Technische Universität Braunschweig

Institut für IDS Dynamik und Schwingungen

Master thesis

Numerical Investigation for Information Tracking of Noisy and Non-smooth data in Large-scale Statistics

submitted by

Avinash Bapu Sreenivas
Matr.-Nr.: 4359023

at

Institute for Dynamics and Vibration
Technical University Braunschweig

Advisor:

Prof. Dr.-Ing. habil. G.-P. Ostermeyer
Dr.-Ing. Tarin Srisupattarawanit

June 2018

Declaration

I, Avinash Bapu Sreenivas, declare that this master thesis titled, “Numerical Investigation for Information Tracking of Noisy and Non-smooth data in Large-scale Statistics” is written independently and no other resources other than those specified were used.

date: signature:

Acknowledgement

I would also take this opportunity to express my profound gratitude and deep regards to my supervisor at TU Braunschweig, Dr.-Ing. Tarin Srisupattarawanit and Prof. Dr.-Ing. habil. G.-P. Ostermeyer for their exemplary guidance, monitoring and constant encouragement throughout the course of this work.

Lastly, I thank my parents and friends for their constant support and encouragement without whom this assignment would not have been possible.

Abstract

The derivative of a function in the field of engineering has numerous applications, one of the major application is in the optimization process. These functions are generally specified by a large scale dataset in almost all engineering disciplines. In many instances, the function may be obtained from a large scale dataset that consists of some type of noise. The analysis of noisy function has ramification on the entire process of data analysis specifically during numerical differentiation. The reason behind this is that the derivative of noisy function tends to amplify the already present noise. Hence hampering the process and it is reflected in the solution of the same.

Overfit condition is a common occurrence while fitting a noisy function. This can be overcome by an additional penalty term. This is known as regularization. The derivatives are slightly treated in a different manner,i.e, the numerical differentiation process itself is regularized to minimizes the amplification of noise. This method is known as total variation regularization and it forms the basis of this thesis.

The successful implementation of this method mainly revolves around a factor known as regularization parameter. Hence, we employ various methods to determine the optimal regularization parameter for different test functions and noise levels.

Some of the methods require the user to provide a range of regularization parameter and based on that and various philosophies behind each method, they provide the optimal value. Therefore, optimization of these methods are vital. This forms the aim of the thesis.

The optimized methods are developed and these are tested using data-driven (sparse regression) methods and specific recommendations are provides to improve the process.

Contents

List of Figures

List of Tables

1 Introduction

In our universe, there is a presence of random bit of disorder in every field that has to be contemplated and understood clearly. This random bit of disorder in a physical system is known as noise. Noise in the field of statistics can be defined as an additional meaningless information that cannot be clearly interpreted which is present in the entire dataset.

In large-scale statistics, noisy data has an adverse effect on the results and it can lead to skewness in any data analysis process, if not properly understood or handled. The adverse effect on the results is mainly due to uncorrelated (zero autocorrelation) property of noise. This makes it completely unpredictable at any given point in time, hence thorough investigation and removal of noise plays a vital role in data analysis process.

In the field of engineering, measurement (experimental) data obtained by using scientific instruments consists of some values that are independent of the experimental setup. One of most widely technique is the optimization methods *viz*, gradient descent, conjugate gradient, Newton's method etc. Most of these methods require the determination of derivative of a function specified by the dataset (using finite-difference approximation). If the noisy data is approximated using a specific finite difference method this results in the amplification of noise present in the data.

In order to overcome the aforementioned problem of amplification of noise in the derivative of a function, various regularization methods are employed. The parameter that plays a vital role in these methods are termed as regularization parameter. One of the most important technique used in the field of regularization is known as total variation regularization.

1.1 Literature review

In the modern field of engineering, we deal with a lot experimental data that may consists of errors. These errors possesses the properties of randomness and non-correlation meaning that they are completely unpredictable in nature. Hence the knowledge behind these errors, proper handling and removal techniques are prioritized during the early phase of data analysis.

Various numerical method for approximating the derivative of functions like finite-difference methods have taken center stage in many engineering interdisciplinary for optimization purposes. Application of these finite-difference methods to the noise contaminated dataset leads to intensification of already present noise [13].

These amplification in the derivatives can be suppressed by applying total variation (TV) regularization technique. TV deals directly with the process of differentiation. This process of regularization assures that the calculated derivative of the function adheres to a certain degree of regularity [13]. The successful implementation of this methods hinges on one aspect,i.e, clearly understanding and determination of regularization parameter.

There are various methods that facilitates the determination of optimal regularization parameter [11, 18, 19]. One of the most important and widely used is the L-curve method. This method provides information on the regularization parameter based on the residual norm (L2) and the solution norm (L1) [18]. The graphical representation between the two for different regularization parameter provides an intersection point that stabilizes the effect of both the residual and the solution. This point is chosen as the the optimal regularization parameter by using curvature plot [11].

A method that completely focuses on extensive analysis of residual vector is the normalized cumulative periodogram [19]. The selection of optimal regularization parameter is based on Kolmogorov-Smirnov test i.e, the cumulative periodogram must strictly lie within the confidence interval of 95% [27]. In certain scenario either the variance of noise nor the exact data is unknown to the user. In these circumstances, the user is generally in a tough spot. Hence the generalized cross validation method [22] is employed to overcome complexities of unknown exact data or the variance of noise.

These optimal parameters can then be used in the data-driven (sparse regression) method in order to determine the PDE of the governing equation [26]. This method provides good approximation of the system as this uses brute-force search and the sparse regression technique for sparse nonlinear time series matrix in order to achieve its goal [26].

1.2 Aim and objective of the project

The aim of this project is to investigate the implications of noisy data in large scale statistics. The presence of such noisy data has a huge impact on the numerical differentiation process. Hence, the regularization of noisy data needs to be performed in order to retrieve vital information. This can be accomplished by achieving the following objective,

1. Understand and implement total variation regularization
2. The complexities involved in the determination of optimal regularization parameter are overcome by employing different methods and they are performed on various test functions to shown the behavior with respect to noise.
3. The information obtained from noisy data are tested using data-driven (sparse regression) method.

1.3 Structure of the report

The report is structured into

i) *Chapter 1* begins with introduction, a brief literature survey on regularization methods and sparse regression (data-driven) method.

ii) *Chapter 2* discusses the theoretical background of steps involved in data analysis, different types noise and a brief description. This chapter ends with the understanding the different types of regularization methods *viz*, L1 and L2 regularization.

iii) *Chapter 3* deals entirely with total variation regularization as this is the focal point of the project.

iv) *Chapter 4* facilitates the thorough implementation procedure and selection of regularization parameter involved in TV regularization.

v) *Chapter 5* encompasses the documentation of results using various test functions and different strengthens of noise (standard deviation). The results of data-driven (sparse regression) are also reported.

vi) *Chapter 6* presents the conclusion of the project.

2 Theoretical background

There are many regularization methods, few of the commonly used in the field of signal processing are,

1. Ridge regression
2. Least Absolute Shrinkage and Selection Operator (LASSO)
3. Total Variation Regularization or Rudin–Osher–Fatemi model

In order to maximize the potential of the aforementioned regularization methods, we shall start with the brief understanding of,

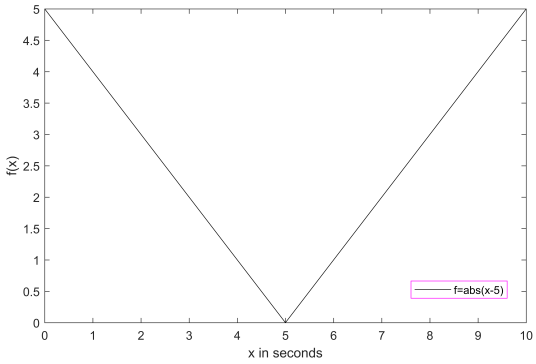
i) Data analysis involved in large scale data

ii) Different types of noise present in a general system described in equation 2.1

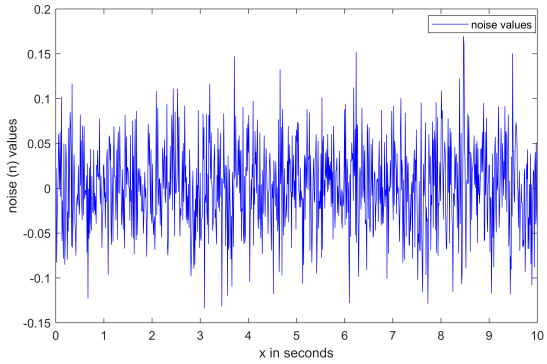
$$s = i + n \tag{2.1}$$

where,
s = Signal
i = Information
n = Noise

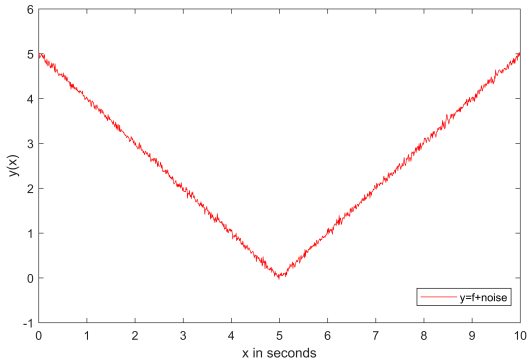
The equation 2.1 is visually represented in the following graphs,



(a)



(b) A graph showing an example for noise(n)



(c) A graph showing signal (s)

Figure 2.1: Graphs depicting the general system in equation 2.1

2.1 Data analysis and steps involved in the process

The process of obtaining raw data and its conversion into information which is useful for decision-making by the user, this is known as data analysis. The various steps involved in data analysis are shown in figure 2.2

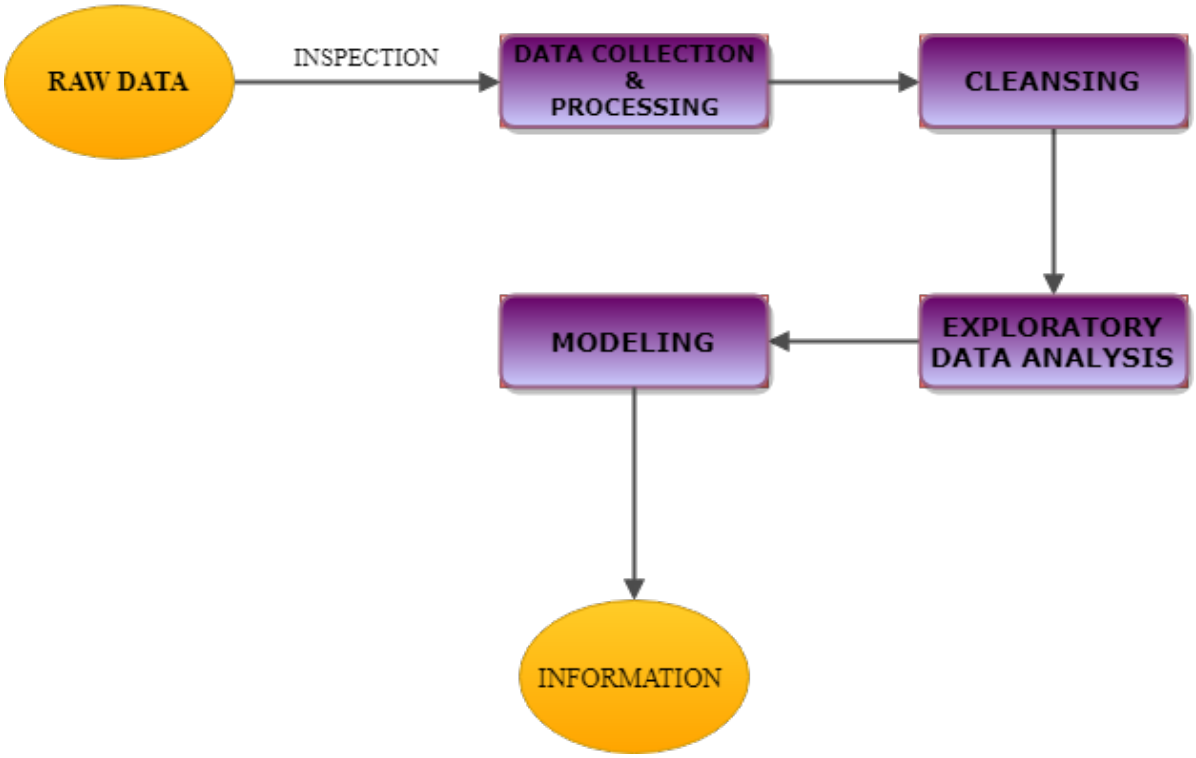


Figure 2.2: A picture showing the steps involved in data analysis

Data collection and processing:

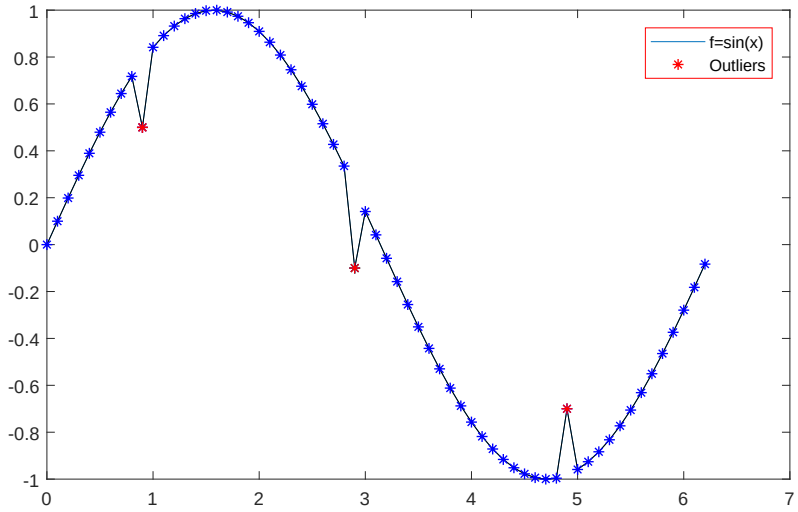
Data in general can be collected from various number of sources. Digital sources of data collection are some of the most convenient and trusted forms. In todays world where technological advancement is at its peak, senors form a large part of data collection. They are reliable, accurate and can transmit data round-the-clock to computers which can then be analyzed by the engineers. Temperature sensors in nuclear power plants, on aircraft to monitor engine temperature, seismic sensors in high earthquake prone regions in world are few examples that can provide engineers and scientists accurate data that can save lives during critical situations.

