The week ahead

- Quiz 0 discussion
  - Focus videos
  - Project seminars
- Office hours start this week!
  - Schedule and sign-up information available on the class website
- Assignment 1 is out! **Due Wednesday, Sep 9th**
  - Early bird special: 2 completed questions by Sep 2nd
- Project seminar 1, **available Thursday, Aug 27th**
- Quiz 1, **Friday, Aug 28th 6:00am until Aug 29th 6:00am**
  - Linear algebra and probability

CS4641B Machine Learning

Lecture 03: Linear algebra

Rodrigo Borela ▸ rborelav@gatech.edu

These slides are based on slides from Le Song, Andres Mendez-Vazquez, Chao Zhang and Mahdi Roozbahani

Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Complementary reading: Bishop PRML - Appendix C

Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Why linear algebra?

- Most data can be represented or stored in matrix-vector form
- Provides compact representation for sets of linear equations

$$\begin{array}{c} 4x_1 - 5x_2 = -13 \\ -2x_1 + 3x_2 = 9 \end{array} \rightarrow \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -13 \\ 9 \end{bmatrix} \rightarrow \mathbf{Ax} = \mathbf{b}$$

- $\mathbf{A} \in \mathbb{R}^{N \times D}$ denotes a matrix with *N* rows and *D* columns, where elements belong to real numbers
- $\mathbf{x} \in \mathbb{R}^{D}$ denotes a vector with *D* real entries. In this class we will follow the convention that that a *D* dimensional vector consists of a matrix with one column and *D* rows.

Linear algebra basics

- Transpose of a matrix results from flipping the rows and the columns. Given $\mathbf{A} \in \mathbb{R}^{N \times D}$, the transpose is $\mathbf{A}^{\mathrm{T}} \in \mathbb{R}^{D \times N}$
- For each element of the matrix, the transpose can be written as $\mathrm{A}_{ij}^{T} = \mathrm{A}_{ji}$
- The following properties of the transposes are easily verified
  - $\left(\mathbf{A}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{A}$
  - $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$
  - $(\mathbf{A} + \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}$
- A square matrix $\mathbf{A} \in \mathbb{R}^{D \times D}$ is symmetric if $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ and it is anti-symmetric if $\mathbf{A} = -\mathbf{A}^{\mathrm{T}}$. Thus each matrix can be written as a sum of symmetric and anti-symmetric matrices

Outline

- Linear Algebra Basics
- **Norms**
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Norms

- Norm of a vector $\|\mathbf{x}\|$ is informally a measure of the length of a vector
- More formally, a norm is any function $f: \mathbb{R}^D \rightarrow \mathbb{R}$ that satisfies:
  - For all $\mathbf{x} \in \mathbb{R}^D, f(\mathbf{x}) \geq \mathbf{0}$ (non-negativity)
  - $f(\mathbf{x}) = \mathbf{0}$ is and only if $x = \mathbf{0}$ (definiteness)
  - For $\mathbf{x} \in \mathbb{R}^D, t \in \mathbb{R}, f(t\mathbf{x}) = |t| f(\mathbf{x})$ (homogeneity)
  - For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D, f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ (triangle inequality)

- Common norms used in machine learning are
  - ℓ_2-norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{D} x_i^2}$
  - ℓ_1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{D} |x_i|$
  - ℓ_∞-norm: $\|\mathbf{x}\|_\infty = \max_i |x_i|$

