ECE 595: Machine Learning I
Tutorial 01: Linear Algebra

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

- Norm
- Cauchy Inequality
- Eigen-decomposition
- Positive Definite Matrices
- Matrix Calculus

Reference:

- Gilbert Strang, Linear Algebra and Its Applications, 5th Edition.
- Carl Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, 2000.
- http://cs229.stanford.edu/section/cs229-linalg.pdf
- https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Basic Notation

- Vector: $\boldsymbol{x} \in \mathbb{R}^n$
- Matrix: $\boldsymbol{A} \in \mathbb{R}^{m \times n}$; Entries are a_{ij} or $[\boldsymbol{A}]_{ij}$.
- Transpose:

$$\boldsymbol{A} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix}, \quad \text{and} \quad \boldsymbol{A}^T = \begin{bmatrix} — & \boldsymbol{a}_1^T & — \\ — & \boldsymbol{a}_2^T & — \\ & \vdots & \\ — & \boldsymbol{a}_n^T & — \end{bmatrix}.$$

- Column: $\boldsymbol{a}_i$ is the i-th column of $\boldsymbol{A}$
- Identity matrix $\boldsymbol{I}$
- All-one vector $\mathbf{1}$ and all-zero vector $\mathbf{0}$
- Standard basis $\boldsymbol{e}_i$.

Norm

- $\|\boldsymbol{x}\|$ is the *length* of $\boldsymbol{x}$.
- We use ℓ_p-norm

Definition

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \tag{1}$$

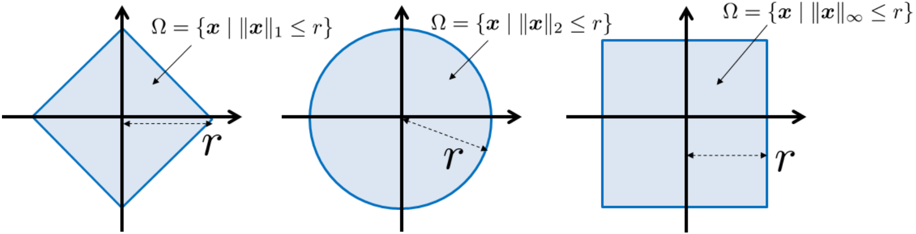


Figure: The shapes of Ω defined using different ℓ_p-norms.

The ℓ_2-norm

Also called the **Euclidean norm**:

Definition

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{2}$$

- The set $\Omega = \{\mathbf{x} \mid \|\mathbf{x}\|_2 \leq r\}$ defines a circle:

$$\Omega = \{\mathbf{x} \mid \|\mathbf{x}\|_2 \leq r\} = \{(x_1, x_2) \mid x_1^2 + x_2^2 \leq r^2\}.$$

- $f(\mathbf{x}) = \|\mathbf{x}\|_2$ is not the same as $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$.
- Triangle inequality holds:

$$\|\mathbf{x} + \mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2.$$

The ℓ_1-norm

Definition

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|. \tag{3}$$

- The set $\Omega = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \leq r\}$ is a diamond.
- $\|\boldsymbol{x}\|_1 = r$ is equivalent to

$$\|\boldsymbol{x}\|_1 = |x_1| + |x_2| = r.$$

- If $x_1 > 0$ and $x_2 > 0$, then the sign has no effect. This is a line in the 1st quadrant.
- MATLAB: `norm(x, 1)`
- Python: `numpy.linalg.norm(x, ord=1)`

Sparsity

- Roughly speaking, a vector $\boldsymbol{x}$ is sparse if it contains many zeros.
- $\|\cdot\|_1$ promotes sparsity:
- If $\boldsymbol{x}$ is the parameter vector, minimizing a cost function over a constraint $\|\boldsymbol{x}\|_1 \le \tau$ leads to a sparse $\boldsymbol{x}$.

