CHAPTER 6

OVERVIEW OF THE MATHEMATICAL BACKGROUND
FOR THE TESIS

6.1 INTRODUCTION

The Fourier transform is a useful tool to analyze the frequency components of the signal. However, if we take the Fourier transform over the whole time axis, we cannot tell at what instant a particular frequency rises. Short Time Fourier Transform (STFT) utilizes a sliding window to find spectrogram, which gives the information of both time and frequency. But still there exists another problem: The length of the window limits the resolution in frequency. Wavelet transforms are based on small wavelets with limited duration.

The big disadvantage of a Fourier expansion is that it has only frequency resolution and no time resolution. This means that even though we might be able to determine all the frequencies present in a signal, we do not know when they are present. In the past decades several solutions have been developed to overcome these problems which are more or less able to represent a signal in the time and frequency domain at the same time to cut the signal of interest into several parts and then analyze the parts separately. It is clear that analyzing a signal this way will give more information about the when and where of different frequency components, but it leads to a fundamental problem of how to cut the signal. The wavelet transform or wavelet analysis is the most recent solution to overcome the shortcomings of the Fourier transform. In wavelet analysis the signal-cutting problem was solved by the use of a fully scalable modulated window. The window is shifted along with the signal and for every position the spectrum is calculated. For every new cycle this process is repeated many times with a slightly shorter (or longer) window. The result will be a collection of time-frequency representations of the signal with different resolutions.

Time-frequency representation is provided by Wavelet transform. It is one of the most important and powerful tools of signal representation. Presently it is being used in image processing, data compression, and signal processing. Conventional Fourier transforms, wavelet transforms are based on small waves, called wavelets. Wavelet transform is one of the best tools for determining the low frequency area [94], [121], [123] & [124].
6.2 FOURIER TRANSFORM

\[ F(w) = \int_{-\infty}^{\infty} f(t) \exp(-jwt) \, dt \] …..(6.1)

Fourier transform converts signal in time domain to frequency domain by integrating over the whole time axis. However, if the signal is not stationary, that is, the frequency composition is a function of time; it is not possible to say when a certain frequency rises.

6.3 STFT (SHORT-TIME FOURIER TRANSFORM)

\[ Sf(u, \epsilon) = \int_{-\infty}^{\infty} f(t)w(t-u) \exp(-j\epsilon t) \, dt \] …… (6.2)

The STFT tries to solve the problem in Fourier transform by introducing a sliding window w(t-u). A small portion of the signal f(t) is extracted by the window and then Fourier transformed. The transformed coefficient has two independent parameters. One is the time parameter, indicating the instant we concern. The other is the frequency parameter, just like that in the Fourier transform. However, there arises another problem. The very low frequency component cannot be detected on the spectrum. It is the reason that we use the window with fixed size. Suppose the window size is 1. If there is a signal with frequency 0.1 Hz, the extracted data in 1 second looks like at (DC) in the time domain.

6.4 WAVELET TRANSFORM

\[ Wf(s, u) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi \left( \frac{t-u}{s} \right) \, dt \] ………(6.3)

Wavelet transform overcomes the previous problem. To strike a balance between time domain (finite length) and frequency domain (finite bandwidth) the wavelet function is designed in such a way that when the mother wavelet is dilated and translated, very low frequency components at large s and very high frequency components precisely at small s are seen.

6.4.1 Continuous Wavelet Transform

The wavelet analysis described is known as the continuous wavelet transform or CWT. More formally it is written as:

\[ \gamma(s, \tau) = \int f(t) \psi^*_s(t) \, dt \] ……….. (6.4)
where * denotes complex conjugation. The above equation shows how a function \( f(t) \) is decomposed into a set of basic functions \( \psi_{s,r}^* \), called the wavelets. After the wavelets transform the variables \( s \) and \( t \) are the new dimensions, scale and translation.

To overcome the resolution problem, the continuous wavelet transform was developed as an alternative approach to the short time Fourier Transforms. The wavelet analysis is done in a similar way to the STFT (Short Term Fourier Transform) as an analysis, in the sense that the signal is multiplied with a function, similar to the window function in the STFT, and the transform is computed separately for different segments of the time-domain signal.

### 6.4.2 Continuous Wavelet Transform with Discrete Coefficients

Signals are usually band-limited, which is equivalent to having finite energy and therefore just a constrained interval of scales is useful. However, the continuous wavelet transform produces redundant information while capturing all the characteristics of the signal. The continuous Wavelet Transform with Discrete Wavelet Transform (CWTDC) has been created to minimize the redundancies produced by CWT [123].

