# Machine Learning Primer *using R*

Ashis Kumer Biswas
February 16, 2015

## 1 Vector Space

Vector space, $V$ over a field $\mathbb{F}$ (e.g., $\mathbb{R}$) is a collection of vectors ($\vec{v} \in V$) allowing the two operations: vector addition and scalar multiplication satisfying the following properties:

- Commutativity: $\vec{a} + \vec{b} = \vec{b} + \vec{a}, \forall \vec{a}, \vec{b} \in V$
- Associativity: $\vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c}, (k_1 k_2)\vec{a} = k_1(k_2\vec{b}), \forall \vec{a}, \vec{b}, \vec{c} \in V, k_1, k_2 \in \mathbb{F}$
- Additive Identity: There exists an element $\vec{0} \in V$ such that $\vec{0} + \vec{a} = \vec{a}, \forall \vec{a} \in V$
- Additive Inverse: For every $\vec{a} \in V$, there exists $\vec{b} \in V$ such that $\vec{a} + \vec{b} = \vec{0}$.
- Multiplicative Identity: $1\vec{a} = \vec{a}, \forall \vec{a} \in V$
- Distributivity: $k_1(\vec{a} + \vec{b}) = k_1\vec{a} + k_1\vec{b}$, and $(k_1 + k_2)\vec{a} = k_1\vec{a} + k_2\vec{a}$, $\forall \vec{a}, \vec{b} \in V$, and $k_1, k_2 \in \mathbb{F}$

Vector space possesses other properties too:

- Every vector $\vec{a} \in V$ has a unique additive identity.
- Every vector $\vec{a} \in V$ has a unique additive inverse.
- $0\vec{a} = \vec{0}, \forall \vec{a} \in V$
- $k_1\vec{0} = \vec{0}, \forall k_1 \in \mathbb{F}$.
- $(-1)\vec{a} = -\vec{a}, \forall \vec{a} \in V$.

## 2 Vector Subspace

Let $V$ be a vector space over $\mathbb{F}$, and let $U \subset V$ be a subset of $V$. Then $U$ is a said to be a subspace of $V$ if $U$ is a vector space over $\mathbb{F}$ under the same operations that make $V$ into a vector space over $\mathbb{F}$. However, it will be enough to check whether the following three conditions hold.

- Additive identity: $\vec{0} \in U$
- Closure under addition: $\vec{a}, \vec{b} \in U$ implies $\vec{a} + \vec{b} \in U$
- Closure under scalar multiplication: $k_1 \in \mathbb{F}, \vec{a} \in U$ implies that $k_1\vec{a} \in U$.

## 3 Linear Span

Let $V$ denote a vector space over $\mathbb{F}$. Given vectors $\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m \in V$, and a vector $\vec{a} \in V$ is a linear combination of $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$ if there exist scalars $k_1, k_2, \cdots, k_m \in \mathbb{F}$ such that $\vec{a} = k_1\vec{a}_1 + k_2\vec{a}_2 + \cdots + k_m\vec{a}_m$.

Thus, a linear span of a set of vectors $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$ is defined as:

$$\text{span}(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m) = \{k_1\vec{a}_1 + k_2\vec{a}_2 + \cdots + k_m\vec{a}_m | k_1, k_2, \cdots, k_m \in \mathbb{F}\}$$

If $V$ is a vector space, and $\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m \in V$, then

- $\vec{a}_i \in \text{span}(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$
- $\text{span}(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$ is a subspace of $V$.
- If $U \subset V$ is a subspace such that $\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m \in U$, then $\text{span}(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m) \subset U$

## 4 Linearly Independent Vectors

Two rows, or two columns (a.k.a. vectors) of a matrix are called linearly dependent to each other if one is a multiple of the other. Three rows, or three columns of a matrix together are linearly dependent if one is linear combination of the other two. And so on... If you can not find this dependence among a set of rows or columns (a.k.a. vectors), then they are called linearly independent vectors.

$$\vec{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \vec{b} = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}, \vec{c} = \begin{pmatrix} -5 \\ 2 \\ 0 \end{pmatrix}, \vec{d} = \begin{pmatrix} 10 \\ -4 \\ 0 \end{pmatrix}, \vec{e} = \begin{pmatrix} 9 \\ -6 \\ -3 \end{pmatrix}$$

Here, $\vec{a}$ and $\vec{b}$ are linearly dependent to each other, because $\vec{b} = \vec{a} + \vec{2}$. $\vec{c}$ and $\vec{d}$ are linearly dependent to each other, because $\vec{d} = -2.\vec{c}$. $\vec{e}$ is linearly dependent to $\vec{a}$ and $\vec{d}$, because $\vec{e} = \vec{d} - \vec{a}$.

The following are the list of sets containing linearly independent vectors from the given five vectors:

- $\{\vec{a}, \vec{c}, \vec{e}\}$
- $\{\vec{a}, \vec{d}\ \}$
- $\{\vec{b}, \vec{c}, \vec{e}\}$
- $\{\vec{b}, \vec{d}, \vec{e}\}$
- $\{\vec{c}, \vec{e}\}$
- $\{\vec{d}, \vec{e}\}$

## 5 Basis

A basis of a vector space $V$ over a field $\mathbb{F}$ is a linearly independent subset of $V$ that spans $V$. All bases for finite-dimensional vector spaces have the same length. This length will then be called the dimension of the vector space.

A list of vectors $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$ is a basis for the finite-dimensional vector space $V$ if $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$ is linearly independent and $V = \text{span}(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$

If $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$ forms a basis of $V$, then every vector $\vec{a} \in V$ can be uniquely written as a linear combination of $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m)$. That is, basis has the following two properties:

- Linear independence property: if $k_1\vec{a}_1 + k_2\vec{a}_2 + \cdots + k_m\vec{a}_m = 0$, then $k_1 = k_2 = \cdots = k_m = 0, \forall k_1, k_2, \cdots, k_m \in \mathbb{F}$
- Spanning property: For every $\vec{a} \in V$ it is possible to choose $k_1, k_2, \cdots, k_m \in \mathbb{F}$ such that $\vec{a} = k_1\vec{a}_1 + k_2\vec{a}_2 + \cdots + k_m\vec{a}_m$.

Every finite-dimensional vector space has a basis.

Every linearly independent list of vectors in a finite-dimensional vector space $V$ can be extended to a basis of $V$.

Concisely, $B$ denotes a subset of a vector space $V$. Then $B$ is a basis iff any of the following equivalent conditions are met:

- $B$ is a minimal generating set of $V$, i.e., it is a generating set and no proper subset of $B$ is also a generating set.
- $B$ is a maximal set of linearly independent vectors, i.e., it is a linearly independent set but no other linearly independent set contains it as a proper subset.
- Every vector in $V$ can be expressed as a linear combination of vectors in $B$ in a unique way.

Figure 1 illustrates a geometric interpretation of the basis concept.

## 6 Rank of a Matrix

The number of linearly independent rows and/or columns is the rank of the matrix. If all of the rows, as well as all of the columns are linearly independent, then the matrix is called full rank.

A set (i.e., family) of vectors is linearly dependent if a member vector is in the linear span of the rest of the family, i.e., a member vector can be represented by a linear combination of the rest of the family.

A set of vectors which is linearly independent and spans some vector space, forms a basis for that vector space. For example, the vector space of all polynomials in $x$ over the reals has the infinite subset $\{1, x, x^2, x^3, \cdots\}$ as a basis.

Here are some more properties of rank of matrices:

- $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$, for any matrix $\mathbf{A}$.
- $\text{rank}(\mathbf{AB}) = \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$
- $\text{rank}(\mathbf{A} + \mathbf{B}) = \text{rank}(A) + \text{rank}(B)$
- $\text{rank}(\mathbf{A}) =$ number of non-zero eigenvalues of $\mathbf{A}$.

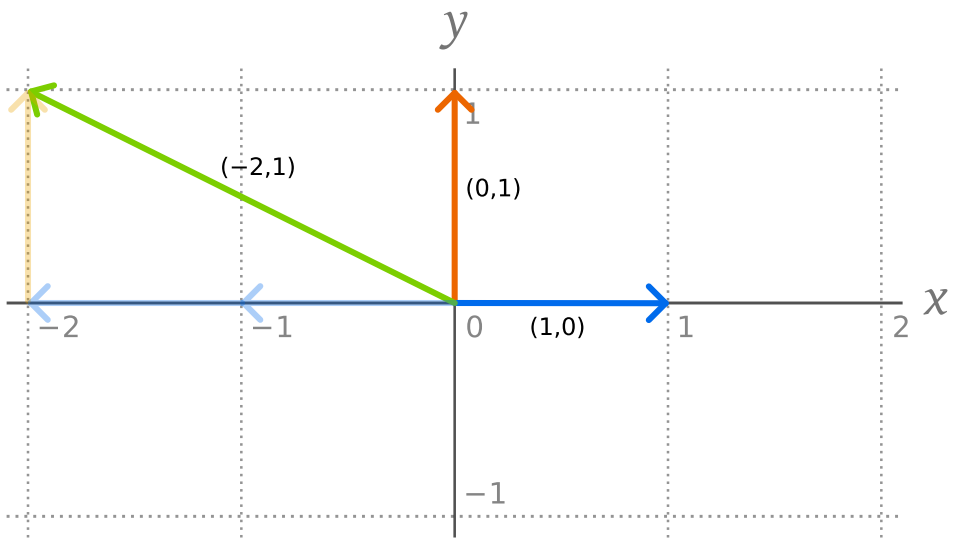



Fig. 1. This picture illustrates the standard basis in $\mathbb{R}^2$. The blue and orange vectors are the elements of the basis; the green vector can be given in terms of the basis vectors, and so is linearly dependent upon them. This picture illustrates how two vectors in $\mathbb{R}^2$ (or $\mathbb{R} \times \mathbb{R}$) can be written in terms of the standard basis. $B = \{(1, 0), (0, 1)\}$. Notice how span$(B) = \mathbb{R}^2$, and how (-2, 1) = (-2)(1,0) + (1)(0,1). Reference https://en.wikipedia.org/wiki/Basis_(linear_algebra) (Last accessed: 10/28/2014, 2:04AM)

## 7 Singular Matrix

The square matrix which is not full rank, is called a singular matrix. That is, one or more rows and/or columns of the matrix are linearly dependent. A singular matrix is not invertible. Only the non-singular matrices are invertible.

Singular matrices are also called the degenerate matrices.

A square matrix is singular iff the determinant is 0.

If a matrix $A$ is singular, so does the $A^T$. Similarly, if a matrix $A$ is non-singular, so does the matrix $A^T$.

Other properties of invertible (i.e., non-singular, or non-degenerate) matrices:

- $(A^{-1})^{-1} = A$
- $(k_1 A)^{-1} = \frac{1}{k_1} A^{-1}$ for any constant $k_1$
- $(A^T)^{-1} = (A^{-1})^T$
- For any invertible matrices $A$ and $B$, $(AB)^{-1} = B^{-1}A^{-1}$. Or, more generally, for the invertible matrices $A_1, A_2, \cdots, A_m$, $(A_1 A_2 \cdots A_m)^{-1} = A_m^{-1} \cdots A_2^{-1} A_1^{-1}$
- $\det(A^{-1}) = (\det(A))^{-1}$

## 8 Idempotent Matrix

It is a matrix when multiplied to itself produces itself [1]. The matrix $A$ is idempotent if $AA = A$. And obviously the matrix $A$ should be a square matrix. Except the identity matrix, all the idempotent matrices are singular.

## 9 Diagonal Matrix

It is a square matrix in which the entries outside the main diagonal (i.e, top-left to bottom-right) are all zero. The diagonal entries themselves may or may not be zero.

Properties of the diagonal matrices are:

- $\det(\text{diag}(a_1, a_2 \cdots, a_n)) = a_1 a_2 \cdots a_n$

## 10 Identity Matrix

It is like 1 in scalar sense. You multiply anything to 1, you get exactly what you multiplied with it. So, for any given square matrix $A \in \mathbb{R}^{n \times n}$, multiplying it with the identity matrix of same dimension would produce the same matrix $A$.

$$IA = AI = A \tag{1}$$

where,

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix}_{n \times n} , \qquad I = I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & 1 & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n \times n}$$

Here, the identity matrix is also known as the "unit matrix" [2]. Identity matrix can also be described using the diagonal matrix operator:

$$diag(I_n) = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}_n$$

If multiplication of two square matrices $X$ and $Y$ yields the identity matrix $I$, then the two matrices $X$ and $Y$ would be called the inverse matrix to each other.

$$XY = I \Rightarrow X = Y^{-1}, \text{ and } Y = X^{-1}$$

The identity matrix is its own inverse.
The identity matrix is an idempotent matrix.
The identity matrix is the only idempotent matrix having full-rank.

```
A = matrix(c(1,2,3,4,5,6,7,8,9),nrow = 3,ncol = 3,byrow= TRUE)
A #The matrix A
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
I = matrix(c(1,0,0,0,1,0,0,0,1),nrow = 3,ncol = 3, byrow =TRUE)
I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```
#You could also use the matlab command eye
library(matlab)
I = eye(3) #dimension 3x3
#Now do the multiplication
A %*% I
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
#Multiply again
I %*% A
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
#Define X matrix
X = matrix(c(1,3,2,-1,5,0,7,7,1),nrow = 3,ncol = 3,byrow= TRUE)
X #The X matrix
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    2
## [2,]   -1    5    0
## [3,]    7    7    1
```

```
#Calculate the inverse of X
Y = solve(X)
#Now, let's multiply X and Y
X %*% Y #It's identity (or very very close to the identity matrix, isn't it?)
```

```
##              [,1]          [,2] [,3]
## [1,] 1.000000e+00  5.551115e-17    0
## [2,] 1.387779e-17  1.000000e+00    0
## [3,] 0.000000e+00 -8.326673e-17    1
```

```
#Let's get the rank of a matrix
library(Matrix)
rankMatrix(A)[1] #rank of matrix A, should be == 2
```

```
## [1] 2
```

```
rankMatrix(I)[1] #rank of matrix I, should be == 3
```

```
## [1] 3
```

```
rankMatrix(X)[1] #rank of matrix X, should be == 3
```

```
## [1] 3
```

```
rankMatrix(Y)[1] #rank of matrix Y, should be == 3
```

```
## [1] 3
```

```
#Done.
```

## 11 Triangular Matrix

A square matrix is called lower triangular matrix if all the entries above the main diagonal (i.e., from top-left to bottom-right) are zero [3]. For instance,

$$L = \begin{pmatrix} l_{1,1} & & & & 0 \\ l_{2,1} & l_{2,2} & & & \\ l_{3,1} & l_{3,2} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n-1} & l_{n,n} \end{pmatrix}$$

A square matrix is called upper triangular if all the entries below the main diagonal are zero. For example,

$$U = \begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,n} \\ & u_{2,2} & u_{2,3} & \cdots & u_{2,n} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & u_{n-1,n} \\ 0 & & & & u_{n,n} \end{pmatrix}$$

A triangular matrix is one that is either lower triangular or upper traingular.

