DS-GA 1014

Optimization and Computational Linear Algebra

Andrew Liao

December 15, 2025

Abstract

This is a graduate level, proof-based linear algebra course taught by Florentin Guth at NYU Center for Data Science. It covers the basics of optimization and computational linear algebra in Data Science applications, with 66% of content being linear algebra and 33% being optimization. This course was taken in Fall of 2025, with the date in title page indicating date last updated.

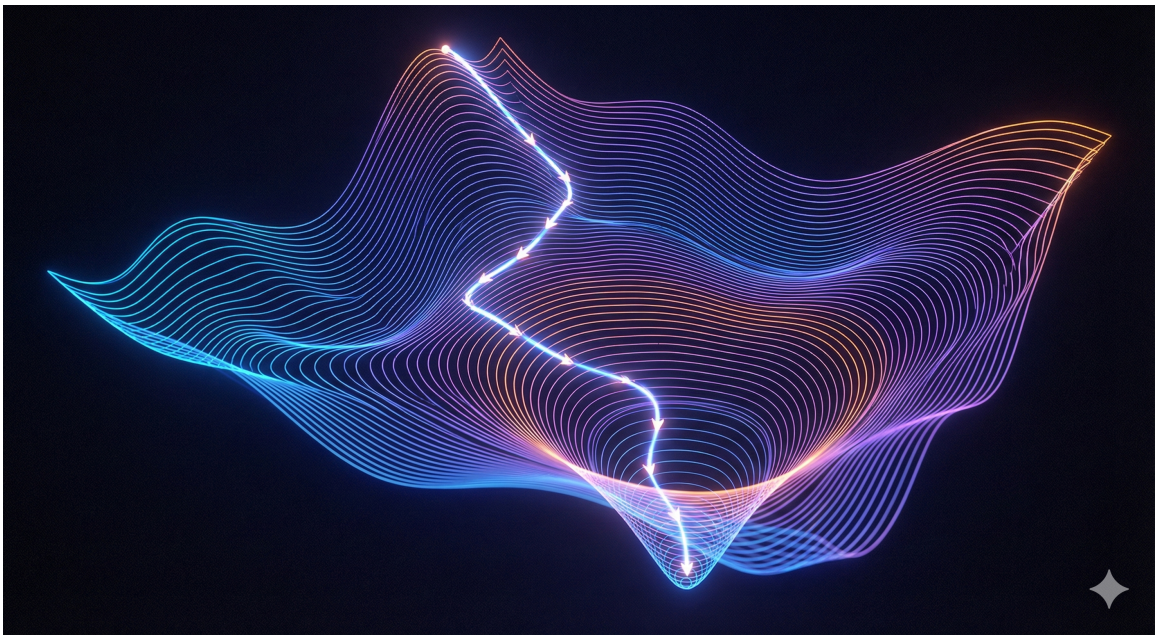

Image generated with Nano Banana Pro with prompt: "cool gradient descent figure".

Contents

Abstract 2

1 Vector Spaces 4

2 Linear Transformations and Matrices 7

3 Matrix Rank 11

4 Norms, Inner Products, and Orthogonality 13

5 Orthogonal Matrices 16

6 Eigenvalues and Markov Chains 19

7 Spectral Theorem and PCA 21

8 SVD and Linear Algebra for Graphs 23

9 Convexity 25

10 Linear Regression 27

11 Optimality Conditions 33

12 Gradient Descent 37

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.

Note that if S is a subspace of V, S is a vector space as well.

Examples

1. $\mathbb{R}^n$ is a subspace of $\mathbb{R}^n$.
2. $\{0\}$ is a subspace of $\mathbb{R}^n$.
3. Any line V going through the origin is a subspace of $\mathbb{R}^2$.
4. Is $\mathbb{R}^2$ a subspace of $\mathbb{R}^3$? No, because the operations are not defined (vectors have different numbers of components). However, the set of vectors $\{(x, y, 0) \mid x, y \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^3$.

1.2 Span and Linear Families

Definition 1.4 (Linear Independence)**.** A family of vectors $x_1, \ldots, x_k$ is **linearly independent** if the only linear combination that equals the zero vector is the trivial solution (all scalars are zero).

$$\sum_{i=1}^{k} \alpha_i x_i = 0 \implies \alpha_1 = \cdots = \alpha_k = 0$$

Examples

1. Are the vectors $e_1 = (1, 0, 0), e_2 = (0, 1, 0), e_3 = (0, 0, 1)$ linearly independent? Let $\alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3 = 0$. $(\alpha_1, \alpha_2, \alpha_3) = (0, 0, 0)$. Thus, $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Yes.

