Constructor University Bremen Autumn Semester 2024

Dr. D. Sinden

MDE-MET-01: Calculus and Linear Algebra for Graduate Students

Summary from 4 December 2024

Notes for the course Calculus and Linear Algebra for Graduate Students, given in autumn 2024. This can be downloaded from:
`https://djps.github.io/docs/gradcalclinalg24/notes`

Contents

Recommended Reading

- G. Strang "*Introduction to Linear Algebra*", Wellesley-Cambridge Press 5th Edition (2016)
- G. Strang "*Linear Algebra and Learning from Data*", Wellesley-Cambridge Press (2019)
- S. Boyd and L. Vandenberghe "*Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*" Cambridge University Press (2018)

1 Vectors

The set of all real numbers is denoted as $\mathbb{R}$ and $\mathbb{R}^N$ as the set of all vectors with N coordinates, whose entries are real numbers. Vectors are denoted by $\boldsymbol{v}$ or $\vec{v}$.

The vector whose every entry is zero is called the zero vector. But note that if $\vec{v} = \vec{0} \in \mathbb{R}^2$ and $\vec{u} = \vec{0} \in \mathbb{R}^3$, then $\vec{v} \neq \vec{u}$.

Definition 1.1 (Linear Combinations)**.** Given some k vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k \in \mathbb{R}^n$, then their **linear combination** is an expression of the form

$$a_1\vec{v}_1 + a_2\vec{v}_2 + \ldots + a_k\vec{v}_k$$

where $a_1, a_2, \ldots, a_k \in \mathbb{R}$, i.e. are scalars.

Definition 1.2 (Dot Product)**.** For two vectors $\vec{a}$, $\vec{b} \in \mathbb{R}^n$, where $\vec{a} = (a_1, a_2, \ldots, a_n)$, the **dot product** is given by

$$\vec{a} \cdot \vec{b} = a_1b_1 + a_2b_2 + \ldots a_nb_n.$$

The result is a scalar.

The *length* of a vector is denoted by $|\vec{v}| = \sqrt{\vec{v} \cdot \vec{v}}$. A unit vector has length one. Any non-zero vector can be transformed to a unit vector by $\vec{v} \mapsto \vec{v}/\,|\vec{v}|$.

One can show that

$$\vec{u} \cdot \vec{v} = |\vec{u}|\,|\vec{v}| \cos\theta.$$

Thus, the angle between two vectors is given by

$$\cos\theta = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}|\,|\vec{v}|}$$

Thus, two vectors, $\vec{v}$ and $\vec{u}$ are said to be orthogonal (or perpendicular) if $\vec{v} \cdot \vec{u} = 0$.

Definition 1.3 (Cauchy Schwarz Inequality)**.** The Cauchy-Schwarz inequality states that the magnitude of the between two vectors is always less than or equal to the product of the magnitudes of the two vectors, i.e.

$$|\vec{v} \cdot \vec{w}| \leq |\vec{v}|\,|\vec{w}|\,.$$

Definition 1.4 (Triangle Inequality)**.** The triangle inequality states that the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side. Formally, this is expressed as

$$|\vec{v} + \vec{w}| \leq |\vec{v}| + |\vec{w}|\,.$$

It can be derived from the Cauchy-Schwarz inequality.

For the Cauchy-Schwarz inequality, when the vectors $\vec{v}$ and $\vec{w}$ lie on the same line, then $|\vec{v} \cdot \vec{w}| = |\vec{v}|\,|\vec{w}|$.

For the triangle inequality, when the vectors point in the same direction, then the $|\vec{v} + \vec{w}| = |\vec{v}| + |\vec{w}|$.

2 Matrices

A matrix is a rectangular array of numbers. Matrices are usually denoted by uppercase letters. A matrix, A, with m rows and n columns, is called an $m \times n$ matrix. If all the entries are real numbers, the matrix is a member of the set of all real matrices with m rows and n columns, i.e. $A \in \mathbb{R}^{m \times n}$.

The entry in the i^{th} row and j^{th} column of the matrix $A \in \mathrm{R}^{m \times n}$ is denoted by a_{ij} or $a_{i,j}$. Hence, the matrix is written as

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}.$$

A matrix is said to be a **square matrix** if the number of rows is equal to the number of columns, i.e. $A \in \mathrm{R}^{n \times m}$ and $n = m$.

Matrix Operations

Definition 2.1 (Matrix-Matrix Multiplication)**.** For two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times l}$, the matrix product $C = AB$ is a $m \times l$ matrix whose entries are given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

for any $i = 1, \ldots, m$ and $j = 1, \ldots, l$.

