DS-GA 1014

Optimization and Computational Linear Algebra

Instructor: Florentin Guth

Notes by: Andrew Liao

November 28, 2025

Contents

1 Vector Spaces

1.1 Vector Spaces

Definition 1.1 (Vector)**.** A **vector** $\bar{v}$ is a mathematical object defined by the operations of vector addition and scalar multiplication. Vectors have a geometric interpretation as arrows in space and a numerical interpretation as ordered lists of real numbers.

Definition 1.2 (Vector Space)**.** A **vector space** V is a set of vectors satisfying the following axioms for all $x, y, z \in V$ and scalars $a, b, c, d \in \mathbb{R}$:

1. Closure under addition: $x + y \in V$
2. Commutativity: $x + y = y + x$
3. Associativity: $(x+y)+z = x+(y+z)$
4. Additive identity: $\exists\, 0 \in V$ such that $x + 0 = x$
5. Additive inverse: $\exists\, (-x) \in V$ such that $x + (-x) = 0$
6. Closure under scalar multiplication: $cx \in V$
7. Compatibility: $c(dx) = (cd)x$
8. Multiplicative identity: $1 \cdot x = x$
9. Distributivity: $a(x + y) = ax + ay$
10. Scalar distributivity: $(a + b)x = ax + bx$

Both vectors and vector spaces are defined through operations rather than representation.

Definition 1.3 (Subspace)**.** A **subspace** $S \subseteq V$ is a subset that:

1. contains the zero vector,
2. is closed under vector addition,
3. is closed under scalar multiplication.

If these conditions hold, S is a vector space.

Examples

- $\mathbb{R}^n$ is a subspace of itself.
- $\{0\}$ is a subspace of any vector space.
- Any line through the origin is a subspace of $\mathbb{R}^2$.

1.2 Span and Linear Dependence

Definition 1.4 (Linear Combination)**.** A vector $y \in V$ is a **linear combination** of $x_1, \ldots, x_k$ if

$$y = \sum_{i=1}^{k} \alpha_i x_i,$$

for scalars $\alpha_i \in \mathbb{R}$.

Definition 1.5 (Span)**.** The **span** of vectors $x_1, \ldots, x_n$ is the set of all linear combinations.

$$\text{span}(x_1, \ldots, x_n) = \{\alpha_1 x_1 + \cdots + \alpha_n x_n : \alpha_i \in \mathbb{R}\},$$

the smallest subspace of V that contains them.

Definition 1.6 (Linear Dependence)**.** Vectors $x_1, \ldots, x_k$ are **linearly dependent** if not all scalars $\alpha_1, \ldots, \alpha_k$ are zero and

$$\alpha_1 x_1 + \cdots + \alpha_k x_k = 0.$$

They are **linearly independent** otherwise.

Examples

1. The set (x) is linearly independent iff $x \neq 0$.
2. The set $(x, -x)$ is always linearly dependent because $x + (-x) = 0$.

1.3 Basis and Dimension

Definition 1.7 (Basis)**.** A family $x_1, \ldots, x_n$ is a **basis** of V if

1. $x_1, \ldots, x_n$ are linearly independent,
2. $\text{span}(x_1, \ldots, x_n) = V$.

Definition 1.8 (Dimension)**.** If every basis of V has n vectors, we say $\dim(V) = n$. If no finite basis exists, $\dim(V) = +\infty$.

1.3.1 Properties

Let $x_1, \ldots, x_n \in V$ s.t. $\dim(V) = n$.

1. If $x_1, \ldots, x_n$ are linearly independent, then $x_1, \ldots, x_n$ is a basis for V (We get span for free).
2. If $Span(x_1, \ldots, x_n) = V$, then $(x_1, \ldots, x_n)$ is the basis of V (we get linear independence for free).

To show that a family of vectors $x_1, \ldots, x_k$ form a basis, we need to show:

1. $k = n$
2. linear independence
3. $Span(x_1, \ldots, x_k) = V$

From above we see that showing any 2 properties implies the third one. We can trivially show $k = n$, so we usually choose to show the easier of 2 and 3.

Definition 1.9 (Line and Hyperplane)**.** Let S be a subspace.

1. If $\dim(S) = 1$, S is a **line**.
2. If $\dim(S) = n - 1$, S is a **hyperplane**.

1.4 Coordinates of a vector in a basis

Theorem 1.1 (Uniqueness of Coordinates)**.** *If $v_1, \ldots, v_n$ is a basis of V, then for every $x \in V$, there exists a unique vector $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ such that $x = \alpha_1 v_1 + \cdots + \alpha_n v_n$. We refer to $(\alpha_1, \ldots, \alpha_n)$ as the **coordinates** of x.*

Proof. $Span(v_1, \ldots, v_n) = V$ since it is the basis of V. Since x is some linear combination of $v_1, \ldots, v_n$, $x \in Span(v_1, \ldots, v_n)$. Therefore, there exists some vector α such that the linear combination of v and α is x. Suppose there is another set of vectors $\beta_1, \ldots, \beta_n$ such that x is a linear combination of v and β.

$$x - x = (\alpha_1 v_1 + \cdots + \alpha_n v_n) - (\beta_1 v_1 + \cdots + \beta_n v_n) = 0$$

Since v is linearly independent, $\alpha_i = \beta_i$ for $1 \leq i \leq n$. Therefore there is a unique vector in each basis to obtain x. □

2 Linear Transformations and Matrices

2.1 Review

There are 2 interpretations of vectors: geometric and numerical. These two interpretations *bridge the abstract world and algorithmic world.*

- The *geometric world* provides the vectors, which contain scale and directionality.
- The *numerical world* provides the basis, which are the coordinates to project the vectors to.
  - Let basis be $e_1, e_2, \ldots, e_n$.
  - $\vec{v} = v_1 e_1 + v_2 e_2 + \cdots + v_n e_n$.

In $\mathbb{R}^n$ with the canonical basis, it is trivial because $\vec{v} = (v_1, \ldots, v_n) = v_1\vec{e_1} + \cdots + v_n\vec{e_n}$. However, when the vectors are functions, since they don't live in $\mathbb{R}^n$, they don't really have components and the $\mathbb{R}^n$ vector geometric interpretation is invalid. However, we can still describe the coordinates of a function in a given basis. It is key to distinguish the two.

2.2 Linear Transformations

Geometric interpretation: operations done on vectors with linear properties. *Numerical interpretation:* arrays of numbers, aka matrices.

2.2.1 Definition

Symmetries and rotations are linear mappings. So they are functions from vectors to vectors.

$$L: \quad \mathbb{R}^2 \to \quad \mathbb{R}^2 \tag{1}$$

$$v \mapsto \quad L(v) \tag{2}$$

Definition 2.1 (Linear Transformation)**.** A function $L : \mathbb{R}^m \to \mathbb{R}^n$ is linear $\iff$

1. Closed under addition: $\forall v, w \in \mathbb{R}^m, L(v + w) = L(v) + L(w)$
2. Closed under scalar multiplication: $\forall v \in \mathbb{R}^m, \alpha \in \mathbb{R}, L(\alpha v) = \alpha L(v)$

Note:

1. They don't have to be in the same vector space!
2. They are transformations that comply with closure under linear combinations.