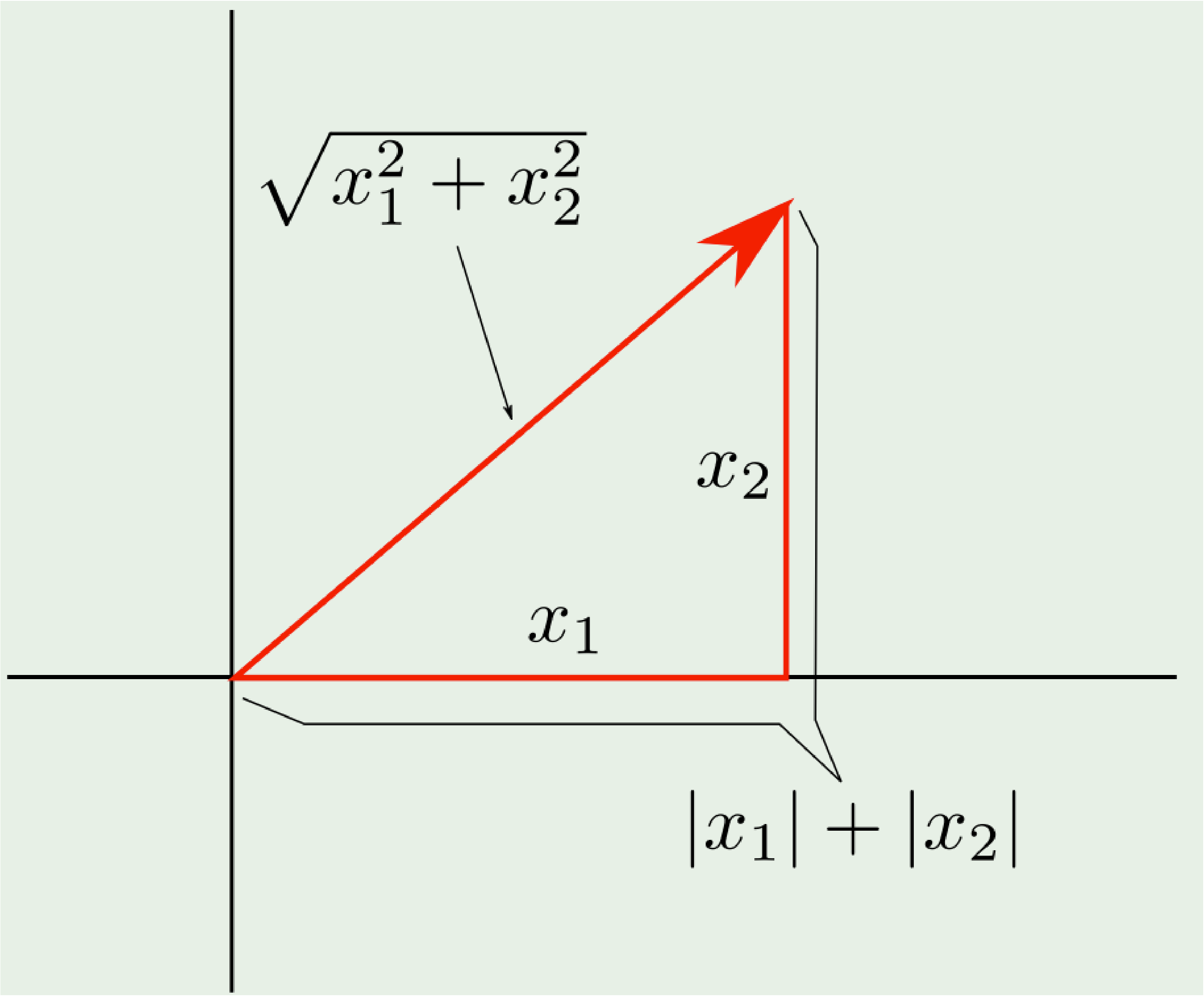
Norms

- All norms presented so far are examples of the family of ℓ_p norms, which are parametrized by a real number $p \geq 1$
  - ℓ_p-norm: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{D} |x_i|^p\right)^{\frac{1}{p}}$

- Norms can be defined for matrices, such as the Frobenius norm
  - $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{D} \mathrm{A}_{\mathrm{ij}}^2} = \sqrt{\mathrm{tr}(\mathbf{A}^{\mathbf{T}}\mathbf{A})}$

Vector norm examples

- ℓ_1-norm and ℓ_2-norm



Special matrices

- The identity matrix, denoted by $\mathbf{I} \in \mathbb{R}^{D \times D}$ is a square matrix with ones on the diagonal and zeros everywhere else
- A diagonal matrix is a matrix where all non-diagonal entries are zero. Typically denoted by $D \in \mathbb{R}^{N \times N}, D = diag(d_1 \quad d_2 \quad \cdots \quad d_N)$
- Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{D}$ are orthogonal if $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y} = \mathbf{0}$ (scalar zero). A square matrix $\mathbf{U} \in \mathbb{R}^{D \times D}$ is orthogonal if all its columns are orthogonal to each other and are normalized
- It follows from orthogonality and normality that
    - $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$
    - $\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$

Outline

- Linear Algebra Basics
- Norms
- **Multiplications**
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Multiplications

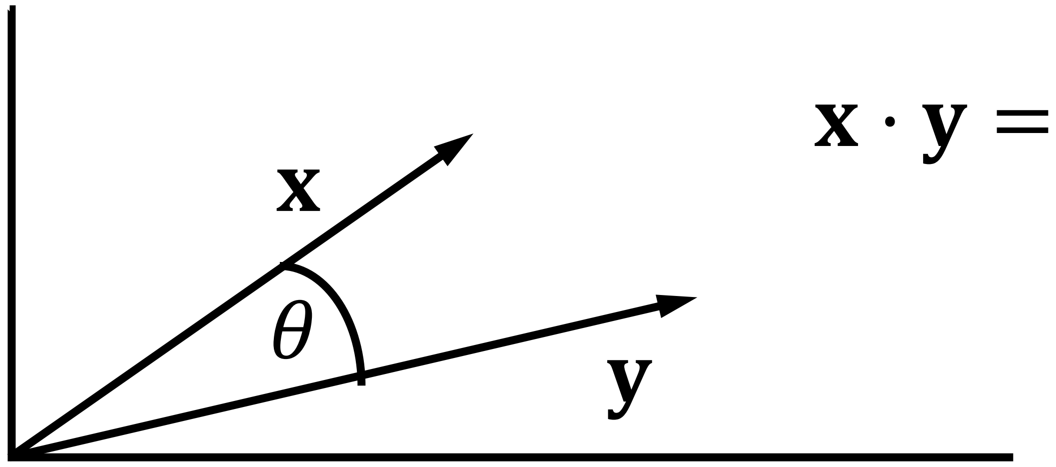
- The product of two matrices $\mathbf{A} \in \mathbb{R}^{N\times D}$ and $\mathbf{B} \in \mathbb{R}^{D\times P}$ is given by $\mathbf{C} \in \mathbb{R}^{N\times P}$, where $C_{ij} = \sum_{k=1}^{D} A_{ik}B_{kj}$

- Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{D}$, the term $\mathbf{x} \cdot \mathbf{y}$ (or $\mathbf{x}^{\mathrm{T}}\mathbf{y}$) is called the inner product or dot product of the vectors, and is a real number given by $\sum_{k=1}^{D} x_i y_i$. For example,