It is possible to compute the wavelet transform for just a proper selection of values of the frequency and time parameters and still not loose any information as recovery original time serious from its transform. The continuous wavelet transform with discrete coefficients is very similar to the continuous wavelet transform, but the parameters \( s \) and \( r \) are fixed.

### 6.4.3 Discrete Wavelet Transform

The continuous wavelet transform maps a one-dimensional time signal to a two-dimensional time-scale joint representation. The time bandwidth product of the continuous wavelet transform output is the square of that of the signal. The goal of signal processing is to represent the signal efficiently with fewer parameters. Using the discrete wavelet transform can reduce the time bandwidth product of the wavelet transform output.
By the term discrete wavelet transform, we mean, in fact, the continuous wavelets with the discrete scale and translation factors. The wavelet transform is then evaluated at discrete scales and translations. The discrete scale is expressed as \( s = s0 \ i \), where \( i \) is integer and \( s0 > 1 \) is a fixed dilation step. The discrete translation factor is expressed as \( \tau = k\tau0s0 \ i \), where \( k \) is an integer. The translation depends on the dilation step, \( s0 \ i \). The corresponding discrete wavelets are written as:

\[
h_{i,k}(t) = S_{0}^{-\frac{i}{2}}h(S_{0}^{-i}(t - k\tau S_{0}^{i})) = S_{0}^{-\frac{i}{2}}h(S_{0}^{-i}(t - k\tau)) \quad \ldots \ldots (6.5)
\]

The foundations of the DWT were laid in 1976 when Croiser, Esteban, and Galand devised a technique to decompose discrete time signals [77]. The discrete wavelets transform (DWT) is a linear transformation that operates on a data vector whose length is an integer power of two, transforming it into a numerically different vector of the same length. It is a tool that separates data into different frequency components, and then studies each component with resolution matched to its scale. DWT [69] & [93] is computed with a cascade of filtering followed by a factor 2 sub sampling as shown in Fig.32. H and L denote high and low-pass filters respectively, \( \downarrow 2 \) denotes sub sampling.

Fig.32 Discrete Wavelet Transform (DWT) Tree

6.4.4 Complex Wavelet Transform

Standard DWT Discrete Wavelet Transform, being non-redundant, is a very powerful tool for many non-stationary Signal Processing applications, but it suffers from three major limitations; 1) shift sensitivity, 2) poor directionality and 3) absence of phase information. Many researchers developed real-valued extensions to the standard DWT such as WPT (Wavelet Packet Transform), and SWT (Stationary Wavelet Transform) to reduce the above limitations.
These extensions are computationally intensive and highly redundant. Complex Wavelet Transform (CWT) is also presented as an alternate complex-valued extension to the standard DWT. The development of CWT was motivated to avail explicitly both magnitude and phase information [93]. A possible implementation of a complex wavelet transform is the one presented by Kingsbury in [68] & [90] as shown in Fig.33.

![Complex Wavelet Tree](image)

**Fig.33 Complex Wavelet Tree**

In figure above are presented four levels of the complex wavelet tree for a 1-D input signal $x$. The real and imaginary parts ($r$ and $j$) of the inputs and outputs are shown separately. Where there is only one input to a block, it is a real signal.

The extension of complex wavelets to 2-D is achieved by separable filtering along rows and then columns. In Fig.34 are presented two levels of the complex wavelet tree for a 2-D input image $x$, giving six directional bands at each level (the directions are shown for level 1). Components of 4-element ‘complex’ vectors are labeled $r$, $j_1$, $j_2$ and $j_1j_2$. 
6.4.5 Double Density Discrete Wavelet Transform

The new version of DWT, known as Double Density DWT (DDDWT) is recently developed. It has the following important additional propieties [43]:

1. It has one scaling function and two distinct wavelets which are designed in such a way as to be offset from one another by one half.

2. The double density DWT is over complete by a factor of two.

3. It is nearly shift invariant where complex wavelets with real and imaginary parts of approximating Hilbert pairs are proposed for denoising the signal.

The double density Discrete Wavelet Transform is constructed with the analysis and synthesis filter bank Fig.35.
In two dimensions, this transform outperforms the standard DWT in terms of enhancement; however, there is need for improvement because not all of the wavelets are directional. That is, although the double density DWT utilizes more wavelets, some Wavelets lack a dominant spatial orientation, which prevents them from being able to isolate those directions.