2. Is $(x, -x)$ linearly dependent? Let $\alpha_1 x + \alpha_2(-x) = 0 \implies (\alpha_1 - \alpha_2)x = 0$. If $x \neq 0$, then $\alpha_1 = \alpha_2$. We can choose $\alpha_1 = 1, \alpha_2 = 1$. Since scalars are non-zero, they are linearly dependent.

Definition 1.5 (Span)**.** The **span** of a set of vectors is the set of all possible linear combinations of those vectors.

$$Span(x_1, \ldots, x_k) = \{\alpha_1 x_1 + \cdots + \alpha_k x_k \mid \alpha_i \in \mathbb{R}\}$$

1.3 Basis and Dimension

Definition 1.6 (Basis)**.** A family of vectors $(x_1, \ldots, x_n)$ is a **basis** of a vector space V if:

1. They are linearly independent.

2. They span V (i.e., $Span(x_1, \ldots, x_n) = V$).

Definition 1.7 (Dimension)**.** The **dimension** of a vector space V, denoted $\dim(V)$, is the number of vectors in its basis. If a basis 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 of V (we get Span $= V$ for free).

2. If $Span(x_1, \ldots, x_n) = V$, then $(x_1, \ldots, x_n)$ is a basis of V (we get linear independence for free).

To show that a family $x_1, \ldots, x_k$ forms a basis, we need to show 3 things:

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 usually choose to show the easier of 2 and 3.

Definition 1.8 (Line and Hyperplane)**.** Let S be a subspace of $\mathbb{R}^n$.

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 matches x.

Uniqueness: Suppose there are two sets of coordinates α and β such that: $x = \sum \alpha_i v_i = \sum \beta_i v_i$. Then $\sum(\alpha_i - \beta_i)v_i = 0$. Since the basis vectors v_i are linearly independent, the coefficients must be 0. $\alpha_i - \beta_i = 0 \implies \alpha_i = \beta_i$. $\square$

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:

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

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

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

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

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

$\therefore L$ is a linear transformation. □

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

$$\begin{aligned} L: \quad \mathbb{R} \to \mathbb{R} && (13) \\ x \mapsto \quad x^2 && (14) \end{aligned}$$

Proof. Using a counter example to demonstrate the square of the sums is not equal to the sum of the squares:

$$\begin{aligned} L(1 + 2) = (1 + 2)^2 = 9 && (15) \\ L(1) + L(2) = 1 + 4 = 5 && (16) \end{aligned}$$

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

2.3 Matrices

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)**.** We can encode a linear map $L : \mathbb{R}^m \to \mathbb{R}^n$ by a $n \times m$ matrix.

$$\tilde{L} = \begin{pmatrix} L_{1,1} & L_{1,2} & \cdots & L_{1,m} \\ L_{2,1} & L_{2,2} & \cdots & L_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & L_{n,m} \end{pmatrix} \tag{17}$$

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

Examples

Identity matrix I:

$$\tilde{L} = \begin{pmatrix} | & | \\ L(1,0) & L(0,1) \\ | & | \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{18}$$

2.4 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
$mathbbR^m$.

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 $L(v) = v$.

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

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

2.5 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. $\theta_1 x_{i,1} + \cdots + \theta_m x_{i,m} = y_i$.

Problem: How do we find θ? Do we need more data? $\exists\theta$ s.t. $X\theta = y$?

- Yes $\iff$ $y \in \mathrm{Im}(X)$.

- If $y \notin \mathrm{Im}(X)$: no solution.
- If $y \in \mathrm{Im}(X)$: at least 1 solution.

If θ is a solution, $X\theta = y$. Let θ_0 be a particular solution, then $X\theta_0 = y$.

$$X\theta - X\theta_0 = y - y = 0 \implies X(\theta - \theta_0) = 0$$

Therefore $\theta - \theta_0 \in Ker(X)$. The general solution is $\{\theta_0 + v, v \in Ker(X)\} = \theta_0 + Ker(X)$.

- If $Ker(X) = \{0\}$: one unique solution θ_0.
- If $Ker(X) \neq \{0\}$: infinite solutions.

3 Matrix Rank

Idea. Motivation: Consider a dataset $x_1, \ldots, x_n \in \mathbb{R}^d$. 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 $x_1, \ldots, x_k$ of vectors in $\mathbb{R}^n$ as the dimension of its span.

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

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

$$\begin{aligned} rank(M) &\stackrel{\text{def}}{=} rank(c_1, \ldots, c_m) && (19) \\ &= dim(\mathrm{Im}(M)) = dim(Span(c_1, \ldots, c_m)) && (20) \end{aligned}$$

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

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. $rank(\begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}) = 2$.

