{r} \sigma_i^2
\end{aligned}$$

$$\|A\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$$

9.11.2 Operator Norm

Definition:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_{\max}(A)$$

9.12 Conditioning Number of a Matrix

Definition: The condition number $\kappa(A)$ measures the numerical stability of solving $Ax = b$.

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

9.12.1 Stability

Derivation: The solution is $x = A^{-1}b$ (if A is invertible). If we change by some small perturbation:

$$b \to b + \Delta b$$
$$x \to x + \Delta x$$

We are interested in comparing:

$$\frac{\|\Delta x\|_2}{\|x\|_2} \quad \text{vs} \quad \frac{\|\Delta b\|_2}{\|b\|_2}$$

- **Observation 1:** If $Ax = b$, and $A(x + \Delta x) = b + \Delta b$, then

$$A \cdot \Delta x = \Delta b \text{ since } Ax = b \implies \Delta x = A^{-1}\Delta b$$

- **Observation 2:** $Ax = b \implies \|b\|_2 = \|Ax\|_2 \leq \sigma_{\max}(A)\|x\|_2$
  - Multiply by the norm of x because x is not assumed to have unit norm (i.e. 1), so there is no limitation on x. Same idea for Δb

  - **Note:** This was proven in the previous section.
- **Observation 3:** $\Delta x = A^{-1}\Delta b \implies \|\Delta x\|_2 = \|A^{-1}\Delta b\|_2 \leq \sigma_{\max}(A^{-1})\|\Delta b\|_2$

Now we can compare the quantity we are interested in:

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \frac{\sigma_{\max}(A^{-1})\|\Delta b\|_2}{\frac{\|b\|_2}{\sigma_{\max}(A)}} = \sigma_{\max}(A)\cdot\sigma_{\max}(A^{-1})\cdot\frac{\|\Delta b\|_2}{\|b\|_2}$$

- **QUESTION:** How is this upper bounded because in the previous derivations one is lower bounded and the other is upper bounded.
  - This is correct, b/c the lower bound is in the denominator, making $\frac{1}{||x||_2}$ being upper bounded by $1/\frac{||b||_2}{\sigma_{\max}(A)}$ NOT lower bounded.

Let $A = U\Sigma V^T$, then

$$A^{-1} = V\Sigma^{-1}U^T = V\begin{bmatrix}\frac{1}{\sigma_1} & 0 & \dots & 0\\ 0 & \frac{1}{\sigma_2} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \frac{1}{\sigma_r}\end{bmatrix}U^T$$

$$\sigma_{\max}(A^{-1}) = \frac{1}{\sigma_{\min}(A)}$$

Thus,

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}\cdot\frac{\|\Delta b\|_2}{\|b\|_2}$$

Define the condition number as:

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

If b changes by 10%, then x changes by at most $\kappa(A)\cdot 10\%$.

- If $\kappa(A)$ is small, solving $Ax = b$ is numerically stable, implying A is a **well-conditioned** matrix.
- If $\kappa(A)$ is large, solving $Ax = b$ is numerically unstable, implying A is an **ill-conditioned** matrix.

Note: The smallest value of κ is 1 (i.e. $\sigma_{\min} = \sigma_{\max}$), which is the best case scenario.

9.13 Latent Semantic Indexing

Example: In homework 1, we compared documents. Assume we have a dictionary of size N. Represent each document by a vector

$$d = \begin{bmatrix}d_1\\ \vdots\\ d_N\end{bmatrix}$$

where d_i = # of occurrences of word i in the document.

In the homework, we compared documents i and j by comparing their corresponding vectors $d^{(i)}$ and $d^{(j)}$. Instead, it would be nice to extract the meaning (i.e., the "semantics") of the words in the document.

Idea: Think about a document as a linear combination of "concepts".

1. Construct a data matrix

$$A = \begin{bmatrix}d^{(1)} & \dots & d^{(m)}\end{bmatrix}_{N\times m}$$

where N represents the number of words in the dictionary and m represents the number of documents, respectively.

2. Perform Singular Value Decomposition (SVD) on the data matrix A:

$$A = U\Sigma V^T = \tilde{U}\tilde{\Sigma}\tilde{V}^T = \begin{bmatrix} u^{(1)} & \dots & u^{(r)} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_r \end{bmatrix} \begin{bmatrix} v^{(1)T} \\ \\ \vdots \\ v^{(r)T} \end{bmatrix}$$

3. To analyze the i-th document:

$$d^{(i)} = Ae^{(i)}$$

where $e^{(i)}$ is a standard basis vector with 1 in the i-th position. Thus,

$$d^{(i)} = \tilde{U}\tilde{\Sigma}\tilde{V}^T e^{(i)}$$

$$= \begin{bmatrix} u^{(1)} & \dots & u^{(r)} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_r \end{bmatrix} \begin{bmatrix} v^{(1)T} \\ \\ \vdots \\ v^{(r)T} \end{bmatrix} e^{(i)}$$

$$= \begin{bmatrix} u^{(1)} & \dots & u^{(r)} \end{bmatrix} \begin{bmatrix} \sigma_1 v^{(1)T} \\ \\ \vdots \\ \sigma_r v^{(r)T} \end{bmatrix} e^{(i)}$$

$$= \begin{bmatrix} u^{(1)} & \dots & u^{(r)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \\ \vdots \\ \alpha_r \end{bmatrix} = \sum_{j=1}^{r} \alpha_j u^{(j)}$$

- **Key:** Each u vector is a concept, so we are writing the document as a linear combination of concepts weighted by α (i.e. strength of the concept in the document).
  - **Note:** This is what we are using to compare documents.

  Reducing Computational Complexity: $r \ll N$, so we are reducing the dimensionality of the problem, making it less computationally expensive to compare the similarity of documents.
  - **Note:** HW1: Compared between length N and here it's only r.

4. **New Document:** When a new document g comes along, we have two options:
   (a) **Bad:** We can add g to A and rerun the SVD, but this is computationally expensive.
   (b) **Good:** We project g onto the span of $\mathcal{R}(A) = \{u^{(1)}, \dots, u^{(r)}\}$ (i.e. projecting onto the range of A):

$$g^* = \sum_{j=1}^{r} \beta_j u^{(j)}, \quad \beta_j = \langle g, u^{(j)} \rangle$$

5. **Comparison:** We then use the vectors $\{\alpha_j\}$ and $\{\beta_j\}$ to compare documents, for example, by finding the angle between them.