The collected data must then be organized for future analysis. This process of organization of collected data is known as data processing. Example of data processing is the placement of data into columns and rows with respective variable names in a statistical software (Microsoft® Excel or Minitab™).

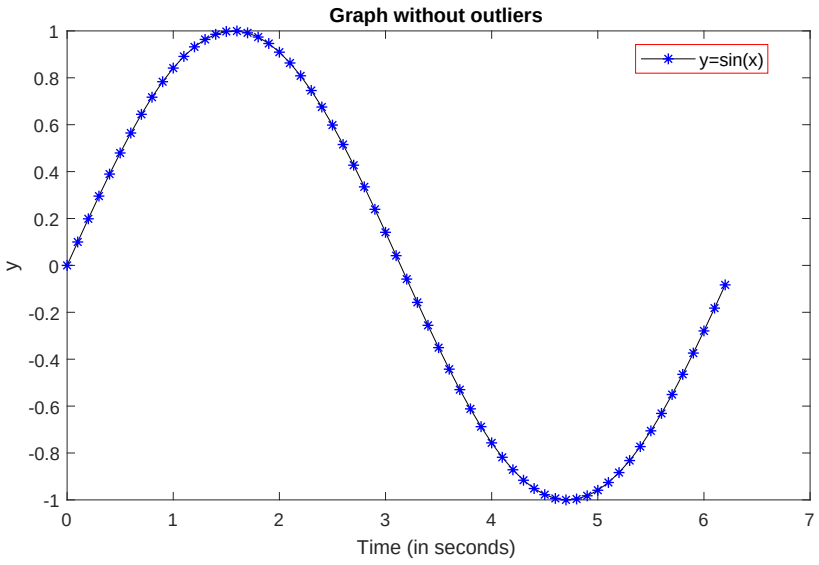
Cleaning the processed data

Data cleaning (cleansing) is the process of understanding, collection and then removal of errors that may be present in the processed data. This process is very critical during the final step of data analysis as it improves the accuracy of results. When dealing with quantitative

processed data using various outliers removal methods forms the part of data cleaning. Outliers are values or observation in processed data that lie far part from the main pattern of the entire dataset. Figure 2.3 shows a process with (figure 2.3(a)) and without outliers2.3(b)).



(a)



s

(b)

Figure 2.3: Representation of Outliers in a process

There are various methods to detect outliers in a process, one of the most commonly used technique is the scatterplot. This is very easy and quick process to detect the number of points lying outside the standard pattern of the whole process. Figure 2.4 shows the scatterplot and the detected outliers.

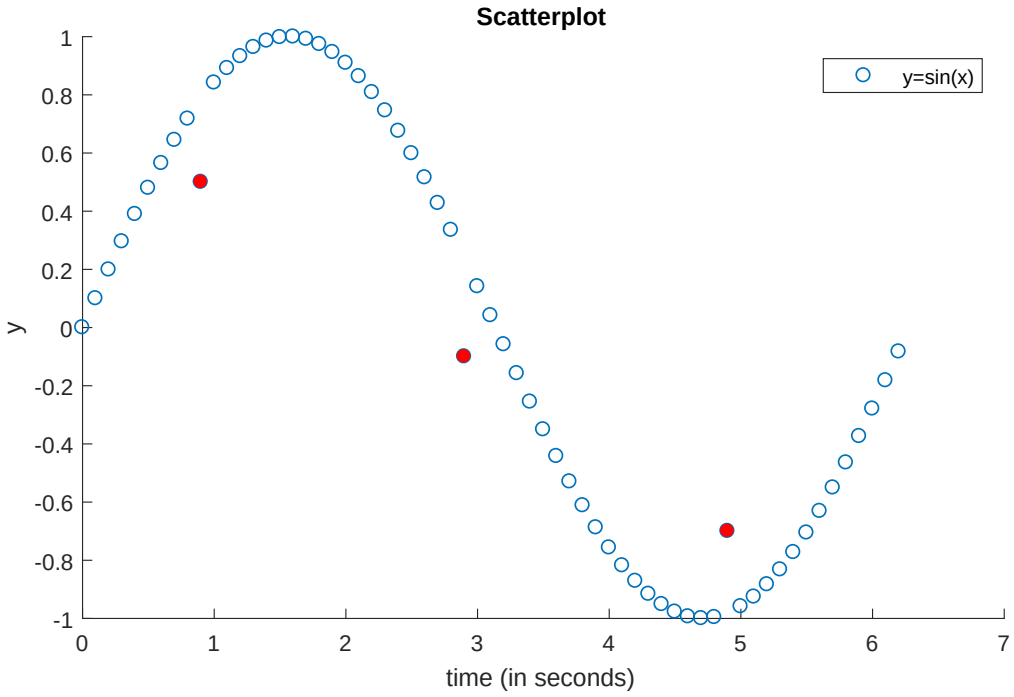


Figure 2.4: A scatterplot showing the process trend and the detected outliers

There are many other techniques like the box plot that are used in the detection of outliers in a process. The advantage of using box plot is that it provides clear information on mild and extreme outliers. Box plot also has the option of detecting outliers by using median, 1st and 3rd quartile principle. A typical boxplot is shown in figure 2.5

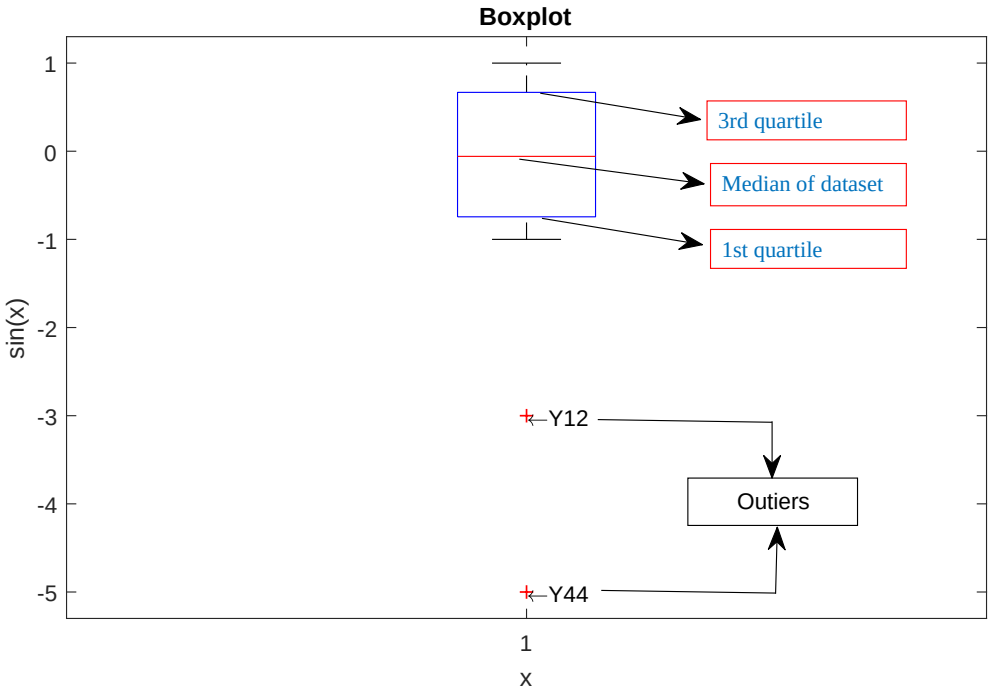


Figure 2.5: A boxplot showing the detected outliers

After the detection of outliers, one can not simply employ univariate and multivariate methods to remove the detected outliers as it can have adverse effect on the entire process [3]. So using robust techniques like "Minkowski error" method helps to reduce the impact of outliers on the dataset (or model). The major advantage of "Minkowski error" over RSS is that it reduces the effect of outliers by taking the power of error terms lesser than 2 [3].

In certain scenarios, processed data and/or processed data after treating outliers may be skewed. This type of skewed data needs to be transformed using certain transformation techniques before analyzing exploratory. The most common method employed for skewed data is the Box-Cox (or power) transformation.

$$x(\lambda) = \frac{(x^{\lambda} - 1)}{\lambda} \qquad \lambda \neq 0 \tag{2.2}$$

$$x(\lambda) = ln(x) \qquad \lambda = 0 \tag{2.3}$$

where,

$x(\lambda)$ = Transformed data

x = Skewed data

λ = Box-Cox parameter

But the best way [1] to select "λ" is by using LLF (logarithm of likelihood function).

This marks the conclusion of cleansing of processed data.

Exploratory data analysis

The process of deciphering the cleaned data extensively by using visualization techniques, calculation of vital descriptive statistics (like mean,median,mode etc) is known as exploratory data analysis. This helps the user to comprehend the meaning behind the obtained dataset. Hence it translates to exploring the cleaning data from all possible angles.

It consists of many sub-tasks like,

i) Re-cleansing (if necessary)

ii) procurement of additional data

iii) Calculation of descriptive statistics

iv) Visualization

Data modeling

The final step in process of data analysis is data modeling. The knowledge obtained from exploratory data analysis steps plays a vital role in the identification of certain relationship between variables. These relationship such as regression analysis, correlation can be obtained by compiling specific algorithms and/or applying specific mathematical formulae. Finally the user can construct descriptive models for analysis. The results obtained can be termed as information, this can help the user to understand the datasets and certain changes can be made in order to improve the efficiency of the process for future studies.

2.2 A brief discussion about noise

This section focuses on the different types of noise and its characteristics encountered in various statistical and signal processing fields. As shown in equation 2.1, noise "n" can be classified as shown below,

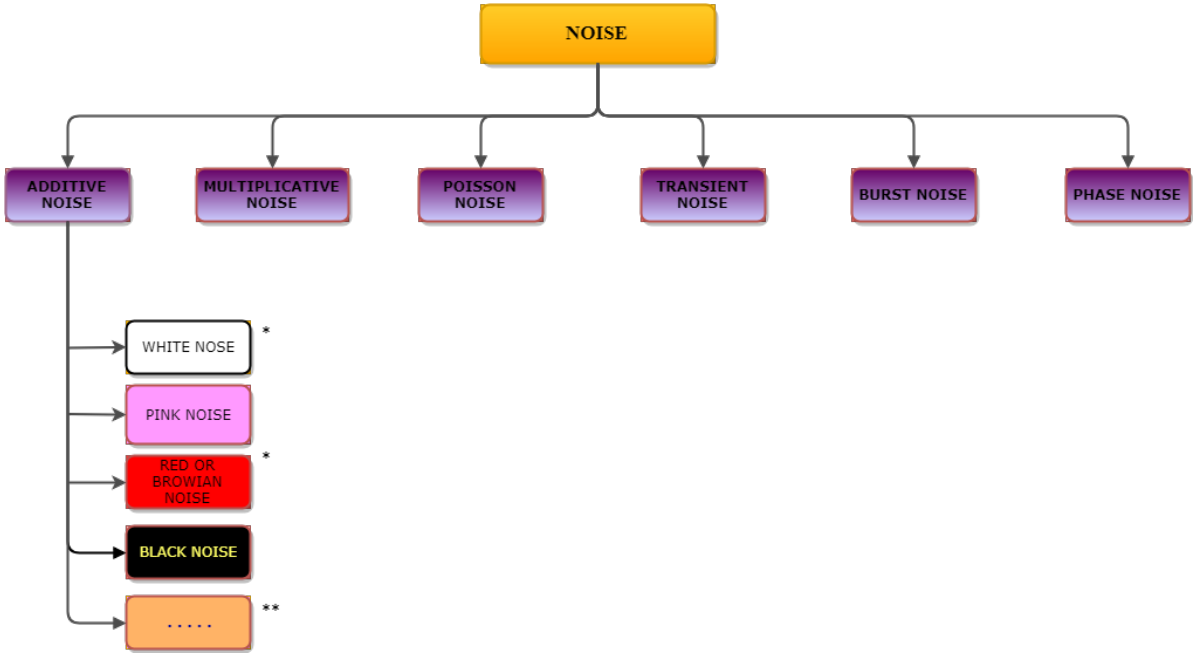


Figure 2.6: Classification of noise

Note: As seen from figure 2.6,

$(*) \implies$ main focal point. Hence it is explicitly described in 2.2.1.

