Linear Algebra + Stats + Probability Notes

Bhargav *github.com/brpy/ml-notes*

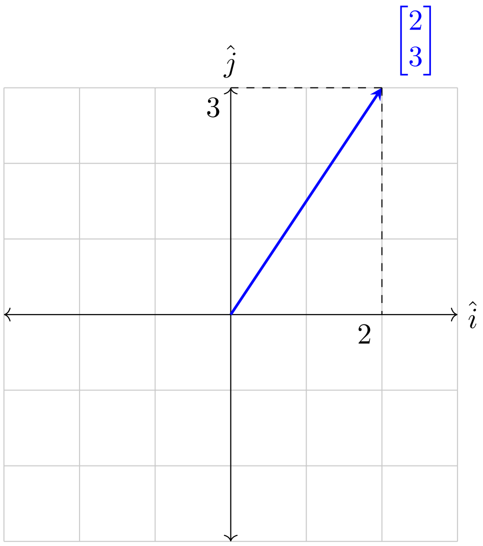
February 28, 2021

1 Linear Algebra

The following notes is primarily made for my revision. I might have skipped some topics that seemed obvious to me. I do not guarantee factual correctness of the notes. If you feel there are any errors, open a github issue/pr. Notes material is collected from various sources. Image credits are given in tex document and in src.csv

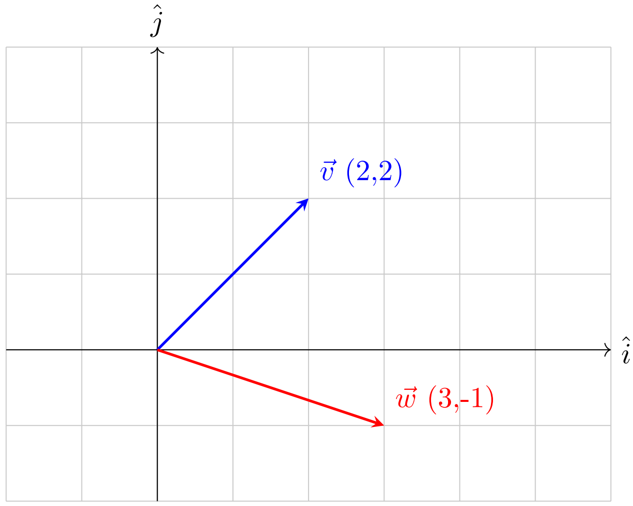
---

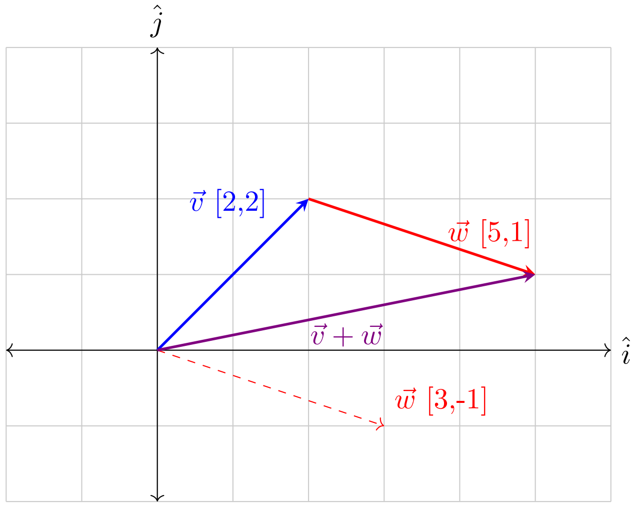
Vectors:



- This vector is represented as $2\hat{i} + 3\hat{j}$ where $\hat{i}$ and $\hat{j}$ are unit vectors perpendicular to each other also known as basis vectors (in 2d).
- It is also represented as a column matrix $\begin{bmatrix}2\\3\end{bmatrix}$
- A vector is represented by its length and direction wrto. basis vectors $\hat{i}$, $\hat{j}$.
- A vector can be freely moved without changing its length and the direction it is pointing to.
- So any vector in space can be represented using a linear combination of $\hat{i}$, $\hat{j}$ by moving the starting point to origin.

Adding 2 Vectors:





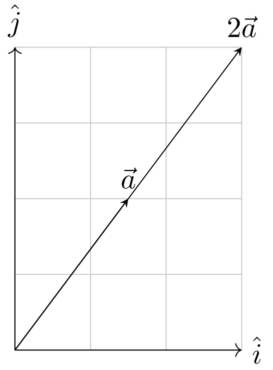
This is also known as Triangular law of vector addition.

- Can also be interpreted as,

$$\begin{bmatrix}5\\1\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}3\\-1\end{bmatrix}$$

- The vector $\vec{w}$ is moved along $\vec{v}$ such that the starting point of $\vec{w}$ meets the end point of $\vec{v}$
- Both $\vec{v}$ and $\vec{w}$ direction and lengths remain unchanged.

Scaling:



- 2 *scales* the vector $\vec{a}$. So, 2 is a Scalar.
- Similarly basis vectors $\hat{i}$ and $\hat{j}$ can be scaled to represent any vector in 2d plane.

Span:

- $a\vec{v} + b\vec{w} \implies$ Linear combination of $\vec{v}$, $\vec{w}$
- Set of all vectors of linear combination of $\vec{v}$, $\vec{w}$; $a\vec{v} + b\vec{w}$ is called **span**.
- For most vectors span consists of all points on the plane.
- If $\vec{v}$, $\vec{w}$ lie on same line, span is a line passin through origin.
- If both $\vec{v}$, $\vec{w}$ are zero, span is zero.