9.14 Principle component analysis

Example: **Example of PCA: Eigenfaces for face recognition (HW3)**

1. Given face 1 to face N, we want to do facial recognition.
2. Vectorize each image: Each face image of size 32×32 pixels is vectorized into a column vector $x^{(i)}$ of dimension $m = 32 \times 32$.
3. Center each vector:

$$\tilde{x}^{(i)} = x^{(i)} - \bar{x}$$

where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

4. Construct the data matrix:

$$\tilde{X} = \begin{bmatrix} \tilde{x}^{(1)} & \dots & \tilde{x}^{(N)} \end{bmatrix}$$

5. Perform spectral decomposition on the covariance matrix:

$$C = \tilde{X}\tilde{X}^T = U\Lambda U^T$$

$$C = \begin{bmatrix} u^{(1)} & \dots & u^{(m)} \end{bmatrix} \Lambda \begin{bmatrix} u^{(1)T} \\ \vdots \\ u_{(m)T} \end{bmatrix}$$

   - **Note:** Each face could be represented as a linear combination of the u's, where the coefficients represent the strength of each u in the face.
6. Alternatively, we could perform SVD on $\tilde{X}$:

$$\tilde{X} = U\Sigma V^T$$

   - **Note:** This results in the same matrix U as obtained from the spectral decomposition of $\tilde{X}\tilde{X}^T$.

Example: **PCA:**

1. Given data points $x^{(i)} \in \mathbb{R}^n$, $i = 1, \dots, m$
2. **Goal of PCA:** Find a direction $z \in \mathbb{R}^n$ that explains most of the variance in $x^{(i)}$'s.
   - **Motivation:** The largest variance gives us the most information about the problem.

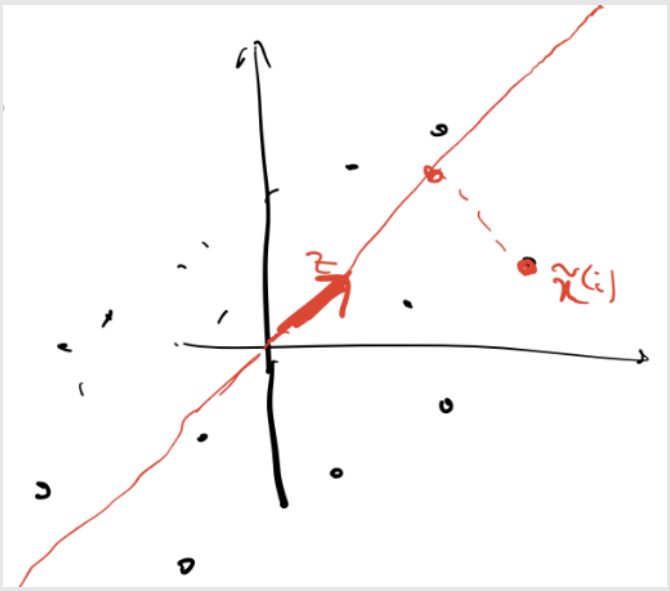

Figure 50: PCA

3. **How to formulate this problem?**
   (a) Center the data:

$$\tilde{x}^{(i)} = x^{(i)} - \bar{x}, \quad \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

   (b) Fix a direction z. Project $\tilde{x}^{(i)}$ onto $\text{span}\{z\}$:

$$\text{Proj}_{\text{span}\{z\}}\tilde{x}^{(i)} = \langle \tilde{x}^{(i)}, z \rangle z = \alpha_i z \quad (\text{assuming } \|z\|_2 = 1)$$

   (c) Look for z that maximizes the variance of the α_i's:

$$\text{Variance} = \frac{1}{m}\sum_{i=1}^{m} \alpha_i^2 = \frac{1}{m}\sum_{i=1}^{m} \left(z^T\tilde{x}^{(i)}\right)\left(\tilde{x}^{(i)T}z\right) \quad \text{Writing } \alpha \text{ in two ways and by defintiion of variance}$$

$$= \frac{1}{m} z^T \left(\sum_{i=1}^{m} \tilde{x}^{(i)}\tilde{x}^{(i)T}\right) z$$

$$= \frac{1}{m} z^T \tilde{X} \tilde{X}^T z$$

- $\tilde{X} = \begin{bmatrix} \tilde{x}^{(1)} & \dots & \tilde{x}^{(m)} \end{bmatrix}$
- $\tilde{X}\tilde{X}^T = \begin{bmatrix} \tilde{x}^{(1)} & \dots & \tilde{x}^{(m)} \end{bmatrix} \begin{bmatrix} \tilde{x}^{(1)T} \\ \vdots \\ \tilde{x}^{(m)T} \end{bmatrix} = \sum_{i=1}^{m} \tilde{x}^{(i)} \tilde{x}^{(i)T}$

4. We want to solve (i.e. maximize the variance to explain most of the variance):

$$\max_z z^T \tilde{X} \tilde{X}^T z \quad \text{subject to} \quad \|z\|_2 = 1$$

This can be expressed as maximizing the Rayleigh quotient from previous lectures.

$$\max_{z \neq 0} \frac{z^T \tilde{X} \tilde{X}^T z}{\|z\|_2^2}$$

5. The solution is given by the eigenvector corresponding to the largest eigenvalue of $\tilde{X}\tilde{X}^T$ from previous lecture i.e. $\lambda_{\max}(\tilde{X}\tilde{X}^T)$, which also corresponds to the left singular vector of $\tilde{X}$ since $\tilde{X}\tilde{X}^T = U\Sigma^2 U^T$.
6. **Alternatively**, we could have used SVD on $\tilde{X}$:

$$\tilde{X} = U\Sigma V^T$$

$$z^T \tilde{X}\tilde{X}^T z = z^T \left(U\Sigma V^T\right)\left(V\Sigma^T U^T\right) z = z^T U \Sigma^2 U^T z = \omega^T \Sigma^2 \omega = \sum_{i=1}^{r} \sigma_i^2 \omega_i^2$$

7. So, the problem becomes (i.e. maximizing the variance):

$$\max \sum_{i=1}^{r} \sigma_i^2 \omega_i^2 \quad \text{subject to} \quad \sum_{i=1}^{r} \omega_i^2 = ||\omega||_2 = 1$$

- **Note:** This holds because ω is transformed by an orthogonal matrix, so the norm is preserved.

8. The solution is $\omega^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, corresponding to the first singular vector, assuming $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$, so the maximizing $z^* = U\omega^* = u^{(1)}$. In addition, the maximum is $\sigma_1^2 = \lambda_{\max}(\tilde{X}\tilde{X}^T)$.
9. What if we're interested in the two directions with largest variance? To find the second principal component, remove the effect of the direction $u^{(1)}$ with the largest variance, and then compute
   (a) **Computation**
   We compute:

$$\begin{aligned} \tilde{\tilde{x}}^{(i)} &= \tilde{x}^{(i)} - \langle \tilde{x}^{(i)}, u^{(1)} \rangle u^{(1)} \\ &= \tilde{x}^{(i)} - u^{(1)} u^{(1)\top} \tilde{x}^{(i)} \\ &= \left(I - u^{(1)} u^{(1)\top}\right) \tilde{x}^{(i)}. \end{aligned}$$

   Next, define:

$$\tilde{\tilde{X}} = \begin{bmatrix} \tilde{\tilde{x}}^{(1)} & \dots & \tilde{\tilde{x}}^{(m)} \end{bmatrix}.$$