A matrix that is both upper and lower triangular matrix is called a diagonal matrix.

Determinant of a triangular matrix ($L$ or $U$) is the product of all the elements in the main diagonal. That is,

$$\det(L) = l_{1,1} \times l_{2,2} \times l_{3,3} \times \cdots \times l_{n,n} = \prod_{i=1}^{n} l_{i,i}$$

$$\det(U) = u_{1,1} \times u_{2,2} \times u_{3,3} \times \cdots \times u_{n,n} = \prod_{i=1}^{n} u_{i,i}$$

## 12 Row Echelon Form

If you apply Gaussian elimination on a matrix, the resulting matrix will be in row echelon form. A matrix is in row echelon form if

- All nonzero rows (i.e.,rows with at least one nonzero element) are above any rows of all zeros (all zero rows, if any, belong at the bottom of the matrix)
- The leading coefficient (i.e., the first nonzero number from the left, also called the pivot) of nonzero row is always strictly to the right of the leading coefficient of the row above it.
- All entries in a column below a leading entry are zeros (implied by the first two criteria).

This is an example of $3 \times 6$ matrix in row echelon form:

$$\begin{pmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix}$$

A matrix is in reduced row echelon form (a.k.a., row canonical form) if it satisfied the following conditions:

- It is in row echelon form
- Every pivot (i.e., leading coefficient of the row) is 1, and is the only nonzero entry in its column.

You can create the reduced row echelon form of a matrix from its echelon form: beginning with the rightmost leading entry, and working upwards to the left, create zeros above each leading entry and scale rows to transform each leading entry into 1.

Let's work with the $3 \times 6$ matrix above:

$$\begin{pmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix} \xrightarrow{9R_2+R_1 \to R_1} \begin{pmatrix} 3 & 0 & -6 & 9 & 0 & -72 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 0 & -6 & 9 & 0 & -72 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix} \xrightarrow{\frac{1}{3}R_1 \to R_1} \begin{pmatrix} 1 & 0 & -2 & 3 & 0 & -24 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix}$$

## 13 Gauss-Jordan Elimination Algorithm

This method is a sequence of operations performed on the coefficient matrix corresponding to a system of linear equations to solve it [4]. However, the method is also useful –

- To find rank of a matrix
- To calculate the determinant of a matrix
- To calculate the inverse of an invertible square matrix.

To perform row reduction on a matrix, one uses a sequence of elementary row operations to modify the matrix until the lower left-hand corner of the matrix is filled with zeros. There are three types of elementary row operations:

1) **Row Switching**: A row within the matrix can be switched to another row. $R_i \leftrightarrow R_j$.
2) **Row Multiplication**: Each element in a row can be multiplied by a non-zero constant. That is, $kR_i \to R_i$, where $k \neq 0$.
3) **Row Addition**: A row can be replaced by the sum of that row and a multiple of another row. That is, $R_i + kR_j \to R_i$, where $i \neq j$

Using these operations, a matrix can always be transformed into an upper triangular matrix, and in fact one that is in row echelon form. The steps required upto this is called Gaussian Elimination. However, the steps of row reduction until the matrix becomes the reduced row echelon form is called Gauss-Jordan Elimination, to distinguish it from stopping after reaching the echelon form.

### *13.1 Solving system of linear equations*

Given the following system of linear equations:

$$2x + y - z = 8 \quad (L_1) \tag{2}$$
$$-3x - y + 2z = -11 \quad (L_2) \tag{3}$$
$$-2x + y + 2z = -3 \quad (L_3) \tag{4}$$

The table below is the row reduction process applied to the associated augmented matrix.

$$\left[\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ -3 & -1 & 2 & -11 \\ -2 & 1 & 2 & -3 \end{array}\right] \xrightarrow[L_3+L_1\to L_3]{L_2+\frac{3}{2}L_1\to L_2} \left[\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 2 & 1 & 5 \end{array}\right]$$

$$\left[\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 2 & 1 & 5 \end{array}\right] \xrightarrow{L_3+(-4)L_2\to L_3} \left[\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & -1 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & -1 & 1 \end{array}\right] \xrightarrow[L_1-L_3\to L_1]{L_2+\frac{1}{2}L_3\to L_2} \left[\begin{array}{ccc|c} 2 & 1 & 0 & 7 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & -1 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccc|c} 2 & 1 & 0 & 7 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & -1 & 1 \end{array}\right] \xrightarrow[-L_3\to L_3]{2L_2\to L_2} \left[\begin{array}{ccc|c} 2 & 1 & 0 & 7 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

$$\left[\begin{array}{ccc|c} 2 & 1 & 0 & 7 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \end{array}\right] \xrightarrow[\frac{1}{2}L_1\to L_1]{L_1-L_2\to L_1} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

So, the solution to the given system is:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$$

### *13.2 Computing Determinant of matrix*

Let's look at how the elementary row operations on a square matrix $A \in \mathbb{R}^{n\times n}$ affect the determinant of that matrix:

- **Row switching**: Everytime you switch any two rows $R_i$ and $R_j$ of matrix $A$, you will have the same determinant with opposite sign, i.e., $\det(A) \xrightarrow[R_j\to R_i]{R_i\to R_j} (-1)\times\det(A)$.
- **Row multiplication**: Multiplying a row by a nonzero scalar $k$ multiplies the determinant by the same scalar. That is, $\det(A) \xrightarrow{kR_i} k\det(A)$
- **Row addition**: Adding to one row a scalar multiple of another does not change the determinant. That is, $\det(A) \xrightarrow{R_i+kR_j\to R_i} \det(A)$

Now let's do two examples.

Example 1: Compute the determinant of

$$A = \begin{bmatrix} 3 & -17 & 4 \\ 0 & 5 & 1 \\ 0 & 0 & -2 \end{bmatrix}$$

Ans: Since the matrix is upper triangular, the determinant is very easy to calculate, which is the product of all the elements of its main diagonal. So, $\det(A) = 3 \times 5 \times (-2) = -30$

Example 2: Compute the determinant of

$$B = \begin{bmatrix} 1 & -3 & 0 \\ -2 & 4 & 1 \\ 5 & -2 & 2 \end{bmatrix}$$

Ans: Let's apply Gaussian Elimination to find the determinant.

$$\det\left(\begin{bmatrix} 1 & -3 & 0 \\ -2 & 4 & 1 \\ 5 & -2 & 2 \end{bmatrix}\right) \xrightarrow[R_3-5R_1\to R_3]{R_2+2R_1\to R_2} \det\left(\begin{bmatrix} 1 & -3 & 0 \\ 0 & -2 & 1 \\ 0 & 13 & 2 \end{bmatrix}\right)$$

$$\det\left(\begin{bmatrix} 1 & -3 & 0 \\ 0 & -2 & 1 \\ 0 & 13 & 2 \end{bmatrix}\right) \xrightarrow{R_3+\frac{13}{2}R_2\to R_3} \det\left(\begin{bmatrix} 1 & -3 & 0 \\ 0 & -2 & 1 \\ 0 & 0 & \frac{17}{2} \end{bmatrix}\right)$$

Since, the echelon form is an upper triangular matrix, then the determinant is given by $1 \times (-2) \times \frac{17}{2} = -17$.
Here are some properties of determinant operations:

- If $\mathbf{A}$ is a square matrix, then $det(\mathbf{A}^T) = det(\mathbf{A})$
- $det(k\mathbf{A}) = k^n det(\mathbf{A})$, for $\mathbf{A}$ is a square matrix and $k$ is a scalar.
- $det(\mathbf{AB}) = det(\mathbf{A})det(\mathbf{B})$, for two square matrices $\mathbf{A}, \mathbf{B}$ of the same dimension.
- A square matrix $\mathbf{A}$ is invertible, i.e., is non-singular if and only if $det(\mathbf{A}) \neq 0$. However, if any row or column of the matrix $\mathbf{A}$ is all zeros, or any two or more rows or columns of $\mathbf{A}$ are all equal, or multiple of each other, then $det(\mathbf{A}) = 0$.
- $det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i = \lambda_1 \lambda_2 \cdots \lambda_n$, where $\lambda_i$ is the $i^{\text{th}}$ eigenvalue of the matrix $\mathbf{A}$.

### *13.3 Computing rank of matrix*

The Gaussian elimination algorithm can be applied to any $m \times n$ matrix $A$. In this way, for example, some $6 \times 9$ matrices can be transformed to a matrix that has a row echelon form like:

$$T = \begin{pmatrix} a & * & * & * & * & * & * & * & * \\ 0 & 0 & b & * & * & * & * & * & * \\ 0 & 0 & 0 & c & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & d & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & e \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the $*$'s are arbitrary entries and $a, b, c, d, e$ are non-zero entries. This echelon matrix $T$ contains a wealth of information about $A$, which are–

- rank$(A) = 5$, since there are 5 non-zero rows in $T$.
- The vector space spanned by the columns of $A$ has a basis consisting of the first, third, fourth, seventh and ninth columns of $A$ (i.e., the columns of $a, b, c, d, e$ in $T$), and the $*$'s tell you how the other columns of $A$ can be written as linear combinations of the basis columns.

Now, let's do an example: Example: compute rank of the following matrix $B$:

$$B = \begin{pmatrix} 2 & 1 & 3 & 2 \\ 3 & 2 & 5 & 1 \\ -1 & 1 & 0 & -7 \\ 3 & -2 & 1 & 17 \\ 0 & 1 & 1 & -4 \end{pmatrix}$$

Ans: We'll apply Gaussian Elimination to find out the rank of the $B$ matrix.

$$\begin{pmatrix} 2 & 1 & 3 & 2 \\ 3 & 2 & 5 & 1 \\ -1 & 1 & 0 & -7 \\ 3 & -2 & 1 & 17 \\ 0 & 1 & 1 & -4 \end{pmatrix} \xrightarrow[R_3+\frac{1}{2}R_1\to R_3]{R_2+3R_3\to R_2,\ R_4+3R_3\to R_4} \begin{pmatrix} 2 & 1 & 3 & 2 \\ 0 & 5 & 5 & -20 \\ 0 & \frac{3}{2} & \frac{3}{2} & -6 \\ 0 & 1 & 1 & -10 \\ 0 & 1 & 1 & -4 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 & 3 & 2 \\ 0 & 5 & 5 & -20 \\ 0 & \frac{3}{2} & \frac{3}{2} & -6 \\ 0 & 1 & 1 & -10 \\ 0 & 1 & 1 & -4 \end{pmatrix} \xrightarrow[R_4+(-1)R_5\to R_4,\ R_5+(-1)R_4\to R_5]{R_3+\left(\frac{-3}{2}\right)R_4\to R_3} \begin{pmatrix} 2 & 1 & 3 & 2 \\ 0 & 5 & 5 & -20 \\ 0 & 0 & 0 & 15 \\ 0 & 0 & 0 & -6 \\ 0 & 0 & 0 & 6 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 & 3 & 2 \\ 0 & 5 & 5 & -20 \\ 0 & 0 & 0 & 15 \\ 0 & 0 & 0 & -6 \\ 0 & 0 & 0 & 6 \end{pmatrix} \xrightarrow[R_4+R_5\to R_4,\ R_5+R_4\to R_5]{R_3+\frac{15}{6}R_4\to R_3} \begin{pmatrix} \mathbf{2} & 1 & 3 & 2 \\ 0 & \mathbf{5} & 5 & -20 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

We see that there are only two pivots in the echelon matrix, so the rank$(B) = 2$. And hence, the first and second columns (i.e., the columns containing the two pivots) form the basis. The two basis vectors (first & second columns of $B$ in this example) spans all the vectors represented in the matrix $B$. We can represent all the other columns of $B$ by linear combination of the two basis vectors, $\vec{v}_1 = \begin{pmatrix} 2 \\ 3 \\ -1 \\ 3 \\ 0 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 1 \\ 2 \\ 1 \\ -2 \\ 1 \end{pmatrix}$. Let's try it:

- The third column of $B$ = $\begin{pmatrix} 3 \\ 5 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \vec{v}_1 + \vec{v}_2 = \begin{pmatrix} 2 \\ 3 \\ -1 \\ 3 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 1 \\ -2 \\ 1 \end{pmatrix}$
- The fourth column of $B$ = $\begin{pmatrix} 2 \\ 1 \\ -7 \\ 17 \\ -4 \end{pmatrix} = 3\vec{v}_1 - 4\vec{v}_2 = 3\begin{pmatrix} 2 \\ 3 \\ -1 \\ 3 \\ 0 \end{pmatrix} - 4\begin{pmatrix} 1 \\ 2 \\ 1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 9 \\ -3 \\ 9 \\ 0 \end{pmatrix} + \begin{pmatrix} -4 \\ -8 \\ -4 \\ 8 \\ -4 \end{pmatrix}$

*13.4 Computing inverse of matrix*

The variant of the Gaussian elimination, called the Gauss-Jordan elimination algorithm can be used to find the inverse of a matrix, $A \in \mathbb{R}^{n\times n}$, if the matrix is invertible.