If A is an $m \times n$ matrix and B is a $n \times l$ matrix, then A has the same number of columns as B has rows, i.e. n. Thus, each entry of $C = AB$, i.e. c_{ij} is the dot product of the i^{th} row of A with the j^{th} column of B, which both have length n.

Definition 2.2 (Matrix-Vector Multiplication)**.** For a matrix $A \in \mathbb{R}^{m \times n}$ and vector $\boldsymbol{x} \in \mathbb{R}^n$, the matrix-vector product $A\boldsymbol{x} = \boldsymbol{y}$ is an n vector whose entries as given by

$$y_j = \sum_{i=1}^{n} a_{ij} x_i.$$

Definition 2.3 (Row and Column Vectors)**.** When matrix $A \in \mathbb{R}^{m \times 1}$ is comprised of a single column of m entries, it is called a column vector. Similarly, a $1 \times n$ matrix with a single row of n entries is called a row vector.

Thus $\boldsymbol{v}^T\boldsymbol{v}$ is a 1×1 matrix, and $\boldsymbol{v}\boldsymbol{v}^T$ is a $n \times n$ matrix, the former is the dot product (or an inner product), whereas the later is an outer product and is often written as $\boldsymbol{v} \otimes \boldsymbol{v}$.

Definition 2.4 (Transpose)**.** For a matrix $A \in \mathbb{R}^{m \times n}$, the transpose of the matrix $A^T \in \mathbb{R}^{n \times m}$ where all elements are mapped to $a_{ij} = a_{ji}$.

Thus, if $\boldsymbol{x}$ is a column vector it is possible to perform $A\boldsymbol{x} = \boldsymbol{y}$, then the equivalent multiplication by a row vector is $\boldsymbol{x}^T A^T = \boldsymbol{y}^T$.

Definition 2.5 (Trace)**.** For a square matrix $A \in \mathbb{R}^{n \times n}$, the trace is defined as the sum of the elements on the main diagonal.

$$\operatorname{tr}(A) = \sum_{i=0}^{n} a_{ii}.$$

Properties of Matrix Operations

1. $A + B = B + A$
2. $c(A + B) = cA + cB$
3. $C(A + B) = CA + CB$ – left-hand distributive property
4. $(A + B)C = AC + BC$ – right-hand distributive property
5. $AB \neq BA$
6. $A + (B + C) = (A + B) + C$
7. $A(BC) = (AB)C$ — associative property

Proposition 2.6 (Transpose of Matrix Products)**.** It can be shown that:

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

Useful Types of Matrices

- The **identity matrix** is the square matrix I such that $AI = A$ and $IA = A$, all entries on the main diagonal are one, all others are zero.
- A **diagonal matrix** is a matrix where all entries outside the main diagonal are all zero, i.e. $a_{ij} = 0$ when $i \neq j$.
- A **banded matrix** is a matrix whose non-zero entries are confined to a diagonal band, comprising the main diagonal and zero or more diagonals on either side.

- The **inverse matrix** of a square matrix is the matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$. A matrix is non-singular or invertible if there exists an inverse matrix exists.
- A square matrix A is **symmetric** if $A = A^T$, that is, $a_{i,j} = a_{j,i}$ for all indices i and j.
- A square matrix, with complex elements, is said to be **Hermitian** if the matrix is equal to its *conjugate transpose*, i.e. $a_{i,j} = \overline{a_{j,i}}$ for all indices i and j. A Hermitian matrix is written as $A = A^H$.
- An **orthogonal matrix** Q is a matrix whose columns $\vec{q}_i$ are orthogonal to one another, that is $\vec{q}_i \cdot \vec{q}_j = 0$ for $i \neq j$ and have unit length, i.e. $\|\vec{q}_i\| = 1$.

Proposition 2.7 (Inverse of Orthogonal Matrices)**.** The inverse of an orthogonal matrix Q is its transpose, i.e.

$$Q^{-1} = Q^T.$$

- A **Markov matrix** is a matrix whose elements are all positive and each column sums to one.

Proposition 2.8 (Largest Eigenvalue of Markov Matrices)**.** The largest eigenvalue of a Markov matrix is equal to one.

- A **permutation matrix** is a square matrix that has exactly one entry of 1 in each row and each column and all other entries 0.
- **Rotation matrices** describe rotations by an angle about the axes of a coordinate system. They can be denoted by $R_x(\alpha)$, i.e. rotation by angle θ about the x axis. The product of two rotation matrices is also a rotation matrix. Note that in general $R_x(\theta) R_y(\beta) \neq R_y(\beta) R_x(\theta)$.
- **Reflection Matrices**: $R = I - 2\boldsymbol{u}\boldsymbol{u}^T$.
- A square matrix is said to be **lower triangular matrix** if all the elements above the main diagonal are zero, i.e. $a_{ij} = 0$ when $i < j$. Similarly, a matrix is said to be **upper triangular** if all the entries below the main diagonal are zero, that is $a_{ij} = 0$ when $i > j$. A matrix is said to be strictly upper triangular (or strictly lower triangle) if the main diagonal is also zero.