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \sum_{i=1}^{3} x_i y_i$$

Multiplications

- The dot product also has a geometrical interpretation, for vectors in $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ with an angle θ between them:



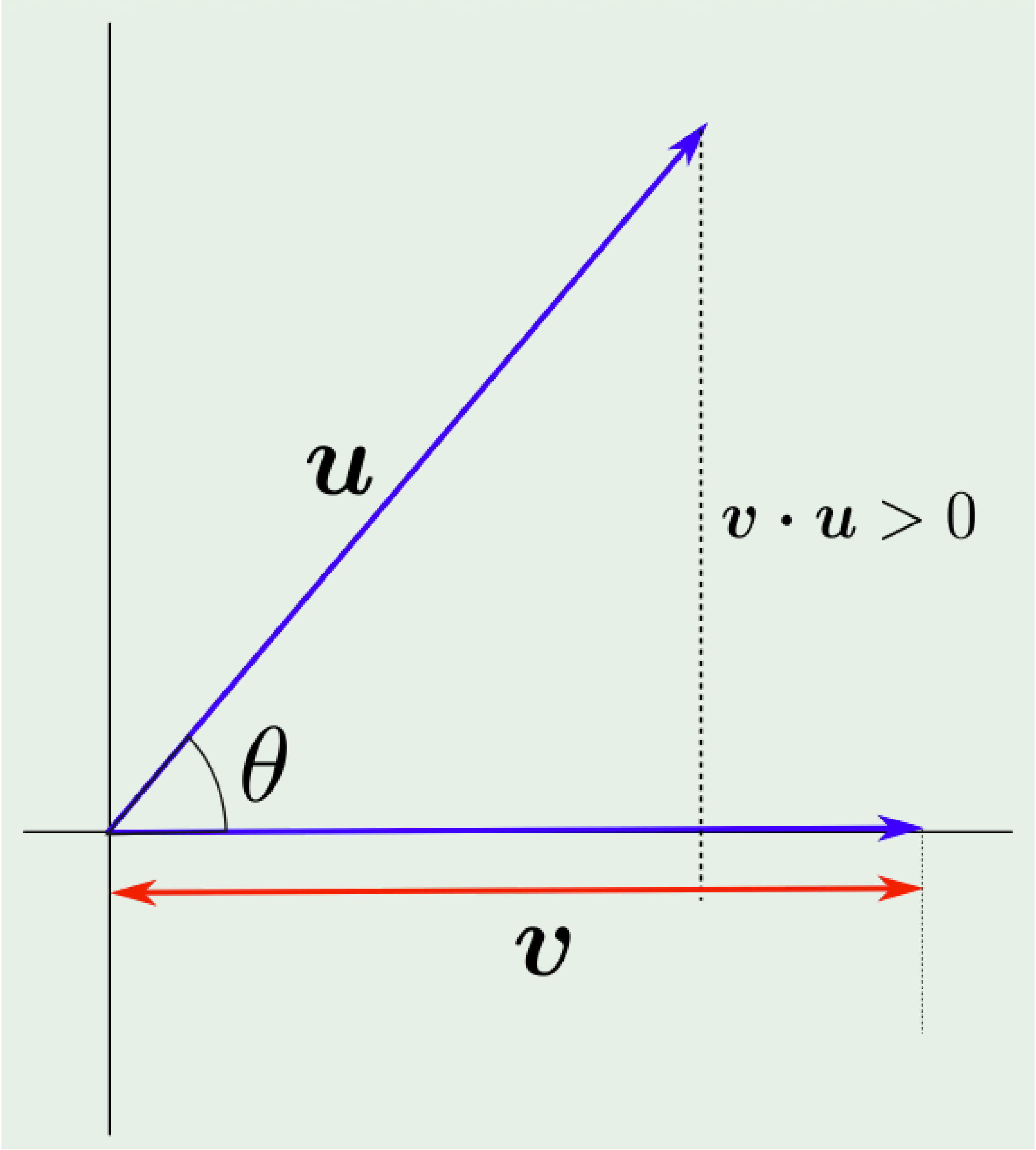
$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta$$

- Given two vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^D$ the term $\mathbf{x}\mathbf{y}^\mathrm{T}$ is called the outer product of the vectors and is a matrix given by $\left(x_i y_j\right)^\mathrm{T} = x_i y_j$. For example,

$$\mathbf{x}\mathbf{y}^\mathrm{T} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & x_1 y_3 \\ x_2 y_1 & x_2 y_2 & x_2 y_3 \\ x_3 y_1 & x_3 y_2 & x_3 y_3 \end{bmatrix}$$