Examples

1. $\mathbb{R}^2 \to \mathbb{R}^3$ linear transformation:

$$L: \quad \mathbb{R}^2 \to \quad \mathbb{R}^3 \tag{3}$$
$$(v_1, v_2) \mapsto \quad (5v_1, 0, v_1 + v_2) \tag{4}$$

Proof. **Closure under addition:** Let $(v_1, v_2), (w_1, w_2) \in \mathbb{R}^2$.

$$L(v + w) = L((v_1 + w_1), (v_2 + w_2)) \tag{5}$$
$$= (5(v_1 + w_1), 0, (v_1 + w_1) + (v_2 + w_2)) \tag{6}$$
$$= (5v_1, 0, v_1 + v_2) + (5w_1, 0, w_1 + w_2) \tag{7}$$
$$= L(v) + L(w) \tag{8}$$

Closure under scalar multiplication: Let $(v_1, v_2) \in \mathbb{R}^2, \lambda \in \mathbb{R}$.

$$L(\lambda v) = L(\lambda v_1, \lambda v_2) \tag{9}$$
$$= (5\lambda v_1, 0, \lambda(v_1 + v_2)) \tag{10}$$
$$= \lambda(5v_1, 0, (v_1 + v_2)) \tag{11}$$
$$= \lambda L(v) \tag{12}$$

$\therefore L$ is a linear transformation. □

2. $\mathbb{R} \to \mathbb{R}$ non-linear transformation:

$$L: \quad \mathbb{R} \to \mathbb{R} \tag{13}$$
$$x \mapsto \quad x^2 \tag{14}$$

Proof. Closure under addition: Let $(v_1), (w_1) \in \mathbb{R}$.

$$L(v + w) = (v_1 + w_1)^2 \tag{15}$$
$$= v_1^2 + w_1^2 + 2vw \tag{16}$$
$$L(v) + L(w) = v_1^2 + w_1^2 \tag{17}$$

$\therefore L(v + w) \neq L(v) + L(w)$, so L is not a linear transformation. Or, using a counter example to demonstrate the square of the sums is not equal to the sum of the squares:

$$L(1 + 2) = (1 + 2)^2 = 9 \tag{18}$$
$$L(1) + L(2) = 1 + 4 = 5 \tag{19}$$

$5 \neq 9$, therefore L is not a linear transformation. □

Proposition 2.1 (Properties of Linear Maps)**.** *Let $L : \mathbb{R}^m \to \mathbb{R}^n$ be linear, then:*

1. $L(0) = 0$*:*

$$L(0) = L(a - a) = L(a + (-a)) = L(a) + L(-a) = L(a) - L(a) = 0$$

2. *Distributive over addition and scalar multiplication:*

$$L\left(\sum_{i=1}^{k} \alpha_i v_i\right) = \sum_{i=1}^{k} \alpha_i L(v_i), \quad \forall \alpha_i \in \mathbb{R}, v_i \in \mathbb{R}^m$$

3. *Composition of linear maps is also linear. If* $L : \mathbb{R}^m \to \mathbb{R}^n$ *and* $M : \mathbb{R}^n \to \mathbb{R}^k$ *are both linear, then the composite function is also linear.*

2.3 Matrices

Let $L : \mathbb{R}^m \to \mathbb{R}^n$ be a linear transformation and $(e_1, e_2, \ldots, e_m)$ be the canonical basis of $\mathbb{R}^m$. Then, for all $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$:

$$x = (x_1, 0, \ldots, 0) + (0, x_2, \ldots, 0) + \cdots + (0, \ldots, 0, x_m) \tag{20}$$

$$= x_1 e_1 + x_2 e_2 + \cdots + x_m e_m \tag{21}$$

Then,

$$L(x) = L\left(\sum_{i=1}^{m} x_i e_i\right) = \sum_{i=1}^{m} x_i L(e_i) \tag{22}$$

From the vectors $L(e_1), \ldots, L(e_m) \in \mathbb{R}^n$, we can compute all $L(x) \quad \forall x \in \mathbb{R}^m$, meaning once we know what a linear transformation does to a basis, we know what it does to every possible vector. So, **a linear map is determined by its action on a basis**.
Idea. Why matrices are basically a change in basis: First, decompose any input vector into any basis of the input space. Then, apply the linear map to the basis in the input space. You get vectors in the output space, where you decompose them into any basis of the output space. You end up with an array that tells you everything you need to know about the linear mapping.

Definition 2.2 (Matrix)**.** An $n \times m$ matrix is an array with n rows and m columns. We denote by $\mathbb{R}^{n \times m}$ the set of all $n \times m$ matrices.

Definition 2.3 (Canonical Matrix of a linear map)**.** We can encode a linear map $L : \mathbb{R}^m \to \mathbb{R}^n$ by a $n \times m$ matrix. The canonical matrix of L is the $n \times m$ matrix $\tilde{L}$ whose columns are $L(e_1), \ldots, L(e_m)$.

$$\tilde{L} = \begin{pmatrix} | & | & \cdots & | \\ L(e_1) & L(e_2) & \cdots & L(e_m) \\ | & | & \cdots & | \end{pmatrix} = \begin{pmatrix} L_{1,1} & L_{1,2} & \cdots & L_{1,m} \\ L_{2,1} & L_{2,2} & \cdots & L_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & L_{n,m} \end{pmatrix} \tag{23}$$

Where we write $L(e_j) = (L_{1,j}, L_{2,j}, \ldots, L_{n,j})^\top$.

Examples

1. **Identity matrix I:**

$$\tilde{L} = \begin{pmatrix} | & | \\ L(1,0) & L(0,1) \\ | & | \end{pmatrix} = \begin{pmatrix} 5\times 1 & 5\times 0 \\ 0 & 0 \\ 1+0 & 0+1 \end{pmatrix} \tag{24}$$

$$\tilde{Id} = Id \tag{25}$$

2. **Homothety (Scale):** Let $\lambda \in \mathbb{R}$. The homothety map of ratio λ is linear.

$$H_\lambda(e_1) = \lambda(e_1), \ldots, H_\lambda(e_n) = \lambda e_n \tag{26}$$

$$\tilde{H_\lambda} = \begin{bmatrix} \lambda & 0 & \ldots & 0 \\ 0 & \lambda & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda \end{bmatrix} = \lambda I_n \tag{27}$$

3. **Rotations in $\mathbb{R}^2$:** Let $\theta \in \mathbb{R}$. The rotation $\mathbb{R}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ of angle θ about the origin is linear.

$$\mathbb{R}_\theta(e_1) = (cos\theta, sin\theta) \tag{28}$$

$$\mathbb{R}_\theta(e_2) = (-sin\theta, cos\theta) \tag{29}$$

$$\tilde{\mathbb{R}_\theta} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix} \tag{30}$$

Proposition 2.2 (Matrix-Vector Product)**.** *Consider a linear map $L : \mathbb{R}^m \to \mathbb{R}^n$ and its associated matrix $\tilde{L} \in \mathbb{R}^{n\times m}$.*

$$L(x)_i = L\left(\sum_j x_j e_j\right)_i \tag{31}$$

$$= \left(\sum_j x_j L(e_j)\right)_i = \sum_j x_j L(e_j)_i = \sum_j \tilde{L}_{ij} x_j \tag{32}$$

So, for all $x \in \mathbb{R}^m$ we have $L(x) = \tilde{L}x$. Each element in the output matrix corresponds to the weighted sum of a row in the input matrix and the vector such that the vector is the scaling factor.

2.4 Matrix operations

1. **Addition and Scalar Multiplication:** Since matrices are linear maps, they form a vector space $\mathbb{R}^{n\times m}$ with dimension $n \times m$.

2. **Matrix Product:** The matrix product ML is the $k \times n$ matrix of the linear map $M \circ L$.

$$(ML)_{i,j} = \sum_{l=1}^{m} M_{i,l} L_{l,j} \quad \forall 1 \leq i \leq k, 1 \leq j \leq n \tag{33}$$

3. **Properties:**

   - $(A + B)C = AC + BC$
   - $A(C + D) = AC + AD$
   - $AId_m = A$
   - Not commutative: $AB \neq BA$ usually.
   - No division: If $AB = AC$, it does not imply $B = C$ (unless A is invertible).

2.4.1 Invertible matrices

A square matrix $M \in \mathbb{R}^{n \times n}$ is **invertible** if there exists a unique matrix $M^{-1} \in \mathbb{R}^{n \times n}$ such that $MM^{-1} = M^{-1}M = Id_n$.

2.5 Kernels and Images

Definition 2.4 (Kernel)**.** The **kernel** (or nullspace) of L is the set of all vectors $v \in \mathbb{R}^m$ such that $L(v) = 0$.

$$Ker(L) := \{v \in \mathbb{R}^m \mid L(v) = 0\}$$

$Ker(L)$ is a subspace of $\mathbb{R}^m$.

Definition 2.5 (Image)**.** The **image** (aka range, column space) of L is the set of all vectors $v \in \mathbb{R}^n$ such that there exists $v \in \mathbb{R}^m$ such that $L(v) = v$.

$$Im(L) = \{L(v) \mid v \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$$

$\text{Im}(L)$ is a subspace of $\mathbb{R}^n$.