First, the $n \times n$ identity matrix $I$ is augmented to the right of $A$, forming a block matrix $[A|I] \in \mathbb{R}^{n\times 2n}$.

Then by applying elementary row operations on the block matrix, we find the "reduced row echelon form".

The matrix $A$ is invertible if and only if the left block can be reduced to the identiy matrix $I$; in this case the right block of the final matrix is the result inverse matrix of $A$, which is $A^{-1}$. If the algorithm can not reduce the left block to $I$, then $A$ must not be invertible.

Let's work with an example. Consider the matrix $A$:

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

We need to find out $A^{-1}$ using the Gauss-Jordan Elimination Algorithm. Let's first build the augmented block matrix, and perform the elementary row operations on it:

$$
\begin{aligned}
[A|I] = \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right] &\xrightarrow{\frac{1}{2}R_1+R_2\to R_2} \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & \frac{1}{2} & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right] \\
&\xrightarrow{\frac{3}{2}R_3+R_2\to R_3} \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & \frac{1}{2} & 1 & 0 \\ 0 & 0 & 2 & \frac{1}{2} & 1 & \frac{3}{2} \end{array}\right] \\
&\xrightarrow{\frac{1}{2}R_3+R_2\to R_2} \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & \frac{3}{2} & 0 & \frac{3}{4} & \frac{3}{2} & \frac{3}{4} \\ 0 & 0 & 2 & \frac{1}{2} & 1 & \frac{3}{2} \end{array}\right] \\
&\xrightarrow{\frac{2}{3}R_2+R_1\to R_1} \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & \frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{3}{2} & 0 & \frac{3}{4} & \frac{3}{2} & \frac{3}{4} \\ 0 & 0 & 2 & \frac{1}{2} & 1 & \frac{3}{2} \end{array}\right] \\
&\xrightarrow[\frac{1}{2}R_3\to R_3]{\substack{\frac{1}{2}R_1\to R_1 \\ \frac{2}{3}R_2\to R_2}} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{3}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 1 & 0 & \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} \end{array}\right] = [I|A^{-1}]
\end{aligned}
\tag{5}
$$

So, the inverse of $A$ is:

$$A^{-1} = \begin{bmatrix} \frac{3}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{3}{4} \end{bmatrix}$$

## 14 Inverse of Matrix

Suppose, we have two matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times n}$, then if $AB = BA = I_n$ holds, that is, their multiplication yields identity matrix, then we say $A$ is the inverse of $B$, or vice versa.

For instance, $A = \begin{pmatrix} 4 & 3 \\ 3 & 2 \end{pmatrix}$, and its inverse (written as $A^{-1}$) is $A^{-1} = \begin{pmatrix} -2 & 3 \\ 3 & -4 \end{pmatrix}$ because, you can verify that $AA^{-1} = A^{-1}A = I$.

We can use the Gauss-Jordan Elimination method to compute inverse of a matrix.

Let $A \in \mathbb{R}^{n\times n}$ over a field $K$ (e.g., the field $\mathbb{R}$ of real numbers), then below are the properties either all TRUE, or all FALSE [5]:

- $A$ is invertible (i.e., non-singular, or non-degenerate)
- $A$ has $n$ pivot positions
- $\det(A) \neq 0$
- $A$ has full rank, i.e., $\text{rank}(A) = n$
- The equation $A\vec{x} = 0$ has only the trivial solution $\vec{x} = \vec{0}$
- The equation $A\vec{x} = \vec{b}$ has exactly one solution for each $\vec{b}$ in the $n$-dimensional field $K^n$.
- The columns of $A$ are linearly independent
- The columns of $A$ span $K^n$
- The columns of $A$ form a basis of $K^n$
- The transpose matrix $A^T$ is also an invertible matrix
- The number 0 is not an eigenvalue of $A$.

Here are some other properties of invertible matrix $A$:

- $(A^{-1})^{-1} = A$
- $(kA)^{-1} = k^{-1}A^{-1} = \frac{1}{k}A^{-1}$ for any nonzero scalar $k$.
- $(A^T)^{-1} = (A^{-1})^T$
- For any invertible $n \times n$ matrices $A$ and $B$: $(AB)^{-1} = B^{-1}A^{-1}$. More generally, if $A_1, A_2, \cdots, A_k$ are invertible $n \times n$ matrices, then $(A_1A_2 \cdots A_k)^{-1} = A_k^{-1}A_{k-1}^{-1} \cdots A_2^{-1}A_1^{-1}$
- $\det(A^{-1}) = (\det(A))^{-1} = \frac{1}{\det(A)}$

A matrix that is its own inverse, i.e., $A = A^{-1}$ and $A^2 = I$, is called an involution.

## 15 System of Linear Equations

A set of linear equations can be formulated using Matrices. For instance, consider the two equations of two variables $x$ and $y$.

$$3x + 2y = 7 \tag{6}$$

$$-6x + 6y = 6 \tag{7}$$

Left side of the equal signs are the linear combination of the two variables, and the right side of the equal signs are the constants. Let's prepare the coefficient matrix (i.e., the coefficients of the variables in each equation), and the constant vector. So, the matrix system becomes:

$$\begin{pmatrix} 3 & 2 \\ -6 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

Then, multiplying both sides by the inverse of the coefficient matrix, we get

$$\begin{pmatrix} 3 & 2 \\ -6 & 6 \end{pmatrix}^{-1} \begin{pmatrix} 3 & 2 \\ -6 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ -6 & 6 \end{pmatrix}^{-1} \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

Then, we have:

$$I. \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ -6 & 6 \end{pmatrix}^{-1} \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.2 & -0.067 \\ 0.2 & 0.1 \end{pmatrix} \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

So, the final solution of the linear equations becomes:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

```r
A = matrix(c(3,2,-6,6),nrow=2,ncol=2,byrow=T) #the coefficient Matrix
A
```

```
##      [,1] [,2]
## [1,]    3    2
## [2,]   -6    6
```

```r
b = as.matrix(c(7,6))
b
```

```
##      [,1]
## [1,]    7
## [2,]    6
```

```r
Ainv = solve(A) #Inverse of A
Ainv
```

```
##      [,1]        [,2]
## [1,]  0.2 -0.06666667
## [2,]  0.2  0.10000000
```

```r
solution = Ainv %*% b
solution #each row is the solution to the corresponding variable
```

```
##      [,1]
## [1,]    1
## [2,]    2
```

```r
#You could use the solve() function directly here:
solve(A,b) #solution
```

```
##      [,1]
## [1,]    1
## [2,]    2
```

## 16 Dot Product of two vectors

It is an operation that takes two equal-length vecors and returns a scalar. It is also known as scalar product, or sometimes inner product in the context of the Euclidean space.

The dot product of two vectors $\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$ is defined as:

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i$$

In Euclidean space, the dot product of the vectors $\vec{a}$ and $\vec{b}$ is defined as:

$$\vec{a} \cdot \vec{b} = ||\vec{a}||||\vec{b}||\cos\theta$$

where, $\theta$ is the angle between the two vectors $\vec{a}$ and $\vec{b}$.

If $\vec{a}$ and $\vec{b}$ are orthogonal, then the angle between them is $90^o$, and their dot product becomes zero. That is,

$$\vec{a} \cdot \vec{b} = ||\vec{a}||||\vec{b}||\cos 90^o = ||\vec{a}||||\vec{b}||0 = 0$$

On the other hand, if the two vectors are co-directional, then the angle between them is $0^o$, and

$$\vec{a} \cdot \vec{b} = ||\vec{a}||||\vec{b}||\cos 0^o = ||\vec{a}||||\vec{b}||1 = ||\vec{a}||||\vec{b}||$$

This co-directionality of the two vectors implies that,

$$\vec{a} \cdot \vec{a} = ||\vec{a}||||\vec{a}|| = ||\vec{a}||^2$$

This implies:

$$||\vec{a}|| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

This the fomula to compute the Euclidean length of a vector.

## 17 Unit vector

A vector having length (i.e., the Euclidean length) equal to 1. The normalized vector of a non-zero vector $\hat{\mathbf{v}}$ is the unit vector co-directional with $\mathbf{v}$, i.e.,

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{||\mathbf{v}||}$$

## 18 Matrix Product

If $A \in \mathbb{R}^{m \times n}$, and $B \in \mathbb{R}^{n \times k}$, their product $AB \in \mathbb{R}^{m \times k}$

Properties of matrix product:

- Matrix products are associative: $AB\mathbf{x} = A(B\mathbf{x})$, or $(AB)C = A(BC)$ for matrices $A, B, C$ and vector $\mathbf{x}$.
- Matrix products are NOT commutative: $AB \neq BA$.
- Matrix products are distributive: $A(B + C) = AB + AC$, or $(B + C)A = BA + CA$

## 19 Transpose of a Matrix

Let's consider a matrix $A \in \mathbb{R}^{m \times n}$,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \text{then the transpose of A, } A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix}$$

That is, first row of $A$ becomes first column of $A^T$, and so on.

```
A = matrix(c(1,2,3,4,5,6,7,8,9),nrow=3,ncol=3,byrow=T)
A
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
t(A) #transpose of A
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

Properties of transpose operator:

- For any matrix $A$, $(A^T)^T = A$
- $\det(A) = \det(A^T)$
- $(AB)^T = B^T A^T$, $(ABC)^T = C^T B^T A^T$, or more generally: for matrices $A_1, A_2, \cdots, A_k$, $(A_1 A_2 \cdots A_k)^T = A_k^T \cdots A_2^T A_1^T$
- $(A + B)^T = A^T + B^T$
- $(A^{-1})^T = (A^T)^{-1}$

## 20 Eigenvalues and Eigenvectors

An eigenvector of a square matrix $A$ is a non-zero vector $\vec{v}$ that, when the matrix multiplies $\vec{v}$, yields the same result as when some scalar multiplies $\vec{v}$. The scalar multiplier is denoted by $\lambda$, which is called the eigenvalue of $A$ corresponding to the eigenvector $\vec{v}$. That is:

$$A\vec{v} = \lambda\vec{v}$$

The set of all eigenvectors of a matrix, each paired with its corresponding eigenvalue, is called the eigensystem of that matrix [6].

Any multiple of an eigenvector is also an eigenvector, with the same eigenvalue.

An eigenspace of a matrix $A$ is the set of all eigenvectors with the same eigenvalue, together with the zero vector. An eigenbasis for $A$ is any basis for the set of all vectors that consists of linearly independent eigenvectors of $A$. Not every matrix has an eigenbasis, but every symmetric matrix does.

The eigenvalue equation for a matrix $A$ is

$$A\vec{v} = \lambda\vec{v}$$

which is equivalent to

$$A\vec{v} - \lambda\vec{c} = 0$$

$$(A - \lambda I)\vec{v} = 0$$

where, $I$ is an $n \times n$ identity matrix. We know that in an equation like $M\vec{v} = 0$ has a non-zero solution for $\vec{v}$ iff the determinant of the matrix $M$ is zero. It follows that the eigenvalues of $A$ are precisely the real numbers $\lambda$ that satisfy the equation:

$$\det(A - \lambda I) = 0$$

The left-side of the equation can be seen as a polynomial function of the variable $\lambda$. The degree of this polynomial is $n$, the order of the matrix. This polynomial $\det(A - \lambda I)$ is called the characteristic polynomial of $A$, and the above equation is the characteristic equation of $A$.

For example, consider finding the eigenvalues and eigenvectors for the square matrix $A$:

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}$$

Let's first build the characteristic equation:

$$\begin{aligned}
\det(A - \lambda I) &= 0 \\
\det(\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}) &= 0 \\
\det(\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}) &= 0 \\
\det(\begin{pmatrix} 1-\lambda & 2 \\ 4 & 3-\lambda \end{pmatrix}) &= 0 \\
3 - 4\lambda + \lambda^2 - 8 &= 0 \\
\lambda^2 - 4\lambda - 5 &= 0 \\
\lambda^2 - 5\lambda + \lambda - 5 &= 0 \\
\lambda(\lambda - 5) + 1(\lambda - 5) &= 0 \\
(\lambda + 1)(\lambda - 5) &= 0
\end{aligned}$$

So, there are two eigenvalues, $\lambda_1 = 5$, and $\lambda_2 = -1$ (sorted by descending order of the values).

Now, let's find out the two eigenvectors corresponding to the two eigenvalues we got.

$$A\vec{v} = \lambda\vec{v}$$
$$A\vec{v} - \lambda\vec{v} = 0$$
$$(A - \lambda I)\vec{v} = 0 \tag{8}$$
$$\tag{9}$$

For the eigenvalue $\lambda_1 = 5$, we get:

$$(A - (5)I)\vec{v} = 0$$
$$(\begin{pmatrix}1 & 2\\ 4 & 3\end{pmatrix} - \begin{pmatrix}5 & 0\\ 0 & 5\end{pmatrix})\begin{pmatrix}v_x\\ v_y\end{pmatrix} = 0$$
$$\begin{pmatrix}-4 & 2\\ 4 & -2\end{pmatrix}\begin{pmatrix}v_x\\ v_y\end{pmatrix} = 0$$
$$\begin{pmatrix}-4v_x + 2v_y\\ 4v_x - 2v_y\end{pmatrix} = 0$$

The two equations $-4v_x + 2v_y = 0$ and $4v_x - 2v_y = 0$ have infinitely many solutions. And, $v_x = \frac{1}{2}v_y = 0$. If $v_x = t, v_y = 2t$. So, one eigenvector for the eigenvalue $\lambda_2 = 5$ is $\begin{pmatrix}1\\ 2\end{pmatrix}$ And the eigenspace $E_{\lambda=5} = \text{span}(\begin{pmatrix}1\\ 2\end{pmatrix})$

Similarly, for the eigenvalue $\lambda_2 = -1$, from equation 8 we get,

$$(A - (-1)I)\vec{v} = 0$$
$$(\begin{pmatrix}1 & 2\\ 4 & 3\end{pmatrix} - \begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix})\begin{pmatrix}v_x\\ v_y\end{pmatrix} = 0$$
$$\begin{pmatrix}2 & 2\\ 4 & 4\end{pmatrix}\begin{pmatrix}v_x\\ v_y\end{pmatrix} = 0$$
$$\begin{pmatrix}2v_x + 2v_y\\ 4v_x + 4v_y\end{pmatrix} = 0$$

The two equations $2v_x + 2v_y = 0$ and $4v_x + 4v_y = 0$ have infinitely many solutions. And, $v_x + v_y = 0 \Rightarrow v_x = -v_y$. If $v_x = t, v_y = -t$. So, one eigenvector for the eigenvalue $\lambda_2 = -1$ is $\begin{pmatrix}1\\ -1\end{pmatrix}$ And the eigenspace $E_{\lambda=-1} = \text{span}(\begin{pmatrix}1\\ -1\end{pmatrix})$