$(**) \implies$ additive noise includes many other slightly less significant subdivision.

Let us now begin to understand the different types of noise as seen in figure 2.6.

1. **Multiplicative noise**
   In a given system , if the random terms depends on the state of that system , this type of noise is termed as multiplicative noise. In terms of dataset, we can say that the noisy data is the resultant of noise multiplied to the data vector. This can be clearly interpreted with the help of a following system (model).

$$s = i \cdot n \tag{2.4}$$

   where,

   s = Signal
   i = Information (true signal)
   n = Noise

   Denoising of multiplicative noise requires a transformation of the model in equation 2.4 into additive noise. Logarithmic transformation is very helpful tool in denoising

multiplicative noise as this provides an additive form.

$$log(s) = log(i \cdot n) \tag{2.5}$$
$$log(s) = log(i) + log(n) \tag{2.6}$$

where,

s = Signal
i = Information (true signal)
n = Noise

Now, equation 2.6 clearly represents an additive system and various denoising techniques can be applied. Finally , inverse logarithm (log^{-1}) of the denoised signal provides the solution to the original system.

2. **Poisson noise**
   Poisson noise is also termed as shot noise. Shot noise is mainly observed in electronic devices. This type of noise is generated when a charge carrier such as electrons or ions travel through a gap results in random fluctuation in electric current. This random fluctuation is known as shot noise.
3. **Transient noise**
   This type of noise is very common in the field of communication systems like mobile phones and hearing aids. The background noise that hinders communication in the field of communication systems is termed as transient noise.
4. **Burst noise**
   Burst noise is also termed as Random Telegraph Signal (RTS) and "popcorn" noise. It is very similar to the shot noise and generated at low frequencies. When a single charger carrier is captured by a single trapping center, this leads to the generation of burst noise as shown in figure 2.7.

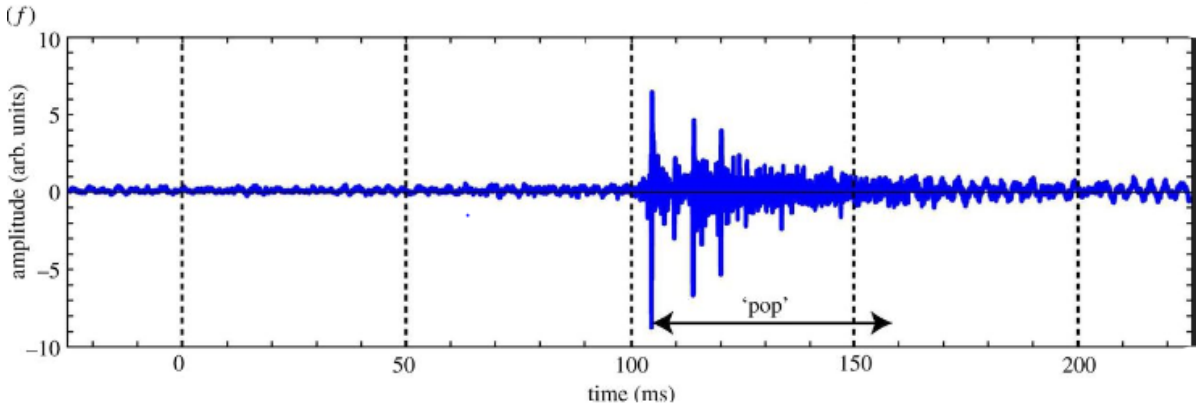


Figure 2.7: A graph showing the generation of pop (burst) noise [28, p. 5]

5. **Phase noise**

   In order to understand the meaning and definition of phase noise , let us define the term "phase". Phase in a waveform cycle is defined as the position of a point in time. Three types of phases in a wave is shown in figure 2.9.

   Square,triangle,sinusoidal,complex are a few examples of different types of waveforms shown in figure 2.8. The random and rapid variation of phase in a signal (waveform) caused by time domain instability is known as phase noise.

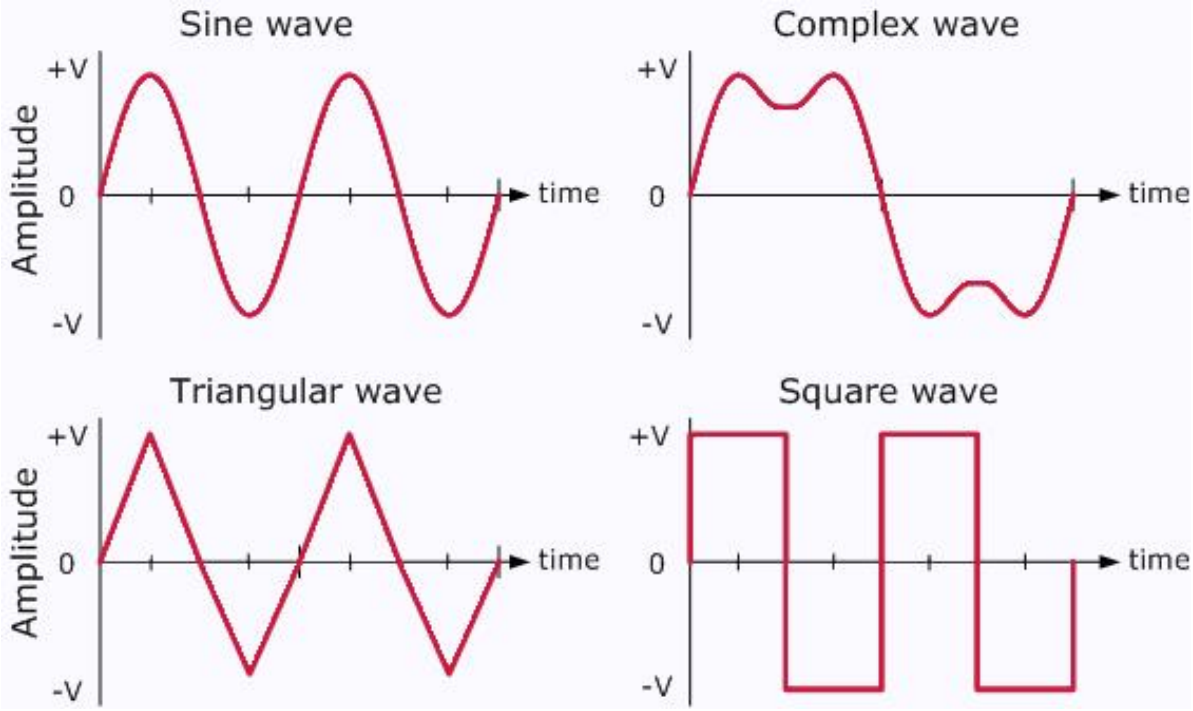
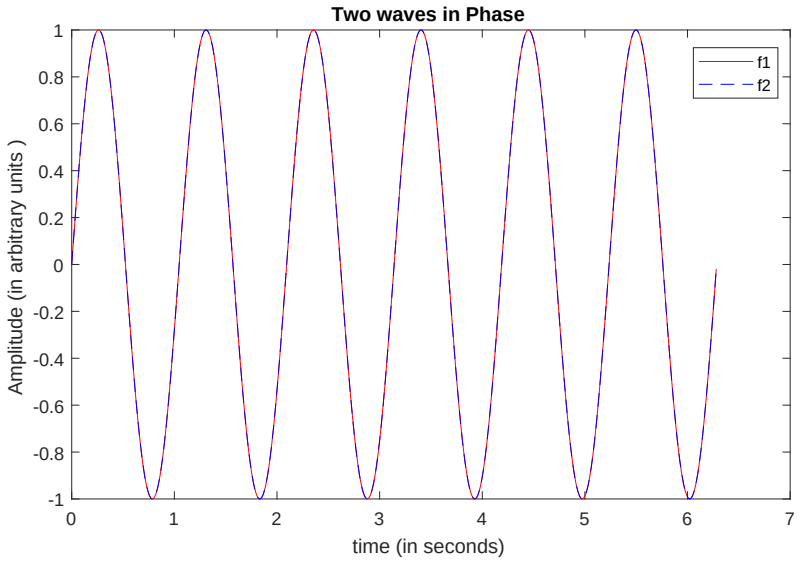
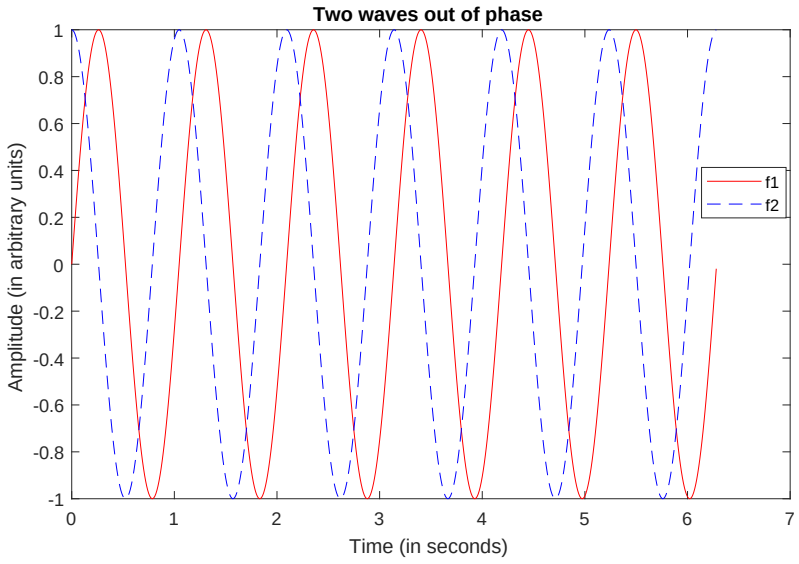


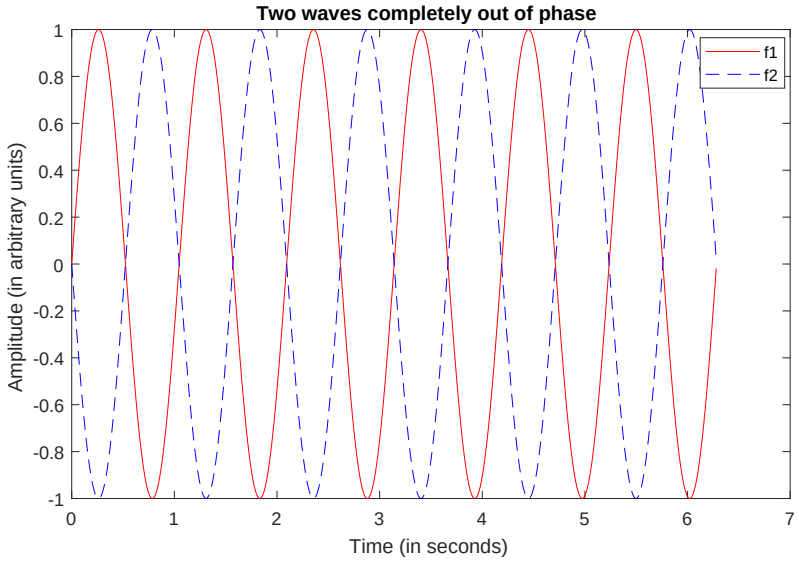
Figure 2.8: A picture showing common types of waveforms [8]



(a)



(b)



(c)

Figure 2.9: Graphs showing two waves in phase (a), out of phase (b) & completely out of phase (c)

2.2.1 Additive White Gaussian Noise (AWGN)

Before jumping into the deep end regarding the explanation of AWGN, let us first breakdown and understand the terminology "*Additive White Gaussian Noise*".

1. *Additive* $\Longrightarrow$ This type of noise are additive in nature
   This means that the received signal is the resultant of information added with some noise as shown in equation 2.1

2. *White* $\Longrightarrow$ It is mixture of all types or colors of noise
   White light is mixture of all the frequencies or wavelength of visible spectrum (shown in figure 2.10). This definition of white light is literally translated into white noise.

