01 Linear Algebra

May 26, 2021

Information: *A brief review of Linear Algebra needed in Machine Learning.*

Written by: *Zihao Xu*

Last update date: *05.26.2021*

1 Basic Concepts

1.1 Scalars, Vectors, Matrices and Tensors

1.1.1 Scalars

- Single number
- Denoted as italic lowercase letter such as a, b, c

1.1.2 Vectors

- Array of numbers
- Usually consider vectors to be "column vectors"
- Denoted as lowercase letter (often bolded) $>$ $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$
- Dimension is often denoted by d, D, or p $>$ $\mathbf{x} \in \mathbb{R}^d$
- Access elements via subscript $>$ x_i is the i-th element

1.1.3 Matrices

- 2D array of numbers
- Denoted as uppercase letter (often bolded) $>$ $\mathbf{A} = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}$
- Dimension is often denoted by $m \times n$ $>$ $\mathbf{A} \in \mathbb{R}^{m \times n}$
- Access elements by double subscript $>$ $X_{i,j}$ or $x_{i,j}$ is the i, j-th entry of the matrix
- Access rows or columns via subscript or numpy notation $>$ $X_{i,:}$ is the i-th row, $X_{:,j}$ is the j-th column

1.1.4 Tensors

- n-D array, array with more than two axes $>$ $\mathbf{A} \in \mathbb{R}^{c \times w \times h}$

- Other notations are similar with Matrices

1.1.5 Addition of matrices, scalar multiplication and addition

- When $\mathbf{A} = [A_{i,j}]$ and $\mathbf{B} = [B_{i,j}]$ have the same shape, the sum of them is written as $\mathbf{C} = \mathbf{A}+\mathbf{B}$ where $C_{i,j} = A_{i,j} + B_{i,j}$.
  - In general, matrices of different sizes cannot be added.
  - However, in the context of Deep Learning, notations like $\mathbf{C} = \mathbf{A} + \mathbf{b}$ is allowed where $C_{i,j} = A_{i,j} + b_j$, which means the vector $\mathbf{b}$ is added to each row of the matrix. This is to avoid the need to define a matrix with $\mathbf{b}$ copied into each row before doing the addition, This implicit copying is called **broadcasting**.
- The product of any $m \times n$ matrix $\mathbf{A} = [A_{i,j}]$ and any scalar c is written as $\mathbf{C} = c\mathbf{A}$ where $C_{i,j} = c \cdot A_{i,j}$.
- Similarly, the addition of any $m \times n$ matrix $\mathbf{A} = [A_{i,j}]$ and any scalar b is written as $\mathbf{C} = \mathbf{A}+b$ where $C_{i,j} = A_{i,j} + b$.
- Common calculation rules
  - $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
  - $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
  - $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$
  - $(c + k)\mathbf{A} = c\mathbf{A} + k\mathbf{A}$
  - $c(k\mathbf{A}) = ck\mathbf{A}$

1.1.6 Multiplication (Standard Product)

- The product $\mathbf{C} = \mathbf{AB}$ of an $m \times n_1$ matrix $\mathbf{A} = [A_{i,j}]$ times an $n_2 \times p$ matrix $\mathbf{B} = [B_{i,j}]$ is defined if and only if $n_1 = n_2$ and then $\mathbf{C}$ will be an $m \times p$ matrix $\mathbf{C}$ with entries

$$C_{i,j} = \sum_{k}^{n} A_{i,k} Bk, j$$

- Called standard product or matrix product.
- Common calculation rules
  - $(k\mathbf{A})\mathbf{B} = k(\mathbf{AB}) = \mathbf{A}(k\mathbf{B})$
  - $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
  - $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$
  - $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$

1.1.7 Element-wise product

- A matrix containing the product of the individual elements from two matrix have the same size.
- Denoted by $\mathbf{C} = \mathbf{A} \odot \mathbf{B}$ where $C_{i,j} = A_{i,j} \cdot B_{i,j}$
- Also called Hadamard product

1.1.8 Transposition of Matrices and Vectors

- Denoted as $\mathbf{A}^T$
- The transpose of an $m \times n$ matrix $\mathbf{A} = [A_{i,j}]$ is the $n \times m$ matrix $\mathbf{A}^T$ that has the first row of $\mathbf{A}$ as its first column, the second row as its second column, and so on. > $\mathbf{A}^T = [A_{j,i}] =$

$$\begin{bmatrix} A_{1,1} & \cdots & A_{m,1} \\ \vdots & \ddots & \vdots \\ A_{1,n} & \cdots & A_{m,n} \end{bmatrix}$$

- For vector $\mathbf{v}$, the transpose changes it from a column vector to a row vector. > $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$, $\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}$
- Rules for transposition
  - $(\mathbf{A}^T)^T = \mathbf{A}$
  - $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
  - $(c\mathbf{A})^T = c\mathbf{A}^T$
  - $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

1.1.9 Special Matrices

- Symmetric matrix: $\mathbf{A}^T = \mathbf{A}$, $A_{i,j} = A_{j,i}$
- Skew-symmetric matrix: $\mathbf{A}^T = -\mathbf{A}$
- Triangular matrix:
  - Upper triangular matrix can have non-zero entries only **on and above** the diagonal
  - Lower triangular matrix can have non-zero entries only **on and below** the diagonal
- Identity matrix:
  - Identity matrix of size n is the $n \times n$ square matrix with ones on the main diagonal and zeros elsewhere. It is denoted by $\mathbf{I}_n$ or simply by $\mathbf{I}$ if the size is immaterial or can be trivially determined by the context.
  - Some times called unit matrix (depends on the context).
- Scalar matrix:
  - Any multiple of an Identity matrix.
- Diagonal matrix:
  - A square matrix in which the entries outside the diagonal are all zero.