Properties of Nonsingular Matrices

For a nonsingular matrix, the following all hold:

- Nonsingular matrices have full rank.
- A square matrix is nonsingular if and only if the determinant of the matrix is non-zero.
- If a matrix is singular, both versions of Gaussian elimination (with and without pivoting) will fail due to division by zero, yielding a floating exception error. Another way to understand this is that the number of pivots is equal to the rank, so if the matrix does not have full rank, so there will not be enough pivots in order to transform the matrix in to row-echelon form.

3 Linear Systems

Definition 3.1 (Systems of Linear Equations)**.** A system of linear equations (or a linear system) is a collection of one or more linear equations involving the same variables. If there are m equations with n unknown variables to solve for, i.e.

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

then the system of linear equations can be written in matrix form $A\boldsymbol{x} = \boldsymbol{b}$, where

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix},$$

with $A \in \mathbb{R}^{m \times n}$, $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^m$.

Direct Methods

Algorithm (Gaussian Elimination)**.**
Gaussian elimination is a method to solve systems of linear equations based on forward elimination (a series of row-wise operations) to convert the matrix, A, to upper triangular form (echelon form), and then back-substitution to solve the system. The row operations are:

- row swapping
- row scaling, i.e. multiplying by a non-zero scalar
- row addition, i.e. adding a multiple of one row to another

procedure Forward Elimination
  for $k = 1$ to $n - 1$ **do**
    for $i = k + 1$ to n **do**
      for $j = k$ to n **do**
        $a_{i,j} = a_{i,j} - \frac{a_{i,k}}{a_{k,k}} a_{k,j}$
      end for
      $b_i = b_i - \frac{a_{i,k}}{a_{k,k}} b_k$
    end for
  end for
end procedure
procedure Back Substitution
  $x_n = \frac{b_n}{a_{n,n}}$
  for $i = n - 1$ to 1 **do**
    $y = b_i$
    for $j = n$ to $i + 1$ **do**
      $y = y - a_{i,j} x_j$
    end for
    $x_i = \frac{y}{a_{i,i}}$
  end for
end procedure

Algorithm (Gaussian Elimination with Scaled Partial Pivoting)**.** A pivot element is the element of a matrix which is selected first to do certain calculations. Pivoting helps reduce errors due to rounding during forward elimination.
To use partial pivoting to produce a matrix in row-echelon form

Find maximal absolute values vector $\boldsymbol{s}$ with entries $s_i = \max j = 1, \ldots, n \, |a_{i,j}|$
for $k = 1$ to $n - 1$ **do**
  for $i = k$ to n **do**
    Compute $\left| \frac{a_{i,k}}{s_i} \right|$
  end for
  Find row with largest relative pivot element, denote this as row j
  Swap rows k and j in the matrix A
  Swap entries k and j in the vector $\boldsymbol{s}$
  Do forward elimination on row k
end for

Theorem 1 (LU-Decomposition). Let $A \in \mathbb{R}^{n \times n}$ be invertible. Then there exists a decomposition of A such that $A = LU$, where L is a lower triangular matrix and U is an upper triangular matrix, and

$$L = U_1^{-1} U_2^{-1} \cdots U_{n-1}^{-1}$$

where each matrix U_i is a matrix which describes the i^{th} step in forward elimination part of Gaussian elimination.

The upper triangular matrix U is given by

$$U = U_{n-1} \cdots U_2 U_1 A.$$

4 Vector Spaces

Definition 4.1 (Linear Independence)**.** A set of vectors, $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ are linearly independent if none of the vectors can be expressed as a linear combination of the remaining $n-1$ vectors.

An alternative definition is that if

$$c_1\vec{v}_1 + c_2\vec{v}_2 + \ldots c_n\vec{v}_n = \vec{0},$$

then the only set of values for all the c_i which satisfies this is

$$c_1 = c_2 = \ldots = c_n = 0.$$

Thus a matrix A has linearly independent columns if and only if the equation $A\vec{x} = \vec{0}$ has exactly one solution. Conversely, if the columns of A are not linearly independent, then $A\vec{x} = \vec{0}$ will have infinitely many solutions.

