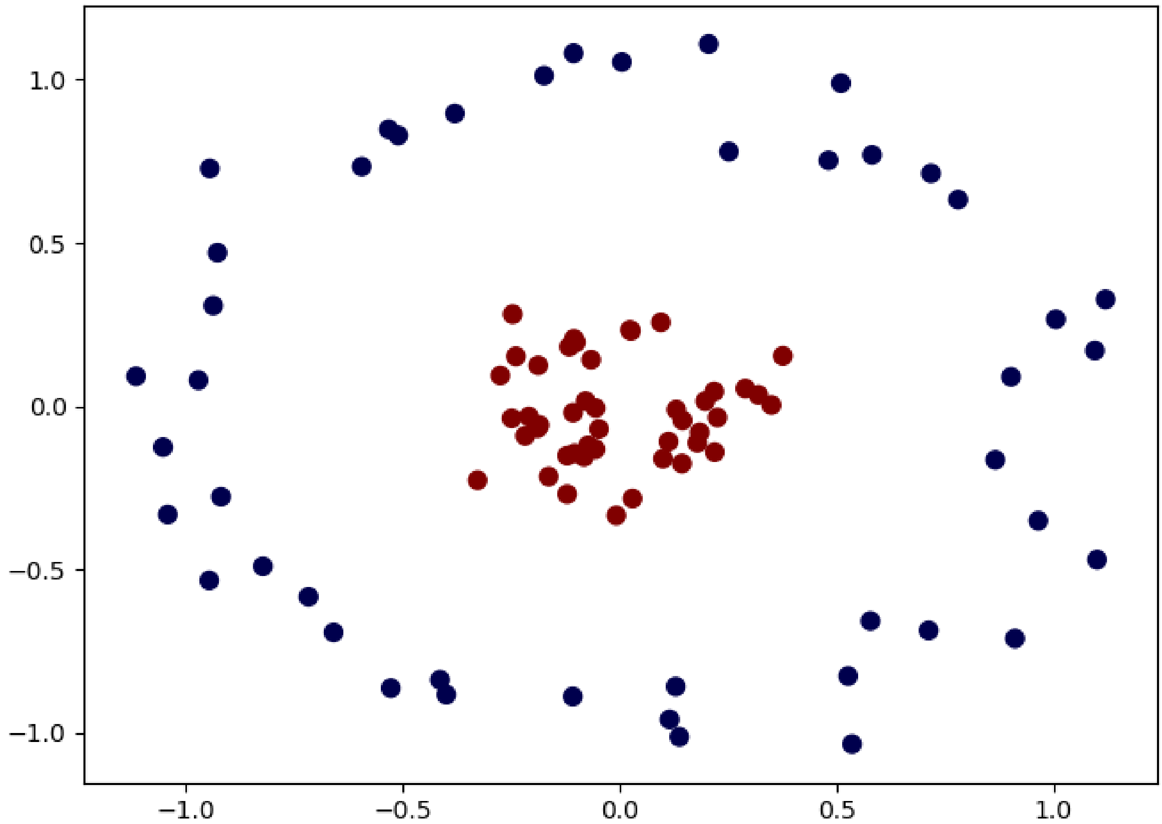
Reproducing kernel Hilbert space

Introduction

The aim of this report is to provide a summary of Reproducing Kernel Hilbert Spaces (RKHS) with its fundamental features. In functional analysis, RKHS is a Hilbert space of functions in which the evaluation of a function at any given point is a continuous linear functional. The theory of RKHS is used in various areas, including machine learning, operator algebra, approximation theory, and quantum mechanics. An RKHS is associated with a reproducing kernel which was first introduced by Stanisław Zaremba in 1907 with an aim to reproduce every function in a space so that for any value on which the function has defined the evaluation on any point can be taken with the help of an inner product. The framework provided in this report is for real-valued Hilbert spaces though the theory can be extended to complex-valued functional spaces as well.

Let's start with a simple example :

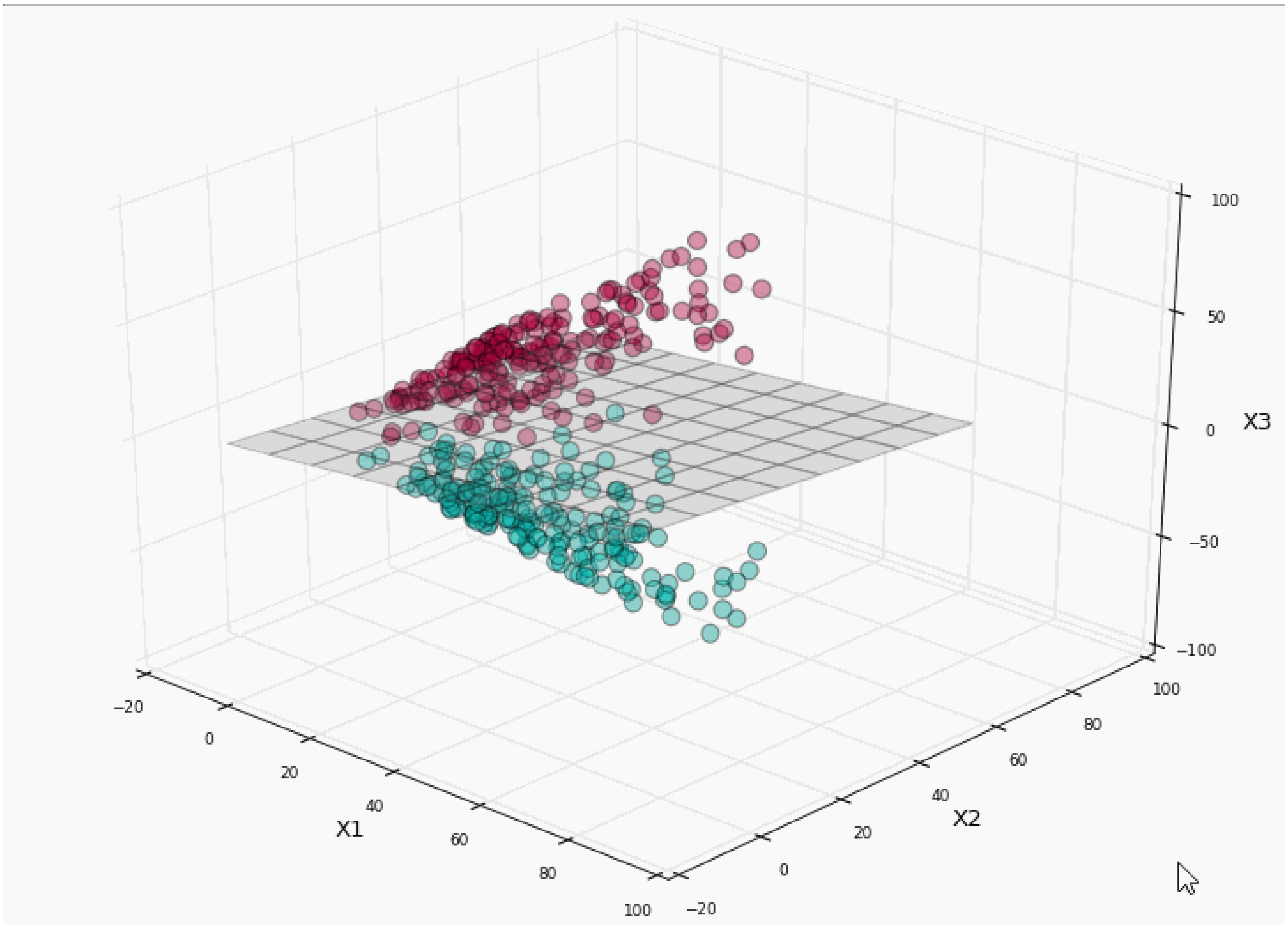
What if we are given this dataset with a task to classify these two classes of red and blue dots.



