Linear Algebra

Onkar Pandit

30 November 2021

Last updated on October 27, 2022.

Contents

1 Introduction

Linear algebra primarily deals with linear systems of equations and approaches to solve them. This also leads to matrix which is presents the transformation or operation which transforms vectors. The natural consequence is the study of these matrices so as to understand more about them and manipulation of these matrices to use them effectively. The subject focuses more on solving the linear system of equations $Ax = b$ where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^{n\times 1}$, and $b \in \mathbb{R}^{m\times 1}$. x denotes the vector of variables which needs to obtained at the end of the process. Then the subject develops different approaches to solve them also studies when this system is solvable and what are the conditions. (**What could be the non-linear system of equations? Does something like that exist?**)

Further, it studies different operation on matrices so as to gain as much insights from the matrices as possible. It also studies different types of matrices which are useful in solving equations as well as in other studies. The main advantage of these matrix studies is that whenever we can convert any physical system into mathematical systems, these operations or insights can be used to retrieve as much information as possible.

The study of these matrices and different operations make this subject more powerful as a lot of real world problems can be converted into matrix forms and the methods established in this subject can be applied over them so as to solve those problems.

2 Vectors and Vector Spaces

2.1 Vectors

Vector is a special case of matrix which contains only one column and n rows. A vector can have different meanings and interpretations. It can only mean the sequence of numbers in computer science, in physics it can represent the direction and magnitude of the object and mathematically it is a point in a n-dimensional space. For example:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \tag{1}$$

A **matrix** in the cruder sense contains a number of rows and columns of numbers. This constituents two dimensions of matrix one is row and other is column, therefore it is generally denoted as $m \times n$ mathematical entity where m denotes number of rows and n denotes number of columns of the matrix. For example,

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \; B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \tag{2}$$

Tensor is a generalized matrix which can be of multiple dimensions such as $m_1 \times m_2 \times \cdots m_k$.

2.2 Vector Spaces

In broader sense, vector spaces are the spaces where normal multiplication addition operation can be performed over the elements present in those spaces. Vector spaces are also known as linear spaces.

Definition 2.1. Let V be the nonempty set of elements. Then V is called vector spaces or linear spaces if it follows ten axioms. We list down two axioms which are basic the other ones are commutative, distributive laws of addition and multiplications.

1. *Closure under addition*: For any elements $x, y \in V$ there exists $x + y$ in V.

2. *Closure under multiplication by scalars*: For any $x \in V$ there exists $ax \in V$ where a is scalar value.

The popular examples are real numbers where addition and multiplication gives rise to another real number. Also, tuple of n-numbers i.e. R^n space, is also vector space. In addition to suppose matrix space of $n \times n$ matrices, space of functions, etc. are also some examples of vector spaces. These examples and many others illustrate how the linear space concept permeates algebra, geometry, and analysis.

2.3 Vector subspace

Let S be a nonempty subset of a linear space V. Then S is a vector subspace if it also satisfies the axioms of closure under addition and multiplication by scalars. The vector subspaces are vector spaces only they are a part of bigger vector space. For example, any line passing through 0 will be a vector subspace of R^2 plane.

2.3.1 Column space

When we multiply a matrix by a vector we take the linear combination of columns of the matrix therefore the result always lies in the column space of the matrix columns. The column space is the space spanned by columns of the matrix and denoted as $C(A)$ for matrix A.

Suppose, $A = \begin{bmatrix} 3 & 4 & 2 & 7 \\ 1 & 1 & 5 & 5 \\ 1 & 3 & 2 & 11 \end{bmatrix}$, then Ax is nothing but the linear combination of columns of A:

$$Ax = \begin{bmatrix} 3 & 4 & 2 & 7 \\ 1 & 1 & 5 & 5 \\ 1 & 3 & 2 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3x_1 + 4x_2 + 2x_3 + 7x_4 \\ x_1 + x_2 + 5x_3 + 5x_4 \\ x_1 + 3x_2 + 2x_3 + 11x_4 \end{bmatrix}$$
$$= x_1 \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 4 \\ 1 \\ 3 \end{bmatrix} + x_3 \begin{bmatrix} 2 \\ 5 \\ 2 \end{bmatrix} + x_4 \begin{bmatrix} 7 \\ 5 \\ 11 \end{bmatrix}$$

Further, suppose we want to solve the equation $Ax = b$ for x then, b should be in column space of Ax, otherwise a solution will not exist. The size of vectors in the column space of any vector is decided by the number of rows in the matrix. **For a matrix of size $m \times n$, the column space contains vectors of $m \times 1$ vectors.**

2.3.2 Row space

A row space is a space spanned by rows of the matrix. It is also a column space of transpose of the matrix. For matrix A, the row space is denoted as $C(A^T)$. **For a matrix of size $m \times n$, the row space contains vectors of $n \times 1$ vectors.**

Consider matrix A from the previous section, then the row space is given as:

$$x_1 \begin{bmatrix} 3 \\ 4 \\ 2 \\ 7 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 1 \\ 5 \\ 5 \end{bmatrix} + x_3 \begin{bmatrix} 1 \\ 2 \\ 2 \\ 11 \end{bmatrix}$$

2.3.3 Null space

For any given matrix A, null space consists of all the vectors which satisfies the condition $Ax = 0$. By default, the zero vector always lies in the null space even though if there are no other vectors which produce 0 vector i.e. if columns of the matrix are linearly independent then the null space of that matrix contains only a single 0 vector. It is denoted as $N(A)$ for matrix A. The ***left null space*** of A is null space of A^T i.e. we find the solution for $A^T x = 0$, here we are dealing with the rows of the matrix A.

For an invertible matrix, the nullspace contains only $x = 0$ as multiplication of A^{-1} to $Ax = 0$ gives $x = 0$. Also, the column space is a whole space as the columns are independent and span entire space.

I guess all these four spaces can be called as vector spaces as they contain zero vector, and linear combination of them lies in the same space. Is it true?

2.4 Basis

Before looking at basis, we first define the concept of dependence and independence.

Definition 2.2. A set of elements $x_1, x_2, \cdots, x_n$ are called linearly dependant if and only if $\sum_{i=1}^{n} c_i x_i = 0$ for some values of c_i which are non-zero. In other words, some elements of the set can be represented with use of other elements. Those sets which are not linearly dependant are called independent sets. Formally, $\sum_{i=1}^{n} c_i x_i = 0$ is true only for $\forall c_i = 0$.

For example, columns of $A = \begin{bmatrix} 3 & 4 & 2 \\ 0 & 1 & 5 \\ 0 & 0 & 2 \end{bmatrix}$ are linearly independent as

$$\textbf{Solve Ac=0} \qquad c_1 \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix} + c_3 \begin{bmatrix} 2 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3}$$

If in these equations, c_1, c_2, c_3 are 0 then we say that the columns of A are linearly independent. We can see that, A is already in row-echelon form, so by back-substituting, first we get $c_3 = 0$ then, $c_2 = 0$ and $c_1 = 0$. Therefore, we say that columns of A are linearly independent.

Definition 2.3. A finite set S of elements in a linear space V is called a finite basis for V if all the elements of S are linearly independent and span entire V. The space V is called finite-dimensional if it has a finite elements in S. Otherwise infinite-dimensional space.

A basis is a minimal set of vectors required to generate entire vector space. The number of elements in the basis set is called as a **dimension** of the vector space. For a given vector space, there can be infinitely many sets of basis. Dimension of vector space is an unique number so every finite basis set has same number of elements in them.

For example, R^n has one of the basis as set of n elements containing unit vectors corresponding to each n axis. Another example, consider space of polynomials $p(t)$ which is $\geq n$ has one basis as $n+1$ elements in it i.e. $\{1, t, t^{,} \cdots, t^n\}$ can produce any polynomial of degree $\leq n$ with linear combination of these elements.

How to identify if the given set of vectors are basis or not?

- First determine whether they are independent or not. This can be done with Gaussian Elimination. If we can obtain a echelon form then we can say that the set is linearly independent.

- Now, we have to see that the set can span entire vector space. For that we can go a step further and obtain reduced row echelon form this can also ensure that we can obtain entire vector space.

- Another simple approach can be to check if we can obtain canonical basis set from these echelon form vectors or from the original vector set.
- Another way can be to take determinant of the matrix formed with the vectors. If the determinant is non-zero then it is linearly independent other dependent.

The dimension of the column space $C(A)$ equals the rank r, which also equals the dimension of the row space: The number of independent columns equals the number of independent rows. The basis for $C(A)$ is formed by the r columns of A that correspond in U (echelon form of A), to the columns containing pivots.

Change of Basis Suppose certain vector is represented with the canonical basis i.e. standard basis, then its coordinates give the magnitude of the vector in each x, y, z direction, suppose basis is of R^3. Now, suppose we have coordinates x_1, y_1, z_1 for vector v with standard basis, then the vector remains there itself even with the other basis but the coordinates change with the new basis.

For example, suppose the basis $B = [b_1, b_2, b_3]$ is a new basis and (x_1, y_1, z_1) are coordinates of vector v with standard basis. The new coordinates $[v]_B = (c_1, c_2, c_3)$ with basis B are given as

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = c_1 \begin{bmatrix} | \\ b_1 \\ | \end{bmatrix} + c_2 \begin{bmatrix} | \\ b_2 \\ | \end{bmatrix} + c_3 \begin{bmatrix} | \\ b_3 \\ | \end{bmatrix} \tag{4}$$

With the above equations we can also find out the coordinates with standard basis given the coordinates with the non-standard basis.