```
A = matrix(c(1,2,4,3),nrow=2,ncol=2,byrow=T)
A
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    4    3
```

```
result = eigen(A) #compute eigenvalue & eigenvectors of A
result$val #eigenvalues of A
```

```
## [1]  5 -1
```

```
result$vec #eigenvectors of the corresponding eigenvalues of A
```

```
##            [,1]       [,2]
```

```
## [1,] -0.4472136 -0.7071068
## [2,] -0.8944272  0.7071068
```

Other properties of the eigenvalues and eigenvectors of matrix $A$ with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$:

- trace$(A) = \sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i = \lambda_1 + \lambda_2 + \cdots + \lambda_n$
- det$(A) = \prod_{i=1}^{n} \lambda_i = \lambda_1 \lambda_2 \cdots \lambda_n$
- The eigenvalues of $A^k$ for any positive integer $k$ are: $\lambda_1^k, \lambda_2^k, \cdots, \lambda_n^k$.
- If $A$ is invertible, then the eigenvalues of $A^{-1}$ are: $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots, \frac{1}{\lambda_n}$

## 21 Norm of vectors

A norm is a function that assigns a strictly positive length or size to each vector in a vector space, other than the zero vector which has zero length assigned to it.

### *21.1 Euclidean norm of a vector*

A.k.a., $L^2$-norm, $L^2$-distance, $\ell^2$-norm, $\ell^2$-distance, Euclidean length. And is defined for an $n$-dimensional vector $\vec{x} = (x_1, x_2, \cdots, x_n)^T$ on $\mathbb{R}^n$ as:

$$||\vec{x}|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{\vec{x}^T \vec{x}}$$

Thus, the Euclidean norm of the vector $\vec{x}$ is the square root of the inner-product of the vector $\vec{x}$ to itself. So, square of the Euclidean norm $||\vec{x}||^2 = \vec{x}^T \vec{x}$, simply the inner-produc of the vector $\vec{x}$ to itself.

### *21.2 Manhattan norm of a vector*

A.k.a., taxicab norm, $L_1$-norm, $L_1$-distance, Manhattan-distance.

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

The name relates to the distance a taxi has to drive in a rectangular street grid to get from the origin to the point $x$.

### *21.3 p-Norm of a vector*

For any real number $p \geq 1$:

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

### *21.4 Maximum Norm of a vector*

A.k.a., special case of inifinity norm, uniform norm or supremum norm. It is defined as:

$$||\vec{x}||_\infty = \max\left(|x_1|, |x_2|, \cdots, |x_n|\right)$$

## 22 Norm of Matrices

The norm of a square matrix $A$ is a non-negative real number denoted by $||A||$. There are several different ways of defining a matrix norm, but they all share the following properties:

- $||A|| \geq 0$, for any square matrix $A$.
- $||A|| = 0$ iff $A = 0$
- $||kA|| = |k|||A||$, for any scalar $k$ and any square matrix $A$.
- $||A + B|| \leq ||A|| + ||B||$
- $||AB|| \leq ||A||||B||$

The norm of a matrix is a measure of how large its elements are. It is a way of determining the size of a matrix which is not necessarilty related to how many rows or columns the matrix has. So, the norm of a matrix is a non-negative real number that is a measure of the magnitude of the matrix.

### *22.1 $L_1$-norm of a matrix*

It is the maximum absolute column sum.

$$||A||_1 = \max_{1 \le j \le n} \left( \sum_{i=1}^{m} |a_{ij}| \right)$$

### *22.2 Inifinity-norm of a matrix*

It is the maximum absolute row sum.

$$||A||_1 = \max_{1 \le i \le m} \left( \sum_{j=1}^{n} |a_{ij}| \right)$$

### *22.3 Euclidean-norm of a matrix*

A.k.a., Frobenius norm of a matrix. It is the square root of the sum of all the squares of the elements.

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij})^2} = \sqrt{\text{trace}(AA^T)} = \sqrt{\text{trace}(A^T A)}$$

And the square of the Frobenius norm of the matrix $A$ is:

$$||A||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij})^2 = \text{trace}(AA^T) = \text{trace}(A^T A)$$

### *22.4 $L_{2,1}$-norm of a matrix*

Let $(\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_n)$ be the columns of matrix $A$. The $L_{2,1}$ norm [7] is a sum of Euclidean norm of columns:

$$||A||_{2,1} = \sum_{j=1}^{n} ||\vec{a_j}||_2 = \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

$L_{2,1}$ norm can be generalized into $L_{p,q}$ norm:

$$||A||_{p,q} = \left[ \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |a_{ij}|^p \right)^{\frac{q}{p}} \right]^{\frac{1}{q}}$$

```
A = matrix(c(1,2,3,4,5,6,7,8,9),nrow=3,ncol=3,byrow=T)
A
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
frobNorm = sqrt(sum(rowSums(A^2))) #Frobenius norm
frobNorm
```