   Then:

$$\tilde{\tilde{X}} = \left(I - u^{(1)} u^{(1)\top}\right) \tilde{X}.$$

   (b) **Using Singular Value Decomposition (SVD)**
   Let $\tilde{X} = U\Sigma V^\top$, where:

$$\tilde{X} = \sum_{i=1}^{r} \sigma_i u^{(i)} v^{(i)\top}.$$

Compute $\tilde{\tilde{X}}$:

$$\begin{aligned}\tilde{\tilde{X}} &= \left(I - u^{(1)}u^{(1)\top}\right)\tilde{X} \\ &= \tilde{X} - u^{(1)}u^{(1)\top}\tilde{X} \\ &= \sum_{i=1}^{r}\sigma_i u^{(i)}v^{(i)\top} - u^{(1)}u^{(1)\top}\sum_{i=1}^{r}\sigma_i u^{(i)}v^{(i)\top}.\end{aligned}$$

Simplify:

$$\tilde{\tilde{X}} = \sum_{i=1}^{r}\sigma_i u^{(i)}v^{(i)\top} - \sum_{i=1}^{r}\sigma_i\left(u^{(1)}u^{(1)\top}u^{(i)}\right)v^{(i)\top}.$$

Since $u^{(1)\top}u^{(i)} = 0$ for $i \neq 1$, we get:

$$\tilde{\tilde{X}} = \sum_{i=1}^{r}\sigma_i u^{(i)}v^{(i)\top} - \sigma_1 u^{(1)}v^{(1)T} = \sum_{i=2}^{r}\sigma_i u^{(i)}v^{(i)\top}.$$

So, the next strongest direction that explains the most variance is the singular vector corresponding to the second-largest singular value.

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

10.1 Low-rank approximation

Definition: A rank-k approximation of A is:

$$\sum_{i=1}^{k} \sigma_i u^{(i)} v^{(i)\top}.$$

Derivation: Let $A \in \mathbb{R}^{m\times n}$:

$$\begin{aligned} A &= U\Sigma V^\top \\ &= \begin{bmatrix} u^{(1)} & \cdots & u^{(r)} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} v^{(1)\top} \\ \vdots \\ v^{(r)\top} \end{bmatrix} \\ &= \sum_{i=1}^{r} \sigma_i u^{(i)} v^{(i)\top}. \end{aligned}$$

Assume $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.
A rank-k approximation of A is:

$$\sum_{i=1}^{k} \sigma_i u^{(i)} v^{(i)\top}.$$

It turns out that this is the **best rank-k approximation**. It minimizes the Frobenius norm:

$$\min_X \|A - X\|_F^2 \quad \text{s.t. } \operatorname{rank}(X) = k.$$

The solution is:

$$\begin{aligned} X^* &= \sum_{i=1}^{k} \sigma_i u^{(i)} v^{(i)\top} \\ &= \begin{bmatrix} u^{(1)} & \cdots & u^{(k)} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{bmatrix} \begin{bmatrix} v^{(1)\top} \\ \vdots \\ v^{(k)\top} \end{bmatrix}. \end{aligned}$$

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

11.1 Least squares

Definition: The least squares problem is:

$$\min_x \|Ax - b\|_2^2.$$

11.1.1 Motivation:

Intuition: Recall the system of linear equations:

$$Ax = b$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are given.

The goal is to solve for the unknown $x \in \mathbb{R}^n$.

If A is a square matrix and has full rank (i.e., A is invertible), then the solution is:

$$x^* = A^{-1}b.$$

What if A is not square or not full-rank? What can we say about the solutions to the system of linear equations?

There are two cases:
1. Overdetermined ($m > n$)
2. Underdetermined ($m < n$)

11.2 Overdetermined linear equation

Definition: Solve

$$\min_x \|Ax - b\|_2^2$$

The solution is $x^* = (A^\top A)^{-1} A^\top b = A^\dagger b$.

Derivation: **Projection Method:**

1. **Setup:**

$$[A_{m \times n}]\,[x_{n \times 1}] = [b_{m \times 1}]$$

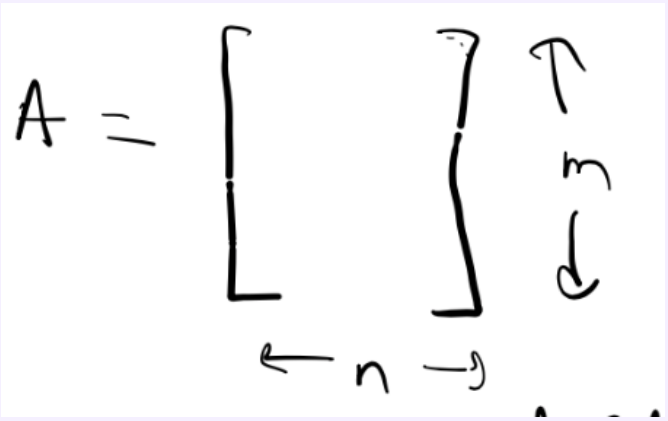

Figure 51: Tall matrix.

This represents a case where there are more equations than unknowns.

2. **Note:** Assume A has full column rank (columns of A are linearly independent). This implies:

$$\mathcal{N}(A) = \{x \mid Ax = 0\} = \{x \mid x_1 a^{(1)} + \cdots + x_n a^{(n)} = 0\} = \{0\},$$

where $\mathcal{N}(A)$ is the null space of A.

If A does not have full column rank, we can remove some of the columns of A to achieve linear independence.

3. **Projection Problem:** In general, there are no solutions to $Ax = b$, instead

$$\min_x \|Ax - b\|_2^2$$

Recall the range of A, denoted as $\mathcal{R}(A)$, is defined as:

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}.$$

Let $Ax^* \in \mathcal{R}(A)$, where x^* minimizes the norm:

$$\min_x \|Ax - b\|_2$$

or equivalently:

$$\min_{y \in \mathcal{R}(A)} \|y - b\|_2.$$

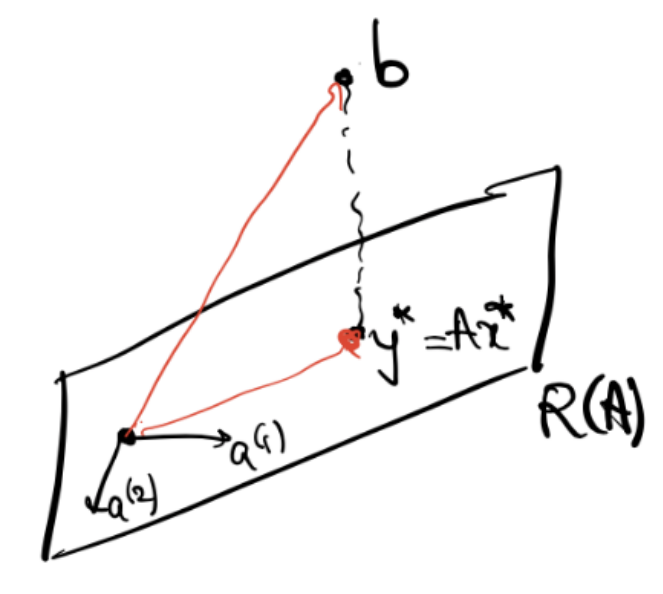

Figure 52: Least squares projection problem

Let $x^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_n^* \end{bmatrix}$, then:

$$Ax^* = \sum_{i=1}^{n} x_i^* a^{(i)},$$

where $a^{(i)}$ are the columns of A.