2.6 Application to ML

Goal. Given data of input-output pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$, predict y for a new x.

Idea. Hypothesis: $\exists(\theta_1, \ldots, \theta_m)$ s.t. $\theta_1 x_{i,1} + \cdots + \theta_m x_{i,m} = y_i$. This forms a system $X\theta = y$.

- If $y \notin \text{Im}(X)$: No solution.

- If $y \in \text{Im}(X)$: At least 1 solution.
- If $Ker(X) = \{0\}$: One unique solution.
- If $Ker(X) \neq \{0\}$: Infinite solutions (Affine space $\theta_0 + Ker(X)$).

2.6.1 Gaussian Elimination

Goal: To get matrices into row echelon form. **Rules:** You are allowed 3 operations: 1. Row swap. 2. Row scale. 3. Row addition.

3 Matrix Rank

Idea. Motivation: Consider a dataset $x_1, \dots, x_n \in \mathbb{R}^d$. What is its dimensionality? There are many notions of dimensionality. The rank is one of them.

3.1 The Rank

Definition 3.1 (Rank of a Family)**.** We can define the rank of a family $x_1, \dots, x_k$ of vectors in $\mathbb{R}^n$ as the dimension of its span.

$$rank(x_1, \dots, x_k) \stackrel{\text{def}}{=} dim(Span(x_1, \dots, x_k))$$

Definition 3.2 (Rank of a Matrix)**.** Let $M \in \mathbb{R}^{n \times m}$ and $c_1, \dots, c_m \in \mathbb{R}^n$ be its columns.

$$rank(M) \stackrel{\text{def}}{=} rank(c_1, \dots, c_m) \tag{34}$$

$$= dim(\text{Im}(M)) = dim(Span(c_1, \dots, c_m)) \tag{35}$$

$rank(M) \leq \min(m, n)$.

Examples

1. $rank(Id_n) = dim(\mathbb{R}^n) = n$.
2. $rank(\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}) = 2$ (vectors are linearly independent).
3. $rank(\begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}) \leq 2$.

Proposition 3.1 (Rank of columns = Rank of rows)**.** *Let $r_1, \dots, r_n$ be the rows and $c_1, \dots, c_m$ be the columns of M.*

$$rank(r_1, \dots r_n) = rank(c_1, \dots c_m) = rank(M)$$

3.2 The Rank-Nullity Theorem

Theorem 3.2 (Rank-Nullity)**.** *Let $L : \mathbb{R}^m \to \mathbb{R}^n$ be a linear transformation. Then:*

$$rank(L) + dim(Ker(L)) = m$$

Proof. Let $v_1, \dots v_k$ be a basis of $Ker(L)$ ($k = dim(Ker(L))$). Complete it with $v_{k+1}, \dots, v_m$ into a basis of $\mathbb{R}^m$.

$$\text{Im}(L) = \{L(x), x \in \mathbb{R}^m\} \tag{36}$$

$$= \{\alpha_{k+1} L(v_{k+1}) + \dots + \alpha_m L(v_m) \mid \alpha_{k+1}, \dots, \alpha_m \in \mathbb{R}\} \tag{37}$$

$$= Span(L(v_{k+1}), \dots, L(v_m)) \tag{38}$$

To show $L(v_{k+1}), \ldots, L(v_m)$ are linearly independent: Assume $\sum_{i=k+1}^{m} \alpha_i L(v_i) = 0$. Then $L(\sum \alpha_i v_i) = 0$, so $\sum \alpha_i v_i \in Ker(L)$. This implies $\sum_{i=k+1}^{m} \alpha_i v_i = \sum_{j=1}^{k} \beta_j v_j$. Since $v_1, \ldots, v_m$ is a basis, all coefficients must be zero. Thus $L(v_{k+1}), \ldots, L(v_m)$ are independent. So $rank(L) = m - k = m - dim(Ker(L))$. □

3.3 Inequalities

Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$.

1. $rank(A) \leq \min(n, m)$.
2. $rank(AB) \leq \min(rank(A), rank(B))$.

Proof. **Showing** $rank(AB) \leq rank(A)$**:** $\text{Im}(AB) = \{ABx, x \in \mathbb{R}^k\} \subseteq \text{Im}(A)$. Since the subspace is contained, the dimension is smaller or equal.

Showing $rank(AB) \leq rank(B)$**:** We show $Ker(B) \subseteq Ker(AB)$. Let $x \in Ker(B) \implies Bx = 0 \implies ABx = A(0) = 0 \implies x \in Ker(AB)$. Since $Ker(B) \subseteq Ker(AB)$, $dim(Ker(B)) \leq dim(Ker(AB))$. By Rank-Nullity: $rank(AB) = k - dim(Ker(AB)) \leq k - dim(Ker(B)) = rank(B)$. □

3.4 Rank of invertible matrices

Theorem 3.3. *Let $M \in \mathbb{R}^{n \times n}$. The following points are equivalent:*

1. *M is invertible.*
2. *$rank(M) = n$.*
3. *$Ker(M) = \{0\}$.*
4. *$\forall y \in \mathbb{R}^n, \exists$ unique $x \in \mathbb{R}^n$ s.t. $Mx = y$.*

Proof. $1 \to 2$**:** Let $y \in \mathbb{R}^n$. $y = M(M^{-1}y)$, so $y \in \text{Im}(M)$. Thus $\text{Im}(M) = \mathbb{R}^n$, so $rank(M) = n$. $2 \to 3$**:** From Rank-Nullity, $dim(Ker(M)) = n - rank(M) = 0$.

$3 \to 4$**:** Existence: $dim(\text{Im}(M)) = n - dim(Ker(M)) = n$, so $\text{Im}(M) = \mathbb{R}^n$. Uniqueness: If $Mx = Mx'$, then $M(x - x') = 0 \implies x - x' \in Ker(M) = \{0\} \implies x = x'$. $4 \to 1$**:** Property 4 implies M is a bijection. Construct M^{-1} using the unique pre-images of the canonical basis. □

3.5 Transpose of a matrix

Let $M \in \mathbb{R}^{n \times m}$. We define its transpose $M^\top \in \mathbb{R}^{m \times n}$ by $M^\top_{i,j} = M_{j,i}$.

- $\forall A \in \mathbb{R}^{n \times m}, rank(A) = rank(A^\top)$.
- $(AB)^\top = B^\top A^\top$.
- A matrix is **symmetric** if $A = A^\top$.

4 Norms, Inner Products, and Orthogonality

4.1 Norms

Definition 4.1 (Euclidean Norm)**.** We define the Euclidean norm of $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ as:

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$$

4.1.1 General norms

For it to be a measure of length, it must satisfy 3 qualities. Let V be a vector space.

Definition 4.2 (Norm)**.** A **norm** $\|\cdot\|$ on V is a function from V to $\mathbb{R}_{\geq 0}$ that verifies:

1. **Homogeneity:** $\|\alpha v\| = |\alpha| \times \|v\| \quad \forall \alpha \in \mathbb{R}, v \in V$.
2. **Positive definiteness:** If $\|x\| = 0$ for some $x \in V$, then $x = 0$.
3. **Triangular inequality:** $\|u + v\| \leq \|u\| + \|v\| \quad \forall u, v \in V$.

Other norms:

- ℓ_1 norm (Manhattan): $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- ℓ_∞ norm: $\|x\|_\infty = \max(|x_1|, \ldots, |x_n|)$.
- ℓ_p norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$.

4.1.2 Application to regularized linear regression

Suppose this is an under-constrained (over-parameterized) solution. Thus the solution occupies an affine space $\{\theta \mid X\theta = y\}$.
Idea. How do you pick which θ if there are infinite solutions? By **Occam's Razor**, let's take the simplest one, or $\min\|\theta\|$ among solutions. To visualize this, let's suppose on a 2D graph there is a line that doesn't cross the origin. Starting from the origin, we gradually increase the norm until some point of the norm intersects with the solution set.

4.2 Inner products

Definition 4.3 (Euclidean Dot Product)**.** We define the Euclidean dot product of two vectors x and y of $\mathbb{R}^n$ as:

$$x \cdot y = \sum_{i=1}^n x_i y_i = x_1 y_1 + \cdots + x_n y_n \tag{39}$$

$$= \|x\|\|y\| \cos\theta \tag{40}$$

If they are aligned, the inner product is large; if orthogonal, it is 0.