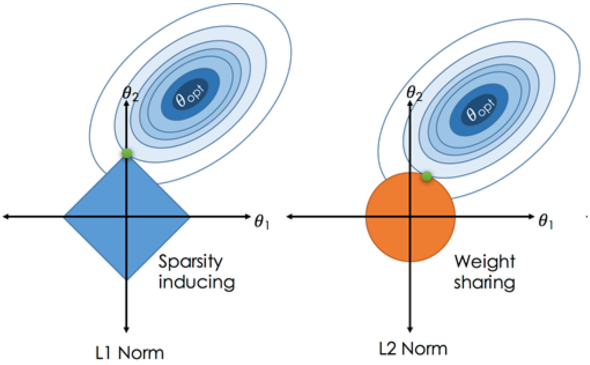


Figure: ℓ_1-norm promotes sparsity whereas ℓ_2-norm leads to weight sharing. Figure is taken from http://www.ds100.org/

The ℓ_∞-norm

Definition

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|. \tag{4}$$

- A hand-waving argument: If we set $p \to \infty$

$$\lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{5}$$

  then the largest term $|x_i|^p$ will dominate eventually.
- The set $\Omega = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_\infty \leq r\}$ is a square
- We can show the following inequality

$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1, \tag{6}$$

  and $\Omega_1 \subseteq \Omega_2 \subseteq \Omega_\infty$.

Holder's Inequality and Cauchy-Schwarz Inequality

Theorem (Holder's Inequality)

Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$. Then,

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q \tag{7}$$

for any p and q such that $\frac{1}{p} + \frac{1}{q} = 1$, where $p \geq 1$. Equality holds if and only if $|x_i|^p = \alpha |y_i|^q$ for some scalar α and for all $i = 1, \ldots, n$.

Corollary (Cauchy-Schwarz Inequality)

Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$. Then,

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2, \tag{8}$$

where the equality holds if and only if $\mathbf{x} = \alpha\mathbf{y}$ for some scalar α.

Geometry of Cauchy-Schwarz Inequality

- $\boldsymbol{x}^T\boldsymbol{y}/(\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2)$ defines the cosine angle between the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.
- Cosine is always less than 1. So is $\boldsymbol{x}^T\boldsymbol{y}/(\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2)$.
- The equality holds if and only if the two vectors are parallel.

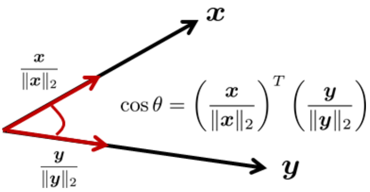


Figure: Pictorial interpretation of Cauchy-Schwarz inequality. The inner product defines the cosine angle, which by definition must be less than 1.

Eigenvalue and Eigenvector

Definition

Given a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the vector $\boldsymbol{u} \in \mathbb{R}^n$ (with $\boldsymbol{u} \neq \mathbf{0}$) is called the **eigenvector** of $\boldsymbol{A}$ if

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}, \tag{9}$$

for some $\lambda \in \mathbb{R}$. The scalar λ is called the **eigenvalue** associated with $\boldsymbol{u}$.

The following conditions are equivalent

- There exists $\boldsymbol{u} \neq 0$ such that $\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}$;
- There exists $\boldsymbol{u} \neq 0$ such that $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{u} = \mathbf{0}$;
- $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is not invertible;
- $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$;

Exercise: Prove these results.

Eigen-Decomposition for Symmetric Matrices

- Not all matrices have eigenvalues.
- For example, the matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ does not have an eigenvalue.
- If $\boldsymbol{A}$ is symmetric, then eigenvalues exist and are real.