Linear (in)dependent:

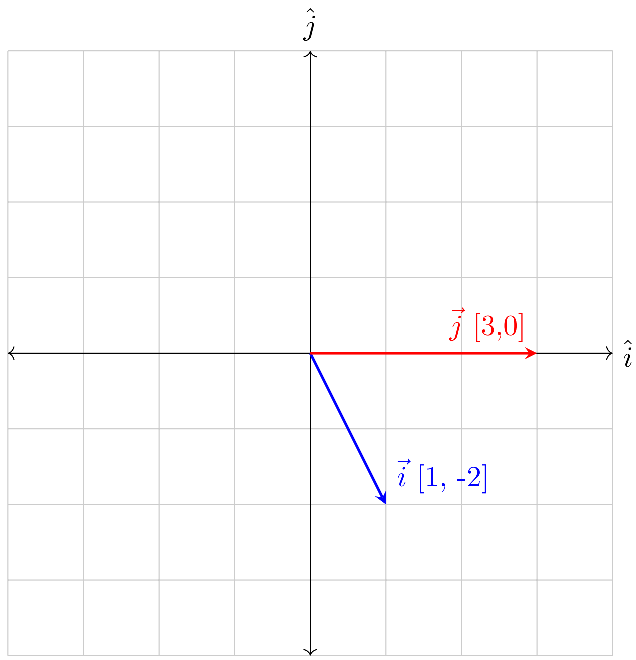
- If one vector can be represented as a linear combination of other, then the vectors are linearly dependent.
- A linearly dependent vector doesn't add to span of a vector.
- If one vector adds a dimension to a span of a vector, they are linearly independent.

Transformation:

- Takes a vector and gives a output vector.

Linear Transformation:
The conditions for the tranformation to be linear are:

- All lines must remain lines.
- Origin must be remain inplace.
- This makes all equidistant parallel lines remain equidistant and parallel.
- Matrices = Transformation of space.
- For example, the linear transform $\begin{bmatrix} 1 & 3 \\ -2 & 0 \end{bmatrix}$ transforms the 2d plane such a way that the new $\hat{i}$ lands at $\begin{bmatrix} 1\hat{i} \\ -2\hat{j} \end{bmatrix}$ and $\hat{j}$ lands at $\begin{bmatrix} 3\hat{i} \\ 0\hat{j} \end{bmatrix}$



The new basis vectors are blue $\vec{i}$ and red $\vec{j}$.

Now in this new transformed space, every vector has to be represented as a linear combination of these two new basis vectors.

The vector $\vec{v}$ = -1$\hat{i}$ + 2$\hat{j}$ after transformation lands at the point $\begin{bmatrix} 5 \\ 2 \end{bmatrix}$ in the new vector space.

Even after the transformation, the linear combination doesn't change. So, $\hat{i}$ and $\hat{j}$ are replaced by $\begin{bmatrix} 1\hat{i} \\ -2\hat{j} \end{bmatrix}$ and $\begin{bmatrix} 3\hat{i} \\ 0\hat{j} \end{bmatrix}$ respectively.

So the new transformed $\vec{v}$ becomes,

$$\vec{v}_{new} = -1\hat{i}_{new} + 2\hat{j}_{new} \tag{1}$$

$$\vec{v}_{new} = -1 \begin{bmatrix} 1\hat{i} \\ \text{-}2\hat{j} \end{bmatrix} + 2 \begin{bmatrix} 3\hat{i} \\ 0\hat{j} \end{bmatrix} \tag{2}$$

$$\vec{v}_{new} = \begin{bmatrix} \text{-}1\hat{i} \\ 2\hat{j} \end{bmatrix} + \begin{bmatrix} 6\hat{i} \\ 0\hat{j} \end{bmatrix} \tag{3}$$

$$\vec{v}_{new} = \begin{bmatrix} 5\hat{i} \\ 2\hat{j} \end{bmatrix} \tag{4}$$

For any vector $x\hat{i} + y\hat{j}$, after applying the transformation $\begin{bmatrix} 1 & 3 \\ \text{-}2 & 0 \end{bmatrix}$ becomes,

$$\begin{bmatrix} x\hat{i} \\ y\hat{j} \end{bmatrix} \overset{lr.\ transform}{\Longrightarrow} x \begin{bmatrix} 1\hat{i} \\ \text{-}2\hat{j} \end{bmatrix} + y \begin{bmatrix} 3\hat{i} \\ 0\hat{j} \end{bmatrix} = \begin{bmatrix} 1x + 3y\hat{i} \\ -2x + 0y\hat{j} \end{bmatrix} \tag{5}$$

Removing $\hat{i}$, $\hat{j}$ for legibility,

$$\begin{bmatrix} x \\ y \end{bmatrix} \overset{lr.\ transform}{\Longrightarrow} x \begin{bmatrix} 1 \\ \text{-}2 \end{bmatrix} + y \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1x + 3y \\ -2x + 0y \end{bmatrix} \tag{6}$$

This is the basis for Matrix multiplication.

- $\begin{bmatrix} 1 \\ \text{-}2 \end{bmatrix}$ is where $\hat{i}$ lands and $\begin{bmatrix} 3 \\ 0 \end{bmatrix}$ is where $\hat{j}$ (basis vectors) lands after the transformation.
- So the 2X2 matrix $\begin{bmatrix} 1 & 3 \\ \text{-}2 & 0 \end{bmatrix}$ itself can represent the transformation.
- This explains the *rules* for multiplication and why matrix multiplication is not commutative.
- This also explains why the matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is an identity matrix. Since, this *transform* actually does nothing, $\hat{i}$ and $\hat{j}$ remain unchanged.