By combining the characteristics of the double density DWT and the dual tree DWT, double-density complex DWT provides a solution to this problem. The double density complex DWT is based on two scaling functions and four distinct wavelets, each of which is specifically designed in such a way that the two wavelets of the first pair are offset from one other by one half, and the other pair of wavelets form an approximate Hilbert transform pair.

The double density complex DWT possesses improved directional selectivity by ensuring the above two properties. It can be used to implement directional and complex wavelet transforms in multiple dimensions. We construct the filter bank structures for both the double density DWT and the double-density complex DWT using finite impulse response (FIR) perfect reconstruction filter banks. These filter banks are then applied recursively to the low pass subband, using the analysis filters for the forward transform and the synthesis filters for the inverse transform. While doing this, it is then possible to evaluate each transform’s performance in several applications including signal and image enhancement.
6.4.5.1 1-D Double Density DWT

The double density DWT is implemented by recursively applying the 3 channel analysis filter bank to the low pass sub band. This process is illustrated in Fig.36.

![Fig.36 Three Stage Recursion of the 1-D Double-Density DWT](image)

Conversely, the inverse double density DWT is obtained by iteratively applying the synthesis filter bank.

6.4.5.2 2-D Double Density DWT

To use the double density discrete wavelet transform for 2-D signal and image processing, we must implement a two-dimensional analysis and synthesis filter bank structure. This can be done simply by alternatively applying the transform first to the rows, then to the columns of an image. This gives rise to nine 2 D sub bands, one of which is the 2 D low pass scaling filter, and the other eight of which make up the eight 2 D wavelet filters as shown in Fig.37 [43] & [68].

The differences between the double density DWT and the dual tree DWT can be clarified with the following comparisons.

1. In the dual tree DWT, the two wavelets form an approximate Hilbert transform pair, whereas in the double density DWT, the two wavelets are offset by one half.
2. For the dual tree DWT here is fewer degrees of freedom for s design (achieving the Hilbert pair property adds constraints), whereas for the double density DWT, there are more degrees of freedom for design.

3. Uses different filter bank structures are used to implement the dual tree and double density DWTs.

![An Oversampled Filter Bank for 2-D Images](image)

**Fig.37** An Oversampled Filter Bank for 2-D Images

4. The dual tree DWT can be interpreted as a complex valued wavelet transform which is useful for signal modeling and denoising (the double density DWT cannot be interpreted as such).

5. The dual tree DWT can be used to implement 2-D transforms with directional Gabor like wavelets, which is highly desirable for image processing (the double density DWT cannot be, although it can be used in conjunction with specialized post-filters to implement a complex wavelet transform with low-redundancy, as developed.
6.4.6 Limitation of Discrete Wavelet Transform

Digital signals and images require efficient tools for their analysis. Some attention has to be paid to the use of robust methods in industrial applications since they are often corrupted by noises, errors or disturbances from different sources. The latter should take also into account classical properties of the data. Amongst these tools, the standard discrete wavelet transform (DWT) has been shown to be very effective both in theory and practice [73] in the processing of certain types of signals, for instance piecewise smooth signals, having a finite number of discontinuities.

The DWT [106] implements a form of multiscale analysis based on successive average/detail type approximations of data. Its most traditional structure relies on a critically sampled or decimated 2-channel filter bank with perfect reconstruction. In the ongoing standardization of the JPEG 2000 format for image compression this decimated structure is particularly efficient for coding applications. While mean, other processing applications such as data analysis, denoising, detection or restoration often require improvements over the DWT, to overcome its limitations.

One of the first most striking drawbacks of the DWT relates in one dimension (1D) to the time origin of signals: due to the presence of non shift-invariant operators (the down- and up-samplers), integer time-shifted versions of a signal yield wavelet coefficients which are not shifted accordingly, except at power-of-two time-shifts, depending on the number of resolution levels in the transform. After subsequent processing in the wavelet domain these results in shift-variant artifacts near jumps or edges in signals or images, which are not desirable in real-world applications since time-shifts are not controlled. In 1D, these artifacts can be avoided or limited by suppressing all or part of the decimators in the transform.