Theorem

If $\boldsymbol{A}$ is symmetric, then all the eigenvalues are real, and there exists $\boldsymbol{U}$ such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$:

$$\boldsymbol{A} = \underbrace{\begin{bmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_n \\ | & | & & | \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} — & \boldsymbol{u}_1^T & — \\ — & \boldsymbol{u}_2^T & — \\ & \vdots & \\ — & \boldsymbol{u}_n^T & — \end{bmatrix}}_{\boldsymbol{U}^T}. \tag{10}$$

Basis Representation

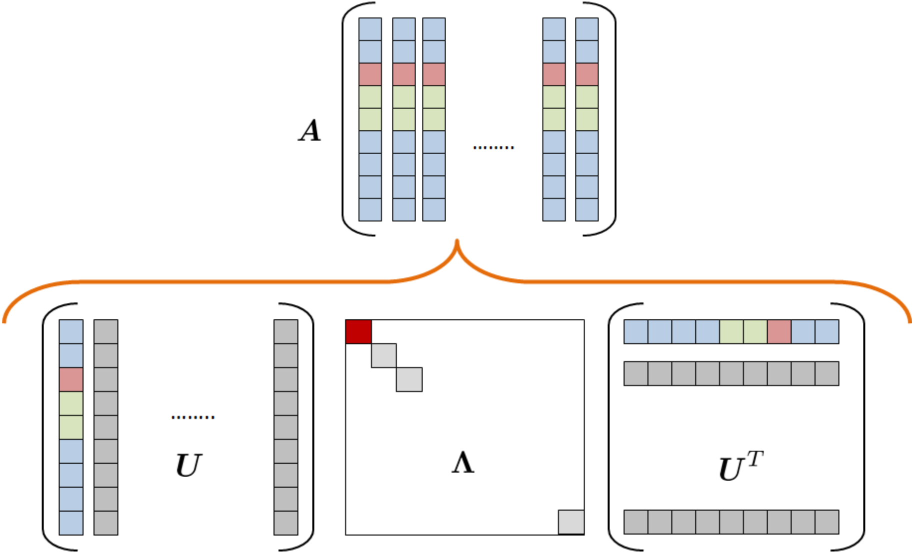
```
% MATLAB Code:
A = randn(100,100);
A = (A + A')/2;      % symmetrize because A is not symmetric
[U,S] = eig(A);      % eigen-decomposition
s = diag(S);         % extract eigen-value
```

- Eigenvectors satisfy $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$.
- This is equivalent to $\boldsymbol{u}_i^T\boldsymbol{u}_j = 1$ if $i = j$ and $\boldsymbol{u}_i^T\boldsymbol{u}_j = 0$ if $i \neq j$.
- $\boldsymbol{U}$ can be served as basis

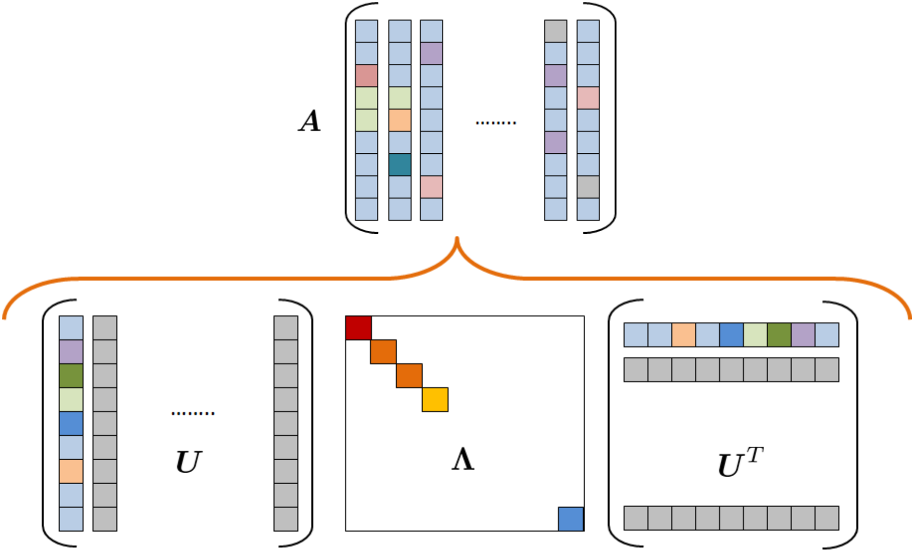
$$\boldsymbol{x} = \sum_{j=1}^{n} \alpha_j \boldsymbol{u}_j, \tag{11}$$

- $\alpha_j = \boldsymbol{u}_j^T\boldsymbol{x}$ is called the **basis coefficient**.