Definition 4.4 (Inner Product)**.** An **inner product** on V is a function $\langle \cdot, \cdot \rangle$ from $V \times V$ to $\mathbb{R}$ that verifies:

1. **Symmetry:** $\langle u, v \rangle = \langle v, u \rangle \quad \forall u, v \in V$.
2. **Bilinearity:** $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$ and $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$.
3. **Positive definiteness:** $\langle v, v \rangle \geq 0$ with equality iff $v = 0$.

Proposition 4.1 (Induced norm)**.** *If $\langle \cdot, \cdot \rangle$ is an inner product on V then,*

$$\|v\| \stackrel{def}{=} \sqrt{\langle v, v \rangle}$$

is a norm on V. We say that the norm $\|\cdot\|$ is induced by the inner product.

4.2.1 Example: Random Variables

Consider the set V of all random variables (on a probability space Ω) that have a finite second moment, with the inner product:

$$\langle X, Y \rangle \stackrel{\text{def}}{=} \mathbb{E}[XY]$$

The induced norm is $\|X\| = \sqrt{\mathbb{E}[X^2]}$ (Standard deviation if zero-mean).

Theorem 4.2 (Cauchy-Schwarz inequality)**.** *If $\|\cdot\|$ is the norm induced by the inner product $\langle \cdot, \cdot \rangle$ on V, then $\forall x, y \in V$:*

$$|\langle x, y \rangle| \leq \|x\| \|y\|$$

Equality holds iff x and y are linearly dependent.

4.3 Orthogonality

Definition 4.5 (Orthogonality)**.** We say that vectors x, y are orthogonal if $\langle x, y \rangle = 0$. We write then $x \perp y$.

Definition 4.6 (Orthogonal and Orthonormal Families)**.** A family of vectors $(v_1, \ldots, v_k)$ is:

- **Orthogonal** if $\langle v_i, v_j \rangle = 0 \quad \forall i \neq j$.
- **Orthonormal** if it is orthogonal and $\|v_i\| = 1$.

Proposition 4.3. *Assume that $dim(V) = n$ and let $(v_1, \ldots, v_n)$ be an orthonormal basis of V. Then the coordinates of a vector $x \in V$ are $(\langle v_1, x \rangle, \ldots, \langle v_n, x \rangle)$:*

$$x = \langle v_1, x \rangle v_1 + \cdots + \langle v_n, x \rangle v_n$$

Theorem 4.4 (Pythagorean theorem)**.** *Let $\|\cdot\|$ be the norm induced by $\langle\cdot,\cdot\rangle$. For all $x, y \in V$:*

$$x \perp y \iff \|x+y\|^2 = \|x\|^2 + \|y\|^2$$

Proof.

$$\begin{aligned}
\|x+y\|^2 &= \langle x+y, x+y\rangle && (41)\\
&= \langle x,x\rangle + \langle y,x\rangle + \langle x,y\rangle + \langle y,y\rangle && (42)\\
&= \|x\|^2 + |y|^2 + 2\langle x,y\rangle && (43)\\
&= \|x\|^2 + |y|^2 \iff x \perp y && (44)
\end{aligned}$$

□

4.3.1 Orthogonal Projection

Let S be a subspace of $\mathbb{R}^n$. The **orthogonal projection** of x onto S is the vector $P_S(x)$ in S that minimizes the distance to x:

$$P_S(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in S} \|x-y\|$$

Proposition 4.5. *Let $(v_1,\ldots,v_k)$ be an orthonormal basis of S. Then for all $x \in \mathbb{R}^n$:*

$$P_S(x) = \langle v_1, x\rangle v_1 + \cdots + \langle v_k, x\rangle v_k$$

Let V gather the orthonormal basis vectors of S. Then $P_S(x) = VV^\top x$.

Corollary 4.6. *1. $x - P_S(x)$ is orthogonal to S.*

2. $\|P_S(x)\| \le \|x\|$.

4.3.2 Orthogonal Complement

We define the orthogonal complement of S as $S^\perp = \{x \in V \mid x \perp S\}$.

- $S^\perp$ is a subspace of V.
- $dim(S^\perp) = dim(V) - dim(S)$.

5 Orthogonal Matrices

5.1 Gram-Schmidt Algorithm

Goal. Purpose: How to create an orthonormal basis? The Gram-Schmidt process takes as:

- **Input:** A linearly independent family $(x_1, \ldots, x_k)$ of $\mathbb{R}^n$.
- **Output:** An orthonormal basis $v_1, \ldots, v_k$ of $Span(x_1, \ldots, x_k)$.

Corollary 5.1. *Every subspace of $\mathbb{R}^n$ admits an orthonormal basis.*

We need to do 2 things: normalize each vector, and remove the projection of each vector onto every other vector (project onto the complement).
Solution. For example, let there be x_1, x_2.

$$v_1 = \frac{x_1}{\|x_1\|} \tag{45}$$

$$\tilde{v_2} = x_2 - \langle x_2, v_1 \rangle v_1 \tag{46}$$

$$v_2 = \frac{\tilde{v_2}}{\|\tilde{v_2}\|} \tag{47}$$

We check orthogonality: $\langle \tilde{v_2}, v_1 \rangle = \langle x_2, v_1 \rangle - \langle x_2, v_1 \rangle \langle v_1, v_1 \rangle = 0$. $\tilde{v_2}$ cannot be 0 because it is 0 iff $v_2 \not\perp v_1$ (which implies dependence).

Idea. Using this, we can find $A = QR$ where Q is an orthogonal matrix and R is an upper triangular matrix.

5.1.1 Gram-Schmidt construction

The Gram-Schmidt process is a recursive algorithm that constructs $v_1, \ldots, v_k$ such that for all $i \in \{1, \ldots, k\}$:

$$Span(v_1, \ldots, v_i) = Span(x_1, \ldots, x_i)$$

where $(v_1, \ldots, v_i)$ is an orthogonal family.

5.2 Orthogonal Matrices

Definition 5.1 (Orthogonal Matrix)**.** A matrix $A \in \mathbb{R}^{n \times n}$ is called an **orthogonal matrix** if its columns are an orthonormal family.

Proposition 5.2. *Let $A \in \mathbb{R}^{n \times n}$. The following points are equivalent:*

1. *A is orthogonal.*
2. $A^\top A = Id_n$.
3. $AA^\top = Id_n$.

5.2.1 Orthogonal matrices and norm

Proposition 5.3 (Preservation of Norms)**.** *Let $A \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Then A preserves the dot product:*

$$\forall x, y \in \mathbb{R}^n, \quad \langle Ax, Ay\rangle = \langle x, y\rangle$$

In particular if we take $x = y$ we see that A preserves the Euclidean norm: $\|Ax\| = \|x\|$.

Proof.

$$\begin{aligned}\langle Ax, Ay\rangle &= (Ax)^\top (Ay) &(48)\\ &= x^\top A^\top Ay &(49)\\ &= x^\top y &(50)\\ &= \langle x, y\rangle &(51)\end{aligned}$$

□

Example: Rotations and Reflections

Let R_θ be a rotation by angle θ:

$$R_\theta = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$$

Let S be a reflection:

$$S = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

These are the only types of orthogonal matrices in $\mathbb{R}^2$.

5.3 Orthonormal Bases

Let $(a_1, \ldots, a_n)$ be an orthonormal basis of $\mathbb{R}^n$, and A the matrix collecting these vectors. Consider x where coordinates are in the canonical basis.

Proposition 5.4 (Change of Basis)**.** *The coordinates of x in the $(a_1, \ldots, a_n)$ basis are given by $x' = A^\top x$.*

Idea. Usually finding the orthonormal bases requires solving a linear system (finding the inverse). The general formula is $x' = A^{-1}x$, but $A^{-1} = A^\top \iff A$ is an orthogonal matrix.

5.4 Eigenvalues and Eigenvectors

Definition 5.2 (Eigenvector and Eigenvalue)**.** Let $A \in \mathbb{R}^{n \times n}$. A non-zero vector $v \in \mathbb{R}^n$ is said to be an **eigenvector** of A if $\exists \lambda \in \mathbb{R}$ such that $Av = \lambda v$. The scalar λ is then called an **eigenvalue** of A.

