

# Linear Algebra Primer

Note: the slides are based on CS131 (Juan Carlos et al) and EE263 (by Stephen Boyd et al) at Stanford. Reorganized, revised, and typed by Hao Su

# Outline

- ▶ Geometry of Linear Algebra
  - ▶ Vector spaces
  - ▶ Basis, dimension
  - ▶ Nullspace, range
- ▶ Spectral Decomposition
  - ▶ Eigenpairs
  - ▶ Spectral theory
- ▶ Singular Value Decomposition
  - ▶ Geometry of linear maps
  - ▶ Singular values, singular vectors
  - ▶ Pseudo-inverse
- ▶ Matrix Calculus
  - ▶ Gradient

# Outline

- ▶ Geometry of Linear Algebra
  - ▶ Vector spaces
  - ▶ Basis, dimension
  - ▶ Nullspace, range
- ▶ Spectral Decomposition
  - ▶ Eigenpairs
  - ▶ Spectral theory
- ▶ Singular Value Decomposition
  - ▶ Geometry of linear maps
  - ▶ Singular values, singular vectors
  - ▶ Pseudo-inverse
- ▶ Matrix Calculus
  - ▶ Gradient

# Vector spaces

a **vector space** or **linear space** (over the reals) consists of

- ▶ a set  $\mathcal{V}$
- ▶ a vector sum  $+: \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$
- ▶ a scalar multiplication:  $\mathbb{R} \times \mathcal{V} \rightarrow \mathcal{V}$
- ▶ a distinguished element  $0 \in \mathcal{V}$

which satisfy a list of properties

# Vector space axioms

- ▶  $x + y = y + x, \forall x, y \in \mathcal{V}$
- ▶  $(x + y) + z = x + (y + z), \forall x, y, z \in \mathcal{V}$
- ▶  $0 + x = x, x \in \mathcal{V}$
- ▶  $\forall x \in \mathcal{V} \quad \exists(-x) \in \mathcal{V} \text{ s.t. } x + (-x) = 0$
- ▶  $(\alpha\beta)x = \alpha(\beta x), \quad \forall \alpha, \beta \in \mathbb{R} \quad \forall x \in \mathcal{V}$
- ▶  $\alpha(x + y) = \alpha x + \alpha y, \quad \forall \alpha \in \mathbb{R} \quad \forall x, y \in \mathcal{V}$
- ▶  $(\alpha + \beta)x = \alpha x + \beta x, \quad \forall \alpha, \beta \in \mathbb{R} \quad \forall x \in \mathcal{V}$
- ▶  $1x = x, \quad \forall x \in \mathcal{V}$

$+$  is commutative

$+$  is associative

0 is additive identity

existence of additive inverse

scalar mult. is associative

right distributive rule

left distributive rule

1 is mult. identity

# Examples

- ▶  $\mathcal{V}_1 = \mathbb{R}^n$ , with standard (componentwise) vector addition and scalar multiplication
- ▶  $\mathcal{V}_2 = \{0\}$  (where  $0 \in \mathbb{R}^n$ )
- ▶  $\mathcal{V}_3 = \text{span}(v_1, v_2, \dots, v_k)$  where
$$\text{span}(v_1, v_2, \dots, v_k) = \{\alpha_1 v_1 + \dots + \alpha_k v_k \mid \alpha_i \in \mathbb{R}\}$$
and  $v_1, \dots, v_k \in \mathbb{R}^n$

# Subspaces

- ▶ a **subspace** of a vector space is a **subset** of a vector space which is itself a vector space
- ▶ roughly speaking, a subspace is closed under vector addition and scalar multiplication
- ▶ examples  $\mathcal{V}_1, \mathcal{V}_3, \mathcal{V}_3$  above are subspaces of  $\mathbb{R}^n$

# Basis and dimension

set of vectors  $\{v_1, v_2, \dots, v_k\}$  is called a **basis** for a vector space  $\mathcal{V}$  if

$$\mathcal{V} = \text{span}(v_1, v_2, \dots, v_k) \\ \text{and} \\ \{v_1, v_2, \dots, v_k\} \text{ is independent}$$

- ▶ equivalently, every  $v \in \mathcal{V}$  **can be uniquely** expressed as

$$v = \alpha_1 v_1 + \dots + \alpha_k v_k$$

- ▶ for a given vector space  $\mathcal{V}$ , the number of vectors in any basis is the same
- ▶ number of vectors in any basis is called the **dimension** of  $\mathcal{V}$ , denoted  **$\dim \mathcal{V}$**