3. *Gaussian* $\Longrightarrow$ This type of noise follows normal probability distribution function (pdf). classified as shown below,

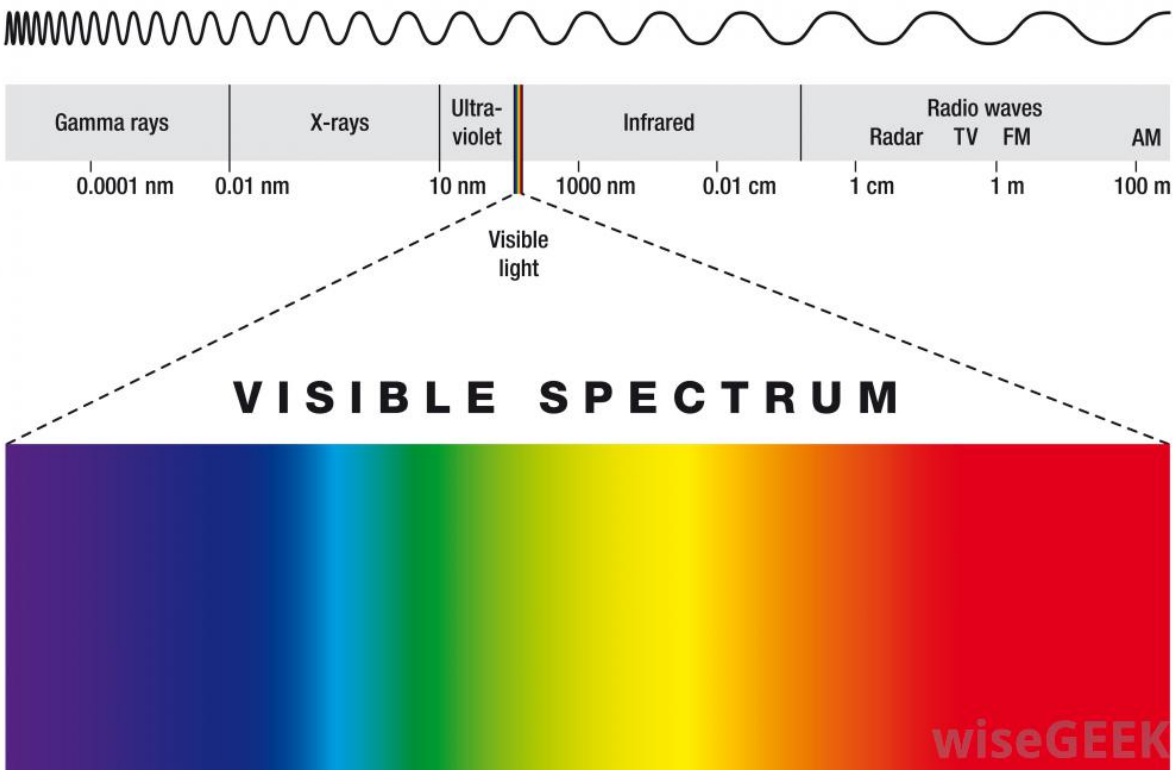
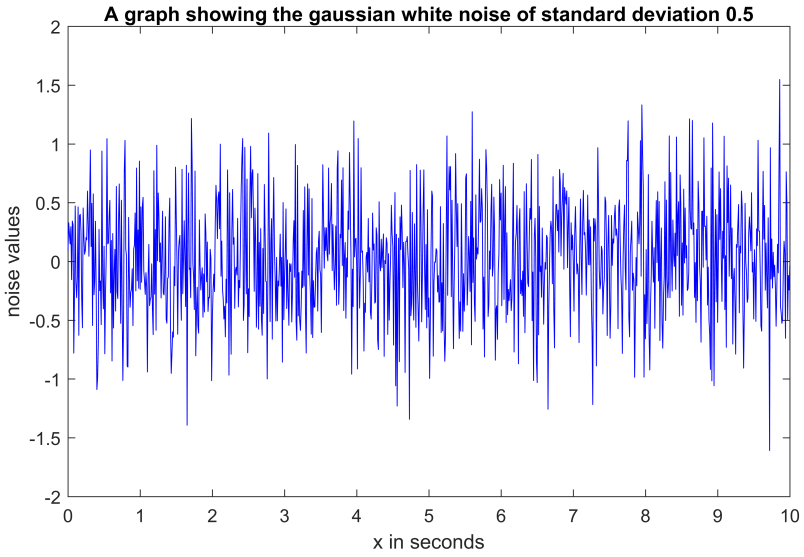


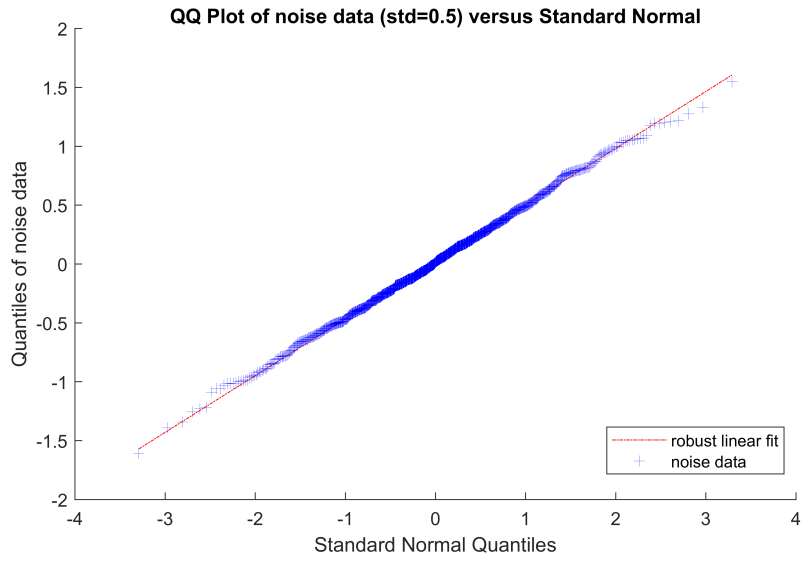
Figure 2.10: A picture showing the visible spectrum [9]

White noise with respect to a signal and its source is a statistical model having constant power spectral density (PSD), which means that it is a random noise having equal intensity for different frequencies.

An example of the Gaussian white noise is shown below,



(a)



s

(b)

Figure 2.11: Representation of Gaussian white noise and its quantile-quantile plot

2.2.2 Brownian noise

Brownian noise is also known as,

i) Red noise –Longer wavelength produces stronger noise similar to radio waves shown in figure 2.10, hence the term "*red*" noise

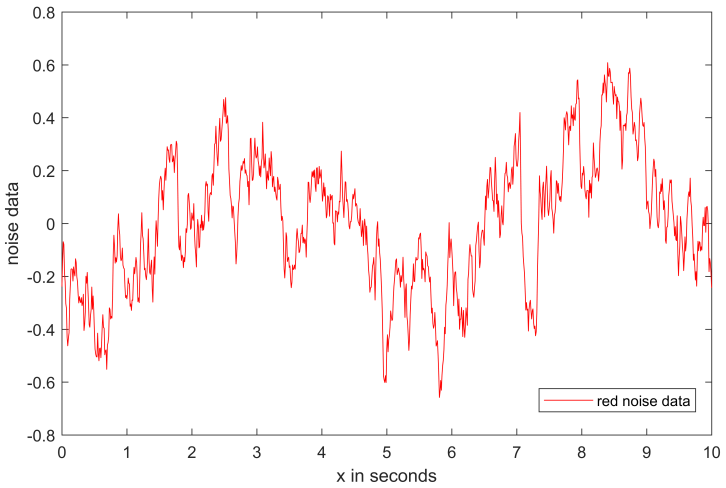
ii) Brown noise –Robert Brown discovered Brownian motion. Hence its also coined as "*brown*" noise.

The characteristics of red are briefly discussed below,

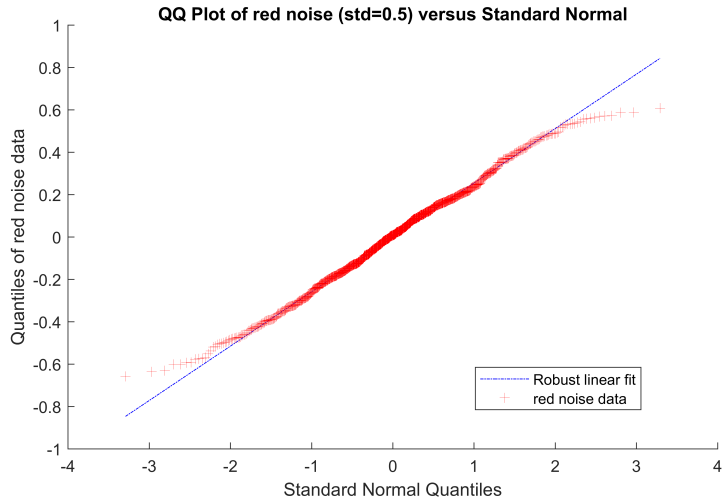
i) Red noise has more energy at lower frequencies $\implies P(f) \propto 1/f^2$.
Power spectrum is denoted by $P(f)$
Frequency is denoted by f

ii) Integration of white noise $\rightarrow$ Red noise

An example of the Brownian or red noise is shown below,



(a)



s

(b)

Figure 2.12: Representation of Brownian/red noise and its quantile-quantile plot

With this brief understanding of different types of noise, let us now dive into the concepts surrounding important regularization methods.

2.3 A brief discussion regarding regularization methods

As mentioned earlier , the 3 widely used regularization techniques are

1. Ridge regression or Tikhonov regularization method
2. Least Absolute Shrinkage and Selection Operator (LASSO)
3. Total Variation Regularization or Rudin–Osher–Fatemi model

Before we step into each of the aforementioned regularization techniques, let us define the term *regularization*.

Regularization is defined as a method that helps to overcome the problem surrounding over-fitting of penalized regularization coefficients [15]. This aim of regularization is achieved by the introduction of additional information to solve ill-poised problems. Due to the fact that minimization of residual sum square are highly unstable in nature, regularization methods proves to be all the more important in many scientific fields.

2.3.1 Ridge regression (L2 regularization)

The aim of ridge regression is to minimize the ordinary least square with an added penalty term.This penalty term is the square of the magnitude of the coefficients.This explanation is summarized in equation 2.8.
The ridge regression solution "$\hat{x}_{ridge}$" solves the following minimization problem for a given system $Ax = b$,

$$\underset{x\in\mathbb{R}^m}{\operatorname{argmin}} \sum_{i}^{n} \left(\sum_{j}^{m} a_{ij}x_j - b_i^2 \right)^2 + \alpha \sum_{j}^{m} x_j^2 \tag{2.7}$$

The equation 2.7 can be represented in a simpler form as,

$$\underset{x\in\mathbb{R}^m}{\operatorname{argmin}} \underbrace{\|Ax - b\|_2^2}_{Residual} + \alpha \underbrace{\left\|x^2\right\|_2^2}_{Penalty} \tag{2.8}$$

where,

$b \in \mathbb{R}^n$ = Response vector

$A \in \mathbb{R}^{n \times m}$ = Predictor matrix

α = Regularization parameter

In matrix notation equation 2.8 becomes,

$$C_{ridge} = (A\,x - b)^T (A\,x - b) + \alpha\, x^T x \tag{2.9}$$

Expanding and simultaneous simplification of equation 2.9 results in the following [6],

$$C_{ridge} = x^T A^T A\,x - b^T A\,x - x^T A^T b + b^T b + \alpha\, x^T x \tag{2.10}$$

$$= x^T A^T A\,x - x^T A^T b - x^T A^T b + b^T b + x^T \alpha I\, x \tag{2.11}$$

$$= b^T b - 2\, x^T A^T b + x^T A^T A\,x + x^T \alpha I\, x \tag{2.12}$$

$$C_{ridge} = b^T b - 2\, x^T A^T b + x^T (A^T A + \alpha I) x \tag{2.13}$$

The objective function in 2.7 can be minimized by taking the partial derivative of 2.13 with respect to "x" .