Proposition 3.1 (Rank of columns = Rank of rows)**.** *Let $r_1, \ldots, r_n \in \mathbb{R}^m$ be the rows of M.*

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

3.2 Rank-Nullity Theorem

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

$$\underbrace{dim(\mathbb{R}^k)}_{domain} = \underbrace{dim(Ker(L))}_{nullity} + \underbrace{dim(\mathrm{Im}(L))}_{rank}$$

Proof. **Inequality 2:** $rank(AB) \leq \min(rank(A), rank(B))$. Showing $rank(AB) \leq rank(B)$: Let L_A, L_B be linear maps for A and B. $Ker(B) \subseteq Ker(AB)$. Thus $dim(Ker(B)) \leq dim(Ker(AB))$. By Rank-Nullity: $rank(AB) = k - dim(Ker(AB)) \leq k - dim(Ker(B)) = rank(B)$. □

3.3 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 $y \in \mathrm{Im}(M)$. Thus $\mathrm{Im}(M) = \mathbb{R}^n$, so $rank(M) = n$. $2 \to 3$**:** From Rank-Nullity, $dim(Ker(M)) = n - rank(M) = 0$. $3 \to 4$**:** Existence: $dim(\mathrm{Im}(M)) = n - dim(Ker(M)) = n$, so $\mathrm{Im}(M) = \mathbb{R}^n$. Uniqueness: If $Mx = Mx'$, then $M(x - x') = 0 \implies x - x' \in Ker(M) = \{0\}$. □

3.4 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 $M^\top_{i,j} = M_{j,i}$ for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$.

1. $(M^\top)^\top = M$.
2. The mapping $M \mapsto M^\top$ is linear, so it is closed under addition and scalar multiplication.

3.4.1 Properties

1. $\forall A \in \mathbb{R}^{n\times m}, rank(A) = rank(A^\top)$. (Rank of A = Rank of columns of A = Rank of rows of A transpose = Rank of A transpose).
2. Let $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times p}$. Then $(AB)^\top = B^\top A^\top$.

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: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$.

Exercise

For each of the norms $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$, draw the "sphere" $S = \{x \in \mathbb{R}^2 \mid \|x\| = 1\}$.

1. ℓ_1 **Norm:** The 4 vectors are $(0, 1), (1, 0), (0, -1), (-1, 0)$. Connecting these 4 vectors forms a square ($|x_1| + |x_2| = 1$).
2. ℓ_2 **Norm:** Forms the standard circle ($x_1^2 + x_2^2 = 1$).
3. ℓ_∞ **Norm:** Forms a square that envelops both ℓ_1 and ℓ_2. ($\max(|x_1|, |x_2|) = 1$).

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 picking the one with the smallest norm.

$$\text{minimize } \|\theta\| \text{ such that } X\theta = y$$

Different norms induce different solutions. For ℓ_1, we prefer solutions on the axes (sparse solutions). For ℓ_2, we prefer solutions distributed across components.

4.2 Inner Products

Definition 4.3 (Inner Product)**.** Let V be a vector space. An **inner product** on V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies:

1. **Linearity in the first argument:** $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$.
2. **Symmetry:** $\langle v, w \rangle = \langle w, v \rangle$.
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.

4.3 Orthogonality

Definition 4.4 (Orthogonality)**.** Two vectors x, y are orthogonal ($x \perp y$) if $\langle x, y \rangle = 0$. We say x is orthogonal to a set S if $x \perp v$ for all $v \in S$.

Theorem 4.2 (Pythagorean Theorem)**.** *If $x \perp y$, then:*

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

Proof.

$$\begin{aligned}
\|x + y\|^2 &= \langle x + y, x + y \rangle && (21)\\
&= \langle x, x \rangle + \langle y, x \rangle + \langle x, y \rangle + \langle y, y \rangle && (22)\\
&= \|x\|^2 + |y|^2 + 2\langle x, y \rangle && (23)\\
&= \|x\|^2 + |y|^2 \iff x \perp y && (24)
\end{aligned}$$

□

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 4.3. *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 $V = \begin{pmatrix} | & & | \\ v_1 & \ldots & v_k \\ | & & | \end{pmatrix}$ *gather the orthonormal basis vectors of S. Then $P_S(x) = VV^\top x$.*

1. P_S is a linear transform.
2. $VV^\top$ is its matrix.
3. $x - P_S(x) \perp S$.

5 Orthogonal Matrices

Preliminaries

Two useful formulas:

1. Let $x, y \in \mathbb{R}^n$. Then $\langle x, y \rangle = x^\top y$.
2. Let $A \in \mathbb{R}^{n \times m}, x \in \mathbb{R}^m, y \in \mathbb{R}^n$. Then $\langle Ax, y \rangle = \langle x, A^\top y \rangle$. This tells us that we can put symmetric matrices on either side and the answer stays the same.