1.2 Linear System of equations

1.2.1 Represent linear set of equations in matrix equations

- Linear set of equations can be compactly represented as matrix equation
- In general:

$$\begin{gathered} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2 \\ \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n = b_m \end{gathered}$$

  is **equivalent** to:

$$\mathbf{Ax} = \mathbf{b}$$

  where $\mathbf{A} \in \mathbb{R}^{m\times n}$,$\mathbf{x} \in \mathbb{R}^n$,$\mathbf{b} \in \mathbb{R}^m$
- Augmented matrix

$$\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{b}]$$

1.2.2 Gaussian Elimination

- **Goal**: Bring system to a triangular form
- **Step**:
  - Elementary operations on equations $\longleftrightarrow$ Operation on matrices
  - Interchange of two equations $\longleftrightarrow$ Interchange two rows in a matrix
  - Addition of a constant $\longleftrightarrow$ Addition of a constant
- Row equivalent
  - We call a linear system S_1 row-equivalent to a linear system S_2 if S_1 can be obtained by finitely many row operations from S_2.
- Theorem
  - Row-equivalent linear systems have the same set of solutions.
- Solution by Gaussian Elimination
  - System:
    - $\mathbf{Ax} = \mathbf{b}$ with augmented matrix $\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{b}]$
    - $\mathbf{A} \in \mathbb{R}^{m \times n}$,$\mathbf{x} \in \mathbb{R}^n$,$\mathbf{b} \in \mathbb{R}^m$
  - Step 1:
    - Pivot row: First row of $\tilde{\mathbf{A}}$
    - Pivot: Coefficient of the x_1 term in pivot row
    - Use pivot row to eliminate x_1 term in all other rows below
  - Step 2:
    - First equation remains as it is
    - Pivot row: Second row of $\tilde{\mathbf{A}}$
    - Pivot: Coefficient of the x_2 term in pivot row
    - Use pivot row to eliminate x_2 term in all other rows below
  - Step 3:
    - Repeat the procedure which moves the pivot row from s to $s+1$ and set pivot to be the coefficient of x_{s+1} term in pivot row in each step, until $\mathbf{A}$ is in upper triangular form
  - Step 4:
    - Back-substitution to get x_n, x_{n-1}, ..., x_2, x_1 sequentially

1.2.3 Classification of solutions of Linear Systems

- At the end of Gaussian elimination, $\mathbf{A}$ is in upper triangular form (row echelon form)
  - r = number of non-zero rows in $\tilde{\mathbf{A}}$ = **rank** of $\tilde{\mathbf{A}}$, $r \leq m$
- In general, three possible cases
  - Consistent if $r = m$ or $r < m$ but $\tilde{b}_{r+1}$, ..., $\tilde{b}_m$ are all zero
    - One unique solution if consistent and $r = n$
    - Infinite many solution if consistent and $r < n$. In this case, choose x_{r+1}, ...,x_n arbitrarily.
  - Inconsistent if $r < m$ and at least one of $\tilde{b}_{r+1}$, ..., $\tilde{b}_m$ is non-zero
    - No solution

1.3 Linear Independence, Rank of Matrix, Vector Space

1.3.1 Linear Independence

- Given: Set of vectors $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \cdots, \mathbf{v}^{(n)}\}$

- With $c_1, c_2, \cdots, c_n$ are scalars, a linear combination of these vectors is of the form:

$$c_1\mathbf{v}^{(1)} + c_2\mathbf{v}^{(2)} + \cdots + c_n\mathbf{v}^{(n)}$$

- Consider $c_1\mathbf{v}^{(1)} + c_2\mathbf{v}^{(2)} + \cdots + c_n\mathbf{v}^{(n)} = 0$ true for $c_1 = c_2 = \cdots = c_n = 0$
  - If this is the only solution:**This set of vectors form a linear independent set**
  - Otherwise: **Linear Dependent**

1.3.2 Rank of a matrix

- The rank of a matrix $\mathbf{A}$ is the number of linearly independent row vectors of $\mathbf{A}$,
- Denoted by **rank A**
- Determine the rank of a matrix
  - Observation: Number of linearly independent row vectors does not change by elementary row operations
  - **Theorem 1**:
    * Row equivalent matrices have the same rank
  - Strategy: Reduce the matrix to row-echelon form (upper triangular form) and read off the rank directly
  - **Theorem 2**:
    * p vectors with n components each are independent if the matrix with these vectors as row vectors has rank p, but linearly dependent if that rank is less than p
  - **Theorem 3**:
    * The rank of a matrix $\mathbf{A}$ equals the maximum number of linearly independent column vectors of $\mathbf{A}$. Hence $\mathbf{A}$ and its transpose $\mathbf{A}^T$ have the same rank
  - **Theorem 4**:
    * p vectors with $n < p$ components are always linearly dependent.