Minimization condition $\implies$ the gradient of the objective function must be equal to zero.

$$\frac{\partial C_{ridge}}{\partial x} = 0 \tag{2.14}$$

$$\implies \frac{\partial C_{ridge}}{\partial x} = -2\, A^T b + \underbrace{2(A^T A + \alpha I)x}_{*} \tag{2.15}$$

$$\implies -2\, A^T b + 2(A^T A + \alpha I)x = 0 \tag{2.16}$$

$*$ indicates that the specific part of the equation was achieved by successfully applying matrix (symmetric) differentiation rule

Simplification of the equation 2.16 leads to the ridge regression solution i.e, "$\hat{x}_{ridge}$"

$$\hat{x}_{ridge} = (A^T A + \alpha\, I)^{-1} A^T b \tag{2.17}$$

where,

I = Identity matrix $(n \times m)$

$\alpha\, I$ = Ridge term

Advantages ridge term,

i) Facilitates invertibility of resultant matrix and it gets added to the principle diagonal

ii) consistently achieves a unique solution

The equation 2.8 can be interpreted geometrically as shown below,

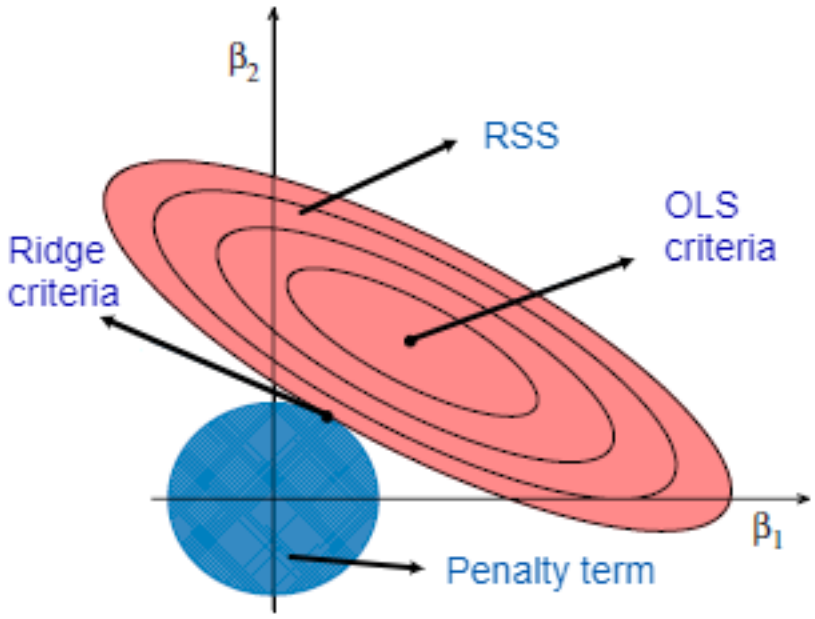


Figure 2.13: Geometric representation of ridge regression [10]

The figure 2.13 clearly depicts the aim of ridge (L2-regularization) regression i.e, minimization occurs simultaneously between the RSS (ellipse) and the penalty term (circle) mentioned in equation 2.8.The simultaneous minimization occurs at "$\hat{x}_{ridge}$" shown in equation 2.17.

2.3.2 LASSO

LASSO aims to minimize the ordinary least square with an added penalty term.In case of L1-regularization ,the penalty term is the sum of the absolute value of the regression coefficients. Hence LASSO is also known as the L1-regularization [4].

$$\underset{x\in\mathbb{R}^m}{\operatorname{argmin}} \sum_{i}^{n} \left(\sum_{j}^{m} a_{ij}x_j - b_i^2 \right)^2 + \alpha \sum_{j}^{m} |x_j| \tag{2.18}$$

The equation 2.7 can be represented in a simpler form as,

$$\underset{x\in\mathbb{R}^m}{\operatorname{argmin}} \underbrace{\|Ax-b\|_2^2}_{Residual} + \alpha \underbrace{|x|_1}_{Penalty} \tag{2.19}$$

where,

$b \in \mathbb{R}^n$ = Response vector
$A \in \mathbb{R}^{n\times m}$ = Predictor matrix
α = Regularization parameter

The first part of the derivation is similar to L2-regularization , In matrix notation equation 2.8 becomes,

$$C_{lasso} = (A\,x-b)^T(A\,x-b) + \alpha\ |x|_1 \tag{2.20}$$

Expanding and simultaneous simplification of equation 2.20 results in the following,

$$C_{lasso} = x^T A^T A\,x - b^T A\,x - x^T A^T b + b^T b + \alpha\ |x|_1 \tag{2.21}$$
$$= x^T A^T A\,x - x^T A^T b - x^T A^T b + b^T b + \alpha\,|x|_1 \tag{2.22}$$
$$C_{lasso} = b^T b - 2\,x^T A^T b + x^T A^T A\,x + \alpha\,|x|_1 \tag{2.23}$$

Next,taking the derivative of equation 2.23,we get,

$$\nabla C_{lasso} = -2A^T b + 2A^T Ax + \nabla(\alpha\,|x|_1) \tag{2.24}$$

Due the face that equation 2.24 consists of the term "$\nabla(\alpha\,|x|_1)$" ,sub-differential helps us to arrive at the final solution.But before we step into sub-differential,let us assume that the $A^T A$ is equal to I and multiply "2" to the penalty term.

Equation 2.24 becomes,

$$\nabla C_{lasso} = -2A^T b + 2x + 2\nabla(\alpha\,|x|_1) \tag{2.25}$$

Now,the sub-differential becomes,

$$\nabla(C_{lasso}) = \begin{cases} 2x - 2A^T b + 2\alpha, & x > 0 \\ [-2\alpha, 2\alpha] - 2A^T b, & x = 0 \\ 2x - 2A^T b - 2\alpha, & x < 0 \end{cases} \tag{2.26}$$

Breaking down each of the 3 conditions mentioned in equation 2.26,
Case 1: when $x > 0$

$$2x - 2A^T b + 2\alpha = 0 \tag{2.27}$$

Equation 2.27 must be satisfied.
Therefore,we get,

$$x = 2A^T b - \alpha \tag{2.28}$$

Case 2: when $x = 0$

$$0 \in [-2\alpha, 2\alpha] - 2A^T b \tag{2.29}$$

Therefore,we now have 2 sub-cases,i.e,

$$-2\alpha - 2A^T b < 0 \implies \alpha > -A^T b \tag{2.30}$$

$$2\alpha - 2A^T b > 0 \implies \alpha > A^T b \tag{2.31}$$

The sub-cases mentioned in equation 2.31 becomes ,

$$\alpha > \left|A^T b\right| \tag{2.32}$$

when $x = 0$

Case 3: when $x < 0$

$$2x - A^T b - 2\alpha = 0 \tag{2.33}$$

Equation 2.33 must be satisfied.
Therefore,we get,

$$x = A^T b + \alpha \tag{2.34}$$

The aforementioned cases helps us to arrive at the solution for LASSO and it summarized in the equation below,

$$\hat{x}_{lasso} = \begin{cases} 0, & x_j > \left|A^T b\right| \\ A^T b - sign(A^T b) \cdot \alpha, & x_j \leq \left|A^T b\right| \end{cases} \tag{2.35}$$

The equation 2.19 can be interpreted geometrically as shown below,

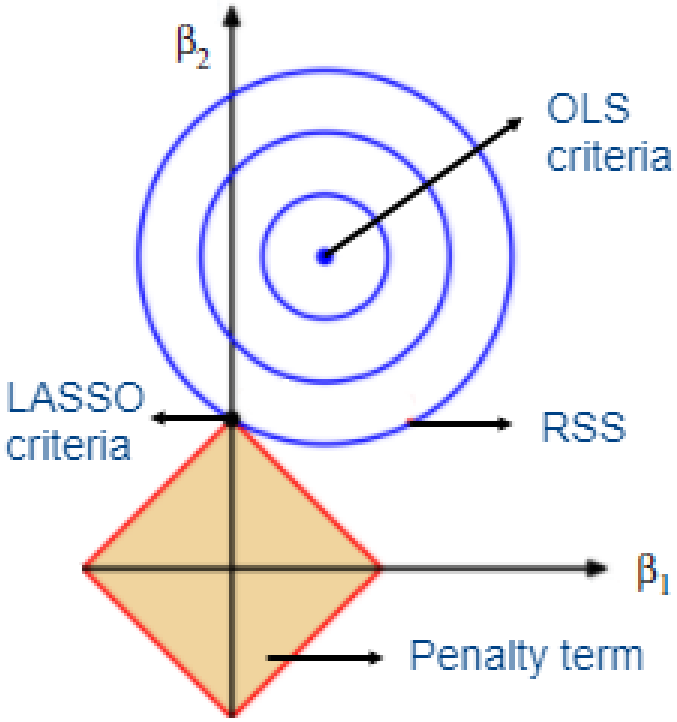


Figure 2.14: Geometric representation of LASSO regression [7]

The figure 2.14 clearly depicts the aim of LASSO (L1-regularization) regression i.e, minimization occurs simultaneously between the RSS (circle) and the penalty term (square) mentioned in equation 2.19.The simultaneous minimization occurs at "$\hat{x}_{lasso}$" shown in equation 2.35.

The differences between L1 and L2 regularization are summarized in the table below,

Properties	L1 regularization	L2 regularization
Robustness	Penalty term $\|x\|_1 \implies$ Absolute value of coefficients Outliers are affected linearly This method is more robust	Penalty term $\|x^2\|_2^2 \implies$ square of the coefficients Outliers are affected exponentially This method is less robust
Computational effort	Penalty term $\|x\|_1 \implies$ Non-differentiable term This methods requires more computational effort	Penalty term $\|x^2\|_2^2 \implies$ Closed form of solution Solution are obtained by using matrix form This method requires less computational effort
Sparsity	This method has the ability to shrink coefficients to zero Sparse solution.	This method spreads the error hindering sparsity

Table 2.1: Comparison between L1 and L2 regularization

The third and the most important regularization technique employed in this paper is total variation regularization.Hence,we are going to discuss in depth the theory and implementation process involved in TV regularization in chapters 3 and 4.

3 Total Variation Regularization

This chapter deals with theory and approach involved in TV regularization [13, 23] and it's importance is discussed in section 3.1 and 3.3 respectively.

The prior knowledge regarding the vital aspects about AWGN and regularization methods gained from sections 2.2.1 and 2.3 respectively helps us grasp the basics of TV regularization in this chapter a bit quicker.

3.1 Fundamentals and Approach

Total Variation Regularization is a common technique used in the field of engineering and scientific computing.The basic principle of this method[13] is to determine the derivative of function "f",which can be obtained by minimizing the following equation 3.1,

$$F(u) = \alpha R(u) + DF(Au - f) \tag{3.1}$$

where,

R(u)	= Regularization or penalty term
A(u)	= Anti-differentiation term which is given by
Au(x)	= $\int_0^x u$
α	= Regularization parameter
DF(Au-f)	= Data Fidelity term

Note:

- The role of regularization term is to penalize the irregularities in the "u" (solution)
- The role of the data fidelity term is to penalize the discrepancy between Au and f
- The role of the regularization parameter is to maintain a balance between regularization and data fidelity terms

Now,equation 3.1 is the general form of a regularization technique.Similar to the equation 3.1,in total variation regularization the derivative of the function is determined by minimizing the functional[13] of the length [0,L] as shown below,

$$F(u) = \frac{1}{2}\int_0^L \|Au - f\|_2^2 + \alpha \int_0^L |u'|_1 \tag{3.2}$$

where,

F = Functional defined on bounded variation $[0, L]$

f = Given function & $f \epsilon L^2$

u = Solution

α = Regularization parameter

Note:

- The regularization parameter plays a vital role in balancing the regularization term and data fidelity term as explained in equation 3.1.
- The choice of the regularization parameter is generally based on "eye-balling" technique for very small data set.
- Real world application usually deals with large data set,hence the "eye-balling" technique is computationally time consuming and reduces the efficiency of the process.

Due to the disadvantages of "eye-balling" technique ,automated techniques are designed and understood in order to improve the efficiency and reduce the computational time for large data set.

There are 3 techniques which are employed to overcome the aforementioned problem and they are,

1. L-curve method
2. Normalized Cumulative Periodogram (NCP) method
3. Generalized cross validation (GCV)

A brief explanation of the above 3 techniques are discussed in the section 4.2,where as the methodology and implementation of the same are discussed in section 4.1 .

3.2 Techniques involved in the determination of regularization parameter

L-curve method:

- In this method,the solution for each specific regularization parameter say,"u_α" and the residual vector $\|Au_\alpha - f\|$ are determined for various values of regularization parameter ($\alpha > 0$).
- A graph of $(log\,\|Au_\alpha - f\|\,, log\,|u_\alpha|)$ is plotted.The nature of this graph takes shape of an "L".Hence,this is coined as "L-curve" method.
- The flat part of the "L-curve" as seen in figure 3.1 signifies the dominance of regularization part and the steep part of the "L-curve" signifies the dominance of perturbation error.
- The "corner" of the "L-curve" as seen in figure 3.1 indicates the balance between the 2 errors,i.e,regularization and perturbation errors [21].