Definition 4.2 (Vector Spaces)**.**
A vector space is a set V with two operations:

- Addition of the elements of V, i.e. if $\vec{v}$, $\vec{u} \in V$, then $\vec{v} + \vec{u} \in V$
- Multiplication of the elements of V by a scalar, i.e. if $\vec{v} \in V$, and $\alpha \in \mathbb{R}$ then $\alpha\vec{v} \in V$,

which satisfies all of the following conditions:

1. $\vec{v} + \vec{u} = \vec{u} + \vec{v}$
2. $\vec{u} + (\vec{v} + \vec{w}) = (\vec{u} + \vec{v}) + \vec{w}$
3. There exists a vector $\vec{0} \in V$ such that $\vec{u} + \vec{0} = \vec{u}$ for all $\vec{u} \in V$
4. There exists a vector $\vec{1} \in V$ such that $\vec{u}\vec{1} = \vec{u}$ for all $\vec{u} \in V$
5. For any vector $\vec{v} \in V$, there exists a vector $\vec{u} \in V$ such that $\vec{v} + \vec{u} = \vec{0}$, which is denoted as $\vec{u} = -\vec{v}$
6. For any a, $b \in \mathbb{R}$ and $\vec{v} \in V$, a $(b\vec{v}) = (ab)\,\vec{v}$
7. $a\,(\vec{v} + \vec{u}) = a\vec{v} + a\vec{u}$
8. $(a + b)\,\vec{v} = a\vec{v} + b\vec{v}$

Definition 4.3 (Subspaces)**.** If V is a vector space and $W \subset V$ and W is also a vector space, then it is called a subspace of V.

Definition 4.4 (Span)**.** If $\mathcal{A} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ where each vector $\boldsymbol{v}_i \in \mathbb{R}^n$, then the span of $\mathcal{A}$ is the set of all possible linear combinations of the vectors in $\mathcal{A}$.

Definition 4.5 (Basis)**.** A basis of a vectors span is the maximal collection of linearly independent vectors from that vector space.

Thus, if you add another vector from the vectors space to the basis set, it will be a linear combination of the vectors from the basis.

The number of vectors in the basis is the **dimension of the vector space**.

Definition 4.6 (Orthogonal Basis)**.** A basis is orthogonal if the collection of all basis vectors are orthogonal.

Lemma 4.7. Any orthogonal collection of vectors is linearly independent, hence is a basis of its span.

For a vector space with basis vectors $\vec{b_i}$, the coordinates of $\vec{x}$ with respect to the basis B is the vectors of coefficients $\vec{c} = (c_1, \ldots, c_n)$ such that $\vec{x} = c_1\vec{b}_1 + \ldots + c_n\vec{b}_n$. Thus, coordinates are by solving $B\vec{c} = \vec{x}$. This requires the inverse, but if the basis vectors form an orthogonal basis, so the matrix B is orthogonal and $B^{-1} = B^T$.

Definition 4.8 (Rank)**.** The rank of a matrix A is the dimension of the column space of A.

The rank is also the number of pivots in A.

Definition 4.9 (Column and Row Spaces)**.** The column space of a matrix A is the span of all columns of A. Similarly, the row space is the span of the rows of A or the column space of A^T.

Definition 4.10 (Null spaces)**.** The null space of a matrix A is the collection of all solutions to $A\boldsymbol{x} = \boldsymbol{0}$.

Definition 4.11 (Projections)**.** The projection of a vector $\boldsymbol{v}$ onto a nonzero vector $\boldsymbol{u}$ is given by

$$\text{proj}_{\boldsymbol{u}}(\boldsymbol{v}) = \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}}\boldsymbol{u}.$$

Definition 4.12 (Gram-Schmidt)**.** Given k vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$ the Gram–Schmidt process defines the vectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_k}$ as follows:

$$\begin{aligned}
\boldsymbol{u_1} &= \boldsymbol{v_1} \\
\boldsymbol{u_2} &= \boldsymbol{v_2} - \operatorname{proj}_{\boldsymbol{u_1}}(\boldsymbol{v_2}) \\
&\vdots \\
\boldsymbol{u_k} &= \boldsymbol{v_k} - \sum_{j=1}^{k-1} \operatorname{proj}_{\boldsymbol{u_j}}(\boldsymbol{v_k})
\end{aligned}$$

The set of vectors $\boldsymbol{u_k}$ are orthogonal. Normalizing the vectors as $\boldsymbol{e_j} = \dfrac{\boldsymbol{u_j}}{\|\boldsymbol{u_j}\|}$ is a set of orthornormal vectors.

5 Eigenvalues & Eigenvectors

Determinants

For a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

the determinant, denoted by $|A|$ or $\det A$ is given by $ad - bc$.