# Nullspace of a matrix

the **nullspace** of  $A \in \mathbb{R}^{m \times n}$  is defined as

$$\mathbf{null}(A) = \{x \in \mathbb{R}^n | Ax = 0\}$$

- ▶  $\mathbf{null}(A)$  is set of vectors mapped to zero by  $y = Ax$
- ▶  $\mathbf{null}(A)$  is set of vectors orthogonal to all rows of  $A$

$\mathbf{null}(A)$  gives **ambiguity** in  $x$  given  $y = Ax$ :

- ▶ if  $y = Ax$  and  $z \in \mathbf{null}(A)$ , then  $y = A(x + z)$
- ▶ conversely, if  $y = Ax$  and  $y = A\tilde{x}$ , then  $\tilde{x} = x + z$  for some  $z \in \mathbf{null}(A)$

$\mathbf{null}(A)$  is also written  $\mathcal{N}(A)$

# Zero nullspace

$A$  is called **one-to-one** if  $0$  is the only element of its null space

$$\text{null}(A) = \{0\}$$

Equivalently,

- ▶  $x$  can always be uniquely determined from  $y = Ax$  (i.e., the linear transformation  $y = Ax$  doesn't 'lose' information)
- ▶ mapping from  $x$  to  $Ax$  is one-to-one: different  $x$ 's map to different  $y$ 's
- ▶ columns of  $A$  are independent (hence, a basis for their span)
- ▶  $A$  has a **left inverse**, i.e., there is a matrix  $B \in \mathbb{R}^{n \times m}$  s.t.  $BA = I$
- ▶  $A^T A$  is invertible

# Range of a matrix

the **range** of  $A \in \mathbb{R}^{m \times n}$  is defined as

$$\text{range}(A) = \{Ax | x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$$

**range**( $A$ ) can be interpreted as

- ▶ the set of vectors that can be 'hit' by linear mapping  $y = Ax$
- ▶ the span of columns of  $A$
- ▶ the set of vectors  $y$  for which  $Ax = y$  has a solution

**range**( $A$ ) is also written  $\mathcal{R}(A)$

# Outline

- ▶ Geometry of Linear Algebra
  - ▶ Vector spaces
  - ▶ Basis, dimension
  - ▶ Nullspace, range
- ▶ Spectral Decomposition
  - ▶ Eigenpairs
  - ▶ Spectral theory
- ▶ Singular Value Decomposition
  - ▶ Geometry of linear maps
  - ▶ Singular values, singular vectors
  - ▶ Pseudo-inverse
- ▶ Matrix Calculus
  - ▶ Gradient

# Eigenvector and Eigenvalue

- ▶ an eigenvector  $x$  of a linear transformation  $A$  is a non-zero vector that, when  $A$  is applied to it, does not change direction

$$Ax = \lambda x, \quad x \neq 0.$$

- ▶ applying  $A$  to the vector only scales the vector by the scalar value  $\lambda$ , called an **eigenvalue**.

# Eigenvector and Eigenvalue

- ▶ we want to find all the eigenvalues of  $A$ :

$$Ax = \lambda x, \quad x \neq 0.$$

- ▶ which can be written as:

$$Ax = (\lambda I)x, \quad x \neq 0.$$

- ▶ therefore:

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

# Eigenvector and Eigenvalue

- ▶ we can solve for eigenvalues by solving :

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

- ▶ above means that  $\lambda I - A$  is not full rank, thus we can instead solve the above equation as:

$$|(\lambda I - A)| = 0.$$

- ▶ this is called **characteristic polynomial** of an  $n \times n$  matrix

# Diagonalization

- ▶ if matrix  $A$  can be diagonalized, that is,

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

- ▶ then:

$$AP = P \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

- ▶ write  $P = [\vec{\alpha}_1, \dots, \vec{\alpha}_n]$ , the above equation can be rewritten as

$$A\vec{\alpha}_i = \lambda_i\vec{\alpha}_i$$



# Diagonalization by Spectral Decomposition

- ▶ here is a sufficient (but not necessary) condition
- ▶ assuming all  $\lambda_i$ 's are unique, by eigenvalue equation:

$$AV = VD$$

$$A = VDV^{-1}$$

- ▶ why?
  - ▶ eigenvectors associated with different eigenvalues are linearly independent, thus  $A$  invertible
  - ▶ in fact, if  $A$  is symmetric,  $V$  would be orthonormal and  $A = VDV^T$

# Diagonalization (Summary)

- ▶ an  $n \times n$  matrix  $A$  is diagonalizable if it has  $n$  linearly independent (in fact, orthogonal) eigenvectors.
- ▶ matrices with  $n$  distinct eigenvalues are diagonalizable