What can be done here?

As, it can be seen that the given data is in 2-d what if we can take it to some higher dimensions.

For an instance from 2-d to 3-d.

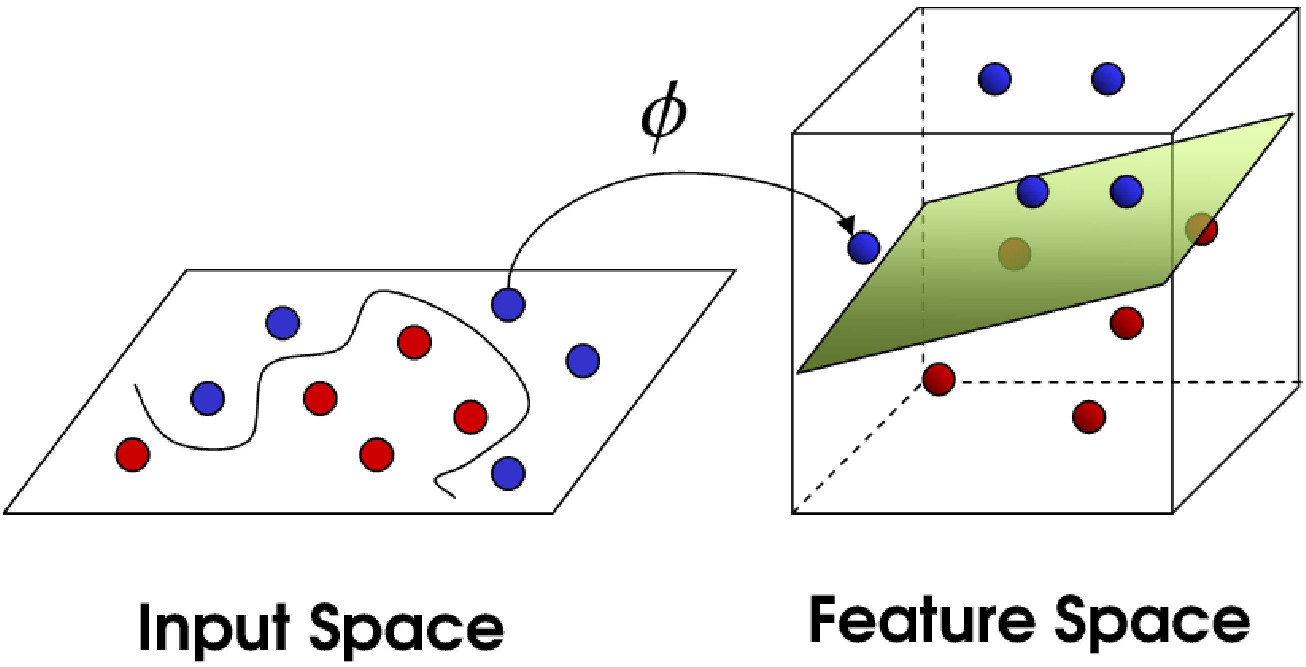


It seems like a fair solution, but the only problem is **how to do it**?

what we did here is just mapped data from 2-d to 3-d space.

So, the question is what does **mapping** mean and how to do it?

Suppose we have a number, and we pass it through a function to get some output. This gives us a new set of numbers, now instead of using our original set of numbers we use this new set with higher dimensions for performing the respective task.



Now the question arises that from **where to get this Φ ?**

Assumption : Output is a real number

H = Space of hypothesis: To start with let's say, it is a subset space of all possible functions from input to output.

$$H \subset \{f | f : x \to \mathbb{R}\}$$

what are the properties, we are looking for, to define this hypothesis space?

> 1. It must be simple (too complex can lead to overfitting). But not too simple, as it might not overfit the data but will give bad results.

```
Example: constant functions, extremely stable but not flexible enough.
```

> 2. It should be simple but flexible.

> 3. Computationally efficient.

One simple reasonable exampe of this is : **Linear Functions**

$$H = \{f : x \to \mathbb{R} | 3\ w\varepsilon\mathbb{R}^D, f(x) = w^T x, \forall x\varepsilon\mathbb{R}^d\}$$

- It is a continuous collection of inputs, each function here is parametrized by W.

- For a given function there is only one w and for a given w there is only one function.

- Inner product: As linear space means we can add functions or multiple them by numbers. We can find the norm by finding the inner product of associated vectors.

$$d(f, \bar{f}) = \|w - \bar{w}\|$$

- It states that if we have 2 functions, defined by 2 vectors, and if these vectors are in the vicinity then the function is also going to be vicinity at every possible value.

- So, the good thing here is that functions become vectors and we know how to deal with vectors i.e taking projection, norms, etc.

So, if our hypothesis space contains a linear set of functions, data that is linearly separable can easily be classified.

What if data is not linearly separable, as with linear function at most we can only create a bunch of straight lines.

So, what we are lacking here is flexibility.

We need our hypothesis space with some more flexible functions.

To solve this, let's understand the concept of a finite dictionary:

Finite Dictionary: These are the set of functions. Every element inside is called **atoms or features**.

$$D = \{\phi, x \to \mathbb{R} | i = 1,2 \ldots P\} p \in \mathbb{N}$$

Φ is just a function from the input to the output, and let's say we got p of them.

p $\geqslant$ d (dimension of input space)

Let's take a function from this space now , By taking the linear combination.

$$H = \left\{ f : x \to \mathbb{R} | 3w\varepsilon\mathbb{R}^P, f = \sum_{j=1}^{P} w_j \phi_j \right\}$$

For instance, if your data (x) is a set that contains the graphs then D can be a dictionary of 10 functions which are amazing for working with graphs.

Properties of this space of functions:

> - It is a linear space which means we can associate each function with a vector.

> - We can define the inner product.

> - So, now we have a dictionary of linearly dependent features.

> - What we have until now is actually good, we can perform operations like computing gradients in case we want to minimize the function. Moreover, as we can define the inner product which give us the notion of orthogonality.

Q. What do we gain here ?

A. Now, we can build non-linear features and for doing that Φ doesn't have to be linear. As the function, itself is linear in parameters, but with respect to input, it is non-linear because we are seeing the input through Φ.

Example: Φ could be Sine, Cosine, exponential, or any polynomial function. So, by taking linear combination we can build non-linear functions.

So far now we are at a good place, we have a parametrization that is linear.

Q. What are we still missing ?

> We would like to kill the assumption that we have finite p. we have to find something which can allow us to work in a larger space. what if we can have infinite-dimensional H?

From where we will get Φ ?