For any transformation $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$,

$$\underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{Transformation} \overbrace{\begin{bmatrix} x \\ y \end{bmatrix}}^{old\ vector} = x \overbrace{\begin{bmatrix} a \\ c \end{bmatrix}}^{where\ \hat{i}\ lands} + y \underbrace{\begin{bmatrix} b \\ d \end{bmatrix}}_{where\ \hat{j}\ lands} = \overbrace{\begin{bmatrix} ax + by \\ cx + by \end{bmatrix}}^{Transformed\ matrix} \tag{7}$$

- $\begin{bmatrix} a \\ c \end{bmatrix}$ and $\begin{bmatrix} b \\ d \end{bmatrix}$ are the new basis vectors; where old $\hat{i}$ and $\hat{j}$ land. These are the new transformed basis vectors.
- Now these new basis vectors have to be used to represent all the vectors in it's span. In other words the linear combination of these two basis vectors.

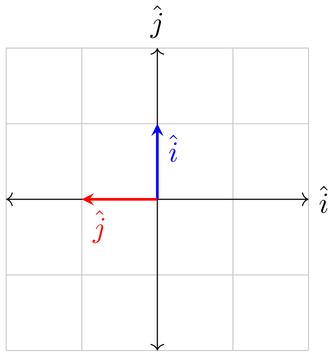
- From equation 7 it can be observed that the scalars x and y scale the corresponding new basis vectors.
- For no transformation (or) Identity transform / Multiplication,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \text{x} \\ \text{y} \end{bmatrix} = \begin{bmatrix} \text{x} \\ \text{y} \end{bmatrix} \tag{8}$$

Which makes sense !!

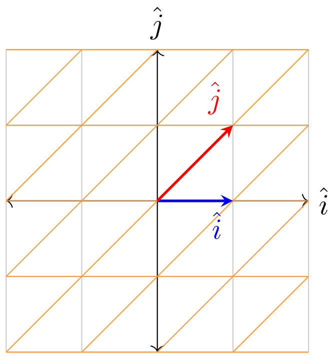
Counterclock Transform:

$$\begin{bmatrix} 0 & \text{-}1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \text{x} \\ \text{y} \end{bmatrix}$$



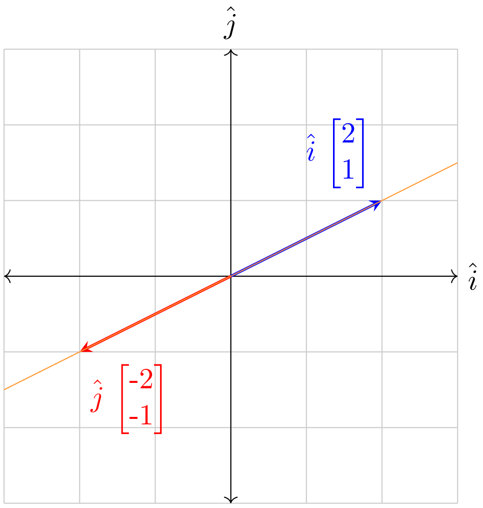
Shear Transform:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \text{x} \\ \text{y} \end{bmatrix}$$



Few notable points:

- In transformations, order matters. $M_1M_2 \neq M_2M_1$ because $f(g(x)) \neq g(f(x))$
- The associative property $(AB)C = A(BC)$ holds true because the order is C, B then A regardless.
- For a $\begin{bmatrix} 2 & \text{-}2 \\ 1 & \text{-}1 \end{bmatrix}$ transformation all 2d space is squished into a line.



These two vectors $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -2 \\ -1 \end{bmatrix}$ are linearly dependent vectors.

- Two transforms for ex. shear and rotation can be composed into a single transform. This composition transform is nothing but a multiplication of shear and rotation matrix.

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{Shear} \overbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}^{Rotation} \begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{Compostion} \begin{bmatrix} x \\ y \end{bmatrix} \tag{9}$$

- Product or composition of two matrices/ transforms,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix} \tag{10}$$

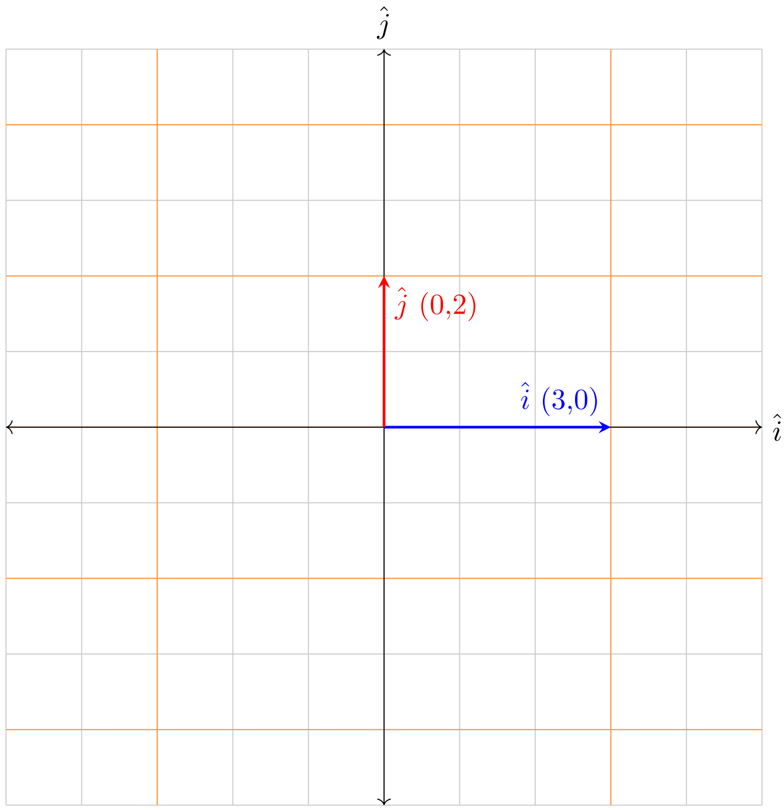
Determinant:

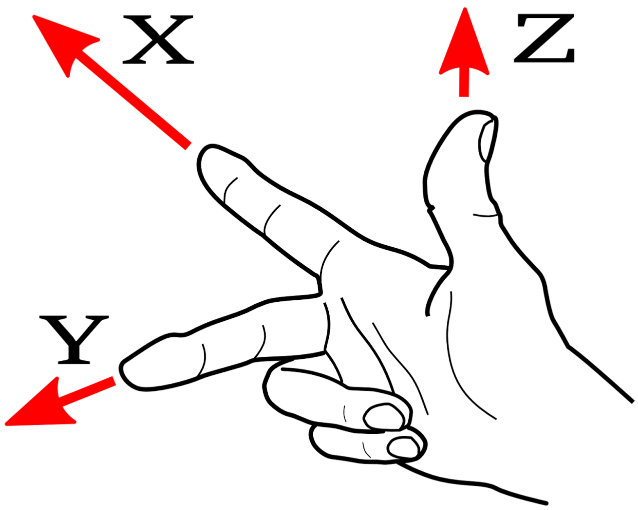
- Transformation $\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$ scales $\hat{i}$ by factor 3 $\hat{j}$ by a factor 2.

- We observe that the unit square in transformed space is scaled from 1 sq. unit to 6 sq. units. So the determinant of the transform/matrix $\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$ is 6. In case of 3d, volume is scaled.

- For a matrix/transform $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the determinat is $ad - bc$.

- The determinant can be $-ve$ if the axis cross each other during the transformation. It has the effect of *flipping* or *inverting* the space.





- This can be checked using right hand rule to check if the axis still lie in the same orientation after transformation. If not, the determinant is $-ve$.

- For 3d space, determinant is measured by unit volume instead of area.

- Determinant can be zero if the space is squished. For ex. if transformed to a line or point in 2d or to a plane, line or a point in 3d. This has the effect of reduction in number of dimensions.

$$det(M_1 M_2) = det(M_1)det(M_2) \tag{11}$$

System of equations:

$A.\vec{X} = \vec{V}$

$3x + 1y + 4z = 1$
$3x + 9y + 2z = 6$
$3x + 3y + 3z = 8$

- Solving these system of equations imply, finding the vector $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ that if applied the transformation $\begin{bmatrix} 3 & 1 & 4 \\ 3 & 9 & 2 \\ 3 & 3 & 3 \end{bmatrix}$, would land on the vector $\begin{bmatrix} 1 \\ 6 \\ 8 \end{bmatrix}$.
- This can only be solved if A^{-1} exists. Since space cannot be unpacked since there is/are lost dimension(s) if $det(A) = 0$. So $det(A) \neq 0$ has to be true for the system of equations to be solved using,

$$\vec{X} = A^{-1}\vec{V} \tag{12}$$

Rank:

- Rank is the number of dimensions of the transformed space.
- The Rank of a matrix ; output space.

  Rank : 1 $\implies$ output transformation : Line

  Rank : 2 $\implies$ output transformation : Plane

  Rank : 0 $\implies$ output transformation : Point

Column space:

- Set of all possible linear combinations or span of column vectors of A. Or,
- Set of all possible $A\vec{V}$

Null space:

- Set of vectors that get squished into origin after transformation.

Dot Product:

- $\vec{a}$ and $\vec{b}$ are two vectors and angle between them is θ.

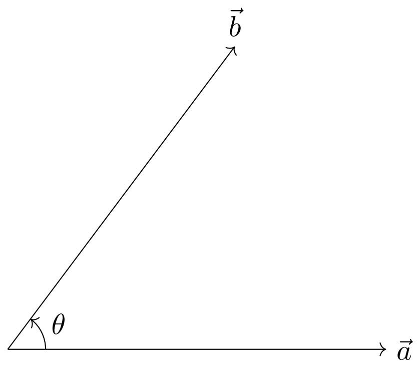
$$\vec{a} = [a_1, a_2, ...a_n] \tag{13}$$

$$\vec{b} = [b_1, b_2, ...b_n] \tag{14}$$

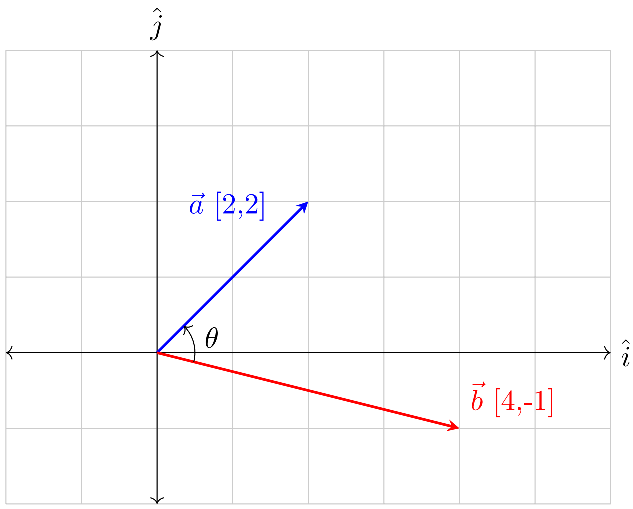
$$\vec{a}.\vec{b} = (a_1)*(b_1)+(a_2)*(b_2)+...+(a_n)*(b_n) \tag{15}$$

- Geometric representation:

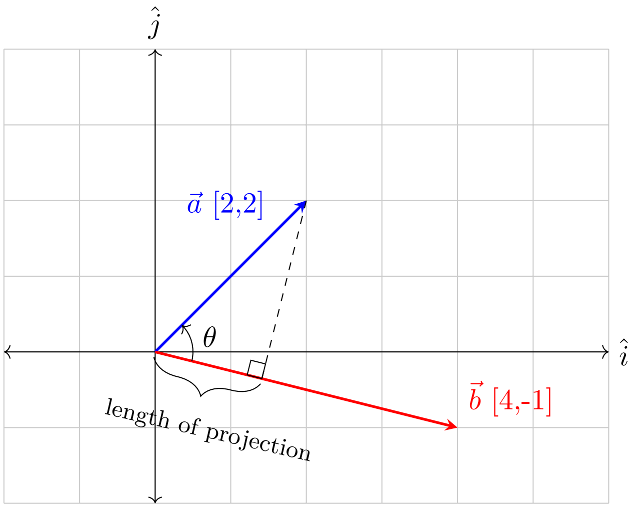
$$\vec{a}.\vec{b} = ||a|| * ||b|| * cos\theta \tag{16}$$



dot product of two vectors = (length of any vector) x (length of projection made on the vector)



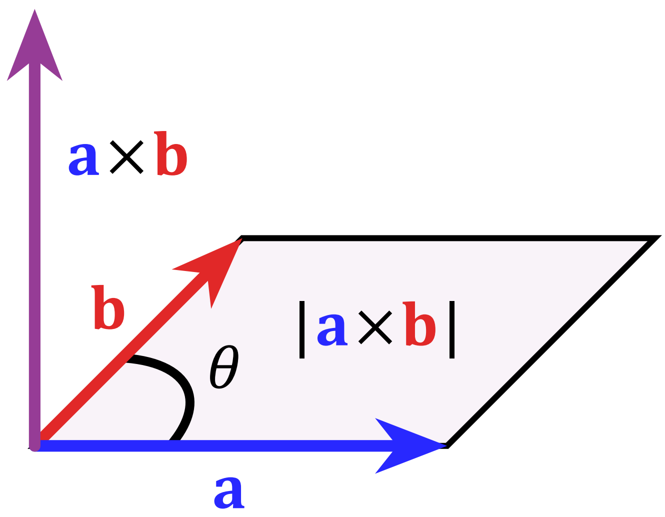
Projecting $\vec{a}$ on $\vec{b}$:



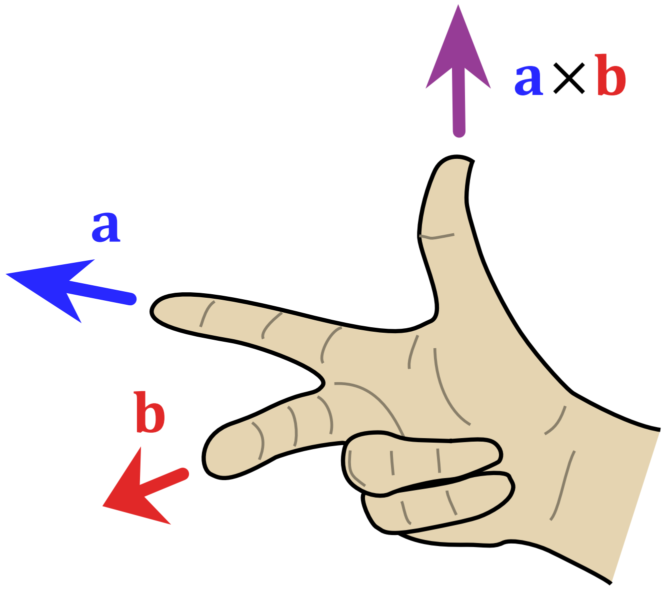
- Length of projection $= ||a||cos\theta$
- Dot product $= ||b||.||a||cos\theta$
- The length of projection does not depend on the length of $\vec{b}$. It only depends on length of $\vec{a}$ and θ.
- Projecting $\vec{b}$ on $\vec{a}$ will result in the same dot product result. The order does not matter.
- If projection does not lie betweeon origin and the end point of the other vector, you can extend the other vector since length of projection is not affected by it.
- Dot product is -ve if projection is on -ve side.

Cross Product:

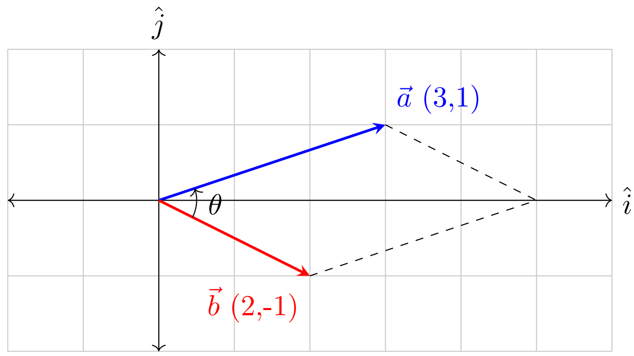
- Directed area product.



- $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$



- Length of $\vec{a} \times \vec{b} = det(\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix})$
- The cross product vector is normal to the area.



- The area of this parallelogram is $det(\begin{bmatrix} 3 & 2 \\ 1 & -1 \end{bmatrix})$
- The direction of the cross product is normal to the area, i.e, in the direction of $\hat{k}$
- The magnitude of this vector is the area of this parallelogram i.e, determinant.

$$\vec{a} \times \vec{b} = ||a||.||b||sin\theta \tag{17}$$

Eigen:

- For a few transformations, vectors only scale. i.e, only stretch or compress.
- Such vectors are called eigen vectors of the transformation.
- Such vectors are transformed into it's own span.

- The scaled value of such vectors (scalar) is the eigen value. It can be +ve or -ve.

$$A\vec{V} = \lambda\vec{V} \tag{18}$$

where:

A is the transformation,

$\vec{V}$ is the eigen vector and,

λ is the eigen value.

Vector from origin:

- As discussed previously, any vector can be moved to origin for reference, without changing it's direction it is pointing to and it's length.

- To make a vector starting from $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and ending at $\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ as a reference vector, we can apply triangular law of addition to get a vector which starts from origin and is equivalent with the vector starting and ending at the above two points.