If Columns are Similar:



If Columns are Different:



Positive Semi-Definite

Definition (Positive Semi-Definite)

A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive semi-definite if

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0 \tag{12}$$

for any $\boldsymbol{x} \in \mathbb{R}^n$. $\boldsymbol{A}$ is positive definite if $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$.

Theorem

A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive semi-definite if and only if

$$\lambda_i(\boldsymbol{A}) \geq 0 \tag{13}$$

for all $i = 1, \ldots, n$, where $\lambda_i(\boldsymbol{A})$ denotes the i-th eigenvalue of $\boldsymbol{A}$.

Positive Semi-Definite

Proof.

By definition of eigenvalue and eigenvector, we have that $\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ where λ_i is the eigenvalue and $\boldsymbol{u}_i$ is the corresponding eigenvector. If $\boldsymbol{A}$ is positive semi-definite then $\boldsymbol{u}_i^T \boldsymbol{A}\boldsymbol{u}_i \geq 0$ since $\boldsymbol{u}_i$ is a particular vector in $\mathbb{R}^n$. So we have $0 \leq \boldsymbol{u}_i^T \boldsymbol{A}\boldsymbol{u}_i = \lambda \|\boldsymbol{u}_i\|^2$ and hence $\lambda_i \geq 0$. Conversely, if $\lambda_i \geq 0$ for all i, then since $\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ we can conclude that $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T \left(\sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T\right) \boldsymbol{x} = \sum_{i=1}^n \lambda_i (\boldsymbol{u}_i^T \boldsymbol{x})^2 \geq 0$. □

Corollary

If a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ *is positive definite (not semi-definite), then* $\boldsymbol{A}$ *must be invertible, i.e., there exist* $\boldsymbol{A}^{-1} \in \mathbb{R}^{n \times n}$ *such that*

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}. \tag{14}$$

Matrix Calculus

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a scalar field. The gradient of f with respect to $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}. \tag{15}$$

Example 1. $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$. In this case, the gradient is

$$\nabla_{\boldsymbol{x}}\left(\boldsymbol{a}^T\boldsymbol{x}\right) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_{j=1}^n a_j x_j \\ \vdots \\ \frac{\partial}{\partial x_n} \sum_{j=1}^n a_j x_j \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \boldsymbol{a}. \tag{16}$$

More Examples

Example 2. $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}$. Then,

$$\nabla_{\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_{i,j=1}^n a_{ij} x_i x_j \\ \vdots \\ \frac{\partial}{\partial x_n} \sum_{i,j=1}^n a_{ij} x_i x_j \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^n a_{1,j} x_j \\ \vdots \\ \sum_{j=1}^n a_{n,j} x_j \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^n a_{i,1} x_i \\ \vdots \\ \sum_{i=1}^n a_{i,n} x_i \end{bmatrix} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^T\boldsymbol{x}$$

If $\boldsymbol{A}$ is symmetric so that $\boldsymbol{A} = \boldsymbol{A}^T$ then $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}$

More Examples

Example 3. $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2$. The gradient is

$$
\begin{aligned}
\nabla_{\boldsymbol{x}}\left(\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2\right) &= \nabla_{\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{y}\right) \\
&= \nabla_{\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}\right) - 2\nabla_{\boldsymbol{x}}\left(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}\right) + \nabla_{\boldsymbol{x}}\left(\boldsymbol{y}^T\boldsymbol{y}\right) \\
&= 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^T\boldsymbol{y} + 0 = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}).
\end{aligned}
$$

Definition

The Hessian of f with respect to $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$
\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix}. \tag{17}
$$