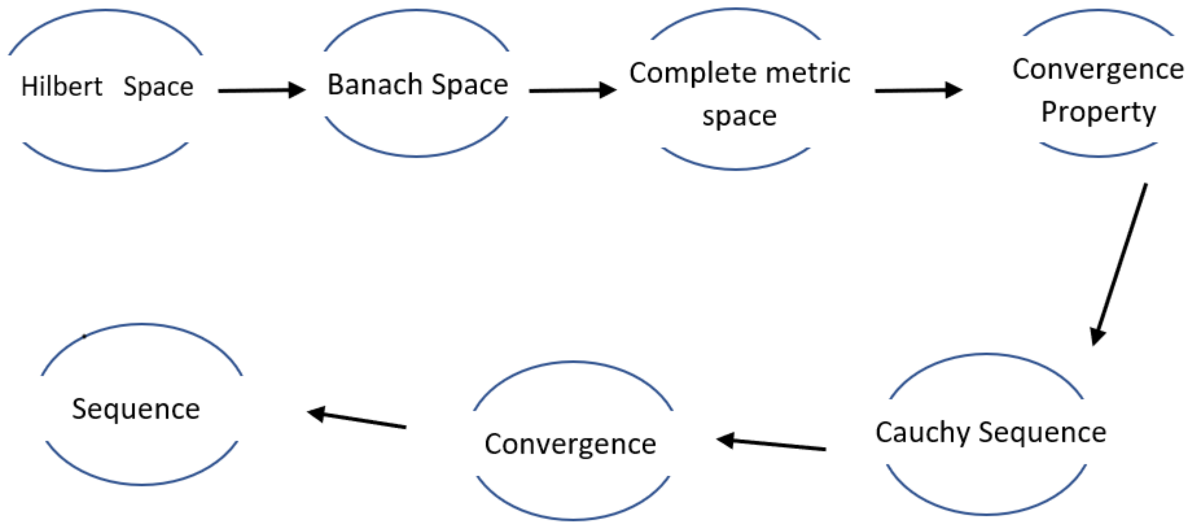
if we choose polynomial we have to stick with polynomial, if we choose sine or cosine we have to stick with that. so, from where we will get Φ. it is important to get the right Φ because it holds all the information about the function we want to estimate.

What if we can invent or reproduce this Φ depending on the data.

Seems like there is a mathematical concept called **Hilbert Space** that can solve our problem.

If we have a linear space with the inner product and we want to generalize it to infinite dimension, we have some technical conditions to fulfill like we want series of functions to converge and then we are good to go. The name we give to the space with these properties **(Linear space + inner product + completeness)** is Hilbert space. Let's say this is a paragraph **X** for future references

Moving away from our original problem, let's first understand what a **Hilbert Space is ?**



For understanding Hilbert Spaces let's start it from scratch, so the first thing to understand is that what is a sequence?

Sequence: is just a ordered set inside the metric space X.if a_1 , a_2 , a_3 ...etc. denote the terms of the sequence then 1,2,3,4.. denotes the position of those terms.

Convergence: A sequence (X_n) is a metric space (X,d) is called **convergent** if there is a limit point we call $\bar{x}$

What this means is that if we take a ball of whatever arbitrary size, we will always find some point in the sequence say N after which, all the points will fall within that ball. So, the sequence remains within that ball.

Let's take **example of a series that does not converge.**

X = (0,1]

$$s: N \rightarrow x$$

$$s: n \rightarrow \frac{1}{n}$$

$$s_1 = \frac{1}{1}, s_2 = \frac{1}{2}, s_3 = \frac{1}{3} \ldots . Sn = \frac{1}{n}$$

Considering 0 is not the element of set X, where will this series converge?

In the case of real line this series converges at zero, but zero is not an element of a given metric space. So, there is no element in a given metric space at which the series converges.

Cauchy Sequence

If the terms of the given sequence become very close to each other when it progresses we call it a Cauchy sequence. it means after certain points all the elements are so close that they seem at the same point.

Banach Space : To understand the Banach space, it's necessary to understand the concept of norm first.

A norm is a mathematical object which describes the quantity in some sense(possible abstract), it tells about the size, length, or extent of the given object(vector, matrice). Normally single bars |x| is used to denote the vector or quaternion norms while ||x|| double bar is used for matrix norms.

Properties of a norm :

> 1. Positive definite : ||x|| ≥ 0 for all x ∈ R^d and also ||x|| = 0 iff x = 0

> 2. Absolutely Homogenious : ||αx|| = |α| · ||x|| for all α ∈ R, x ∈ R^d

> 3. Triangle inequality : ||x + y|| ≤ ||x|| + ||y|| for all x, y ∈ R^d

---

Normed Space: Vector space witn norm is called normed space.(X,||.||)

Complete Metric Space: We saw the property of convergence, in simpler words what it meant is that if the sequence is not converging means there is a hole in space or it can be said that the space is not complete.

X = [0,3] with ($d_{(x,y)}$) = |x-y| is a complete because sequence is converging at 0 and 3 given that it's a closed set.

Now we can define **Banach Space**

if (X,$d_{||.||}$) is a complete metric space, than the normed space is called a **Banach Space**

So, until now we know that by norm we mean to measure distances and length, and by inner product, it means <x,y>. we are measuring distances, length, angles. So it gives the geometry of a plane.

A map <. , .> : X*X $\longrightarrow F$ is called an inner product if :

> 1. <x,x> $\geq$ 0 for all x ∈ X and <x,x> = 0

> 2. <x,y> = <y,x> for $F = R$

> 3. <x,y_1 +y_2 > = <x,y_1 > + <x,y_2 > for all x,y_1 ,y_2 ∈ X

f <.,.> is a inner product, then

$$||x||_{<.,.>} = \sqrt{< x, x >}$$

Definition : (x, <.>) is called **Hilbert Space** if

$$X, ||.||_{<.,.>}$$

is a **Banach Space.**

It means that it is a vector space where we can measure length, angles and it is also a complete metric space. So, Hilbert space is a real or complete vector space with an inner product such that the associated metric space is complete.

Ex: 1) $\mathbb{R}^n$, $\mathbb{C}^n$ with $\langle x, y \rangle = \sum_{i=1}^{n} \bar{x}_i y_i$ ← euclidean geometry.

2) This ex is generalization for infinite dim

$l^2(\mathbb{N}, \mathbb{F})$ with $\langle x,y \rangle = \sum_{i=1}^{\infty} \bar{x}_i y_i$ ← generalize this to infinite dim.

c) Continuous f^n defined on unit interval

$C([0,1], \mathbb{F})$ with $\langle f,g \rangle = \int \overline{f(t)} g(t) dt$ ← geometry for continuous f^n

domain co-domain

Not a Hilbert Space

we have inner product but completness fails here

So, now we have space that is linear, has an inner product, and complete. If you see above we discussed a problem referenced as paragraph **X** we were looking for creating an infinite dictionary. We wanted p to be infinite, which means we want our space to be of infinite dimensions. So, we can say we need a **Hilbert Space**

Now we have a complete, linear space with infinite dimensions but this is not enough we also need to have some sort of linear structure. it turns out there are some small changes that can help us in getting linear structure.

There is one property which can help :

Property of having evaluation functionals continous.

Let's see this in detail.

Functionals are mathematical objects which can take input as a function and return a number.

Evaluation Functionals are the functions that return the value of the function at a point that evaluates a function.

So the evaluation functionals are continuous means that if two functions are close then the evaluation functions are also going to be close. They just map functions into values.

if

$$||f - f^{`}|| \leq \delta$$

then there exists a

$$\epsilon_\delta$$

and

$$|f(x) - f`(x)| < \epsilon_\delta$$

So, we are looking for something which has these two properties :

> 1. It should have the property of evaluation functionals are continuous.

> 2. It should be from the Hilbert space of functions.