Leibniz Method : Consider all possible choices of n elements from a matrix such that there is precisely one element chosen from each row and column. Let such a choice be denoted by σ and let $\text{sgn}\sigma$ be -1 to the power of the number of row swaps required to turn the choice σ into the diagonal. Then $\det \sigma$ is the sum of the products of each σ multiplied by its sign.

Laplace Expansion : delete the i^{th} row and the j^{th} column from a matrix to yield a $(n-1) \times (n-1)$ matrix, called a minor, denoted by M_{ij}, then the cofactor is

$$C_{ij} = (-1)^{i+j} \det M_{ij}$$

and the determinant is given by

$$\det A = a_{i1}C_{i1} + a_{i2}C_{i2} + \ldots + a_{in}C_{in}$$

which is the cofactor expansion along row i. The computation of the determinant can be performed for any row. The cofactor expansion can also be performed along any column, i.e.

$$\det A = a_{1j}C_{1j} + a_{2j}C_{2j} + \ldots + a_{nj}C_{nj}.$$

The components of the inverse of a matrix can be found via

$$\left(A^{-1}\right)_{ij} = \frac{C_{ji}}{\det A}.$$

Algorithm (Cramer's Rule). For $A\boldsymbol{x} = \boldsymbol{b}$, define the matrix B_j as the matrix A with the j^{th} column replaced by $\boldsymbol{b}$, then, if $\det A \neq 0$,

$$x_j = \frac{\det B_j}{\det A}.$$

Properties of Determinants

1. A^{-1} exists if and only if $\det A \neq 0$.
2. $\det A$ is equal to the product of the pivots of A up to a sign.
3. The determinant changes sign when two rows or columns are interchanged

4. $\det I = 1$
5. The determinant is a linear function of each row of the matrix, i.e. if one row is scaled so is the determinant, and if a row is translated by the row vector, so the determinant is also translated, i.e. Let A be a row matrix and the row $r_i \mapsto \alpha r_i + \beta \bar{r}$, then the determinant of the new matrix is given by $\det(A) \mapsto \alpha \det(A) + \beta \det\left(\bar{A}\right)$, where

$$A = \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix} \quad \text{and} \quad \bar{A} = \begin{pmatrix} r_1 \\ \vdots \\ \bar{r} \\ \vdots \\ r_n \end{pmatrix}.$$

6. If any two rows are equal the determinant is zero.
7. Subtracting one row from another does not change the determinant.
8. As a consequence, if a row is zero, the determinant is zero.
9. If a matrix is triangular, then the determinant is the product of the diagonal entries.
10. $\det(A) = \det\left(A^T\right)$.
11. $\det(AB) = \det(A) \cdot \det(B)$.

Eigenvalues and Eigenvectors

Definition 5.1 (Eigenvalues & Eigenvectors)**.** An eigenvector is a non-zero vector that has its direction unchanged by a given linear transformation. More precisely, an eigenvector, $\boldsymbol{v}$, of a linear transformation, A, is scaled by a constant factor, λ, when the linear transformation is applied to it:

$$A\boldsymbol{v} = \lambda \boldsymbol{v}.$$

Then $\boldsymbol{v}$ is called an eigenvector of A, and λ is the corresponding eigenvalue. Thus $A\boldsymbol{v}$ and $\boldsymbol{v}$ are *collinear*.

Thus,

$$A\boldsymbol{v} = \lambda \boldsymbol{v} \Leftrightarrow A\boldsymbol{v} - \lambda \boldsymbol{v}\,(A - \lambda I)\,\boldsymbol{v} = \boldsymbol{0}$$

i.e. there is a matrix B such that

$$B\boldsymbol{v} = \boldsymbol{0} \quad \text{where} \quad B = A - \lambda I.$$

As, by definition, $\boldsymbol{v}$ is not a zero vector and $B\boldsymbol{v} = \boldsymbol{0}$, then the determinant of B is zero. Finding the roots of the polynomial, called the **characteristic**

equation, $|A - \lambda I|$ yields the eigenvalues, whose eigenvectors are in the span of the nullspace of B.

For a matrix A, the column matrix of its eigenvectors $X = (\boldsymbol{v_1}\, \boldsymbol{v_2}\, \cdots\, \boldsymbol{v_n})$ and the diagonal matrix of the eigenvalues $D = \operatorname{diag} \lambda_i$ then

$$A\boldsymbol{v} = X^{-1}DX\boldsymbol{v}$$

thus, as this is true for any $\boldsymbol{v}$, then a matrix can be expressed in terms of its eigenvalues and its eigenvectors as

$$A = X^{-1}DX.$$

Then $A^2 = X^{-1}DXX^{-1}DX = X^{-1}D^2X$. Hence powers of matrices, such as A^k, can be computed for any k, as $A^k = X^{-1}D^kX$.