- We need to remember that the above vectors themselves start at origin. So, we can get the vector by doing $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$

Matrices:

Diagonal Matrix:

- A matrix with all 0's except diagonal elements.

Identity Matrix:

- A Matrix with all diagonal elements as 1.

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Scalar Matrix:

- A Matrix with only equal diagonal elements.

- A scalar matrix can be represented by a single scalar.

$$K.I = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

Upper Triangular Matrix:

- A Matrix with all elements below diagonal are 0

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 2 & 7 \\ 0 & 0 & 1 \end{bmatrix}$$

Lower Triangular Matrix:

- A Matrix with all elements above diagonal are 0

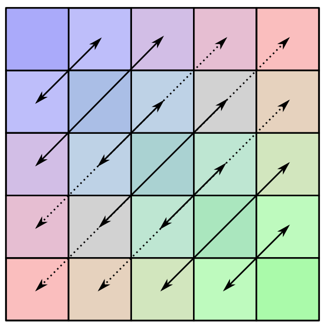
$$\begin{bmatrix} 3 & 0 & 0 \\ 7 & 2 & 0 \\ 5 & 3 & 1 \end{bmatrix}$$

Symmetric Matrix:

- A square matrix which is equal to its transpose.

- $A = A^T$

- $a_{ij} = a_{ji}$ for every i and j.



Skew Symmetric Matrix:

- A square matrix which is equal to negative of its transpose.

- $A = -A^T$

- $a_{ij} = -a_{ji}$ for every i and j.

Orthogonal Matrix:

- A square matrix whose column vectors and row vectors are orthonormal.
- $AA^T = A^T A = I$ or,
- $A^T = A^{-1}$
- $a_{ij} = -a_{ji}$ for every i and j.

Lines, Planes and Hyperplanes:

- **Line** - A line has 2 degrees of freedom. You need a point and a slope or two points to uniquely identify a line.
- Equation of a line
  1. $y = mx + c$; slope = m ; y - intercept = c ; passes through $(0, c)$
  2. $ax + by + c = 0$ is a simple linear form with 2 unknowns
  3. $y - y_1 = m(x - x_1) + c$; slope = m ; passes through x_1 and y_1
- Slope is the *tan* of angle x axis makes with the line in anti-clock wise direction.
- Any 2 points on a line can uniquely identify the line.
- An object on a line can move in 1 dimension.

- **Plane** - A plane is a flat 2d surface.
- A xy plane is an example of a plane.
- To represent a plane you need 3 non collinear points or a point and a normal.
- A normal is a vector that is perperndicular to all vectors on the surface.
  1. So, by definition a plane is set all (x, y, z) points that satisfy,

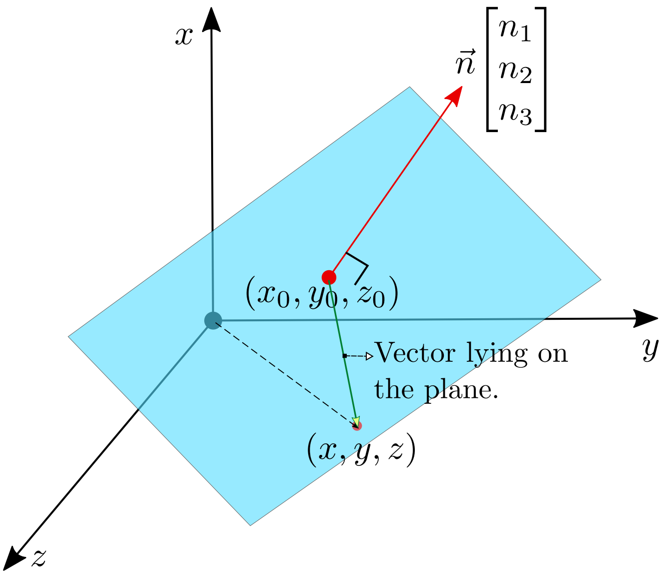
$$dot(\begin{bmatrix} x \\ y \\ z \end{bmatrix} - \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}, \vec{n}) = 0 \qquad (19)$$

Where:

$\vec{n} = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}$ is the normal to the plane and starts at point $\begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}$ which is located on the plane.

- The equation subsequently solves to,

$$n_1(x - x_0) + n_2(y - y_0) + n_3(z - z_0) = 0 \quad (20)$$



- **Hyperplane** - Is an extension to Line and a plane in higher dimensions.

- A **Linear regression** model can successfully represent a line, plane or a hyperplane.

Miscellaneous:

- Distance between two points (x_1, y_1) and (x_2, y_2) is

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Length of a vector $a\hat{i} + b\hat{j} + c\hat{k}$ is

$$\sqrt{a^2 + b^2 + c^2}$$

also known as L2 Norm. (Check knn notes for info on norms.)

- Distance from a plane $ax + by + cz = 0$ to a point (x_1, y_1) is

$$\frac{|ax_1 + by_1 + cz_1|}{\sqrt{a^2 + b^2 + c^2}}$$

Credits:

- Most of this notes is taken from 3Blue1Brown's Essence of Linear Algebra playlist. Some images are taken from wikipedia.

2 Probability

Population and Sample:

- (Statistical) population is a set of all the items or events which is of interest for a question or experiment.
- (Statistical) sample is a subset of population. The aim of sampling is that our sample represents the population.
  The advantage is that it is faster and cheaper to collect data than for the entire population.