Change of Basis Matrix Suppose there are two bases B and A. The vector coordinates are given with basis B and suppose we want to obtain coordinates with basis A, then we have to obtain a basis changing matrix $P_{B \to A}$. This matrix is obtained as $P_{B \to A} = [[b_1]_A \;\; [b_2]_A \;\; [b_3]_A]$ where $[b_1]_A$ are basis vectors represented with basis A. Each vector $[b_i]_A$ is obtained by solving the above linear system of equations separately i.e.

$$[b_1]_A : \begin{bmatrix} | \\ b_1 \\ | \end{bmatrix} = c_1 \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} + c_2 \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} + c_3 \begin{bmatrix} | \\ a_3 \\ | \end{bmatrix} \tag{5}$$

From this we can calculate (c_1, c_2, c_3) which is nothing but the $[b_1]_A$. Similar, procedure we have to repeat to obtain $[b_2]_A$ and $[b_3]_A$.

2.5 Inner product

Let V be a vector space then $\Omega : V \times V \to R$ is a function which maps two elements to a real number. The function Ω is called an *inner product* if it satisfies following conditions:

1. Symmetry: $\Omega(x, y) = \Omega(y, x) \;\; \forall x, y \in V$
2. Positive definite: $\forall x \in V$
   $0 \;\; \Omega(x, x) \geq 0$ and $\Omega(0, 0) = 0$
3. Distributive: $\Omega(x, y + z) = \Omega(x, y) + \Omega(x, z)$
4. Associative: $\Omega(cx, y) = c\Omega(x, y)$

In general, inner product is written as $\langle x, y \rangle$. The *dot product* defined as

$$x.y = x^T y = \sum_{i=1}^{n} x_i y_i = |x||y| \cos\theta$$

is a special type of inner product where θ is angle between x and y. The vector space V with a dot product $(V, \langle \cdot, \cdot \rangle)$ is called an *Euclidean vector space.* Intuitively, it tells us something about how much two vectors point in the same direction or what is the projection of a vector on another vector and it returns a single number. Note that the dot product of two vectors is a scalar, not another vector. Because of this, the dot product is also called the scalar product.

For each inner product there is an associated *positive definite matrix A* such that we can write inner product between x, y as

$$\langle x, y \rangle = x^T A y$$

So, suppose a function is given and we want to verify whether it actually represents inner product or not, then we can verify whether the matrix A is positive definite or not. It is a necessary and sufficient condition for being the inner product.

2.6 Norms

The inner product can be used to introduce the metric concept of length: *norm.*

$$||x|| = \sqrt{x^T x}$$

This is called as L_2-norm. It calculates distance of vector from origin. The more general definition is given as

$$||x||_n = \sqrt[n]{\sum_{i=1}^{l} x_i^n}$$

if x has l elements.

Norm is a type of inner product applied only to itself, because of that, the norm of two elements is always positive except 0.
Similarly, Manhattan distance can be found by putting $l = 1$ as:

$$||v||_1 = |v_1| + |v_2| + \cdots + |v_n| \tag{6}$$

Not all the norms are obtained with inner product as Manhattan norm does not involve inner product.

2.7 Matrix norms

As with vectors, we can measure the size of a matrix by computing its norm. There are multiple ways to define the norm for a matrix, as long it satisfies the same properties defined for vectors norms: (1) absolutely homogeneous, (2) triangle inequality, (3) positive definite (see vector norms section).

2.7.1 Frobenius norm

The Frobenius norm is an element-wise norm named after the German mathematician Ferdinand Georg Frobenius. We denote this norm as ‖A‖F. You can thing about this norm as flattening out the matrix into a long vector. For instance, a 3×3 matrix would become a vector with n=9 entries. For matrix $A \in R^{m \times n}$, we define the Frobenius norm as:

$$\|A\|_F = \sqrt{\sum_i^m \sum_j^n a_{ij}^2} \tag{7}$$

2.7.2 Max Norm

The max norm or infinity norm of a matrix equals to the largest sum of the absolute value of row vectors. We denote the max norm as $\|A\|_{max}$. Consider $A \in R^{m \times n}$. We define the max norm for A as:

$$\|A\|_{max} \coloneqq max_i \sum_j^n |a_{ij}| \tag{8}$$

This equals to go row by row, adding the absolute value of each entry, and then selecting the largest sum.

2.7.3 Spectral norm

The spectral norm of a matrix equals to the largest singular value σ_1. We denote the spectral norm as $\|A\|_2$. Consider $A \in R^{m \times n}$. We define the spectral for A as:

$$\|A\|_2 \coloneqq max_x \frac{\|Ax\|_2}{\|x\|_2} \tag{9}$$

2.8 Cross product

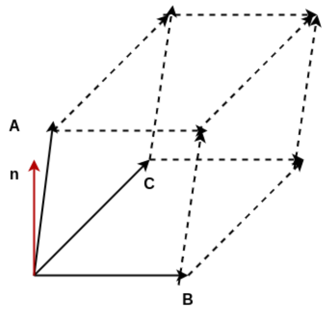
It is unlike dot product produces vectors after the operation. Cross product for vectors A and B is given as :

$$A \times B = \begin{vmatrix} i & j & k \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \tag{10}$$

The calculation is similar to the determinant but it does not give a number as i, j, k are unit vectors which in turn produces new vector from the cross product. Properties of cross product:

1. $|A \times B|$ gives area of the parallelogram created by A and B.
2. Direction of the vector $A \times B$ is perpendicular to the plane of A and B with right hand thumb rule.
3. $A \times B = -B \times A$
4. $A \times A = 0$ as two lines in the determinant are same.

In the determinant section, we saw that determinant of three vectors produces volume, we see that same can be obtained with cross product.



$$\begin{aligned} \text{volume} &= \text{area of base } \times \text{height} \\ &= |B \times C| \, (A.n) \\ &= |B \times C| \, \frac{A.B \times C}{|B \times C|} \\ &= A.(B \times C) \\ &= det(A, B, C) \end{aligned}$$

Application Suppose we want to determine if four points p_1, p_2, p_3, p lie on the same plane. Then we can get the determinant of the three vectors then if the determinant is 0 which shows that there is no volume to it which in turn means they lie on the same plane i.e. $det(p_1p, p_1p_2, p_1p_3) = 0$.

Other way is to consider cross product. We know, that suppose we take cross product of vectors it produces normal vector to the plane in which these two vectors lie. So normal vector $N = p_1p_2 \times p_1p_3$ is perpendicular to the plane. Then, if this vector N is also perpendicular to p_1p then we can say that vector p_1p also lies on the plane. Therefore with orthogonality condition, we can say:

$$\begin{aligned} p_1p.N &= 0 && \text{orthogonality} \\ p_1p.p_1p_2 \times p_1p_3 &= 0 \\ &= det(p_1p, p_1p_2, p_1p_3) && \text{triple product produces determiant} \end{aligned}$$

2.9 Orthogonality

The angle between two non-zero elements is given as:

$$cos\,\theta = \frac{\langle x, y \rangle}{||x||||y||}$$

By cauchy-shwartz inequality we know that, $|\langle x, y\rangle| \leq ||x||||y||$, therefore $cos\theta \in [-1, 1]$ so the $\theta \in (0, \pi)$.

Therefore, when the angle between two elements is 90° the $cos 90$ is 0. Which means $\langle x, y\rangle$ is 0. Orthogonal vectors are perpendicular to each other therefore, angel between them is 90, in turn their dot product is 0.

Orthonormal vectors In addition to the orthogonality, sometimes the vectors magnitude is 1, then these vectors are called orthonormal vectors. Let $q_1, q_2, \cdots, q_n$ are any vectors. They are called orthonormal vectors if:

$$q_i.q_j = \begin{cases} 0 & if \ \ i \neq j \\ 1 & if \ \ i = j \end{cases}$$

Orthogonal matrix If these vectors are arranged in a matrix form, then it is called a *Orthogonal matrix*, suppose it is denoted as Q. An orthogonal matrix is a square matrix whose rows are mutually orthonormal and whose columns are mutually orthonormal.

$$Q = \begin{bmatrix} q_1 & q_2 & \cdots & q_n \end{bmatrix}$$

Orthogonal matrices have some nice properties:

1. $Q^T Q = I$
2. $Q^{-1} = Q^T$
   so orthogonal matrices are of interest because their inverse is very cheap to compute. Pay careful attention to the definition of orthogonal matrices. Counterintuitively, their rows are not merely orthogonal but fully orthonormal. There is no special term for a matrix whose rows or columns are orthogonal but not orthonormal.

Transformations of vector with orthogonal matrix is special as the length of the vector is not changed.

$$\begin{aligned} ||Qx||^2 &= (Qx)^T(Qx) \\ &= x^T Q^T Q x \\ &= x^T I x \\ &= ||x||^2 \end{aligned}$$

Note that though the length of the vector remains same, the orthogonal matrix can change the direction of the vector, thus there is no special connection to eigenvalues or eigenvectors.

This property is intuitive, as multiplication of the vector is with the vector whose length is unit in each of column of the orthogonal matrix.

It is confusing! Why matrix containing orthonormal columns is called orthogonal? Then what is the name for a matrix containing orthogonal columns? This kind of matrix will be slightly different

than the matrix with orthonormal columns, the diagonal entries will be positive number instead of being 1 as diagonals will be squares of the column vectors.