4. **Orthogonality Principle:** By the orthogonality principle:

$$b - Ax^* \perp a^{(j)} \quad \forall j.$$

This implies:

$$\langle b - \sum_{i=1}^{n} x_i^* a^{(i)}, a^{(j)} \rangle = 0 \quad \forall j.$$

Expanding:

$$\langle b, a^{(j)} \rangle = \sum_{i=1}^{n} x_i^* \langle a^{(i)}, a^{(j)} \rangle \quad \forall j.$$

In matrix form, this can be written as:

$$\begin{bmatrix} \langle a^{(1)}, a^{(1)} \rangle & \cdots & \langle a^{(1)}, a^{(n)} \rangle \\ \vdots & \ddots & \vdots \\ \langle a^{(n)}, a^{(1)} \rangle & \cdots & \langle a^{(n)}, a^{(n)} \rangle \end{bmatrix} \begin{bmatrix} x_1^* \\ \vdots \\ x_n^* \end{bmatrix} = \begin{bmatrix} \langle a^{(1)}, b \rangle \\ \vdots \\ \langle a^{(n)}, b \rangle \end{bmatrix} \implies A^\top A x^* = A^\top b.$$

Therefore, the solution to the overdetermined least squares problem is $x^* = (A^T A)^{-1} A^T b$

5. **Full Column Rank:** If A has full column rank, then $A^T A$ is invertible

$$\mathcal{N}(A) = \mathcal{N}(A^\top A) = \{0\}.$$

This implies that $\dim(\mathcal{R}(A^T A)) = n$, which is the number of non-zero eigenvalues.

The inverse can be expressed as:

$$(A^\top A)^{-1} = (U \Lambda U^\top)^{-1},$$

where U and Λ are components of the spectral decomposition, and the rank $r = n$, which means that all the components in the SVD are invertible.

Derivation: **Analytic Method to find the solution (Alterantive Way to Find the Solution):**

1. The objective is to minimize the function:

$$\min_x \|Ax - b\|_2^2,$$

denoted as $f(x)$.

2. Define:

$$f(x) = \|Ax - b\|_2^2 = (Ax - b)^\top (Ax - b).$$

3. Expanding the terms:

$$f(x) = x^\top A^\top A x - 2b^\top A x + b^\top b.$$

4. The gradient of $f(x)$ is:

$$\nabla f(x) = 2A^\top A x - 2A^\top b.$$

5. Setting $\nabla f(x) = 0$:

$$A^\top A x = A^\top b.$$

6. Solving for x^*:

$$x^* = (A^\top A)^{-1} A^\top b.$$

11.2.1 Why the solution has a pseudo inverse form?

Show $A^\dagger = (A^\top A)^{-1} A^\top$

Intuition:

1. **Proving Equality** Given the singular value decomposition (SVD):

$$A = U \Sigma V^\top, \quad \text{where} \quad \Sigma = \begin{bmatrix} \tilde{\Sigma}_{n \times n} \\ 0_{(m-n) \times n} \end{bmatrix}_{m \times n}.$$

2. **LS:** The pseudo-inverse of A is:

$$A^\dagger = V \Sigma^{-1} U^\top.$$

Expanding Σ^{-1}:

$$A^\dagger = V \begin{bmatrix} \tilde{\Sigma}^{-1} & 0 \end{bmatrix} U^\top,$$

where:

$$\tilde{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix}.$$

3. **RS:** For $(A^\top A)^{-1} A^\top$, then

$$\begin{aligned}
(A^\top A)^{-1} A^\top &= \left(V \Sigma^\top U^\top U \Sigma V^\top\right)^{-1} V \Sigma^\top U^\top \\
&= \left(V \Sigma^\top \Sigma V^\top\right)^{-1} V \Sigma^\top U^\top \\
&= \left(V \tilde{\Sigma}^2 V^\top\right)^{-1} V \Sigma^\top U^\top \quad \text{since } \Sigma^T \Sigma = \begin{bmatrix} \tilde{\Sigma} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma} \\ 0 \end{bmatrix} = \tilde{\Sigma}^2. \\
&= V \tilde{\Sigma}^{-2} V^\top V \begin{bmatrix} \tilde{\Sigma} & 0 \end{bmatrix} U^\top \\
&= V \tilde{\Sigma}^{-2} \begin{bmatrix} \tilde{\Sigma} & 0 \end{bmatrix} U^\top \\
&= V \begin{bmatrix} \tilde{\Sigma}^{-1} & 0 \end{bmatrix} U^\top \\
&= V \Sigma^{-1} U^\top = A^\dagger,
\end{aligned}$$

11.2.2 Pseudo Inverse

Intuition:

1. **Key 1:** The pseudo-inverse $A^\dagger$ is the "best" we can do for inverting A in $Ax = b$ (when A is a tall matrix). In fact,
$$A^\dagger A = I \quad (\text{but } AA^\dagger \neq I),$$
2. **Key 2:** Moreover, $AA^\dagger$ is the projection matrix onto $\mathcal{R}(A)$
$$AA^\dagger = \text{projection matrix onto } \mathcal{R}(A).$$
   - Since the projection is Ax^*, therefore, for a certain b and you want to project onto the range space of A, then the projection of b onto $\mathcal{R}(A)$ is:
   $$Ax^* = AA^\dagger b.$$
   - The solution x^* to the least squares problem is:
   $$x^* = (A^\top A)^{-1} A^\top b = A^\dagger b.$$

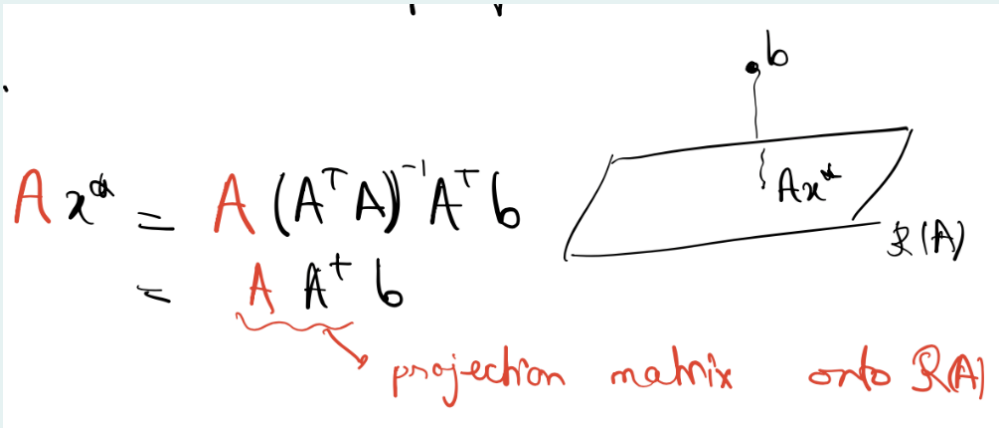

Figure 53: Projection matrix onto $\mathcal{R}(A)$

11.2.3 Application: Linear Regression

Example:

1. Given data points $\{(x_i, y_i)\}_{i=1}^n$, we want to fit a straight line $y_i = ax_i + b$ by minimizing the squared error:
$$\min_{(a,b)} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

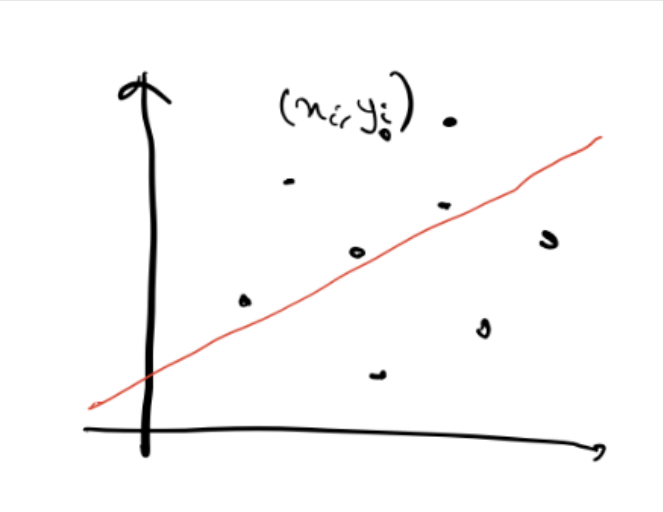

Figure 54: Linear regression

2. Rewriting in matrix form (i.e. reformulating to a least squares problem):

$$\min_{a,b} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2,$$

where:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad A = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad x = \begin{bmatrix} a \\ b \end{bmatrix}.$$

3. Solution:

$$\begin{bmatrix} a^* \\ b^* \end{bmatrix} = (A^\top A)^{-1} A^\top y,$$

assuming A has full column rank and $n > 2$.

4. **Gradient Derivation (Alternative way to solve the problem)**
We minimize:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = f(a, b).$$

Compute the partial derivatives:

$$\frac{\partial f}{\partial a} = \sum_{i=1}^{n} 2(y_i - ax_i - b)(-x_i) = 0,$$

$$\frac{\partial f}{\partial b} = \sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 0.$$

This yields two equations and two unknowns, which can be solved for a and b.

- **HW:** Check this YOURSELF.

Warning: EXAM QUESTION EXAM QUESTION!! He will do a linear regression problem on the exam. So verify the gradient derivation.

11.2.4 Application: Polynomial Regression

Example:

1. **Data:** Given data points $\{(x_i, y_i)\}_{i=1}^n$
2. **Model Representation** For an m-order polynomial:

$$y_i^* = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \cdots + \alpha_m x_i^m$$

3. **Objective Function** The goal is to minimize the squared error:

$$\min \sum_{i=1}^{n} (y_i - y_i^*)^2$$

Substituting the polynomial:

$$\min_{\alpha_0,\ldots,\alpha_m} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \alpha_1 x_i - \alpha_2 x_i^2 - \cdots - \alpha_m x_i^m\right)^2$$

4. **Matrix Formulation** Expressing in matrix form:

$$\min_{\boldsymbol{\alpha}} \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} \right\|_2^2$$

Let:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$$

Then the objective becomes:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2$$

5. **Closed-Form Solution:** If $\mathbf{X}$ is full column rank and $n > m + 1$, the least squares solution is given by:

$$\boldsymbol{\alpha} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

11.2.5 Regularized LS:

Example:

1. Consider the overdetermined case:

$$\min_x \|Ax - b\|_2^2.$$

2. In certain cases, we prefer a "smaller" x. To achieve this, we add a regularization term:

$$\min_x \|Ax - b\|_2^2 + \gamma\|x\|_2^2,$$

where $\gamma > 0$ controls the strength of regularization.
   - γ helps tune the trade off between $Ax - b$ and x.

3. Rewriting:

$$\begin{aligned} \min_x \|Ax - b\|_2^2 + \gamma\|x\|_2^2 &= \min_x \left\| \begin{bmatrix} Ax - b \\ \sqrt{\gamma}x \end{bmatrix} \right\|_2^2, \\ &= \min_x \left\| \begin{bmatrix} A \\ \sqrt{\gamma}I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2, \\ &= \min_x \|A'x - b'\|_2^2, \end{aligned}$$

where:

$$A' = \begin{bmatrix} A \\ \sqrt{\gamma} I \end{bmatrix}, \quad b' = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

4. The solution is:

$$\begin{aligned} x^* &= (A'^\top A')^{-1} A'^\top b', \\ &= \left(\begin{bmatrix} A^\top & \sqrt{\gamma} I \end{bmatrix} \begin{bmatrix} A \\ \sqrt{\gamma} I \end{bmatrix} \right)^{-1} \begin{bmatrix} A^\top & \sqrt{\gamma} I \end{bmatrix} \begin{bmatrix} b \\ 0 \end{bmatrix}, \\ &= \left(A^\top A + \gamma I \right)^{-1} A^\top b. \end{aligned}$$

This is known as **ridge regression**.

Example: **Regularized LS:** Analytical Method We aim to solve the regularized least squares problem:

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \gamma \|\mathbf{x}\|_2^2,$$

where A is an $m \times n$ matrix, $\mathbf{b}$ is an $m \times 1$ vector, $\gamma > 0$ is the regularization parameter, and $\mathbf{x}$ is the $n \times 1$ vector of variables.

1. **Define the objective function:** The cost function can be written as:

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \gamma \|\mathbf{x}\|_2^2.$$

Expanding the terms:

$$f(\mathbf{x}) = (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) + \gamma \mathbf{x}^T \mathbf{x}.$$

Expanding further:

$$f(\mathbf{x}) = \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} + \mathbf{b}^T \mathbf{b} + \gamma \mathbf{x}^T \mathbf{x}.$$

Simplify:

$$f(\mathbf{x}) = \mathbf{x}^T (A^T A + \gamma I)\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} + \mathbf{b}^T \mathbf{b},$$

where I is the identity matrix.