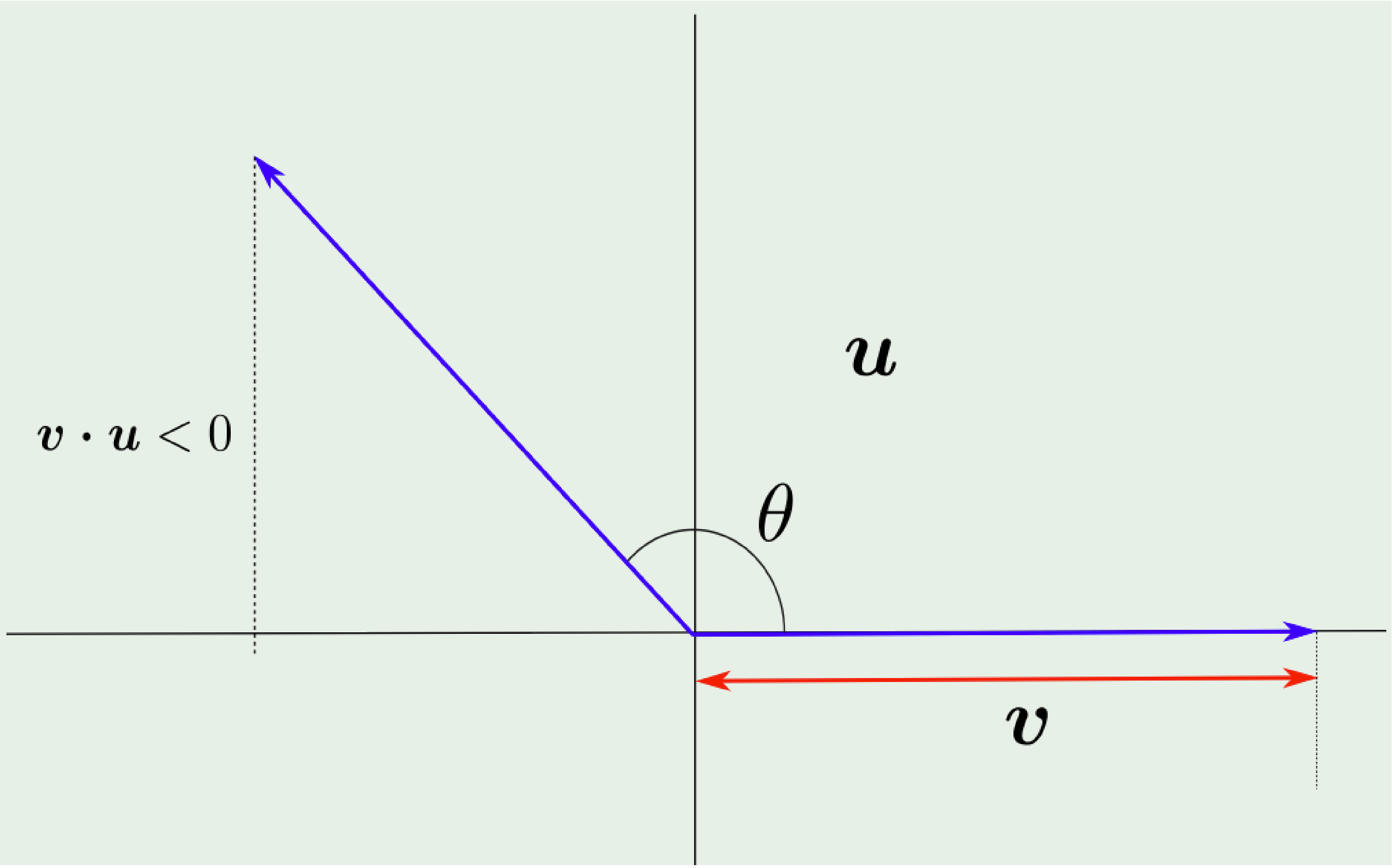
Inner product properties

- The inner product is a measure of correlation between two vectors scaled by the norms of the vectors