Orthonormal Basis Consider a n-dimensional vector space V and a basis $\{b_1, b_2, \cdots, b_n\}$ of V. If

$$\langle b_i, b_j \rangle = 0 \qquad \forall i \neq j \tag{11}$$
$$\langle b_i, b_i \rangle = 1 \tag{12}$$

then the basis is called orthonormal basis. If only Eq.**??** is satisfied then it is called orthogonal basis.
Suppose, basis vectors which are non-orthogonal and unnormalized are given then gram-schmidt procedure is iteratively applied to obtain orthonormal basis.

Gram-Schmidt approach to get a orthonormal vectors from independent vectors. Suppose a, b, c are three vectors and we want to obtain q_1, q_2, q_3 which are orthonormal vectors derived from them. Then, we get first vector $q_1 = \frac{a}{||a||}$ which just makes it unit vector. Then we find a vector from b which is perpendicular to a. This is similar to what we did in projections, the only difference was that in projections we were interested in vector p which is in same direction as a but now we are interested in error vector $e = b - p$ which is perpendicular to a. Therefore, from projection formulas we write $e_b = b - \frac{q_1^T b}{q_1^T q_1} q_1$ and then it is normalized to get $q_2 = \frac{e_b}{||e_b||}$. Further we want to find a vector which is both perpendicular to q_1 and q_2. It can be obtained by subtracting both the components from vector c as $e_c = c - \frac{q_1^T c}{q_1^T q_1} q_1 - \frac{q_2^T c}{q_2^T q_2} q_2$. Then q_3 is obtained as: $q_3 = \frac{e_c}{||e_c||}$.

We can verify that these vectors are actually orthogonal by taking dot product, first we check e_b

$$\begin{aligned}
q_1^T e_b &= q_1^T (b - \frac{q_1^T b}{q_1^T q_1} q_1) \\
&= q_1^T b - q_1^T \frac{q_1^T b}{q_1^T q_1} q_1 \\
&= q_1^T b - \frac{q_1^T b}{\cancel{q_1^T q_1}} \cancel{q_1^T q_1} \\
&= q_1^T b - q_1^T b \\
&= 0
\end{aligned}$$

Further we check with e_c

$$\begin{aligned}
q_1^T e_c &= q_1^T(c - \frac{q_1^T c}{q_1^T q_1} q_1 - \frac{q_2^T c}{q_2^T q_2} q_2) \\
&= q_1^T c - q_1^T \frac{q_1^T c}{q_1^T q_1} q_1 - q_1^T \frac{q_2^T c}{q_2^T q_2} q_2 \\
&= q_1^T c - \frac{q_1^T c}{\cancel{q_1^T q_1}} \cancel{q_1^T q_1} - q_1^T \frac{q_2^T c}{q_2^T q_2} q_2 \\
&= -\frac{q_2^T c}{q_2^T q_2} q_1^T q_2 \\
&= 0 && \text{as we already know } q_1 \text{ and } q_2 \text{ are perpendicular}
\end{aligned}$$

QR decomposition In the last section, we saw how to obtain orthogonal matrix from a given matrix A. Now we see decomposition of A with Q. In this approach, another vector R is obtained such that

$$A = QR \tag{13}$$

The procedure for obtaining Q is described in the previous section. We see how to obtain R. We know, $Q^{-1} = Q^T$ because Q is an orthogonal matrix, with the use of this fact we can get $Q^T A = Q^T QR$ which is $R = Q^T A$.

3 Solving linear system

A general form of systems of equation can be given as:

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
&\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{aligned}$$

Here, $(x_1, \cdots, x_n)$ are n-tuple, which are unknown and value of them should be found so that the above equations are satisfied. The compact notation of the same linear systems is given as $Ax = b$. Thus, the main goal in this section is to find out the solution for $Ax = b$ equations where vector x is of $n \times 1$-dimension represents the unknown variable whose values need to be obtained. A is $m \times n$ matrix which transforms x linearly into m-dimensional space such that b is $m \times 1$ dimensional vector.

To analyse how many solutions the equation has, we can think of the columns of A as specifying different directions we can travel from the origin (the point specified by the vector of all zeros), and determine how many ways there are of reaching b. In this view, each element of x specifies how far we should travel in each of these directions, with x_i specifying how far to move in the direction of column i.

Many problems can be formulated as systems of linear equations, and linear algebra gives us the tools for solving them. Sometimes, there can be no solution for the system, then it is called *singular case*, and non-singular case is the opposite of that which may have one or more solutions. Let us look at some of the popular approaches of solving them.

3.1 Geometrical approach

In a *system of linear equations* with two variables x_1, x_2 , each linear equation defines a line on the x_1, x_2-plane. Since a solution to a system of linear equations must satisfy all equations simultaneously, the solution set is the intersection of these lines. This intersection set can be a line (if the linear equations describe the same line), a point, or empty (when the lines are parallel).

Suppose we want to solve the following equations:

$$4x_1 + 4x_2 = 5$$
$$2x_1 - 4x_2 = 1$$

The graphical representation of these equations can be given as shown in Figure **??**. As these, two lines intersect at a single point $(1, \frac{1}{4})$, those are the values of x_1, x_2 which solve these equations.

Similarly, for three variables, each linear equation determines a plane in three-dimensional space. When we intersect these planes, i.e., satisfy all linear equations at the same time, we can obtain a solution set that is a plane, a line, a point or empty (when the planes have no common intersection).

It is still possible to draw them and identify where these planes interest in 3-dimensional but it becomes increasingly difficult if the number of unknowns increase beyond 3. Therefore, to solve such equations sophisticated approaches are designed which are discussed below.

3.2 Gaussian Elimination approach

The method starts by subtracting multiples of some equations from the other equations so as to eliminate all variables except one from the last row i.e. to get 0 everywhere except one variable, then at most two variables in the last but one row and so on. The best way to illustrate this process is to solve a system of linear e equations.

Solve the equations and obtain the values of $u, v,$ and w.

$$\begin{aligned} 2u + v + w &= 5 \\ 4u - 6v &= -2 \\ -2u + 7v + 2w &= 9 \end{aligned}$$

Now this can be written as:

$$\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & -7 & 2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix} \tag{14}$$

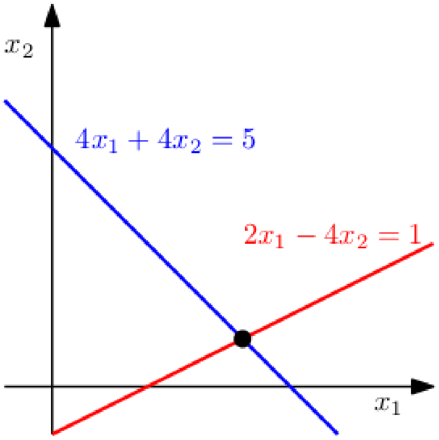


Figure 1: The two lines interest at a single point, producing an unique solution.

For the forward elimination step, we write it as where the vector of values is concatenated in the matrix:

$$\begin{bmatrix} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & -7 & 2 & 9 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 8 & 3 & 14 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 0 & 1 & 2 \end{bmatrix} \tag{15}$$

Once this form row-echelon form is obtained the value is back substituted to get the actual values of $u, v,$ and w. Always, the variables value is obtained first as it does not depend any other variable. Then the second last and so on. By this process, we first obtain $w = 2$, then $v = 1$, and finally $u = \frac{3}{2}$.

3.3 LU decomposition

Lower–upper (LU) decomposition or factorization factors a matrix as the product of a lower triangular matrix and an upper triangular matrix. The product sometimes includes a permutation matrix as well. LU decomposition can be viewed as the matrix form of Gaussian elimination. Computers usually solve square systems of linear equations using LU decomposition, and it is also a key step when inverting a matrix or computing the determinant of a matrix. Traditionally applicable to only square matrix A, although rectangular matrices can be applicable

Let A be a square matrix. An LU factorization refers to the factorization of A, with proper row and/or column orderings or permutations, into two factors – a lower triangular matrix L and an upper triangular matrix U:

$$A = LU \tag{16}$$

In the lower triangular matrix all elements above the diagonal are zero, in the upper triangular matrix, all the elements below the diagonal are zero. For example, for a 3 × 3 matrix A, its LU decomposition looks like this:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \tag{17}$$

To get this decomposition, gaussian elimination process which was described in the previous section is applied with trick of placing identity matrix with it and applying similar operations on the identity matrix. Let us see this with an example.

Consider $A = \begin{bmatrix} 2 & 4 & 3 & 5 \\ -4 & -7 & -5 & -8 \\ 6 & 8 & 2 & 9 \\ 4 & 9 & -2 & 14 \end{bmatrix}$. Then the decomposition is started with identity matrix which ends up being a lower triangular matrix and the other one after gaussion elimination becomes upper triangular matrix. Let us see:

$$\begin{aligned} \begin{bmatrix} 2 & 4 & 3 & 5 \\ -4 & -7 & -5 & -8 \\ 6 & 8 & 2 & 9 \\ 4 & 9 & -2 & 14 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & 3 & 5 \\ -4 & -7 & -5 & -8 \\ 6 & 8 & 2 & 9 \\ 4 & 9 & -2 & 14 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & 3 & 5 \\ 0 & 1 & 1 & 2 \\ 0 & -4 & -7 & -6 \\ 0 & 1 & -8 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & -4 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & 3 & 5 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & -3 & 2 \\ 0 & 0 & -9 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & -4 & 1 & 0 \\ 7 & -13 & 3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & 3 & 5 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & -3 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix} \end{aligned}$$

which is of the form $A = LU$.