```
## [1] 16.88194
```

```
norm(A,type="F") #Alternatively, use R function norm()
```

```
## [1] 16.88194
```

```
#L_{p,q}-norm of the matrix
p = 2
```

```
q = 1
L_pq_norm = sum(colSums(abs(A)^(q/p))^(1/q)  #L_p,q norm
L_pq_norm

## [1] 19.306
```

## 23 Derivative

The fundamental definition of derivative of the function $f(x)$ is:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{10}$$

OR,

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

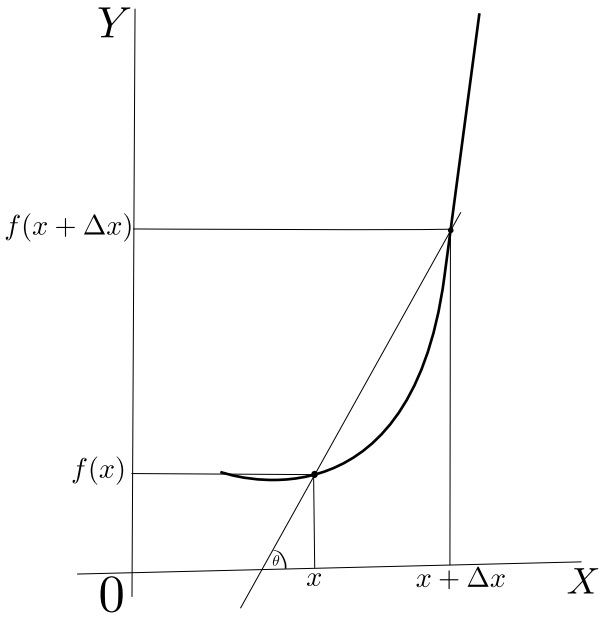



Fig. 2. Understanding the derivative of a function as the slope of the secant line going through the two points on the function $(x, f(x))$ and $((x + \Delta x), f(x + \Delta x))$.

Let's take a look at Figure 2. The slope of the line connecting two points $(x, f(x))$ and $((x + \Delta x), f(x + \Delta x))$ is:

$$\tan(\theta) = m = \frac{\text{change in } y}{\text{change in } x} = \frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - x} = \frac{f(x + \Delta x) - f(x)}{(\Delta x)}$$

But, the derivative tends to find the slope of the function (i.e., the rate of change of the function in y) on a given point in $x$, unlike the two given points in the example above. How can we accomplish that?

We can find that value using the concept of limits. That is, we shrink the small difference $\Delta x$ towards 0 (Equation 10). It's that simple! Thus, the derivative at a point $x$ of a function $f(x)$ is the slope of the function at that point $x$ (Figure 3).

More intuitive explanations about the derivative can be found in MathIsFun blog [8] and the Khalid Azad's blog [9].

Example 1: What is $\frac{d}{dx}x^3$? In other words, let's say, $f(x) = x^3$, then what is $f'(x)$?

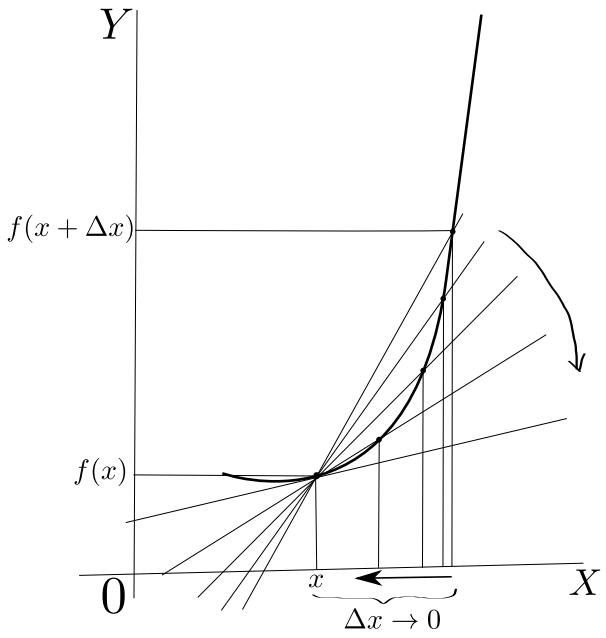



Fig. 3. Deriviative is the slope of the function $f(x)$ at point $x$.

$$\begin{aligned} f'(x) &= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{(x+\Delta x)^3 - x^3}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\cancel{x^3} + 3x^2\Delta x + 3x\Delta x^2 + \Delta x^3 - \cancel{x^3}}{\Delta x} \\ &= \lim_{\Delta x \to 0} 3x^2 + 3x\Delta x + \Delta x^2 \\ &= 3x^2 + 0 + 0 \\ &= 3x^2 \end{aligned}$$

**The first and the second derivatives** [10]: The first derivative of the function $f(x)$, which is $f'(x)$ or $\frac{df}{dx}$, is the slope of the tangent line to the function at the point $x$. Essentially, the first derivative tells us whether a function is increasing or decreasing, and by how much it is increasing or decreasing in terms of slope. A positive slope tells us that as $x$ increases, $f(x)$ also increases. Negative slope tells us that, as $x$ increases, $f(x)$ decreases. However, a zero slope does not tell us anthing in particular: the function may be increasing, decreasing, or at a local maximum or a local minimum at that point. So, we find:

- if $f'(p) = \left.\frac{df}{dx}\right|_{x=p} > 0$, then $f(x)$ is an increasing function at $x = p$.
- if $f'(p) = \left.\frac{df}{dx}\right|_{x=p} < 0$, then $f(x)$ is a decreasing function at $x = p$.
- if $f'(p) = \left.\frac{df}{dx}\right|_{x=p} = 0$, then $f(x)$ is at critical point at $x = p$. We do not know anything new about the behavior of $f(x)$ at $x = p$.

If $f(x)$ is differentiable at point $p$ and $f'(p) = 0 = \tan\theta$ (i.e., the tangent is parallel to $X$ axis, Figure 4), then we call $p$ a critical point or stationary point of $f(x)$. A point $p$ at which the derivative of $f(x)$ is not defined is called a singular point of $f(x)$.

The second derivative of a function is the derivative of the derivative of that function. We write it as $f''(x)$ or $\frac{d^2f}{dx^2}$. While the first derivative can tell us if the function is increasing or decreasing, the second derivative tells us

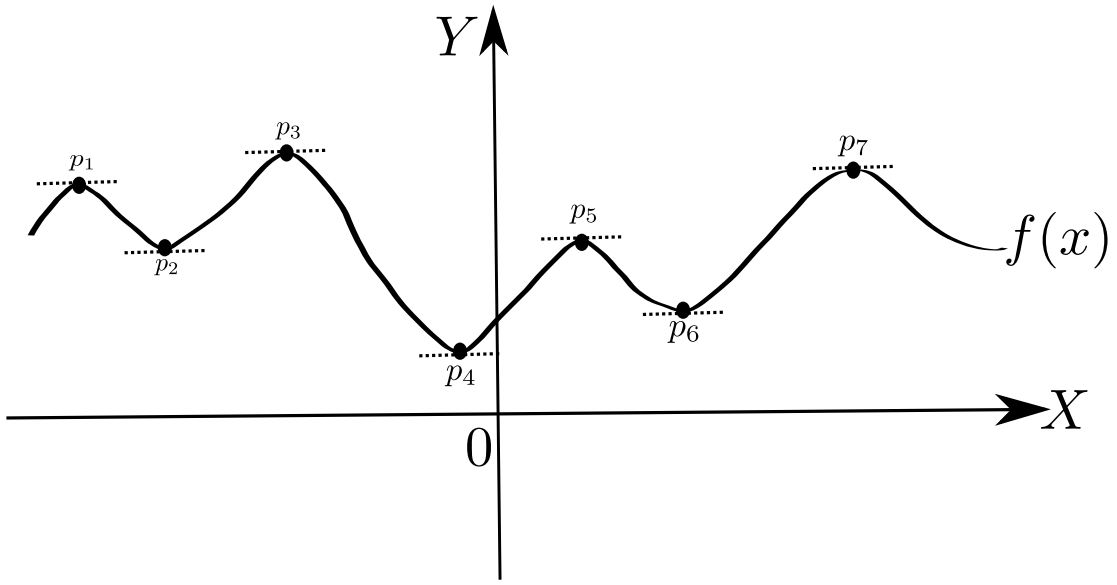



Fig. 4. The plot of a function $f(x)$. The first derivatives of the function at points $p_1, p_2, p_3, p_4, p_5, p_6, p_7$ are zero; we see that the tangent lines on these points have slope 0, i.e., are parallel to the $X$ axis, $\tan\theta = \left.\frac{df}{dx}\right|_{x=p_i} = 0$, for $i = 1 \cdots 7$. So, these 7 points are called the critical points of the function. We can't say anything further than this phenomenon whether the critical points are either local minimum, or local maximum by looking at the critical points alone. That's why we need the second derivative test.

if the first derivative is increasing or decreasing, i.e., the slope of the tangent line to the function is increasing or decreasing. If the second derivative is positive, then the first derivative is increasing, so that the slope of the tangent line to the function is increasing as $x$ increases. We see this phenomenon graphically as the curve of the graph being concave up. Likewise, if the second derivative is negative, then the first derivative is decreasing, so that the slope of the tangent line to the function is decreasing as $x$ increases. Graphically, we see this as the curve of the graph being concave down. At the points where the second derivative is zero, we do not learn anything about the shape of the graph: it may be concave up or concave down, or it may be changing from concave up to concave down or changing from concave down to concave up. So, let's summarize these:

- if $\left.\frac{d^2f}{dx^2}\right|_{x=p} > 0$ at $x = p$, then the function $f(x)$ is concave up at $x = p$.
- if $\left.\frac{d^2f}{dx^2}\right|_{x=p} < 0$ at $x = p$, then $f(x)$ is concave down at $x = p$.
- if $\left.\frac{d^2f}{dx^2}\right|_{x=p} = 0$ at $x = p$, then we do not know anything new about the behavior of $f(x)$ at $x = p$.

**Critical Points and the second derivative test:** We learned that, when $x$ is a critical point of the function $f(x)$, we do not learn anything new about the function at that point: it could be increasing, decreasing, a local minimum or a local maximum. We can often use the second derivative of the function, to find out when $x$ is a local maximum or a local minimum.

**Case 1:** Recall that $x$ is a critical point of a function when the slope of the function is zero at that point. Now, suppose that $x$ is a critical point and the second derivative of the function at that point is positive. The positive second derivative at $x$ tells us that the derivative of $f(x)$ is increasing at that point and, graphically, that the curve of the graph is concave up at that point. The only way to sketch the graph of a function at a point where the slope of the function is zero but the graph is concave up is to make that point a local minimum of the function. So, if $x$ is a critical point of $f(x)$ and the second derivative of $f(x)$ is positive, then $x$ is a local minimum of $f(x)$.

**Case 2:** If $x$ is a critical point of $f(x)$, and the second deriviative of $f(x)$ is negative, then the slope of the graph of the function is zero at that point, but the curve of the graph is concave down. The only way to draw a graph like this is to make the point $x$ a local maximum of the function. Hence, we get that if $x$ is a critical point of $f(x)$ and the second derivative of $f(x)$ is negative, then $x$ is a local maximum of $f(x)$.

**Case 3:** If $x$ is a critical point of $f(x)$ and the second derivative is zero, then we learn no new information about

that point. The point $x$ may be a local maximum or a local minimum, and the function may also be increasing or decreasing at that point.

Thus, in a summary:

- If $\left.\frac{df}{dx}\right|_{x=p} = 0$ and $\left.\frac{d^2f}{dx^2}\right|_{x=p} > 0$, then $f(x)$ has a local minimum at $x = p$.
- If $\left.\frac{df}{dx}\right|_{x=p} = 0$ and $\left.\frac{d^2f}{dx^2}\right|_{x=p} < 0$, then $f(x)$ has a local maximum at $x = p$.
- If $\left.\frac{df}{dx}\right|_{x=p} = 0$ and $\left.\frac{d^2f}{dx^2}\right|_{x=p} = 0$, then we learn no new information about the behavior of $f(x)$ at $x = p$.

Example 2: Given $f(x) = x^3 - 9x^2 + 15x - 7$, find critical points of $f(x)$ and check if any of the points are local maximum or local minimum.

Ans: The derivative of $f(x)$ is $f'(x) = 3x^2 - 18x + 15$. We need to solve for $x$ if $f'(x) = 0$. That is,

$$\begin{aligned} f'(x) &= 0 \\ 3x^2 - 18x + 15 &= 0 \\ x^2 - 6x + 5 &= 0 \\ x^2 - x - 5x + 5 &= 0 \\ x(x-1) - 5(x-1) &= 0 \\ (x-1)(x-5) &= 0 \\ x = 1 \text{ or } 5 & \end{aligned}$$

So, $f(x)$ has critical points at $x = 1$ and $x = 5$. We now apply the second derivative test: $f''(x) = 6x - 18$. Then we put the critical points on the $f''(x)$ for the local minimum/maximum test:

At $x = 1$, $f''(1) = 6.1 - 18 = -12 < 0$, so the critical point $x = 1$ is the local maximum of the function $f(x)$.

At $x = 5$, $f''(5) = 6.5 - 18 = 12 > 0$, thus the critical point $x = 5$ is the local minimum of the function $f(x)$.

## 24 Partial Derivative

A partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant [11] [12].

Let's assume $z = f(x, y)$, represents a surface in 3D Euclidean space (Figure 5).

```
#Let's plot the function z = f(x,y) = x^2 + xy + y^2
x <- seq(-10,10,length=30) #generate equally spaced 30 points
                 #on the X axis (between -10 and 10)
y <- x #use the same 30 points for Y axis

f <- function(x,y){ r <- x^2+x*y+y^2} #define the function here
z <- outer(x,y,f) #outer product of the x vector and y vector.
                 #Instead of product, the outer function will apply
                 #the f() function to get the value for z's

#now we can plot the surface
#op <- par(bg = 'white') #white background
#theta & phi control the viewing angle
        #theta moves the viewing angle left and right
        #phi moves it up and down
par(mfrow=c(2,2))
persp(x,y,z,theta=40,phi=-20,col="lightblue")
persp(x,y,z,theta=-40,phi=20,col="lightblue")
persp(x,y,z,theta=-30,phi=-30,col="lightblue")
persp(x,y,z,theta=-30,phi=30,col="lightblue")
```

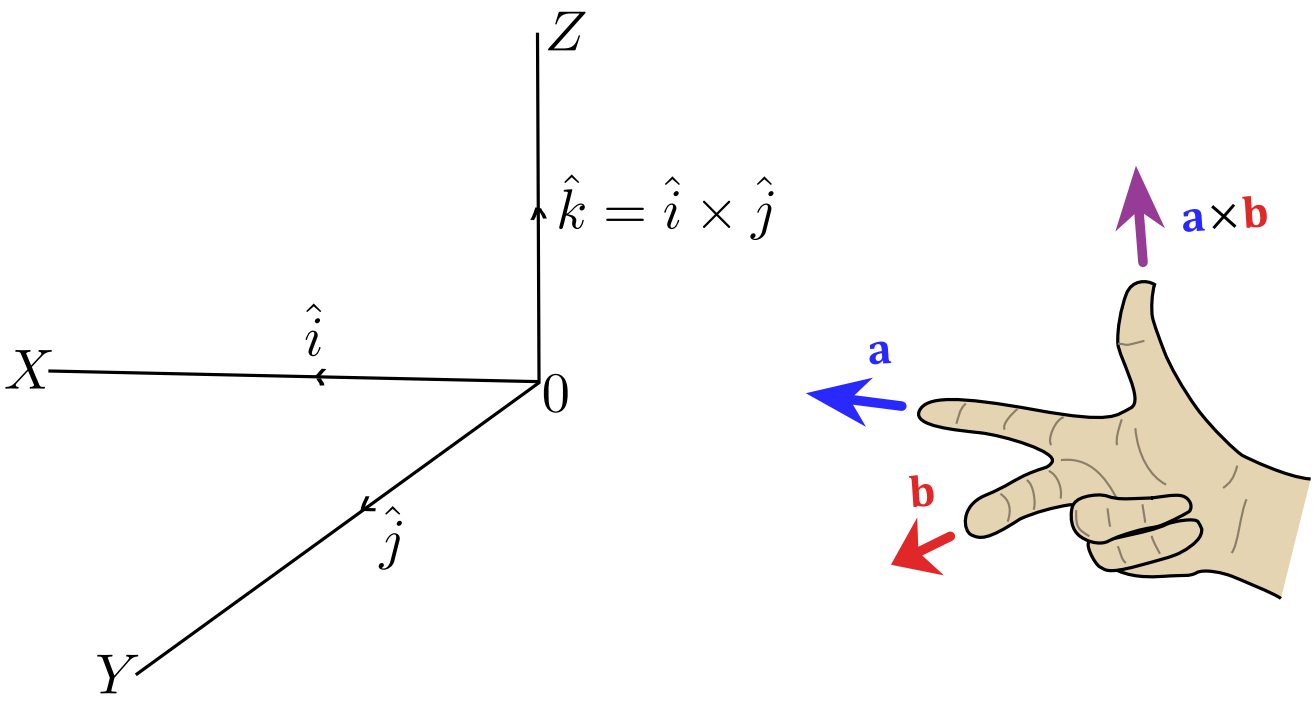
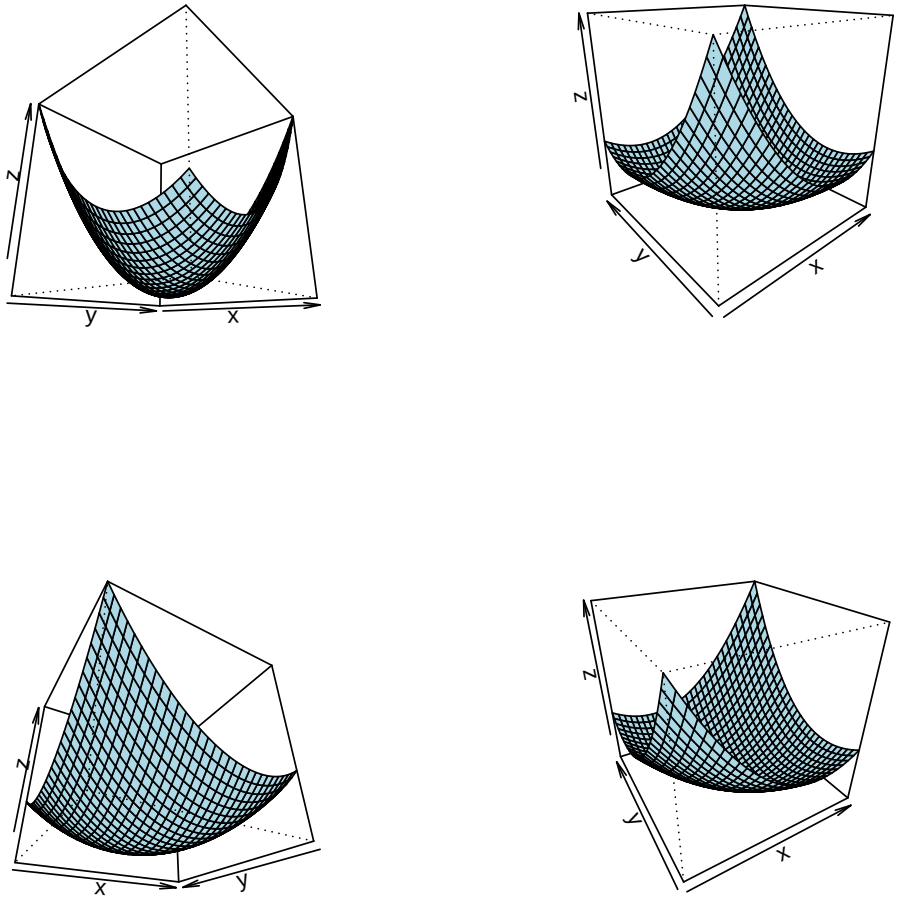



Fig. 5. Visualizing the 3D Euclidean space using the right hand rule. The unit vector acros the $Z$ axis, $\hat{k}$ is equal to the cross product of the unit vectors across the $X$ and $Y$ axes, $\hat{i}$ and $\hat{j}$ respectively. That is, $\hat{k} = \hat{i} \times \hat{j}$



With the function $z = x^2 + xy + y^2$, the first derivative with respect to $x$ by holding $y$ constant, we get:

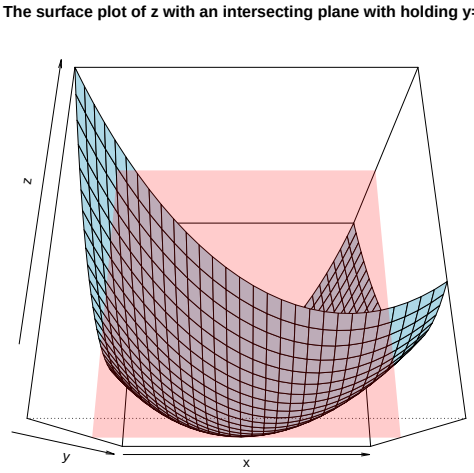
$$\frac{\partial z}{\partial x} = 2x + y$$

Then, we have an intersecting plane for the 3D plot of the $z$ function, for instance $y = 1$. Then at the intersection of the plane and the surface plot, we have a curve, and the slope to that curve is the partial derivative of $z$ with respect to $x$ by holding $y$ constant.