2. **Take the gradient of $f(\mathbf{x})$ with respect to x:** The gradient of $f(\mathbf{x})$ is:

$$\nabla f(\mathbf{x}) = 2(A^T A + \gamma I)\mathbf{x} - 2A^T \mathbf{b}.$$

3. **Set the gradient to zero:** To find the minimizer, set $\nabla f(\mathbf{x}) = 0$:

$$2(A^T A + \gamma I)\mathbf{x} - 2A^T \mathbf{b} = 0.$$

Simplify:

$$(A^T A + \gamma I)\mathbf{x} = A^T \mathbf{b}.$$

4. **Solve for x:** Assuming $A^T A + \gamma I$ is invertible, the solution is:

$$\mathbf{x} = (A^T A + \gamma I)^{-1} A^T \mathbf{b}.$$

11.3 Underdetermined linear equation

Definition: Solve

$$\min_{x,\ \text{s.t. } Ax=b} \|x\|_2^2$$

The solution is $x^* = A^\top (AA^\top)^{-1} b$.

Derivation:

1. Consider a matrix $A \in \mathbb{R}^{m \times n}$ with $m < n$:

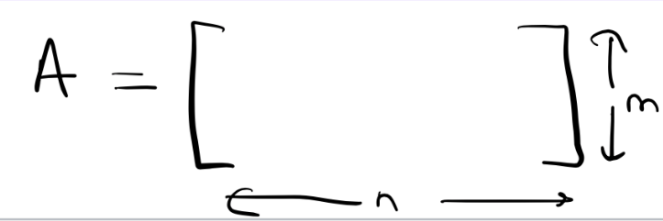

Figure 55: Fat matrix.

- Assume A has full row rank, i.e., $\text{rank}(A) = m$.

2. There are many vectors x that satisfy $Ax = b$ (i.e. infinitely many solutions). Out of all these x's, we aim to find the one with the least norm (i.e. the best solution):

$$\min_{x \text{ s.t. } Ax=b} \|x\|_2^2$$

   - **Note:** More unknowns than equations, so some unknowns can be arbitrary, which can lead to infinitely many solutions. This is because of the full row rank assumption (i.e. rows are linearly independent). If A is not full row rank, then there are no solutions, but it can still be solved (HW5) IS THIS CORRECT?

3. **Note** If $A\bar{x} = b$ where $\bar{x}$ is one solution to $Ax = b$, then for any $x_0 \in \mathcal{N}(A)$ (null space of A):

$$A(\bar{x} + x_0) = A\bar{x} + Ax_0 = A\bar{x} = b.$$

   This means we can write any solution x as:

$$x = \bar{x} + x_0, \quad \text{where } Ax_0 = 0.$$

   Since x_0 is part of the null space.

4. **Projection:** The projection of a vector y onto a set S is defined as:

$$\text{Proj}_S(y) = \arg\min_{z \in S} \|z - y\|_2.$$

   The projection of the zero vector onto S is:

$$x^* = \text{Proj}_S(0) = \arg\min_{z \in S} \|z\|_2.$$

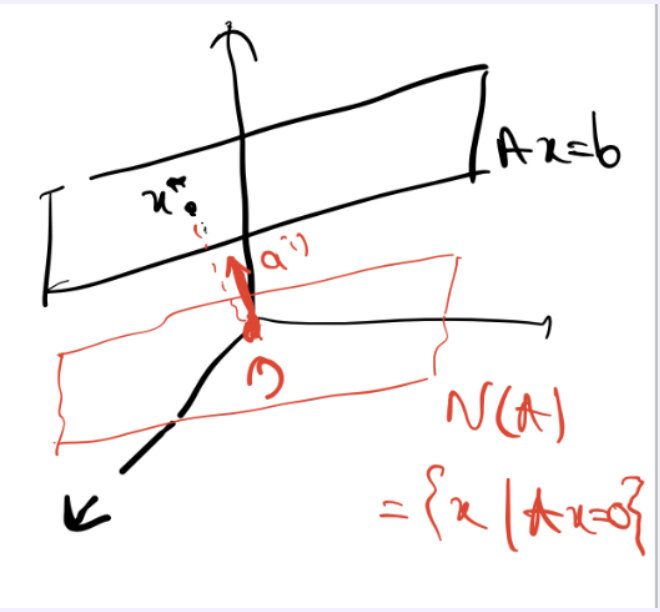

Figure 56: Projection

5. **Orthogonality Principle** By the orthogonality principle, the least-norm solution x^* should belong to the span of the rows of A, i.e.:
$$x^* \in \mathcal{R}(A^\top),$$
where:
$$\mathcal{R}(A^\top) = \mathcal{N}(A)^\perp,$$
the orthogonal complement of the null space of A.
   - This makes sense because you want x^* to be orthogonal to the null space of A as shown in the diagram above.
6. **Solution to the Underdetermined System**
We express the least-norm solution x^* as:
$$x^* = A^\top c \quad \text{for some } c.$$
To satisfy $Ax^* = b$, we substitute:
$$AA^\top c = b.$$
Since $AA^\top$ is invertible (because A has full row rank), we solve for c:
$$c = (AA^\top)^{-1} b.$$
Substituting c back, we find:
$$x^* = A^\top c = A^\top (AA^\top)^{-1} b.$$
Thus, the solution to the underdetermined system is:
$$x^* = A^\top (AA^\top)^{-1} b.$$
7. **Note:** The expression $A^\top (AA^\top)^{-1} = A^\dagger$ is the pseudo-inverse of A, denoted as $A^\dagger$. That is:
$$A^\dagger = A^\top (AA^\top)^{-1}.$$
Using the singular value decomposition (SVD), where $A = U\Sigma V^\top$ with $\Sigma = \begin{bmatrix} \tilde{\Sigma}_{m\times m} & 0 \end{bmatrix}$, we can show that:
$$A^\top (AA^\top)^{-1} = A^\dagger.$$
   - **Note:** This is similar to how we derived the pseudo-inverse for the overdetermined case.
   - $\tilde{\Sigma}$ is $m \times m$ because it is full row rank.

Derivation: Analytic Method to Solve Underdetermined Least Squares

1. We aim to solve:
$$\min \|x\|_2^2 \quad \text{subject to } Ax = b,$$
where $\|x\|_2^2$ is the objective and $Ax = b$ represents the m constraints. This is a constrained optimization problem.
2. **Lagrangian Formulation:** We form the Lagrangian function:
$$\mathcal{L}(x, \lambda) = \|x\|_2^2 + \lambda^\top (Ax - b),$$
where λ can be interpreted as a "price" that penalizes the deviation of Ax from b.
   - **Note:** The Lagrangian is the objective function plus the constraints multiplied by the Lagrange multipliers.
     - $Ax - b$ is the constraint, which is multiplied by the Lagrange multiplier λ.
3. **Solving the Problem**
For fixed λ, minimize $\mathcal{L}(x, \lambda)$ over x:
$$\nabla_x \mathcal{L} = 2x + A^\top \lambda = 0 \quad \Rightarrow \quad x = -\frac{1}{2} A^\top \lambda.$$
Substitute into the constraint $Ax = b$ to ensure that this x satisfies the constraint:
$$A\left(-\frac{1}{2} A^\top \lambda\right) = b \quad \Rightarrow \quad -\frac{1}{2} AA^\top \lambda = b.$$

Solve for λ:

$$\lambda = -2(AA^\top)^{-1}b.$$

- This is telling me the optimal penalty to minimize the Langrangian function.