A second drawback further limits the DWT and its most natural redundant extension: tensor products of standard wavelets usually possess poor Directional properties. The later problem is sensitive in feature detection or denoising applications. A vast majority of the proposed solutions relies on adding some redundancy to the transform. Redundancy based on shift-in variant wavelet transforms. Several approaches to overcome this limitation have been
developed in the last years, most of them involving a combination of redundancy and improved directionality; for instance: other wavelet frames, bandlets, curvelets, directionlets as well as other “geometrical” wavelets. Some of these approaches bear a relative amount of complexity [31], [73], [74], [110], [126] & [130].

The third drawback concerns design imitations in two-band decompositions, which heavily constrain the properties of filters. In some applications, it is desirable to use real, orthogonal and symmetric filters with compact supports. The Haar wavelet is the only trivial and somewhat restricted solution in the DWT case. Filter banks with improved properties can be obtained with M-band filter banks and wavelets have been proposed [15] & [35].

6.5 DUAL TREE M-BAND WAVELET TRANSFORM (DTMBWT)

The Dual-Tree wavelet transform was initially proposed by Kingsbury [90] and further investigated by Selesnick [106]. As it is a multiresolution analysis, the feature vector is computed by decomposing the given mammogram using DTMBWT at pre defined levels of decomposition with $M$ band filter banks.

6.5.1 General Form of M-Band Filter Banks

Traditional discrete dyadic wavelet transforms can be implemented in several ways. The design of M-band uniform filter banks, which can be viewed as an extension of two-band systems, represents a challenging problem due to the greater number of aliasing components that must be cancelled. The most standard one is akin to the splitting property of the wavelet transform into low- and high-pass components. Loosely speaking, its building block consists in a low-pass and a high-pass filter associated in parallel, each one followed by a two-fold decimation or down sampling operator, as represented in Fig.38.
From this basic one-level wavelet stage, one can derive the necessary and sufficient conditions for the two filters to satisfy the so-called perfect reconstruction properties. Usually, the wavelets transform results from the iteration of one-level stages following each low-pass branch, which yields the well-known multiscale representation, dividing the frequency axis into dyadic frequency intervals. The practical usefulness of this representation in many data processing tasks is attributed in early works to its relationships with audio or visual perception, and to the wavelet ability to detect singularities, at least in 1D [73].

The discrete wavelet decomposition has been very successful in several domains including industrial applications as illustrated by the adoption of wavelets in the forthcoming compression standard JPEG 2000. Nevertheless, in some cases the dyadic structure is not the most appropriate. Amendments to the DWT, like wavelet packets, add flexibility to the octave decomposition, as demonstrated in the FBI compression standard for fingerprints [15], [50] &[129], which are regarded as medium-frequency images. But due to the relatively tight relationship between the low-pass and the high-pass filters, aliasing distortions may appear when processing both the wavelet or wavelet packets coefficients, e.g. in compression or denoising, where some coefficients are cancelled or shrinked [115]. A possible solution lies in relaxing some constraints: some constraints are depicted in Fig.37 by adding more branches with more filters and modifying the sampling factors. These modifications yield more general multirate systems for filter banks[62]&[132] which may comprise several stages of the one-level block depicted in Fig.38 with M branches and arbitrary integer decimation factors km, m∈{0,...,M−1}[31].
In Fig.39 shows general one-level filter bank stages. This general setting includes usual critically sampled filter banks [77 & 105] as well as oversampled ones, which are useful in communications. In many applications, it is more tractable to limit the freedom in the multirate system design. A lot of results can be obtained by simply using M filters in parallel with a decimation factor of M i.e. M-band filter banks Fig.40. These filter banks (FBs) yield a rich family of basis functions, including Discrete Cosine/Sine Transforms, Walsh-Paley-Hadamard functions, Malvar’s Lapped Orthogonal Transforms [65] (LOTs) and their extensions. With appropriate design-band FBs, M>2, generally possess sharper frequency than in the dyadic case, resulting in reduced sensitivity to aliasing problems. Moreover, the filters can be orthogonal, symmetric (or anti-symmetric) and real with compact support. Yet, these attractive properties can be enhanced by adding some multiresolution to the overall transform [73] & [35].

6.5.2 Construction of M-Band Hilbert Pairs
The M-band dual-tree wavelets prove more selective in the frequency domain than their dyadic wavelet transform. The performance of M-band wavelet Transform is demonstrated via de-noising comparisons on several image types with various M-band wavelets and thresholding strategies by Caroline Chaux et al [32].