3.4 Pivot and free variables

Definition 3.1 (Row-echelon form)**.** A matrix is in row-echelon form if

- All rows that contain only zeros are at the bottom of the matrix; correspondingly, all rows that contain at least one nonzero element are on top of rows that contain only zeros.

- Looking at nonzero rows only, the first nonzero number from the left (also called the *pivot* or the leading coefficient) is always strictly to the right of the pivot of the row above it.

The variables corresponding to the pivots in the row-echelon form are called *basic variables* or *pivot variables* and the other variables are *free variables.* The number of pivot variables are also called a ***rank*** of matrix. In other words, rank of matrix is number of linearly independent columns as pivot columns are independent. Also, we know that number of linear independent rows is equal to number of independent columns, hence rank of matrix is equal to number of independent rows as well.

Rank of matrix A: $\rho(A)$ and augmented matrix $(A : b)$: $\rho(A : b)$ give good criterias to understand if the solution for the system exists or not which is as follows:

- $\rho(A) \neq \rho(A : b)$ - solution doe not exist
- $\rho(A) = \rho(A : b) = n$ - unique solution exists (n is number of variables)
- $\rho(A) = \rho(A : b) < n$ - multiple solutions exist

Definition 3.2 (Reduced row-echelon form)**.** An equation system is in *reduced row-echelon form* (also: row-reduced echelon form or row canonical form) if reduced row-echelon form:

- It is in row-echelon form.
- Every pivot is 1.
- The pivot is the only nonzero entry in its column.

$$U = \begin{bmatrix} \bullet & * & * & * & * & * & * & * \\ 0 & \bullet & * & * & * & * & * & * \\ 0 & 0 & 0 & \bullet & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bullet \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad R = \begin{bmatrix} \mathbf{1} & \mathbf{0} & * & \mathbf{0} & * & * & * & \mathbf{0} \\ 0 & \mathbf{1} & * & \mathbf{0} & * & * & * & \mathbf{0} \\ 0 & 0 & 0 & \mathbf{1} & * & * & * & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 2: U denotes echelon form of 5×8 matrix and R denotes the reduced row echelon form. Note the difference, in U the pivot can have any value other than 1 but R should have only 1 also, after pivot variable there can be any value in echelon form (U) but reduced row echelon form R contains only zeros.

The elimination process can give solutions for only special type of matrices i.e. those matrices which have linearly independent columns and in turn which are invertible. Thus, they also have same number of pivots as the number of variables, but in those cases where pivots are lesser than number of variables i.e. if matrix contains number of free variables we can not apply the previous method. In those cases we first find a particular solution x_p for *non-homogeneous system* $Ax_p = b$ and x_n for *homogeneous system* $Ax_n = 0$. The solutions to all linear equations have this form, $x = x_p + x_n$:

Complete solution $\quad Ax_p = b$ and $Ax_n = 0$ **produce** $A(x_p + x_n) = b$

A possibly surprising fact, which has not yet been proved, is that no matter what sequence of row operations you use to put the system into echelon form, you always get the same number of free variables. This means that the number in the system is uniquely determined by the system.

A ***homogeneous system*** of linear equations is one in which all of the constant terms are zero. A homogeneous system always has at least one solution, namely the zero vector. When a row operation is applied to a homogeneous system, the new system is still homogeneous. It is important to note that when we represent a homogeneous system as a matrix, we often leave off the final column of constant terms, since applying row operations would not modify that column. So, we use a regular matrix instead of an augmented matrix. Of course, when looking for a solution, it's important to take the constant zero terms into account.

A ***nonhomogeneous system*** has at least one constant term as non-zero and it always has an associated homogeneous system, which you get by replacing the constant term in each equation with zero.

Let us see this with an example,

$$\begin{bmatrix} 2 & 4 & 6 & 4 \\ 2 & 5 & 7 & 6 \\ 2 & 3 & 5 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 5 \end{bmatrix}$$

We want to solve this system of equations. First we find a particular solution corresponding to the non-homogeneous system of equation. For that we form a augmented matrix $A|b$ and get the echelon form (row-reduced echelon form is not needed)

$$\begin{bmatrix} [cccc|c]2 & 4 & 6 & 4 & 4 \\ 2 & 5 & 7 & 6 & 3 \\ 2 & 3 & 5 & 2 & 5 \end{bmatrix} \xrightarrow{R_2-R_1, R_3-R_1} \begin{bmatrix} [cccc|c]2 & 4 & 6 & 4 & 4 \\ 0 & 1 & 1 & 2 & -1 \\ 0 & -1 & -1 & -2 & 1 \end{bmatrix}$$

$$\xrightarrow{R_3+R_2} \begin{bmatrix} [cccc|c]2 & 4 & 6 & 4 & 4 \\ 0 & 1 & 1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

This is in echelon form where column 1 and 2 are pivot as they have first non-zero entries which also means x_1, x_2 are pivot variables and x_3, x_4 are free variables. The free variables can take any values but pivot variables depend on these free variables. By back substitution we get:

$$\begin{aligned} 2x_1 + &4x_2 + 6x_3 + 4x_4 = 4 \\ &x_2 + x_3 + 2x_4 = -1 \end{aligned}$$

Suppose, we set free variables at 0, then $x_2 = -1$ and $x_1 = 4$ which gives following particular solution:

$$x_p = \begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

Now let us solve for nullspace solutions x_n, for that we start from echelon form where we put constants as 0s: $\begin{bmatrix} [cccc|c]2 & 4 & 6 & 4 & 0 \\ 0 & 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{R_1 - 4R_2}$ $\begin{bmatrix} 2 & 0 & 2 & -6 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ which yields following equations.

$$\begin{aligned} 2x_1 + 2x_3 - 6x_4 = 0 &\rightarrow x_1 = -x_3 + 3x_4 \\ x_2 + x_3 + 2x_4 = 0 &\rightarrow x_2 = -x_3 - 2x_4 \end{aligned}$$

Then the solution for nullspace is given as:

$$x_n = \begin{bmatrix} -x_3 + 3x_4 \\ -x_3 - 2x_4 \\ x_3 \\ x_4 \end{bmatrix}$$

Now, we set free variables x_3, x_4 to 0s simeltatneously which produces following:

$$x_n = x_3 \begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 3 \\ -2 \\ 0 \\ 1 \end{bmatrix}$$

Then finally, we combine particular and nullspace solution to produce complete solution as:

$$\begin{aligned} x_c &= x_p + x_n \\ x_c &= \begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 3 \\ -2 \\ 0 \\ 1 \end{bmatrix} \end{aligned}$$

4 Transformations

A vector can be transformed into other vector with some function applied to it. We look at two popular types of transformations.

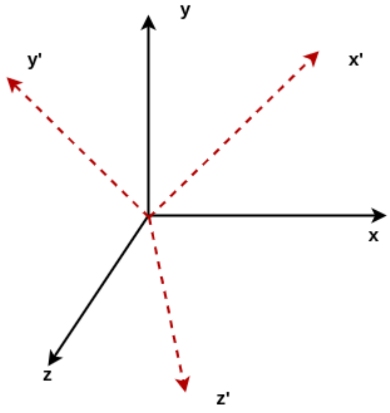


Figure 3: With some arbitrary transformation original axes are transformed to new axes. The vectors are transformed to new vectors which lie in the new axes planes.

4.1 Linear transformation

A linear transformation can be represented with a matrix. **Can we say that each matrix produces a linear transformation?** Some of the popular transformations are shown in the following equations.

$$\begin{array}{ccc}
\begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} & \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} & \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \\
Scaling & \text{Rotation clockwise} & \text{Rotation anti-clockwise} \\
\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \\
\text{Reflection (x-axis)} & \text{Reflection (y-axis)} & \text{Reflection (origin)} \\
\begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix} & \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & d \\ c & 1 \end{bmatrix} \\
\text{Shearing (x-axis)} & \text{Shearing (y-axis)} & \text{Shearing (xy-axis)}
\end{array}$$

Definition 4.1. If V, W are two linear spaces (or subspaces) then the transformation $\Psi : V \rightarrow W$ is called *linear transformation* or *homomorphism* if it follows the laws of addition and multiplication i.e. $\Psi(ax + by) = a\Psi(x) + b\Psi(y)$.

A transformation or mapping is in other words is a function which takes some input and produces transformed output. Let $\mathcal{V}, \mathcal{W}$ be some arbitrary sets, let $\Phi : \mathcal{V} \rightarrow \mathcal{W}$ be some transformation. Then,

1. If $\forall x, y \in \mathcal{V}$ if $\Phi(x) = \Phi(y) \implies x = y$, then the transformation Φ is called injective. In other words, for an injective transformation, there should be a unique map for a distinct input.

2. If $\Phi(\mathcal{V}) = \mathcal{W}$, then the transformation Φ is called surjective. In other words, each element of the output set $\mathcal{W}$ has some corresponding input in the set $\mathcal{V}$. This may give an impression that $|\mathcal{V}| = |\mathcal{W}|$, but this does not have to be true. As, there can be multiple inputs mapping to a same output. Because of this, there may not be a inverse mapping as a single output can not be uniquely mapped to input.