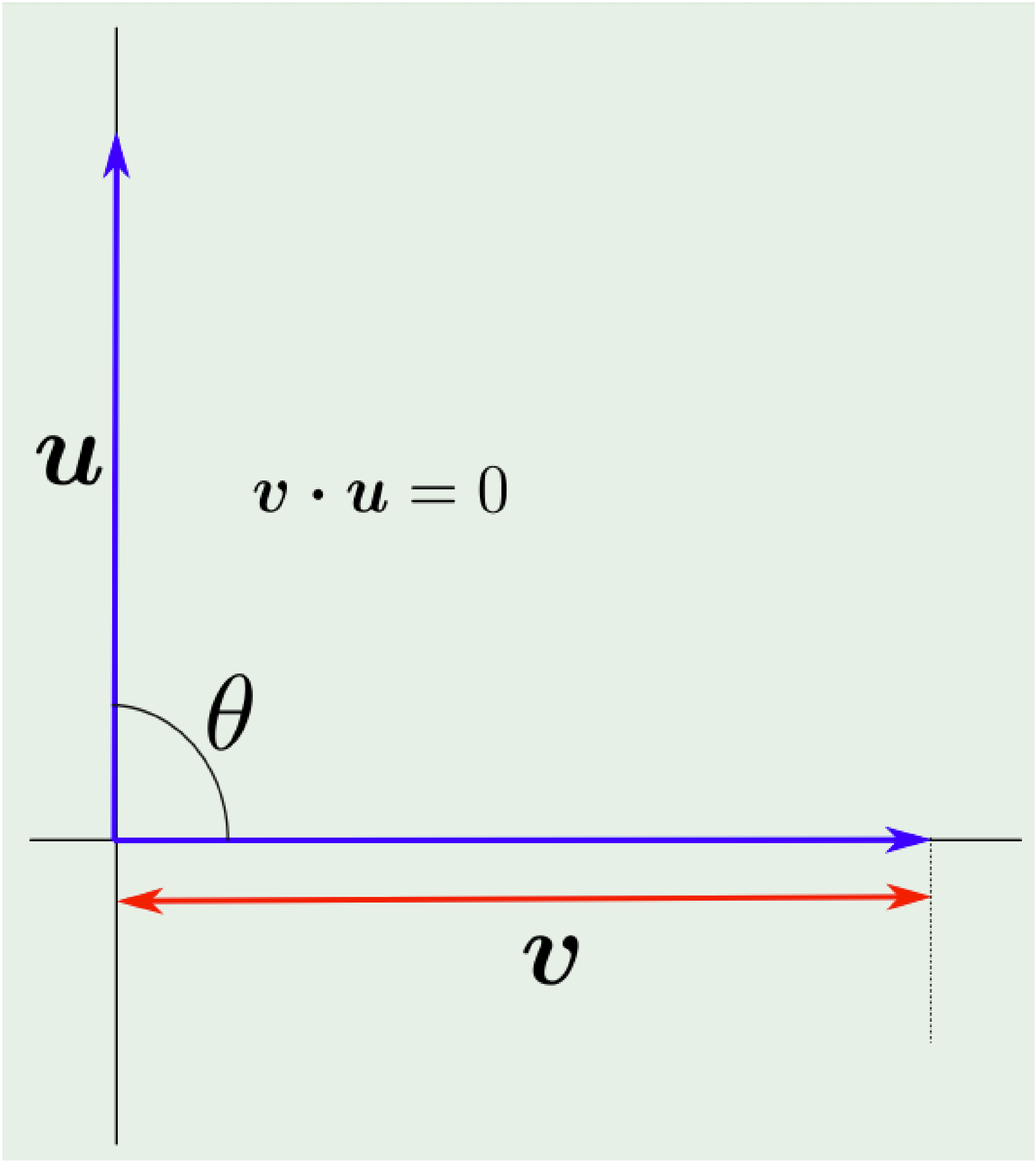
Inner product properties

- The inner product is a measure of correlation between two vectors, scaled by the norms of the vectors



Inner product properties

- The inner product is a measure of correlation between two vectors scaled by the norms of the vectors



Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Linear independence and matrix rank

- A set of vectors $\{\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$ are said to be **linearly independent** if no vector can be represented as a linear combination of the remaining vectors. Thus, if $\mathbf{x}_N = \sum_{i=1}^{N-1} \alpha_i \mathbf{x}_i$ for some scalar values $\alpha_1, \alpha_2, \ldots, \alpha_{N-1} \in \mathbb{R}$ then we say that the vectors are linearly dependent; otherwise the vectors are linearly independent

- The **column rank** of a matrix $\mathbf{A} \in \mathbb{R}^{N \times D}$ is the size of the largest subset of columns of $\mathbf{A}$ that constitute a linearly independent set. **Row rank** of a matrix is defined similarly for rows of a matrix. It can be easily shown that the row and column ranks are equivalent, therefore we shall refer only to the **rank** of a matrix.

Matrix rank examples

- What are the ranks for the following matrices?

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

Matrix inverse

- The inverse of a square matrix $\mathbf{A} \in \mathbb{R}^{N \times D}$ is denoted $\mathbf{A}^{-\mathbf{1}}$ and is the unique matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$
- For some square matrices $\mathbf{A}^{-1}$ may not exist, and we say that $\mathbf{A}$ is singular or non-invertible. In order for $\mathbf{A}$ to have an inverse, $\mathbf{A}$ must be full-rank
- For non-square matrices the inverse, denoted by $\mathbf{A}^{+}$ is given by $\mathbf{A}^{+} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-\mathbf{1}}\mathbf{A}^{\mathrm{T}}$ called the pseudo-inverse

Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Matrix trace

- The trace of a matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ denoted as $\mathbf{tr}(\mathbf{A})$ is the sum of the diagonal elements in the matrix

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{D} \mathrm{A}_{ii}$$

- The trace has the following properties
  - For $\mathbf{A} \in \mathbb{R}^{D\times D}$, $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^{\mathrm{T}})$
  - For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{D\times D}$, $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$
  - For $\mathbf{A} \in \mathbb{R}^{D\times D}, \mathrm{t} \in \mathbb{R}, \mathrm{tr}(t\mathbf{A}) = \mathrm{t} \cdot \mathrm{tr}(\mathbf{A}^{\mathrm{T}})$
  - For $\mathbf{A}, \mathbf{B}, \mathbf{C}$ such that $\mathbf{ABC}$ is a square matrix $\mathbf{tr}(\mathbf{ABC}) = \mathbf{tr}(\mathbf{BCA}) = \mathbf{tr}(\mathbf{CAB})$

- The trace of a matrix helps us easily compute norms and eigenvalues of matrices we will see later

Matrix determinant

- The determinant of a square matrix $\mathbf{A}$, denoted by $|\mathbf{A}|$, is defined as

$$\det(\mathbf{A}) = \sum_{j=1}^{D} (-1)^{i+j} a_{ij} M_{ij}$$

where M_{ij} is determinant of matrix $\mathbf{A}$ without row i and column j.