Let us consider 1D signals belonging to the space $L^2(\mathbb{R})$. The M-band multi-resolution analysis of $L^2(\mathbb{R})$ (with $M \geq 2$) is defined by one scaling function (or father wavelet) $\psi_0 \in L^2(\mathbb{R})$ and $(M-1)$ mother wavelets $\psi_m \in L^2(\mathbb{R})$, $m \in \{1,\ldots,M-1\}$. These functions are solutions of the following scaling eqn.

$$\frac{1}{\sqrt{M}} \psi_m \left( \frac{t}{M} \right) = \sum_{k=\infty}^{\infty} h_m[k] \psi_0(t-k) \quad \text{ .......... (6.6)}$$

where the sequence $(h_m[k])_{k \in \mathbb{Z}}$ are square integral and are real-valued. The Fourier transform of $(h_m[k])_{k \in \mathbb{Z}}$ is a $2\pi$ periodic function, denoted by $H_m$. Hence, the above equation can be expressed in frequency domain as

$$\sqrt{M} \tilde{\psi}_m(M\omega) = H_m(\omega) \tilde{\psi}_0(\omega) \quad \text{ .......... (6.7)}$$

where $\tilde{a}$ is the Fourier transform of a function $a$. For the set of

$$\bigcup_{m=1}^{M-1} \left\{ M^{-j/2} \psi_m(M^{-j} t - k), (j,k) \in \mathbb{Z}^2 \right\} \quad \text{ .......... (6.8)}$$

function corresponding to an orthonormal basis of $L^2(\mathbb{R})$ the following para-unitarity conditions must hold:

$$\sum_{p=0}^{M-1} H_m\left(\omega + p \frac{2\pi}{M}\right) H_{m'}\left(\omega + p \frac{2\pi}{M}\right) = M\delta_{m-m'} \quad \text{ .......... (6.9)}$$

Where $\delta_m = 1$ if $m = 0$ and 0 otherwise. $H_0$ is low-pass, whereas usually the frequency response $H_m$, $m \in \{1,\ldots,M-2\}$ is a band-pass filter. In this case, cascading the M-band para-unitary analysis and synthesis filter banks depicted in the upper branch in Fig.41 allow to decompose to reconstruct perfectly a given signal.
From the 1 D dual-tree decomposition, Two-dimensional separable M-band wavelet bases can be derived. In a 2D case, there are two bases of \( L^2(\mathbb{R}^2) \); the first one corresponds to the classical 2D separable wavelet basis while the second one results from tensor products of the dual wavelet basis functions. A discrete implementation of these wavelet decompositions starts from first level to go up to the coarsest resolution level.

### 6.5.3 Application of Denoising

The 2-band multidimensional dual-tree complex wavelets transform has already been proved to be useful in denoising problems, in particular for video processing [36] or satellite imaging [65]. M-band dual-tree wavelet transforms also demonstrate good performances in image denoising and outperform existing methods such as those relying on classical M-band wavelet transforms (M\(\geq 2\)) or even 2-band dual-tree wavelet transforms.

### 6.5.4 Advantage of Dual Tree M-Band Wavelet Transform (DTMBWT)
Advantages of Hilbert pairs are [32]:-

1. Redundancy of only $2^d$ for $d$-dimensional signals, with a much lower shift sensitivity and better directionality in 2D than the DWT.
2. The double-density DWT and combined both frame approaches.
3. The phaselet extension of the dual tree DWT has been recently introduced by Gopinath. R [97]
4. A more recently proposed, projection scheme with an explicit control of the redundancy or with specific filter bank structures.
5. Analytic signals and wavelets must be mentioned, in the context of denoising.
6. Recent developments based on “geometrical” wavelets.

6.5.5 Extension to 2D Dual Tree M-Band Wavelet Analysis

A 2D generalization to the M-band case of the dual tree structure (initially proposed by N. Kingsbury [90] and further investigated by I. Selesnick [106] based on a Hilbert pair of Wavelets. We address the construction of the dual basis and the resulting directional analysis. We revisit the necessary pre-processing stage in the M-band case. While many reconstructions are possible because of the redundancy of the representation, a new optimal signal reconstruction technique is proposed which minimizes potential estimation errors. The effectiveness of the proposed M-band decomposition is demonstrated via image denoising comparisons in [43] & [55].