Figure 1: *A random sampling process*

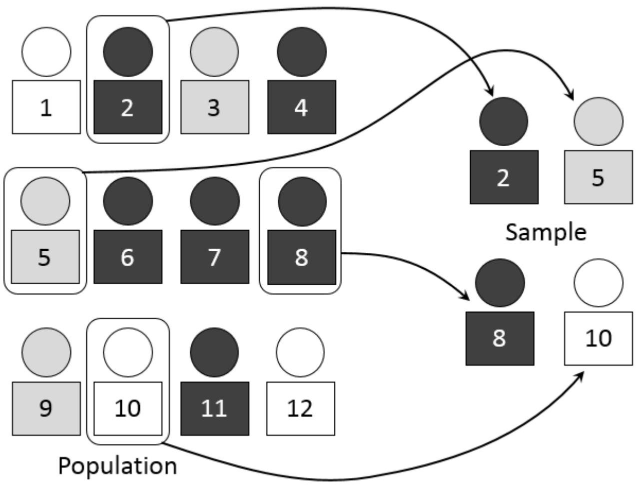
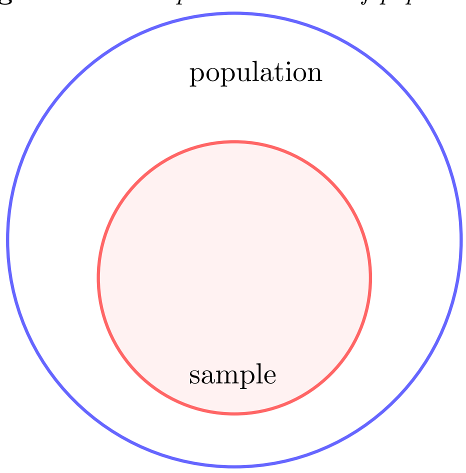


Figure 2: *A sample is a subset of population*



Random variable: (X)

- A Random variable is a measurable function defined on probability space that maps from the sample space to real numbers.
- It is called a **Discrete Random variable** if the range of the function is countable.
- It is called a **Continous Random variable** if the range of the function is uncountably infinite or an interval.
- Both these types of Random variable has a distribution.
  - If X is a **Discrete Random variable**, it's distribution is a discrete probability distribution i.e. can be described by a probability mass function (pmf) that assigns a probability to each value in the range of X.
  - If X is a **Continous Random variable**, and *absolutely continuous*, its distribution can be described by a probability density function (pdf).
    Pdf assigns probabilities to an interval and since the range is absolutely continuous, each individual point must have a probability of 0.
- Not all continuous random variables are absolutely continuous.
  A mixture distribution is one such counterexample; such random variables cannot be described by a probability density or a probability mass function.
- But all Random variable can be described by a Cumulative distribution function (cdf).
- Ex. for a discrete rv is dice rolls and for continuous rv is height or incomes of people.

Probability Distribution:

- A Probability Distribution is the mathematical function that gives probabilities of occurences of different possible outcomes for an experiment.
- It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

Probability Density Function (pdf):

- Pdf is a function used to specify the random variable falling within a particular range of values, as opposed to taking any particular value(which is 0, since there are infinite number of values).
- Probablity is the integral or area under the curve over an interval.
- It's value at any given point can be interpreted as providing the relative likelyhood that the value of the random variable would equal that sample.
- In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there is an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample.

Properties of pdf:

- Area under the curve or AUC(pdf) = 1
- Value at any point of pdf curve has to be ≥ 0
- x-axis : Random variable X
  y-axis : Relative likelyhood/ probability density.
- For a continuous rv, it's pdf has to also be continuous.
- A pdf doesn't have to be differentiable. For ex. Laplace distribution is not differentiable.

$$\frac{1}{2}e^{-|x|}$$

- But a CDF has to be differentiable.
- A point on pdf provides relative likelyhood. Whereas, area under an interval gives probability.
- Few random variables may not have pdf. They are called Singular distributions.

Gaussian/ Normal Distribution: (N)

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2} \tag{21}$$

where:
μ - Mean or expectation
σ - Standard deviation
σ^2 - Variance

- The significance of Normal distribution comes from various factors like,
  - Central limit theorem
  - Many natural properties follow a roughly Normal distribution
  - For a fixed mean and std. deviation, Normal disitribution has maximum entropy.
- Though Normal distribution is commonly referred to as a bell curve, there are also other distributions in the bell curve family.

Figure 3: *pdf of a normal distribution*

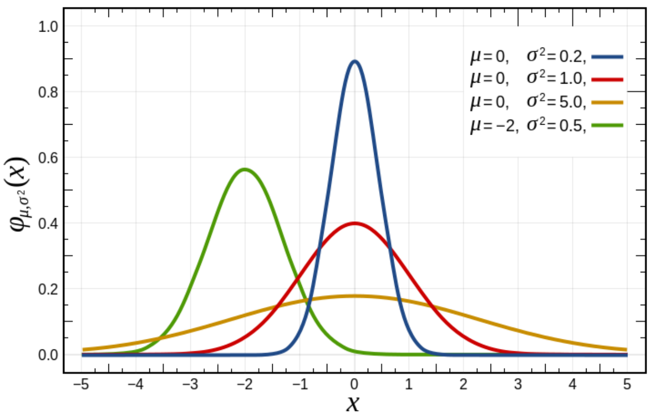
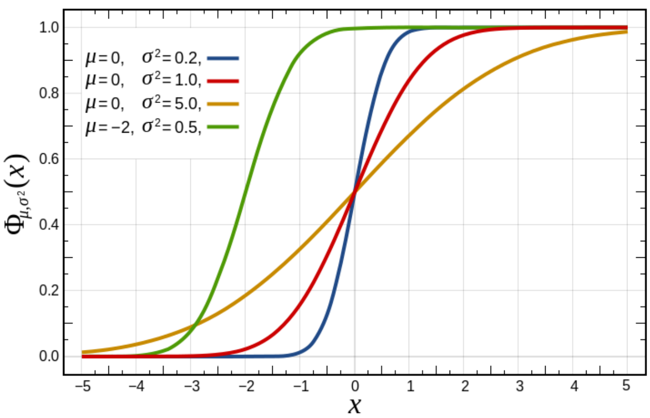