- For a 2×2 matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $|\mathbf{A}| = ad - bc$

Properties of matrix determinant

- $|\mathbf{A}| = |\mathbf{A}^{\mathrm{T}}|$
- $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$
- $|\mathbf{A}| = 0$ if and only if $\mathbf{A}$ is not invertible
- If $\mathbf{A}$ is invertible, then $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$

Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Eigenvalues and eigenvectors

- Given a square matrix $\mathbf{A} \in \mathbb{R}^{D \times D}$ we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x} \in \mathbb{C}^D$ is an eigenvector if

$$\mathbf{Ax} = \lambda \mathbf{x}, \qquad \mathbf{x} \neq \mathbf{0}$$

- Intuitively this means that upon multiplying the matrix $\mathbf{A}$ with a vector $\mathbf{x}$ we get the same vector, but scaled by the parameter λ

- Geometrically, we are transforming the matrix $\mathbf{A}$ from its original orthonormal basis/coordinates to a new set of orthonormal basis $\mathbf{x}$ with magnitude as λ

Computing eigenvalues and eigenvectors

- We can rewrite the original equation in the following manner

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \qquad \mathbf{x} \neq \mathbf{0}$$

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \qquad \mathbf{x} \neq \mathbf{0}$$

- This is only possible if $(\mathbf{A} - \lambda\mathbf{I})$ is singular, that is $|\mathbf{A} - \lambda\mathbf{I}| = 0$

- Thus eigenvalues and eigenvectors can be computed
  1. Compute the determinant of $\mathbf{A} - \lambda\mathbf{I}$
     - This results in a polynomial of degree *D*
  2. Find the roots of the polynomial equating it to zero
     - The *D* roots are the *D* eigenvalues of $\mathbf{A}$. They make $\mathbf{A} - \lambda\mathbf{I}$ singular
  3. For each eigenvalue λ solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ to find eigenvector $\mathbf{x}$

Eigenvalue example

Matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$

1. Compute the determinant of $\mathbf{A} - \lambda \mathbf{I}$

$$\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 1-\lambda & 2 \\ 3 & -4-\lambda \end{bmatrix}$$

$$|\mathbf{A} - \lambda \mathbf{I}| = (1-\lambda)(-4-\lambda) - 6$$

2. Find the roots of the polynomial equating it to zero

$$|\mathbf{A} - \lambda \mathbf{I}| = 0 \rightarrow (1-\lambda)(-4-\lambda) - 6 = 0 \rightarrow \begin{cases} \lambda_1 = -5 \\ \lambda_2 = 2 \end{cases}$$

Eigenvalue example

3. For each eigenvalue λ solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ to find eigenvector $\mathbf{x}$

$$\begin{bmatrix} 1-\lambda & 2 \\ 3 & -4-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow \begin{cases} (1-\lambda)x_1 + 2x_2 = 0 \\ 3x_1 - (4+\lambda)x_2 = 0 \end{cases}$$

Eigenvector for $\lambda_1 = -5$

$$\begin{cases} 6x_1 + 2x_2 = 0 \\ 3x_1 + x_2 = 0 \end{cases} \rightarrow \mathbf{x}_1 = \begin{bmatrix} 1 \\ -3 \end{bmatrix} \text{ or } \begin{bmatrix} 0.3162 \\ -0.9487 \end{bmatrix}$$

Eigenvector for $\lambda_2 = 2$

$$\begin{cases} -x_1 + 2x_2 = 0 \\ 3x_1 - 6x_2 = 0 \end{cases} \rightarrow \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \text{ or } \begin{bmatrix} 0.8944 \\ 0.4472 \end{bmatrix}$$

Matrix eigen-decomposition

- All the eigenvectors can be written together as $\mathbf{AX} = \mathbf{X\Lambda}$ where the columns of $\mathbf{X}$ are the eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda}$ is a diagonal matrix whose elements are eigenvalues of $\mathbf{A}$
- If the eigenvectors of $\mathbf{A}$ are invertible, then $\mathbf{A} = \mathbf{X\Lambda X}^{-1}$
- There are several properties of eigenvalues and eigenvectors
  - $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{D} \lambda_i$
  - $\det(\mathbf{A}) = \prod_{i=1}^{D} \lambda_i$
  - Rank of $\mathbf{A}$ is the number of non-zero eigenvalues of $\mathbf{A}$
  - If $\mathbf{A}$ is non-singular then $\frac{1}{\lambda_i}$ are the eigenvalues of $\mathbf{A}^{-1}$
  - The eigenvalues of a diagonal matrix are the diagonal elements of the matrix itself

Properties of eigendecomposition

- For a symmetric matrix $\mathbf{A}$ it can be shown that eigenvalues are real and the eigenvectors are orthonormal. Thus it can be represented as $\mathbf{U\Lambda U}^{\mathrm{T}}$

- Considering quadratic form of $\mathbf{A}$,
  - $\mathbf{x}^{\mathrm{T}}\mathbf{Ax} = \mathbf{x}^{\mathrm{T}}\mathbf{U\Lambda U}^{\mathrm{T}}\mathbf{x} = \mathbf{y}^{\mathrm{T}}\mathbf{\Lambda y} = \sum_{i=1}^{N} \lambda_i y_i^2$ where ($\mathbf{y} = \mathbf{U}^{\mathrm{T}}\mathbf{x}$)

- Since y_i^2 is always positive the sign of the expression always depends on λ_i. If $\lambda_i > \mathbf{0}$ then the matrix is $\mathbf{A}$ positive definite, if $\lambda_i \geq \mathbf{0}$ then the matrix $\mathbf{A}$ is positive semidefinite

Covariance matrix

- For a dataset $\mathbf{A}$ we can define the covariance matrix as $\mathbf{C} = \frac{\bar{A}^{\mathrm{T}} \bar{A}}{N}$ for large *N* and $\mathbf{C} = \frac{\bar{A}^{\mathrm{T}} \bar{A}}{N-1}$ for small *N*. $\bar{\mathbf{A}}$ is the matrix $\mathbf{A}$ centered around its mean

- Application example: organizing products in a grocery store

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 2 \\ 4 & 3 & 2 \\ 2 & 1 & 1 \end{bmatrix}, \qquad \boldsymbol{\mu}^T = [2.5 \quad 2.0 \quad 1.5]$$