3. If a mapping is both injective and surjective then it is bijection. It also means that the inverse mapping can be found, as in this case we can uniquely map an element from $\mathcal{W}$ to $\mathcal{V}$.

Given these some special cases of transformations, we look at the special cases associated specifically with *linear* transformations.

1. Isomorphism: $\Phi : \mathcal{V} \to \mathcal{W}$ if Φ is linear transformation and bijection.

2. Endomorphism: $\Phi : \mathcal{V} \to \mathcal{V}$ if Φ is linear transformation.

3. Automorphism: $\Phi : \mathcal{V} \to \mathcal{V}$ if Φ is linear transformation and bijection.

4.2 Affine transformation

Affine transformation is closely related to linear transformation so both of them are used interchangeably in the machine learning literature. However, both of them have very different meanings

Definition 4.2. If V, W are two linear spaces (or subspaces) then the transformation $\Phi : V \to W$ is called linear transformation if it is of the form: $\Phi(x) = \Psi(x) + u$ where $u \in W$ and Ψ is a linear transformation. The element u is called translation vector of Ψ. Every affine transformation or mapping is always composed of linear mappings Φ and translation $\tau : W \to W$ such that $\Phi = \tau \circ \Psi$.

Composition of two affine transformations $\Phi' \circ \Phi$ is also an affine transformation. These affine transformations keep the geometric structure invariant. They also preserve the dimension and parallelism.

4.3 Projections

It is important to find out what will be the corresponding vector for any vector in the given subspace. For example, some times we require to find out a *projection* of a vector on a line. We see this pictorially in Figure **??**.

The goal in the projection is to find a vector living the given plane which is corresponding to a original vector. The projection plane can be a plane, a line or a point. For the sake of simplicity we start with a line. From figure **??**, we can see that, we intend to project vector x on line l, the line is nothing but a set of all vectors which are multiples of unit vector v i.e. $l = \{cv | c \in R\}$. The goal is to find vector p which is a projection of x on line l. To get the projection on the line, we put the perpendicular on the line from a given vector which is

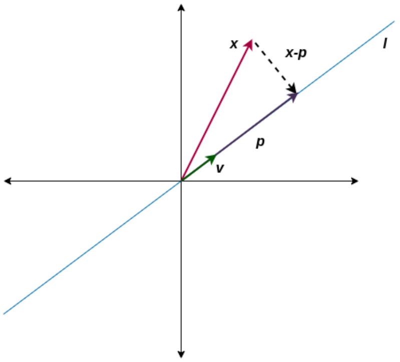


Figure 4: Projection of vector x on line l with unit vector v. The projection is given as vector p.

a nearest point from the vector to the projection line. This vector is given as $x - p$. We know that $x - p$ and p are orthogonal which means their dot product is 0. One important point to consider is that vector p lies on line l, therefore it will be some multiple c of unit vector v i.e. $p = vc$.

$$\begin{aligned}(x-p).p &= x.p - p.p \\ 0 &= x.cv - cv.cv \\ cx.v &= c^2||v||^2 \\ c &= \frac{x.v}{||v||^2} \\ \therefore \boxed{p = v \ \frac{v^T x}{||v||^2}}\end{aligned}$$

Note that the term $\frac{x.v}{||v||^2}$ is a constant and not a vector as it is a division of two numbers. We can further simplify it. We know that vector v is a unit vector, so $||v||^2 = 1$, so we can write

$$\boxed{p = v\ (v^T x)}$$

Projection can be seen as a transformation which produces a new vector for a given vector. Further, it can also be shown easily that, it is a linear transformation which means a matrix can be found to get a projection. This is an interesting thing, as we can find a compact matrix and then multiplying it with any vector, we can find out its projection on the line. Let vector $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be a unit vector which defines the line on which we project, then we know that projection vector is given as following:

$$\begin{aligned}
p &= v \ \frac{v^T x}{v^T v} \\
&= \frac{v v^T}{v^T v} \ x && \text{adjusting terms to get multiples of x} \\
&= \frac{1}{||v||^2} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix} \ x && \text{actually putting values of v} \\
&= \begin{bmatrix} v_1^2 & v_1 v_2 \\ v_1 v_2 & v_2^2 \end{bmatrix} \ x && \text{with the assumption that v is a unit vector} \\
&= \mathcal{P} x
\end{aligned}$$

where projection matrix is:

$$\mathcal{P} = \begin{bmatrix} v_1^2 & v_1 v_2 \\ v_1 v_2 & v_2^2 \end{bmatrix}$$

The projection matrix has some nice properties.

1. Its column space is a line on which we project any vector.
2. It is a symmetric matrix which means: $\mathcal{P}^T = \mathcal{P}$
3. Suppose we project a vector twice, which is like taking square, we will land on the same line. This means $\mathcal{P}^2 = \mathcal{P}$, in generic sense, we can write $\mathcal{P}^n = \mathcal{P} \ \forall n \geq 1$

Find a projection matrix for a line passing through vector $v = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.
First, we find the unit vector associated with v as $u = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ Now the projection matrix from this vector is given as: $\frac{1}{5} \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$

Higher dimension projections So far, we have seen a projection on a line, now we extend our study for higher dimensions. One of the need of this projections is to solve a linear system of equations which is unsolvable i.e. if there are more equations than variables and the vector b does not lie in the columns space of A while solving equation $Ax = b$. Then in that case, we project b into most probable vector p which is equal to $p = A\hat{x}$. Then in this, our goal is to get $\hat{x}$.

We see that, $e = b - p$ is the error which incurred because we could not find the actual vector. This error vector e is always perpendicular to the plane i.e. $b - A\hat{x}$ is perpendicular to the basis of plane A. which can be written as:

$$a_1^T (b - A\hat{x}) = 0 \tag{18}$$

$$a_2^T (b - A\hat{x}) = 0 \tag{19}$$

In a matrix form it can be written as: $A^T(b - A\hat{x}) = 0$, then we solve for $\hat{x}$,

$$\begin{aligned} A^T b &= A^T A\hat{x} \\ \hat{x} &= (A^T A)^{-1}\ \ A^T b \\ p &= A\,(A^T A)^{-1} A^T b \end{aligned}$$

This shows that the projection matrix is $\boxed{\mathcal{P} = A(A^T A)^{-1} A^T}$.
Now, if we can invert A then, we can simplify projection matrix $\mathcal{P}$ as:

$$\begin{aligned} \mathcal{P} &= A(A^T A)^{-1} A^T \\ &= A\,A^{-1}\,(A^T)^{-1} A^T \\ &= I \end{aligned}$$

This seems weird, because we were expecting some matrix which can actually project b but now we got projection matrix as identity matrix. The main flaw in this argument is that, we assumed A invertible which is not the case. In case A is invertible we will get the whole space is spanned A, so the projection b into that plane is b itself which is produced by identity.

There are two possibilities for b:

1. If b is in column space of A then $\mathcal{P}b = b$.
2. If b is perpendicular column space of A then $\mathcal{P}b = 0$. This follows from the observation that, if b is perpendicular to the column space of A it must be in null space of A which means $A^T b = 0$, so the whole term becomes 0.

Let us go back to the projection matrix: $\mathcal{P} = A(A^T A)^{-1} A^T$. Suppose A is an orthogonal matrix. Then we know that $A^T A = I$ which simplifies the projection matrix greatly and becomes: $\mathcal{P} = AA^T$. This simplification is the motivation behind converting A into an orthogonal matrix so that calculation becomes easy.

5 Matrix

A **matrix** in the cruder sense contains a number of rows and columns of numbers. This constituents two dimensions of matrix one is row and other is column, therefore it is generally denoted as $m \times n$ mathematical entity where m denotes number of rows and n denotes number of columns of the matrix. For example,

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \tag{20}$$

Vector is a special case of matrix which contains only one column and n rows. A vector can have different meanings and interpretations. It can only mean the sequence of numbers in computer science, in physics it can represent the direction and magnitude of the object and mathematically it is a point in a n-dimensional space. For example:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \tag{21}$$

Tensor is a generalized matrix which can be of multiple dimensions such as $m_1 \times m_2 \times \cdots m_k$.

5.1 Different operations on Matrix

5.1.1 Addition

Two matrices can be added only if they are of same dimension. If $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{m\times n}$ then addition of A and B is $C \in \mathbb{R}^{m\times n}$ where $c_{ij} = a_{ij} + b_{ij}$.

5.1.2 Multiplication

Two matrices can be multiplied only if they are of right dimensions. If we want to multiply two matrices A and B then A should have same number of columns as B having number of rows. Formally, $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times q}$ then matrix multiplication AB exists. Even though, multiplication of AB exists, BA is not possible. This is because $AB \neq BA$.

Also, any matrix can be multiplied by a scalar. Such scalar multiplication scales each element of matrix by its value.

What does the matrix multiplication indicate? A matrix can be seen as a linear transformation over vectors so when we multiply two matrices what do we get?

5.2 Hadamard Multiplication

It is an element-wise multiplication of same dimensional matrices producing the same dimensional matrix. For matrix $A, B \in R^{m\times n}$ the matrix obtained with Hadamard product is $C = A \odot B$ such that $c_{ij} = a_{ij}b_{ij}$. For matrices of different dimensions ($m \times n$ and $p \times q$, where $m \neq p$ or $n \neq q$), the Hadamard product is undefined.