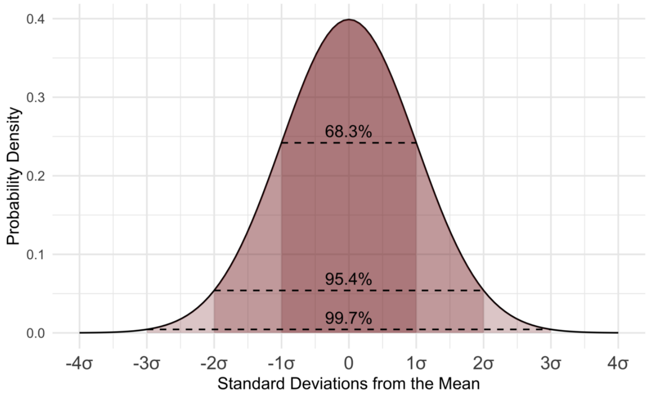
Figure 4: *cdf of a normal distribution*



68-95-99.7 Rule:

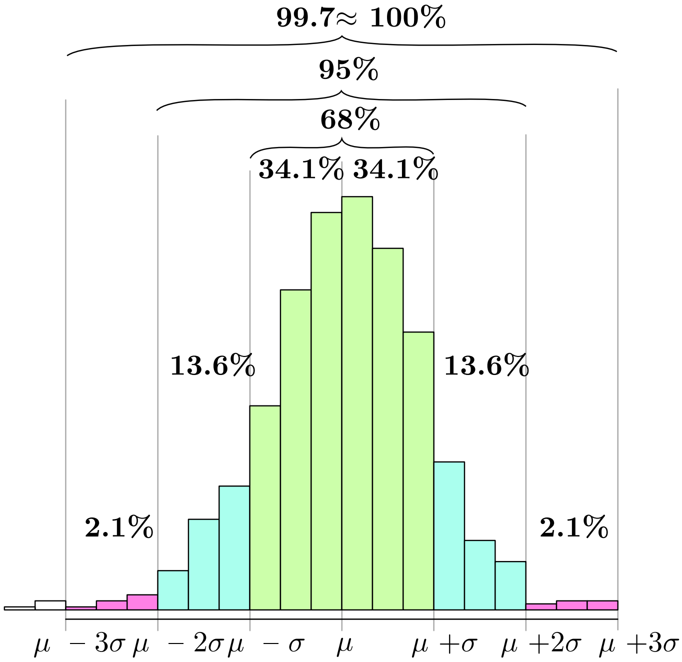
- It is an empirical rule that a percentage of values lie within a band around the mean for a normal distribution.

Figure 5: *pdf of a standard normal distribution*



- The percentage of values are,
  - 68% of values lie within $\mu - 1\sigma$ and $\mu + 1\sigma$
  - 95% of values lie within $\mu - 2\sigma$ and $\mu + 2\sigma$
  - 99.7% of values lie within $\mu - 3\sigma$ and $\mu + 3\sigma$
- Is valid for any value of μ and σ

Figure 6: *pdf of a normal distribution with mean μ and std-dev. σ*



Cumulative distribution function: (CDF)

- The probability that rv X will take a value less than x. $P(X < x)$.
- Fig 4 on page 10, represents cdf of some normal distributions.
- CDF of a distribution has to differentiable and non decreasing.
- x-axis : Random variable X
  y-axis : $P(X < x)$
- Slope at a point on cdf gives pdf.

$$f(x) = \frac{dF(x)}{dx} \tag{22}$$

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{23}$$

where:

$F(x)$ is cdf.

$f(x)$ is pdf.

- All random variable have cdf.
- On observing CDFs it is easy to get an idea on percentiles.

Probability mass function: (PMF)

- Probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.
- It is associated with discrete random variable.
- Unlike pdf's there's no need to integrate to get probabilities.
- Some examples are Bernoulli distribution and Binomial distribution.
- Sum of all probabilities has to equal 1, and no values can be negative.

Figure 7: *pmf of a fair dice*

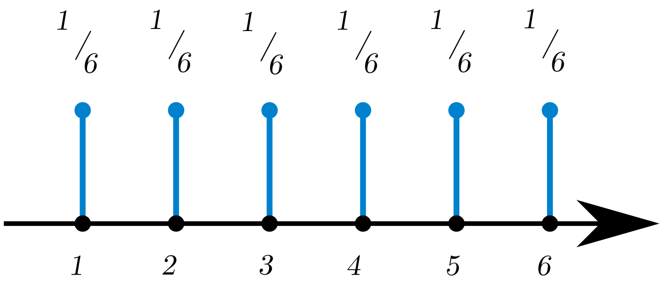
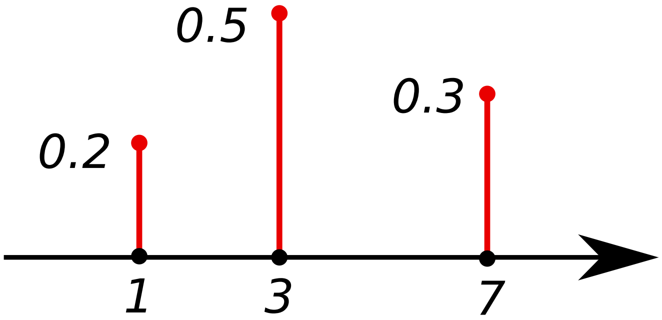


Figure 8: *A discrete probaility distribution function*



Mean, Median and Mode:

- Mean (μ) :
$$\frac{1}{n}\sum_{i=1}^{n} x_i$$
where:
μ - mean of the population
n - size of the population
x_i - value from population
- Median :
Central most observation after sorting. If two values are in middle, median is the average of the two.
- Mode: Most frequent observation.
- These are the measures of central tendency.

Variance: