Linear Algebra Review
(Adapted from Punit Shah's slides)

Introduction to Machine Learning (CSC 311)
Spring 2020

University of Toronto

Basics

- A scalar is a number. $x \in \mathbb{R}$ 0.1

- A vector is a 1-D array of numbers. The set of vectors of length n with real elements is denoted by $\mathbb{R}^n$.
  - Vectos can be multiplied by a scalar.
  - Vector can be added together if dimensions match.

- A matrix is a 2-D array of numbers. The set of $m \times n$ matrices with real elements is denoted by $\mathbb{R}^{m \times n}$.
  - Matrices can be added together or multiplied by a scalar.
  - We can multiply Matrices to a vector if dimensions match.

- In the rest we denote scalars with lowercase letters like a, vectors with bold lowercase $\mathbf{v}$, and matrices with bold uppercase $\mathbf{A}$.

Norms

- Norms measure how "large" a vector is. They can be defined for matrices too.

require: norm $p: V \to \mathbb{R}$

① $p(u+v) \le p(u) + p(v)$

② $p(av) = |a|\, p(v)$

③ $p(v) = 0 \Rightarrow v = 0$

- The ℓ_p-norm for a vector $\mathbf{x}$:

$p \in [1, \infty)$

$$\|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}.$$

- The ℓ_2-norm is known as the Euclidean norm.
- The ℓ_1-norm is known as the Manhattan norm, i.e., $\|\mathbf{x}\|_1 = \sum_i |x_i|$.
- The ℓ_∞ is the max (or supremum) norm, i.e., $\|\mathbf{x}\|_\infty = \max_i |x_i|$.

ℓ_0 - "norm" is how many non-zero elements $[1\ 2\ 0\ 0\ 0 \dots 0]$

↳ not a real norm

- Dot product is defined as $\mathbf{v} \cdot \mathbf{u} = \mathbf{v}^\top \mathbf{u} = \sum_i u_i v_i$.
- The ℓ_2 norm can be written in terms of dot product: $\|\mathbf{u}\|_2 = \sqrt{\mathbf{u}.\mathbf{u}}$.

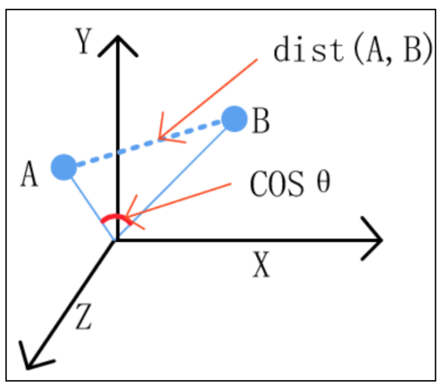
$\sqrt{\sum u_i^2}$

- Dot product of two vectors can be written in terms of their ℓ_2 norms and the angle θ between them:

$$\mathbf{a}^\top \mathbf{b} \ = \|\mathbf{a}\|_2 \, \|\mathbf{b}\|_2 \cos(\theta).$$

$\int \|f\|_2$ for functional

$\int_X |f(x)| \, dx$



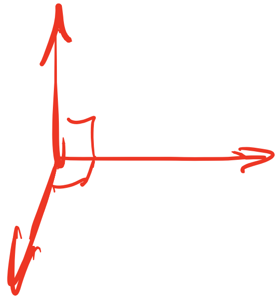
Cosine Similarity

- Cosine between two vectors is a measure of their similarity:

$\in (-\infty, \infty)$

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}.$$

- **Orthogonal Vectors**: Two vectors **a** and **b** are orthogonal to each other if $\mathbf{a} \cdot \mathbf{b} = 0$.



Vector Projection

- Given two vectors $\mathbf{a}$ and $\mathbf{b}$, let $\hat{\mathbf{b}} = \frac{\mathbf{b}}{\|\mathbf{b}\|}$ be the unit vector in the direction of $\mathbf{b}$.
- Then $\mathbf{a}_1 = a_1 \cdot \hat{\mathbf{b}}$ is the orthogonal projection of $\mathbf{a}$ onto a straight line parallel to $\mathbf{b}$, where

$$a_1 = \|\mathbf{a}\| \cos(\theta) = \mathbf{a} \cdot \hat{\mathbf{b}} = \mathbf{a} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|}$$

point p

p

$\mathrm{proj}_S(p)$

S

set S

$\mathrm{proj}_S(p)$

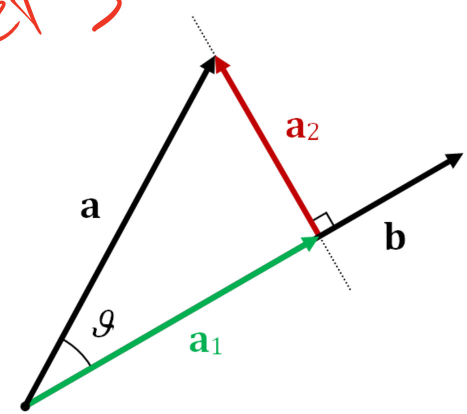


Image taken from wikipedia.

Trace

- Trace is the sum of all the diagonal elements of a matrix, i.e.,

$$\mathrm{Tr}(\mathbf{A}) = \sum_i A_{i,i}.$$

- Cyclic property:

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA}).$$

others...

e.g.,

$\log(\det(A)) = \mathrm{tr}(\log(A))$ A sym

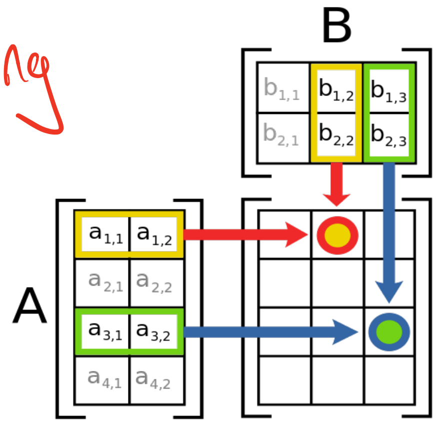
Multiplication

- Matrix-vector multiplication is a linear transformation. In other words,

$$\mathbf{M}(v_1 + a v_2) = \mathbf{M}v_1 + a\mathbf{M}v_2 \implies (\mathbf{M}v)_i = \sum_j M_{i,j} v_j.$$

- Matrix-matrix multiplication is the composition of linear transformations, i.e.,
$(\mathbf{AB})v = \mathbf{A}(\mathbf{B}v) \implies (\mathbf{AB})_{i,j} = \sum_k A_{i,k} B_{k,j}.$

Cost of computing $(AB)v$ can be very different than $A(Bv)$



Cost of $(M_1 M_2)$ is $O(n \cdot m \cdot k)$

if $M_1 \in \mathbb{R}^{n\times m}$, $M_2 \in \mathbb{R}^{m \times k}$

- $\mathbf{I}$ denotes the identity matrix which is a square matrix of zeros with ones along the diagonal. It has the property $\mathbf{IA} = \mathbf{A}$ ($\mathbf{BI} = \mathbf{B}$) and $\mathbf{Iv} = \mathbf{v}$

$I = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$

- A square matrix $\mathbf{A}$ is invertible if $\mathbf{A}^{-1}$ exists such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

inverse very expensive! $O(n^3)$ in general

- Not all non-zero matrices are invertible, e.g., the following matrix is not invertible:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Transposition

- Transposition is an operation on matrices (and vectors) that interchange rows with columns. $(\mathbf{A}^\top)_{i,j} = \mathbf{A}_{j,i}$.
- $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.
- $\mathbf{A}$ is called symmetric when $\mathbf{A} = \mathbf{A}^\top$. $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \to \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$
- $\mathbf{A}$ is called orthogonal when $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$ or $\mathbf{A}^{-1} = \mathbf{A}^\top$.