5.2.1 Division or Inverse

For some matrices, we can find a matrix which can make the original matrix an identity matrix, such matrix is the inverse of a matrix. **Inverse of the matrix exists only for the square matrix ($n \times n$. I think it makes sense as we**

know $AA^{-1} = A^{-1}A = I$). A matrix for which its inverse exists, it is called as invertible matrix and its determinant is non-zero. For square matrices, the left inverse and right inverse are equal.

To obtain matrix inverse, number of row operations are performed to go from original matrix to the inverse. **Reduced row echelon form**

5.2.2 Determinants

It finds a value of a matrix. For a 2×2 matrix A, determinant $|A|$ is given as:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad |A| = a_{11}a_{22} - a_{21}a_{12} \tag{22}$$

1. Identity matrix has determinant is 1.

2. If we exchange the row, the sign changes. This means determinants of permutation matrices is either 1 or -1 as if we exchange rows odd times we get -1 and with even exchanges we get 1.

3. If we multiply by some constant to the first row and leave n-1 rows as it is then the constant can be factored out. $\begin{vmatrix} ta & tb \\ c & d \end{vmatrix} = t\begin{vmatrix} a & b \\ c & d \end{vmatrix}$

4. Similar with the addition to the first row elements

$$\begin{vmatrix} a+a' & b+b' \\ c & d \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a' & b' \\ c & d \end{vmatrix}$$

5. If the 2 rows of the matrix are same then the determinant is 0. This comes from the first property where we know, row exchanges changes sign of determinant. But, with the two same rows, exchange of rows produces same matrix which has opposite determinant signs, which means its determinant is 0.

6. Subtracting l times from any row does not change the determinant. Let us see this with 2×2 matrix:

$$\begin{aligned} \begin{vmatrix} a & b \\ c-la & d-lb \end{vmatrix} &= \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a & b \\ -la & -lb \end{vmatrix} \\ &= \begin{vmatrix} a & b \\ c & d \end{vmatrix} - l\begin{vmatrix} a & b \\ a & b \end{vmatrix} \\ &= \begin{vmatrix} a & b \\ c & d \end{vmatrix} \end{aligned}$$

7. Rows of zeros produce determinant 0.

8. If the determinant is 0 then matrix is non-invertible and singular otherwise invertible.

9. Determinant of triangular matrix is multiplication of diagonal elements. $|U| = \begin{vmatrix} d_1 & * & \cdots & * \\ 0 & d_2 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & d_n \end{vmatrix} = d_1 d_2 \cdots d_n$ This is true for both upper and lower triangular matrices as well as diagonal matrices. With this rule, we can find determinant of any matrix. First we get the matrix and obtain its triangular form or row-echelon form. To get this form we do some row operations and we know that determinant does not change from the row operations,so determinant of the original matrix and reduced matrix is same. Let us see this with simple matrix:

$$\begin{aligned} |A| &= \begin{vmatrix} a & b \\ c & d \end{vmatrix} \\ &= \begin{vmatrix} a & b \\ 0 & d - b\frac{c}{a} \end{vmatrix} \\ &= a(d - b\frac{c}{a}) \\ &= ad - bc \end{aligned}$$

10. det(AB)=det(A)det(B). Also, from that det(A^{-1})=$\frac{1}{det(A)}$. Further, det(A^2)=det(A)^2, det(2A)=2^ndet(A).

11. det(A^T)= det(A). This means, column with 0s produce determinant is 0. This can be proved easily by the assumption that most of the matrices A can be written as product of lower and upper triangular matrices $A = LU$ then from the property 10, we can see that determinant of these triangular matrices is just multiplication of their diagonals which remains same even after transpose.

We figured out the formula for getting determinants of 2×2 matrices but have not generalized to $n \times n$ matrices. We intend to do it now. For that, we derive the determinant of 2×2 and 3×3 matrices from the three basic properties (1,2,3 mentioned in the previous paragraphs).

Suppose, $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we take a systematic approach to get its determinant which uses only first three properties of determinants.

$$\begin{aligned} |A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} &= \begin{vmatrix} a & 0 \\ c & d \end{vmatrix} + \begin{vmatrix} 0 & b \\ c & d \end{vmatrix} && \cdots \text{property 4} \\ &= \begin{vmatrix} a & 0 \\ c & 0 \end{vmatrix} + \begin{vmatrix} a & 0 \\ 0 & d \end{vmatrix} + \begin{vmatrix} 0 & b \\ c & 0 \end{vmatrix} + \begin{vmatrix} 0 & b \\ 0 & d \end{vmatrix} && \cdots \text{Same operation on the se} \\ &= 0 + \begin{vmatrix} a & 0 \\ 0 & d \end{vmatrix} + \begin{vmatrix} 0 & b \\ c & 0 \end{vmatrix} + 0 && \cdots \text{Columns are 0s} \\ &= ad - bc \cdots \text{One row exchange is needed to get the diagonal matrix in the} \\ &\qquad \text{second term, hence negative sign} \end{aligned}$$

Now suppose, $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$, we do the similar expansion as we did for the 2×2 matrix. But we neglect all those matrices which have 0 columns. Thus, we are left with very interesting permutations of the matrices, as we choose one element from each row and keep other 0 also, we maintain only one non-zero term in the column as well. So these terms are given as:

$$\begin{aligned}
|A| &= \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \\
&= \begin{vmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & 0 & 0 \\ 0 & 0 & a_{23} \\ 0 & a_{32} & 0 \end{vmatrix} + \begin{vmatrix} 0 & a_{12} & 0 \\ a_{21} & 0 & 0 \\ 0 & 0 & a_{33} \end{vmatrix} + \begin{vmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ a_{31} & 0 & 0 \end{vmatrix} \\
&+ \begin{vmatrix} 0 & 0 & a_{13} \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{vmatrix} + \begin{vmatrix} 0 & 0 & a_{13} \\ 0 & a_{22} & 0 \\ a_{31} & 0 & 0 \end{vmatrix} \\
&= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}
\end{aligned}$$

Terms in the determinants show very interesting pattern. The second index of all the terms are permutations of (1,2,3), hence there are $6 = 3! = 3 \times 2 \times 1$ terms out of that half are positive and half negative. Therefore the determinant terms of generalized $n \times n$ matrix is given as $\sum_{n!} \pm\, a_{1\alpha}\; a_{2\beta}\; a_{3\gamma}\; \cdots\; a_{n\omega}$. The values of $(\alpha, \beta, \gamma \cdots \omega) =$ permutations of $1, 2, \cdots, n$. The sign of the term is determined by how many times we have to exchange position of numbers to get them in proper descending order. For example, (3,2,1) will be of negative sign as there is one operation to get (1,2,3).

Further, the determinant can be viewed as multiplication of terms and cofactors. For example,

$$|A| = a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \quad (23)$$

This can be generalized to all the terms and cofactor formula can be given as:

$$|A| = a_{11}C_{11} + a_{12}C_{12} + \cdots + a_{1n}C_{1n} \quad (24)$$

Now, with the use of this we can find A^{-1} which is given as:

$$A^{-1} = \frac{1}{|A|}C^T \quad (25)$$

where C is a cofactor matrix for A. This formula can be used to find the $x = A^{-1}b$.

Area and Volume Suppose there are two vectors, $A = (a_1, a_2)$ and $B = (b_1, b_2)$, then $det(A, B)$ gives area of parallelogram formed by these vectors which is equal to $\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}$. Suppose there are three vectors in a space $A = (a_1, a_2, a_3)$,$B = (b_1, b_2, b_3)$, and then $C = (c_1, c_2, c_3)$ $det(A, B, C)$ gives volume of parallelepiped (box) as $\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}$

5.2.3 Transpose

This operation exchanges row by columns and columns by row of the matrix which means dimension of the matrix are changed. For a matrix A of $m \times n$ dimension, transpose B matrix is given as $b_{ij} = a_{ji}$ and denoted as A^T. The dimensions of the new matrix are $n \times m$.

$$A = \begin{bmatrix} 1 & 2 & 3 & 12 \\ 4 & 5 & 6 & 11 \\ 7 & 8 & 9 & 10 \end{bmatrix} \; B = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \\ 12 & 11 & 10 \end{bmatrix} \tag{26}$$

Few properties of transpose

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $x^T y = y^T x$ dot product is commutative.

5.3 Types of Matrix

5.3.1 Symmetric matrix

The matrix which has same elements over and above the diagonal i.e. $a_{ij} = a_{ji}$ $\forall i, j \in n$ for a $n \times n$ matrix. For example, a 3×3 symmetric matrix is given as:

$$A = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} \tag{27}$$

For any symmetric matrix A: $A^T = A$ as only other than diagonal elements are shifted in transpose. Also, the symmetric matrix has to be *square matrix* i.e. a matrix which has same number of rows and columns.

The eigenvalues of the symmetric matrix are real and their eigenvectors are perpendicular.

Symmetric matrices often arise when the entries are generated by some function of two arguments that does not depend on the order of the arguments. For example, if A is a matrix of distance measurements, with $A_{i,j}$ giving the distance from point i to point j, then $A_{i,j} = A_{j,i}$ because distance functions are symmetric.