Examples

1. **Identity matrix:** Any non-zero vector v is an eigenvector of Id with $\lambda = 1$.
2. **Matrix with Kernel:** The eigenvectors for eigenvalue 0 are $Ker(A) \setminus \{0\}$. If 0 is an eigenvalue, A is not invertible.

5.4.1 Orthogonal Projection

Let $P_S(x)$ be an orthogonal projection onto subspace S.

- If $x \in S \setminus \{0\}$, x is an eigenvector with eigenvalue 1 ($P_S(s) = x$).
- If $x \in S^{\perp} \setminus \{0\}$, x is an eigenvector with eigenvalue 0 ($P_S(x) = 0$).

6 Eigenvalues and Markov Chains

6.1 Eigenvalues and Eigenvectors

Definition 6.1 (Eigenvalues and Eigenvectors)**.** Let $A \in \mathbb{R}^{n\times n}$. A non-zero vector $v \in \mathbb{R}^n$ is an eigenvector if $\exists \lambda \in \mathbb{R}$ s.t. $Av = \lambda v$.

Proposition 6.1 (Eigenvalue Properties)**.** *1. $\alpha\lambda$ is an eigenvalue of αA.*

2. $\lambda + \alpha$ is an eigenvalue of $A + \alpha Id$.

3. λ^k is an eigenvalue of A^k.

4. If A is invertible, $\frac{1}{\lambda}$ is an eigenvalue of A^{-1}.

Definition 6.2 (Eigenspace)**.** The eigenspace $E_\lambda(A) = Ker(A - \lambda Id)$ is the set of all eigenvectors for λ (plus the zero vector). The dimension of $E_\lambda(A)$ is the **multiplicity** of λ.

Definition 6.3 (Spectrum)**.** The set of all eigenvalues is the **spectrum** $Sp(A)$.

Proposition 6.2 (Bounds on Spectrum)**.** *A $n \times n$ matrix A admits at most n distinct eigenvalues. If $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues with multiplicities m_i, then $\sum m_i \leq n$.*

6.2 Markov Chains

Whenever you have a graph of nodes and edges, you can encode it as an adjacency matrix.

Definition 6.4 (Stochastic Matrix)**.** A matrix $P \in \mathbb{R}^{n\times n}$ is stochastic if:

1. $P_{i,j} \geq 0 \quad \forall i, j$.
2. $\sum_{i=1}^{n} P_{i,j} = 1 \quad \forall j$ (Columns sum to 1).

Definition 6.5 (Probability Vector)**.** $x_t \in \mathbb{R}^n$ is a probability vector if entries are non-negative and sum to 1.

Proposition 6.3 (The Key Equation)**.** $x^{t+1} = Px^t \implies x^t = P^t x^0$.

Definition 6.6 (Invariant Measure)**.** We define μ as an **invariant measure** (or stationary distribution) if $\mu = P\mu$. This means μ is an eigenvector of P associated with eigenvalue 1.

6.2.1 Perron-Frobenius Theorem

Theorem 6.4 (Perron-Frobenius)**.** *Let P be a stochastic matrix such that there exists $k \geq 1$ where all entries of P^k are strictly positive (Regular/Ergodic). Then:*

1. *1 is an eigenvalue of P with a unique eigenvector $\mu \in \nabla_n$.*
2. *The eigenvalue 1 has multiplicity 1.*
3. *All other eigenvalues satisfy $|\lambda| < 1$.*
4. *For any initial x^0, $P^t x^0 \to \mu$ as $t \to \infty$.*

6.3 PageRank: Ordering the Web

Goal. Find interesting pages. **Idea:** Interesting pages have links from other interesting pages.

Idea. Random Surfer: Suppose someone clicks on a random link every time. The page they spend the most time on is likely the most interesting.

This defines a Markov chain. The "importance" of a webpage is its value in the invariant measure μ. To ensure convergence (satisfy Perron-Frobenius), we introduce a "teleportation" probability α (Random Spectator model), where the surfer sometimes picks a completely random page instead of following a link.

7 Spectral Theorem and PCA

7.1 Spectral Theorem

Theorem 7.1 (Spectral Theorem)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a **symmetric** matrix. Then there is an **orthonormal basis** of $\mathbb{R}^n$ composed of eigenvectors of A.*

That means if A is symmetric, then there exists an orthonormal basis $(v_1, \ldots, v_n)$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that:

$$Av_i = \lambda_i v_i \quad \forall i \in \{1, \ldots, n\} \tag{52}$$

$$AV = V\Lambda \qquad (\Lambda = diag(\lambda_1, \ldots, \lambda_n))$$

$$A = V\Lambda V^\top \qquad (\text{Since } V^\top = V^{-1})$$

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top \tag{53}$$

Geometrically, $\lambda_i v_i v_i^\top$ is the orthogonal projection onto v_i scaled by λ_i.

Theorem 7.2 (Matrix Formulation)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a **symmetric** matrix. Then there exists an orthogonal matrix P and a diagonal matrix D such that $A = PDP^\top$.*

7.1.1 The spectral orthonormal basis

Why are eigenvectors orthogonal when A is symmetric? Let v, w be eigenvectors of A for eigenvalues $\lambda \neq \mu$.

$$v^\top Aw = v^\top(\mu w) = \mu v^\top w \tag{54}$$

$$= v^\top A^\top w = (Av)^\top w = \lambda v^\top w \tag{55}$$

Therefore $(\mu - \lambda)v^\top w = 0$. Since $\mu \neq \lambda$, $v^\top w$ must be 0.

7.1.2 Geometric Interpretation

Ax is equivalent to the orthogonal projection of x onto the eigenvector basis, scaled by its corresponding eigenvalue.

$$Ax = A\left(\sum_{i=1}^{n} \langle x, v_i\rangle v_i\right) = \sum_{i=1}^{n} \lambda_i \langle x, v_i\rangle v_i \tag{56}$$

Operations (Right to Left in $V\Lambda V^\top$):

1. Decompose into eigenvector basis ($V^\top$).
2. Scale by eigenvalues (Λ).
3. Transform back to standard basis (V).

7.1.3 Consequences

If $A = PDP^\top$ for orthogonal P:

1. The rank of A equals the number of non-zero λ_i.
2. A is invertible iff $\lambda_i \neq 0$ for all i. Then $A^{-1} = P\Lambda^{-1}P^\top$.
3. $Tr(A) = \sum \lambda_i$.

7.2 Principal Component Analysis (PCA)

7.2.1 Empirical mean and covariance

Consider a dataset $a_1, \ldots, a_n \in \mathbb{R}^d$.

- **Mean:** $\mu = \frac{1}{n}\sum_{i=1}^n a_i$.
- **Covariance Matrix:** $\Sigma = \frac{1}{n}\sum_{i=1}^n (a_i - \mu)(a_i - \mu)^\top$.

It describes the variance in all possible unit directions.

7.2.2 PCA Algorithm

Goal. Find a lower dimensional representation $\tilde{a}_1, \ldots \tilde{a}_n \in \mathbb{R}^k$ where $k \ll d$.

Assume centered data ($\mu = 0$). Then $\Sigma \propto \sum a_i a_i^\top = A^\top A$, which is symmetric. We can apply the Spectral Theorem.

Idea. Direction of maximal variance: We aim to find u where variance is maximal. This corresponds to the eigenvector associated with the largest eigenvalue of the covariance matrix.

The j-th direction of maximal variance is v_j, the eigenvector corresponding to the j-th largest eigenvalue. The dimensionally reduced dataset is the projection onto these first k eigenvectors.

7.2.3 Choosing k

1. **Elbow rule:** Eyeballing where there is a sharp decrease in explained variance.
2. **Percentage threshold:** Keep enough components to explain $x\%$ of total variance.

8 SVD and Linear Algebra for Graphs

8.1 Singular Value Decomposition (SVD)

> "This is the single most powerful matrix decomposition, and once you know it, well, there's a before and after." - Florentein Guth

Goal. Motivation: How do we generalize the spectral theorem decomposition to generic matrices $A \in \mathbb{R}^{n\times m}$? Since A is now rectangular, $Av_i = \lambda_i v_i$ no longer has any meaning since the inner spaces no longer match up between A, v_i.