# Symmetric matrices

## Properties

- ▶ for a real symmetric matrix  $A$ , all the eigenvalues are real
- ▶  $A$  is diagonalizable
- ▶ the eigenvectors of  $A$  are orthonormal

$$A = VDV^T$$

# Symmetric matrices

- ▶ therefore

$$x^T A x = x^T V D V^T x = y^T D y = \sum_{i=1}^n \lambda_i y_i^2$$

where  $y = V^T x$

- ▶ so, if we wanted to find the vector  $x$  that

$$\max_{x \in \mathbb{R}^n} x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

# Symmetric matrices

- ▶ therefore

$$x^T A x = x^T V D V^T x = y^T D y = \sum_{i=1}^n \lambda_i y_i^2$$

where  $y = V^T x$

- ▶ so, if we wanted to find the vector  $x$  that

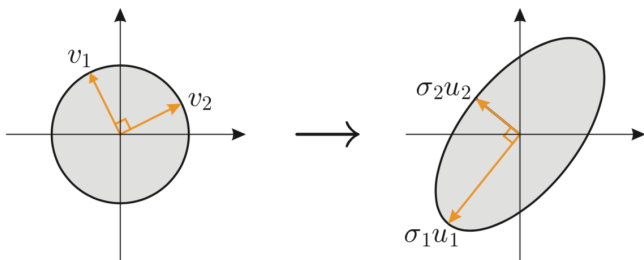
$$\max_{x \in \mathbb{R}^n} x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

is the same as finding the eigenvector that corresponds to the largest eigenvalue.

# Outline

- ▶ Geometry of Linear Algebra
  - ▶ Vector spaces
  - ▶ Basis, dimension
  - ▶ Nullspace, range
- ▶ Spectral Decomposition
  - ▶ Eigenpairs
  - ▶ Spectral theory
- ▶ Singular Value Decomposition
  - ▶ Geometry of linear maps
  - ▶ Singular values, singular vectors
  - ▶ Pseudo-inverse
- ▶ Matrix Calculus
  - ▶ Gradient

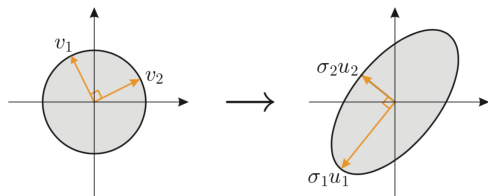
# Geometry of linear maps



every matrix  $A \in \mathbb{R}^{m \times n}$  maps the unit ball in  $\mathbb{R}^n$  to an ellipsoid in  $\mathbb{R}^m$

$$S = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\} \quad AS = \{Ax \mid x \in S\}$$

# Singular values and singular vectors



- ▶ first, assume  $A \in \mathbb{R}^{m \times n}$  is skinny and full rank
- ▶ the numbers  $\sigma_1, \dots, \sigma_n > 0$  are called the **singular values** of  $A$
- ▶ the vectors  $u_1, \dots, u_n$  are called the **left** or **output singular vectors** of  $A$ . These are **unit vectors** along the principal semiaxes of  $AS$
- ▶ the vectors  $v_1, \dots, v_n$  are called the **right** or **input singular vectors** of  $A$ . These map to the principal semiaxes, so that

$$Av_i = \sigma_i u_i$$



# Thin singular value decomposition

$$Av_i = \sigma_i u_i \text{ for } 1 \leq i \leq n$$

For  $A \in \mathbb{R}^{m \times n}$  with  $\text{rank}(A) = n$ , let

$$U = [u_1 \ u_2 \ \dots \ u_n] \quad \Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \quad V = [v_1 \ v_2 \ \dots \ v_n]$$

the above equation is  $AV = U\Sigma$  and since  $V$  is orthogonal

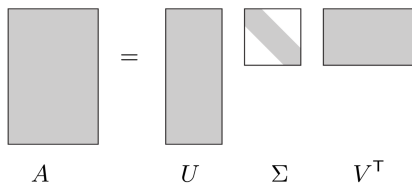
$$A = U\Sigma V^T$$

called the **thin SVD** of  $A$

# Thin SVD

For  $A \in \mathbb{R}^{m \times n}$  with  $\text{rank}(A) = r$ , the **thin SVD** is

$$A = U \Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$$



here

- ▶  $U \in \mathbb{R}^{m \times r}$  has orthonormal columns,
- ▶  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ , where  $\sigma_1 \geq \dots \geq \sigma_r > 0$
- ▶  $V \in \mathbb{R}^{n \times r}$  has orthonormal columns

# SVD and eigenvectors

$$A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^2 V^T$$

hence:

- ▶  $v_i$  are eigenvectors of  $A^T A$  (corresponding to nonzero eigenvalues)
- ▶  $\sigma_i = \sqrt{\lambda_i(A^T A)}$  (and  $\lambda_i(A^T A) = 0$  for  $i > r$ )
- ▶  $\|A\| = \sigma_1$