5.3.2 Diagonal Matrix:

Other than diagonal elements all other elements are 0. For example:

$$A = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \tag{28}$$

Diagonal matrices are of interest in part because multiplying by a diagonal matrix is very computationally efficient. To compute $\text{diag}(v)x$, we only need to scale each element x_i by v_i. In other words, $\text{diag}(v)x = v \odot x$ (hadamard multiplication i.e. element-wise multiplication). Inverting a square diagonal matrix is also efficient. The inverse exists only if every diagonal entry is nonzero, and in that case, $\text{diag}(v)^{-1} = \text{diag}([1/v_1, \cdots, 1/v_n]^T)$. In many cases, we may derive some very general machine learning algorithm in terms of arbitrary matrices, but obtain a less expensive (and less descriptive) algorithm by restricting some matrices to be diagonal.

In general diagonal matrix are considered to be square matrices but there is a possibility that we can call rectangular matrix as diagonal matrix where $a_{ij} = 0$ for $i \neq j$. In rectangular matrix also we have concept of diagonal where row and column are equal and those diagonals are called main or primary or leading or major diagonals. In a rectangular matrix, the total number of elements in a row is not equal to the total number of entries in a column but there is an element whose row and column are equal in every row or column of a rectangular matrix, those elements can be connected by a straight path diagonally and it is called a main diagonal of a rectangular matrix.

5.3.3 Antidiagonal matrix

Matrices are said to be antidiagonal when all the entries are zero but the antidiagonal (i.e., the diagonal starting from the bottom left corner to the upper right corner). For instance:

$$B = \begin{bmatrix} 0 & 0 & a \\ 0 & b & 0 \\ c & 0 & 0 \end{bmatrix} \tag{29}$$

5.3.4 Upper and lower triangular matrix

: Upper triangular matrix has only those elements which are above diagonal as 0 and lower triangular matrix is opposite of upper triangular matrix. If any type of matrix out of these two is transposed, it produces the other type of matrix.

$$A_U = \begin{bmatrix} a & d & e \\ 0 & b & f \\ 0 & 0 & c \end{bmatrix} \quad A_L = \begin{bmatrix} a & 0 & 0 \\ b & d & 0 \\ c & e & f \end{bmatrix} \tag{30}$$

5.3.5 Identity matrix

: It is a diagonal matrix with all diagonal entries as 1. It is most special type of matrix which has determinant 1 and other interesting properties like multiplication with any other matrix produces the other matrix, its inverse is itself, and so on. It is a square matrix $n \times n$ and denoted as I_n. A 3×3 identity matrix is given as:

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{31}$$

5.3.6 Permutation matrix

:A permutation matrix is a square binary matrix that has exactly one entry of 1 in each row and each column and 0s elsewhere. Identity matrix is a special type of permutation matrix. The other permutation matrices of 3×3 sizes are given as:

$$\begin{gathered} P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \\ P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{gathered} \tag{32}$$

For any $n \times n$ matrices, there can be $n!$ permutation matrices. I think it is true. But I have to verify.

5.3.7 Positive definite matrix

A $n \times n$ symmetric matrix A is called positive definite iff $x^T\ A\ x > 0$ for any non-zero vector x. For example, consider the identity matrix $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, then

$$x^T I x = [a\ \ b] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = a^2 + b^2 > 0 \ \ \forall a, b \in R \tag{33}$$

It shows that I is a positive definite matrix.

Following conditions are sufficient to prove that a symmetric matrix is a positive definite matrix

1. All the eigenvalues of the matrix are real and positive i.e. $\lambda_i > 0$. The condition on eigenvalues being real is true even in the case of symmetric matrix, the condition of positivity is further extended for positive definite matrices. This means that the determinant is also positive as it is nothing but the product of eigenvalues. (Note: sum of diagonals (trace) is also sum of eigenvalues.)

This is interesting to see the connection between convex function and eigenvalues. A function is said to be convex if its Hessian matrix, the matrix of second order derivatives, is positive definite. This means if the eigenvalues of Hessian matrix are positive and real then the function is convex. This is really beautiful concept which connects linear algebra and calculus.

2. Another thing which is equal to determinant is product of pivots. Thus pivots of the positive definite matrix should also be positive. There is another theorem which states that number of signs of eigenvalues are equal to number of signs of pivots, because of that also we can say that all the pivots are positive.

3. The energy based equation $x^TAx > 0$ holds in case of A being positive definite as seen earlier. (Frequently in physics the energy of a system in state x is represented as $x^TAx \ \ (\frac{1}{2}x^TAx)$ and so this is frequently called the energy-based definition of a positive definite matrix.)

4. If the positive definite matrix can be written as $A = B^TB$ (independent columns in B).

5. All the leading determinants of the matrix are positive. This means, for $n \times n$ matrix, all the sub-determinants obtained by considering matrices of $n, n-1, n-2, \cdots 1$ should be positive.

6. An inner product between vectors x, y, $\langle x, y \rangle$ always can be expressed as x^TAy where matrix A determines the value of the inner product. This matrix always has to be positive (semi)definite.

Few facts about positive definite matrices

1. If two matrices S, T are positive definite then $S+T$ is also positive definite. We can prove this with simply putting them in the energy based definition.

2. If matrix S is positive definite then its inverse S^{-1} is also positive definite as it will have positive and real eigenvalues.

3. For any $m \times n$ matrix A, matrix obtained as A^TA is positive definite. Because $x^TA^TAx = (Ax)^T(Ax) = |Ax|^2$ which is always greater 0 as it is the length of the vector (equivalent of squaring the terms).

If the condition of positivity is relaxed and if it can be equal to 0 then it is called positive semi-definite i.e. iff $x^T \ A; x \geq 0$ for any non-zero vector x. Opposite to the positive definite matrix, negative definite matrix A is obtained if $x^T \ A; x < 0$. Also, A is called negative semi-definite or non-positive definite if $x^T \ A; x \leq 0$.

5.3.8 Similar matrices

Two matrices A and B of size $n \times n$ are called similar if they can be written as

$$A = M^{-1}BM \tag{34}$$

where M is any matrix.

The other important part is that they have same eigenvalues.

5.4 Rank of matrix

Earlier we briefly mentioned about the rank of a matrix, it is the number of linearly independent columns of a matrix. It equals number of independent rows of the matrix. The rank of matrix A is denoted as $\text{rk}(A)$. Following are some of the properties of the rank:

- $\text{rk}(A) = \text{rk}(A^T)$ **It means that in case of rectangular matrix where $m \neq n$ we should have $|m - n|$ free variables as the rank of the row and column matrix should be same, so some of the columns (rows) have to be linearly independent.**
- The columns of $A \in R^{m \times n}$ span the subspace $U \subset R^m$ with $dim(U) = \text{rk}(A)$. This means that basis required to span U are independent columns, if we want to get these independent columns, first we use Gaussian elimination to obtain echelon form. Further, if we want to obtain orthonormal basis, we use Gram-schmit approach to get them.
- Similarly, rows of matrix A span subspace $W \subset R^n$ with $dim(W) = \text{rk}(A^T)$ and these basis can be found by applying gaussin elimination over A^T.
- A matrix $A \in R^{n \times n}$ is invertible iff it has $\text{rk}(A) = n$, it means it has pivot variables on the diagonal. If one of pivot elements is 0 then, the determinant will be zero as the row echelon form will produce upper triangular matrix and determinant of these matrices is multiplication of diagonals.
- A matrix $A \in R^{m \times n}$ has *full rank* if its rank equals largest possible rank for a matrix of the same dimension. This means that the rank of a full matrix is $min(m, n)$.

5.4.1 Rank-Nullity theorem

For that first we define Image and Kernel.

Definition 5.1. For $\Phi : V \rightarrow W$, we define kernel/null space as

$$ker(\Phi) := \Phi^{-1}(\mathbf{0}_W) = \{\mathbf{v} \in V : \Phi(v) = \mathbf{0}_W\} \tag{35}$$

This is a set of vectors which produces 0 vector in the domain.

Image/range of Φ

$$Im(\Phi) := \Phi(V) = \{\mathbf{w} \in W | \mathbf{v} \in V : \Phi(\mathbf{v}) = w\} \tag{36}$$

Then the rank-nullity theorem says

$$\dim(ker(\Phi)) + \dim(Im\Phi)) = \dim(V) \tag{37}$$

This theorem is also known as fundamental theorem of linear mappings.

6 Eigenvalues and Eigenvectors

In general, multiplication of a matrix transforms a vector into some other space. But, eigenvectors are a special type of vectors associated with that matrix which do not change the direction. In fact, they either stretched or contracted. They are of the form:

$$\boxed{Ax = \lambda x} \tag{38}$$

There are some interesting properties of them:

1. The Eigenvalue for rectangular matrix does not exist, there is no possibility to generate vector of a same dimension by multiplying it with matrix which is not square matrix. Consider this $A \in R^{m\times n}$ and $x \in R^{n\times 1}$, then $Ax \in R^{m\times 1}$. Hence it is not possible to generate vector which is just scalar multiplication.

2. trace(A) $= \sum_i \lambda_i$ i.e. sum of diagonals of A is equal to sum of all the eigenvalues of independent eigenvectors.

3. $|A| = \Pi_i \lambda_i$ i.e. multiplication of eigenvalues is equal to the determinant of the matrix.

4. If $Ax = \lambda x$ then $A^2 x = \lambda Ax = \lambda^2 x$, which means the eigenvalues of A^2 is λ^2 with the same eigenvectors.