Idea. Let $Au_i = \sigma_i v_i$ for another orthonormal family $v_1, \dots, v_m$. So, we know A is an operation from $\mathbb{R}^m \to \mathbb{R}^n$ such that $Av_i = \sigma_i u_i$.

$$Ax = A\left(\sum_{i=1}^{m} \langle x, v_i\rangle v_i\right) = \sum_{i=1}^{m} \langle x, v_i\rangle Av_i \tag{57}$$

$$= \sum_{i=1}^{m} \sigma_i \langle x, v_i\rangle u_i = \sum_{i=1}^{m} \sigma_i (v_i^\top x) u_i \tag{58}$$

$$= \left(\sum_{i=1}^{m} \sigma_i u_i v_i^\top\right) x \tag{59}$$

Therefore, $A = \sum_{i=1}^{m} \sigma_i u_i v_i^\top$.

Steps for linear combination:

1. Decompose into eigenvector basis (rotate): $(\langle x, v_i\rangle)_{i=1:m}$
2. Scale by eigenvalues: $(\sigma_i \langle x, v_i\rangle)_{i=1:m}$
3. Transform output space: $(\sum_{i=1}^{m} \sigma_i \langle x_i, v_i\rangle u_i)$

Theorem 8.1 (Singular Value Decomposition)**.** *Let $A \in \mathbb{R}^{n\times m}$. Then there exists two orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ and a matrix $\Sigma \in \mathbb{R}^{n\times m}$ such that $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq 0$ and $\Sigma_{i,j} = 0$ for $i \neq j$ that verify $A = U\Sigma V^\top$.*

Notice that Σ is a rectangular matrix. Since A is rectangular, it isn't invertible and has a non-trivial kernel. The zero columns correspond to directions in $Ker(A)$ (Rank-Nullity theorem).

8.1.1 Proof Sketch

1. Establish orthogonality of Av_i: $\langle Av_i, Av_j\rangle = \sigma_i^2$ if $i = j$, else 0.
2. Show v_i are eigenvectors of $A^\top A$: $A^\top A v_1 = \lambda_1 v_1$.

3. Apply spectral theorem to symmetric $A^\top A$ to get orthonormal basis $\{v_i\}$.

4. Define left singular vectors $u_i = \frac{Av_i}{\|Av_i\|}$ and singular values $\sigma_i = \|Av_i\|$.

Remark 8.1.
1. Right singular vectors v_i are eigenvectors of $A^\top A \in \mathbb{R}^{m\times m}$, with eigenvalues $\lambda_i = \sigma_i^2$.

2. Left singular vectors u_i are eigenvectors of $AA^\top \in \mathbb{R}^{n\times n}$, with eigenvalues $\lambda_i = \sigma_i^2$.

3. If A is symmetric, then $u_i = v_i$.

8.1.2 Non-negative singular values and sign instability

Singular values $\sigma_i = \|Av_i\|$ are defined by norms and measure scaling, thus are non-negative. Since $u_i = \frac{Av_i}{\sigma_i}$, flipping the sign of v_i flips the sign of u_i. Thus (v_i, u_i) and $(-v_i, -u_i)$ are both valid pairs.

8.2 Graphs and Graph Laplacian

Note that we are assuming *undirected graphs*.

Definition 8.1 (Adjacency Matrix)**.** The adjacency matrix A of graph G is the $n \times n$ matrix with entries:

$$A_{ij} = \begin{cases} 1 & \text{if edges between nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

Definition 8.2 (Degree Matrix)**.** The degree matrix $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix with $D_{i,i} = deg(i)$.

Definition 8.3 (Graph Laplacian)**.** The Laplacian matrix of G is defined as $L = D - A$.

Proposition 8.2. *For the graph Laplacian L:*

$$\forall x \in \mathbb{R}^n, \quad x^\top Lx = \sum_{i\sim j} (x_i - x_j)^2$$

This measures smoothness w.r.t. the graph (nodes linked by an edge are similar).

8.2.1 Properties of the Laplacian

1. **Symmetric:** Since D and A are symmetric.

2. **Positive Semi-Definite (PSD):** $x^\top Lx \geq 0$, so $\lambda_i \geq 0$.

3. **Non-invertible:** L always has eigenvalue 0 with eigenvector $v_1 = \frac{1}{\sqrt{n}}(1, \ldots, 1)$ (constant vector). Thus $Ker(L)$ is not empty.

Theorem 8.3 (Algebraic Connectivity)**.** *The multiplicity of the eigenvalue 0 of L is equal to the number of connected components of G. G is connected $\iff$ $\lambda_2 > 0$. λ_2 is known as the* ***algebraic connectivity****.*

8.3 Application: Spectral Clustering

8.3.1 Two clusters: Minimal cut problem

Definition 8.4 (Cut)**.** The cut of $S \subset \{1, \ldots, n\}$, denoted $cut(S)$, is the number of edges between S and S^C.

Goal. Find 2 clusters (S, S^C) such they are balanced and minimize edges across.

Using sign vectors $x \in \{+1, -1\}^n$, we find $cut(S) = \frac{1}{4}x^\top Lx$. Minimizing $x^\top Lx$ subject to $x \in \{-1, 1\}^n$ and $x \perp 1$ is NP-hard.
Idea. Relax the constraints to $v \in \mathbb{R}^n$ with $\|v\|^2 = n$. This becomes the problem of finding the eigenvector for the second smallest eigenvalue, v_2.

8.3.2 Spectral Clustering Algorithm (2 Clusters)

1. Compute v_2, the second eigenvector of L.
2. Set $x_i = v_2(i)$.
3. Define $S = \{i \mid v_2(i) \geq 0\}$ and $S^C = \{i \mid v_2(i) < 0\}$.

8.3.3 Spectral Clustering Algorithm (k Clusters)

1. Compute $v_1, \ldots, v_k$ of L.
2. Embed nodes as $x_i = (v_2(i), \ldots, v_k(i))$.
3. Cluster $x_1, \ldots, x_n$ using k-means.

9 Convexity

Goal. Motivation: In machine learning, we often minimize functions $f(\theta) = Loss(data, model_\theta)$. For higher dimensions, we rely on local searches (derivatives). Local information gives global information when the function is *convex*.

9.1 Functions of n Variables

9.1.1 Single Variable ($n = 1$)

We define the tangent using the derivative. **First-Order Taylor Approximation:**

$$f(x_0 + h) = f(x_0) + h \cdot f'(x_0) + o(h)$$

9.1.2 General Case (n Variables)

$f : \mathbb{R}^n \to \mathbb{R}$. We define the gradient:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)$$

Taylor approximation:

$$f(x + h) \approx f(x) + \langle \nabla f(x), h \rangle$$

Definition 9.1 (Jacobian Matrix)**.** For vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$, the **Jacobian J**$(x)$ is an $m \times n$ matrix where the i-th row is $\nabla f_i(x)$.

Definition 9.2 (Hessian Matrix)**.** For scalar-valued functions $f : \mathbb{R}^n \to \mathbb{R}$, the **Hessian** is the matrix of second derivatives:

$$H(x) = \nabla^2 f(x) \in \mathbb{R}^{n \times n} \quad \text{where} \quad H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Theorem 9.1 (Schwarz's Theorem)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable with continuous second partial derivatives, then the Hessian is symmetric:*

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

9.1.3 Taylor's Formula (Second Order)

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x) h + o(\|h\|^2)$$

9.2 Convexity Definitions

9.2.1 Convex Sets

Definition 9.3 (Convex Set)**.** A set $S \subset \mathbb{R}^n$ is **convex** if $\forall x, y \in S, \forall \alpha \in [0,1]$,

$$\alpha x + (1-\alpha)y \in S$$

Example 9.1. Subspaces of $\mathbb{R}^n$ and Norm balls $B(r) = \{x \mid \|x\| \leq r\}$ are convex sets.

9.2.2 Convex Functions

Definition 9.4 (Convex Function)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if $\forall x, y \in \mathbb{R}^n, \forall \alpha \in [0,1]$:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

Intuitively, the function always curves upwards. f is concave if $-f$ is convex.