Substitute back to find x:

$$x = -\frac{1}{2}A^\top\lambda = A^\top(AA^\top)^{-1}b.$$

4. **Conclusion** The solution matches the least-norm solution derived earlier:

$$x = A^\top(AA^\top)^{-1}b.$$

- **General:** Very difficult to find a closed form solution of x in terms of λ, but in this case, we can find it. This is one of the few problems where we can find a closed form solution.

Warning: For constrained optimization, you cannot just set the gradient to zero because it's not gauranteed to satisfy the constraints. However, if you use the Lagrangian method, you are effectively solving an unconstrainted optimization problem, so that you can set the gradient to zero.

Process:

1. Form the lagrangian function
2. Take the gradient of the lagrangian function with respect to x and set it to zero
3. Solve for x in terms of λ
4. Substitute x back into the constraint to solve for λ
5. Substitute λ back into x to solve for x

11.3.1 Application: Optimal Control

Example:

1. **State of the Mass**

Assume $m = 1$. The state of the mass is represented as:

$$\begin{bmatrix} x[n] \\ v[n] \end{bmatrix},$$

where $x[n]$ is the position and $v[n]$ is the velocity.

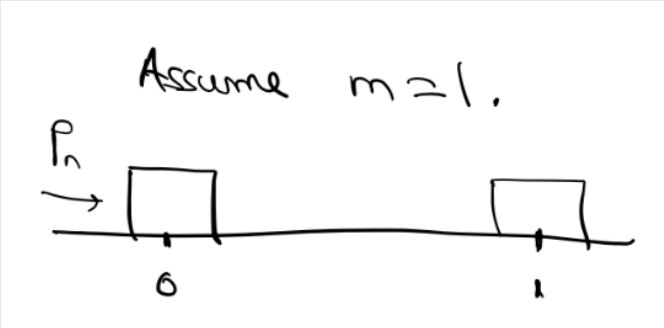

Figure 57: Mass system

- Pushing the mass with force p_n at time n from position 0 to position 1.

2. The equations of motion are:

$$x(t) = x(0) + vt + \frac{1}{2}at^2, \quad v(t) = v(0) + at.$$

- $p_n = ma = a$ since the mass is 1 (i.e. applied force is equal to the acceleration)

3. **Goal:** We want to find the optimal force p_n (i.e. least amount of force) to push the mass from position 0 to position 1.

4. Discretizing, we have:

$$x[n] = x[n-1] + v[n-1] + \frac{1}{2}p_n,$$
$$v[n] = v[n-1] + p_n.$$

Therefore, in matrix form, we can write it as:

$$\begin{bmatrix} x[n] \\ v[n] \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x[n-1] \\ v[n-1] \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} p_n,$$

Define:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}.$$

The system evolves as:

$$x_n = Ax_{n-1} + Bp_n.$$

5. **Iterative Evolution** Suppose $x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and the final state is $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ after 10 time steps. Then:

$$\begin{bmatrix} x[10] \\ v[10] \end{bmatrix} = A \begin{bmatrix} x[9] \\ v[9] \end{bmatrix} + Bp_{10}.$$

Expanding recursively:

$$\begin{bmatrix} x[10] \\ v[10] \end{bmatrix} = A \left(A \begin{bmatrix} x[8] \\ v[8] \end{bmatrix} + Bp_9 \right) + Bp_{10}.$$

Simplify to:

$$\begin{bmatrix} x[10] \\ v[10] \end{bmatrix} = \begin{bmatrix} A^{10}B & A^9B & A^8B & \cdots & AB & B \end{bmatrix} \begin{bmatrix} p_0 \\ \vdots \\ p_{10} \end{bmatrix}.$$

Let:

$$C = \begin{bmatrix} A^{10}B & A^9B & A^8B & \cdots & AB & B \end{bmatrix}_{2\times 11}, \quad p = \begin{bmatrix} p_0 \\ \vdots \\ p_{10} \end{bmatrix}_{11\times 1}.$$

6. **Optimization Problem:** We want to solve:

$$\min \|p\|_2^2 \quad \text{subject to } Cp = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

- $Cp = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ because we want to reach the final state of position 1 with velocity 0.
- Minimize the force applied to the mass subject to the constraint of reaching the final state.

7. This is an underdetermined least squares problem, with the solution:

$$p^* = C^\top (CC^\top)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

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

12.1 Regularized least-squares

Derivation:

1. Recall the overdetermined least squares (LS):

$$\min_x \|Ax - b\|_2$$

where $A \in \mathbb{R}^{m \times n}$, with $m > n$.

2. Regularized Least Squares (LS):

$$\min_x \|Ax - b\|_2^2 + \gamma \|x\|_2^2$$

- This is a tradeoff between minimizing the norm of x and solving the LS problem.

The solution is given by:

$$x^* = \left(A^\top A + \gamma I\right)^{-1} A^\top b$$

3. To understand regularized LS, we can formulate it as a constrained optimization problem, rather than an unconstrainted optimization problem:

$$\min_{x \text{ s.t. } \|x\|_2^2 \leq t} \|Ax - b\|_2^2$$

The optimal point x^{**} can be represented as:

$$x^{**} = \text{argmin}_{\text{s.t. } \|x\|_2^2 \leq t} \|Ax - b\|_2^2$$

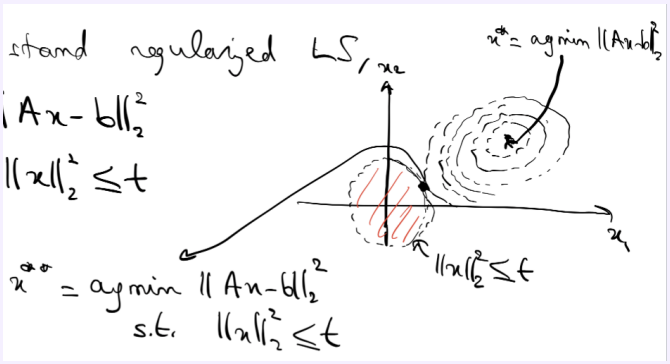

Figure 58: Regularized Least Squares

- The constraint is the unit circle in the 2D case, and the $||Ax - b||_2^2$ is the contour ellipse plot that grows in size until it intersects with the constraint, which is the optimal point.

Derivation:

1. In many cases, we are looking for a sparse solution x, i.e., we aim to solve:

$$x^* = \min_{x \text{ s.t. } \|x\|_0 \leq t} \|Ax - b\|_2^2$$

Here, $\|x\|_0$ represents the number of non-zero entries in x.

2. **Convex Relaxation** The ℓ_0-norm constraint is difficult to handle. Instead, we replace it with the ℓ_1-norm:

$$\min_{x \text{ s.t. } \|x\|_1 \leq t} \|Ax - b\|_2^2$$

- $l1$ norm encourages sparsity, but $l2$ norm does not.