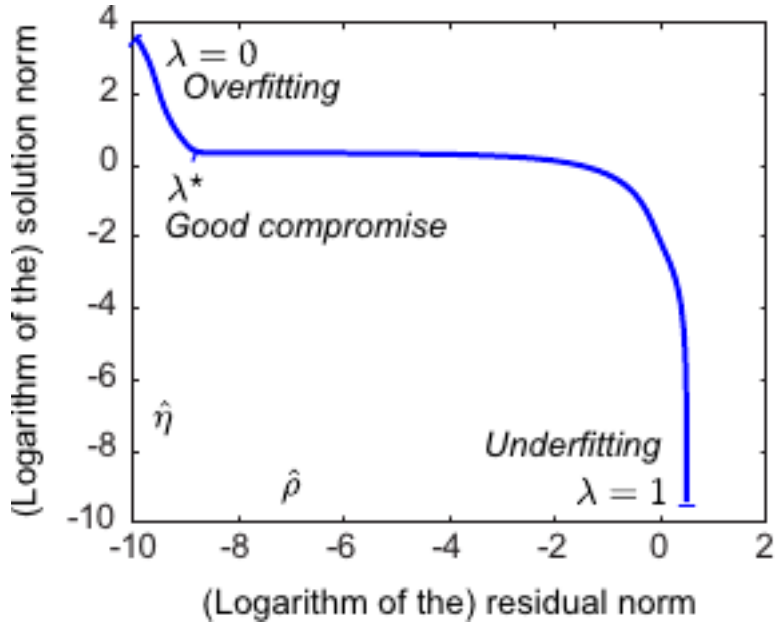


Figure 3.1: A graph showing the general form of the "L-curve" [5]

Normalized Cumulative Periodogram (NCP)**:**

- NCP solely requires the determination of residual vector as this methods aims to perform statistical analysis that focuses completely on the determined residual.

- NCP methods helps to isolate the noise from relevant information in the dataset.Regularization parameter which succeeds at achieving the aforementioned task is chosen as the optimal parameter [21].

General form of the NCP [20] is as follows,

$$\hat{r} = fft(r) \tag{3.3}$$

where,

r = Residual vector

$fft(r)$ = discrete Fourier transform of the residual vector

Then,the periodogram of residual vector ,"r" is represented as,

$$p = (|\hat{r_1}|^2, |\hat{r_2}|^2, |\hat{r}|^2, ... |\hat{r_q}|^2)^T \tag{3.4}$$

where,

q = $[n/2] + 1$

n = length of residual vector "r"

NCP for residual vector "r" is shown below,

$$c(r)_k = \frac{\|p(2:k+1)\|_1}{\|p(2:q)\|_1}, \qquad k = 1, 2, ...q-1 \tag{3.5}$$

where,

c(r) = Normalized cumulative periodogram of residual vector "r"

p = Periodogram or power spectrum of residual vector "r"

Generalized Cross Validation(GCV)**:**

- Generalized cross validation is an intuitive method to estimate the optimal regularization parameter.
- The aim of the GCV method is to calculate the target function "$\hat{G}(\alpha)$" as shown in equation 3.7.

$$f = Au + \eta \tag{3.6}$$

where,

f $=$ Noisy function
A $=$ Integral operator (with bounded kernel)
u $=$ Given data (or function)
η $=$ Gaussian noise of known standard deviation

$$\hat{G}(\alpha) = \frac{\frac{1}{N}\|f - AB_\alpha f\|_2^2}{[tr(I - AB_\alpha)]^2} \tag{3.7}$$

where,

B_α $= (A^T A + \alpha L(u^m))^{-1} A^T$
$L(u^m)$ $=$ Differential operator
tr $=$ Trace of a matrix
I $=$ Identity matrix
α $=$ Regularization parameter

- The target function "$G(\alpha)$" is determined for each regularization parameter

  $$\alpha = \alpha_1, \alpha_2, ...\alpha_n \qquad \alpha_n ... \alpha_2 > \alpha_1 > 0$$

  and then plot a graph of $(\alpha, G(\alpha))$.

- The optimal α value is selected for the corresponding minimum most value of "$G(\alpha)$" as seen in figure 3.2.

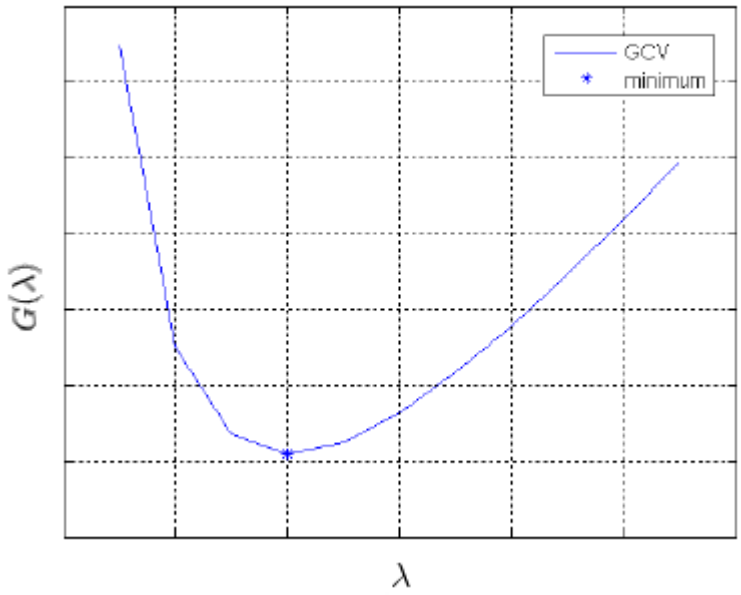


Figure 3.2: A graph showing the general form of generalized cross validation curve [24]

3.3 Importance of Total Variation Regularization

This section allows us to understand the importance employing TV regularization.

TV regularization can be applied to a process irrespective of the type of noise it contains. Hence, this method can be employed in various fields. Generally the regularization or penalty term shown in equation 3.1 is given by,

$$R(u) = \int_0^L |u'|^2 \tag{3.8}$$

where,
$R(u)$=Regularization or penalty term
Equation 3.8 constraints the minimizer to be continuous.Hence it leads to inaccurate differentiation of the given function.

In order to overcome and avoid the aforementioned difficulties ,the total variation regularization method represented in equation 3.2 is employed [13].The advantages of this methods are described below,

- This method helps to keep a check on the noise present in the data as it has a large total variation.

- Unlike in equation 3.8,total variation regularization considers the (jump) discontinuities.
- Total variation regularization facilitates the computation of discontinuous derivatives and characterizes the noisy data clearly.

The sections 3.1 and 3.3 provide the basic understanding of the total variation regularization.This is then translated into methodology of total variation regularization in the following Chapter 4.

4 Methodology of total variation regularization

In this chapter,we are going to discuss the process of numerical implementation involved in total variation regularization by using the brief understanding of the same from previous chapters.

4.1 Implementation of total variation regularization

In section 3.1, our aim is to determine the derivative of "f" by minimizing the equation 3.2. This can be achieved by using the gradient descent method. Gradient descent is a method employed to determine the local minimum of function. The aim of gradient descent is to initialize the step size that is proportional to the negative of the approximated gradient of the given function at each current iteration point. Therefore, for each iteration the gradient slowly tends towards the local minimum of the given function and once the convergence criteria is satisfied, the point at which this occurs is labeled as final value or minimum value of the given function as shown in figure 4.1.

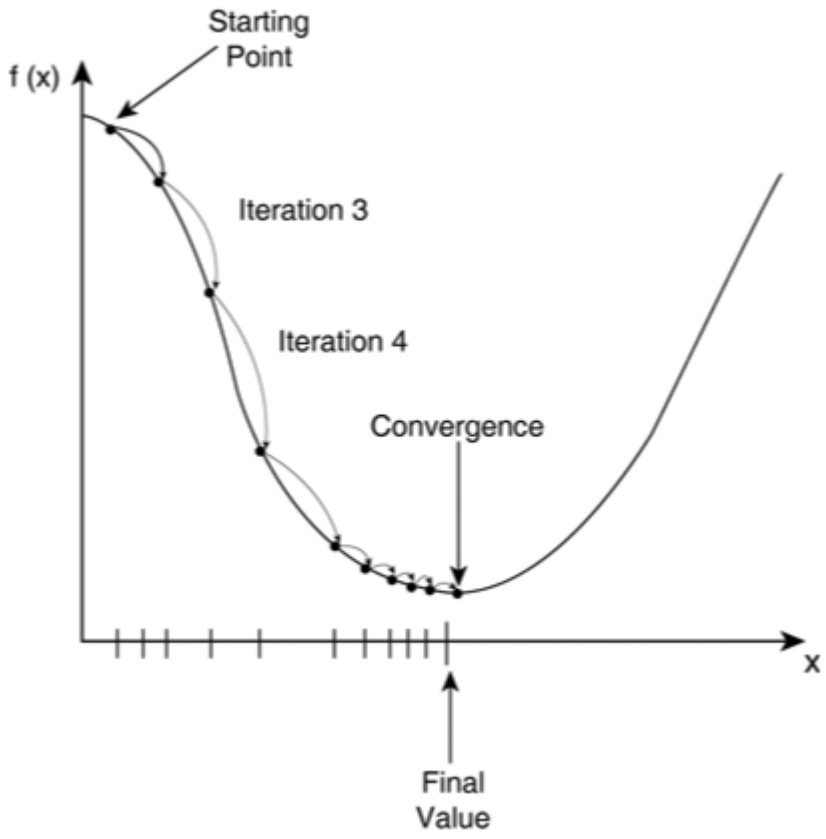


Figure 4.1: A graph depicting the convergence of a function using gradient descent method

Gradient descent method uses the principle of Euler-Lagrange differential equation.A general form of Euler-Lagrange equation is represented in the following equation,

$$J = \int f(t, y, \dot{y}) dt \tag{4.1}$$

where,

J $=$ Functional

f $=$ Given function depending on t,y,$\dot{y}$

$\dot{y} = \frac{\mathrm{d}y}{\mathrm{d}t}$

If the following Euler-Lagrange differential equation is satisfied,then "J" has a stationary value,

$$\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial f}{\partial \dot{y}}\right) = 0 \tag{4.2}$$

Note:

The time derivative in equation 4.2 can be replaced by space derivative as shown below,

$$\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial \dot{y}}\right) = 0 \tag{4.3}$$

The Euler-Lagrange differential equation using space derivative [30], as shown in equation 4.3 is applied to the total variation regularization functional in equation 3.2.This leads to the following system of equations in order to understand the gradient descent method.

$$\partial F(u) = A^T(Au - f) - \alpha \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{u'}{|u'|}\right) \tag{4.4}$$

The above equation can be solved by using steepest gradient descent method.In order to achieve the aim of method, the minimum can be determined by applying the following condition,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\partial F(u) \tag{4.5}$$

Therefore, we arrive at the final equation,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \alpha \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{u'}{|u'|}\right) - A^T(Au - f) \tag{4.6}$$

Note:

- In equation 4.6,in order to overcome the problem of getting undetermined solution (division by zero) the denominator $|u'|$ is substituted with "$\sqrt{|u'|^2 + \epsilon}$.The characteristics of ϵ are as follows,
  - $\epsilon > 0$
  - very small constant
- Numerical approximation of the solution in 4.6 is determined by employing the explicit method.
- Discretization of u_t in equation 4.6 is performed using forward difference method for a fixed time step Δt as shown below,

$$\frac{(u_{n+1} - u_n)}{\Delta t} \tag{4.7}$$

One of the disadvantage of using the gradient descent method for the determination of the minimum of the functional in equation 3.2 is the slow rate of convergence.Hence,to overcome

the aforementioned disadvantage the nonlinear differential operator $u \mapsto \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)\left(\frac{u'}{|u'|}\right)$ is substituted with a linear operator $u \mapsto \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)\left(\frac{u'}{|u'_n|}\right)$ for each iteration in equation 4.6.
This method is known as lagged diffusivity and it utilizes two types of algorithm,

- for smaller problems
- for larger problems

4.1.1 Lagged diffusivity fixed point method for smaller problems

We are now going to discuss the method of lagged diffusivity for smaller problems [13] in detail.The three principles of this method are,

i) "u" is constructed on an uniform grid i.e,
$\{x_i\ \}_0^L = \{0, \Delta x, \Delta 2x, \Delta 3x.....L\ \}$

ii) Derivative of "u" is determined halfway between the grid using the forward difference method i.e,
$Du(x_i + \Delta x/2) = u(x_{i+1}) - u(x_i)$

iii) Similarly,the integral of "u" is determined halfway between the grid using trapezoidal rule.

The table 4.1 summarizes the formula employed during various phases in the lagged diffusivity fixed point algorithms and the pseudo code is presented in 4.1.1,

Variable names	**Formula**
E_n	$\sqrt{((u_n(x_i) - u_n(x_{i-1})^2 + \epsilon)}$
L_n	$\Delta x D^T E_n D$
H_n	$K^T K + \alpha L_n$
g_n	$K^T(Ku_n - f) + \alpha L_n u_n$

Table 4.1: A table summarizing the important formula required in the lagged diffusivity method

By utilizing the formula in table 4.1,the updated value shown in equation 4.8 forms the solution to equation 4.9 required over each iteration point,

$$s_n = u_{n+1} - u_n \tag{4.8}$$

where,

$$s_n = -H_n^{-1} g_n \tag{4.9}$$

Algorithm 4.1.1: Lagged diffusivity fixed point pseudo code

1. Niter $\leftarrow$ Initialize %Specify number of iteration
2. $\alpha \leftarrow$ Initialize %Specify regularization parameter
3. $u \leftarrow$[0;diff (Data);0] % Naive derivative
4. for n $\leftarrow$ 1:Niter

   $g_n \leftarrow K^T(Ku_n - f) + \alpha L_n u_n$ % Gradient

   $H_n \leftarrow K^T K + \alpha L_n$ % Cost function (Hessian approximation)

   $s_n \leftarrow -H^{-1} g_n$ % Determined using preconditioned conjugate gradient

   $u_n \leftarrow u_n - s_n$ % Update

   end

Note:

- The second and third principles of lagged diffusivity method leads to the formation of a differentiation matrix "D" and "A" of the size $L \times (L+1)$ respectively.
- The advantage of this method is that it prevents the need to deal with the boundary conditions required during differentiation process.
- Computationally ,this method provides better results.