To understand it better, let's take two examples, one which has the property of evaluation functionals are continuous but not the Hilbert space of functions and vice versa

---

Example 1 :

Let's take a real line x ($X \longrightarrow R$)

C(x) = space of continous function

What is the natural norm in the case of continuous function :

If we want to attach a length to a continous function we use **supremum norm** :

$$||f||_\infty = sup_x|f(x)|$$

This is not a Hilbert space because Norm is not coming from the inner product, this is a L^1 norm.

Note: if the norm of the function is not coming from the inner product then it's not a Hilbert space.

What about the property of **evaluation functions**, can we evaluate this continuous function at a given point?

Here the norm is actually evaluating the function at an extremely positive point. So, if the functions are close then the points are also going to be close.

This example is not a part of Hilbert's space but following the property of evaluation functionals.

Example 2:

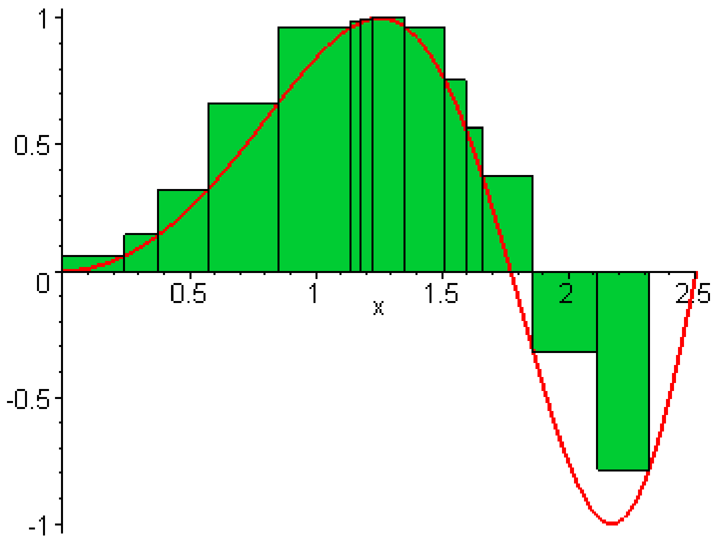
Take L^2 of R, which is just space of function such that it is finite.

$$L^2(R) = \{f \int |f(x)|^2 dx < x\}$$

As, the norm comes from the inner product it is a Hilbert space.

What about the second requirement:

If we take L^2, does it has continuous evaluation functions. No, because it is defined from the equivalence class of functions.



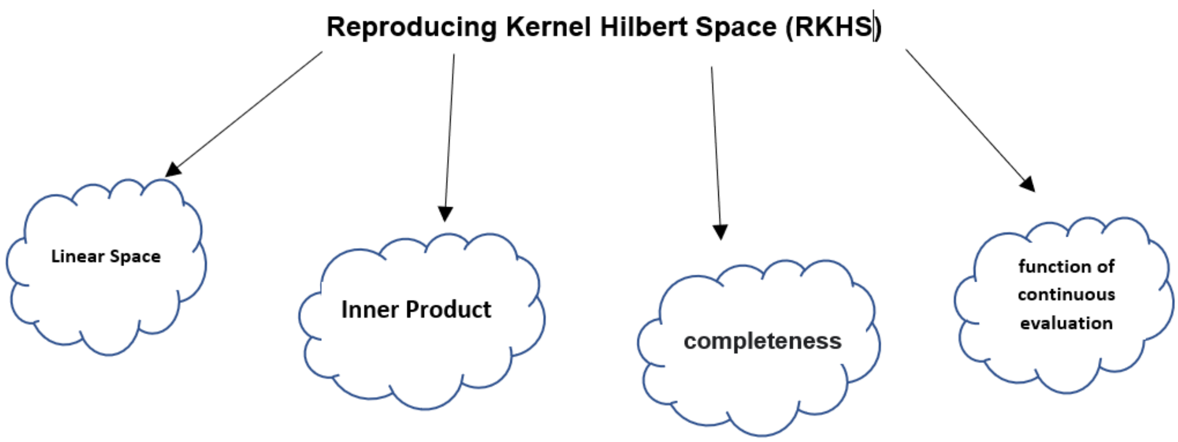
This function is completely defined by integral of functions and integral of functions does not care about what we do at one single point. So, there is no way to control :

$$|f(x) - |f(x)`|$$

So, L^2 is a space that satisfies the assumption of Hilbert space **(linear space, inner product, completeness)** but does not satisfy the evaluation functional assumption.

We need a function space with a norm but with the power to control what the function does at every point.

So, we need a function space where norm controls what the function does at every point, this is what we call a **Reproducing Kernel Hilbert Space (RKHS)**



Reproducing kernel Hilbert space (RKHS) is a space satisfying these four properties :

> 1. It is a linear space.

> 2. It has a norm defined by the inner product.

> 3. It is a complete place.

> 4 It has evaluation functional property.

Whenever all these assumptions are satisfied, we get a function which is called a **Reproducing Kernel**

Reproducing kernel is a function K x*x $\longrightarrow$ R goes from input to output and has two properties.

> If we fix one entry K(x,.) ε H, this function belongs to Hilbert space and if we take the inner product of any function with this function K(x,.) then this gives us f(x) back.

$$< f, K(x, . >) \geqslant f(x)$$

If H is RKHS $\Longrightarrow$ $\exists$ reproducing kernel

If H has reproducing kernel $\Longrightarrow$ $\exists$ RKHS

Summary:

RKHS is just a general mathematical object. it is just Hilbert space of function + evaluation functional requirement. Then we saw two examples of continuous functions for one reason or another they gave rise to RKHS.

Then we saw a theorem states whenever we have reproducing kernel Hilbert space we have some special function associated with it what we call a reproducing kernel.

Example :

$$H : \{f : X \longrightarrow R | \exists w \epsilon R^p, f = \sum_{j=1}^{p} w^j \phi_j\}$$

is it a hilbert space ?

It is a linear, how do we have inner product

$$< f, f`>_H = < w, w`>_{R^p}$$

$$||f(x) \leq ||w|| \sqrt{\sum_{j=1}^{p} (\phi_j(x)^2)}$$

As, the inner product define the norm so, it is a Hilbert space.

Is this the reproducing kernel for this space?

Remove one entry and check if it is still a function in H space.

$$K(x, .) = \sqrt{\sum_{j=1}^{p} (\phi_j . (x) \phi_j)}$$

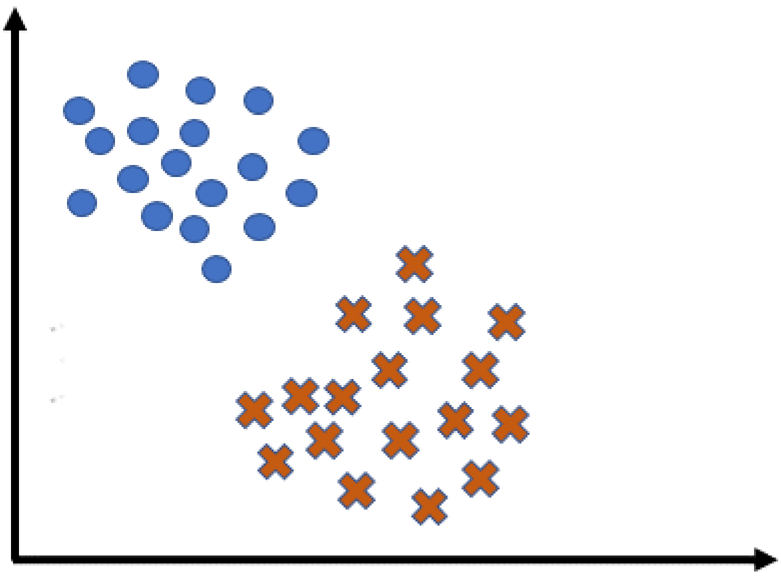
this is just a linear combination of a bunch of numbers, so this still belongs to H.