Proposition 9.2. *1. Linear maps are both convex and concave.*

2. Norms are convex (Triangle Inequality).

3. Sum of convex functions is convex.

9.3 Convexity and Derivatives

Proposition 9.3 (Tangents)**.** *A differentiable function f is convex $\iff \forall x, y \in \mathbb{R}^n$:*

$$f(y) \geq f(x) + \langle \nabla f(x), (y-x) \rangle$$

(The function is always above its tangent).

Corollary 9.4. *If f is convex and differentiable: x is a global minimizer $\iff \nabla f(x) = 0$.*

Proposition 9.5 (Hessian Condition)**.** *Let f be twice-differentiable. Then f is convex $\iff \nabla^2 f(x)$ is Positive Semi-Definite (PSD) for all x.*

9.4 Jensen's Inequality

Theorem 9.6 (Jensen's Inequality)**.** *Let f be a convex function. For $x_i \in \mathbb{R}^n$ and weights $\alpha_i \geq 0$ summing to 1:*

$$f\left(\sum \alpha_i x_i\right) \leq \sum \alpha_i f(x_i)$$

For a random variable X:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Example 9.2 (Entropy)**.** For a random variable X, entropy is $H(X) = \sum p_i \log(1/p_i)$. Since log is concave, Jensen's inequality applies in reverse, showing $H(X) \leq \log(k)$ (Uniform distribution maximizes entropy).

10 Linear Regression

10.1 Introduction

Goal. Given feature vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ and outputs $y_1, \ldots, y_n \in \mathbb{R}$, predict y for a new input $a \in \mathbb{R}^d$.

Subgoal. Find $x \in \mathbb{R}^d$ such that $y_i = \langle x, a_i \rangle + b$.

We can include the additive constant b by making the first coordinate 1:

$$a_i \to \tilde{a}_i = (a_i, 1) \in \mathbb{R}^{d+1} \tag{60}$$

$$\tilde{x} \in \mathbb{R}^{d+1} \tag{61}$$

$$\langle \tilde{x}, \tilde{a}_i \rangle = \sum_{j=1}^{d} (\tilde{x}_j \cdot \tilde{a}_{i,j}) + \tilde{x}_{d+1} \tag{62}$$

$$= \langle x, a \rangle + b \tag{63}$$

Subgoal. Solve for x in $y = Ax$ where A is the feature matrix (including bias) and x is the parameters.

We assume that A is **full rank:** $rank(A) = \min(n, d)$. 3 things could happen:

- $n = d$ (square system): single solution $x = A^{-1}y$.
- $n > d$ (overdetermined system): typically no solution. Since $rank(A) = d < n$, $dim(Im(A)) = d < n$. So the image of A is a d-dimensional subspace of $\mathbb{R}^{n \times d}$, so y is typically not in $Im(A)$ because there is noise. The challenge here is to find an **approximate solution**.
- $n < d$ (underdetermined system): infinitely many solutions. By the Rank-Nullity theorem, $dim(Ker(A)) = d - n > 0$. The challenge is to find the **best solution among many**.

10.2 Ordinary least squares

10.2.1 Overdetermined case ($n > d$)

In the overdetermined case, how do you choose the approximate solution if there is no solution? The usual way is to minimize the error. This translates to minimizing some norm. Typically, we minimize the square norm because it is a smooth, differentiable, convex function and mathematically equivalent to minimizing the norm itself.
Goal. (OLS) $\min_x f(x) = \|Ax - y\|^2$ w.r.t. $x \in \mathbb{R}^d$.

f is convex $(\nabla^2 f(x)) = 2A^\top A$, which is PSD. Therefore we can find the global minima by solving for x where the gradient is 0.

$$\text{x minimizes } f \iff \nabla f(x) = 0 \tag{64}$$
$$\iff 2A^\top(Ax - y) = 0 \tag{65}$$
$$\iff A^\top Ax = A^\top y \tag{66}$$

Since $n > d$, and $A^\top A \in \mathbb{R}^{d \times d}$, $A^\top A$ is full rank, and thus has an inverse. Thus:
Subgoal. Solve closed form solution $x = (A^\top A)^{-1} A^\top Ay$ (when $n > d$).

10.2.2 Underdetermined case ($n < d$)

In the underdetermined case, there are infinitely many solutions. How then do we pick the "best solution"? Note that this is an empirical solution since we're grappling with philosophical approximations here. Good ol' Occam's Razor.
Idea. Pick the most "stable solution". Stable meaning the solution changes the least while undergoing some perturbation. Among all x s.t. $Ax = y$, pick the least normed one. All solutions lie on an affine subspace (a line in 2D, a hyperplane in general). The minimum norm solution x^* is the point on this subspace **closest to the origin**, found by dropping a perpendicular from the origin to the solution set.

Let δa denote some perturbation of a.

$$y_{pred} = \langle a + \delta a, x \rangle = \langle a, x \rangle + \langle \delta a, x \rangle \tag{67}$$
$$\|y_{pred} - \langle a, x \rangle\| = |\langle \delta a, x \rangle| \leq \|\delta a\| \cdot \|x\| \qquad \text{(Cauchy-Schwarz)} \tag{68}$$

We see that the error is bounded by the Cauchy Schwarz inequality. Thus the way to minimize errors under perturbation is to minimize $\|x\|$. **In general,** The solution set $\{x : Ax = y\}$ forms an affine subspace (hyperplane shifted away from origin). The minimum norm solution x^* is found by projecting the origin **orthogonally** onto this hyperplane.
Goal. (Underdetermined OLS) Minimize $\|x\|^2$ subject to $\|Ax - y\|^2$ being minimum.

Solution. We can use the solution from the previous section, $A^\top Ax = A^\top y$, except now $A^\top A$ is no longer invertible.

$$\text{Let } A = USV^\top \tag{SVD}$$

$$A^\top Ax = A^\top y \to (USV^\top)^\top (USV^\top)x = (USV^\top)^\top y \tag{69}$$

$$VS^\top U^\top USV^\top x = VS^\top U^\top y \qquad \text{(U is orthogonal)}$$

$$S^2 V^\top x = SU^\top y \tag{70}$$

$$\text{so each individual component: } \sigma_i^2 \langle v_i, x\rangle = \sigma_i \langle u_i, y\rangle \tag{71}$$

$$\langle v_i, x\rangle = \begin{cases} \frac{1}{\sigma_i}\langle u_i, y\rangle & \text{if } \sigma_i > 0 \\ \text{free} & \text{if } \sigma_i = 0 \end{cases} \tag{72}$$

Since $\|x\|^2 = \sum \langle x, v_i\rangle^2$, using the most "stable" approach, we minimize the norm by setting $\langle x, v_i\rangle$ to 0 if the singular value is zero. Since the system is underdetermined, A cannot have full column rank. In its SVD $A = USV^\top$, the diagonal matrix S therefore has some zeros on the diagonal and is not invertible. To derive a closed form solution, we introduce **the pseudoinverse** $S^\dagger$ to get $x = VS^\dagger U^\top y$.

Definition 10.1 (Moore-Penrose Pseudoinverse)**.** Let $A = USV^\top$ be the SVD of A. The matrix $A^\dagger \stackrel{\text{def}}{=} VS^\dagger U^\top$ is called the **(Moore-Penrose) pseudoinverse** of A, where $S^\dagger \in \mathbb{R}^{d\times n}$ is defined as

$$S^\dagger_{i,i} = \begin{cases} \frac{1}{S_{i,i}} & \text{if } S_{i,i} \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad S^\dagger_{i,j} = 0 \text{ for } i \neq j \tag{73}$$

The pseudoinverse has the following properties:

- If A is invertible, then $A^\dagger = A^{-1}$.
- If $A^\top A$ is invertible, then $A^\dagger = (A^\top A)^{-1}A^\top$.
- Generally, since this is a pseudoinverse, $AA^\dagger \neq I$.
- $AA^\dagger A = A$ and $A^\dagger AA^\dagger = A^\dagger$.

Subgoal. Therefore, we can write the general solution of linear regression as solving for $x = A^\dagger y$.

10.3 Penalized linear regression

In general, there is a tradeoff between fitting the data well ($\|Ax - y\|^2$ small) and having a stable solution ($\|x\|^2$ small). **Ridge regression** combines these 2 objectives.