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.

$$v_1 = \frac{x_1}{\|x_1\|} \tag{25}$$

$$\tilde{v_2} = x_2 - \langle x_2, v_1 \rangle v_1 \tag{26}$$

$$v_2 = \frac{\tilde{v_2}}{\|\tilde{v_2}\|} \tag{27}$$

We check orthogonality: $\langle \tilde{v_2}, v_1 \rangle = \langle x_2, v_1 \rangle - \langle x_2, v_1 \rangle \langle v_1, v_1 \rangle = 0$.

5.2 Orthogonal Matrices

Definition 5.1 (Orthogonal Matrix)**.** A matrix $Q \in \mathbb{R}^{n \times n}$ is **orthogonal** if its columns form an orthonormal basis of $\mathbb{R}^n$.

$$Q^\top Q = I_n$$

5.2.1 Properties

Let Q be an orthogonal matrix.

1. **Isometry (Norm preserving):** $\|Qx\| = \|x\|$ for all $x \in \mathbb{R}^n$.

2. **Angle preserving:** $\langle Qx, Qy \rangle = \langle x, y \rangle$.

3. **Orthonormal rows:** The rows of Q also form an orthonormal basis ($QQ^\top = I_n$).

4. **Inverse:** $Q^{-1} = Q^\top$.

Proposition 5.2 (Change of Basis)**.** *Let $(a_1, \ldots, a_n)$ be an orthonormal basis of $\mathbb{R}^n$, and A the matrix collecting these vectors. 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.

5.3 Eigenvalues and Eigenvectors

5.3.1 Introduction

- **Numerical:** Let A be a product of simpler matrices.

- **Geometric:** Suppose we have a linear transformation $L : V \to V$. Is there a basis of V where the matrix is diagonal?

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.

5.3.2 Matrix with no eigenvalues

Consider a rotation matrix $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. If θ is not a multiple of π, there are no real eigenvalues, because no vector stays on the same line (span) after rotation.

5.3.3 Orthogonal Projection

Let P_S be an orthogonal projection onto subspace S.

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

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. The eigenvector is unchanged.*

$$(\alpha A)x = \alpha(Ax) = \alpha(\lambda x) = (\alpha\lambda)x$$

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

$$(A + \alpha Id)x = Ax + \alpha x = \lambda x + \alpha x = (\lambda + \alpha)x$$

3. λ^k is an eigenvalue of A^k. The eigenvector is unchanged.

$$A^k x = A^{k-1}(Ax) = \lambda A^{k-1}x = \cdots = \lambda^k x$$

4. If A is invertible, $\frac{1}{\lambda}$ is an eigenvalue of A^{-1}. (An invertible matrix cannot have a zero eigenvalue).

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 $d_1, \ldots, d_k$, then $\sum d_i \leq n$.*

6.2 Markov Chains

Goal. Motivation: PageRank. How to order web pages by importance?

1. **Naive attempt:** Rank according to incoming links.
2. **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 with transition matrix:

$$P_{i,j} = \begin{cases} \frac{1}{\deg(j)} & \text{if } j \to i \\ 0 & \text{otherwise} \end{cases}$$

Note: Columns sum to 1 (Stochastic matrix). $x^{t+1} = Px^t \implies x^t = P^t x^0$.

Definition 6.4 (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 Example: Random Spectator (Google Matrix)

Imagine a "random spectator" watching tennis:

- At time t, the spectator believes player j is the best ($X_t = j$).
- They pick a game of player j uniformly at random.
- If j wins, they still believe j is best.
- If i beat j, they change their mind to $X_{t+1} = i$.
- With probability α, they pick a new favorite uniformly at random.

This defines a transition matrix P. We rank the players according to the stationary distribution μ of the matrix:

$$M = (1 - \alpha)P + \frac{\alpha}{N}J$$

Where J is the matrix of all ones (teleportation).

6.3 Perron-Frobenius Theorem

Theorem 6.3 (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 $\mu \in \nabla_n$ (probability simplex).*
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 $t \to \infty$.*

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:

$$
\begin{aligned}
Av_i &= \lambda_i v_i \quad \forall i \in \{1, \ldots, n\} && (28)\\
AV &= V\Lambda && (\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n))\\
A &= V\Lambda V^\top && (\text{Since } V^\top = V^{-1})\\
A &= \sum_{i=1}^{n} \lambda_i v_i v_i^\top && (29)
\end{aligned}
$$

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 Why are eigenvectors orthogonal?

