Mathematical Preliminaries

SIE 449/549: Optimization for Machine Learning

Afrooz Jalilzadeh

Department of Systems and Industrial Engineering
University of Arizona

The space $\mathbb{R}^n$

- Vector space $\mathbb{R}^n$ is the set of n-dimensional column vectors with real components

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix} \qquad \alpha \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

- **Zero vector** denoted by **0** or simply 0 is a vector with all elements being zero
- $\mathbf{e} = [1, 1, \ldots, 1]^T$ is a vector with all elements being 1
- $\mathbf{e}_i$ is the n-length vector with i-th element is 1 and all others are zeros

Important Subsets of $\mathbb{R}^n$

- Nonnegative orthant:

$$\mathbb{R}^n_+ = \{(x_1, x_2, \ldots, x_n)^T : x_1, x_2, \ldots, x_n \geq 0\}$$

- Positive orthant:

$$\mathbb{R}^n_{++} = \{(x_1, x_2, \ldots, x_n)^T : x_1, x_2, \ldots, x_n > 0\}$$

- Closed line segment between $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ is given by:

$$[x, y] = \{\alpha x + (1-\alpha)y : \alpha \in [0,1]\}$$

- Open line segment (x, y) is similarly defined as:

$$(x, y) = \{\alpha x + (1-\alpha)y : \alpha \in (0,1)\}$$

for $x \neq y$ and $(x, x) = \emptyset$

- Unit Simplex:

$$\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \underbrace{e^T x = 1}_{\sum_{i=1}^n x_i = 1}\}$$

The space $\mathbb{R}^{m\times n}$

- **Matrix** $A \in \mathbb{R}^{m\times n}$ has m rows and n columns and we use a_{ij} to refer to the entry in the ith row and jth column.

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

- A matrix with one row (or one column) is a vector. $u = [u_1, \dots, u_n]$ or $v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$

- Given $A, B \in \mathbb{R}^{m\times n}$, entries of the matrix sum $A+B$ are given by the sum of entries, i.e. $(A+B)_{ij} = a_{ij} + b_{ij}$. Multiplying a matrix by scalar $\alpha \in \mathbb{R}$ involves scaling each entry by α, i.e. $(\alpha A)_{ij} = (A\alpha)_{ij} = \alpha a_{ij}$.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}, \quad 2\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$$

Matrix

- if $A \in \mathbb{R}^{m \times n}$, then we denote the **transpose** of A by A^T, where $A^T \in \mathbb{R}^{n \times m}$ and $a_{ij}^T = a_{ji}$. Matrix A is **symmetric**, if $A = A^T$.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}, \quad \begin{bmatrix} 7 & 8 \\ 9 & 10 \end{bmatrix}^T = \begin{bmatrix} 7 & 9 \\ 8 & 10 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}^T = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

Symmetric

- **Multiplication**: Given $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times k}$ then $C = A \times B$ has m rows and k columns, $C \in \mathbb{R}^{m \times k}$, and $c_{ij} = \sum_{z=1}^{n} a_{iz} b_{zj}$.

$$\underset{2\times 3}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}} = \underset{2\times 1}{\begin{bmatrix} 14 \\ 32 \end{bmatrix}}, \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix}$$

Matrix

- Matrix multiplication is **associative**: $A(BC) = (AB)C$.
- Matrix multiplication is **distributive**: $A(B + C) = AB + AC$.
- Some important matrices and vectors:
  - **Identity matrix** of size n is the $n \times n$ square matrix with ones in the main diagonal and zeros elsewhere and we denoted by I_n.

eye(n)

$$I_1 = \begin{bmatrix} 1 \end{bmatrix}, \quad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ldots$$

  - $x_i, x_2, \ldots, x_n \geq 0$ and $x = [x_i, x_2, \ldots, x_n]^T$: $x \geq \mathbf{0}$.

Inner products

- The inner product (also called dot product) of two n-vectors is defined as the scalar

$$\langle x, y \rangle = x \cdot y = x^T y = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$

$$\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

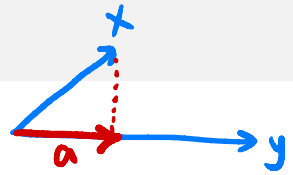
$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \text{ for any } x, y \in \mathbb{R}^n$$

General Examples:

- Unit vector: $\langle e_i, x \rangle = e_i^T x = \begin{bmatrix} 0 & 0 & \cdots & \underset{\text{ith}}{1} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_i$
- Sum: $\langle e, x \rangle = e^T x = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i$
- Sum of squares: $\langle x, x \rangle = x^T x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i^2$