3. **LASSO Optimization:** The equivalent formulation is:

$$\min_x \|Ax - b\|_2^2 + \gamma \|x\|_1$$

- The LASSO optimization problem does not have an analytic solution, but it can be solved numerically since it belongs to the class of convex optimization problems. For $1D$, it does have an analytic solution.

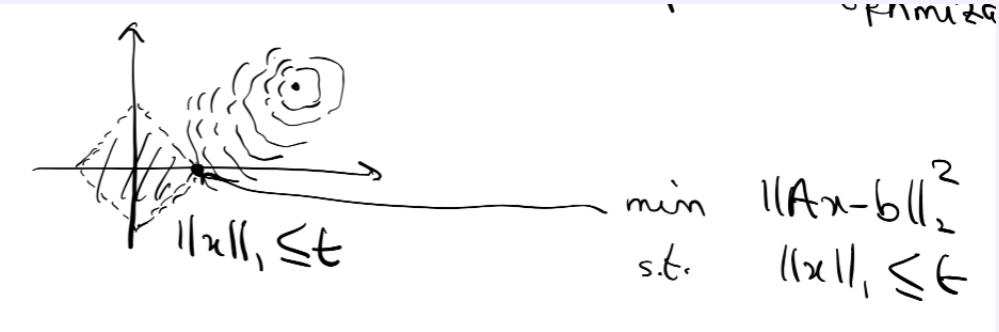

Figure 59: Optimization solution for $l1$ norm

- **Note:** The solution occurs with one of the entries being 0, indicating the sparse solution that can be achieved with $l1$ norm.
- **Why Sparse:** Reduces the memory and computational complexity since we only need to store and compute the non-zero entries.

12.1.1 Application: Sparse Coding of Images

Example:

- Suppose $M \in \mathbb{R}^{n\times n}$ is an image.
- We aim to encode M into $\tilde{M} = HMH^\top$, where H is an orthogonal matrix chosen such that $\tilde{M}$ is sparse.
- Instead of encoding M exactly, we allow some loss.

We aim to solve the following optimization problem:

$$\min_X \frac{1}{2}\|H^\top XH - M\|_F^2 + \lambda\|X\|_1,$$

where $\|X\|_1 = \sum_{ij} |x_{ij}|$.

- **Intuition:** First term is the loss, which is comparing the decoding of X with the original image M. The second term is the sparsity constraint.
- Note: $\tilde{M} = HMH^\top \implies M = H^\top \tilde{M} H$.
- It turns out that this is equivalent to the LASSO optimization in the 1-dimensional case:

$$\min_u \frac{1}{2}(u-a)^2 + \lambda|u|,$$

where $f(u) = \frac{1}{2}(u-a)^2 + \lambda|u|$.

Solving the Optimization Problem

1. If $u \ge 0$ and $a > \lambda$, then:
$$\frac{\partial f(u)}{\partial u} = 0 \implies u - a + \lambda = 0 \implies u = a - \lambda.$$
2. If $u \le 0$ and $a < -\lambda$, then:
$$\frac{\partial f(u)}{\partial u} = 0 \implies u - a - \lambda = 0 \implies u = a + \lambda.$$
3. If $0 < a < \lambda$ and $u \ge 0$, then:
$$\frac{\partial f(u)}{\partial u} = u - a + \lambda > 0,$$
indicating that f is increasing on $u > 0$. Thus, $u^* = 0$ is the solution in this case.

4. Similarly, for $-\lambda < a < 0$ and $u < 0$, the solution is $u^* = 0$.

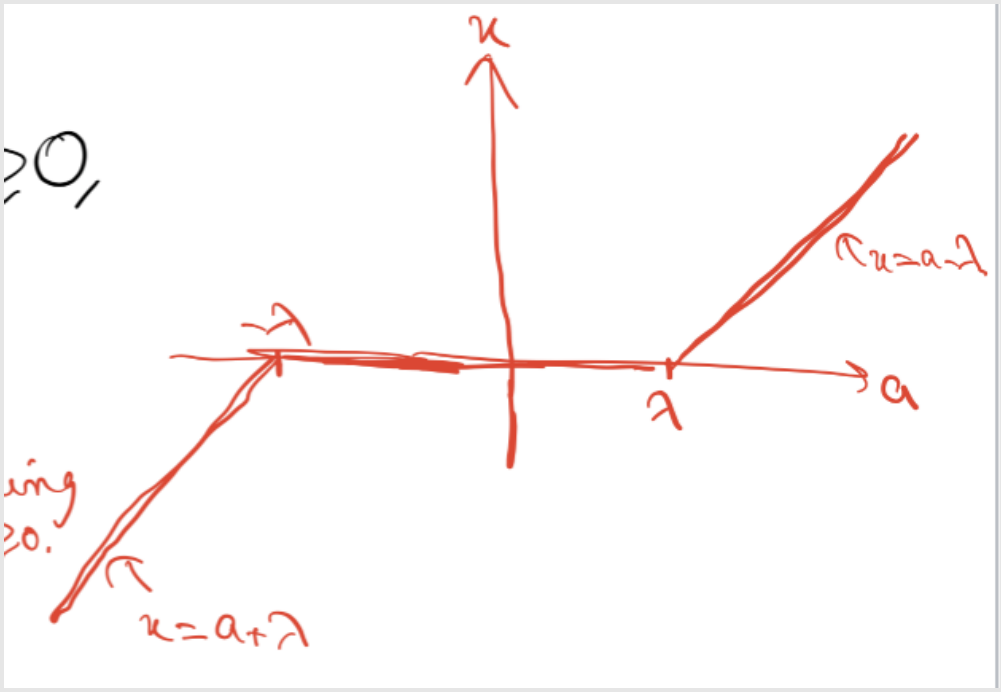

Figure 60: Optimization solution for LASSO in 1D

12.2 Convex sets and convex functions

12.2.1 General Form of an Optimization Problem

Definition:

- The general form of an optimization problem is:

$$\min f_o(x) \quad \text{subject to} \quad \begin{array}{ll} f_i(x) \leq 0, & i = 1, \ldots, m \quad \text{(inequality constraints)} \\ h_i(x) = 0, & i = 1, \ldots, p \quad \text{(equality constraints)} \end{array}$$

- The **feasible set** is defined as:

$$C = \{x \mid f_i(x) \leq 0,\ i = 1, \ldots, m,\ h_i(x) = 0,\ i = 1, \ldots, p\}.$$

- The **optimal value** p^* is:

$$p^* = \min f_o(x) \quad \text{subject to} \quad x \in C.$$

- The **optimal point/optimal solution** x^* is:

$$x^* = \arg\min f_o(x) \quad \text{subject to} \quad x \in C.$$

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

13.1 Lagrangian method for constrained optimization

13.2 Linear programming and quadratic programming

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

14.1 Numerical algorithms for unconstrained optimization

14.2 Numerical algorithms for constrained optimization