Covariance matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 2 \\ 4 & 3 & 2 \\ 2 & 1 & 1 \end{bmatrix} \rightarrow \bar{\mathbf{A}} = \begin{bmatrix} 0.5 & 0.0 & -0.5 \\ -1.5 & 0.0 & 0.5 \\ 1.5 & 1.0 & 0.5 \\ -0.5 & -1.0 & -0.5 \end{bmatrix}$$

$$\mathbf{C} = \frac{\bar{A}^{\mathrm{T}}\bar{A}}{N-1} = \frac{1}{4-1}\begin{bmatrix} 0.5 & -1.5 & 1.5 & -0.5 \\ 0.0 & 0.0 & 1.0 & -1.0 \\ -0.5 & 0.5 & 0.5 & -0.5 \end{bmatrix}\begin{bmatrix} 0.5 & 0.0 & -0.5 \\ -1.5 & 0.0 & 0.5 \\ 1.5 & 1.0 & 0.5 \\ -0.5 & -1 & -0.5 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 1.7 & 0.7 & 0.0 \\ 0.7 & 0.7 & 0.3 \\ 00. & 0.3 & 0.3 \end{bmatrix}$$

Correlation matrix

- Given that the different features may not be on the same order of magnitude, the covariance matrix can be standardized based on the standard deviation of the individual features to yield the correlation matrix, such that

$$\mathbf{corr} = \frac{covariance(X, Y)}{\sigma_x \sigma_y}$$

Correlation matrix

- Back to our example…

$$\mathbf{corr} = \begin{bmatrix} \frac{1.7}{1.7} & \frac{0.7}{\sqrt{1.7}\sqrt{0.7}} & \frac{0.0}{\sqrt{1.7}\sqrt{0.3}} \\ \frac{0.7}{\sqrt{1.7}\sqrt{0.7}} & \frac{0.7}{0.7} & \frac{0.3}{\sqrt{0.7}\sqrt{0.3}} \\ \frac{0.0}{\sqrt{1.7}\sqrt{0.3}} & \frac{0.3}{\sqrt{0.7}\sqrt{0.3}} & \frac{0.3}{0.3} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.6 & 0.0 \\ 0.6 & 1.0 & 0.7 \\ 00. & 0.7 & 1.0 \end{bmatrix}$$

Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- **Singular Value Decomposition**
- Matrix Calculus

Singular value decomposition

- $\mathbf{X}_{N\times D}$, *N* is the number of dataset instances, *D* is the dimensionality of each instance (i.e. the number of features) and $\mathbf{X}$ is a centered matrix
- The singular value decomposition is given by

$$\mathbf{X} = \mathbf{U\Sigma V}^{\mathrm{T}}$$

Where

- $\mathbf{U}_{N\times N} \rightarrow unitary\ matrix \rightarrow \ \mathbf{UU}^{\mathrm{T}} = \mathbf{I}$
- $\mathbf{\Sigma}_{N\times D} \rightarrow diagonal\ matrix$
- $\mathbf{V}_{D\times D} \rightarrow unitary\ matrix \rightarrow \mathbf{VV}^{\mathrm{T}} = \mathbf{I}$

$$\underset{U}{\begin{bmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & \ddots & \vdots \\ u_{N1} & \cdots & u_{NN} \end{bmatrix}_{N\times N}} \quad \underset{\Sigma \quad (D < N)}{\begin{bmatrix} s_{11} & \cdots & 0 \\ 0 & \ddots & \vdots \\ \vdots & 0 & s_{DD} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{N\times D}} \quad \underset{V^T}{\begin{bmatrix} v_{11} & \cdots & v_{1D} \\ \vdots & \ddots & \vdots \\ v_{D1} & \cdots & v_{DD} \end{bmatrix}_{D\times D}}$$

Singular value decomposition

$$
\underbrace{\begin{bmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & \ddots & \vdots \\ u_{N1} & \cdots & u_{NN} \end{bmatrix}_{N\times N}}_{U}
\underbrace{\begin{bmatrix} s_{11} & \cdots & 0 \\ 0 & \ddots & \vdots \\ \vdots & 0 & s_{DD} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{N\times D}}_{\Sigma}
\underbrace{\begin{bmatrix} v_{11} & \cdots & v_{1D} \\ \vdots & \ddots & \vdots \\ v_{D1} & \cdots & v_{DD} \end{bmatrix}_{D\times D}}_{V^T}
$$

$$(D < N)$$

Covariance matrix and SVD

- Starting with the covariance matrix expression $\mathbf{C}_{D\times D} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{N}$ and replacing $\mathbf{X} = \mathbf{U\Sigma V}^{\mathrm{T}}$ into the expression for the covariance, we obtain:

$$\mathbf{C} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{N} \rightarrow \mathbf{C} = \frac{\mathbf{V\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U\Sigma V}^{\mathrm{T}}}{N} = \frac{\mathbf{V\Sigma}^{2}\mathbf{V}^{\mathrm{T}}}{N}$$

- Multiplying the result by $\mathbf{V}$ on the right hand side

$$\mathbf{CV} = \mathbf{V}\frac{\Sigma^2}{N}\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\frac{\mathbf{\Sigma}^2}{N}$$

