Lecture 2: Linear Algebra Review

Modified from notes of cs229

Ailin Zhang

Agenda

- Basic Concepts and Notation
- Matrix Multiplication
- Operations and Properties
- Matrix Calculus
  - Strategy: Whenever possible, you can think of matrices or vectors as scalars or numbers in your calculations. You only need to take care of matching dimensionalities. We shall define expectations, variances, and derivatives for vectors as a matter of conveniently packaging their elements, so that we can avoid subindices in our calculations.

Basic Notation

- By $x \in \mathbb{R}^n$, we denote a vector with n entries.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}$$

The Identity Matrix

The identity matrix, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

It has the property that for all $A \in \mathbb{R}^{m \times n}$,

$$AI = A = IA$$

Diagonal matrices

A diagonal matrix is a matrix where all non-diagonal elements are 0.

This is typically denoted $D = \text{diag}(d_1, d_2, \ldots, d_n)$, with

$$D_{ij} = \begin{cases} d_i, & i = j \\ 0, & i \neq j \end{cases}$$

Clearly, $I = \text{diag}(1, 1, ..., 1)$.

Matrix Multiplication

- Vector-Vector inner product (dot product)

  $x, y \in \mathbb{R}^n$

  $$< x, y >= x^T y \in \mathbb{R} = \sum_{i=1}^{n} x_i y_i$$

- Vector-Vector outer product

  $x \in \mathbb{R}^m, y \in \mathbb{R}^n$

  $$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

Matrix Multiplication

- Matrix-Vector Product

  If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}$$

Matrix Multiplication

- Matrix-Vector Product

  If we write A by columns, then we have:

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \\ a_1 \\ \\ \end{bmatrix} x_1 + \begin{bmatrix} \\ a_2 \\ \\ \end{bmatrix} x_2 + \ldots + \begin{bmatrix} \\ a_n \\ \\ \end{bmatrix} x_n$$

  y is a linear combination of the columns of A.

Matrix Multiplication

Discussion: how about multiply on the left by a row vector?

$x^T A =$

Matrix Multiplication

- Matrix-Matrix Multiplication

$$AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \dots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \dots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \dots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \dots & a_m^T b_p \end{bmatrix}$$

Matrix Multiplication

- Matrix-Matrix Multiplication

$$AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_p \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^T & — \\ — & b_2^T & — \\ & \vdots & \\ — & b_p^T & — \end{bmatrix} = \sum_{i=1}^{p} a_i b_i^T$$

- Other views

Matrix-Matrix Multiplication (properties)

- Associative: $(AB)C = A(BC)$.
- Distributive: $A(B + C) = AB + AC$.
- In general, not commutative; that is, it can be the case that $AB \neq BA$.

Operations and Properties - Transpose

The ***transpose*** of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}.$$

The following properties of transposes are easily verified:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

Operations and Properties - Trace

The ***trace*** of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}A$, is the sum of diagonal elements in the matrix:

$$\mathrm{tr}A = \sum_{i=1}^{n} A_{ii}.$$

The trace has the following properties:

- For $A \in \mathbb{R}^{n \times n}$, $\mathrm{tr}A = \mathrm{tr}A^T$.
- For $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(A + B) = \mathrm{tr}A + \mathrm{tr}B$.
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$, $\mathrm{tr}(tA) = t\ \mathrm{tr}A$.
- For A, B such that AB is square, $\mathrm{tr}AB = \mathrm{tr}BA$.
- For A, B, C such that ABC is square, $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$, and so on for the product of more matrices.

Operations and Properties - Norms

A norm of a vector $||x||$ is informally a measure of the "length" of the vector.

More formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negative).
2. $f(x) = 0$ if and only if $x = 0$ (definite).
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t|f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^n$,$f(x + y) \leq f(x) + f(y)$ (triangle inequality).

l_p norm:

$$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$$

Example of Norms:

The commonly-used Euclidean or ℓ_2 norm,

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

The ℓ_1 norm,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

The ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

Linear Independence

A set of vectors $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^m$ is said to be **(linearly) dependent** if one vector belonging to the set can be represented as a linear combination of the remaining vectors; that is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$; otherwise, the vectors are **(linearly) independent**.

Rank of a Matrix

- The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.

Rank of a Matrix

- The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.
- The row rank is the largest number of rows of A that constitute a linearly independent set.

Rank of a Matrix

- The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.
- The row rank is the largest number of rows of A that constitute a linearly independent set.
- For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (prove it yourself!), and so both quantities are referred to collectively as the rank of A, denoted as rank(A).

Rank of a Matrix

- The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.
- The row rank is the largest number of rows of A that constitute a linearly independent set.
- For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (prove it yourself!), and so both quantities are referred to collectively as the rank of A, denoted as rank(A).
- For $A \in \mathbb{R}^{m \times n}$, rank(A) $\leq$ min(m, n). If rank(A) = min(m, n), then A is said to be full rank.

Rank of a Matrix

- The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.
- The row rank is the largest number of rows of A that constitute a linearly independent set.
- For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (prove it yourself!), and so both quantities are referred to collectively as the rank of A, denoted as rank(A).
- For $A \in \mathbb{R}^{m \times n}$, rank(A) $\leq$ min(m, n). If rank(A) = min(m, n), then A is said to be full rank.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$

Rank of a Matrix

- The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.
- The row rank is the largest number of rows of A that constitute a linearly independent set.
- For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (prove it yourself!), and so both quantities are referred to collectively as the rank of A, denoted as rank(A).
- For $A \in \mathbb{R}^{m \times n}$, rank(A) $\leq$ min(m, n). If rank(A) = min(m, n), then A is said to be full rank.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$
- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times n}$, $\text{rank}(AB) \leq min(\text{rank}(A), \text{rank}(B))$

Inverse of a Square matrix

- The ***inverse*** of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1}, and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

- We say that A is ***invertible*** or ***non-singular*** if A^{-1} exists and ***non-invertible*** or ***singular*** otherwise.