Let v_1, v_2 be eigenvectors for distinct eigenvalues λ_1, λ_2.

$$
\begin{aligned}
\lambda_1 \langle v_1, v_2\rangle &= \langle Av_1, v_2\rangle && (30)\\
&= \langle v_1, A^\top v_2\rangle && (31)\\
&= \langle v_1, Av_2\rangle && (\text{Symmetry } A = A^\top)\\
&= \langle v_1, \lambda_2 v_2\rangle && (32)\\
&= \lambda_2 \langle v_1, v_2\rangle && (33)
\end{aligned}
$$

$(\lambda_1 - \lambda_2)\langle v_1, v_2\rangle = 0$. Since $\lambda_1 \neq \lambda_2$, $\langle v_1, v_2\rangle = 0$.

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 $\tilde{a}_1, \dots \tilde{a}_n \in \mathbb{R}^k$ where $k \ll d$.

Assume centered data ($\mu = 0$). Then $\Sigma = \frac{1}{n} \sum a_i a_i^\top = \frac{1}{n} A^\top A$, which is symmetric. We can apply the Spectral Theorem.

Idea. Direction of maximal variance: We aim to find u where variance is maximal: $\max_{\|u\|=1} u^\top \Sigma u$. 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 solution to maximizing $v^\top \Sigma v$ subject to $\|v\| = 1$ and $v \perp v_1, \dots, v \perp v_{j-1}$. This is the j-th eigenvector.

The dimensionally reduced dataset in k dimensions is:

$$\begin{pmatrix} \langle v_1, a_1 \rangle \\ \vdots \\ \langle v_k, a_1 \rangle \end{pmatrix}, \dots, \begin{pmatrix} \langle v_1, a_n \rangle \\ \vdots \\ \langle v_k, a_n \rangle \end{pmatrix}$$

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.

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

Derivation: $A^\top A v_i = \sigma_i A^\top u_i$. From the transpose, $A^\top u_i = \sigma_i v_i$. Thus $A^\top A v_i = \sigma_i^2 v_i$.

- v_i are eigenvectors of $A^\top A$.
- $\sigma_i = \sqrt{\lambda_i(A^\top A)}$.
- $u_i = \frac{Av_i}{\sigma_i}$.

Theorem 8.1 (SVD)**.** *Any matrix $A \in \mathbb{R}^{n \times m}$ can be written as $A = U \Sigma V^\top$.*

8.2 Graphs and Graph Laplacian

Let $G = (V, E)$ be a graph with n vertices.

- **Adjacency Matrix** A: $A_{ij} = 1$ if $(i, j) \in E$, else 0.
- **Degree Matrix** D: Diagonal matrix with $D_{ii} = \deg(i)$.
- **Laplacian** L: $L = D - A$.

8.2.1 Properties of L

1. L is symmetric.
2. L is Positive Semi-Definite (PSD).
3. $\forall x \in \mathbb{R}^n, x^\top L x = \sum_{(i,j) \in E} (x_i - x_j)^2$.
4. $Ker(L)$ contains the vector $\mathbf{1}$ (all ones), since sum of rows is 0.

Theorem 8.2 (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 8.1 (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.

Algorithm (2 Clusters):

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

8.3.2 Spectral Clustering: k clusters

1. Compute $v_1, \ldots, v_k$, the first k eigenvectors of L.
2. Embed every node i into $\mathbb{R}^k$ using the rows of $V = [v_1, \ldots, v_k]$.
3. Run K-means clustering on these points in $\mathbb{R}^k$.

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** $\mathbf{J}(x)$ 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** $\nabla^2 f(x)$ is the $n \times n$ matrix of partial derivatives:

$$(\nabla^2 f(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

9.2 Convexity Definitions

Definition 9.3 (Convex Function)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$:

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

Definition 9.4 (Strictly Convex)**.** A function is **strictly convex** if the inequality is strict for $x \neq y$ and $\alpha \in (0, 1)$.

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

Proposition 9.1. *We say that:*

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 9.2 (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 9.3. *If f is convex and differentiable: x is a global minimizer $\iff$ $\nabla f(x) = 0$.*

Proposition 9.4 (Hessian Condition)**.** *Let f be twice-differentiable. Then f is convex $\iff$ $\nabla^2 f(x)$ is Positive Semi-Definite (PSD) for all x. For strict convexity, we require $\nabla^2 f(x)$ to be Positive Definite.*

9.4 Jensen's Inequality

Theorem 9.5 (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 taking values in $\mathbb{R}^n$:

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

One result of Jensen's inequality is that the variance is always non-negative ($Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$ since x^2 is convex).