Covariance matrix

- According to the eigendecomposition definition $\mathbf{CV} = \mathbf{V\Lambda}$, therefore the eigenvalues of the covariance matrix are:

$$\lambda_i = \frac{\Sigma_{\mathrm{i}}^2}{N}$$

- λ_i: eigenvalue of $\mathbf{C}$ or covariance matrix
- Σ_i: singular value of $\mathbf{X}$ matrix

So we can directly calculate eigenvalue of a covariance matrix by having the singular values of matrix $\mathbf{X}$

Geometric meaning of SVD

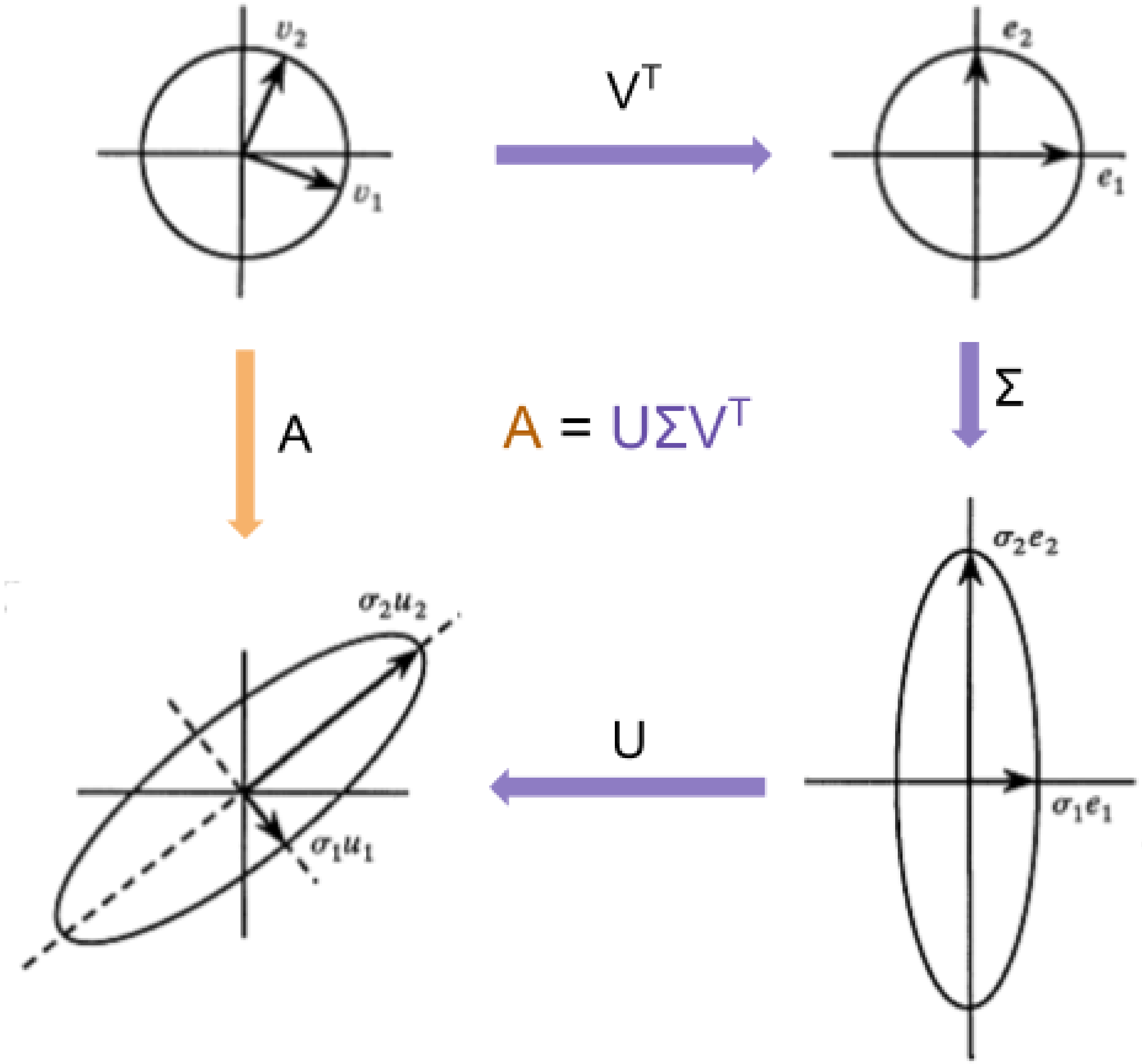


Image credit: Kevin Binz

Outline

- Linear Algebra Basics
- Norms
- Multiplications
- Matrix Inversion
- Trace and Determinant
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- **Matrix Calculus**

Matrix calculus

- For a vector $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$, let $f(\mathbf{x}) = \mathbf{b}^\mathrm{T}\mathbf{x}$, then $\nabla_\mathbf{x}\mathbf{b}^\mathrm{T}\mathbf{x} = \mathbf{b}$
  - $\frac{\partial f(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\sum_{i=1}^{D} b_i x_i = b_k$
- Now for a quadratic function $f(\mathbf{x}) = \mathbf{x}^\mathrm{T}\mathbf{A}\mathbf{x}$, with $\mathbf{A} \in \mathbb{S}^D$, $\nabla_\mathbf{x}\left(\mathbf{x}^\mathrm{T}\mathbf{A}\mathbf{x}\right) = 2\mathbf{A}\mathbf{x}$
  - $\frac{\partial f(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\sum_{i=1}^{D}\sum_{j=1}^{D} A_{ij}x_i x_j = \sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j + 2A_{kk}x_k = 2\sum_{i=1}^{D} A_{ki}x_i$