Note the following properties:

- If A is triangular, its eigenvalues are the entries on the diagonal.
- For an arbitrary n by n matrix A, the product of the n eigenvalues is equal to the determinant of A.
- The sum of the n eigenvalues is equal to the trace of A.

Principal Component Analysis

Theorem 2 (Eigenvalues of Symmetric Matrices). All eigenvalues of a symmetric matrix are real and positive. Furthermore, the eigenvectors can be chosen to be pairwise orthogonal.

Definition 5.2 (Principle Component Analysis)**.** Principal Component Analysis is a linear transformation of a dataset, Z, onto a new coordinate system, X, such that the directions (principal components) capturing the largest variation in the data.

Algorithm (Finding the Principal Components)**.** Shift so that the mean of each column is zero, i.e. subtract the mean of each column of Z from itself, yielding a new matrix, X.

Let $X\boldsymbol{w}$ be the projection of each data row on the direction $\boldsymbol{w}$.

Given the mean of each column is zero, so the variance of the set of column

vectors is given by

$$\begin{aligned}\operatorname{var} X &= \frac{1}{n-1}\left(x_1^2 + \ldots + x_n^2\right) \\ &= \frac{1}{n-1}\left(X\boldsymbol{w}\right)^T\left(X\boldsymbol{w}\right) \\ &= \frac{1}{n-1}\boldsymbol{w}^T X^T X\boldsymbol{w}.\end{aligned}$$

Now find the vector $\boldsymbol{w}$ so that variance is maximal.
Note that matrices of the form $A = X^T X$ are symmetric, i.e. $A^T = A$, then all eigenvalues are real and there exists a orthonormal basis given by the eigenvectors, $\boldsymbol{q_i}$ of A.

Let $Q = (\boldsymbol{q_1}\, \boldsymbol{q_2} \cdots \boldsymbol{q_n})$, with $Q^T = Q^{-1}$ and $D = \operatorname{diag}\lambda$, with eigenvalues λ_i. Then A can be expressed using the eigenvalues and eigenvectors, thus

$$X^T X = A = QDQ^T.$$

Then, the expression for the variance is given by

$$\begin{aligned}\boldsymbol{w}^T X^T X\boldsymbol{w} &= \boldsymbol{w}^T A\boldsymbol{w} \\ &= \boldsymbol{w}^T QDQ^T\boldsymbol{w} \\ &= \left(\boldsymbol{w}^T Q\right) D\left(Q^T\boldsymbol{w}\right) \\ &= \left(Q^T\boldsymbol{w}\right)^T D\left(Q^T\boldsymbol{w}\right), \quad \text{let} \quad \boldsymbol{y} = Q^T\boldsymbol{w} \\ &= \boldsymbol{y}^T D\boldsymbol{y}.\end{aligned}$$

Since $\boldsymbol{w}$ is a unit vector and Q is orthogonal, so $\boldsymbol{y}$ is also a unit vector. It can easily be shown that the vector $\boldsymbol{y} = (1, 0 \ldots 0)^T$ maximizes the variance. Thus, the corresponding principal component $\boldsymbol{w}$ is recovered from $\boldsymbol{y} = Q^T\boldsymbol{w}$, i.e.

$$\begin{aligned}\boldsymbol{w} &= \left(Q^T\right)^{-1}\boldsymbol{y} \\ &= \left(Q^T\right)^T\boldsymbol{y} \\ &= Q\boldsymbol{y}.\end{aligned}$$

6 Taylor Series

The Taylor series, or the Taylor expansion of a function, is defined as

Definition 6.1 (Taylor Series). For a function $f : \mathbb{R} \mapsto \mathbb{R}$ which is infinitely differentiable at a point c, the Taylor series of $f(c)$ is given by

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!} (x-c)^k$$

where $f^{(k)} = \dfrac{\mathrm{d}^k f}{\mathrm{d}x^k}$ is the k^{th} derivative.

This is a power series, which is convergent for some radius.

Theorem 3 (Taylor's Theorem). For a function $f \in C^{n+1}([a,b])$, i.e. f is $(n+1)$-times continuously differentiable in the interval $[a,b]$, then for some c in the interval, the function can be written as

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}$$

for some value $\xi \in [a,b]$ where

$$\lim_{\xi \to c} \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1} = 0.$$

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ which is differentiable around $\boldsymbol{a}$, then the Taylor expansion can be generalised as

$$f(\boldsymbol{x}+\boldsymbol{a}) = f(\boldsymbol{x}) + \boldsymbol{a} \cdot J + \frac{1}{2} \boldsymbol{a}^T H \boldsymbol{a} + \dots$$

where J is the Jacobian and H is the Hessian.