Example: Entropy

Consider a random variable X taking values in $\{1, \ldots, k\}$. Let $p_i = P(X = i)$. The entropy is defined as $H(X) = -\sum p_i \log(p_i) = \mathbb{E}[-\log(p(X))]$. Since $-\log(x)$ is convex, we can apply Jensen's inequality to bound the 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:

$$a_i \to \tilde{a}_i = (a_i, 1) \in \mathbb{R}^{d+1} \tag{34}$$

$$\tilde{x} \in \mathbb{R}^{d+1} \tag{35}$$

$$\langle \tilde{x}, \tilde{a}_i \rangle = \sum_{j=1}^{d} (\tilde{x}_j \cdot \tilde{a}_{i,j}) + \tilde{x}_{d+1} \tag{36}$$

$$= \langle x, a \rangle + b \tag{37}$$

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.

$$\begin{aligned} \text{x minimizes } f &\iff \nabla f(x) = 0 && (38)\\ &\iff 2A^\top(Ax - y) = 0 && (39)\\ &\iff A^\top Ax = A^\top y && (40) \end{aligned}$$

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.

$$y_{pred} = \langle a + \delta a, x\rangle = \langle a, x\rangle + \langle \delta a, x\rangle \tag{41}$$

$$\|y_{pred} - \langle a, x\rangle\| = |\langle \delta a, x\rangle| \leq \|\delta a\| \cdot \|x\| \qquad \text{(Cauchy-Schwarz)} \tag{42}$$

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.

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

$$A^\top Ax = A^\top y \to (USV^\top)^\top (USV^\top)x = (USV^\top)^\top y \tag{43}$$

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

$$S^2 V^\top x = SU^\top y \tag{44}$$

$$\text{so each individual component: } \sigma_i^2 \langle v_i, x\rangle = \sigma_i \langle u_i, y\rangle \tag{45}$$

$$\langle v_i, x\rangle = \begin{cases} \frac{1}{\sigma_i}\langle u_i, y\rangle & \text{if } \sigma_i > 0 \\ \text{free} & \text{if } \sigma_i = 0 \end{cases} \tag{46}$$

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

$$S^\dagger_{i,i} = \begin{cases} \frac{1}{S_{i,i}} & \text{if } S_{i,i} \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad S^\dagger_{i,j} = 0 \text{ for } i \neq j \tag{47}$$

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.

$$A = \begin{bmatrix} — & u_1^\top & — \\ & \vdots & \\ — & u_n^\top & — \end{bmatrix} \cdot \begin{bmatrix} | & & | \\ v_1 & \dots & v_m \\ | & & | \end{bmatrix} = U^\top V \text{ where d small} \tag{48}$$

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).

$$\sigma(A) = (\sigma_1(A), \dots, \sigma_{min(n,m)A}) \tag{49}$$

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

11.1 Unconstrained Optimization

Definition 11.1 (Minimizers and Critical Points)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. We say that $x \in \mathbb{R}^n$ is:

- A **global minimizer** of f if for all $x' \in \mathbb{R}^n$, $f(x) \leq f(x')$.
- A **local minimizer** of f if $\exists \delta > 0$ s.t. we have $f(x) \leq f(x')$ for all x' verifying $\|x - x'\| \leq \delta$.
- A **critical point** of f if $\nabla f(x) = 0$.

Proposition 11.1. *For convex functions:*

1. *x is a local minimizer of $f \implies \nabla f(x) = 0$.*
2. *$\nabla f(x) = 0 \iff x$ is a global minimizer of f.*

Heuristic Proof. We define a local minimizer as a point where, under some interval, there is no direction to go from x that ends in a function value lesser than x. Using the first order Taylor expansion to approximate $f(x+h)$:

$$\begin{aligned} f(x+h) = f(x) + h^\top \nabla f(x) + o(\|h\|) &\geq f(x) \\ h^\top \nabla f(x) + o(\|h\|) &\geq 0 \end{aligned}$$

Let $h = -\nabla f(x)$.

$$\begin{aligned} -\|\nabla f(x)\|^2 + o(\|-\nabla f(x)\|) &\geq 0 && (o(\|h\|) \to 0) \\ -\|\nabla f(x)\|^2 &\geq 0 \\ -\|\nabla f(x)\|^2 \geq 0 &\iff \nabla f(x) = 0 \end{aligned}$$

□

In non-convex functions, $\nabla f(x) = 0$ only indicates a critical point. To determine if it is a minimizer, we must examine the second-order term.

Proposition 11.2 (Second Order Conditions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Let $x \in \mathbb{R}^n$ be a critical point of f (i.e., $\nabla f(x) = 0$). Then:*

1. *If $H_f(x)$ admits strictly positive eigenvalues, then x is a **local minimizer** of f.*
2. *If $H_f(x)$ admits strictly negative eigenvalues, then x is a **local maximizer** of f.*

3. *If $H_f(x)$ admits strictly positive AND strictly negative eigenvalues, then x is a* ***saddle point*** *of f.*

Heuristic Proof. Let $f(x)$ be a critical point with $\nabla f(x) = 0$.

