Basic Concepts and Notatio

000000000000

Operations and Properties

Matrix Calculus

CS229 Section: Linear Algebra

Griffin Young

Slides adapted from past CS229 teams

October 1, 2021



- Basic Concepts and Notation
- 2 Matrix Multiplication
- 3 Operations and Properties
- 4 Matrix Calculus

Basic Concepts and Notation

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.

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

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,

$$I_{ij} = \left\{ \begin{array}{ll} 1 & i = j \\ 0 & i \neq j \end{array} \right.$$

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 $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$, with

$$D_{ij} = \left\{ \begin{array}{ll} d_i & i = j \\ 0 & i \neq j \end{array} \right.$$

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

- 1 Basic Concepts and Notation
- Matrix Multiplication
- 3 Operations and Properties
- 4 Matrix Calculus

4□ > 4□ > 4□ > 4□ > 4□ > 4□ > 4□

Vector-Vector Product

• inner product or dot product

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i.$$

outer product

$$xy^{T} \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}.$$

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

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

• If we write A by columns, then we have:

$$y = Ax = \begin{bmatrix} \begin{vmatrix} & & & & & | \\ a^1 & a^2 & \cdots & a^n \\ & & & \end{vmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a^1 \\ x_1 \end{bmatrix} x_1 + \begin{bmatrix} a^2 \\ x_2 \end{bmatrix} x_2 + \ldots + \begin{bmatrix} a^n \\ x_n \end{bmatrix} x_n .$$

$$(1)$$

y is a *linear combination* of the *columns* of A.

It is also possible to multiply on the left by a row vector.

• If we write A by columns, then we can express x^TA as,

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a^1 & x^T a^2 & \cdots & x^T a^n \end{bmatrix}$$

It is also possible to multiply on the left by a row vector.

• expressing A in terms of rows we have:

$$y^{T} = x^{T}A = \begin{bmatrix} x_{1} & x_{2} & \cdots & x_{m} \end{bmatrix} \begin{bmatrix} - & a_{1}^{T} & - \\ - & a_{2}^{T} & - \\ & \vdots & \\ - & a_{m}^{T} & - \end{bmatrix}$$
$$= x_{1} \begin{bmatrix} - & a_{1}^{T} & - \end{bmatrix} + x_{2} \begin{bmatrix} - & a_{2}^{T} & - \end{bmatrix} + \dots + x_{m} \begin{bmatrix} - & a_{m}^{T} & - \end{bmatrix}$$

 y^T is a linear combination of the *rows* of A.



1. As a set of vector-vector products (dot product)

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \cdots & b^p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b^1 & a_1^T b^2 & \cdots & a_1^T b^p \\ a_2^T b^1 & a_2^T b^2 & \cdots & a_2^T b^p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b^1 & a_m^T b^2 & \cdots & a_m^T b^p \end{bmatrix}.$$

2. As a sum of outer products

$$C = AB = \begin{bmatrix} | & | & & | \\ | & | & & | \\ | & | & a^2 & \cdots & a^p \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_p^T & - \end{bmatrix} = \sum_{i=1}^p a^i b_i^T .$$

3. As a set of matrix-vector products.

$$C = AB = A \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \cdots & b^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ Ab^1 & Ab^2 & \cdots & Ab^n \\ | & | & & | \end{bmatrix}. \tag{2}$$

Here the *i*th column of C is given by the matrix-vector product with the vector on the right, $c_i = Ab_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection.

4. As a set of vector-matrix products.

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \vdots \\ - & a_m^T & - \end{bmatrix} B = \begin{bmatrix} - & a_1^T B & - \\ - & a_2^T B & - \\ \vdots & & \vdots \\ - & a_m^T B & - \end{bmatrix}.$$

Matrix-Matrix Multiplication (properties)

- Associative: (AB)C = A(BC).
- Distributive: A(B+C) = AB + AC.
- ullet In general, not commutative; that is, it can be the case that AB
 eq BA. (For example, if $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times q}$, the matrix product BA does not even exist if m and q are not equal!)

- 1 Basic Concepts and Notation
- Matrix Multiplication
- 3 Operations and Properties
- 4 Matrix Calculus

Operations and Properties



The 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:

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

Trace

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

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

The trace has the following properties:

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



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 $x \in \mathbb{R}^n$, $f(x) \ge 0$ (non-negativity).
- 2. f(x) = 0 if and only if x = 0 (definiteness).
- 3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, f(tx) = |t|f(x) (homogeneity).
- 4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \le f(x) + f(y)$ (triangle inequality).