4.1.2 Lagged diffusivity for large problems

The lagged diffusivity for large problems differs from the small problems in following aspects,

i) The differentiation matrix "D" is constructed as,
   - Sparse and square matrix of size $L \times L$
   - Derivative of "u" is determined using the forward difference method
   - The determination of the differentiation matrix "D" also requires the knowledge of periodic boundary condition

ii) The integral of "u" uses the predefined cumulative sum operator from MATLAB®.

iii) This method uses the same formula in table 4.1 and concept as explained in section 4.1.1 but the only major difference is the inclusion of the periodic boundary condition while calculating the "E_n" matrix.

iv) Preconditioned conjugate gradient method is employed to solve the equation 4.9 in large problems.

In this thesis ,we are mainly going to concentrate on the lagged diffusivity for small problems.

A very crucial element which has to be addressed in total variation regularization is the selection of the regularization parameter "α" in equation 3.2. The selection process of the "α" is clearly discussed in the following section.

4.2 Selection of regularization parameter

In the field of engineering "the eye-balling" technique does not prove to be computationally viable mainly due to large data set and other complexities.Therefore,the aim is to design algorithms to determine the optimal regularization parameter to help denoise the dataset for further numerical analysis.

In this section we are going to discuss in depth the selection process of regularization parameter by using the brief understanding already explained in section 3.2.

4.2.1 L-curve method

The "L-curve" method is implemented by using the principles of the total variation regularization as explained in section 4.1.

The solution is determined for each of the user defined regularization parameter,

$$\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, ...\alpha_n \ \}$$

This solution is denoted as "u_α".Then, for each of the respective "u_α", residual between the solution and given function is estimated.This is represented by "$Au_\alpha - f$".

Finally,the L1-norm of the solution i.e,"$|u_\alpha|_1$" and the L2-norm of the residual ,i.e," $\|Au_\alpha - f\|_2^2$" are determined.A graph of ($\|Au_\alpha - f\|$,$|u_\alpha|$) is plotted.This results in graph characterized by the "L-curve".

In order to extract the optimal regularization parameter from the aforementioned graph, the curvature is determined using the following equation,

$$\hat{C} = 2\frac{\xi\rho}{\xi'}\left[\frac{\alpha^2\xi'\rho + 2\alpha\xi\rho + \alpha^4\xi\xi'}{(\alpha^2\xi^2 + \rho^2)^{\frac{3}{2}}}\right] \tag{4.10}$$

where,
$\hat{C}$ = Curvature
$\xi = |u_\alpha|_1$
$\rho = \|Au_\alpha - f\|_2^2$
$\xi' = \frac{\partial\xi}{\partial\alpha}$

The entire process of "L-curve" method is summarized in the following flowchart.

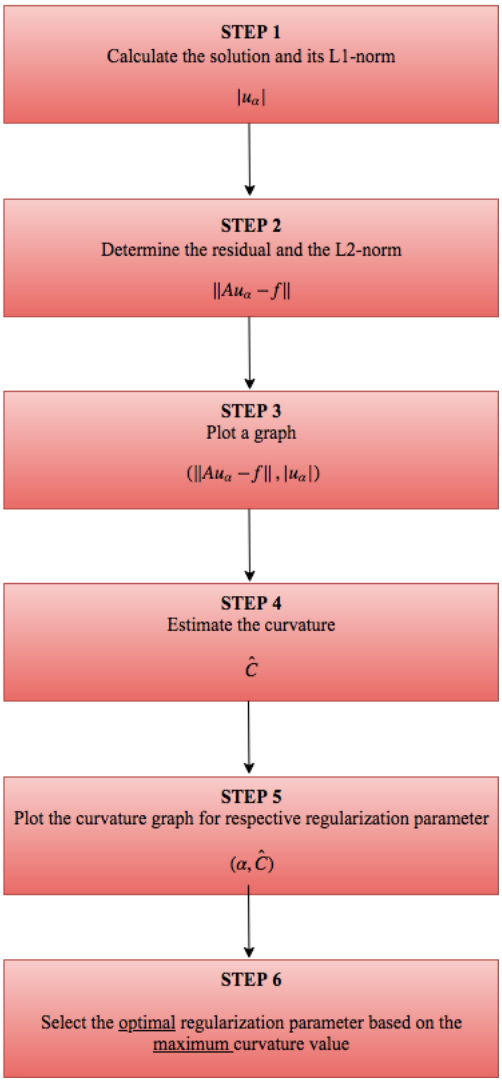


Figure 4.2: A flowchart describing the process involved in L-curve method

Note:
The equation 4.10 deals with the determination of "ξ'" which signifies the derivative with respect to "α", hence making it difficult to calculate the curvature values.To overcome the aforementioned complexity we employ an alternative method (based on coordinates) which is described below,

$$
\begin{aligned}
a &= \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \\
b &= \sqrt{(x2 - x3)^2 + (y2 - y3)^2} \\
c &= \sqrt{(x3 - x1)^2 + (y3 - y1)^2} \\
A &= 0.5 \cdot |(x1 - x2) \cdot (y3 - y2) - (y1 - y2) \cdot (x3 - x2)|
\end{aligned}
$$

$$\hat{C} = \left(\frac{4A}{a \cdot b \cdot c}\right) \tag{4.11}$$

4.2.2 Normalized Cumulative Periodogram (NCP)

NCP is a method that concentrates mainly on the residual vector "r",where

$$r = Au_\alpha - f$$

Based on the perception of the residual vector ,we can draw the following conclusion,

i) All information in the given function "f" is not extracted when "α" is large.

ii) Only noise remains in the residual vector when "α" is small.

As explained in section 3.2,we have the following system of equations,

$$\hat{r} = fft(r) \tag{4.12}$$

where,

r = Residual vector

$fft(r)$ = discrete Fourier transform of the residual vector

Then, periodogram of the residual vector ,"r" is given as,

$$p = (|\hat{r_1}|^2, |\hat{r_2}|^2, |\hat{r_3}|^2, ... |\hat{r_q}|^2)^T \tag{4.13}$$

where,

q = [n/2]+1

n = length of residual vector "r"

NCP for the residual vector "r" is shown below,

$$c(r)_k = \frac{\|p(2:k+1)\|_1}{\|p(2:q)\|_1}, \qquad k = 1, 2, ...q-1 \tag{4.14}$$

where,

c(r) = Normalized cumulative periodogram of residual vector "r"

p = Periodogram or power spectrum of residual vector "r"

The main objective of NCP method is to determine whether the residual contains only "white noise". A graph of (k,$c(r)$) is plotted. In order to verify if the objective of NCP is met,the graph of NCP clearly indicates a straight line with coordinates $(0, 0)$ and $(q, 1)$. *Note*:
The objective of NCP is extensively documented in the results section 5 which facilitates in better understanding the entire process.

4.2.3 Generalized Cross Validation (GCV)

The GCV method is mainly used in nonlinear regularization methods to estimate optimal regularization parameter as it successfully considers/addresses even nonlinear terms. This is shown below,

$$F(u) = \alpha \int_0^L |u'|_1 + \frac{1}{2} \int_0^L \|Au - f\|_2^2 \tag{4.15}$$

Minimization of the above equation 4.15 with respect to "u" results in the Euler-Lagrange equation as follows,

$$\frac{\partial F(u)}{\partial u} = \alpha \frac{\mathrm{d}}{\mathrm{d}x} \left(\frac{u'}{|u'|} \right) - A^T(Au - f) = 0 \tag{4.16}$$

Note:

$\frac{\mathrm{d}}{\mathrm{d}x} \left(\frac{u'}{|u'|} \right)$ represents the divergence i.e, $div \left(\frac{u'}{|u'|} \right)$.

The nonlinear part of equation 4.16 can be expressed (similar to the second equation in table 4.1 of lagged diffusivity method) as,

$$div \left(\frac{u'^{m+1}}{|u'^{m+1}|} \right) = \underbrace{L(u^m)}_{part\ 1} u^{m+1} \tag{4.17}$$

where,

m = arbitrary iteration step

In equation 4.17, the "part 1" is expressed below in 1D,

$$L(u^m) = Dx^T \cdot Du^m \cdot Dx \tag{4.18}$$

In 2D as,

$$L(u^m) = Dx^T \cdot Du^m \cdot Dx + Dy^T \cdot Du^m \cdot Dy \tag{4.19}$$

where,

Dx = Difference operator in x-direction

Du^m = Diagonal matrix with principle axis as $\left(\frac{1}{\sqrt{(u')^2+\epsilon}}\right)$

Dy = Difference operator in y-direction

ϵ = Small constant (user-defined)

With the help of above understanding ,we again apply the fixed point iteration method in 4.17 to evaluate the equation 4.16 and this leads to ,

$$(A^T A + \alpha L(u^m)u^{m+1} = A^T f \tag{4.20}$$

Finally ,the improvised GCV can be formulated by using the 4.18 and 4.20 as,

$$\hat{G}(\alpha^{m+1}) = \frac{\frac{1}{N}\left\|f - A\hat{B}_\alpha f\right\|_2^2}{[tr(I - A\hat{B}_\alpha)]^2} \tag{4.21}$$

where,

$\hat{B}_\alpha$ = $(A^T A + \alpha L(u^m))^{-1}A^T$

$L(u^m)$ = Differential operator

tr = Trace of a matrix

I = Identity matrix

α = Regularization parameter

The following flowchart facilitates the better understanding of improvised GCV and sheds light on the steps involved to achieve the aim i.e, to calculate the target function $\hat{G}(\alpha^{m+1})$,

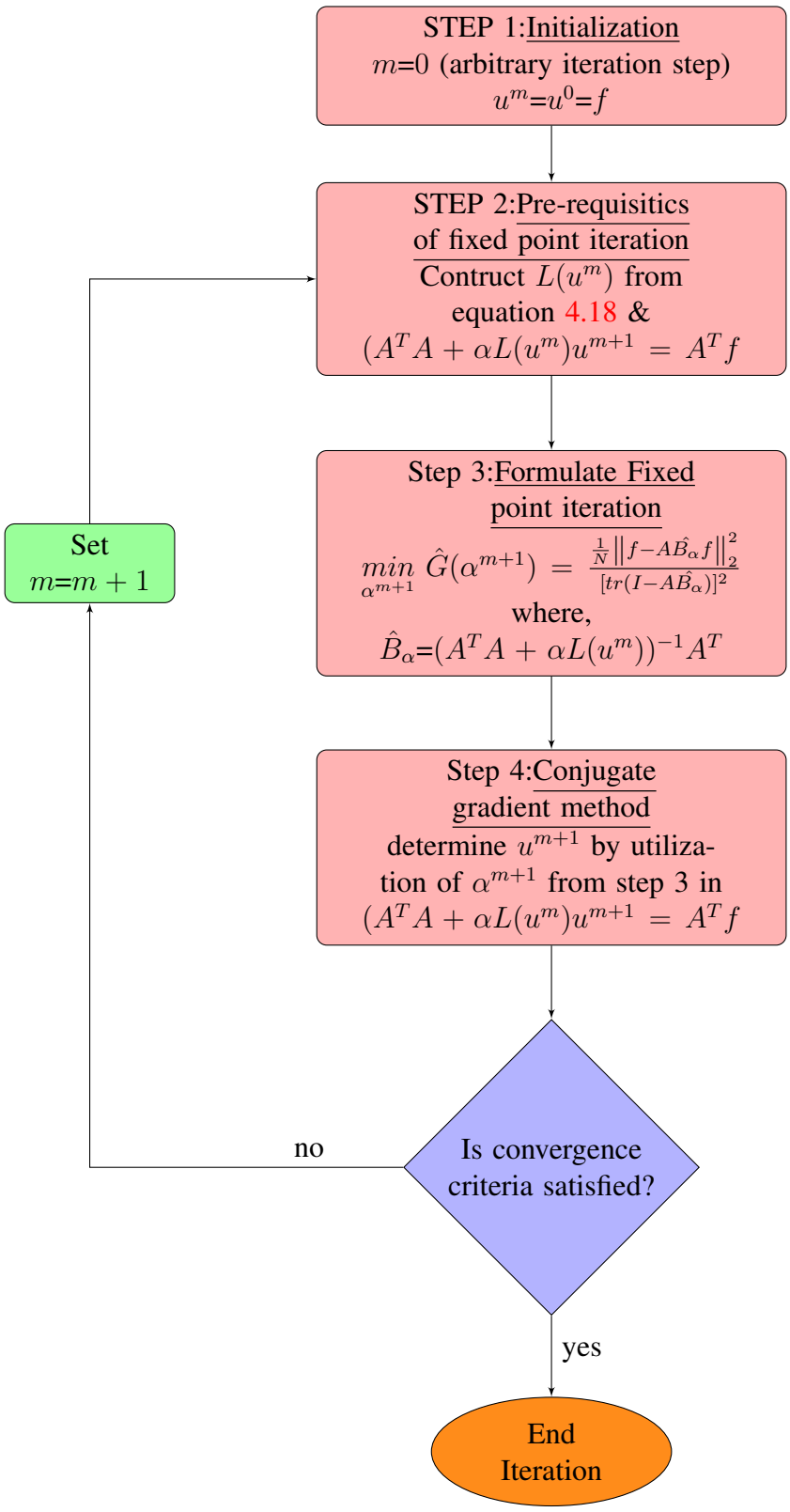


Figure 4.3: A flowchart describing the process involved in improvised GCV

4.2.4 Consideration of mean squared error

In the field of statistics, MSE is simply defined as the average of the square of the errors .The error is the difference between estimator and estimate of a required attribute in a process.The MSE is calculated using the following equation,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{4.22}$$

where,

$\left(Y_i - \hat{Y}_i\right) = Error$

n = total number of samples

Properties of MSE:

i) MSE measures the quality of the estimator.

ii) MSE values are always positive (MSE>0).

iii) In theory,MSE=0 indicates a perfect fit between the predicted ($\hat{Y}$) and the true values (Y) . But in practical applications this condition is virtually impossible to achieve,hence the values closer to zero signifies better accuracy between $\hat{Y}$ and Y.

iv) The squaring of the error terms ensures non-negative values but provides higher weights to large error.

