SDU Summer School

Deep Learning

Summer 2018

Linear Algebra

Scalar and Vector

Scalar

- Single number
- Normally: $x \in \mathbb{R}$ or $x \in \mathbb{N}$

Vector

- An array of numbers
  - Arranged in order
  - Each no. identified by an index
- $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \dots \\ x_d \end{bmatrix}$ and $\boldsymbol{x}^T = [x_1, \dots, x_d]$, $\boldsymbol{x} \in \mathbb{R}^d$
- We think of vectors as points in space
  - Each element gives coordinate along an axis

Matrix

- 2-D array of numbers
- Each element identified by two indices
- Denoted by bold typeface $\boldsymbol{A}$
- Elements indicated as $A_{m,n}$

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- $A_{i:}$ is ith row of $\boldsymbol{A}$, $A_{:j}$ is jth column
- $\boldsymbol{A}$ has m rows and n columns, then $A \in \mathbb{R}^{m \times n}$

Tensor

- Sometimes need an array with more than two axes
- An array arranged on a regular grid with variable number of axes is referred to as a **Tensor**

- We denote a tensor with bold non-italic typeface: $\mathbf{A}$ in comparison to a normal Matrix $\boldsymbol{A}$
- I try to keep the notation consistent, but double-check ☺
- An Element (i, j, k) of tensor denoted by $\mathrm{A}_{i,j,k}$

Transpose of a Matrix

- Mirror image across principal diagonal

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix}, A^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

- Vectors are then either a Matrix with a single column or a single row; For convenience often written in a line: $\boldsymbol{x}^T = [x_1, x_2, x_3]$

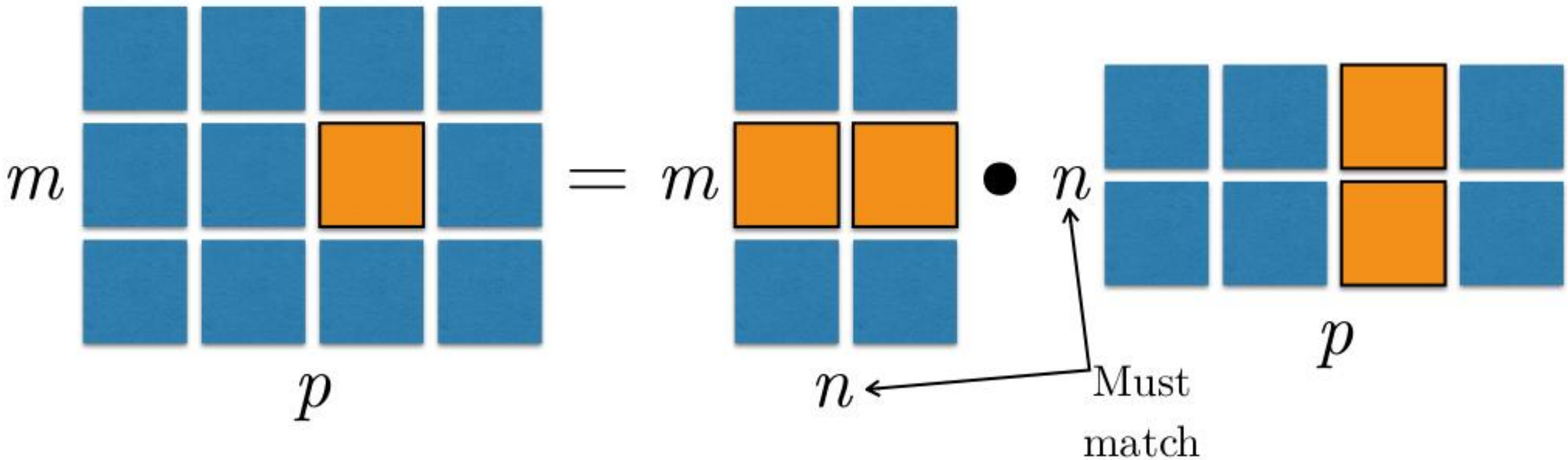
- Scalar is a Matrix with just one element, thus:

$$x^T = x$$

Matrix Product

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}.$$

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}.$$



Matrix Product Properties

- Distributivity over addition:

$$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{AC}$$

- Associativity:

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$$

- Not commutative: $\boldsymbol{AB} = \boldsymbol{BA}$ is not always true

- Dot product between vectors is commutative:

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}$$

- Transpose of a matrix product has a simple form:

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$$

Linear Transformations

$$\boldsymbol{Ax} = \boldsymbol{b}$$

- Where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$
- More Explicitly:

$$\begin{aligned}
&A_{1,1}x_1 + \cdots + A_{1,n}x_1 = b_1 \\
&A_{2,1}x_1 + \cdots + A_{2,n}x_2 = b_2 \\
&\ldots \\
&A_{n,1}x_1 + \cdots + A_{n,n}x_n = b_n
\end{aligned}$$

- In total: n equations with n unknowns
- Can view $\boldsymbol{A}$ as a linear transformation of vector $\boldsymbol{x}$ to vector b

- More interesting: We want to solve for unknowns $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ with $\boldsymbol{A}$ and $\boldsymbol{b}$ being constraints.

Determinant of a Matrix

- Determinant of a square matrix det(A) is a mapping to a scalar
- It is equal to the product of all eigenvalues of the matrix
- Measures how much multiplication by the matrix expands or contracts space

Matrix Decomposition

- Matrices can be decomposed into factors to learn universal properties about them not discernible from their representation
  - E.g., from decomposition of integer into prime factors 12=2x2x3 we can discern that
    - 12 is not divisible by 5 or
    - any multiple of 12 is divisible by 3
- Analogously, a matrix is decomposed into Eigenvalues and Eigenvectors to discern universal properties

How To Calculate a Determinant

- For a 1x1 matrix:

$$\det(a_{11}) = a_{11}$$

- For a 2x2 matrix:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

- For a 3x3 matrix:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} =$$

$$a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} =$$
$$a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

More on Determinants

- The determinant of a matrix can be calculate by multiplying each element of one of its lines by the determinant of a sub-matrix formed by the elements that stay when one suppress the line and column containing this element. One give to the obtained product the sign $(-1)^{i+j}$.

- Determinant exists only for square matrices

- The determinant of a matrix is zero if and only if there exist a linear relationship between the lines or the columns of the matrix

Identify and Inverse

- Matrix inversion is a powerful tool to analytically solve
$$\boldsymbol{Ax} = \boldsymbol{b}$$
- Needs concept of Identity matrix
- Identity matrix does not change value of vector when we multiply the vector by identity matrix
  - Denote identity matrix that preserves n-dimensional vectors as $\boldsymbol{I}_n$
  - For Example:
$$\boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- The inverse of a **square** Matrix $\boldsymbol{A}$ is defined as:
$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

Solving for x

- Using the inverse of $\boldsymbol{A}$, we can easily solve for $\boldsymbol{x}$:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$
$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$
$$\boldsymbol{I}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$
$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

- Numerically unstable, but useful for abstract analysis

- Matrix can’t be inverted if...
  - More rows than columns
  - More columns than rows
  - Redundant rows/columns (“linearly dependent”, “low rank”)

Linear Equations: Closed-Form Solutions

1. Matrix Formulation: $Ax=b$
Solution: $x=A^{-1}b$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad \vdots \qquad\qquad \vdots \quad &\ \ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

2. Gaussian Elimination
followed by back-substitution

$$\begin{aligned} x + 3y - 2z &= 5 \\ 3x + 5y + 6z &= 7 \\ 2x + 4y + 3z &= 8 \end{aligned}$$

L_2-$3L_1 \rightarrow L_2$ $\qquad$ L_3-$2L_1 \rightarrow L_3$ $\qquad$ -$L_2/4 \rightarrow L_2$

$$\left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 3 & 5 & 6 & 7 \\ 2 & 4 & 3 & 8 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & -4 & 12 & -8 \\ 2 & 4 & 3 & 8 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & -4 & 12 & -8 \\ 0 & -2 & 7 & -2 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & 1 & -3 & 2 \\ 0 & -2 & 7 & -2 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 1 & 2 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & 0 & 9 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 0 & -15 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

Disadvantages of Closed Form Solutions

- If $\boldsymbol{A}^{-1}$ exists, the same $\boldsymbol{A}^{-1}$ can be used for any given $\boldsymbol{b}$
  - But $\boldsymbol{A}^{-1}$ cannot be represented with sufficient precision
  - It is not used in practice

- Gaussian elimination also has disadvantages
  - numerical instability (division by small no.)
  - $O(n^3)$ $n \times n$ for matrix

- Software solutions use value of $\boldsymbol{b}$ in finding $\boldsymbol{x}$
  - E.g., difference (derivative) between $\boldsymbol{b}$ and output is used iteratively

Norms

- Used for measuring the size of a vector
- Norms map vectors to non-negative values
- Norm of vector $\boldsymbol{x}$ is distance from origin to $\boldsymbol{x}$

- A norm is any function f that satisfies:

1. $f(x) = 0 \Rightarrow x = 0$
2. $f(x + y) \leq f(x) + f(y)$
3. $\forall \alpha \in \mathbb{R}, f(\alpha x) = |\alpha| f(x)$

L^p Norms

$$L^p = \left(\sum_i |x_i|^p\right)^{1/p}$$

- L^2 Norm
  - Called Euclidean norm, written simply as $||\boldsymbol{x}||$
  - Squared Euclidean norm is same as $\boldsymbol{x}^T\boldsymbol{x}$
- L^1 Norm
  - Useful when 0 and non-zero have to be distinguished (since L^2 increases slowly near origin, e.g., $0.1^2 = 0.01$)
- L^∞ Norm
  - $||\boldsymbol{x}||_\infty = \max |x_i|$
  - Called max norm

Size of a Matrix

- Frobenius norm

$$\|\boldsymbol{A}\| = \left(\sum_{i,j} A_{i,j}^2\right)^{1/2}$$

- It is analogous to L^2 norm of a vector

Angle between Vectors

- Dot product of two vectors can be written in terms of their L^2 norms and angle θ between them

$$\boldsymbol{x}^T\boldsymbol{y} = \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2 \cos\theta$$

Special Matrices & Vectors

- **Unit Vector**
  - A vector with unit norm

$$||\boldsymbol{x}||_2 = 1$$

- **Orthogonal Vectors**
  - A vector x and a vector y are orthogonal to each other if

$$\boldsymbol{x}^T \boldsymbol{y} = 0$$

  - Vectors are at 90 degrees to each other

- **Orthogonal Matrix**
  - A square matrix columns and rows are orthogonal unit vectors

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T$$

Special Matrices & Vectors

- **Diagonal Matrix**
  - Mostly zeros, with non-zero entries in diagonal
  - $diag(\boldsymbol{v})$ is a square diagonal matrix with diagonal elements given by entries of vector $\boldsymbol{v}$
  - Multiplying $diag(\boldsymbol{v})$ by vector $\boldsymbol{x}$ only needs to scale each element x_i by v_i

- **Symmetric Matrix**
  - Is equal to its transpose:

$$\boldsymbol{A} = \boldsymbol{A}^T$$

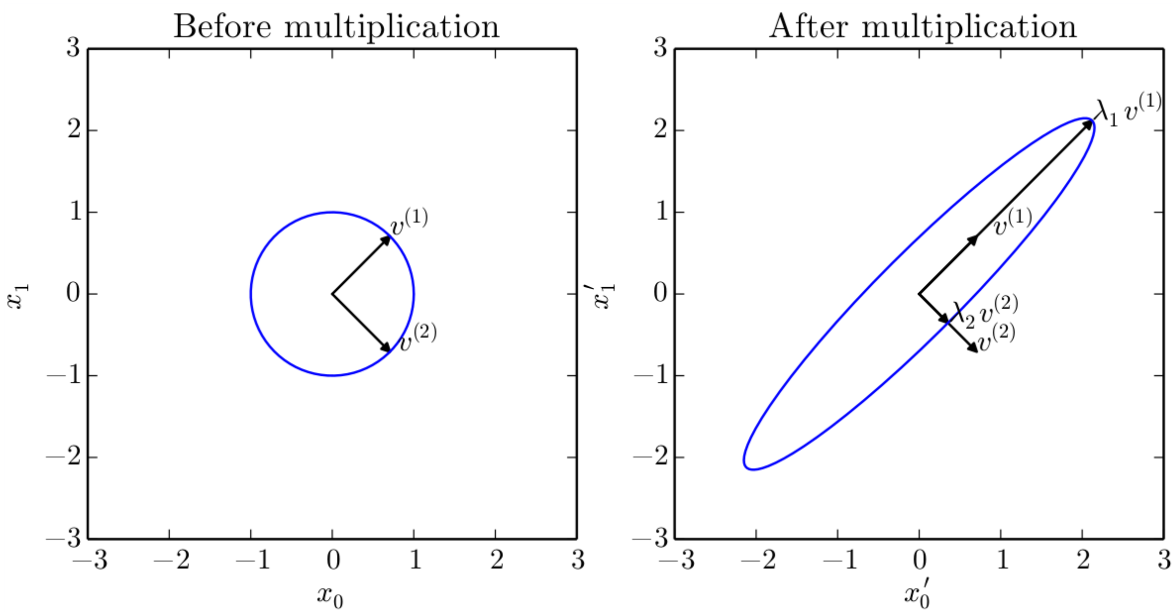
  - E.g., a distance matrix is symmetric with $A_{ij} = A_{ji}$

Eigenvector

- An eigenvector of a square matrix $\boldsymbol{A}$ is a non-zero vector $\boldsymbol{v}$ such that multiplication by $\boldsymbol{A}$ only changes the scale of $\boldsymbol{v}$

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

- The scalar λ is known as eigenvalue
  - If $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{A}$, so is any rescaled vector $s\boldsymbol{v}$.
  - $s\boldsymbol{v}$ still has the same eigen value.
  - Thus, the unit Eigenvector is used



Eigenvalue and Characteristic Polynomial

- Consider $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{w}$

- If $\boldsymbol{v}$ and $\boldsymbol{w}$ are scalar multipliers, then

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$
$$(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v} = 0$$

- This has a non-zero solution if

$$|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$$

- The nulls of the polynomial of degree n are the eigenvalues of $\boldsymbol{A}$

Example

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

- Then, the characteristic polynomial is

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = 3 - 4\lambda + \lambda^2$$

- The polynomial has the roots $\lambda = 1$ and $\lambda = 3$ which are the eigenvalues of $\boldsymbol{A}$.
- The eigenvectors can be found by solving the equation $Av = \lambda v$ using the different eigenvectors:

$$v_{\lambda=1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, v_{\lambda=3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Eigendecomposition

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

- Concatenate eigenvectors to form matrix $\boldsymbol{V}$

- Concatenate eigenvalues to form vector $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]$ (normally in descending order)

- The Eigendecomposition of $\boldsymbol{A}$ is given by

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

Decomposition of Symmetric Matrix

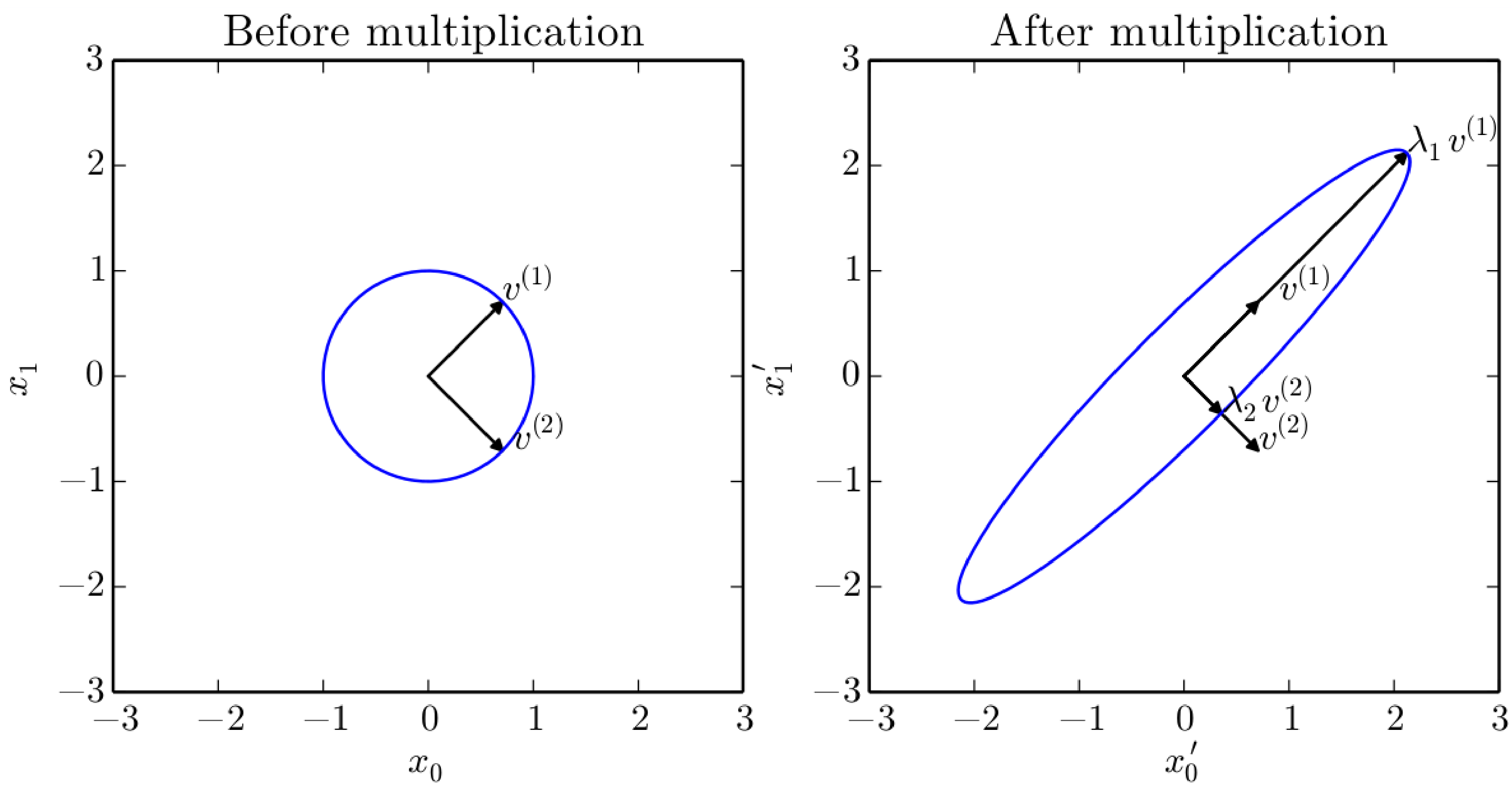
- Every real symmetric matrix $\boldsymbol{A}$ can be decomposed into real-valued eigenvectors and eigenvalues

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$$

- where $\boldsymbol{Q}$ is an orthogonal matrix composed of the eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix of eigenvalues.

- We can think of $\boldsymbol{A}$ as scaling space by λ_i in direction $v^{(i)}$

Effect of Eigenvectors and Eigenvalues



Effect of Eigenvectors and Eigenvalues

- A matrix whose eigenvalues are
  - all positive is called **positive definite**
  - all positive or zero-valued is called **positive semidefinite**.
  - all negative is called **negative definite**
  - all negative or zero-valued is called **negative semidefinite**. Positive

- Semidefinite matrices are interesting because they guarantee that

$$\forall \boldsymbol{x}, \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \ \geq \ 0$$

- Positive definite matrices additionally guarantee that

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0$$

Principal Component Analysis

- see other slide set