Examples of Norms

The commonly-used Euclidean or ℓ_2 norm,

$$||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}.$$
 $||x||_1 = \sum_{i=1}^n |x_i|$

The ℓ_1 norm,

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

The ℓ_{∞} norm,

$$||x||_{\infty} = \max_{i} |x_{i}|.$$

Examples of Norms

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \ge 1$, and defined as

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

Matrix Norms

Norms can also be defined for matrices, such as the Frobenius norm,

$$||A||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2} = \sqrt{\operatorname{tr}(A^T A)}.$$

Many other norms exist, but they are beyond the scope of this review.

Linear Independence

A set of vectors $\{x_1, x_2, \dots x_n\} \subset \mathbb{R}^m$ 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*.

Linear Independence

A set of vectors $\{x_1, x_2, \dots x_n\} \subset \mathbb{R}^m$ 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 $\alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}$; otherwise, the vectors are *(linearly) independent*. **Example**:

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$
 $x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix}$ $x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$

are linearly dependent because $x_3 = -2x_1 + x_2$.



• 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).



• For $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank}(A) \leq \min(m, n)$. If $\operatorname{rank}(A) = \min(m, n)$, then A is said to be **full** rank.

- For $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank}(A) \leq \min(m, n)$. If $\operatorname{rank}(A) = \min(m, n)$, then A is said to be **full** rank.
- For $A \in \mathbb{R}^{m \times n}$, rank $(A) = \text{rank}(A^T)$.

- For $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank}(A) \leq \min(m, n)$. If $\operatorname{rank}(A) = \min(m, n)$, then A is said to be **full** rank.
- For $A \in \mathbb{R}^{m \times n}$, rank $(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, $\operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B))$.

- 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 $A \in \mathbb{R}^{m \times n}$, rank $(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, $\operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}$, $rank(A + B) \le rank(A) + rank(B)$.

The 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

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

The 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

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

Operations and Properties

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



The 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

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

Operations and Properties

- 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.



The 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 Matrices

- 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*).

Orthogonal Matrices

- 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.

$$U^TU = I = UU^T$$
.



Orthogonal Matrices

- 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.

$$U^TU = I = UU^T$$
.

▶ Operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$||Ux||_2 = ||x||_2$$

for any $x \in \mathbb{R}^n$, $U \in \mathbb{R}^{n \times n}$ orthogonal.



Span and Projection

• The *span* of a set of vectors $\{x_1, x_2, \dots x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \dots, x_n\}$. That is,

$$\operatorname{span}(\{x_1,\ldots x_n\}) = \left\{v: v = \sum_{i=1}^n \alpha_i x_i, \ \alpha_i \in \mathbb{R}\right\}.$$

Span and Projection

• The *span* of a set of vectors $\{x_1, x_2, \dots x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \dots, x_n\}$. That is,

$$\operatorname{span}(\{x_1,\ldots x_n\}) = \left\{v: v = \sum_{i=1}^n \alpha_i x_i, \ \alpha_i \in \mathbb{R}\right\}.$$

• The **projection** of a vector $y \in \mathbb{R}^m$ onto the span of $\{x_1, \ldots, x_n\}$ is the vector $v \in \text{span}(\{x_1, \ldots, x_n\})$, such that v is as close as possible to y, as measured by the Euclidean norm $\|v - y\|_2$.

$$\operatorname{Proj}(y; \{x_1, \dots x_n\}) = \operatorname{argmin}_{v \in \operatorname{span}(\{x_1, \dots, x_n\})} \|y - v\|_2.$$

Range

• The *range* or the column space of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the the span of the columns of A. In other words,

$$\mathcal{R}(A) = \{ v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n \}.$$

Range

• The *range* or the column space of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the the span of the columns of A. In other words,

$$\mathcal{R}(A) = \{ v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n \}.$$

• Assuming A is full rank and n < m, the projection of a vector $y \in \mathbb{R}^m$ onto the range of A is given by,

$$\operatorname{Proj}(y; A) = \operatorname{argmin}_{v \in \mathcal{R}(A)} ||v - y||_2.$$

Null space

The *nullspace* of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$ is the set of all vectors that equal 0 when multiplied by A, i.e.,

$$\mathcal{N}(A) = \{ x \in \mathbb{R}^n : Ax = 0 \}.$$

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 $\det A$.