Two-dimensional separable M-band wavelet bases can be derived from the 1D dual-tree decomposition. Thus, we obtain two bases of $L^2(\mathbb{R}^2)$: the first one corresponds to the classical 2D separable wavelet basis while the second one results from tensor products of the dual wavelet basis functions. A discrete implementation of these wavelet decompositions starts from level $j=1$ to go up to the coarsest resolution level $J \in \mathbb{N}^*$. The decomposition onto the former 2D wavelet basis yields 2D coefficients $c_{j,m,m'}[k,l]$, whereas the decomposition onto the dual basis generates coefficients $C^H_{j,m,m'}[k,l]$. As pointed out in the seminal works of Kingsbury [90] and Selesnick[106], it is advantageous to add some pre- and post-processing to this decomposition.

6.6 SUPPORT VECTOR MACHINE
SVMs were introduced in COLT-92 by Boser, Guyon & Vapnik [137]. It is a theoretically well motivated algorithm, developed from Statistical Learning Theory by Vapnik & Chervonenkis [92] introduced since the 60s. It yields empirically good performance and has many successful applications in many fields (bioinformatics, text, image recognition, etc.). It is very large and diverse communities work in them, on machine learning, optimization, statistics, neural networks, functional analysis etc.

6.6.1 Mathematical Background for Support Vector Machine

SVM is a set of supervised machine learning approaches used for classification, regression and outlier’s detection. It is a non linear classifier and trained to automatically detect the presence of microcalcifications in a mammogram by El-Naqa et al [55]. It classifies the binary classes by computing a class boundary hyper plane and maximizing the margin in the given training data. It is used for wide range applications such as modern statistical learning theory, digital recognition, object recognition, speaker identification, face recognition and detection, cancer diagnosis, glaucoma diagnosis, gene data analysis and text categorization.

A classification task normally involves training and testing data which consist of some data instances. Each instance in the training set contains one “target value” (class labels) and several “attributes” (features). Let us consider the training samples \((x_1,y_1),(x_2,y_2),...,(x_n,y_n)\) where \(x_i\) in \(R^d\), d-dimensional feature space, and \(y\) in \{-1,+1\}, the class label for \(n\) samples. Mathematically, SVM finds the optimal values for the hyper plane parameters \(w\) (e.g. \(w_0\)) and \(b\) (e.g. \(b_0\)) that separates the samples. After finding the optimal separating hyper plane, such as \(w_0X+b_0=0\), an unseen pattern \(x_t\), can be classified by the decision rule:

\[
f(x) = \text{sign}(w_0X+b_0) \quad \text{......... (6.10)}
\]

where \(x\) is a vector of the dataset mapped to a high dimensional space. Each \(x_i\), belonging as it does to one of two classes, has a corresponding value \(y_i\), classes, while \(w\) and \(b\) are parameters of the hyper plane that the SVM will estimate. The nearest data points to the maximum margin hyper plane lie on the planes:

\[
(w.x)+b=+1 \quad \text{for} \quad y=+1
\]
\[
(w.x)+b=-1 \quad \text{for} \quad y=-1
\]

\[
\text{......... (6.11)}
\]
By rescaling \( w \) and \( b \), with no loss of generality, and grouping the above constraints in a single formula:

\[
\forall i, y_i f(x_i) \geq 1 \quad \text{\ldots\ldots\ldots (6.12)}
\]

where \( y = +1 \) for class \( w_1 \) and \( y = -1 \) for class \( w_2 \). The optimal separating hyper plane is enforced to separate the two classes of examples with the largest margin because, intuitively, a classifier with a larger margin is more noise-resistant. SVMs identify the data points near the optimal separating hyper plane which are called support vectors. The distance between the separating hyper plane and the nearest of the positive and negative data points is called the margin of the SVM classifier.

Intuitively, we want the hyperplane that maximizes the geometric distance to the closest data points and is shown in Fig.42 [47], [52], [120] & [128].

![Fig.42 Choosing the Hyperplane that Maximizes the Margin](image)

### 6.6.2 SVM Uses Different Kernel Functions

In this quadratic optimization task the kernel function used i.e. ‘K’ which maps the input features to a high dimensional feature space. Mathematically the kernel function is defined as:
\[ K(X_n, X_i) = \Phi^T(x_n) \Phi^T(x_i) \] ........... (6.13)

Classification used to various kernel functions are described in SVM literature for this Non-linear mapping of the input features. These are linear, quadratic, polynomial, Radial Basis Function (RBF) and Multilayer Perception (MLP). Mathematically these functions are expressed as below:

1. **Linear kernel**: \( K(x_n, x_i) = x_n, x_i \)

2. **Polynomial Kernel**: \( K(x_n, x_i) = (x_n, x_i + a)^b \), where \( b>2 \) and \( a>0 \).

