M2177.0043 Introduction to Deep Learning

Lecture 2: Linear algebra review[1]

Hyun Oh Song[1]

[1] Dept. of Computer Science and Engineering, Seoul National University

March 19, 2020

[1] Many slides and figures adapted from Stephen Boyd

Last time

- Logistics
- Overview

Outline

Euclidean norm

for $x \in \mathbb{R}^n$ we define the Euclidean norm as

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{x^\mathsf{T} x}$$

$\|x\|$ measures length of vector (from origin)

important properties:

- $\|\alpha x\| = |\alpha| \|x\|$ (homogeneity)
- $\|x + y\| \leqslant \|x\| + \|y\|$ (triangle inequality)
- $\|x\| \geqslant 0$ (nonnegativity)
- $\|x\| = 0 \iff x = 0$ (definiteness)

Inner product

$$\langle x, y \rangle := x_1 y_1 + \cdots + x_n y_n = x^\mathsf{T} y$$

important properties:

- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle x, x \rangle \geqslant 0$
- $\langle x, x \rangle = 0 \iff x = 0$

Cauchy-Schwarz inequality

Theorem 1

For any $a, b \in \mathbb{R}^n,\ |a^\intercal b| \leqslant \|a\| \|b\|$

Proof of triangular inequality

Proof of triangular inequality via Cauchy-Schwarz

$$\|a+b\|^2 = \|a\|^2 + 2a^\intercal b + \|b\|^2 \leqslant \|a\|^2 + 2\|a\|\ \|b\| + \|b\|^2 = (\|a\| + \|b\|)^2$$

Proof.

- It's trivially true if either a or b is 0.
- so assume $\alpha = \|a\|$ and $\beta = \|b\|$ are nonzero
- we have

$$\begin{aligned} 0 &\leqslant \|\beta a - \alpha b\|^2 \\ &= \|\beta a\|^2 - 2(\beta a)^\mathsf{T}(\alpha b) + \|\alpha b\|^2 \\ &= \beta^2\|a\|^2 - 2\beta\alpha(a^\mathsf{T} b) + \alpha^2\|b\|^2 \\ &= 2\|a\|^2\|b\|^2 - 2\|a\|\|b\|(a^\mathsf{T} b) \end{aligned}$$

- divide by $2\|a\|\|b\|$ to get $a^\mathsf{T} b \leqslant \|a\|\|b\|$
- apply to $-a, b$ to get the other half of Cauchy-Schwarz inequality

■

Example: Cauchy-Schwarz

Given $x, y \in \mathbb{R}$, if $2x + 3y = 4$, find the value of x, y s.t. $x^2 + y^2$ has the minimum value.

Example: Cauchy-Schwarz

Given $x, y \in \mathbb{R}$, if $2x + 3y = 4$, find the value of x, y s.t. $x^2 + y^2$ has the minimum value.

solution

Consider two vectors $v_1 = [x, y], v_2 = [2, 3]$. From C-S,
$|2x + 3y| \leqslant \sqrt{x^2 + y^2}\sqrt{13}$.
Given $2x + 3y = 4$, $x^2 + y^2 \geqslant \frac{16}{13}$

Furthermore, C-S holds with equality iff v_1 is parallel to v_2.
Let $x = 2z, y = 3z$, then $13z = 4$, so
$x = 8/13, y = 12/13$.
This is the unique minimizing solution.

Other norms

for $x \in \mathbb{R}^n$ we define

- p-norm
$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$$
- 1-norm
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$
- ∞-norm
$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$
- Hölder's inequality states,
$$|x^\mathsf{T} y| \leqslant \|x\|_p \|y\|_q \text{ for } 1/p + 1/q = 1$$

Outline

Eigenvectors and eigenvalues

$\lambda \in \mathbb{C}$ is an *eigenvalue* of $A \in \mathbb{C}^{n \times n}$ if

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

equivalent to:

- there exists nonzero $v \in \mathbb{C}^n$ s.t. $(\lambda I - A)v = 0$, *i.e.*

  $$Av = \lambda v$$

  any such v is called an *eigenvector* of A associated with eigenvalue λ

- if v is an eigenvector of A with eigenvalue λ, then so is αv, for any $\alpha \in \mathbb{C}, \alpha \neq 0$

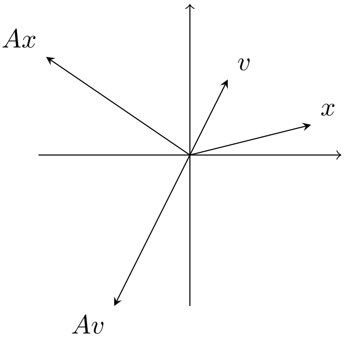
- even when A is real, eigenvalue λ and eigenvector v can be complex

- when A and λ are real, we can always find a real eigenvector v associated with λ

assume A is real from now on..

Scaling interpretation

if v is an eigenvector, effect of A on v is very simple: scaling by λ



(what is λ here?)

Diagonalization I

suppose $v_1, \ldots, v_n$ is a *linearly independent* set of eigenvectors of $A \in \mathbb{R}^{n \times n}$:

$$Av_i = \lambda_i v_i, \;\; i = 1, \ldots, n$$

expressed as

$$A \left[v_1 \cdots v_n \right] = \left[v_1 \cdots v_n \right] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

define $T = [v_1 \cdots v_n]$ and $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, so

$$AT = T\Lambda$$

and finally

$$T^{-1}AT = \Lambda$$

Diagonalization II

- T invertible since $v_1, \ldots, v_n$ linearly independent
- similarity transformation by T diagonalizes A

conversely if there is a $T = [v_1 \cdots v_n]$ s.t.

$$T^{-1}AT = \Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$$

then $AT = T\Lambda$, *i.e.*

$$Av_i = \lambda_i v_i, \ \ i = 1, \ldots, n$$

so $v_1, \ldots, v_n$ is a linearly independent set of eigenvectors of A
we say A *diagonalizable* if

- there exists T s.t. $T^{-1}AT = \Lambda$ is diagonal
- A has a set of linearly independent eigenvectors

Distinct eigenvalues

fact: if A has distinct eigenvalues, *i.e.* $\lambda_i \neq \lambda_j$ for $i \neq j$, then A is diagonalizable

(the converse is false – A can have repeated eigenvalues but still be diagonalizable)

Eigenvectors of symmetric matrices

fact: there is a set of orthonormal eigenvectors of A, *i.e.* $q_1, \dots, q_n$ s.t. $Aq_i = \lambda_i q_i, q_i^\mathsf{T} q_j = \delta_{ij}$ in matrix form: there is an orthogonal Q s.t.

$$Q^{-1}AQ = Q^\mathsf{T} AQ = \Lambda$$

hence we can express A as

$$A = Q\Lambda Q^\mathsf{T} = \sum_{i=1}^{n} \lambda_i q_i q_i^\mathsf{T}$$

Outline

Inequalities for quadratic forms I

suppose $A = A^\mathsf{T}$, $A = Q\Lambda Q^\mathsf{T}$ with eigenvalues sorted so $\lambda_1 \geqslant \cdots \geqslant \lambda_n$

$$\begin{aligned}
x^\mathsf{T} A x &= x^\mathsf{T} Q \Lambda Q^\mathsf{T} x \\
&= (Q^\mathsf{T} x)^\mathsf{T} \Lambda (Q^\mathsf{T} x) \\
&= \sum_{i=1}^{n} \lambda_i (q_i^\mathsf{T} x)^2 \\
&\leqslant \lambda_1 \sum_{i=1}^{n} (q_i^\mathsf{T} x)^2 = \lambda_1 \|Q^\mathsf{T} x\|^2 && (\lambda_1 \geqslant \lambda_i, \forall i) \\
&= \lambda_1 \|x\|^2 && (QQ^\mathsf{T} = I)
\end{aligned}$$

i.e. we have that $x^\mathsf{T} A x \leqslant \lambda_1 x^\mathsf{T} x$

Inequalities for quadratic forms II

similar argument shows $x^\mathsf{T} Ax \geqslant \lambda_n \|x\|^2$, so we have

$$\lambda_n x^\mathsf{T} x \leqslant x^\mathsf{T} Ax \leqslant \lambda_1 x^\mathsf{T} x$$

sometimes λ_1 is called λ_{max}, λ_n is called λ_{min}

note also that

$$q_1^\mathsf{T} Aq_1 = \lambda_1 \|q_1\|^2, \; q_n^\mathsf{T} Aq_n = \lambda_n \|q_n\|^2,$$

so the inequalities are tight

Positive semidefinite and positive definite matrices

suppose $A = A^\mathsf{T} \in \mathbb{R}^{n \times n}$
we say A is *positive semidefinite* if $x^\mathsf{T} A x \geqslant 0$ for all x

- denoted $A \geqslant 0$ and sometimes $A \succeq 0$
- $A \geqslant 0$ iff $\lambda_{\min}(A) \geqslant 0$ *i.e.* all eigenvalues are nonnegative
- not the same as $A_{ij} \geqslant 0$ for all i, j

we say A is *positive definite* if $x^\mathsf{T} A x > 0$ for all $x \neq 0$

- denoted $A > 0$
- $A > 0$ iff $\lambda_{\min} > 0$ *i.e.* all eigenvalues are positive

Matrix norm I

the maximum gain

$$\max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

is called the *matrix norm* or *spectral norm* of A and is denoted $\|A\|$

$$\max_{x\neq 0} \frac{\|Ax\|^2}{\|x\|^2} = \max_{x\neq 0} \frac{x^\mathsf{T} A^\mathsf{T} Ax}{\|x\|^2} = \lambda_{\mathsf{max}}(A^\mathsf{T} A)$$

so we have $\|A\| = \sqrt{\lambda_{\mathsf{max}}(A^\mathsf{T} A)}$ similarly the minimum gain is given by

$$\min_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sqrt{\lambda_{\mathsf{min}}(A^\mathsf{T} A)}$$

Matrix norm II

note that

- $A^\mathsf{T} A \in \mathbb{R}^{n \times n}$ is symmetric and $A^\mathsf{T} A \geqslant 0$ so $\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}} \geqslant 0$
- 'max gain' input direction is $x = q_1$, eigenvector of $A^\mathsf{T} A$ associated with λ_{max}
- 'min gain' input direction is $x = q_n$, eigenvector of $A^\mathsf{T} A$ associated with λ_{min}

Properties of matrix norm

- consistent with vector norm: matrix norm of $a \in \mathbb{R}^{n\times 1}$ is $\sqrt{\lambda_{\max}(a^\intercal a)} = \sqrt{a^\intercal a}$
- for any x, $\|Ax\| \leqslant \|A\|\|x\|$
- scaling: $\|aA\| = |a|\|A\|$
- triangle inequality: $\|A+B\| \leqslant \|A\| + \|B\|$
- definiteness: $\|A\| = 0 \implies A = 0$
- norm of product: $\|AB\| \leqslant \|A\|\|B\|$

Outline

Singular value decomposition I

singular value decomposition (SVD) of A:

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

where

- $A \in \mathbb{R}^{m\times n}, \mathbf{rank}(A) = r$
- $U \in \mathbb{R}^{m\times r}, U^\intercal U = I$, *i.e.* unitary matrix
- $V \in \mathbb{R}^{n\times r}, V^\intercal V = I$, *i.e.* unitary matrix
- $\Sigma = \mathbf{diag}(\sigma_1, \dots, \sigma_r)$, where $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$

Singular value decomposition II

with $U = [u_1 \cdots u_r], V = [v_1 \cdots v_r]$,

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

- σ_i are the (nonzero) *singular values* of A
- v_i are the *right* singular vectors of A
- u_i are the *left* singular vectors of A

Singular value decomposition III

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

hence:

- v_i are eigenvectors of $A^\intercal A$ (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(A^\intercal A)}$ (and $\lambda_i(A^\intercal A) = 0$ for $i > r$)
- $\|A\| = \sigma_1$. In words, the matrix norm is equal to the largest singular value.

Pseudo-inverse

- if $A \neq 0$ has SVD $A = U\Sigma V^\mathsf{T}$,

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

  is the *pseudo-inverse* or *Moore-Penrose inverse* of A.

- If A is skinny and full rank,

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

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

- IF A is fat and full rank,

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

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

Example: Generate correlated normal samples from std samples

Given n independent $N(0,1)$ random variables $z_1, \ldots, z_n$, generate correlated random variables that follow a n-dimensional multivariate normal distribution $X = [X_1, \ldots, X_n]^\intercal \sim N(\mu, \Sigma)$.

Example: Generate correlated normal samples from std samples

Given n independent $N(0,1)$ random variables $z_1, \dots, z_n$, generate correlated random variables that follow a n-dimensional multivariate normal distribution $X = [X_1, \dots, X_n]^\intercal \sim N(\mu, \Sigma)$.

solution

Take eigendecomposition of the target covariance matrix $\Sigma = QDQ^\intercal = (QD^{1/2})(D^{1/2}Q^\intercal) = SS^\intercal$. Then $X = \mu + SZ$ generates the samples from $N(\mu, \Sigma)$.

Proof. Let $\bar{X} = X - \mu$

$$E[\bar{X}\bar{X}^\intercal] = E[SZZ^\intercal S^\intercal] = SE[ZZ^\intercal]S^\intercal = SIS^\intercal = \Sigma$$

Shifting the mean by μ gives the desired samples $X = [X_1, \dots, X_n]$