10.3.1 Ridge

Goal. (Ridge) $\min_x f(x) = \|Ax - y\|^r + \lambda\|x\|^2$ w.r.t. $x \in \mathbb{R}^d$ for some fixed $\lambda > 0$.

We can scale λ proportionally with how much we care about stability. As a result of this penalty term, $\|Ax - y\|^2$ is almost never zero. When $\lambda \to 0$, we recover ordinary least squares.
Solution. This can be solved in closed form by $x = (A^\top A + \lambda I)^{-1} A^\top y$. $(A^\top A + \lambda I)$ is always invertible because $A^\top A$ is positive semidefinite and λI is positive definite. The sum of a positive semidefinite matrix and a positive definite matrix is always positive definite, and positive definite matrices are always invertible. This also means that $(A^\top A + \lambda I)^{-1} A^\top \to_{\lambda \to 0} A^\dagger$.

10.3.2 Lasso

Beyond stability, sparsity is also a property to be tuned. Sparsity means having many coordinates of x equal to zero, only selecting a few features for the prediction. The primary difference of regularizations is the norm to penalize with.
Goal. (Sparse) $\min_x f(x) = \|Ax - y\|^2 + \lambda\|x\|_0$ w.r.t. $x \in \mathbb{R}^d$ where $\|x\|_0$ is a pseudonorm that is the number of non-zero coordinates of x. In practice, this is a **NP-hard** problem.

Goal. (Lasso) $\min_x f(x) = \|Ax - y\|^2 + \lambda\|x\|_1$ w.r.t. $x \in \mathbb{R}^d$.
This is a non-smooth optimization problem ($L1$ is absolute value) that does not have a closed form solution and must be solved numerically. It's primary difference from ridge is that it sometimes gives you 0 for features whereas Ridge never reaches 0. It is called Lasso in linear regression context, but it is also known as sparse coding, or compressed sensing.

10.3.3 Soft vs hard regularization

In practice, there are 2 paths to go about regularization. You can either minimize OLS with some penalizing term (soft regularization), or you can minimize OLS subject to some determined norm c of x (hard regularization). It turns out that these 2 forms of regularization are equivalent for some mapping $c \leftrightarrow \lambda$. For implementation, usually we use the soft regularization because it is simpler computationally.

10.4 Matrix norms

10.4.1 Motivation: matrix completion

The Netflix challenge: Given n users and m movies. Each user has only rated a few movies. Can we predict how a user would rate an unseen movie? Formally, we have matrix $A \in \mathbb{R}^{n \times m}$ and we only observe entries $A_{i,j}$ for $(i,j) \in \Omega$. Can we recover the full matrix A?

The winners of the challenge represented each user ($v_i \in \mathbb{R}^d$) and movie ($u_i \in \mathbb{R}^d$) with a vector embedding. The hope is that similar movies would have be represented with similar vectors (small angle). Likewise for users. Furthermore, the model A_{ij} would return high values for inner product $v_i^\top v_j$ if you would likely rate high.

$$A = \begin{bmatrix} — & u_1^\top & — \\ & \vdots & \\ — & u_n^\top & — \end{bmatrix} \cdot \begin{bmatrix} | & & | \\ v_1 & \dots & v_m \\ | & & | \end{bmatrix} = U^\top V \text{ where d small} \tag{74}$$

d represents the latent state of the user, and we make d as low as we can. If d is low, then it means that A is a low rank matrix.
Goal. (Low Rank Approximation) $\min_B rank(B)$ s.t. $B_{ij} = A_{ij} \forall i,j \in \Omega$.
This is NP-hard!

10.4.2 Schatten norms

The rank of A is the ℓ^0 norm of its singular values (number of eigenvalues).

$$\sigma(A) = (\sigma_1(A), \dots, \sigma_{min(n,m)A}) \tag{75}$$

This suggests the following family of norms:

Definition 10.2 (Schatten p-norm)**.** $\|A\|_p = \|\sigma(A)\|_p = \begin{cases} (\Sigma_i \sigma_i(A)^p)^{\frac{1}{p}} & \text{if } 1 \le p < \infty \\ \max_i \sigma_i(A) & \text{if } p = \infty \end{cases}$

This norm describes the size of the singular values, which describes how much eigenvectors shrink or stretch. When $p = 1$, all singular values contribute equally. However, as $p \to \infty$, the norm approaches σ_1, the largest singular value. Proof of valid norm requires homogeneity, non-negativity and triangular inequality. It is trivial to show homogeneity and non-negativity, however, it is difficult to prove $\|A+B\|_p \le \|A\|_p + \|B\|_p$ because $\sigma_i(A+B) \ne \sigma_i(A) + \sigma_i(B)$.

While $p \in [1, \infty)$, we discuss $p = 1, 2, \infty$ since it has some nice properties.

Proposition 10.1 (Schatten 2-norm = Frobenius norm)**.** $\|A\|_2 = \sqrt{\sum_{i=1}^{min(n,m)} \sigma_i(A)^2}$
$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m A_{i,j}^2} = \sqrt{Tr(A^\top A)}$

Proof.

$$
\begin{aligned}
A &= U\Sigma V^\top && \text{(SVD decomposition)} \\
\|A\|_F^2 = Tr(A^\top A) &= Tr((U\Sigma V^\top)^\top (U\Sigma V^\top)) \\
&= Tr(V\Sigma^\top U^\top U\Sigma V^\top) \\
&= Tr(V^\top V\Sigma^\top U^\top U\Sigma) && \text{(Tr invariance to cyclical ops)} \\
&= Tr(\Sigma^\top \Sigma) \\
&= \sum_{i=1}^{n} \sigma_i^2
\end{aligned}
$$

□

Proposition 10.2 (Schatten ∞-norm = Operator norm)**.** $\|A\|_\infty = \max_{1\le i\le n} \sigma_i(A)$
$\|A\|_{op} = \max_{\|x\|=1} \|Ax\|$

Proof.

($\geq$) Operator norm is at least Schatten ∞-norm:

$$
\begin{aligned}
\text{Take } x &= v_1 \text{ (right singular vector for } \sigma_1) \\
\|Ax\| &= \|Av_1\| = \|\sigma_1 u_1\| = \sigma_1 \\
\text{Since } \|v_1\| = 1, \text{ we have } &\|A\|_{op} \geq \sigma_1 = \|A\|_\infty
\end{aligned}
$$

We can also write it as: $\|A\|_{op} = \max_{\|x\|=1} \sqrt{x^\top A^\top A x} = \sqrt{\max_{\|x\|=1} x^\top A^\top A x} = \sqrt{\lambda_{max}(A^\top A)}$

($\leq$) Operator norm is at most Schatten ∞-norm:

$$
\begin{aligned}
\text{For any } x \text{ with } \|x\| = 1: & \\
\|Ax\| &= \|\sum_i \sigma_i (u_i v_i^T) x\| = \|\sum_i \sigma_i \langle v_i, x\rangle u_i\| \\
&\leq \sum_i \sigma_i |\langle v_i, x\rangle| \leq \sigma_1 \sum_i |\langle v_i, x\rangle| \leq \sigma_1 = \|A\|_\infty
\end{aligned}
$$

□

Definition 10.3 (Schatten 1-norm / Nuclear norm)**.** The Schatten 1-norm, also called the **nuclear norm**, is defined as:

$$\|A\|_1 = \|A\|_* = \sum_{i=1}^{\min(n,m)} \sigma_i(A)$$

This is the norm used in the Netflix problem! Reminder that for sparse linear regression, ℓ^0 is NP-hard, so we relax it to ℓ^1, which is a convex problem. min-rank = "Schatten-0 norm" is a NP-hard problem, which becomes the nuclear norm. The Netflix solution is $\min\|B\|_*$ s.t. $B_{i,j} = A_{i,j} \in \Omega$.

11 Optimality Conditions

If you have a convex problem, we have the global minimizer iff this point is a critical point. But what about non-convex problems?

11.1 Critical points (Unconstrained optimization)

Def. 11.1 (differentiable function)
Let f :

11.2 Constrained optimization and Lagrange multipliers

11.3 Convex constrained optimization