```
#Let's plot the function z = f(x,y) = x^2 + xy + y^2
x <- seq(-10,10,length=30) #generate equally spaced 30 points
                #on the X axis (between -10 and 10)
y <- x #use the same 30 points for Y axis

f <- function(x,y){
        r <- x^2+x*y+y^2
} #define the function here
z <- outer(x,y,f) #outer product of the x vector and y vector.
                #Instead of product, the outer function will apply
                #the f() function to get the value for z's

res <- persp(x,y,z,phi=-20,col="lightblue",
        main="The surface plot of z with an intersecting plane with holding y=1")
polygon(trans3d(c(-10,10,10,-10,-10), c(1,1,1,1,1),c(0,0,300,300,0),res),
                        col=rgb(1,0,0,0.2),border=NA)
```




Similarly, the first derivative of $z$ with respect to $y$ would be:

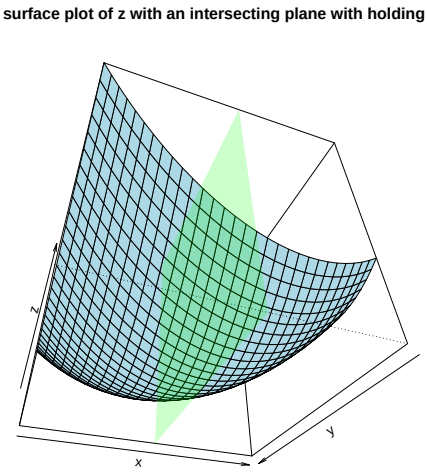
$$\frac{\partial z}{\partial y} = x + 2y$$

```
#Let's plot the function z = f(x,y) = x^2 + xy + y^2
x <- seq(-10,10,length=30) #generate equally spaced 30 points
                #on the X axis (between -10 and 10)
y <- x #use the same 30 points for Y axis

f <- function(x,y){
```

```
        r <- x^2+x*y+y^2
} #define the function here
z <- outer(x,y,f) #outer product of the x vector and y vector.
                  #Instead of product, the outer function will apply
                  #the f() function to get the value for z's