1.3.3 Vector Space

- **Vector Space**:
  - Denoted by V
  - Also called a **linear space**
  - Nonempty set of vectors with the same number of components such that with any two vectors $\mathbf{a}$ and $\mathbf{b}$, all linear combinations $\alpha\mathbf{a} + \beta\mathbf{b}$ (α, β are real numbers) are elements of V and these vectors satisfy the rules for vector addition and scalar multiplication.
- **Dimension** of V:
  - Maximal number of linearly independent vectors
- **Basis**:
  - Linear independent set of maximally possible vectors
  - Number of vectors in the basis = dim V
- **Span**:
  - Set of all linear combinations given vectors $\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_p}$
- **Subspace**:
  - Nonempty set of vectors which forms itself a vector space with respect to addition and scalar multiplication
- **Theorem 5**:
  - The vector space $\mathbb{R}^n$ consisting of all vectors with n components (real) has dimension n
- **Theorem 6**:

- The row space and the column space of a matrix $\mathbf{A}$ have the same dimension, equal to rank $\mathbf{A}$

1.4 Solution of linear systems: Existence, Uniqueness

1.4.1 Submatrix of a matrix A

- Any matrix obtained from $\mathbf{A}$ by omitting some rows or columns

1.4.2 Theorems for linear systems(homogeneous systems)

- **Homogeneous systems**
  - A linear system of m equations and n unknowns in the form

$$\mathbf{A}\mathbf{x} = 0$$

    where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$
- Always has the trivial solution $\mathbf{x} = 0$
- Nontrivial solutions exist if and only if rank $\mathbf{A} = r < n$
- If $r < n$, the solution, together with $\mathbf{x} = 0$, form a vector space of dimension $n - r$, called the solution space of the system
- In particular, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are solution vectors, so is $\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2$
- Solution space of the system is called **Null Space**, $\mathbf{A}\mathbf{x} = 0$ for every $\mathbf{x}$ from this solution space N
  - dim N = **Nullity**
  - rank $\mathbf{A}$ + nullity $\mathbf{A} = n$
- A homogeneous system with fewer equations than unknowns always has non-trivial solution
  - rank $\mathbf{A} = r \leq m < n$

1.4.3 Theorems for linear systems (non-homogeneous systems)

- **Non-homogeneous systems**:
  - A linear system of m equations and n unknowns in the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

    where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{b} \neq 0$
- **Existence**:
  - A non-homogeneous linear system is consistent (i.e. has solutions) if and only if the coefficient matrix $\mathbf{A}$ and the augmented matrix $\tilde{\mathbf{A}}$ have the same rank.
- **Uniqueness**:
  - The system has precisely one solution if and only if the common rank r of $\mathbf{A}$ and $\tilde{\mathbf{A}}$ equals n
- **Infinite many solutions**
  - If this common rank is less than n, the system has infinitely many solutions. All the solutions can be obtained by determining r unknowns in terms of the remaining $n - r$ unknowns.
- **Solution**
  - If a non-homogeneous system is consistent, then all the solutions are obtained as $\mathbf{x} = \mathbf{x}_o + \mathbf{x}_h$
    - $\mathbf{x}_o$: Fixed solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$
    - $\mathbf{x}_h$: Run through all solutions of $\mathbf{A}\mathbf{x} = 0$

1.5 Determinants, Cramer's Rule

1.5.1 Determinant of order n

- **Only defined for a square matrix**
- $D = \det\mathbf{A} = \begin{vmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \cdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{vmatrix}$
- $n = 1$, $D = a_1$
- $n \geq 2$,expand by $i - th$ rows $(i = 1, 2, \cdots, n)$
  - $D = a_{i,1}C_{i,1} + a_{i,2}C_{i,2} + \cdots + a_{i,n}C_{i,n}$
    - $C_{i,j} = (-1)^{i+j}M_{i,j}$
    - $M_{i,j}$ is the determinant of order $n - 1$, of a submatrix of $\mathbf{A}$ obtained from $\mathbf{A}$ by deleting the i-th row and the j-th column as indicated by the entry $a_{i,j}$
  - $D = \sum\limits_{j=1}^{n} a_{i,j}C_{i,j}$
  - Or alternatively expand by $j - th$ column: $D = \sum\limits_{i=1}^{n} a_{i,j}C_{i,j}$ where $j = 1, 2, \cdots, n$
  - **Remark**: Easier for n upper triangular matrix

1.5.2 General properties of determinants

- Behavior of n-th order determinant under elementary row operations
  - Interchange of two rows or two columns multiplies the determinant by -1
  - Addition of a multiple of one row/column to another row/column doesn't alter the value of the determinant
  - Multiplication of a row/column by a constant c multiplies the value of the determinant by c
    - $\det(c\mathbf{A}) = c^n\det(\mathbf{A})$
    - $\det(\mathbf{A}^T) = \det(\mathbf{A})$
    - $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$
    - $\det(\mathbf{A} + \mathbf{B}) \neq \det(\mathbf{A}) + \det(\mathbf{B})$ (In general)
  - Transposition leaves determinant the same
  - A zero row or zero column renders the value of $\det = 0$
  - Proportional rows or columns render the value of $\det = 0$
- For practical purposes, to evaluate a determinant of n-th order:
  - reduce the matrix to upper triangular form, which need to keep track of operations that change the determinant
  - multiply the elements on the diagonal to calculate the determinant
- Relationship between **Rank** and **Determinant**
  - An $m \times n$ matrix $\mathbf{A} = [A_{i,j}]$ has rank $r \geq 1$ if and only if it has an $r \times r$ submatrix with non-zero determinant
  - In particular, if $\mathbf{A}$ is square with size $n \times n$, it has rank $= n$ if and only if $\det \neq 0$

1.5.3 Cramer's rule (Solution of linear system by determinants)

- If a linear system of n equations for n unknowns:

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$ has non-zero coefficient determinant ($\det(\mathbf{A}) = D \neq 0$), it has precisely one solution

- The solution is given by $x_1 = \frac{D_1}{D}, x_2 = \frac{D_2}{D}, \cdots, x_n = \frac{D_n}{D}$ where D_k is the determinant of a matrix obtained from $\mathbf{A}$ by replacing the j-th column by a column with entries $b_1, b_2, \cdots, b_n$
- If the system is homogeneous and $D \neq 0$, it has only the trivial solution. If $D = 0$, the system has non-trivial solutions.