Inner products

$\langle x, y \rangle = x \cdot y = x^T y$



$x \cdot y =$ (length of $\vec{a}$) . (length of $\vec{y}$)

Definition 1 (Inner Products)

An inner product on R^n is a map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:

1. Symmetry: $\langle x, y \rangle = \langle y, x \rangle$ for any $x, y \in \mathbb{R}^n$
2. Additivity: $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ for any $x, y, z \in \mathbb{R}^n$
3. Homogeneity: $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$ for $\lambda \in \mathbb{R}$ and any $x, y \in \mathbb{R}^n$
4. Positive definiteness: $\langle x, x \rangle \geq 0$ for any $x \in \mathbb{R}^n$ and $\langle x, x \rangle = 0$ iff $x = 0$

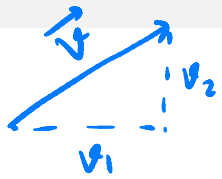
- Weighted dot product: $\langle x, y \rangle_w = \sum_{i=1}^n w_i x_i y_i$ where $w \in \mathbb{R}^n_{++}$

Vector Norms

length of $v = \sqrt{x^2 + y^2} = \|v\|$

- The **norm** of a nonzero vector is a positive number $\|x\|$
- That number measures the “length" of the vector
- There are many useful measures of length (many different norms)
- A norm $\|\cdot\|$ on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ satisfying
  - For all $x \in \mathbb{R}^n$, $\|x\| \geq 0$ (non-negativity).
  - $\|x\| = 0$ if and only if $x = 0$ (definiteness).
  - For all $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$, $\|\lambda x\| = |\lambda|\|x\|$ (homogeneity).
  - For all $x, y \in \mathbb{R}^n$, $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

Vector Norms



$\|v\| = \sqrt{v_1^2 + v_2^2}$

$\|v\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$

- One natural way to generate a norm on $\mathbb{R}^n$ is to take any inner product $\langle \cdot, \cdot \rangle$ defined on $\mathbb{R}^n$, and define the associated norm

$$\|x\| = \sqrt{\langle x, x \rangle}, \quad \text{for all } x \in \mathbb{R}^n$$

$\langle x, x \rangle = \sum x_i^2$

- The norm associated with the dot-product is the so-called Euclidean norm or ℓ_2-norm:

$$\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \text{for all } x \in \mathbb{R}^n$$

Note that $\|x\|_2^2 = x^T x$

Vector Norms

Important vector norms are the following:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_\infty = \max_{1 \le i \le n} |x_i|$$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \text{ for a real number } p \ge 1$$

Compute the following norms for $x = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$:

$\|\cdot\|_{1/2}$ is not a norm!

$\|x\|_1 = |1| + |-2| = 3$

$\|x\|_2 = \sqrt{1^2 + 2^2} = \sqrt{5}$

$\|x\|_\infty = \max\{|1|, |-2|\} = 2$

Vector Norms

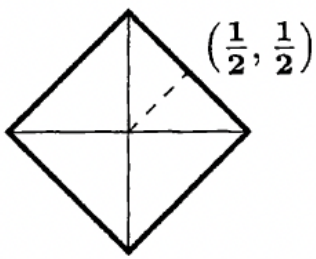
$\|v\|_\infty \leq \|v\|_2 \leq \|v\|_1$

- The all-ones vector $v = (1, 1, \ldots, 1)$ has norms:

$$\|v\|_2 = \sqrt{n} \qquad \|v\|_1 = n \qquad \|v\|_\infty = 1$$

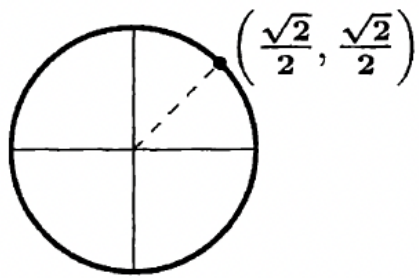
$\|v\|_1 \leq 1$

ℓ^1 **norm**
$|v_1| + |v_2| \leq 1$
diamond

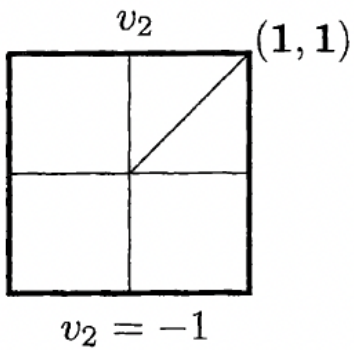


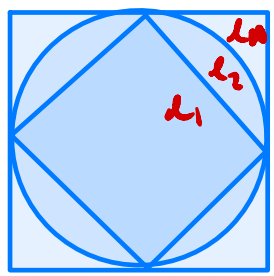
$\|v\|_2 \leq 1$

ℓ^2 **norm**
$v_1^2 + v_2^2 \leq 1$
circle



ℓ^∞ **norm**
$|v_1| \leq 1, |v_2| \leq 1$
square





$\|v\|_\infty \leq \|v\|_2 \leq \|v\|_1$

Vector Norms

- Which point on a diagonal line like $3v_1 + 4v_2 = 1$ is closest to $(0, 0)$?
- The answer (and the meaning of “closest") will depend on the norm
- This is another way to see important differences between norms

$$\min_{v} \; \|v - 0\|_{?}$$

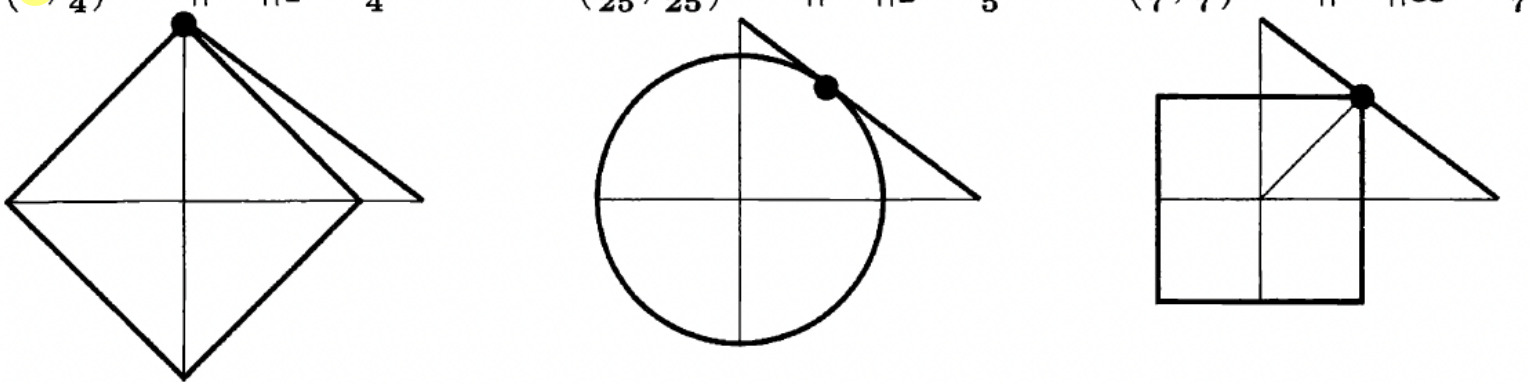
$$\text{s.t.} \quad 3v_1 + 4v_2 = 1$$

Vector Norms

- Which point on a diagonal line like $3v_1 + 4v_2 = 1$ is closest to $(0, 0)$?
- The answer (and the meaning of “closest") will depend on the norm
- This is another way to see important differences between norms

Minimize $||v||_p$ among vectors (v_1, v_2) on the line $3v_1 + 4v_2 = 1$

$(0, \frac{1}{4})$ has $||v^*||_1 = \frac{1}{4}$ $\left(\frac{3}{25}, \frac{4}{25}\right)$ has $||v^*||_2 = \frac{1}{5}$ $\left(\frac{1}{7}, \frac{1}{7}\right)$ has $||v^*||_\infty = \frac{1}{7}$



- The first figure displays a highly important property of the minimizing solution to the ℓ_1 problem: **That solution v^* has zero components $\implies$ v^* is sparse**
- This is because a diamond touches a line at a sharp point

The Cauchy-Schwartz Inequality

Lemma 2 (Cauchy-Schwartz Inequality)

For any $x, y \in \mathbb{R}^n$: $|x^T y| \leq \|x\|\|y\|$

One can use Cauchy-Schwartz Inequality to prove Triangle Inequality.

Lemma 3 (Triangle Inequality)

For any $x, y \in \mathbb{R}^n$: $\|x + y\| \leq \|x\| + \|y\|$

$\langle x, x \rangle = \|x\|^2$

Proof.

$$\|x+y\|^2 = \langle x+y, x+y \rangle$$

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

$$= \|x\|^2 + \langle x, y \rangle + \langle y, x \rangle + \|y\|^2$$

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

$$= (\|x\| + \|y\|)^2. \quad \square$$

Linear Combination

- A **linear combination** of the vectors in V is any vector of the form

$$c_1 v_1 + c_2 v_2 + \ldots + c_k v_k$$

where $c_1, c_2, \ldots, c_k$ are arbitrary scalars.

- We call the linear combination of vectors in V for which $c_1 = c_2 = \ldots = c_k = 0$ the **trivial linear combination** of vectors in V.
- A set V of n-dimensional vectors is **linearly independent** if the only linear combination of vectors in V that equals **0** is the trivial linear combination.

Example 1

Show that $V = \{[1, 0], [0, 2]\}$ *is a linearly independent set of vectors.*

$$c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2\begin{bmatrix}0\\2\end{bmatrix} = 0$$

$$\begin{bmatrix}c_1\\2c_2\end{bmatrix} = 0 \Rightarrow \begin{cases} c_1 = 0 \\ c_2 = 0 \end{cases} \Rightarrow \text{linearly indep.}$$

Linear Combination

- A set V of n-dimensional vectors is **linearly dependent** if there is a nontrivial linear combination of the vectors in V that adds up to $\mathbf{0}$.

Example 2

Show that $V = \{[0,0],[1,0],[0,2]\}$ *is a linearly dependent set of vectors.*

$$c_1\begin{bmatrix}0\\0\end{bmatrix} + c_2\begin{bmatrix}1\\0\end{bmatrix} + c_3\begin{bmatrix}0\\2\end{bmatrix} = 0$$

$$\begin{bmatrix}c_2\\2c_3\end{bmatrix} = 0 \Longrightarrow c_2 = 0 \quad c_3 = 0 \Longrightarrow \text{linearly dep.}$$

$$c_1 \in \mathbb{R}$$

Rank of Matrix

- The **column rank** of a matrix A is the largest number of columns of A that constitute linearly independent set. The **row rank** is the largest number of rows of A that constitute a linearly independent set.
- for any matrix A, columnrank(A) = rowrank(A), this quantity is simply refereed to as the **rank of** A, denoted as rank(A).

Example 3

What is the rank of the following matrices?

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

$c_1\begin{bmatrix}1\\2\end{bmatrix} + c_2\begin{bmatrix}1\\2\end{bmatrix} = 0 \Rightarrow c_1 = -c_2 \Rightarrow$ dep. rank(A) = 1

$c_1\begin{bmatrix}1\\0\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\\0\end{bmatrix} + c_3\begin{bmatrix}0\\0\\1\end{bmatrix} = 0 \Rightarrow c_1 = c_2 = c_3 \Rightarrow$ indep. rank(B) = 3

- For $A \in \mathbb{R}^{m\times n}$, rank($A$)$\leq \min(m, n)$.
- If rank(A) = $\min(m, n)$, then A is said to be **full rank**.

Inverse of Matrix

- For a given $m \times m$ matrix A, the $m \times m$ matrix B is the **inverse** of A if:

$$AB = BA = I_m.$$

- A^{-1} exists if and only if A is full rank.
- Given a matrix square $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a **quadratic form**. Written explicitly as:

$$x^T Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

$$\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \cdots A_{1n} \\ \vdots & \\ A_{n1} & \cdots A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Positive Definitie Matrix

$a > 0 \quad A \succ 0$

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **positive definite** (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$. This is usually denoted by $A \succ 0$.
- If $x^T Ax \geq 0$, we call A a **positive semidefinite** (PSD) matrix, $A \succeq 0$.
- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called **indefinite** if there exist $x, y \in \mathbb{R}^n$ such that $x^T Ax > 0$ and $y^T Ay < 0$.
- If A is positive definite, then $-A$ is negative definite and vice versa.
- If A is positive semidefinite then $-A$ is negative semidefinite and vice versa.

$\langle x, y \rangle = x^T y \qquad \langle x, x \rangle = \|x\|^2$

Theorem 4

Let $A \in \mathbb{R}^{n \times n}$, the matrix $A^T A$ is symmetric and PSD.

Proof.

$x^T A^T A x = \langle Ax, Ax \rangle = \|Ax\|^2 \geq 0$

Positive Definitie Matrix

Example. Show that $A = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ is positive definite.

$$X^T A X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 2x_1^2 - 2x_1x_2 + x_2^2 = x_1^2 + (x_1 - x_2)^2 \geq 0$$

Also, $x_1^2 + (x_1 - x_2)^2 = 0$ iff $x_1 = x_2 = 0$.

Eigenvalues and Eigenvectors

- Given an $n \times n$ matrix A, an **eigenvalue** for A is a number λ which, for some nonzero vector v, satisfies

$$Av = \lambda v \quad \Longleftrightarrow \quad (A - \lambda I)v = 0,$$

where I is the identity matrix.

- The vector v is the **eigenvector** associated with the eigenvalue λ.
- If we know the eigenvalue λ, then we can get the associated eigenvector by solving the simultaneous linear equations.
- How many eigenvectors can you find? infinite.

$$Av = \lambda v \xrightarrow{c \text{ is a scalar}} Acv = \lambda cv \Rightarrow Av = \lambda v$$

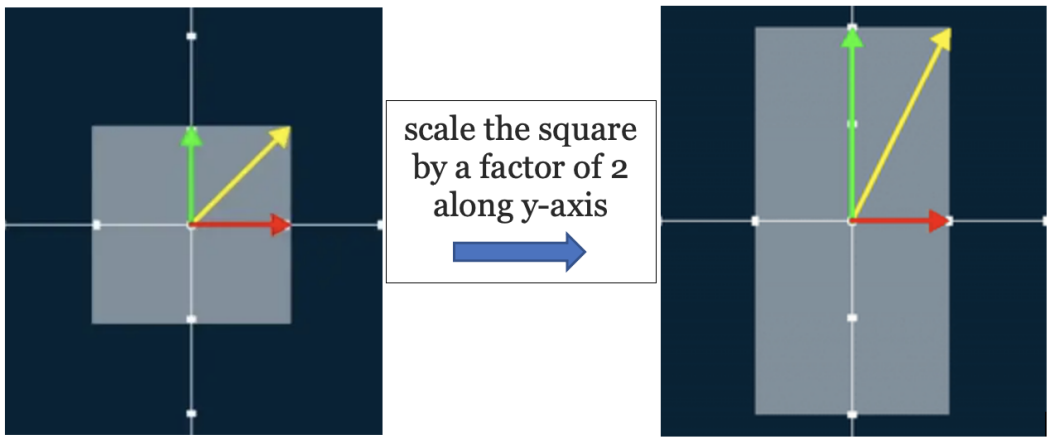
- How to find eigenvalues? $\det(A - \lambda I) = 0$

Meaning of Eigenvalues and Eigenvectors

- Suppose we have a square represented in 2d space where every point on the square is a vector:

$Av = \lambda v$

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



- red vector has the same scale and direction after the linear transformation
- Green vector changes in scale but still has the same direction
- Yellow vector has a different scale and direction
- We can say the red and green vector are special and they are characteristic of this linear transform
- Red and Green vectors are **eigenvectors**

$\lambda_1 = 1 \qquad \lambda_2 = 2$

- Their change in scale due to the transformation is called their **eigenvalue**

Why eigenvalues are important?

- The Tacoma Bridge was built in 1940 in state of Washington.



Video: https://www.youtube.com/watch?v=XggxeuFDaDU

- The oscillations of the bridge were caused by the frequency of the wind being too close to the natural frequency of the bridge.
- The natural frequency of the bridge is the eigenvalue of smallest magnitude of a system that models the bridge.
- **Other applications:** Face recognition, Rank pages in Google, ...

Why eigenvalues are important?

- Consider the following square matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

- The trace of A is $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$
- Can we find the trace of a matrix using eigenvalues?

$$tr(A) = \lambda_1 + \lambda_2 + \dots + \lambda_n$$

- How about and determinant?

$$det(A) = \lambda_1 \times \lambda_2 \times \dots \times \lambda_n$$

Eigenvalue Decomposition

$Av = \lambda v$

- If $A = A^T$, i.e., A is symmetric, and $A \in \mathbb{R}^{n\times n}$, then it can be shown that there exist n pairs of eigenvalue-eigenvectors $(\lambda_i, v_i) \in \mathbb{R} \times \mathbb{R}^n$ such that $v_i^T v_j = 0$ for all $i \neq j$, $\|v_i\| = 1$ for all i and

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T.$$

Theorem 5

A is PSD $\iff \lambda_i \geq 0 \ \forall i$

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is a positive semidefinite (PSD) if and only if all eigenvalues of A are non-negative. (Similarly, it is PD iff all eigenvalues are positive.)

Proof. "$\Longrightarrow$" By contradiction, suppose $\exists v \in \mathbb{R}^n$ $\lambda < 0$ s.t. $Av = \lambda v$.

$v^T A v = \lambda v^T v = \lambda \|v\|^2 < 0 \Longrightarrow A$ is not PSD.

"$\Longleftarrow$" $A = \sum_{i=1}^{n} \lambda_i v_i v_i^T \Longrightarrow x^T A x = \sum_{i=1}^{n} \lambda_i x_i^T v_i v_i^T x_i = \sum_{i=1}^{n} \lambda_i (v_i^T x_i)^2 \geq 0$

$\Longrightarrow A$ is PSD. $\square$

Theorem 6

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is indefinite if and only if it has at least one positive eigenvalue and at least one negative eigenvalue.

Singular Value Decomposition (SVD)

- If A is a symmetric and $n \times n$ matrix, then:
  - Eigenvalues are always real
  - Eigenvectors are orthogonal to each other, i.e., the dot product of the two eigenvectors is zero
- If A is not symmetric, then the eigenvalues are complex or the eigenvectors are not orthogonal
- If A is not square then $Av = \lambda v$ is impossible and eigenvectors fail (left side in $\mathbb{R}^m$, right side in $\mathbb{R}^n$)
- We need an idea that succeeds for every matrix
- The Singular Value Decomposition (SVD) fills this gap in a perfect way

$$AV = U\Sigma \qquad A \begin{bmatrix} v_1 \,..\, v_r \,..\, v_n \end{bmatrix} = \begin{bmatrix} u_1 \,..\, u_r \,..\, u_m \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$

Singular Value Decomposition (SVD)

- A is $m \times n$ matrix with rank r, then:
  - We will have r positive singular values in descending order

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$$

  - We have n right singular vectors $(v_1, \ldots, v_n)$
  - We have m left singular vectors $(u_1, \ldots, u_m)$

$$Av_1 = \sigma_1 u_1, \quad \ldots, \quad Av_r = \sigma_r u_r$$

$$Av_{r+1} = 0, \quad \ldots, \quad Av_n = 0$$

  - All of the right singular vectors v_1 to v_n go in the columns of V
  - The left singular vectors u_1 to u_m go in the columns of U
  - V and U are square orthogonal matrices, i.e., $V^T = V^{-1}$ and $U^T = U^{-1}$
  - Then we have: $AV = U\Sigma$

The Singular Value Decomposition of A is: $A = U\Sigma V^T$

Matrix Norm

$|xy| \leq \|x\| \, \|y\|$.

- A matrix norm $\|\cdot\| : K^{m\times n} \to \mathbb{R}$ is a function satisfying the following properties for all scalar $\lambda \in \mathbb{R}$ and matrices $A, B \in \mathbb{R}^{m\times n}$:
  - $\|A\| \geq 0$ (non-negativity).
  - $\|A\| = 0$ if and only if $A = 0_{m\times n}$ (definiteness).
  - $\|\lambda A\| = |\lambda|\|A\|$ (homogeneity).
  - $\|A + B\| \leq \|A\| + \|B\|$ (triangle inequality).
- Matrix norms are particularly useful if they are also sub-multiplicative:

$$\|AB\| = \|A\|\|B\|$$ ← similar to Cauchy-Schwartz

- **Frobenius norm:** treats matrices as long vectors: $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad \|A\|_F = \sqrt{(1)^2 + (2)^2 + (3)^2 + (4)^2} = \sqrt{30}$$

Matrix norms induced by vector p-norms

- When we compare $\|Av\|$ to $\|v\|$, this measures the growth factor – the increase or decrease in size produced when we multiply by A
- If we choose the vector v with the largest growth factor, that gives an important matrix norm $\|A\|$:

$$\|A\| = \max_{v \neq 0} \frac{\|Av\|}{\|v\|}$$

- How to compute $\|A\|_2$, $\|A\|_1$ and $\|A\|_\infty$?

$\|A\|_2$ = largest singular value of $A = \sigma_1$

$\|A\|_1$ = " ℓ_1 norm of columns of A.

$\|A\|_\infty$ = " ℓ_1 " " rows of A.

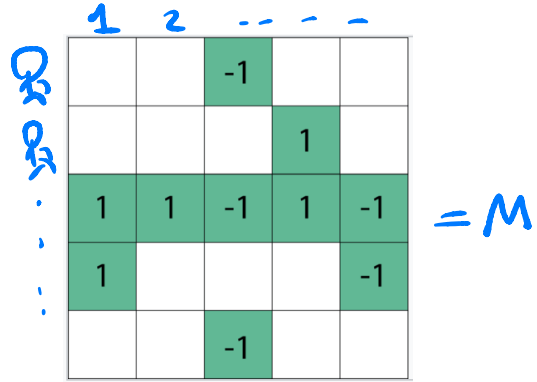
$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$ $\|A\|_1 = 8$ $\|A\|_\infty = 9$

- **Nuclear norm** or trace norm: $\|A\|_{nuclear} = \sigma_1 + \ldots + \sigma_r$
- **Frobenius norm**: $\|A\|_F = \sigma_1^2 + \ldots + \sigma_r^2$

Example: Matrix Completion

- **Netflix Challenge**: An automated recommendation system
- Consider the dataset of user-media preference pairs as a partially-observed matrix
- Every row represents a person, and every column represents a movie

$$M_{ij} = \begin{cases} -1, & \text{person i dislikes movie j} \\ 1, & \text{person i likes movie j} \\ \star, & \text{preference unknown} \end{cases}$$



- **Goal:** Complete the matrix
- There are only k factors that determine a persons preference over movies, such as genre, director, actors and so on
- Define $\|\cdot\|_{OB}$ as the norm only on the observed (non starred) entries of M, i.e., $\|X\|_{OB} = \sum_{M_{ij} \neq \star} X_{ij}^2$

$$\min_X \|X - M\|_{OB} \quad \text{s.t.} \quad \text{rank}(X) \leq k \qquad \text{or} \qquad \min_X \text{rank}(X) \quad \text{s.t.} \quad X_{ij} = M_{ij}$$

$$\min \|X\|_{\text{nuclear}}$$

Example: Principal Component Analysis (PCA)

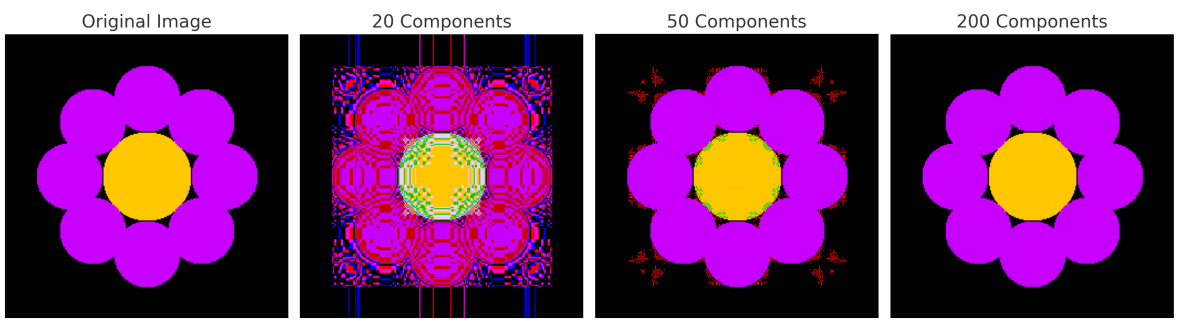
- What is the closest rank k matrix to matrix A?
- **Principal Component Analysis (PCA)** uses the largest σs connected to the first us and vs to understand the information in a matrix of data:

$$A_k = \sigma_1 u_1 v_1^T + \ldots + \sigma_k u_k v_k^T, \quad \text{s.t.} \quad rank(A) = k$$

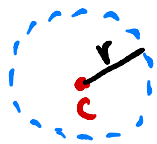
- A_k solves an optimization problem:

$$\min_{A_k} \|A - A_k\| \quad \text{s.t.} \quad rank(A_k) = k$$

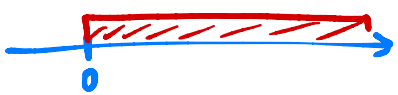
- **Applications:** Dimensionality Reduction, Data Visualization, ...



Basic Topological Concepts



- The **open ball** with center $c \in \mathbb{R}^n$ and radius r: $B(c, r) = \{x \mid \|x - c\| < r\}$
- The **Closed ball** with center $c \in \mathbb{R}^n$ and radius r: $B[c, r] = \{x \mid \|x - c\| \leq r\}$
- Given a set $U \subseteq \mathbb{R}^n$, a point $c \in U$ is called an **interior point** of U if there exists $r > 0$ for which $B(c, r) \subseteq U$.



$\text{int}(U) = \{x \in U \mid B(x, r) \subseteq U \text{ for some } r > 0\}$

$\text{int}(\mathbb{R}^n_+) = \mathbb{R}^n_{++}$

Basic Topological Concepts

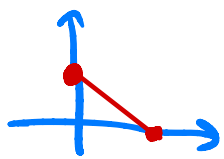
- An **open set** is a set that contains only interior points. Meaning that $U = \text{int}(U)$.

$B(c, r)$, $\mathbb{R}^n_{++}$

- A set $U \subseteq \mathbb{R}^n$ is **closed** if it contains all the limits of convergent sequences of vectors in U, that is, if $\{x_i\}_{i=1}^{\infty} \subseteq U$ satisfies $x_i \to x^*$ as $i \to \infty$, then $x^* \in U$.

$B[c, r]$, $\mathbb{R}^n_+$, Δ_n

$\Delta_2 = \{x \in \mathbb{R}^2 \mid x_i \geq 0, \ \sum x_i = 1\}$



- A set $U \subseteq \mathbb{R}^n$ is called **bounded** if there exists $M > 0$ which $U \subseteq B(0, M)$.
- A set $U \subseteq \mathbb{R}^n$ is called **compact** if it is closed and bounded.

$\mathbb{R}^n_+$ is not compact. b/c not bounded.

Big O Notation

x x^2

- Imagine that you're working on an investment strategy that aims to grow a principal amount of money as fast as possible. Would you be more likely to pick a linear growth strategy or a quadratic growth strategy?
- To make the most amount of money as quickly as possible, it would be wise to pick the quadratic growth strategy.
- We can use Big O notation to describe how fast your money will grow over time.
- In general, Big O notations are used to describe the limiting behavior of functions when $x \to \infty$.

Big O Notation

$O(1/k)$ $O(1/k^2)$

Definition 7

Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two real valued functions. We denote $f(x) = \mathcal{O}(g(x))$, as $x \to \infty$ if there exist $\alpha, x_0 > 0$ such that $|f(x)| \leq \alpha |g(x)|$ for all $x > x_0$.

Example 4

Consider $x > 1$*:*

(a) $x + 1 = \mathcal{O}(x)$, *(b)* $2x^4 + 2x^2 - 1 = \mathcal{O}(x^4)$, *(c)* $\frac{1}{x} + \frac{1}{x^2} = \mathcal{O}(\frac{1}{x})$

Gradient

- The **gradient** of a function $f : \mathbb{R}^n \to \mathbb{R}$ at point x is a vector valued function $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$ whose components are the partial derivatives of of f at point x.

Example 5

$$f(x) = x_1 + 2x_2 - x_3 \implies \nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \dots \\ \frac{\partial f(x)}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$$

- Gradient of linear function $f(x) = a^T x + b$ is $\nabla f(x) = a$, where $a, x \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

$$f(x) = a^T x + b = \sum_{i=1}^{n} (a_i x_i + b)$$

$$\implies \frac{\partial \sum_{i=1}^{n} (a_i x_i + b)}{\partial x_k} = a_k \implies \nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \dots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a$$

Gradient of quadratic function

- $f(x) = x^T Ax \implies \nabla f(x) = (A^T + A)x$, for $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{13} \\ a_{21} & a_{22} & \dots & a_{23} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

$$f(x) = x^T Ax = x^T \begin{bmatrix} \sum_{i=1}^n a_{1i} x_i \\ \sum_{i=1}^n a_{2i} x_i \\ \vdots \\ \sum_{i=1}^n a_{ni} x_n \end{bmatrix} = \sum_{j=1}^n \sum_{i=1}^n x_j a_{ji} x_i = \sum_{j=1}^n \left(a_{jj} x_j^2 + \sum_{i \neq j} x_j a_{ji} x_i \right).$$

$$\frac{\partial}{\partial x_k} \left[\sum_{j=1}^n \left(a_{jj} x_j^2 + \sum_{i \neq j} x_j a_{ji} x_i \right) \right] = 2a_{kk} x_k + \sum_{i \neq k} x_i a_{ik} + \sum_{i \neq k} a_{ki} x_i = \sum_{i=1}^n x_i a_{ik} + \sum_{i=1}^n a_{ki} x_i.$$

$$\nabla f(x) = \begin{bmatrix} \sum_{i=1}^n x_i a_{i1} \\ \vdots \\ \sum_{i=1}^n x_i a_{in} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^n a_{1i} x_i \\ \vdots \\ \sum_{i=1}^n a_{ni} x_i \end{bmatrix} = A^T x + Ax = (A^T + A)x.$$

Gradient and Hessian

- $f(x) = x^T Ax + b^T x + c \implies \nabla f(x) = (A^T + A)x + b$
- If matrix A is symmetric ($A = A^T$) then for $f(x) = x^T Ax + b^T x + c \implies \nabla f(x) = 2Ax + b$
- The **Hessian** of a function $f : \mathbb{R}^n \to \mathbb{R}$ with respect to x is a $n \times n$ matrix of partial derivatives:

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

Example 6

$$f(x) = x_1^2 - x_1 x_2 \implies \nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 - x_2 \\ -x_1 \end{bmatrix}$$

$$\implies \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}$$

Taylor Expansion

$$f(x) = f(y) + \frac{f'(y)}{1!}(x-y) + \frac{f''(y)}{2!}(x-y)^2 + \cdots$$

- We can use Taylor expansions to approximate functions.
- A first order approximation uses the first derivative of a function f evaluated at x to approximate the value of f at a neighborhood of x:

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

- Similarly, a second order approximation uses the first and second derivatives of a function f evaluated at x to approximate the value of f at a neighborhood of x:

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

Mean-value Theoem

Theorem 8 (Mean-value)

Let f be a continuously differentiable function and $x, y \in \mathbb{R}^n$. There exist η, ζ in the line segment between x and y such that

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

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(\zeta)(y - x)$$