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

Let $h = \varepsilon e_i$ (where e_i is an eigenvector of $\nabla^2 f(x)$ for eigenvalue λ_i).

$$f(x + \varepsilon e_i) = f(x) + \frac{\varepsilon^2}{2} \lambda_i + o(\varepsilon^2)$$

Thus, the sign of the eigenvalues determines how $f(x+h)$ compares with $f(x)$. Note that **strict** signage is important!

Edge case: If eigenvalues are 0, the second order term vanishes, and we must rely on third-order information. □

11.2 Constrained Optimization and Lagrange Multipliers

Constrained optimization problems take the form:

$$\begin{aligned} \text{minimize } f(x) \text{ with variable } x \in \mathbb{R}^n \\ \text{subject to } g_i(x) \leq 0, \quad i = 1, \ldots, m \\ h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

Examples

- $\min_x x^\top A x$ s.t. $\|x\| = 1$. Here $h_1(x) = \|x\| - 1$.
- $\min_x \|Ax - y\|^2$ s.t. $\|x\| \leq r$. Here $g_1(x) = \|x\| - r$.

11.2.1 Feasible Points

Definition 11.2 (Feasible Set)**.** A point $x \in \mathbb{R}^n$ is **feasible** if it satisfies all constraints. The set of all feasible points is called the **feasible set**, denoted F.

Idea. Can we be sure that the *unconstrained minimizer* is always in the feasible set? No. Consider $\min_x x^2$ s.t. $x \geq 1$. The unconstrained minimizer is $x = 0$. However, the feasible set is $[1, \infty)$, thus the *constrained minimizer* is $x = 1$.

When the optimality condition $\nabla f(x) = 0$ doesn't hold, what conditions characterize solutions?

11.2.2 First Order Optimality Condition

Consider the problem: $\min_x f(x)$ subject to $g(x) \leq 0$. The optimal point x^* can be either inside the feasible set or on the boundary.

Case 1: x^* inside (Constraint Inactive). There exists $\delta > 0$ s.t. $B(x^*, \delta) \subset F$. Since $g(x^*) < 0$ and we assume continuity, we can locally minimize $f(x)$ over the ball $B(x^*, \delta)$. This is equivalent to an unconstrained local minimizer, so $\nabla f(x^*) = 0$.

Case 2: x^* on boundary (Constraint Active). At a boundary minimum, the gradient of the objective function is orthogonal to the tangent of the feasible set and points outward (away from the unconstrained optimal). Formally, if $g(x^*) = 0$, then $f(x^*) + \lambda \nabla g(x^*) = 0$ for some $\lambda \geq 0$.

Theorem 11.3 (Lagrange Multipliers)**.** *If x is a solution and if $\nabla h_1(x), \ldots, \nabla h_p(x), \{\nabla g_i(x) \mid g_i(x) = 0\}$ are linearly independent, then there exists $\lambda_1, \ldots, \lambda_m \geq 0$ and $\nu_1, \ldots, \nu_p \in \mathbb{R}$ such that:*

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

Moreover, for all $i \in \{1, \ldots, m\}$, if $g_i(x) < 0$, then $\lambda_i = 0$.

11.2.3 Example

Minimize $\langle x, u \rangle$ subject to $\|x\|^2 = 1$ ($u \neq 0$). Let $h(x) = \|x\|^2 - 1$, so $\nabla h(x) = 2x$.

$$u + \nu 2x = 0 \implies x = -\frac{1}{2\nu} u$$

Using the constraint $\|x\| = 1$:

$$1 = \|-\frac{1}{2\nu} u\| = \frac{1}{2|\nu|} \|u\| \implies |\nu| = \frac{\|u\|}{2}$$

Solution: $x = \pm \frac{u}{\|u\|}$. Checking values: $f(x) = \langle \pm \frac{u}{\|u\|}, u \rangle = \pm \|u\|$. The minimum is $-\|u\|$ at $x = -\frac{u}{\|u\|}$.

11.3 Convex Constrained Optimization

A constrained optimization problem is **convex** when $f, g_1, \ldots, g_m$ are convex functions and $h_1, \ldots, h_p$ are affine functions.

11.3.1 Karush-Kuhn-Tucker (KKT) Theorem

Theorem 11.4 (KKT Conditions)**.** *Assume the problem is convex and Slater's condition holds (exists a feasible point where strict inequality holds). Then x is a solution* ***if and only if*** *x is feasible and there exists Lagrange multipliers $\lambda_1, \ldots, \lambda_m \geq 0, \nu_1, \ldots, \nu_p \in \mathbb{R}$ such that:*

1. ***Primal feasibility:*** *$g_i(x) \leq 0, \forall i$ and $h_j(x) = 0, \forall j$.*

2. ***Dual feasibility:*** $\lambda_i \geq 0, \forall i$.