3. **RBF Kernel**: \( K(x_n, x_i) = \frac{1}{\sigma^2} \exp \left( -\frac{|x_n - x_i|^2}{2\sigma^2} \right) \), where \( \sigma \) is the Standard deviation of Gaussian curve. The new test data evaluated with respect to the kernel trick is given as: \( y(x) = \sum \alpha_i y_i k(x_n, x_i) + b \)

4. **Multilayer Perception**: \( E = \sum_{i=1}^{N} \left( \frac{1}{2} \sum_{j=1}^{J} (d_{ij} - y_{ij})^2 \right) \), Where \( i \) indexes each pattern in the training set and \( j \) indexes each output variable (\( N \) patterns in the training set; each pattern has \( J \) outputs); \( d_{ij} \) is the desired value of output \( j \) as given by example pattern \( i \); \( y_{ij} \) and is the actual output value from the model.

5. **Quadratic Kernel**: \( K(X, Z) = (X^T Z)^2 \) or \( (1 + X^T Z)^2 \)

### 6.7 ARCHITECTURE OF SVM CLASSIFIER WITH LINEAR OR NON-LINEAR KERNEL FUNCTION

The support vector machine (SVM), can be characterized as a supervised learning algorithm capable of solving linear and non-linear classification problems. In comparison with neural networks we may describe SVM as a feed-forward neural net with one hidden layer and is shown in Fig.43. The main building blocks of SVM’s are structural risk minimization, originating from statistical learning theory which was mainly developed by VAPNIK and CHERVONENKIS. Non-linear optimization and duality and kernel induced features spaces, underline the technique with an exact mathematical framework. Meanwhile, many extensions to the basic SVM have been introduced, e.g. for multi-class classification as well as regression and clustering problems, making the technique broadly applicable in the data mining area.
The main idea of support vector classification is to separate examples with a linear decision surface and maximize the margin between the different classes. The Lagrange multiplier measures the influence of the i’th learning example on the functional $W$. Examples for which is positive are called support vectors, as they define the separating hyperplane. $C$ is a constant cost parameter, controlling the number of support vectors and enabling the user to control the trade-off between learning error and model complexity, regarded by the margin of the separating hyperplane. As complexity is considered directly during the learning stage, the risk of over fitting the training data is less severe for SVM. The separation rule is given by the indicator function using the dot product between the pattern to be classified ($x$), the support vectors and a constant threshold $b$.

For constructing more general non-linear decision functions, SVMs implement the idea to map the examples from input space $X$ into a high-dimensional feature via a priori chosen non-linear mapping function. The construction of $\Psi$ space a separating hyperplane in the features space leads to a non-linear decision surface in the original space see Fig.44. By introducing a kernel function, expensive calculation of dot products in a high-dimensional space can be avoided. Leaving the algorithms almost unchanged, this reduces numerical complexity significantly and allows efficient support vector learning for up to hundreds of thousands examples. The modified decision function is given in Fig.44.
Thus, the method is very flexible as a variety of learning machines can be constructed simply by using different kernel functions. Common kernels include polynomials of degree $d$ and radial basis function classifiers with smoothing parameter.

![Diagram](image)

**Fig.44 Non-Linear Mapping from Two-Dimensional Input Space**

In Fig.44 shows Non-Linear mapping from two-dimensional input space with Non-Linear class boundaries into a three-dimensional feature space with linear separation by a hyperplane. Compared with neural networks the SVM method offers a significantly smaller number of parameters. The main modelling freedom consists in the choice of a kernel function and the corresponding kernel parameters, influencing the speed of convergence and the quality of results. Furthermore, the choice of the cost parameter $C$ is vital to obtain good classification results, although algorithmic modifications can further simplify this task.

### 6.7.1 Linearly Separable Data Hard Margin SVM

A training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathbb{R}^n \times \{-1, 1\})^m$ is said to be linearly separable if there exists a linear classifier $h(x) = \text{sign}(w.x + b)$ which classifies all examples in $S$ correctly, i.e. for which $y_i(w.x_i + b) > 0 \; \forall i \in \{1, \ldots, m\}$. For example, Fig.45 (left) shows a training sample in $\mathbb{R}^2$ shows a training sample in $\mathbb{R}^2$ that is linearly separable, together with two possible linear classifiers that separate the data correctly (note that the decision surface of a linear classifier in 2 dimensions is a line, and more generally in $n > 2$ dimensions is a hyperplane).
A Linearly Separable Data Set, with two possible linear classifiers that separate the data. Blue circles represent class label 1 and red crosses -1; the arrow represents the direction of positive classification. Right: The same data set and classifiers, with margin of separation shown. Although both classifiers separate the data, the distance or margin with which the separation is achieved is different; this is shown in Fig.45 (right). The SVM algorithm selects the maximum margin classifier the linear classifier that separates the training data with the largest margin. More precisely, define the (geometric) margin of a linear classifier.