1.6 Inverse of matrix, Gauss-Jordan eliminations

1.6.1 Inverse of matrix

- Consider only square matrices
- Inverse of an $n \times n$ matrix $\mathbf{A} = [A_{i,j}]$ is $\mathbf{A}^{-1}$ such that:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

- If $\mathbf{A}$ has inverse: $\mathbf{A}$ is non-singular, otherwise $\mathbf{A}$ is singular
  - Singular matrices are similar to zeros (similar to the idea that 0 does not have an inverse)
  - Called "singular" because a random matrix is unlikely to be singular just like choosing a random number is unlikely to be 0
- Motivation:
  - $\mathbf{A}\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ (usually not suitable for numerical calculation)
- **Theorem**: Existence of $\mathbf{A}^{-1}$
  - The inverse $\mathbf{A}^{-1}$ of an $n \times n$ matrix $\mathbf{A}$ exists if and only if the rank$\mathbf{A} = n$, thus if and only if det$\mathbf{A} \neq 0$
- Formula for Inverse of $\mathbf{A}$
  - $\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}}[C_{i,j}]^T$
  - $C_{i,j} = (-1)^{i+j} M_{i,j}$
  - $M_{i,j}$ is the determinant of order $n-1$, of a submatrix of $\mathbf{A}$ obtained from $\mathbf{A}$ by deleting the i-th row and the j-th column as indicated by the entry $a_{i,j}$
  - Usually used on only 2×2 matrix

1.6.2 Gauss-Jordan elimination

- Method to find the inverse
- Build an matrix $[\mathbf{A}|\mathbf{I}]$ containing $\mathbf{A}$ and identity matrix $\mathbf{I}$
- Perform Gaussian elimination on $\mathbf{A}$, but do the same steps on $\mathbf{I}$, until get the result $[\mathbf{I}|\mathbf{B}]$. Thus, $\mathbf{B} = \mathbf{A}^{-1}$

1.7 Norms

1.7.1 Definition

- The "size" of a vector or matrix.
- Intuitively,the norm of a vector $\mathbf{x}$ measures the distance from the origin to the point $\mathbf{a}$.
- Functions mapping vectors or matrices to non-negative values
- Formally, a norm is any function f that satisfies the following properties:
  - $f(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = 0$
  - $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ (the triangle inequality)
  - $\forall \alpha \in \mathbb{R}, f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$

1.7.2 L^p norm

- $||\mathbf{x}||_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$
- $p = 2$: **Euclidean Norm**, used so frequently in machine learning that it is often denoted simply as $||\mathbf{x}||$ with the subscript 2 omitted. It is also common to measure the size of a vector using the squared L^2 norm, which can be calculated simply as $\mathbf{x}^T\mathbf{x}$
  - In most machine learning cases, the squared L^2 norm is more convenient to work with mathematically and computationally than the L^2 norm itself. On example is that each derivative of the squared L^2 norm with respect to each element of $\mathbf{x}$ depends only on the corresponding element of $\mathbf{x}$.
- $p = 1$: commonly used in machine learning when the difference between zero and nonzero elements is very important. Every time an element of $\mathbf{x}$ moves away from 0 by ϵ, the L^1 norm increases by ϵ
  - Sometimes used to count the number of nonzero entries
- $p = \infty$:**Max Norm**, simplifies to the absolute value of the element with the largest magnitude in the vector

$$||\mathbf{x}||_\infty = \max_i |\mathbf{x}_i|$$

1.7.3 Frobenius Norm

- Used to measure the size of a matrix
- $||\mathbf{A}||_F = \sqrt{\sum_{i,j} A_{i,j}^2}$
- Analogous to the L^2 norm of a vector

1.8 Inner Product Space, Linear Transformations

1.8.1 Inner Product

- A binary operation associates each pair of vectors in the space with a scalar quantity known as the inner product of the vectors, often denoted using angle brackets (as in $\langle \mathbf{a}, \mathbf{b} \rangle$).
- **Dot product**: One widely used inner product on a finite dimensional Euclidean space. Apply for two vectors with the same length.
  - $\langle \mathbf{a}, \mathbf{b} \rangle = (\mathbf{a}, \mathbf{b}) = \mathbf{a} \bullet \mathbf{b} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{n} a_i b_i$
  - Two vectors $\mathbf{a}, \mathbf{b}$ are called **orthogonal** if $\mathbf{a} \bullet \mathbf{b} = 0$
  - Can be written in terms of norms: $\mathbf{a}^T\mathbf{b} = ||a||_2 ||b||_2 \cos\theta$

1.8.2 Abstract Real Inner Product Space

- Real vector space V is called real inner product space V together with an inner product $(\mathbf{a}, \mathbf{b})$ satisfying
  - Linearity: $(q_1\mathbf{a} + q_2\mathbf{b}, \mathbf{c}) = q_1(\mathbf{a}, \mathbf{c}) + q_2(\mathbf{b}, \mathbf{c})$ where $\mathbf{a}, \mathbf{b} \in V, q_1, q_2 \in \mathbb{R}$
  - Symmetry: $(\mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{a})$
  - Positive-definite: $(\mathbf{a}, \mathbf{a}) \geq 0$, $(\mathbf{a}, \mathbf{a}) = 0$ if and only if $\mathbf{a} = 0$