- In order for a square matrix A to have an inverse A^{-1}, then A must be full rank.

- Properties (Assuming $A, B \in \mathbb{R}^{n \times n}$ are non-singular):
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T}.

Orthogonal and orthonormal matrix

- Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$.
- A vector $x \in \mathbb{R}^n$ is ***normalized*** if $\|x\|_2 = 1$.
- A square matrix $U \in \mathbb{R}^{n \times n}$ is ***orthogonal*** if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being ***orthonormal***).

- **Properties:**
  - The inverse of an orthogonal matrix is its transpose.

$$U^T U = I = U U^T.$$

The Determinant

The determinant of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \to \mathbb{R}$, and is denoted $|A|$ or $\det A$.

$$
\begin{aligned}
|[a_{11}]| &= a_{11} \\
\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| &= a_{11}a_{22} - a_{12}a_{21} \\
\left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right| &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\
&\quad - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}
\end{aligned}
$$

The Determinant: Properties

Algebraically, the determinant satisfies the following three properties:

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).
2. Given a matrix $A \in \mathbb{R}^{n \times n}$,if we multiply a single row in A by ascalar $t \in \mathbb{R}$,then the determinant of the new matrix is $t|A|$, (Geometrically, multiplying one of the sides of the set S by a factor t causes the volume to increase by a factor t.)
3. If we exchange any two rows a_i^T and a_j^T of A, then the determinant of the new matrix is $-|A|$

The Determinant: Properties

1. For $A \in \mathbb{R}^{n \times n}$, $|A| = |A^T|$.
2. For $A, B \in \mathbb{R}^{n \times n}$, $|AB| = |A||B|$.
3. For $A \in \mathbb{R}^{n \times n}$, $|\mathrm{A}| = 0$ if and only if A is singular (i.e., non-invertible). (If A is singular then it does not have full rank, and hence its columns are linearly dependent.)
4. For $A \in \mathbb{R}^{n \times n}$ and A non-singular, $|A^{-1}| = 1/|A|$.

Quadratic Forms

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a ***quadratic form***. Written explicitly, we see that

$$x^T Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \quad .$$

We often implicitly assume that the matrices appearing in a quadratic form are symmetric.

$$x^T Ax = (x^T Ax)^T = x^T A^T x = x^T \left(\frac{1}{2} A + \frac{1}{2} A^T \right) x,$$

Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$,we say that $\lambda \in \mathbb{C}$ is an eigenvalue of A and $x \in \mathbb{C}^n$ is the corresponding eigenvector if

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

Intuitively, this definition means that multiplying A by the vector x results in a new vector that points in the same direction as x, but scaled by a factor λ.

Properties of eigenvalues and eigenvectors

- The trace of a A is equal to the sum of its eigenvalues,

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues,

$$|A| = \prod_{i=1}^{n} \lambda_i.$$

- The rank of A is equal to the number of non-zero eigenvalues of A.
- Suppose A is non-singular with eigenvalue λ and an associated eigenvector x. Then $1/\lambda$ is an eigenvalue of A^{-1} with an associated eigenvector x, i.e., $A^{-1}x = (1/\lambda)x$.
- The eigenvalues of a diagonal matrix $D = \mathrm{diag}(d_1, \ldots d_n)$ are just the diagonal entries $d_1, \ldots d_n$.

The Gradient

Suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $A \in \mathbb{R}^{m \times n}$) is the matrix of partial derivatives, defined as:

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

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

The Gradient

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It follows directly from the equivalent properties of partial derivatives that:

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

Matrix Calculus

Suppose $Y = (y_i)_{m\times 1}$, and $X = (x_j)_{n\times 1}$. Suppose $Y = h(X)$. We can define

$$\frac{\partial Y}{\partial X^T} = \left(\frac{\partial y_i}{\partial x_j}\right)_{m\times n}.$$

- we can treat $\partial Y = (\partial y_i, i = 1, ..., m)^T$ as a column vector
- $1/\partial X = (1/\partial x_j, j = 1, ..., m)^T$ as another column vector.

The product of the elements of the two vectors is understood as composition of the two operators, i.e., $\partial y_i(1/\partial x_j) = \partial y_i/\partial x_j$. Then $\partial Y/\partial X^T$ is a matrix according to the matrix multiplication rule.

Second Derivative (Hessian)

By the same reasoning, if Y is a scaler, then the gradient $h'(X) = \partial Y/\partial X$ is a $n \times 1$ column vector, and $\partial Y/\partial X^T$ is a $1 \times n$ row vector. For scaler Y, we can define the Hessian or second derivative

$$h''(X) = \frac{\partial^2 Y}{\partial X \partial X^T} = \left(\frac{\partial^2 Y}{\partial x_i \partial x_j}\right)_{n \times n}.$$

- Again, we can treat $\partial/\partial X$ as a column vector. Then $\partial^2/\partial X \partial X^T = (\partial/\partial X)(\partial/\partial X)^T$ is a squared matrix following the rule of vector multiplication.
- Note that the Hessian is always symmetric

Examples

1. If $Y = AX$, then $y_i = \sum_k a_{ik} x_k$. Thus $\partial y_i / \partial x_j = a_{ij}$. So $\partial Y / \partial X^T = A$.
2. If $Y = X^T S X$, where S is symmetric, then $\partial Y / \partial X = 2SX$, and $\partial^2 Y / \partial X \partial X^T = 2S$.
3. If $S = I$, $Y = \|X\|^2$, $\partial Y / \partial X = 2X$.

The above results generalize the scalar results with almost no change in notation.