# SVD and eigenvectors

similarly,

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma^2 U^T$$

hence:

- ▶  $u_i$  are eigenvectors of  $AA^T$  (corresponding to nonzero eigenvalues)
- ▶  $\sigma_i = \sqrt{\lambda_i(AA^T)}$  (and  $\lambda_i(AA^T) = 0$  for  $i > r$ )

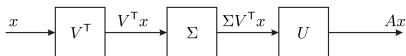
## SVD and range

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

- ▶  $u_1, \dots, u_r$  are orthonormal basis for **range**( $A$ )
- ▶  $v_1, \dots, v_r$  are orthonormal basis for **null**( $A$ )<sup>⊥</sup>

# Interpretations

$$A = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$$



linear mapping  $y = Ax$  can be decomposed as

- ▶ compute coefficients of  $x$  along input directions  $v_1, \dots, v_r$
- ▶ scale coefficients by  $\sigma_i$
- ▶ reconstitute along output directions  $u_1, \dots, u_r$

difference with eigenvalue decomposition for symmetric  $A$ : input and output directions are **different**

# General pseudo-inverse

if  $A \neq 0$  has SVD  $A = U\Sigma V^T$ , the **pseudo-inverse** or **Moore-Penrose inverse** of  $A$  is

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

- ▶ if  $A$  is skinny and full rank,

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

gives the least-squares approximate solution  $x_{ls} = A^\dagger y$

- ▶ if  $A$  is fat and full rank,

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

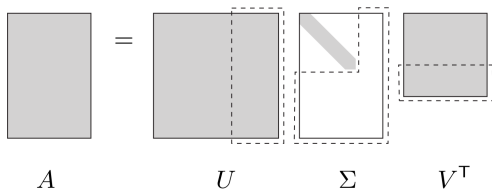
gives the least-norm solution  $x_{ln} = A^\dagger y$

# Full SVD

SVD of  $A \in \mathbb{R}^{m \times n}$  with  $\text{rank}(A)=r$

$$A = U_1 \Sigma_1 V_1^T = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_r^T \end{bmatrix}$$

Add extra columns to  $U$  and  $V$ , and add zero rows/cols to  $\Sigma_1$





# Full SVD

- ▶ find  $U_2 \in \mathbb{R}^{m \times (m-r)}$  such that  $U = [U_1 \ U_2] \in \mathbb{R}^{m \times m}$  is orthogonal
- ▶ find  $V_2 \in \mathbb{R}^{n \times (n-r)}$  such that  $V = [V_1 \ V_2] \in \mathbb{R}^{n \times n}$  is orthogonal
- ▶ add zero rows/cols to  $\Sigma_1$  to form  $\Sigma \in \mathbb{R}^{m \times n}$

$$\Sigma = \left[ \begin{array}{c|c} \Sigma_i & 0_{r \times (n-r)} \\ \hline 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{array} \right]$$

then the full SVD is

$$A = U_1 \Sigma_1 V_1^T = \left[ \begin{array}{c|c} U_1 & U_2 \end{array} \right] \left[ \begin{array}{c|c} \Sigma_i & 0_{r \times (n-r)} \\ \hline 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{array} \right] \left[ \begin{array}{c} V_1^T \\ V_2^T \end{array} \right]$$

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

# Image of unit ball under linear transformation

full SVD:

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

gives interpretation of  $y = Ax$

- ▶ rotate (by  $V^T$ )
- ▶ stretch along axes by  $\sigma_i$  ( $\sigma_i = 0$  for  $i > r$ )
- ▶ zero-pad (if  $m > n$ ) or truncate (if  $m < n$ ) to get  $m$ -vector
- ▶ rotate (by  $U$ )

# Outline

- ▶ Geometry of Linear Algebra
  - ▶ Vector spaces
  - ▶ Basis, dimension
  - ▶ Nullspace, range
- ▶ Spectral Decomposition
  - ▶ Eigenpairs
  - ▶ Spectral theory
- ▶ Singular Value Decomposition
  - ▶ Geometry of linear maps
  - ▶ Singular values, singular vectors
  - ▶ Pseudo-inverse
- ▶ Matrix Calculus
  - ▶ Gradient

# Matrix calculus – the gradient

- ▶ let a function  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  take as input a matrix  $A$  of size  $m \times n$  and returns a real value
- ▶ then the **gradient** of  $f$ :

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

# Matrix calculus – the gradient

## Properties

- ▶  $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$
- ▶ For  $t \in \mathbb{R}$ ,  $\nabla_x(tf(x)) = t\nabla_x f(x)$