So, it is a reproducing kernel for this space.

Let's see a application of reproducing kernel and how they solve the problem of support vector machine.

We will be using **Scikit-learn library** for implementing the kernel.

Support vector machine SVM is a supervised machine learning classification algorithm. if the data is linearly separable in 2-d then the learning algorithm learns a boundary that divides the data with the aim to minimize the misclassification error.



Our example here is about predicting if the bank currency note is authentic or not.

We will be using 4 features :

> 1. Image variance.

> 2. Image entropy.

> 3. Curtosis of image.

> 4. Wavelet transformed skewness of image.

Importing libraries

In [6]:

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
```

Loading the Dataset

Link of dataset : [https://archive.ics.uci.edu/ml/datasets/banknote+authentication (https://archive.ics.uci.edu/ml/datasets/banknote+authentication)](https://archive.ics.uci.edu/ml/datasets/banknote+authentication)

In [7]:

```
bdata = pd.read_csv("bill_authentication.csv")
```

Exploring the dataset

In [9]:

```
print(bdata.shape)
print(bdata.head())
```

```
(1372, 5)
   Variance  Skewness  Curtosis  Entropy  Class
0   3.62160    8.6661   -2.8073 -0.44699      0
1   4.54590    8.1674   -2.4586 -1.46210      0
2   3.86600   -2.6383    1.9242  0.10645      0
3   3.45660    9.5228   -4.0112 -3.59440      0
4   0.32924   -4.4552    4.5718 -0.98880      0
```

Data Preprocessing

In [33]:

```
# dividing the data into train and test set

X = bdata.drop('Class', axis=1)
y = bdata['Class']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
```

Making the model

In [34]:

```
from sklearn.svm import SVC
model = SVC(kernel='linear')
model.fit(X_train, y_train)
```

Out[34]:

```
SVC(kernel='linear')
```

Making Predictions

In [35]:

```
y_pred = model.predict(X_test)

from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

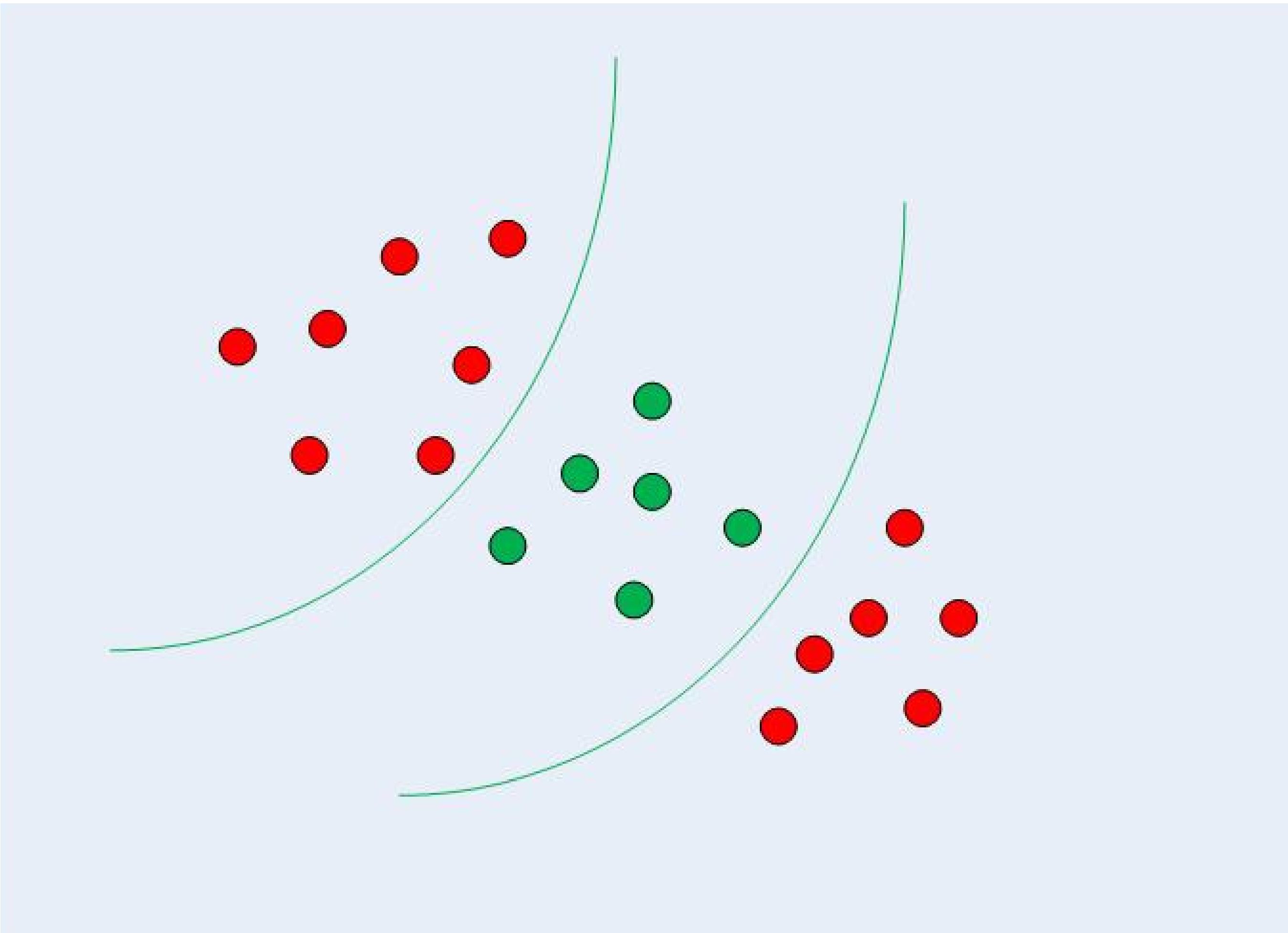
```
[[213   6]
 [  1 192]]
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       219
           1       0.97      0.99      0.98       193

    accuracy                           0.98       412
   macro avg       0.98      0.98      0.98       412
weighted avg       0.98      0.98      0.98       412
```

The Model is giving 7 misclassifications. Let's see if we can make it better with some kernel trick.

What if our data is not linearly seperable



In case of non-linearity , the straight line can not give us the decision boundary.

Instead of using SVM we need to use Kernel SVM. As , kernel projects the non-linearity of data from lower dimensions to linearly seperable in higher dimensions.