res <- persp(x,y,z,theta=-20,phi=-40,col="lightblue",
        main="The surface plot of z with an intersecting plane with holding x=1")
polygon(trans3d(c(1,1,1,1,1), c(-10,10,10,-10,-10),c(0,0,300,300,0),res),
                        col=rgb(0,1,0,0.2),border=NA)
```




## 25 Directional Derivative

Let the function $f(x, y)$ be the height of a mountain range at each point $\vec{x} = (x, y)$ [13] (Figure 6). If you stand at some point $\vec{x} = \vec{a}$, the slope of the ground in front of you will depend on the direction you are facing. It might slope steeply up in one direction, be relatively flat in another direction, and slope steeply down in yet another direction.

The partial derivatives of $f$ will give the slope $\frac{\partial f}{\partial x}$ in the positive $x$ direction and the slope $\frac{\partial f}{\partial y}$ in the positive $y$ direction. We can generalize the partial derivatives to calculate the slope in any direction. The result is called the directional derivative.

The first step in taking a directional derivative, is to specify the direction. One way to do this is with a vector $\vec{u} = (u_1, u_2)$ that points in the direction in which we want to compute the slope. For simplicity, we will insist that $\vec{u}$ is a unit vector. We write the directional derivative $f$ in the direction $\vec{u}$ at the point $\vec{a}$ as $D_{\vec{u}}f(\vec{a})$. We could define it with a limit definition just as the ordinary derivative:

$$D_{\vec{u}}f(\vec{a}) = \lim_{h \to 0} \frac{f(\vec{a} + h\vec{u}) - f(\vec{a})}{h} \tag{11}$$

$D_{\vec{u}}f(\vec{a})$ is the slope of $f(\vec{x})$ when standing at the point $\vec{a}$ and facing the direction given by $\vec{u}$. $D_{\vec{u}}f(\vec{a})$ is a number. In fact, the directional derivative is the same as a partial derive if $\vec{u}$ points in the positive $x$ direction (or positive $y$ direction). For example, if $\vec{u} = (1, 0)$, then $D_{\vec{u}}f(\vec{a}) = \frac{\partial f}{\partial x}(\vec{a})$. Similarly if $\vec{u} = (0, 1)$, then $D_{\vec{u}}f(\vec{a}) = \frac{\partial f}{\partial y}(\vec{a})$

In most cases, there is always one direction $\vec{u}$ where the directional derivative $D_{\vec{u}}f(\vec{a})$ is the largest. This is the "uphill" direction. Let's call it the direction $\vec{m}$, and the maximal directional derivative $D_{\vec{m}}f(\vec{a})$ are captured by something called the gradient of $f$, and denoted by $\nabla f(a)$. The gradient is a vector that points in the direction of $\vec{m}$ and whose magnitude is $D_{\vec{m}}f(\vec{a})$. In other word: $\frac{\nabla f(\vec{a})}{||\nabla f(\vec{a})||} = \vec{m}$ and $||\nabla f(\vec{a})|| = D_{\vec{m}}f(\vec{a})$

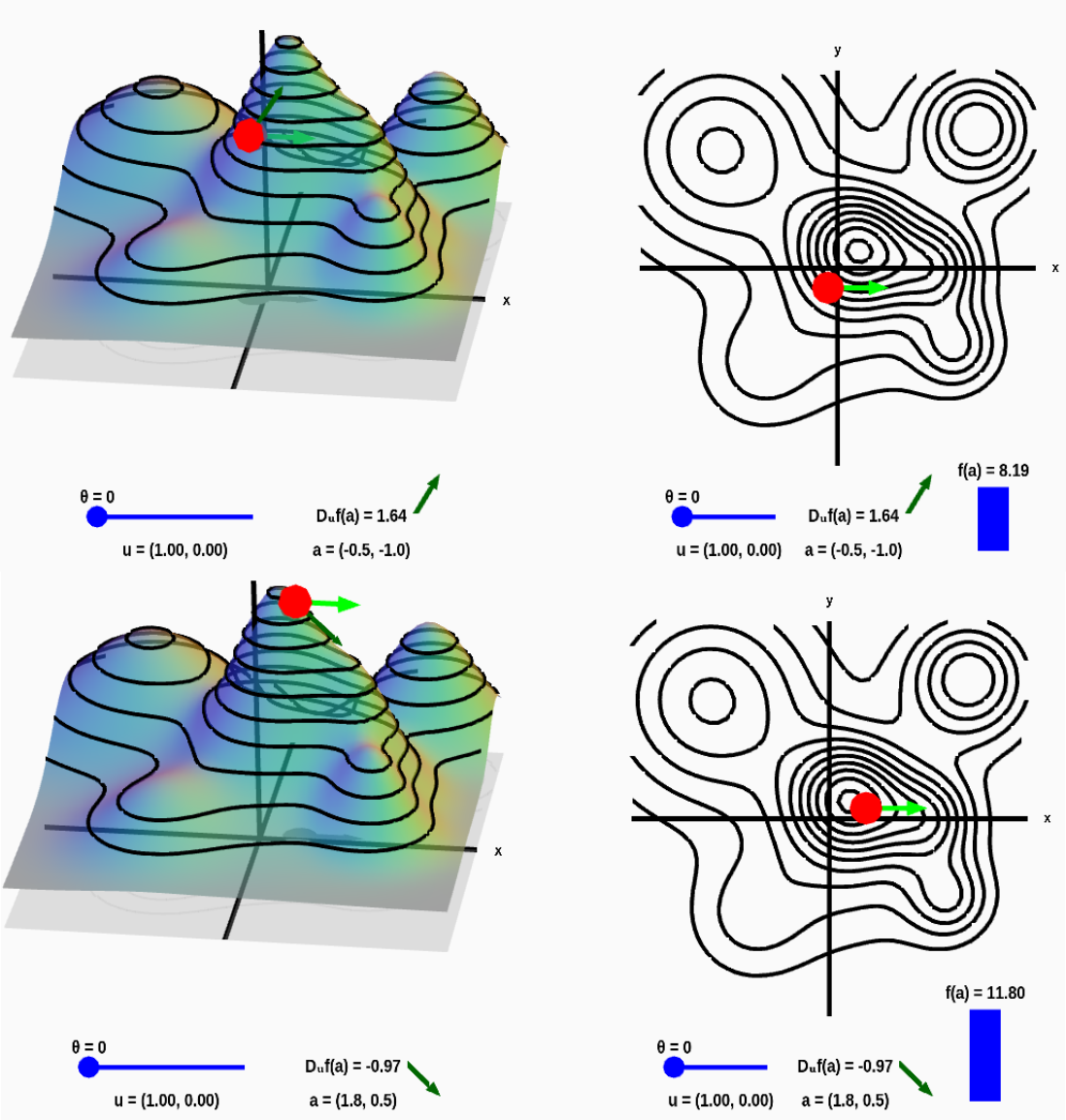



Fig. 6. Directional derivative on a mountain [13]. The height of a mountain range described by a function $f(x, y)$ is shown as surface plot in 3D (left), and a 2-D level curve plot (contour) (right). In each panel, a red point is the viewer, or the point where the directional derivative is to be evaluated. The directional derivative is computed in the direction of the 2D vector $\vec{u}$. The direction is illustrated by the light green vectors as well shown in the lower left. The direction of $\vec{u}$ is determined by the angle $\theta$ it makes with straight east (positive $x$ direction). The 2D point $\vec{a}$ where the directional derivative is computed is illustrated by the shadow of the red point on the $xy$-plane below the surface plot and by the red point itself on the level curve plot. The value of the directional derivative $D_{\vec{u}}f(\vec{a})$ is shown at the bottom of the panel, along with the value of $\vec{a}$ itself. The value of $D_{\vec{u}}f(\vec{a})$ is thes lope of the dark green vector to its right. This dark green vector is also shown emanating from the red point on the surface plot, where it is tangent to the surface, indicating that this slope is indeed the slope of the surface in the direction given by $\vec{u}$. The height of the surface $f(\vec{a})$ is illustrated by the blue bar in the lower right.

Consider $\theta$ is the angle between the gradient $\nabla f(\vec{a})$ and the direction vector $\vec{u}$. When $\theta = 0$ (or $\theta = 2\pi$), then $\vec{u}$ points in the same direction as the gradient. Also, then the directional derivative $D_{\vec{u}}f(\vec{a})$ and the maginitude of the gradient $||\nabla f(\vec{a})||$ are identical, i.e., $D_{\vec{u}}f(\vec{a}) = ||\nabla f(\vec{a})||$

But, when $\theta = \pi$, then $\vec{u}$ points in the opposite direction of the gradient, and $D_{\vec{u}}f(\vec{a}) = -||\nabla f(\vec{a})||$

Thus, for a fixed $\vec{a}$, the maximal value of $D_{\vec{u}}f(\vec{a})$ occurs when $\vec{u}$ and $\nabla f(a)$ point in the same direction (i.e., when $\theta = 0$ or $\theta = 2\pi$), and the minimum value occurs when $\vec{u}$ and $\nabla f(\vec{a})$ point in opposite directions (i.e., when $\theta = \pi$). Hence, $D_{\vec{u}}f(\vec{a})$ always lies between $-||\nabla f(\vec{a})||$ and $||\nabla f(\vec{a})||$.

It turns out that the relationship between the gradient and the directional derivative can be summarized by the

equation:

$$\begin{aligned} D_{\vec{u}}f(\vec{a}) &= \nabla f(\vec{a}) \cdot \vec{u} \\ &= ||\nabla f(\vec{a})||||\vec{u}|| \cos\theta \\ &= ||\nabla f(\vec{a})|| \cos\theta \end{aligned} \tag{12}$$

where $\theta$ is the angle between $\vec{u}$ and the gradient $\nabla f(\vec{a})$, and $\vec{u}$ is the unit vector.

The gradient always points in the direction where the mountain rises most steeply.

In a summary, given a multivariate scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$,

- The directional derivative $D_{\vec{u}}f$ is a generalization of the partial derivative to the slope of $f$ in a direction of an arbitrary unit vector $\vec{u}$.
- The gradient $\nabla f$ is a vector that points in the direction of the greatest upward slope whose length is the directional derivative in that direction,
- The directional derivative is the dot product between the gradient and the unit vector: $D_{\vec{u}}f = \nabla f \cdot \vec{u}$

It is simple to calculate an expression for the gradient of a function, if we can remember what it means for a function to be differentiable. What does it mean for a function $f(\vec{x})$ to be differentiable at the point $\vec{x} = \vec{a}$? The function must be locally be essentially linear, i.e., there must be a linear approximation

$$L(\vec{x}) = f(\vec{a}) + Df(\vec{a})(\vec{x} - \vec{a})$$

that is very close to $f(\vec{x})$ for all $\vec{x}$ near $\vec{a}$. The definition of differentiability means that, for all directions emanating out of $\vec{a}$, $f(\vec{x})$ and $L(\vec{x})$ have the same slope. We can therefore calculate the directional derivatives of $f$ at $\vec{x}$ using $L$ rather than $f$.

Using the definition of directional derivative, we can calculate the directional derivative of $f$ at $\vec{a}$ in the direction of $\vec{u}$:

$$\begin{aligned} D_{\vec{u}}f(\vec{a}) &= D_{\vec{u}}L(\vec{a}) \\ &= \lim_{h\to 0} \frac{L(\vec{a}+h\vec{u}) - L(\vec{a})}{h} \\ &= \lim_{h\to 0} \frac{\cancel{f(\vec{a})} + Df(\vec{a})(\cancel{\vec{a}} + h\vec{u} - \cancel{\vec{a}}) - \cancel{f(\vec{a})} - Df(\vec{a})(\vec{a} - \vec{a})}{h} \\ &= \lim_{h\to 0} \frac{Df(\vec{a})\cancel{h}\vec{u}}{\cancel{h}} \\ &= \lim_{h\to 0} Df(\vec{a})\vec{u} \\ &= Df(\vec{a})\vec{u} \qquad (\text{setting } h = 0) \end{aligned}$$

Since $Df(\vec{x})$ is a $1 \times n$ row vector and $\vec{u}$ is an $n \times 1$ column vector, the result is a scalar. We could rewrite this product as a dot-product between two vectors, by reforming the $1 \times n$ matrix of partial derivatives into a vector. We denote the vector by $\nabla f$ and we call it the gradient. We obtain that the directional derivative is

$$D_{\vec{u}}f(\vec{a}) = \nabla f(\vec{a}) \cdot \vec{u}$$

Now, let's do some practice examples: [14].

**Example 1**: Let $f(x, y) = x^2 y$. (a) Find $\nabla f(3, 2)$. (b) Find the derivative of $f$ in the direction of (1,2) at the point (3,2).

**Solution:** (a) The gradient is the vector of the partial derivatives. That is:

$$\nabla f(x, y) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 2xy \\ x^2 \end{pmatrix}$$

So,

$$\nabla f(3, 2) = \begin{pmatrix} 2.3.2 \\ 3^2 \end{pmatrix} = \begin{pmatrix} 12 \\ 9 \end{pmatrix} = 12\hat{i} + 9\hat{j}$$

(b) We need to find out the directional derivative $D_{\vec{u}}f(\vec{a})$, where the unit vector in the direction of (1,2) is $\vec{u} = \frac{(1,2)}{\sqrt{1^2+2^2}} = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}) = \frac{1}{\sqrt{5}}\hat{i} + \frac{2}{\sqrt{5}}\hat{j}$ and $\vec{a} = (3, 2)$.

So, using the relationship among directional derivative and the gradient, we can obtain the result:

$$\begin{aligned} D_{\vec{u}}f(\vec{a}) &= \nabla f(\vec{a}) \cdot \vec{u} \\ &= (12\hat{i} + 9\hat{j}) \cdot (\frac{1}{\sqrt{5}}\hat{i} + \frac{2}{\sqrt{5}}\hat{j}) \\ &= \frac{12}{\sqrt{5}} + \frac{18}{\sqrt{5}} = \frac{30}{\sqrt{5}} \end{aligned}$$

**Example 2**: For the $f$ in example 1, find the directional derivative of $f$ at the point (3,2) in the direction (2,1).

**Solution**: So, here the unit vector $\vec{u}$ gets changed to $\frac{(2,1)}{\sqrt{2^1+1^2}} = (\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}) = \frac{2}{\sqrt{5}}\hat{i} + \frac{1}{\sqrt{5}}\hat{j}$, and the point $\vec{a} = (3, 2)$ same as in example 1. We already calculated the gradient at that point which is $\nabla f(3, 2) = 12\hat{i} + 9\hat{j}$

So, we can compute the directional derivative by

$$\begin{aligned} D_{\vec{u}}f(\vec{a}) &= \nabla f(\vec{a}) \cdot \vec{u} \\ &= (12\hat{i} + 9\hat{j}) \cdot (\frac{2}{\sqrt{5}}\hat{i} + \frac{1}{\sqrt{5}}\hat{j}) \\ &= \frac{24}{\sqrt{5}} + \frac{9}{\sqrt{5}} = \frac{33}{\sqrt{5}} \end{aligned}$$

**Example 3**: For the $f$ in example 1 at the point (3,2), (a) in which direction is the directional derivative maximal? (b) what is the value of the directional derivative in that direction, i.e., what is the maximal directional derivative of $f$ at point (3,2)?

**Solution**: (a) We know the gradient points are always in the direction of the maximal directional derivative. So since the gradient of $f$ at point (3,2) is $\nabla f(3, 2) = 12\hat{i}+9\hat{j} = (12, 9)$, we will have the maximal directional derivative in this direction. So the answer is the unit vector in that direction, i.e., $\vec{u} = \frac{(12,9)}{\sqrt{12^2+9^2}} = (\frac{12}{15}, \frac{9}{15}) = (\frac{4}{5}, \frac{3}{5}) = \frac{4}{5}\hat{i} + \frac{3}{5}\hat{j}$

(b) The maximal directional derivative in the direction of $\vec{u}$ at point $\vec{a} = (3, 2)$ would be

$$\begin{aligned} D_{\vec{u}}f(\vec{a}) &= \nabla f(\vec{a}) \cdot \vec{u} \\ &= (12\hat{i} + 9\hat{j}) \cdot (\frac{4}{5}\hat{i} + \frac{3}{5}\hat{j}) \\ &= \frac{48}{5} + \frac{27}{5} = \frac{75}{5} = 15 = ||\nabla f(3, 2)|| = ||12\hat{i} + 9\hat{j}|| = \sqrt{12^2 + 9^2} = 15 \end{aligned}$$

### *25.1 Directional Derivatives: Vectors*

Here are the steps to compute directional derivatives of vectors [15]:

- Step 1: Pick an arbitrary unit length $\vec{u}$: $\vec{u}^T\vec{u} = 1$
- Step 2: Setup the standard directional derivative definition (as in Equation 11)
- Step 3: If we can find a $\vec{z}$, such that:

$$D_{\vec{u}}f(\vec{x}) = \lim_{h\to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h} = \vec{u}^T\vec{z}$$

Then, for an arbitrary place $i$, in an arbitrary direction $\langle\vec{u}\rangle_i$: $D_{\vec{u}_{(i)}}f(\vec{x}) = \vec{u}^T\vec{z}$ reduces to just $D_{\vec{u}_{(i)}}f(\vec{x}) = \vec{z}_i$, where $z_i$ is the partial derivative in the $i^{\text{th}}$ place. So, $\vec{z}$ would be the directional derivative of the given function.

### *25.2 Directional Derivatives: Matrices*

Here are the steps to compute directional derivatives of matrices [15]:

- Step 1: $\mathbf{U} = (i, j)$ be a Matrix such that $u_{ij} = 1$ in $\langle\mathbf{U}_{ij}\rangle$, and 0 otherwise.
- Step 2: We extend the idea of directional derivative definition (Equation 11) to matrices:

$$D_{\mathbf{U}}f(\mathbf{X}) = \lim_{h\to 0} \frac{f(\mathbf{X} + h\mathbf{U}) - f(\mathbf{X})}{h}$$

here, we can conclude that

$$D_{\mathbf{U}_{(ij)}} f(\mathbf{X}) = \lim_{h \to 0} \frac{f(\mathbf{X} + h\mathbf{U}_{(ij)}) - f(\mathbf{X})}{h}$$

will "pick off" the partial derivative in the $(i,j)^{\text{th}}$ place.

- Step 3: If we can find a $\mathbf{Z}$, such that:

$$D_{\mathbf{U}} f(\mathbf{X}) = \lim_{h \to 0} \frac{f(\mathbf{X} + h\mathbf{U}) - f(\mathbf{X})}{h} = \text{tr}(\mathbf{U}^T \mathbf{Z})$$

then, for an arbitrary place $(i,j)$, in an arbitrary direction $\langle \mathbf{U} \rangle_{ij}$

$$D_{\mathbf{U}_{(ij)}} f(\mathbf{X}) = \text{tr}(\mathbf{U}_{(ij)}^T \mathbf{Z}) = \sum_j \sum_i (u_{ij} z_{ij}) = z_{ij}$$

so, $\mathbf{Z}$ would be the directional derivative of the given function.

Example 1: Given a function $f(\mathbf{X}) = \text{tr}(AX)$, find $\dfrac{\partial f(\mathbf{X})}{\partial X} = \dfrac{\partial \text{tr}(\mathbf{AX})}{\partial X}$.

Solution: we begin here:

$$\begin{aligned}
D_{\mathbf{U}} f(\mathbf{X}) &= \lim_{h \to 0} \frac{f(\mathbf{X} + h\mathbf{U}) - f(\mathbf{X})}{h} \\
&= \lim_{h \to 0} \frac{\text{tr}(\mathbf{A}(\mathbf{X} + h\mathbf{U})) - \text{tr}(\mathbf{AX})}{h} \\
&= \lim_{h \to 0} \frac{\text{tr}(\mathbf{AX} + \mathbf{A}h\mathbf{U}) - \text{tr}(\mathbf{AX})}{h} \\
&= \lim_{h \to 0} \frac{\cancel{\text{tr}(\mathbf{AX})} + \text{tr}(\mathbf{A}h\mathbf{U}) - \cancel{\text{tr}(\mathbf{AX})}}{h} \\
&= \lim_{h \to 0} \frac{\text{tr}(h\mathbf{AU})}{h} \\
&= \lim_{h \to 0} \frac{\cancel{h}\text{tr}(\mathbf{AU})}{\cancel{h}} \\
&= \lim_{h \to 0} \text{tr}(\mathbf{AU}) \\
&= \text{tr}(\mathbf{AU}) \\
&= \text{tr}((\mathbf{AU})^T) \\
&= \text{tr}(\mathbf{U}^T \mathbf{A}^T)
\end{aligned}$$

So, $\dfrac{\partial \text{tr}(\mathbf{AX})}{\partial X} = \mathbf{A}^T$.

Example 2: find derivative of $f(\mathbf{W}) = \text{tr}(\mathbf{W}^T \mathbf{AW})$ with respect to $\mathbf{W}$.

Solution:

$$\begin{aligned}
D_{\mathbf{U}}f(\mathbf{W}) &= \lim_{h\to 0}\frac{f(\mathbf{W}+h\mathbf{U})-f(\mathbf{W})}{h}\\
&= \lim_{h\to 0}\frac{\text{tr}((\mathbf{W}+h\mathbf{U})^T\mathbf{A}(\mathbf{W}+h\mathbf{U}))-\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}{h}\\
&= \lim_{h\to 0}\frac{\text{tr}((\mathbf{W}+h\mathbf{U})^T(\mathbf{A}\mathbf{W}+h\mathbf{A}\mathbf{U}))-\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}{h}\\
&= \lim_{h\to 0}\frac{\text{tr}(\mathbf{W}^TA\mathbf{W}+h\mathbf{U}^T\mathbf{A}\mathbf{W}+h\mathbf{W}^T\mathbf{A}\mathbf{U}+h^2\mathbf{U}^T\mathbf{A}\mathbf{U})-\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}{h}\\
&= \lim_{h\to 0}\frac{\cancel{\text{tr}(\mathbf{W}^TA\mathbf{W})}+h\text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+h\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{U})+h^2\text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{U})-\cancel{\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}}{h}\\
&= \lim_{h\to 0}\frac{h\text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+h\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{U})+h^2\text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{U})}{h}\\
&= \lim_{h\to 0}\text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{U})+h\text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{U})\\
&= \text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+\text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{U})+0\\
&= \text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+\text{tr}((\mathbf{W}^T\mathbf{A}\mathbf{U})^T)\\
&= \text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+\text{tr}(\mathbf{U}^T\mathbf{A}^T\mathbf{W})\\
&= \text{tr}(\mathbf{U}^T\mathbf{A}\mathbf{W})+\mathbf{U}^T\mathbf{A}^T\mathbf{W})\\
&= \text{tr}(\mathbf{U}^T(\mathbf{A}\mathbf{W}+\mathbf{A}^T\mathbf{W}))
\end{aligned}$$

So, $\dfrac{\partial \text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}{\partial \mathbf{W}} = \mathbf{A}\mathbf{W}+\mathbf{A}^T\mathbf{W} = (\mathbf{A}+\mathbf{A}^T)\mathbf{W}$

## 26 Gradient as a derivative

A Gradient is the derivative of a scalar with respect to a vector. For vector $\vec{x} = (x_1, x_2, \cdots, x_n)^T$.

$$\frac{\partial f(\vec{x})}{\partial \vec{x}} = \begin{pmatrix} \left[\dfrac{\partial f(\vec{x})}{\partial x_1}\right] \\ \left[\dfrac{\partial f(\vec{x})}{\partial x_2}\right] \\ \vdots \\ \left[\dfrac{\partial f(\vec{x})}{\partial x_n}\right] \end{pmatrix}$$

For Example, if we have a function: $f(\vec{x}) = x_1x_3^2 + x_1x_2^2 + x_2x_3^2 + x_1^2 + x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3$, with $\vec{x} = (x_1, x_2, x_3)^T$. Then the Gradient is:

$$\begin{aligned}
\frac{\partial f(\vec{x})}{\partial \vec{x}} &= \begin{pmatrix} \left[\dfrac{\partial f(\vec{x})}{\partial x_1}\right] \\ \left[\dfrac{\partial f(\vec{x})}{\partial x_2}\right] \\ \left[\dfrac{\partial f(\vec{x})}{\partial x_3}\right] \end{pmatrix}\\
&= \begin{pmatrix} x_3^2 + x_2^2 + 2x_1 + 2x_2 + 2x_3 \\ 2x_1x_2 + x_3^2 + 2x_2 + 2x_1 + 2x_3 \\ 2x_1x_3 + 2x_2x_3 + 2x_3 + 2x_1 + 2x_2 \end{pmatrix}
\end{aligned}$$

## 27 JACOBIAN AS A DERIVATIVE

A Jacobian is a derivative of a vector with respect to a transposed vector. For a vector $\vec{x} = (x_1, x_2, \cdots, x_n)^T$, and a $k$ dimensional function $\vec{f} = (f_1, f_2, \cdots, f_k)^T$, we have

$$\frac{\partial \overrightarrow{f(\vec{x})}}{\partial \vec{x}^T} = \begin{pmatrix} \left[\frac{\partial f_1(\vec{x})}{\partial x_1}\right] & \left[\frac{\partial f_1(\vec{x})}{\partial x_2}\right] & \cdots & \left[\frac{\partial f_1(\vec{x})}{\partial x_n}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\frac{\partial f_k(\vec{x})}{\partial x_1}\right] & \left[\frac{\partial f_k(\vec{x})}{\partial x_2}\right] & \cdots & \left[\frac{\partial f_k(\vec{x})}{\partial x_n}\right] \end{pmatrix}$$

For instance, if we have a function

$$\overrightarrow{f(\vec{x})} = \begin{pmatrix} 3x_1^2 + x_2 \\ ln(x_1) + x_3 \\ sin(x_2) + cos(x_4) \end{pmatrix}, \qquad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

Then the Jacobian is

$$\frac{\partial \overrightarrow{f(\vec{x})}}{\partial \vec{x}^T} = \begin{pmatrix} \left[\frac{\partial f_1(\vec{x})}{\partial x_1}\right] & \left[\frac{\partial f_1(\vec{x})}{\partial x_2}\right] & \left[\frac{\partial f_1(\vec{x})}{\partial x_3}\right] & \left[\frac{\partial f_1(\vec{x})}{\partial x_4}\right] \\ \left[\frac{\partial f_2(\vec{x})}{\partial x_1}\right] & \left[\frac{\partial f_2(\vec{x})}{\partial x_2}\right] & \left[\frac{\partial f_2(\vec{x})}{\partial x_3}\right] & \left[\frac{\partial f_2(\vec{x})}{\partial x_4}\right] \\ \left[\frac{\partial f_3(\vec{x})}{\partial x_1}\right] & \left[\frac{\partial f_3(\vec{x})}{\partial x_2}\right] & \left[\frac{\partial f_3(\vec{x})}{\partial x_3}\right] & \left[\frac{\partial f_3(\vec{x})}{\partial x_4}\right] \end{pmatrix} = \begin{pmatrix} 6x_1 & 1 & 0 & 0 \\ \frac{1}{x_1} & 0 & 1 & 0 \\ 0 & cos(x_2) & 0 & -sin(x_4) \end{pmatrix}$$

## 28 HESSIAN AS A DERIVATIVE

The Hessian is the derivative of a Gradient with respect to a transposed vector. For a given vector $\vec{x} = (x_1, x_2, \cdots, x_n)^T$, we have the gradient

$$\frac{\partial f(\vec{x})}{\partial \vec{x}} = \begin{pmatrix} \left[\frac{\partial f(\vec{x})}{\partial x_1}\right] \\ \left[\frac{\partial f(\vec{x})}{\partial x_2}\right] \\ \vdots \\ \left[\frac{\partial f(\vec{x})}{\partial x_n}\right] \end{pmatrix}$$

Then the Hessian would the derivative of this gradient with respect to $\vec{x}^T$.

$$\frac{\partial^2 f(\vec{x})}{\partial \vec{x} \partial \vec{x}^T} = \begin{pmatrix} \left[\frac{\partial f(\vec{x})}{\partial x_1^2}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_1 \partial x_2}\right] & \cdots & \left[\frac{\partial f(\vec{x})}{\partial x_1 \partial x_n}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\frac{\partial f(\vec{x})}{\partial x_n \partial x_1}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_n \partial x_2}\right] & \cdots & \left[\frac{\partial f(\vec{x})}{\partial x_n^2}\right] \end{pmatrix}$$

For instance, for function $f(\vec{x}) = x_1x_3^2 + x_1x_2^2 + x_2x_3^2 + x_1^2 + x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3$, with $\vec{x} = (x_1, x_2, x_3)^T$, the Gradient we computed previously wass

$$\frac{\partial f(\vec{x})}{\partial \vec{x}} = \begin{pmatrix} x_3^2 + x_2^2 + 2x_1 + 2x_2 + 2x_3 \\ 2x_1x_2 + x_3^2 + 2x_2 + 2x_1 + 2x_3 \\ 2x_1x_3 + 2x_2x_3 + 2x_3 + 2x_1 + 2x_2 \end{pmatrix}$$

Hence, the Hessian would be

$$\frac{\partial^2 f(\vec{x})}{\partial \vec{x} \partial \vec{x}^T} = \begin{pmatrix} \left[\frac{\partial f(\vec{x})}{\partial x_1^2}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_1 \partial x_2}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_1 \partial x_3}\right] \\ \left[\frac{\partial f(\vec{x})}{\partial x_2 \partial x_1}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_2^2}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_2 \partial x_3}\right] \\ \left[\frac{\partial f(\vec{x})}{\partial x_3 \partial x_1}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_3 \partial x_2}\right] & \left[\frac{\partial f(\vec{x})}{\partial x_3^2}\right] \end{pmatrix} = \begin{pmatrix} 2 & 2x_2+2 & 2x_3+2 \\ 2x_2+2 & 2x_1+2 & 2x_3+2 \\ 2x_3+2 & 2x_3+2 & 2x_1+2x_2+2 \end{pmatrix}$$

## 29 Hadamard Product

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimensions $n \times n$, the Hadamard product $\mathbf{A} \circ \mathbf{B}$ (a.k.a elementwise product, entrywise product, Schur product) is a matrix of the same dimension, the $i, j$ element of $\mathbf{A}$ is multiplied with $i, j$ element of $\mathbf{B}$, that is:

$$(\mathbf{A} \circ \mathbf{B})_{ij} = A_{ij} B_{ij}$$

And can be displayed fully by:

$$\mathbf{A} \circ \mathbf{B} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} \circ \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nm} \end{pmatrix} = \begin{pmatrix} A_{11}B_{11} & A_{12}B_{12} & \cdots & A_{1m}B_{1m} \\ A_{21}B_{21} & A_{22}B_{22} & \cdots & A_{2m}B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}B_{n1} & A_{n2}B_{n2} & \cdots & A_{nm}B_{nm} \end{pmatrix}$$

## 30 Kronecker Product

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of any different dimensions $m \times n$ and $p \times q$ respectively, the Kronecker product is the matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} A_{11}\mathbf{B} & A_{12}\mathbf{B} & \cdots & A_{1n}\mathbf{B} \\ A_{21}\mathbf{B} & A_{22}\mathbf{B} & \cdots & A_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}\mathbf{B} & A_{m2}\mathbf{B} & \cdots & A_{mn}\mathbf{B} \end{pmatrix}$$

with dimensions $mp \times nq$. For instance, Suppose $\mathbf{X} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} s & t & u \\ v & w & z \end{pmatrix}$, then

$$\mathbf{X} \otimes \mathbf{Y} = \begin{pmatrix} a\mathbf{Y} & b\mathbf{Y} \\ c\mathbf{Y} & d\mathbf{Y} \end{pmatrix} = \begin{pmatrix} as & at & au & bs & bt & bu \\ av & aw & az & bv & bw & bz \\ cs & ct & cu & ds & dt & du \\ cv & cw & cz & dv & dw & dz \end{pmatrix}$$

## 31 Trace of a Matrix

The trace of an $n \times n$ square matrix $A$ is defined to be the sum of elements on the main diagonal (i.e., the diagonal from the upper left to the lower right corner):

$$tr(A) = \sum_{i=1}^{n} A_{i,i}$$

Here are some properties of trace:

- It is a linear operator. That is for two square matrices $\mathbf{A}$ and $\mathbf{B}$:
  1) $tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$
  2) $tr(c\mathbf{A}) = c \cdot tr(\mathbf{A})$ for any scalar $c$
  3) $\partial tr(\mathbf{X}) = tr(\partial X)$

- Transposition of dependent variable: $tr(\mathbf{Y}) = tr(\mathbf{Y}^T)$, and thus $\dfrac{\partial tr(\mathbf{Y})}{\partial \mathbf{X}} = \dfrac{\partial tr(\mathbf{Y}^T)}{\partial \mathbf{X}}$
- Cyclic permuation: $tr(\mathbf{ABCD}) = tr(\mathbf{BCDA}) = tr(\mathbf{CDAB}) = tr(\mathbf{DABC})$
- If $\mathbf{U}, \mathbf{H}$ are two squared matrices of same dimension, and you want to multiply all the paired indices $(u_{ij}h_{ij})$ and add all the multiplications, you can use $tr(\mathbf{U}^T\mathbf{H}) = \sum_{j=1}^{n}\sum_{i=1}^{n} u_{ij}h_{ij} = \sum_{ij}(\mathbf{U} \circ \mathbf{Y})_{ij}$ (using Hadamard product)
- If $\mathbf{A}$ is a square $n \times n$ matrix with real or complex entries and if $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of $\mathbf{A}$, then $tr(\mathbf{A}) = \sum_i \lambda_i$. This follows from the fact that $\mathbf{A}$ is always similar to its Jordan form, an upper-triangular matrix having $\lambda_1, \cdots, \lambda_n$ on their main diagonal, and trace operator would return their sum. [In contrast, the determinant of $\mathbf{A}$ is the product of its eigenvalues, i.e., $det(\mathbf{A}) = \prod_i \lambda_i$]. And more generally, $tr(\mathbf{A}^k) = \sum_i \lambda_i^k$

## REFERENCES


[1] A. Chiang and K. Wainwright, "Fundamental methods of mathematical economics," *McGraw-Hill, New York*, 2005. 8

[2] M. A. Akivis and V. V. Goldberg, *An introduction to linear algebra and tensors*. Courier Dover Publications, 2012. 10

[3] (2014, Nov.) Triangular matrix. [Online]. Available: http://en.wikipedia.org/wiki/Triangular_matrix 11

[4] (2014, Nov.) Gaussian elimination. [Online]. Available: http://en.wikipedia.org/wiki/Gaussian_elimination 13

[5] (2014, Nov.) Invertible matrix. [Online]. Available: http://en.wikipedia.org/wiki/Invertible_matrix 14

[6] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The art of scientific computing*. Cambridge University Press London, 1987, vol. 2. 20

[7] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 281–288. 22.4

[8] (2014, Nov.) Introduction to derivatives. [Online]. Available: http://www.mathsisfun.com/calculus/derivatives-introduction.html 23

[9] (2014, Nov.) Calculus: Building intuition for the derivative. [Online]. Available: http://betterexplained.com/articles/calculus-building-intuition-for-the-derivative/ 23

[10] (2014, Nov.) The first and second derivatives. [Online]. Available: https://math.dartmouth.edu/opencalc2/cole/lecture8.pdf 23

[11] (2014, Nov.) Partial derivative. [Online]. Available: http://en.wikipedia.org/wiki/Partial_derivative 24

[12] (2015, Jan.) Partial derivative. [Online]. Available: https://www.khanacademy.org/math/multivariable-calculus/partial_derivatives_topic/partial_derivatives/v/partial-derivatives 24

[13] (2015, Jan.) Directional derivative. [Online]. Available: http://mathinsight.org/directional_derivative_gradient_introduction 25, 6

[14] (2015, Jan.) Directional derivative and gradient examples. [Online]. Available: http://mathinsight.org/directional_derivative_gradient_examples 25

[15] (2012, May) With(out) a trace, matrix derivatives the easy way. [Online]. Available: http://www.tc.umn.edu/~nydic001/docs/unpubs/Schonemann_Trace_Derivatives_Presentation.pdf 25.1, 25.2

# INDEX