7 Calculus of Function of a Single Variable

Definition 7.1 (Derivative of a Function)**.** The derivative of a function $f(x)$ is given by

$$\frac{\mathrm{d}f}{\mathrm{d}x} = f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

Note the following

1. $f(x) = c \quad \Rightarrow f'(x) = 0$
2. $f(x) = x^a \quad \Rightarrow f'(x) = ax^{a-1}$
3. $f(x) = a^x \quad \Rightarrow f'(x) = a^x \ln a$
4. $f(x) = \log_b x \quad \Rightarrow f'(x) = \dfrac{1}{x \log_e b}$
5. $f(x) = \sin(x) \quad \Rightarrow f'(x) = \cos(x)$
6. $f(x) = \cos(x) \quad \Rightarrow f'(x) = -\sin(x)$

Thus for $f(x) = a^x$, when $a = e$ then, $f = e^x$ and $f'(x) = f(x) = e^x$.
Similarly, for $f(x) = \log_b x$ when $b = e$, i.e. $f(x) = \log_e x = \ln x$, so $f'(x) = \dfrac{1}{x}$.

Definition 7.2 (Summation Rule)**.** For a function of the form $f = g + h$,

$$\frac{\mathrm{d}(h+g)}{\mathrm{d}x} = h'(x) + g'(x)$$

Definition 7.3 (Product Rule)**.** For a function which is of the form $f = gh$

$$\frac{\mathrm{d}(hg)}{\mathrm{d}x} = h'(x)\,g(x) + h(x)\,g'(x)$$

Definition 7.4 (Quotient Rule)**.** For a function which is of the form $f = g/h$

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{h(x)\,g'(x) - h'(x)\,g(x)}{(h'(x))^2}$$

Definition 7.5 (Chain Rule)**.** The chain rule enables the derivative of a function which can be expressed as a composition of two differentiable functions. For a function which is of the from $f = g(h(x))$

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = g'(h)\,h'(x)$$

Another expression is

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{\mathrm{d}g}{\mathrm{d}h}\frac{\mathrm{d}h}{\mathrm{d}x}$$

This form can be understood as stating that if a function f is written in terms of g, which itself depends on the variable x (that is both f and g are dependent variables), then f depends on x as well, via the intermediate variable g.

Definition 7.6 (Critical Points)**.** If $f'(x_0) = 0$ for some x_0, then this point is called a critical point of f.

Critical points are candidates for being local maxima or minima for the function.

8 Calculus of a Function of Several Variables

Definition 8.1 (Partial Derivatives)**.** For a function with multiple input arguments, $z = f(x, y)$, the partial derivative of f with respect to x can be expressed as

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}$$

similarly, the partial derivative with respect to y can be expressed as

$$\frac{\partial f}{\partial y} = \lim_{h \to 0} \frac{f(x, y+h) - f(x, y)}{h}.$$

Note that the partial derivative is often denoted as $\frac{\partial f}{\partial x} = f_x$.

Definition 8.2 (Gradient of a Function)**.** The gradient of a function $z = f(x_1, x_2, \ldots, x_n)$ is the column vector

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)^T.$$

As the gradient ∇f depends on the point at which it is evaluated, it is denoted by $\nabla f(x_1, x_2, \ldots, x_n)$.

The gradient is the analogue to the derivative, but, as a vector, has a direction.

Definition 8.3 (Critical Point of a Function)**.** If $f(x, y)$ has a local minima or maxima at (x_0, y_0), then $\nabla f(x_0, y_0) = \vec{0}$.

Such points are called **critical points**.

Note that not all critical points are either maxima or minima. The classification of the critical points high-order derivatives.

Definition 8.4 (Second Derivatives)**.** Each partial derivative can be differentiated again to yield a second-order partial derivative,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \lim_{h \to 0} \frac{f_{x_i}(x_1, \ldots, x_j + h, \ldots, x_n) - f_{x_i}(x_1, \ldots, x_n)}{h}.$$

Thus, all second derivatives can be expressed using the **Hessian matrix**

$$H(\vec{x}) = \begin{pmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_1x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_1x_n} \end{pmatrix}.$$

If the second-order partial derivatives are continuous, then the Hessian matrix H is symmetric. Then, as symmetric all eigenvalues are real-valued, (see Theorem 2).

The Taylor expansion of a function of more than one variable is given by

$$f(\vec{x}+\vec{a}) = f(\vec{a}) + \frac{1}{2}\vec{x}^T H(\vec{a})\,\vec{x} + \dots$$

Note that the Hessian matrix can be factorized as

$$H = QDQ^{-1}$$

where D is the diagonal matrix $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and λ_i are the eigenvalues associated with eigenvector $\boldsymbol{q_i}$ of the Hessian matrix. The eigenvector $\boldsymbol{q_i}$ is the i^{th} column of the matrix Q. As the eigenvectors are orthogonal, so Q is an orthogonal matrix, thus $Q^{-1} = Q^T$. Hence,

$$f(\vec{x}+\vec{a}) = f(\vec{a}) + \frac{1}{2}\vec{x}^T QDQ^T\vec{x} + \dots$$

Letting $\vec{y} = Q^T\vec{x}$, then

$$\begin{aligned} f(\vec{x}+\vec{a}) &= f(\vec{a}) + \frac{1}{2}\vec{y}^T D\vec{y} + \dots \\ &= f(\vec{a}) + \frac{1}{2}\left(\lambda_1 y_1^2 + \dots \lambda_n y_n^2\right) + \dots \end{aligned}$$

So:

1. If all $\lambda_1, \lambda_2, \dots, \lambda_n > 0$, then the critical point is a *local minima*.
2. If all $\lambda_1, \lambda_2, \dots, \lambda_n < 0$, then the critical point is a *local maxima*.
3. If some $\lambda_j < 0$, and some $\lambda_i > 0$, then the critical point is neither a local minima nor maxima. If none of the eigenvalues are equal to zero, then the critical point is called a *saddle point*.
4. If all $\lambda_1, \lambda_2, \dots, \lambda_n \geq 0$ and at least is zero, or $\lambda_1, \lambda_2, \dots \lambda_n, \leq 0$ and at least one is zero then the test is inconclusive, as the classification of the point depends on higher derivatives.

The Hessian measures curvature. Curvature can inform how minimization methods converge.

The **Rayleigh quotient** is given by

$$R = \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

where A is Hermitian. It has a minimum at $\lambda_{\min}$, the smallest eigenvalue of A, when $\boldsymbol{x} = \boldsymbol{v}_{\min}$, which is the corresponding eigenvector. Similarly, the Rayleigh quotient has a maximum value at $\lambda_{\max}$, the largest eigenvalue of A. (Note that if $\boldsymbol{x} \in \mathbb{C}$, then the conjugate transpose is used instead).

Jacobians

If the function outputs a vector, i.e. $\boldsymbol{f} : \mathbb{R}^n \mapsto \mathbb{R}^m$, write each component of $\boldsymbol{f} = (f_1, \ldots, f_m)$ and an analogous procedures can be performed on each component of the vector-valued function. That is, a gradient can be computed for each component ∇f_i, critical points must satisfy the vector equation $\boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{0}$.

Definition 8.5 (Jacobian)**.** For a function $\boldsymbol{f} : \mathbb{R}^n \mapsto \mathbb{R}^m$ Jacobian matrix of $\boldsymbol{f}$, denoted $J_f \in \mathbb{R}^{m \times n}$, is defined as

$$J_f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

i.e. the $(i,j)^{\text{th}}$ entry is $\dfrac{\partial f_i}{\partial x_j}$.

For $f(u, v)$, let $u = g(x), v = h(x)$ then the function is a composition $F(x) = f(g(x), h(x))$. Apply the chain rule to compute the gradient

$$\nabla F = \nabla \left(f(g(x)), f(h(x)) \right) = \nabla f \frac{\partial (g, h)}{\partial x}$$

where

$$\nabla f \, \frac{\partial (g, h)}{\partial x} = \begin{pmatrix} \nabla g \\ \nabla h \end{pmatrix}$$

is the Jacobi matrix. Each row of the Jacobi matrix is a gradient of g or h.

If $f : \mathbb{R}^n \mapsto \mathbb{R}$, then

$$J = \left(\nabla f(\boldsymbol{x}) \right)^T.$$

Definition 8.6 (Directional Derivative)**.** For a vector $\boldsymbol{w}$, the **directional de-**

rivative of $f(\boldsymbol{x})$ in the direction of $\boldsymbol{w}$ is given by

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{x}) = \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{w}) - f(\boldsymbol{x})}{h}.$$

If the function is differentiable, then

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) \cdot \boldsymbol{w}.$$

As, by convention, $\nabla f(\boldsymbol{x})$ is a column vector.

The directional derivative can then be expressed as a matrix-vector product, specifically a Jacobian-vector product.

By the Cauchy-Schwarz inequality, the largest value of the directional derivative is when ∇f and $\boldsymbol{w}$ are pointing in the same direction.