5. An important property of the symmetric matrices is that an $n\times n$ symmetric matrix has n linearly independent and orthogonal eigenvectors, and it has n real eigenvalues corresponding to those eigenvectors. It is important to note that these eigenvalues are not necessarily different from each other and some of them can be equal. Another important property of symmetric matrices is that they are orthogonally diagonalizable.

We now want to find these eigenvalues and eigenvectors associated with any given matrix. We can not directly apply the $Ax = b$ formula in equation **??** as it has two unknowns x and λ. Therefore, first we solve for λ and then the value is back substituted to get the eigenvectors.

Equation **??** can be written as $(A - \lambda I)x = 0$. We know from the nullspace knowledge that, if $x = 0$ is a trivial solution which is not so interesting to us. If there has to be x other 0, then the columns of $A - \lambda I$ has to be linearly dependant which means the the matrix $A - \lambda I$ has to be singular. This means rows or columns can be expressed as 0s which in turn means the determinant of this matrix will be 0. This important fact is used to get the values of λs and then back substituted to get eigenvectors.

In the linear algebra literature and software, eigen values are stored in sorted order. The highest eigen values is called as first eigen value and second highest is called second and so on. At the same time associated eigen vectors are also stored in that manner. This is merely a convention and not a rule.

6.1 Eigen decomposition

It is also called as spectral decomposition and can be applied to square matrix A with linearly independent eigenvectors (not necessarily distinct eigenvalues). Suppose A is a matrix with n *independent* eigenvectors $x_1, x_2, \cdots, x_n$ with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$. Suppose S is a eigenvector matrix with eigenvectors as columns.

$$
\begin{aligned}
AS &= A \begin{bmatrix} | & | & \cdots & | \\ | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \\ | & | & \cdots & | \end{bmatrix} \\
&= \begin{bmatrix} | & | & \cdots & | \\ | & | & \cdots & | \\ \lambda_1 x_1 & \lambda_2 x_2 & \cdots & \lambda_n x_n \\ | & | & \cdots & | \\ | & | & \cdots & | \end{bmatrix} \\
&= \begin{bmatrix} | & | & \cdots & | \\ | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \\
AS &= S\Lambda \\
&\boxed{A = S\Lambda S^{-1}} \qquad \text{Another factorization of A} \\
S^{-1}AS &= \Lambda
\end{aligned}
$$

One can also find a left eigenvector such that $v^T A = \lambda v^T$, but we are usually concerned with right eigenvectors.

6.2 Cholesky Decomposition

A positive definite matrix A can be factorized into a product $A = LL^T$ where L is a lower triangular matrix with positive diagonal elements.

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \vdots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix} \tag{39}$$

Here, L is called Cholesky factor of A and L is unique. Obtaining the matrix L is easy as we can just multiply matrices and substitute approapriate values to obtain elements of the matrix L.

7 Singular value decomposition

Many mathematical objects can be understood better by breaking them into constituent parts, finding some properties of them that are universal, not caused by the way we choose to represent them. For example, integers can be decomposed into prime factors. The way we represent the number 12 will change depending on whether we write it in base ten or in binary, but it will always be true that $12 = 2 \times 2 \times 3$. From this representation we can conclude useful properties, such as that 12 is not divisible by 5, or that any integer multiple of 12 will be divisible by 3.

Much as we can discover something about the true nature of an integer by decomposing it into prime factors, we can also decompose matrices in ways that show us information about their functional properties that is not obvious from the representation of the matrix as an array of elements.

We saw in the last section the decomposition of square matrix which has independent eigenvectors. Now, we see decomposition of generic matrix where a matrix can be any $m \times n$ size.

Let $A \in R^{m\times n}$ be a rectangular matrix of rank $r \in [o, min(m, n)]$. The singular value decomposition (SVD) of A is

$$A = U\Sigma V^T \tag{40}$$

with orthogonal matrix $U \in R^{m\times m}$ and an orthogonal matrix $V \in R^{n\times n}$. The matrix $\Sigma \in R^{m\times n}$ where $\Sigma_{ii} = \sigma_i$ and $\Sigma_{ij} = 0$. The Σ matrix contains singular values which is of same size as A that contains diagonal sub-matrix and padding of zeros to make it of size equal to A.

Now, let us see, how we can derive the Eq. **??**. We know that the eigendecomposition of positive definite matrix S can be written as

$$S = S^T = PDP^T \tag{41}$$

Which means if we can somehow form positive definite matrix from the given rectangular matrix then we can apply eigen decomposition to get similar looking expression as SVD, as its eigendecomposition looks in that form.

We know from the spectral therorem that, for any rectangular matrix $A \in R^{m\times n}$ the matrix A^TA is positive semidefinite. Also, we know that the the eigenvectors of a symmetric matrix form an orthonormal basis (this is required as the matrices U, V in SVD are orthonormal), so we can write the following equation

$$A^T A = PDP^T = P \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} P^T \tag{42}$$

Now, let us assume that the SVD exists for A, then we can write

$$A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T \tag{43}$$

Here, we know that U is orthogonal matrix which means $U^T U = I$ and $\Sigma \in R^{m\times n}$ has submatrix containing σ_i values. Therefore, we can write following

$$A^T A = V\Sigma^T \Sigma V^T = V \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} V^T \tag{44}$$

We can see that Eq. **??** is similar to Eq. **??**, which means the values σ are nothing but the eigen values and right-singular matrix V is orthonormal basis formed from the eigenvectors of matrix $A^T A$.

Now, let us obtain left-singular values U. We follow similar steps with slight modification where instead of $A^T A$ we use $AA^T \in R^{m\times m}$:

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^T V\Sigma^T U^T \tag{45}$$

$$= U \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \sigma_m^2 \end{bmatrix} U^T \tag{46}$$

Similar to previous, for AA^T we obtain U from eigenvectors and σ_i from the eigenvalues.

Now, the last part remaining is we combine both the right and left singular values. For the singular value matrices, we need not do anything more as we have obtained them. Only, remaining part is to see how we can deal with the Σ matrix because we get different values and different sizes in Equations **??** and **??**. But, we know that eigenvalues of $A^T A$ and AA^T are same, so we just have to arrange them in matrix Σ.

Eigen decomposition vs SVD For a matrix A, the eigen decomposition $A = PDP^{-1}$ and the SVD $A = U\Sigma V^T$., we compare their properties.

- SVD exists for all $R^{m\times n}$ matrices but eigen decomposition exists only for $R^{n\times n}$ square matrices.

- The vectors in the eigen decomposition matrix P are not necessarily orthogonal i.e. the change of basis is not a simple rotation and scaling. Whereas matrices U, V in SVD are orthogonal, so they do represent rotations.
- In SVD U, V are not inverse to each other necessarily but P and P^{-1} are inverse pair.
- In the SVD, the entries in Σ are all real and non-negative, which is not generally true for the diagonal matrix in eigen decomposition.
- For symmetric matrix $A \in R^{n\times n}$, the eigen decomposition and the SVD are one and the same, which follows from the spectral theorem.

8 Matrix Approximation

In machine learning applications, it is common to find matrices with thousands, hundreds of thousands, and even millions of rows and columns. Although the Eigen decomposition and Singular Value Decomposition make matrix factorization efficient to compute, such large matrices can consume an enormous amount of time and computational resources. One common way to "get around" these issues is to utilize low-rank approximations of the original matrices. By low-rank we mean utilizing a subset of orthogonal vectors instead of the full set of orthogonal vectors, such that we can obtain a "reasonably" good approximation of the original matrix.

There are many well-known and widely use low-approximation procedures in machine learning, like Principal Component Analysis, Factor Analysis, and Latent Semantic analysis, and dimensionality reduction techniques more generally. Low-rank approximations are possible because in most instances, a small subset of vectors contains most of the information in the matrix, which is a way to say the most data points can be computed as linear combinations of a subset of orthogonal vectors.

8.1 Best rank-k approximation with SVD

So far we have represented the SVD as the product of three matrices, U, Σ, and V^T. We can represent this same computation as a the sum of the matching columns of each of these components as:

$$A \coloneqq \sum_{i=1}^{r} \sigma_i u_i u_i^T \tag{47}$$

In other words, the above expression also equals:

$$A \coloneqq \sum_{i=1}^{r} \sigma_i A_i \tag{48}$$

Now, we can approximate A by taking the sum over k values instead of r values. For instance, for a square matrix with r=100 orthogonal vectors, we can compute an approximation with the k=5 orthogonal vectors as:

$$\hat{A} \coloneqq \sum_{i=1}^{k} \sigma_i u_i u_i^T = \sum_{i=1}^{k} \sigma_i A_i \tag{49}$$

In practice, this means that we take k=5 orthogonal vectors from U and V^T, times 5 singular values, which requires considerably less computation and memory than the 100×100 matrix. We call this the best low-rank approximation simply because it takes the 5 largest singular values, which account for most of the information. Nonetheless, we still have a precise way to estimate how good is our estimation, for which we need to compute the norm for $\hat{A}$ and A, and how they differ.

8.2 Best low-rank approximation as a minimization problem

In the previous section, we mentioned we need to compute some norm for $\hat{A}$ and A, and then compare. This can be conceptualized as a error minimization problem, where we search for the smallest distance between A and the low-rank approximation $\hat{A}$. For instance, we can use the Frobenius and compute the distance between $\hat{A}$ and A as:

$$\|A - \hat{A}\|_F \tag{50}$$

Alternatively, we can compute the explained variance for the decomposition, where the highest the variance the better the approximation, ranging from 0 to 1.