\[ h(x) = \text{sign}(w.x + b) \] on an example \((x_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} \) as \\
\[ \gamma = \min_{1 \leq i \leq m} y_i \] \\
where \(|w|\) denotes the Euclidean norm of w. The (geometric) margin of the classifier given by \((w; b)\) on a sample \(S=((x_1, y_1), \ldots, (x_m, y_m))\) is then defined as the minimal margin on examples in \(S\).

\[ \gamma = \min_{1 \leq i \leq m} y_i \]

Given a linearly separable training sample \(S=((x_1, y_1), \ldots, (x_m, y_m)) \in \mathbb{R}^n \times \{-1, 1\})^m\), the hard margin SVM algorithm finds a linear classifier that maximizes the above margin on \(S\). In particular, any linear classifier that separates \(S\) correctly will have margin \(\gamma > 0\); without loss of generality, we can represent any such classifier by some \((w; b)\) such that

\[ \min_{1 \leq i \leq m} y_i(w.x_i + b) = 1 \]

The margin of such a classifier on \(S\) then becomes simply

\[ \gamma = \min_{1 \leq i \leq m} \frac{y_i(w.x_i + b)}{|w|} = \frac{1}{|w|} \]
Thus, maximizing the margin becomes equivalent to minimizing the norm $\|w\|$ subject to the constraints in Eq. (6.17) which can be written as the following optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

Subject to

$$y_i(w.x_i + b) \geq 1, \quad i=1, \ldots, m \quad \ldots (6.18)$$

### 6.7.2 Non-Linearity Separable Data Soft Margin SVM

The above derivation assumed the existence of a linear classifier that can correctly classify all examples in a given training sample $S=((x_1, y_1), \ldots, (x_m, y_m))$.

In this case, one needs to allow for the possibility of errors in classification. This is usually done by relaxing the constraints in Eq. $y_i(w.x_i + b) \geq 1, \quad i=1, \ldots, m.$ through the introduction of slack variables $\epsilon_i \geq 0 (i=1, \ldots, m)$, and requiring only that

$$y_i(w.x_i + b) \geq 1 - \epsilon_i, \quad 1 = 1, \ldots, m \quad \ldots (6.19)$$

$$\epsilon_i \geq 0, \quad i = 1, \ldots, m$$

An extra cost for errors can be assigned as follows:

$$\min_{w,b,\epsilon} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \epsilon_i \quad \ldots \ldots (6.20)$$

Thus, whenever, $y_i(w.x_i+b) < 1$, we pay an associated cost of $C\epsilon_i = C(1 - y_i(w.x_i + b))$ in the objective function; a classification error occurs when $y_i(w.x_i+b) \leq 0$, or equivalently when $\epsilon_i \geq 1$. The parameter $C > 0$ controls the tradeoff between increasing the margin (minimizing $\|w\|$) and reducing the errors (minimizing $\epsilon_i$): a large value of $C$ keeps the errors small at the cost of a reduced margin; a small value of $C$ allows for more errors while increasing the margin on the remaining examples. Forming the dual of the above problem as before leads to the same convex QP as in the linearly separable case, except that the constraints in Eq. $\alpha_i \geq 0, \quad i=1, \ldots, m.$

$$0 \leq \alpha_i \leq C, \quad i=1, \ldots, m \quad \ldots \ldots (6.21)$$
The solution is obtained similarly for the linearly separable case; in this case, there are three types of support vectors with $\alpha_i > 0$ see Fig.46.

**Fig.46  Three types of Support Vectors in the Non-Separable Case**

1. Margin support vector ($\epsilon_i = 0$; these lie on the margin and are correctly classified).
2. Non-margin support vectors with $0 < \epsilon_i < 1$ (these are correctly classified, but lie within the margin).
3. Non-margin support vectors with $\epsilon_i \geq 1$ (these correspond to classification errors). The above formulation of the SVM algorithm for the general (non separable) case is often called the soft margin SVM.