Given a matrix

$$\begin{bmatrix} -& a_1^T & -\\ -& a_2^T & -\\ & \vdots & \\ -& a_n^T & - \end{bmatrix},$$

consider the set of points $S \subset \mathbb{R}^n$ as follows:

$$S = \{ v \in \mathbb{R}^n : v = \sum_{i=1}^n \alpha_i a_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \dots, n \}.$$

The absolute value of the determinant of A is a measure of the "volume" of the set S.

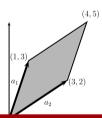
The Determinant: Intuition

For example, consider the 2×2 matrix,

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \tag{3}$$

Here, the rows of the matrix are

$$a_1 = \left[\begin{array}{c} 1 \\ 3 \end{array} \right] \quad a_2 = \left[\begin{array}{c} 3 \\ 2 \end{array} \right]$$



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).

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 a scalar $t \in \mathbb{R}$, 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.)

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 a scalar $t \in \mathbb{R}$, 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_i^T of A, then the determinant of the new matrix is -|A|, for example



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 a scalar $t \in \mathbb{R}$, 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_i^T of A, then the determinant of the new matrix is -|A|, for example



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 a scalar $t \in \mathbb{R}$, 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_i^T of A, then the determinant of the new matrix is -|A|, for example

In case you are wondering, it is not immediately obvious that a function satisfying the above three properties exists. In fact, though, such a function does exist, and is unique (which we will not prove here).



- For $A \in \mathbb{R}^{n \times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n \times n}$, |AB| = |A||B|.
- For $A \in \mathbb{R}^{n \times n}$, |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. In this case, the set S corresponds to a "flat sheet" within the *n*-dimensional space and hence has zero volume.)
- For $A \in \mathbb{R}^{n \times n}$ and A non-singular, $|A^{-1}| = 1/|A|$.

The Determinant: Formula

Let $A \in \mathbb{R}^{n \times n}$, $A_{\setminus i, \setminus j} \in \mathbb{R}^{(n-1) \times (n-1)}$ be the *matrix* that results from deleting the *i*th row and *j*th column from A.

The general (recursive) formula for the determinant is

$$|A| = \sum_{i=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}|$$
 (for any $j \in 1, \dots, n$)
 $= \sum_{j=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}|$ (for any $i \in 1, \dots, n$)

with the initial case that $|A| = a_{11}$ for $A \in \mathbb{R}^{1 \times 1}$. If we were to expand this formula completely for $A \in \mathbb{R}^{n \times n}$, there would be a total of n! (n factorial) different terms. For this reason, we hardly ever explicitly write the complete equation of the determinant for matrices bigger than 3×3 .

The Determinant: Examples

However, the equations for determinants of matrices up to size 3×3 are fairly common, and it is good to know them:

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

Quadratic Forms

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

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

Quadratic Forms

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

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

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

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



Positive Semidefinite Matrices

A symmetric matrix $A \in \mathbb{S}^n$ is:

- positive definite (PD), denoted $A \succ 0$ if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T A x > 0$.
- positive semidefinite (PSD), denoted $A \succeq 0$ if for all vectors $x^T A x \ge 0$.
- negative definite (ND), denoted $A \prec 0$ if for all non-zero $x \in \mathbb{R}^n$, $x^T A x < 0$.
- negative semidefinite (NSD), denoted $A \leq 0$) if for all $x \in \mathbb{R}^n$, $x^T A x \leq 0$.
- *indefinite*, if it is neither positive semidefinite nor negative semidefinite i.e., if there exists $x_1, x_2 \in \mathbb{R}^n$ such that $x_1^T A x_1 > 0$ and $x_2^T A x_2 < 0$.



Positive Semidefinite Matrices

- One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible.
- Given any matrix $A \in \mathbb{R}^{m \times n}$ (not necessarily symmetric or even square), the matrix $G = A^T A$ (sometimes called a **Gram matrix**) is always positive semidefinite. Further, if $m \ge n$ and A is full rank, then $G = A^T A$ is positive definite.



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

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

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 .