1.8.3 Linear Transformations

- a mapping $X \to Y$ between two vector spaces that preserves the operations of vector addition and scalar multiplication

- If F is the mapping between X and $Y (F : X \to Y, F(\mathbf{x}) = \mathbf{y}$ where $\mathbf{x} \in X$ and $\mathbf{y} \in Y)$, then F is called a linear mapping/transformation if for all $\mathbf{x}, \mathbf{x}_2 \in X$ and scalars c:
  - $F(\mathbf{x} + \mathbf{x}_2) = F(\mathbf{x}) + F(\mathbf{x}_2)$
  - $F(c\mathbf{x}) = cF(\mathbf{x})$
- If $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, any real matrix $\mathbf{A} = [A_{i,j}]$ of size $m \times n$ gives a linear transformation

$$\begin{aligned} &\mathbf{y} = \mathbf{A}\mathbf{x} \\ &\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m \\ &\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \end{aligned}$$

- Conversely, every linear transformation $F : \mathbb{R}^n \to \mathbb{R}^m$ can be written in terms of an $m \times n$ matrix

1.9 Trace operator

1.9.1 Definition

- Calculate the sum of all the diagonal entries of a matrix

$$\mathrm{Tr}\left(\mathbf{A}\right) = \sum_i A_{i,i}$$

1.9.2 Properties

- Provides an alternative way of writing the Frobenius norm of a matrix:

$$||\mathbf{A}||_F = \sqrt{\mathrm{Tr}\left(\mathbf{A}\mathbf{A}^T\right)}$$

- Invariant to the transpose operator:

$$\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^T)$$

- Rotational Equivalence: Invariant to the order of factors (if shapes of corresponding matrices allow changing the order):

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA})$$

  - Holds even if the resulting product has a different shape:

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$$

    where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$, which leads to $\mathbf{AB} \in \mathbb{R}^{m \times m}$ and $\mathbf{BA} \in \mathbb{R}^{n \times n}$
- A scalar is its own trace: $a = \mathrm{Tr}(a)$

2 Matrix Eigenvalue Problems

2.1 Determining Eigenvalues and Eigenvectors

2.1.1 Definition of eigenvalues and eigenvectors:

- Let $\mathbf{A} = [A_{i,j}]$ to be an $n \times n$ matrix, then we say $\mathbf{A}$ has an **eigenvector v** corresponding to an **eigenvalue** λ if:

$$\begin{aligned} \mathbf{A}\mathbf{v} &= \lambda\mathbf{v} \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} &= \mathbf{0} \\ \mathbf{v} &\neq \mathbf{0} \end{aligned}$$

  where $\lambda \in \mathbb{R}$
  - A non-trivial solution exists if and only if $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, which gives a polynomial $p(\lambda)$ called the **characteristic polynomial**
  - Eigenvalues are the roots of the characteristic polynomial
    - An $n \times n$ matrix has at least one eigenvalue and at most n numerically different eigenvalues

2.1.2 Theorem:

- The eigenvectors of a matrix $\mathbf{A}$ corresponding to one and the same eigenvalue λ of $\mathbf{A}$, together with $\mathbf{0}$, form a vector space called the **eigenspace** of $\mathbf{A}$
- Eigenvectors are determined only up to a constant $\Rightarrow$ can normalize to get a unit eigenvector
  - $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$
  - Unit eigenvector: $\frac{\mathbf{x}}{||\mathbf{x}||}$
- If $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ are eigenvectors corresponding to different eigenvalues, then $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ are linearly independent
- If a matrix $\mathbf{A}$ has n different eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$, there is a set of eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ which are linearly independent

2.1.3 Remark

- Only applies for square matrix and $\mathbf{v}$ must be non-zero vector
- If $\mathbf{v}$ is an eigenvector, then so is any scaled vector $s\mathbf{v}$ for $s \in \mathbb{R}, s \neq 0$ while sharing the same eigenvalue λ. For this reason, we usually look only fro unit eigenvectors
- Transpose of a square matrix $\mathbf{A}$ has the same eigenvalues as $\mathbf{A}$ but not necessarily same eigenvectors
  - $\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \Rightarrow \det(\mathbf{A} - \lambda\mathbf{I})^T = 0 \Rightarrow \det(\mathbf{A}^T - \lambda\mathbf{I}) = \det(\mathbf{A} - \lambda\mathbf{I})^T = 0$, which means the characteristic polynomials are same
- For real matrices with complex eigenvalues, eigenvectors would come in complex conjugate pairs

2.1.4 Compute eigenvectors

- Solve the characteristic polynomial for eigenvalues λ
- Solve the homogeneous system equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ for each eigenvalue
- **Algebraic multiplicity** M_λ: Order of an eigenvalue λ as a root in characteristic polynomial
  - Sum of all algebraic multiplicity is n

- **Geometric multiplicity** m_λ: Number of linearly independent eigenvectors corresponding to λ
  - $m_\lambda \leq M_\lambda \leq n$
- **Defect of A**: $\Delta_\lambda = M_\lambda - m_\lambda$