3. ***Stationarity:*** $\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{p} \nu_j \nabla h_j(x) = 0$.

4. ***Complementary slackness:*** $\lambda_i g_i(x) = 0, \forall i$.

Example: Ridge Regression

$$\begin{aligned} \text{minimize } & \|Ax - y\|^2 \\ \text{subject to } & \|x\|^2 \leq r^2 \end{aligned}$$

x is a solution if there exists $\lambda \geq 0$ such that:

$$\begin{cases} \|x\|^2 \leq r^2 \\ 2A^\top(Ax - y) + \lambda 2x = 0 \\ \lambda(\|x\|^2 - r^2) = 0 \end{cases}$$

From Stationarity: $(A^\top A + \lambda Id)x = A^\top y$. There are two cases:

1. $\lambda = 0 \implies x = (A^\top A)^{-1} A^\top y$ (Unconstrained solution is feasible).

2. $\lambda > 0 \implies \|x\| = r$.

Example: Projection onto subspace

Let $u, v \in \mathbb{R}^n$ s.t. $\|v\| = 1$.

$$\begin{aligned} \text{minimize } & \|x - u\|^2 \\ \text{subject to } & x \perp v \iff \langle x, v \rangle = 0 \end{aligned}$$

KKT Conditions:

$$\begin{cases} \langle x, v \rangle = 0 \\ 2(x - u) + \nu v = 0 \implies x = u - \frac{\nu}{2} v \end{cases}$$

Substitute x into the constraint:

$$\begin{aligned} \langle u - \frac{\nu}{2} v, v \rangle &= 0 \\ \langle u, v \rangle - \frac{\nu}{2} \|v\|^2 &= 0 \\ \nu &= 2\langle u, v \rangle \end{aligned}$$

Thus, $x = u - \langle u, v \rangle v$. (This is the orthogonal projection of u onto the plane orthogonal to v).

12 Gradient Descent

12.1 Setup and Update Rule

Suppose we have an unconstrained optimization problem (note from last lecture we can turn constrained optimization problems into unconstrained problems via Lagrange multipliers).

Goal. Minimize a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$.

Starting from a point $x_0 \in \mathbb{R}^n$, perform the updates

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t) \quad (\alpha_t = \text{step size})$$

Intuitively, gradient descent moves in the direction of steepest descent under the Euclidean norm.

Recall the first-order Taylor expansion

$$f(x_t + h) = f(x_t) + \langle h, \nabla f(x_t) \rangle + o(\|h\|).$$

Thus the inner product determines the first-order change of the function.

Let $h = -\alpha_t \nabla f(x_t)$. Then

$$f(x_t + h) = f(x_t) - \alpha_t \|\nabla f(x_t)\|^2 + o(\alpha_t \|\nabla f(x_t)\|).$$

Since the squared norm is nonnegative, taking a sufficiently small step size guarantees a decrease in the function value.

In convex cases, gradient descent is guaranteed to find the global minimum. In most real-world problems, the objective is non-convex, so depending on initialization and step size, one is generally only guaranteed convergence to a local critical point.

12.2 Convergence Analysis for Convex Functions

Assume f is convex and L-smooth ($\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$). With a fixed step size $\alpha \leq 1/L$, gradient descent converges at a rate of $O(1/t)$.

$$f(x_t) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\alpha t}$$

12.3 Improvements to Gradient Descent

Momentum. When the condition number $\kappa = \frac{L}{\mu}$ is large, the gradient direction can be poorly aligned with the direction to the minimum.

Momentum mimics a heavy ball:

$$v_t = -\alpha_t \nabla f(x_t) + \beta_t v_{t-1}, \qquad x_{t+1} = x_t + v_t.$$

Newton's Method. Assume f is μ-strongly convex and L-smooth. Newton's method updates

$$x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t).$$

This corresponds to minimizing the second-order Taylor approximation. Solving for the critical point h in the second order expansion yields the direction $h = -\nabla^2 f(x)^{-1} \nabla f(x)$.

A related Newton's method is conceptually similar and is known as Newton's Method for Root Finding. Geometrically, it:

1. Find the tangent of the current point.
2. Find where the tangent intersects 0.
3. Move there: $x_{t+1} = x_t - \frac{g(x_t)}{g'(x_t)}$ (Root finding for $\nabla f(x) = 0$).

Newton's method converges extremely fast (often < 10 iterations), but requires $O(d^2)$ memory and $O(d^3)$ time per step to store and invert the Hessian.

Stochastic Gradient Descent (SGD). In machine learning, objectives often take the form

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i, y_i)$$

Computing the full gradient takes $O(nd)$. SGD approximates the gradient by a single randomly chosen term:

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \ell(\theta, x_{i_t}, y_{i_t})$$