Eigenvalues and Eigenvectors

We can rewrite the equation above to state that (λ, x) is an eigenvalue-eigenvector pair of A if,

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

But $(\lambda I - A)x = 0$ has a non-zero solution to x if and only if $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e.,

$$|(\lambda I - A)| = 0.$$

We can now use the previous definition of the determinant to expand this expression $|(\lambda I - A)|$ into a (very large) polynomial in λ , where λ will have degree n. It's often called the characteristic polynomial of the matrix A.

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

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

• 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 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,

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

ullet The rank of A is equal to the number of non-zero eigenvalues of A.

• 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,

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

- ullet 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 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,

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

- ullet 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 $D = \operatorname{diag}(d_1, \ldots d_n)$ are just the diagonal entries $d_1, \ldots d_n$.

Eigenvalues and Eigenvectors of Symmetric Matrices

Throughout this section, let's assume that A is a symmetric real matrix (i.e., $A^{\top} = A$). We have the following properties:

- 1. All eigenvalues of A are real numbers. We denote them by $\lambda_1, \ldots, \lambda_n$.
- 2. There exists a set of eigenvectors u_1, \ldots, u_n such that (i) for all i, u_i is an eigenvector with eigenvalue λ_i and (ii) u_1, \ldots, u_n are unit vectors and orthogonal to each other.

New Representation for Symmetric Matrices

• Let U be the orthonormal matrix that contains u_i 's as columns:

$$U = \left[\begin{array}{cccc} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{array} \right]$$

New Representation for Symmetric Matrices

• Let U be the orthonormal matrix that contains u_i 's as columns:

$$U = \left[\begin{array}{cccc} | & | & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & | \end{array} \right]$$

• Let $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ be the diagonal matrix that contains $\lambda_1, \dots, \lambda_n$.

$$AU = \begin{bmatrix} & | & & | & & | \\ & Au_1 & Au_2 & \cdots & Au_n \\ & & | & & | \end{bmatrix} = \begin{bmatrix} & | & | & & | \\ & \lambda_1u_1 & \lambda_2u_2 & \cdots & \lambda_nu_n \\ & & | & & | \end{bmatrix} = U\operatorname{diag}(\lambda_1, \dots, \lambda_n) = U\Lambda$$

New Representation for Symmetric Matrices

• Let U be the orthonormal matrix that contains u_i 's as columns:

$$U = \left[\begin{array}{cccc} | & | & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & | \end{array} \right]$$

• Let $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ be the diagonal matrix that contains $\lambda_1, \dots, \lambda_n$.

$$AU = \begin{bmatrix} & | & & | & & | \\ & Au_1 & Au_2 & \cdots & Au_n \\ & & | & & | \end{bmatrix} = \begin{bmatrix} & | & | & & | \\ & \lambda_1u_1 & \lambda_2u_2 & \cdots & \lambda_nu_n \\ & & | & & | \end{bmatrix} = U\operatorname{diag}(\lambda_1, \dots, \lambda_n) = U\Lambda$$

• Recalling that orthonormal matrix U satisfies that $UU^T = I$, we can diagonalize matrix A:

$$A = AUU^{T} = U\Lambda U^{T} \tag{4}$$



Background: representing vector w.r.t. another basis

- Any orthonormal matrix $U = \left[\begin{array}{ccc} | & | & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & | \end{array} \right]$ defines a new basis of \mathbb{R}^n .
- For any vector $x \in \mathbb{R}^n$ can be represented as a linear combination of u_1, \ldots, u_n with coefficient $\hat{x}_1, \ldots, \hat{x}_n$:

$$x = \hat{x}_1 u_1 + \dots + \hat{x}_n u_n = U \hat{x}$$

• Indeed, such \hat{x} uniquely exists

$$x = U\hat{x} \Leftrightarrow U^T x = \hat{x}$$

In other words, the vector $\hat{x} = U^T x$ can serve as another representation of the vector x w.r.t the basis defined by U.

"Diagonalizing" matrix-vector multiplication

- Left-multiplying matrix A can be viewed as left-multiplying a diagonal matrix w.r.t the basic of the eigenvectors.
 - Suppose x is a vector and \hat{x} is its representation w.r.t to the basis of U.
 - Let z = Ax be the matrix-vector product.
 - ▶ the representation *z* w.r.t the basis of *U*:

$$\hat{z} = U^{\mathsf{T}} z = U^{\mathsf{T}} A x = U^{\mathsf{T}} U \Lambda U^{\mathsf{T}} x = \Lambda \hat{x} = \begin{bmatrix} \lambda_1 \hat{x}_1 \\ \lambda_2 \hat{x}_2 \\ \vdots \\ \lambda_n \hat{x}_n \end{bmatrix}$$

• We see that left-multiplying matrix A in the original space is equivalent to left-multiplying the diagonal matrix Λ w.r.t the new basis, which is merely scaling each coordinate by the corresponding eigenvalue. ◆□▶ ◆□▶ ◆■▶ ◆■▶ ■ 夕久で

"Diagonalizing" matrix-vector multiplication

Under the new basis, multiplying a matrix multiple times becomes much simpler as well. For example, suppose q = AAAx.

$$\hat{q} = U^T q = U^T A A A x = U^T U \Lambda U^T U \Lambda U^T U \Lambda U^T x = \Lambda^3 \hat{x} = \begin{bmatrix} \lambda_1^3 \hat{x}_1 \\ \lambda_2^3 \hat{x}_2 \\ \vdots \\ \lambda_n^3 \hat{x}_n \end{bmatrix}$$

"Diagonalizing" quadratic form

As a directly corollary, the quadratic form x^TAx can also be simplified under the new basis

$$x^T A x = x^T U \Lambda U^T x = \hat{x}^T \Lambda \hat{x} = \sum_{i=1}^n \lambda_i \hat{x}_i^2$$

(Recall that with the old representation, $x^TAx = \sum_{i=1,j=1}^n x_i x_j A_{ij}$ involves a sum of n^2 terms instead of n terms in the equation above.)

The definiteness of the matrix *A* depends entirely on the sign of its eigenvalues

- 1. If all $\lambda_i > 0$, then the matrix A is positive definite because $x^T A x = \sum_{i=1}^n \lambda_i \hat{x}_i^2 > 0$ for any $\hat{x} \neq 0.1$
- 2. If all $\lambda_i \geq 0$, it is positive semidefinite because $x^T A x = \sum_{i=1}^n \lambda_i \hat{x}_i^2 \geq 0$ for all \hat{x} .
- 3. Likewise, if all $\lambda_i < 0$ or $\lambda_i \le 0$, then A is negative definite or negative semidefinite respectively.
- 4. Finally, if A has both positive and negative eigenvalues, say $\lambda_i > 0$ and $\lambda_j < 0$, then it is indefinite. This is because if we let \hat{x} satisfy $\hat{x}_i = 1$ and $\hat{x}_k = 0, \forall k \neq i$, then $x^T A x = \sum_{i=1}^n \lambda_i \hat{x}_i^2 > 0$. Similarly we can let \hat{x} satisfy $\hat{x}_j = 1$ and $\hat{x}_k = 0, \forall k \neq j$, then $x^T A x = \sum_{i=1}^n \lambda_i \hat{x}_i^2 < 0$.



¹Note that $\hat{x} \neq 0 \Leftrightarrow x \neq 0$.

- 1 Basic Concepts and Notation
- 2 Matrix Multiplication
- 3 Operations and Properties
- 4 Matrix Calculus

Matrix Calculus

Matrix Calculus



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

$$abla_x f(x) = \left[egin{array}{c} rac{\partial f(x)}{\partial x_1} \ rac{\partial f(x)}{\partial x_2} \ dots \ rac{\partial f(x)}{\partial x_n} \end{array}
ight].$$

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:

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



The Hessian

Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in \mathbb{R}^n and returns a real number. Then the *Hessian* matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_{x}^{2}f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^{2}f(x)}{\partial x_{1}^{2}} & \frac{\partial^{2}f(x)}{\partial x_{1}\partial x_{2}} & \cdots & \frac{\partial^{2}f(x)}{\partial x_{1}\partial x_{n}} \\ \frac{\partial^{2}f(x)}{\partial x_{2}\partial x_{1}} & \frac{\partial^{2}f(x)}{\partial x_{2}^{2}} & \cdots & \frac{\partial^{2}f(x)}{\partial x_{2}\partial x_{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^{2}f(x)}{\partial x_{n}\partial x_{1}} & \frac{\partial^{2}f(x)}{\partial x_{n}\partial x_{2}} & \cdots & \frac{\partial^{2}f(x)}{\partial x_{n}^{2}} \end{bmatrix}.$$

In other words, $\nabla^2_x f(x) \in \mathbb{R}^{n \times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$



The Hessian

Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in \mathbb{R}^n and returns a real number. Then the *Hessian* matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_{x}^{2}f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^{2}f(x)}{\partial x_{1}^{2}} & \frac{\partial^{2}f(x)}{\partial x_{1}\partial x_{2}} & \cdots & \frac{\partial^{2}f(x)}{\partial x_{1}\partial x_{n}} \\ \frac{\partial^{2}f(x)}{\partial x_{2}\partial x_{1}} & \frac{\partial^{2}f(x)}{\partial x_{2}^{2}} & \cdots & \frac{\partial^{2}f(x)}{\partial x_{2}\partial x_{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^{2}f(x)}{\partial x_{n}\partial x_{1}} & \frac{\partial^{2}f(x)}{\partial x_{n}\partial x_{2}} & \cdots & \frac{\partial^{2}f(x)}{\partial x_{n}^{2}} \end{bmatrix}.$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$



Gradients of Linear Functions

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^n b_i x_i = b_k.$$

From this we can easily see that $\nabla_x b^T x = b$. This should be compared to the analogous situation in single variable calculus, where $\partial/(\partial x)$ ax = a.

Now consider the quadratic function $f(x) = x^T A x$ for $A \in \mathbb{S}^n$. Remember that $f(x) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j.$

$$f(x) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j.$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

Now consider the quadratic function $f(x) = x^T A x$ for $A \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j.$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

Now consider the quadratic function $f(x) = x^T A x$ for $A \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j.$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k$$
gebra Review Fall 2022

Stanford University 60 / 64

Now consider the quadratic function $f(x) = x^T A x$ for $A \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j.$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k$$

$$= \sum_{i=1}^n A_{ik} x_i + \sum_{j=1}^n A_{kj} x_j = 2\sum_{i=1}^n A_{ki} x_i,$$

Hessian of Quadratic Functions

Finally, let's look at the Hessian of the quadratic function $f(x) = x^T A x$ In this case,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_\ell} \right] = \frac{\partial}{\partial x_k} \left[2 \sum_{i=1}^n A_{\ell i} x_i \right] = 2A_{\ell k} = 2A_{k\ell}.$$

Therefore, it should be clear that $\nabla_x^2 x^T A x = 2A$, which should be entirely expected (and again analogous to the single-variable fact that $\partial^2/(\partial x^2)$ $ax^2=2a$).

Matrix Calculus

Recap

•
$$\nabla_x x^T A x = 2Ax$$
 (if A symmetric)

•
$$\nabla_x^2 x^T A x = 2A$$
 (if A symmetric)

Basic Concepts and Notation Matrix Multiplication occools occo

Matrix Calculus Example: Least Squares

• Given a full rank matrix $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $||Ax - b||_2^2$.

Matrix Calculus Example: Least Squares

- Given a full rank matrix $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $||Ax - b||_2^2$.
- Using the fact that $||x||_2^2 = x^T x$, we have

$$||Ax - b||_2^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$



Matrix Calculus Example: Least Squares

- Given a full rank matrix $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $||Ax - b||_2^2$.
- Using the fact that $||x||_2^2 = x^T x$, we have

$$||Ax - b||_2^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$

• Taking the gradient with respect to x we have:

$$\nabla_{x}(x^{T}A^{T}Ax - 2b^{T}Ax + b^{T}b) = \nabla_{x}x^{T}A^{T}Ax - \nabla_{x}2b^{T}Ax + \nabla_{x}b^{T}b$$
$$= 2A^{T}Ax - 2A^{T}b$$

Matrix Calculus Example: Least Squares

- Given a full rank matrix $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $||Ax - b||_2^2$.
- Using the fact that $||x||_2^2 = x^T x$, we have

$$||Ax - b||_2^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$

• Taking the gradient with respect to x we have:

$$\nabla_{x}(x^{T}A^{T}Ax - 2b^{T}Ax + b^{T}b) = \nabla_{x}x^{T}A^{T}Ax - \nabla_{x}2b^{T}Ax + \nabla_{x}b^{T}b$$
$$= 2A^{T}Ax - 2A^{T}b$$

• Setting this last expression equal to zero and solving for x gives the normal equations

$$x = (A^T A)^{-1} A^T b$$