Implementing Kernel SVM with Scikit-Learn

SVM with gaussian kernel

In [36]:

```
from sklearn.svm import SVC
model = SVC(kernel='rbf')
model.fit(X_train, y_train)
```

Out[36]:

```
SVC()
```

In [37]:

```
y_pred = model.predict(X_test)

from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

```
[[218   1]
 [  0 193]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       219
           1       0.99      1.00      1.00       193

    accuracy                           1.00       412
   macro avg       1.00      1.00      1.00       412
weighted avg       1.00      1.00      1.00       412
```

Our SVM with gaussian kernel is giving quite a good result with only one missclassification

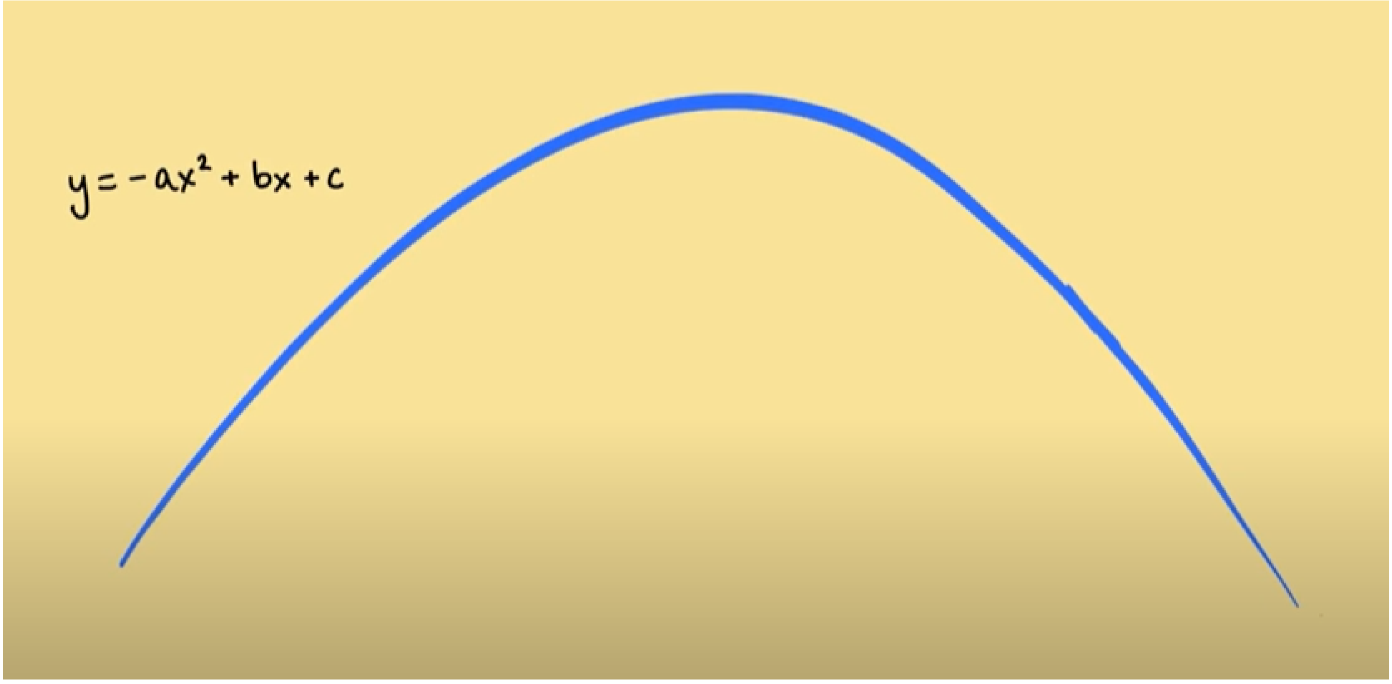
---

Hilbert space in quantum mechanics

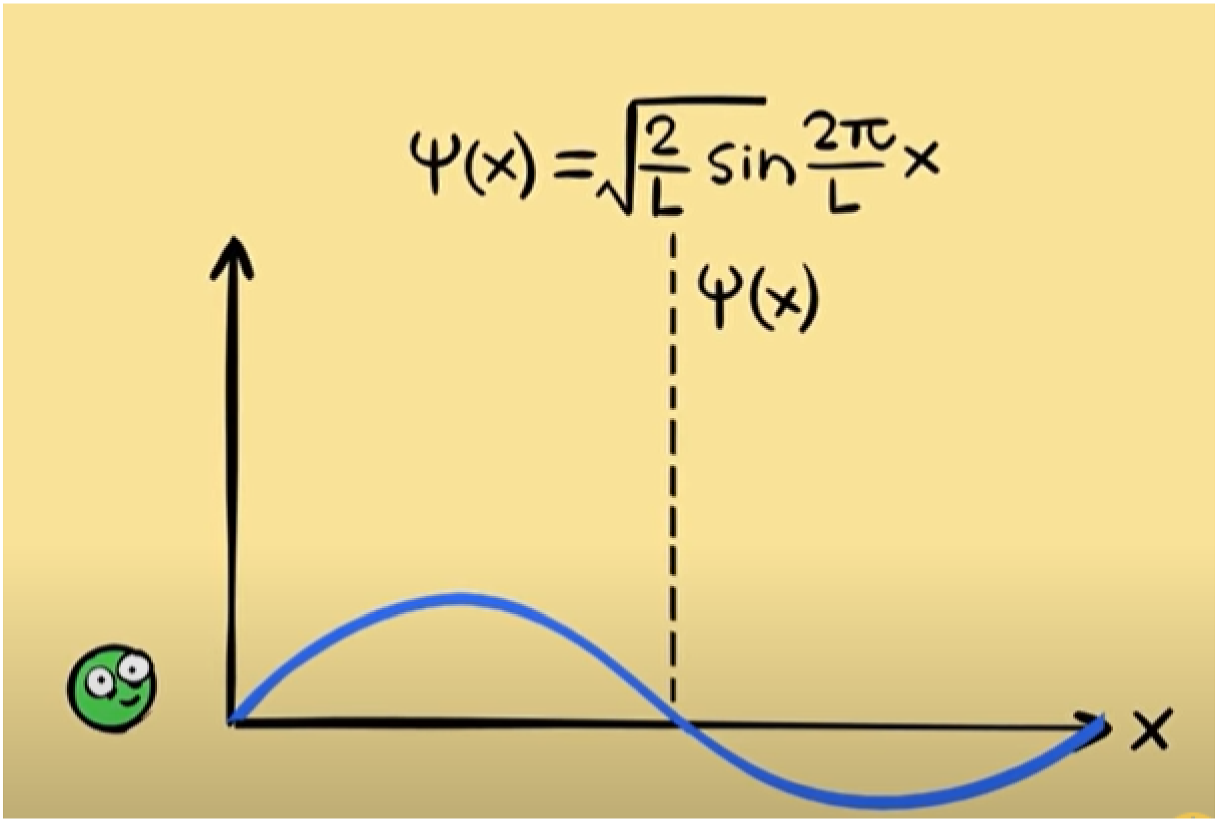
In quantum mechanics, we try to explain the interaction of subatomic particles, wave-particle duality, and the principle of uncertainty.

The most important function to study the subatomic particle is **Wave function**

If we throw a ball, we have a function to give the parabolic path of the ball thrown or we can say the behavior of the thrown object.

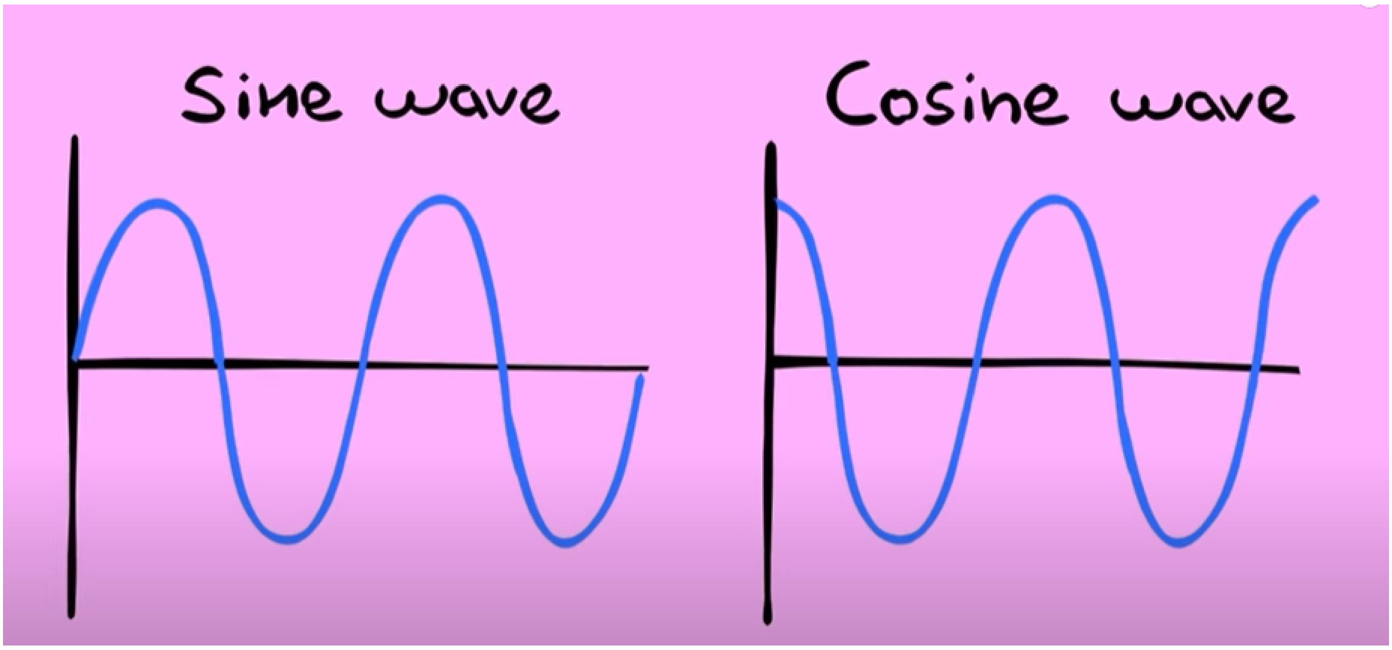


Similarly, wave function to define the behaviour of quantum particle.



So, next question arises is that what is this behaviour like ?

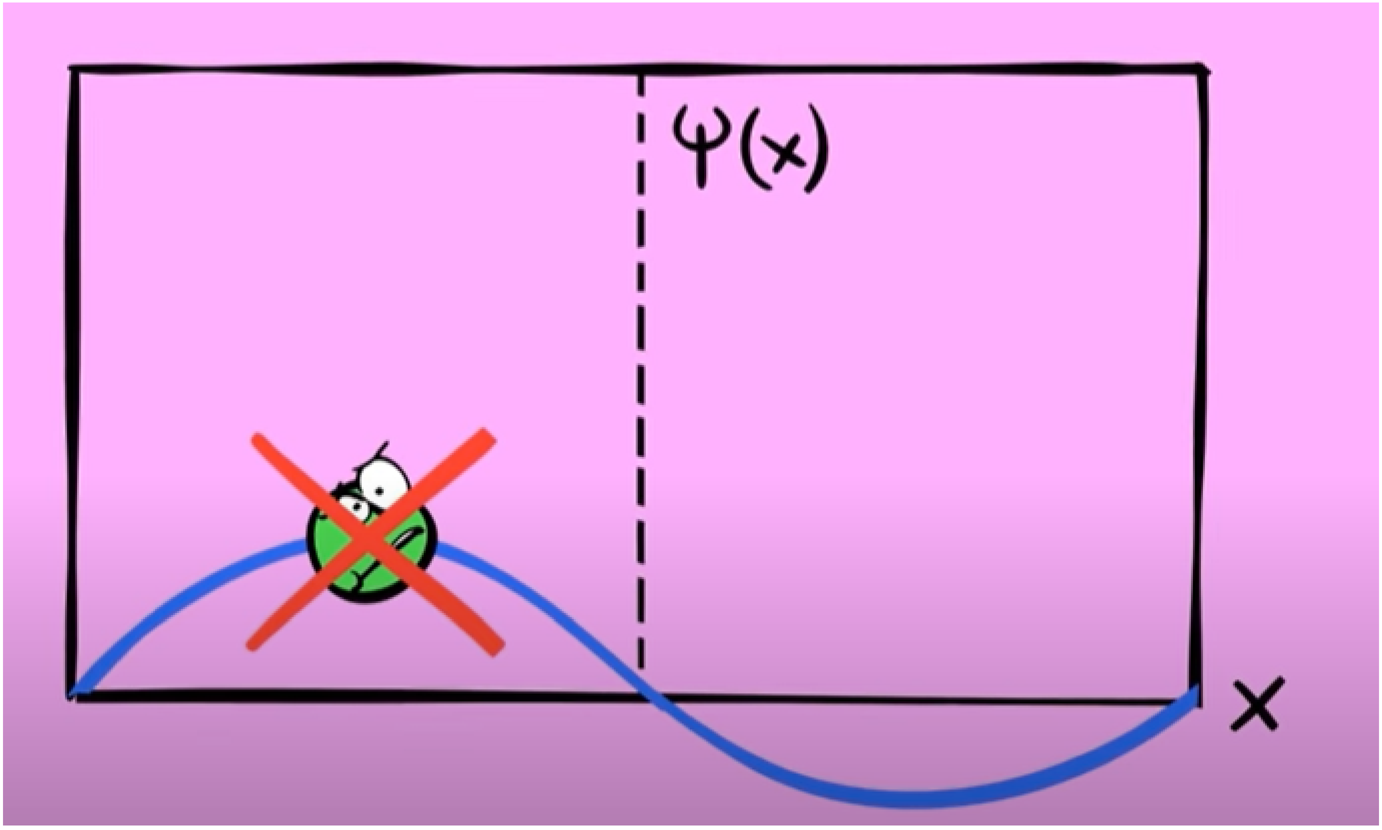
The behaviour of the quantum particle can always be described using the **combination of sine wave and the cosine wave.**



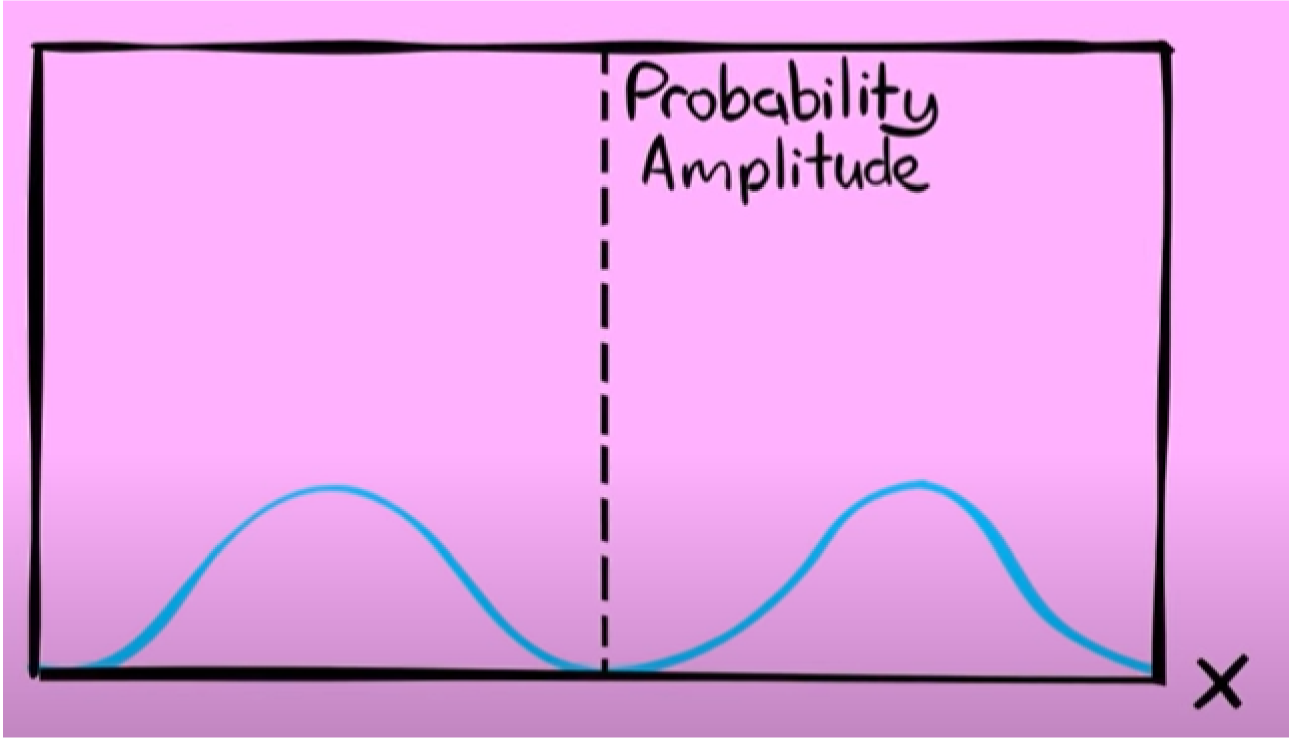
The wave function is the mathematical description of a quantum object takes the form of a wave.

What if our electron is inside a box amd can only move in 1-d ?

Does it move like a wave ?



No , wave function describes something totally different. The x-axis describes where the electron can be in space when we square the value of y-axis it give us something called **probability amplitude**.

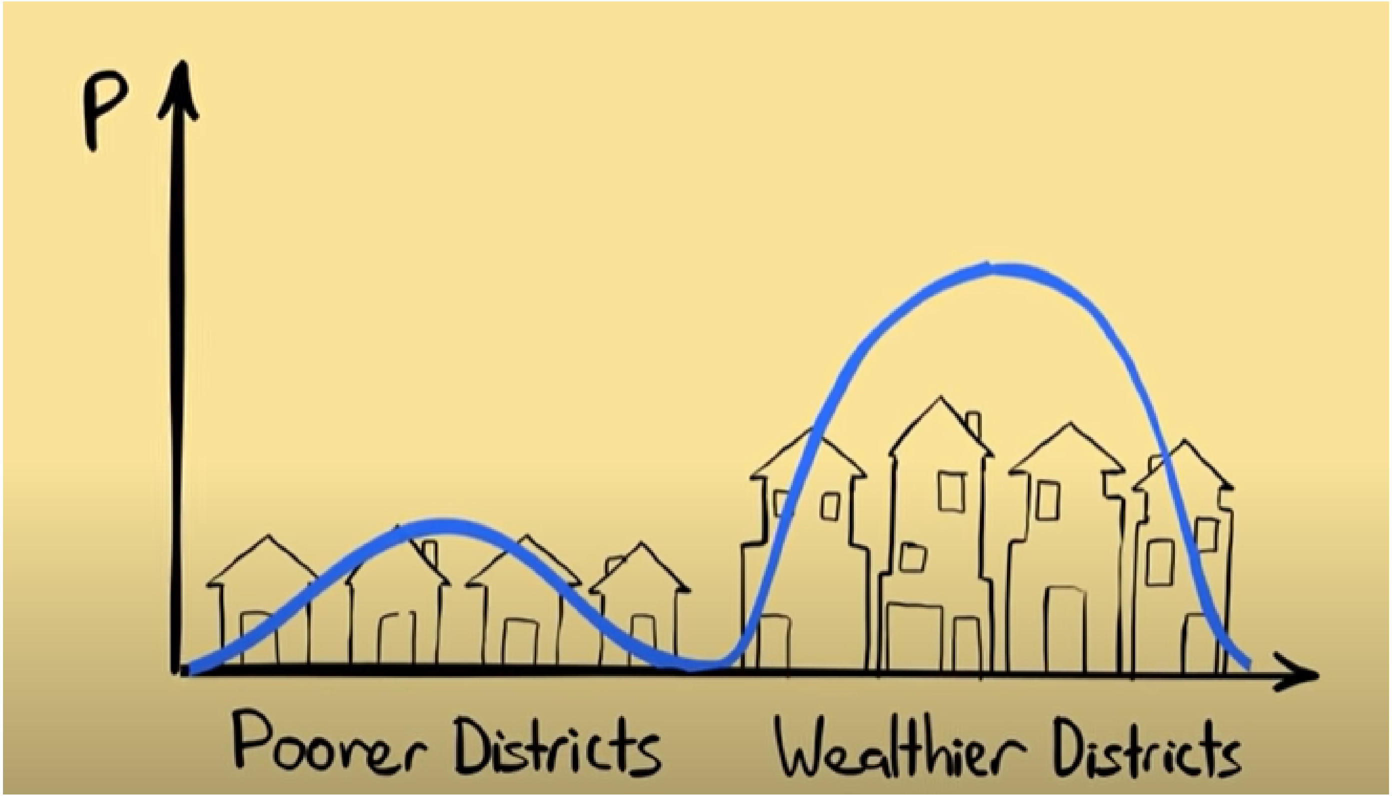


The probability amplitude is like a wave of probability. it describes the likelihood of finding the electron in a given place.

Let's understand it with an example, what is there a thief in town.

Police can catch him if they know where will he strike next.

There is a higher probability of police finding him in a wealthier district.



This can be thought of as a wave of probability, it's not real but just a set of abstract numbers that can be assigned to various parts of the city.

In a similar way, a wave function describes different probabilities of finding the particle at different locations.

If the electron is detected at a certain location, its wave function instantly collapses to a point and there is a zero probability of finding it anywhere else.

What space do we need for defining the wave function ?

Hilbert space is one of the choices of physicists for describing the state of the quantum systems.

The concepts of vectors play a vital role in providing information about any object and Hilbert space is a vector space of all possible functions subjected to some properties we discussed earlier.

The wave function in quantum mechanics tells us about the possible states which a particle can have and there might be infinite possibilities to it. This fits the definition of Hilbert space. Each function can be described as a point in vector space with axes are the basis of vector and co-ordinates are the scalar product between the basis function and wave function. Hilbert space help physicists in reducing it to a finite set of basis functions.

As we saw that the probability amplitude is the measure of the probability for finding the particle in some region of space. So, the wave function is square-integrable in such a way that the integral over all space of modulus squared converges according to the criteria of cauchy convergence. So, the function space should satisfy all the conditions of Hilbert's space.

Research scope of Reproducing kernel Hilbert space

The future scope can include the use of reproducing kernel Hilbert space for proving the optimality of activation functions used in neural networks. The main task of the activation functions in deep learning is to provide non-linearity to the model. we can construct the non-linear functions using the theory of RKHS and activation function constructed from RKHS then it can be considered optimal.

Why we can consider it optimal?

Because the activation function which will be created using RKHS will follow the properties of RKHS and vector spaces. it can lead to an efficient mapping process with the help of the inner product.

The future work can involve finding the possibilities of creating periodic activation functions like sine activation function using RKHS.