$AA^T \neq A^TA$ in general!

only for sym A

Diagonal Matrix

- A diagonal matrix has all entries equal to zero except the diagonal entries which might or might not be zero, e.g. identity matrix.
- A square diagonal matrix with diagonal enteries given by entries of vector $\mathbf{v}$ is denoted by $\mathrm{diag}(\mathbf{v})$.
- Multiplying vector $\mathbf{x}$ by a diagonal matrix is efficient:

$$\mathrm{diag}(\mathbf{v})\mathbf{x} \;=\; \mathbf{v} \odot \mathbf{x},$$

  where $\odot$ is the entrywise product.
- Inverting a square diagonal matrix is efficient

$$\mathrm{diag}(\mathbf{v})^{-1} \;=\; \mathrm{diag}\Big([\frac{1}{v_1}, \dots, \frac{1}{v_n}]^\top\Big).$$

Determinant

- Determinant of a square matrix is a mapping to scalars.

$$\det(\mathbf{A}) \quad \text{or} \quad |\mathbf{A}|$$

- Measures how much multiplication by the matrix expands or contracts the space.

- Determinant of product is the product of determinants:

$$\det(\mathbf{AB}) \;=\; \det(\mathbf{A})\det(\mathbf{B})$$

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} \;=\; ad - bc$$

List of Equivalencies

Assuming that $\mathbf{A}$ is a square matrix, the following statements are equivalent

- $\mathbf{Ax} = \mathbf{b}$ has a **unique** solution (for every b with correct dimension).
- $\mathbf{Ax} = \mathbf{0}$ has a unique, trivial solution: $\mathbf{x} = \mathbf{0}$.
- Columns of $\mathbf{A}$ are linearly independent.
- $\mathbf{A}$ is invertible, i.e. $\mathbf{A}^{-1}$ exists.
- $\det(\mathbf{A}) \neq 0$

Zero Determinant

If $\det(\mathbf{A}) = 0$, then:

- $\mathbf{A}$ is linearly dependent.
- $\mathbf{A}\mathbf{x} = \mathbf{b}$ has infinitely many solutions or no solution. These cases correspond to when b is in the span of columns of $\mathbf{A}$ or out of it.
- $\mathbf{A}\mathbf{x} = \mathbf{0}$ has a non-zero solution. (since every scalar multiple of one solution is a solution and there is a non-zero solution we get infinitely many solutions.)

Matrix Decomposition

- We can decompose an integer into its prime factors, e.g., $12 = 2 \times 2 \times 3$.

- Similarly, matrices can be decomposed into product of other matrices.

kronecker product

$$\mathbf{A} = \mathbf{V}\text{diag}(\boldsymbol{\lambda})\mathbf{V}^{-1}$$

$A = B \otimes C$

$A^{-1} = B^{-1} \otimes C^{-1}$

- Examples are Eigendecomposition, SVD, Schur decomposition, LU decomposition, ...

$A = Q \Lambda Q^T$

Eigenvectors

- An eigenvector of a square matrix $\mathbf{A}$ is a nonzero vector $\mathbf{v}$ such that multiplication by $\mathbf{A}$ only changes the scale of $\mathbf{v}$.

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

$\lambda = 0.1$ $\quad v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

- The scalar λ is known as the **eigenvalue**.

v

λv

- If $\mathbf{v}$ is an eigenvector of $\mathbf{A}$, so is any rescaled vector $s\mathbf{v}$. Moreover, $s\mathbf{v}$ still has the same eigenvalue. Thus, we constrain the eigenvector to be of unit length:

$$||\mathbf{v}||_2 = 1$$

Characteristic Polynomial(1)

- Eigenvalue equation of matrix **A**.

$$\begin{aligned} \mathbf{A}\mathbf{v} &= \lambda\mathbf{v} \\ \lambda\mathbf{v} - \mathbf{A}\mathbf{v} &= \mathbf{0} \\ (\lambda\mathbf{I} - \mathbf{A})\mathbf{v} &= \mathbf{0} \end{aligned}$$

- If nonzero solution for **v** exists, then it must be the case that:

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0$$

- Unpacking the determinant as a function of λ, we get:

$$P_A(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = 1 \times \lambda^n + c_{n-1} \times \lambda^{n-1} + \ldots + c_0$$

- This is called the characterisitc polynomial of A.

Characteristic Polynomial(2)

- If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are roots of the characteristic polynomial, they are eigenvalues of $\mathbf{A}$ and we have $P_A(\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i)$.
- $c_{n-1} = -\sum_{i=1}^{n} \lambda_i = -tr(A)$. This means that the sum of eigenvalues equals to the trace of the matrix.
- $c_0 = (-1)^n \prod_{i=1}^{n} \lambda_i = (-1)^n det(\mathbf{A})$. The determinant is equal to the product of eigenvalues.
- Roots might be complex. If a root has multiplicity of $r_j > 1$ (This is called the algebraic dimension of eigenvalue), then the geometric dimension of eigenspace for that eigenvalue might be less than r_j (or equal but never more). But for every eigenvalue, one eigenvector is guaranteed.

Example

- Consider the matrix:

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

- The characteristic polynomial is:

$$\det(\lambda\mathbf{I} - \mathbf{A}) = \det \begin{bmatrix} \lambda - 2 & -1 \\ -1 & \lambda - 2 \end{bmatrix} = 3 - 4\lambda + \lambda^2 = 0$$

- It has roots $\lambda = 1$ and $\lambda = 3$ which are the two eigenvalues of $\mathbf{A}$.
- We can then solve for eigenvectors using $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$:

$$\mathbf{v}_{\lambda=1} = [1, -1]^\top \quad \text{and} \quad \mathbf{v}_{\lambda=3} = [1, 1]^\top$$

Eigendecomposition

- Suppose that $n \times n$ matrix $\mathbf{A}$ has n linearly independent eigenvectors $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$ with eigenvalues $\{\lambda_1, \dots, \lambda_n\}$.

- Concatenate eigenvectors (as columns) to form matrix $\mathbf{V}$. $\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}$

- Concatenate eigenvalues to form vector $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]^\top$.

- The **eigendecomposition** of $\mathbf{A}$ is given by:

$$\mathbf{A}\mathbf{V} = \mathbf{V}\,diag(\lambda) \implies \mathbf{A} = \mathbf{V}\mathrm{diag}(\boldsymbol{\lambda})\mathbf{V}^{-1}$$

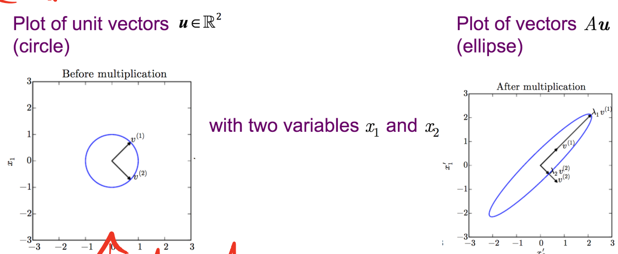
Symmetric Matrices

- Every symmetric (hermitian) matrix of dimension n has a set of (not necessarily unique) n orthogonal eigenvectors. Furthermore, all eigenvalues are real.
- Every real symmetric matrix $\mathbf{A}$ can be decomposed into real-valued eigenvectors and eigenvalues:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$$

- $\mathbf{Q}$ is an orthogonal matrix of the eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues.
- We can think of $\mathbf{A}$ as scaling space by λ_i in direction $\mathbf{v}^{(i)}$.

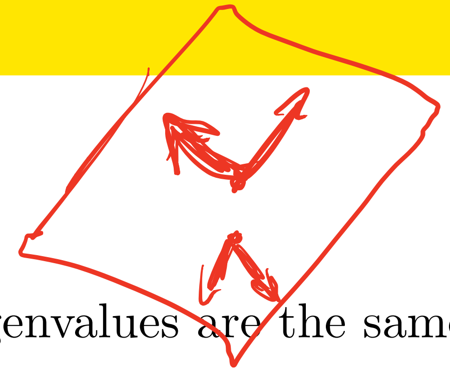
$f(x) = x^T A x \quad x \in \mathbb{R}^{n \times n}$



$\lambda_1 = \lambda_2$

$\lambda_1 < \lambda_2$

Eigendecomposition is not Unique



- Decomposition is not unique when two eigenvalues are the same.
- By convention, order entries of $\boldsymbol{\Lambda}$ in descending order. Then, eigendecomposition is unique if all eigenvalues have multiplicity equal to one.

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

- If any eigenvalue is zero, then the matrix is **singular**. Because if $\mathbf{v}$ is the corresponding eigenvector we have: $\mathbf{A}\mathbf{v} = 0\mathbf{v} = 0$.

Positive Definite Matrix

nxn matrix

C

Semi

PSD

- If a symmetric matrix A has the property:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \quad \text{for any nonzero vector } \mathbf{x}$$

Then A is called **positive definite**.

- If the above inequality is not strict then A is called **positive semidefinite**.

- For positive (semi)definite matrices all eigenvalues are positive(non negative).

Singular Value Decomposition (SVD)

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

- If $\mathbf{A}$ is not square, eigendecomposition is undefined.
- **SVD** is a decomposition of the form $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.
- SVD is more general than eigendecomposition. $U \in \mathbb{R}^{n \times n}$
- Every real matrix has a SVD. $V \in \mathbb{R}^{m \times m}$

SVD Definition (1)

- Write $\mathbf{A}$ as a product of three matrices: $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.
- If $\mathbf{A}$ is $m \times n$, then $\mathbf{U}$ is $m \times m$, $\mathbf{D}$ is $m \times n$, and $\mathbf{V}$ is $n \times n$.
- $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{D}$ is a diagonal matrix (not necessarily square).
- Diagonal entries of $\mathbf{D}$ are called **singular values** of $\mathbf{A}$.
- Columns of $\mathbf{U}$ are the **left singular vectors**, and columns of $\mathbf{V}$ are the **right singular vectors**.

SVD Definition (2)

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

$\cancel{AA^T}$ $\in \mathbb{R}^{m \times m}$

$AA^T \simeq A^2$

- SVD can be interpreted in terms of eigendecompostion.
- Left singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^\top$. $m \times m$
- Right singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A}^\top\mathbf{A}$. $n \times n$
- $m \times n$
- Nonzero singular values of $\mathbf{A}$ are square roots of eigenvalues of $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$.
- Numbers on the diagonal of D are sorted largest to smallest and are non-negative ($\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are semipositive definite.).

Matrix norms

- We may define norms for matrices too. We can either treat a matrix as a vector, and define a norm based on an entrywise norm (example: Frobenius norm). Or we may use a vector norm to "induce" a norm on matrices.
- Frobenius norm:

AB $\|C\|_F$

$$\|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2}.$$

$C = A - B$

- Vector-induced (or operator, or spectral) norm:

$$\|A\|_2 = \sup_{\|x\|_2 = 1} \|Ax\|_2.$$

SVD Optimality

- Given a matrix $\mathbf{A}$, SVD allows us to find its "best" (to be defined) rank-r approximation $\mathbf{A}_r$.
- We can write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ as $\mathbf{A} = \sum_{i=1}^{n} d_i \mathbf{u}_i \mathbf{v}_i^\top$.
- For $r \le n$, construct $\mathbf{A}_r = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^\top$.
- The matrix $\mathbf{A}_r$ is a rank-r approximation of A. Moreover, it is the best approximation of rank r by many norms:
  - When considering the operator (or spectral) norm, it is optimal. This means that $\|A - A_r\|_2 \le \|A - B\|_2$ for any rank r matrix B.
  - When considering Frobenius norm, it is optimal. This means that $\|A - A_r\|_F \le \|A - B\|_F$ for any rank r matrix B. One way to interpret this inequality is that rows (or columns) of A_r are the projection of rows (or columns) of A on the best r dimensional subspace, in the sense that this projection minimizes the sum of squared distances.