2.1.5 Complex Matrices and Forms

- Sometime the complex eigenvalues lead to application of complex matrices
- Extend the concept of dot product to n-component vector with complex entries
  - $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n,\ (\mathbf{u}, \mathbf{v}) = \bar{\mathbf{u}}^T \mathbf{v}$
- Norm is still real
  - $||\mathbf{v}|| = \sqrt{\bar{v}_1 v_1 + \bar{v}_2 v_2 + \cdots + \bar{v}_n v_n}$
- Extend symmetries
  - Called **Hermitian** if $\bar{\mathbf{A}}^T = \mathbf{A}$
    * Real Hermitian matrix is Symmetric matrix
  - Called **Skew-Hermitian** if $\bar{\mathbf{A}}^T = -\mathbf{A}$
    * Real Skew-Hermitian matrix is Skew-Symmetric matrix
  - Called **Unitary** if $\bar{\mathbf{A}}^T = \mathbf{A}^{-1}$
    * Real Unitary matrix is Orthonormal matrix
    * Determinant of a unitary matrix has absolute value 1

2.1.6 Positive Definite Matrix

- **Positive Definite**: A matrix whose eigenvalues are all positive is called positive definite
- **Positive Semidefinite**: A matrix whose eigenvalues are all positive or zero value is called positive semidefinite
- **Negative Definite**: A matrix whose eigenvalues are all negative is called negative definite
- **Negative Semidefinite**: A matrix whose eigenvalues are all negative or zero value is called negative semidefinite
- **Motivation**:
  - A quadratic form in $\mathbb{R}^n$ is an expression $\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i,j=1}^{n} A_{i,j} x_i x_j$ where $\mathbf{x} \in \mathbb{R}^n$
    * This can always be achieved by a symmetric matrix by replacing $A_{i,j}$ and $A_{j,i}$ by their average
  - A positive semidefinite matrix guarantees that $\forall \mathbf{x},\ \mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$
  - A positive definite matrix additionally guarantees that $\mathbf{x}^T \mathbf{A} \mathbf{x} = 0 \Rightarrow \mathbf{x} = 0$

2.2 Eigenvalues and eigenvectors of special matrices

2.2.1 Symmetric/Hermitian

- For symmetric/hermitian square matrices $\mathbf{A} = \mathbf{A}^T$, the eigenvalues are always real
- Symmetry matrices always have an orthogonal basis of eigenvectors for **R**^n
- For hermitian matrices, eigenvectors corresponding to different eigenvalues are orthogonal.
- Hermitian matrices always have a set of n linearly independent eigenvectors, even if there're repeated roots

2.2.2 Skew-symmetric/Skew-hermitian

- For skew-symmetric/skew-hermitian square matrices $\mathbf{A} = -\mathbf{A}^T$, the eigenvalues are always purely imaginary or zero.

2.2.3 Orthonormal/Unitary

- **Orthogonal**: A real square matrix in which the row vectors (and also its column vectors) from an orthogonal system.
  - $A_i \bullet A_j = A_i^T A_j = 0$ if $i \neq j$
  - $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are diagonal matrices
- **Orthonormal**: Orthogonal matrices with all the norms of row vectors and column vectors normalized to 1
  - $A_i \bullet A_j = A_i^T A_j = 0$ if $i \neq j$ and $A_i \bullet A_j = A_i^T A_j = 1$ if $i = j$
  - $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$, which implies $\mathbf{A}^{-1} = \mathbf{A}^T$
  - Determinant of an orthonormal matrix is always $+1$ or -1
- For orthonormal/unitary matrices, the eigenvalues are either real or in complex conjugate pairs and always have absolute value 1
- Unitary matrices always have a set of n linearly independent eigenvectors, even if there're repeated roots

2.3 Eigendecomposition

2.3.1 Motivation

- To understand a matrix better by breaking it into constituent parts or finding some properties that are universal and not caused by the way to represent the matrix
- Analagous to prime factorization of an integer, which allows us to determine whether things are divisible by other integers
- Analagous to representing a signal in the time versus frequency domain, where both time and frequency domain represent the same object but are useful for different computations and derivations.

2.3.2 Similarity

- If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{P}$ is an non-singular $n \times n$ matrix. then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is called a similarity transformation of $\mathbf{A}$ and the resulting matrix $\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is called similar to $\mathbf{A}$
- If $\hat{\mathbf{A}}$ is similar to $\mathbf{A}$, then $\hat{\mathbf{A}}$ and $\mathbf{A}$ have the same eigenvalues

2.3.3 Diagonalization

- If there is a matrix $\mathbf{P}$ (non-singular) such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$ where $\mathbf{D}$ is diagonal, according to similarity, the resulting matrix $\mathbf{D}$ would have the same eigenvalues as $\mathbf{A}$. Therefore, $\mathbf{D}$ has the eigenvalues of $\mathbf{A}$ on diagonal and each λ_i would repeat as many times as its algebraic multiplicity.
  - An $n \times n$ matrix is diagonalizable $\Leftrightarrow$ $\mathbf{A}$ has n linear independent eigenvectors.
- In fact, the columns of $\mathbf{P}$ are the eigenvectors of $\mathbf{A}$

- Let the columns of $\mathbf{P}$ be denoted by $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$:

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$$
$$\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D}$$
$$\mathbf{A}\left[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\right] = \left[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\right] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$
$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

  where $i = 1, 2, \cdots, n$

- General procedures to diagonalize a matrix $\mathbf{A}$:
  1. Find the roots of the characteristic polynomial $\det(\mathbf{A} - \lambda\mathbf{I})$ to get the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$
  2. Find the corresponding eigenvectors $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ by solving the homogeneous equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ corresponding to each eigenvalue
  3. Construct matrix $\mathbf{P} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ and $\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$
  4. Find the inverse of $\mathbf{P}$ (sometimes optional for convenience)
  5. Then the diagonalized form of $\mathbf{A}$ is $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$

2.4 Singular Value Decomposition (SVD)

2.4.1 Motivation

- A more **generally applicable** way to factorize a matrix into **singular vectors** and **singular values**, which reveals the same kind of information as the eigendecompostion does
- Every real matrix (not necessarily to be square) has a singular value decomposition, but the same is not true of the eigenvalue decomposition.

2.4.2 Definition

- For a matrix $\mathbf{A}$ to be decomposed, write it as a product of three matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

  where $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{U}$ is defined to be an $m \times m$ matrix, $\mathbf{D}$ to be an $m \times n$ matrix, and $\mathbf{V}$ to be an $n \times n$ matrix
  - $\mathbf{U}$ and $\mathbf{V}$ are both defined to be orthonormal matrices
  - $\mathbf{D}$ is defined to be a diagonal matrix and is not necessarily square
  - Elements along the diagonal of $\mathbf{D}$ are known as the **singular values** of the matrix $\mathbf{A}$
  - Columns of $\mathbf{U}$ are known as the **left-singular vectors**
  - Columns of $\mathbf{V}$ are known as the **right-singular vectors**

2.4.3 Calculation

- Left-singular vectors (columns of $\mathbf{U}$) are the eigenvectors of $\mathbf{A}\mathbf{A}^T$

- Right-singular vectors (columns of $\mathbf{V}$) are the eigenvectors of $\mathbf{A}^T\mathbf{A}$
- Nonzero singular values (diagonal elements of $\mathbf{D}$) are the square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$, which is the same as that of $\mathbf{A}\mathbf{A}^T$

2.5 Moore-Penrose Pseudoinverse

2.5.1 Motivation

- Matrix inversion is not defined for matrices that are not square
- Find a way to handle the situation where the system matrix is not square

2.5.2 Definition

- Formal Definition: $\mathbf{A}^+ = \lim_{\alpha \searrow 0}(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{I})^{-1}\mathbf{A}^T$
- Practical computation: $\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^T$
  - $\mathbf{U}, \mathbf{D}, \mathbf{V}$ are the singular value decomposition of $\mathbf{A}$
  - The pseudoinverse $\mathbf{D}^+$ is obtained by taking the reciprocal of its nonzero elements then taking the transpose of the resulting matrix

2.5.3 Application

- In the non-homogeneous system $\mathbf{A}\mathbf{x} = \mathbf{y}$, use this pseudoinverse to get the result $\mathbf{x} = \mathbf{A}^+\mathbf{y}$
  - If $\mathbf{A}$ has more columns than rows, the result provides one of the many possible solutions. Specially, it provides the solution $\mathbf{x} = \mathbf{A}^+\mathbf{y}$ with minimal Euclidean norm $||\mathbf{x}||$ among all possible solutions
  - If $\mathbf{A}$ has more row than column, which means it is possible for there to be no solution, the result gives us the $\mathbf{x}$ for which $\mathbf{A}\mathbf{x}$ is as close as possible to $\mathbf{y}$ in terms of Euclidean norm $||\mathbf{A}\mathbf{x} - \mathbf{y}||$

3 Derivatives with vectors and matrices

3.1 Notation

- **Matrix Notation** and **Tensor Index Notation** are two competing notational conventions which split the field of matrix calculus into two separate groups.
  - The two groups can be distinguished by whether they write the derivative of a scalar with respect to a vector as a column vector or a row vector.
  - **Matrix Notation** writes the derivative of a scalar with respect to a vector as a **row vector**, which is used throughout this notebook.
- In **Matrix Notation**:
  - A scalar is denoted with lowercase italic typeface
  - A vector is denoted with a boldface lowercase letter
  - A matrix is denoted with bold capital letters
  - ...

3.2 Derivatives with vectors

3.2.1 Vector-by-scalar

- The derivative of a vector $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^T$, by a scalar x is written as:

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_n}{\partial x} \end{bmatrix}$$

- Notice: Lay out according to $\mathbf{y}$, which is called numerator layout

3.2.2 Scalar-by-vector

- The derivative of a scalar y by a vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$, is written as:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

- Notice: Lay out according to $\mathbf{x}^T$, which is called numerator layout

3.2.3 Vector-by-vector

- The derivative of a vector function $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^T$, with respect to an input vector, $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$, is written as:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

- Notice: Lay out according to $\mathbf{y}$ and $\mathbf{x}^T$, which is sometimes known as the **Jacobian formulation** and is called numerator layout

3.3 Derivatives with matrices

3.3.1 Matrix-by-scalar

- The derivative of a matrix function $\mathbf{Y}$ by a scalar x is known as the **tangent matrix** and is given by:

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{1,1}}{\partial x} & \frac{\partial y_{1,2}}{\partial x} & \cdots & \frac{\partial y_{1,n}}{\partial x} \\ \frac{\partial y_{2,1}}{\partial x} & \frac{\partial y_{2,2}}{\partial x} & \cdots & \frac{\partial y_{2,n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{m,1}}{\partial x} & \frac{\partial y_{m,2}}{\partial x} & \cdots & \frac{\partial y_{m,n}}{\partial x} \end{bmatrix}$$

- Notice: Lay out according to $\mathbf{Y}$, which is called numerator layout

3.3.2 Scalar-by-matrix

- The derivative of a scalar y function of a $p \times q$ matrix $\mathbf{X}$ of independent variables, with respect to the matrix $\mathbf{X}$, is given by:

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{1,1}} & \frac{\partial y}{\partial x_{2,1}} & \cdots & \frac{\partial y}{\partial x_{p,1}} \\ \frac{\partial y}{\partial x_{1,2}} & \frac{\partial y}{\partial x_{2,2}} & \cdots & \frac{\partial y}{\partial x_{p,2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1,q}} & \frac{\partial y}{\partial x_{2,q}} & \cdots & \frac{\partial y}{\partial x_{p,q}} \end{bmatrix}$$

- Notice: Lay out according to $\mathbf{X}^T$, which is called numerator layout

3.3.3 Other matrix derivatives

- Vectors by matrices, matrices by vectors, and matrices by matrices are not widely considered and a notation is not widely agreed upon.

3.4 Identities often used in machine learning

- For complete identities, refer to Matrix calculus

3.4.1 Vector-by-vector identities

- $\mathbf{A}$ is not a function of $\mathbf{x}$:
$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$$
- $\mathbf{A}$ is not a function of $\mathbf{x}$:
$$\frac{\partial \mathbf{x}^T\mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}^T$$
- $v = v(\mathbf{x}), \mathbf{u} = \mathbf{u}(\mathbf{x})$:
$$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} = v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial v}{\partial \mathbf{x}}$$
- $\mathbf{A}$ is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$:
$$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} = \mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$
- $\mathbf{u} = \mathbf{u}(\mathbf{x})$:
$$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

3.4.2 Scalar-by-vector identities

- $u = u(\mathbf{x}), v = v(\mathbf{x})$:
$$\frac{\partial uv}{\partial \mathbf{x}} = u\frac{\partial v}{\partial \mathbf{x}} + v\frac{\partial u}{\partial \mathbf{x}}$$
- $u = u(\mathbf{x})$:
$$\frac{\partial f(g(u))}{\partial \mathbf{x}} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$$
- $\mathbf{u} = \mathbf{u}(\mathbf{x}), \mathbf{v} = \mathbf{v}(\mathbf{x})$:
$$\frac{\partial(\mathbf{u} \bullet \mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}^T\mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^T\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^T\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$
- $\mathbf{a}$ is not a function of $\mathbf{x}$:
$$\frac{\partial(\mathbf{a} \bullet \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^T$$
- $\mathbf{A}$ is not a function of $\mathbf{x}$:
$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$$
- $\mathbf{a}, \mathbf{b}$ are not functions of $\mathbf{x}$:
$$\frac{\partial \mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)$$
- $\mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{D}, \mathbf{e}$ are not functions of $\mathbf{x}$:
$$\frac{\partial(\mathbf{A}\mathbf{x} + \mathbf{b})^T\mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e})}{\partial \mathbf{x}} = (\mathbf{D}\mathbf{x} + \mathbf{e})^T\mathbf{C}^T\mathbf{A} + (\mathbf{A}\mathbf{x} + \mathbf{b})^T\mathbf{C}\mathbf{D}$$
- $\mathbf{a}$ is not a function of $\mathbf{x}$:
$$\frac{\partial||\mathbf{x} - \mathbf{a}||}{\partial \mathbf{x}} = \frac{(\mathbf{x} - \mathbf{a})^T}{||\mathbf{x} - \mathbf{a}||}$$

3.4.3 Vector-by-scalar identities

- $\mathbf{A}$ is not a function of x, $\mathbf{u} = \mathbf{u}(x)$:
$$\frac{\partial \mathbf{A}\mathbf{u}}{\partial x} = \mathbf{A}\frac{\partial \mathbf{u}}{\partial x}$$
- $\mathbf{u} = \mathbf{u}(x), \mathbf{v} = \mathbf{v}(x)$:
$$\frac{\partial(\mathbf{u} + \mathbf{v})}{\partial x} = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$$
- $\mathbf{u} = \mathbf{u}(x)$:
$$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial x} = \frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial x}$$

3.4.4 Scalar-by-matrix identities

- $u = u(\mathbf{X}), v = v(\mathbf{X})$:
$$\frac{\partial uv}{\partial \mathbf{X}} = u\frac{\partial v}{\partial \mathbf{X}} + v\frac{\partial u}{\partial \mathbf{X}}$$
- $u = u(\mathbf{X})$:
$$\frac{\partial f(g(u))}{\partial \mathbf{X}} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{X}}$$

- $\mathbf{a}$ and $\mathbf{b}$ are not function of $\mathbf{X}$:
$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$
- $\mathbf{a}$ and $\mathbf{b}$ are not function of $\mathbf{X}$:
$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$
- $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are not functions of $\mathbf{X}$:
$$\frac{\partial (\mathbf{X}\mathbf{a}+\mathbf{b})^T\mathbf{C}(\mathbf{X}\mathbf{a}+\mathbf{b})}{\partial \mathbf{X}} = \left(\left(\mathbf{C}+\mathbf{C}^T\right)\left(\mathbf{X}\mathbf{a}+\mathbf{b}\right)\mathbf{a}^T\right)^T$$
- $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are not functions of $\mathbf{X}$:
$$\frac{\partial (\mathbf{X}\mathbf{a})^T\mathbf{C}(\mathbf{X}\mathbf{b})}{\partial \mathbf{X}} = \left(\mathbf{C}\mathbf{X}\mathbf{b}\mathbf{a}^T + \mathbf{C}^T\mathbf{X}\mathbf{a}\mathbf{b}^T\right)^T$$