The simplicity of MSE method makes it desirable to implement in numerous engineering fields as well , hence we employ this method in total variation regularization.

The basic definition of the MSE forms the basis for its implementation in TV regularization.In order to obtain the estimator value ($\hat{Y}$) in equation 4.22 the optimal regularization parameter ($\alpha_{optimal}$) needs to determined.Therefore, we utilize the gradient descent method.
The central difference method (with error term) shown in equation 4.23 has many advantages over forward and backward difference method,a few are mentioned below ,

i) Error (truncation) is of the order $O(h^2)$ whereas for both central and backward difference is of the order $O(h)$.This implies, for sufficiently small and fixed value of "h(step size)" the errors are smaller for central difference

ii) No loss in efficiency as all 3 forward ,central ,backward use same number of points.

$$f' = \underbrace{\frac{f(x+h) - f(x-h)}{2h}}_{A} + \underbrace{O(h^2)}_{B} \tag{4.23}$$

where,
Part A of the equation $\Rightarrow$ the central difference formula
Part B of the equation $\Rightarrow$ the error term

Due to the aforementioned advantages,we use the central difference in gradient descent method to find the local gradient during each iteration.The process to obtain the optimal regularization using MSE is summarized in the following process chart,

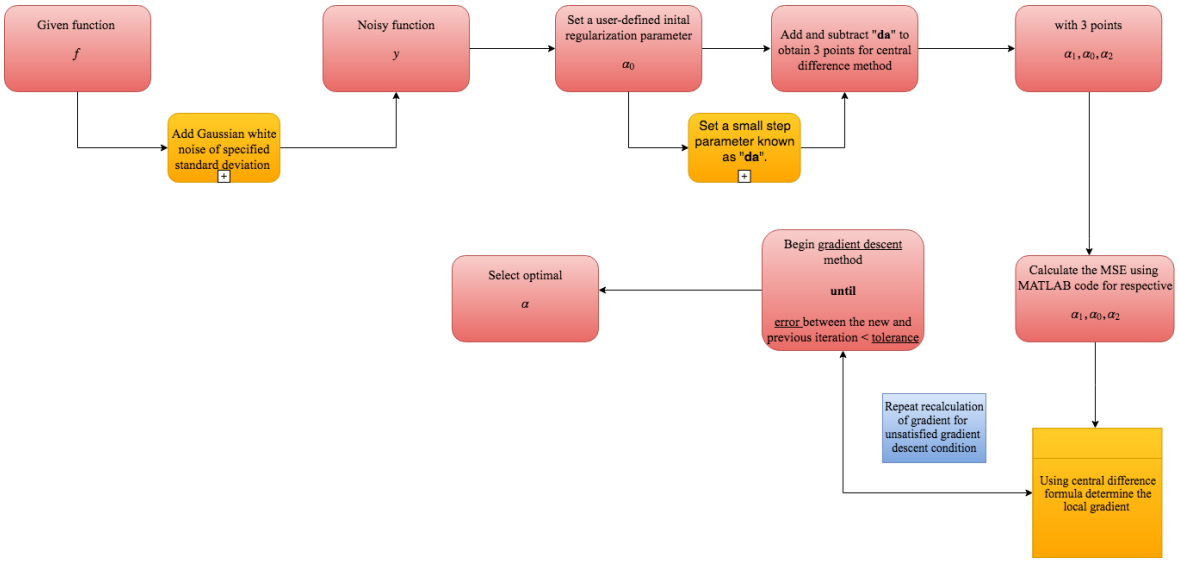


Figure 4.4: A process chart briefly describing the steps involved in gradient descent using total variation regularization

The calculated optimal regularization parameter (α_{opt}) is in-turn used to determine the derivative with the help of TV regularization.Finally,the MSE is estimated using the equation 4.25,

$$error = d_{(known)} - d_{(TV)} \tag{4.24}$$

where,

$d_{(known)}$ = Derivatives values of given function
$d_{(TV)}$ = Derivatives values from TV regularization (using α_{opt})

$$MSE = mean(error) \tag{4.25}$$

4.2.5 Data-driven method (sparse regression)

This method mainly deals with obtaining the governing equation (PDE's) that exists in both time and space (spatiotemporal). Time series data collection forms the main principle behind this method. This principle of data-driven (sparse regression) helps us to extract the coefficients at a fixed spatial location very efficiently. Hence this proves to be of vital importance when dealing with large scale data (e.g. measurement data) [26].

Following are some of the steps in terms of expressions and this is summarized in the flowchart shown in figure 4.5.

A nonlinear form of PDE are represented in the following equation,

$$u_t = N(u, u_x, u_{xx}, u_{xxx}, \ldots . x, t, \mu) \tag{4.26}$$

where,

$x(subscript)$ = Represents partial derivative with respect to time or space
N = unknowns
μ = Encompasses the parameters (coefficients)

The equation 4.26 is discretized into the following system,

$$U_t = \Theta(U, Q)\xi \tag{4.27}$$

where,

U = right hand side mentioned in equation 4.26
$\Theta(U, Q)$ = A matrix consisting of derivative parameters
Q = consists of the corresponding parameters/coefficients of the PDE.

The data-driven method uses brute-force search along all the combinations in $\Theta(U, Q)$ coupled with the sparse regression technique to provide the approximate solution of the the governing equation (PDE).

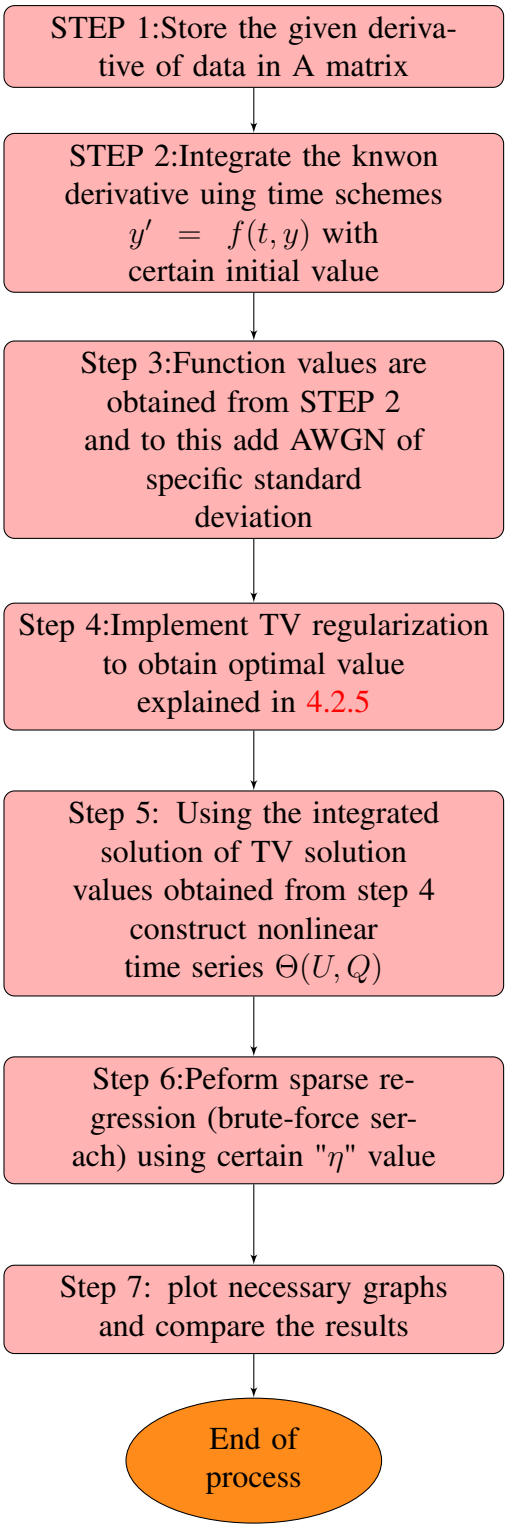


Figure 4.5: A flowchart describing the process involved in data-driven (sparse regression) method

5 Results

The important terminology and the methods explained in previous chapters 1,2,3 and 4 provide a strong foundation (or platform) to apply the acquired knowledge on various test functions.

This chapter is segregated into various case studies,each dealing with different aspects and its results are summarized.

Note: All the graphs were generated using MATLAB®.

5.1 Determination of the regularization parameter

In order to achieve satisfactory results using TV regularization,we must first determine the optimal regularization parameter.Various methods are employed to achieve the aforementioned goal that are presented in this section.To begin with,we consider and perform various methods on three different test functions.They are described below and summarized in tables 5.1,5.2 and 5.3 respectively.

Gaussian white noise of a specific standard deviation is added to a function "f" which yields a noisy "y" function.We then implement various methods to determine the optimal regularization parameter which helps in extracting vital information from a noisy dataset (E.g. measurement data)

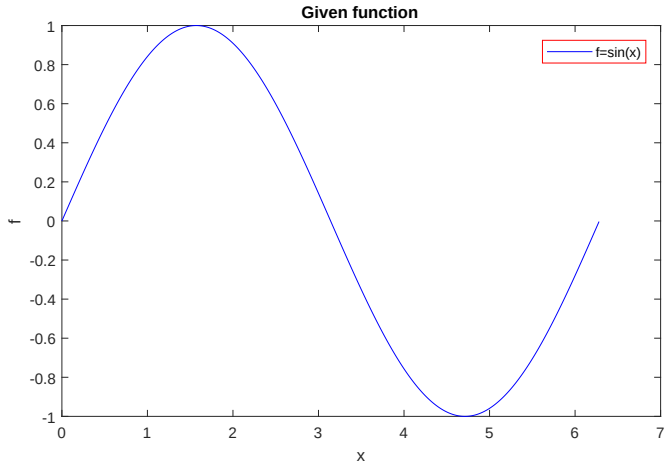
5.1.1 Test function 1

Data points	$x = 0 : 0.01 : 2\pi$
Given function	$f = \sin(x)$
Standard deviation of AWGN	0.05
Noisy function	$y = f + \eta$

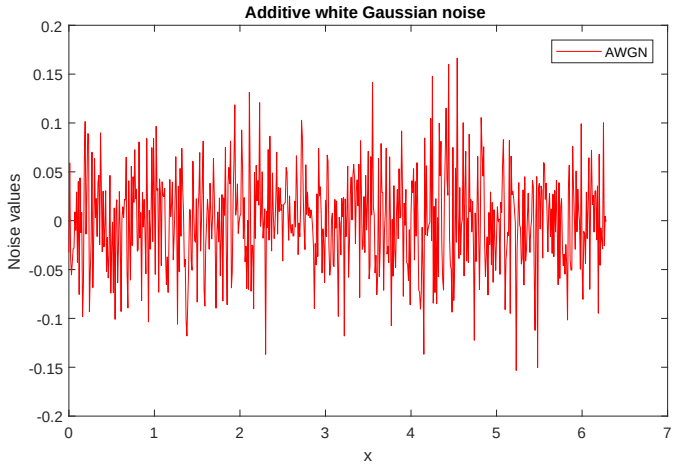
Table 5.1: Given information for test function 1

Note:

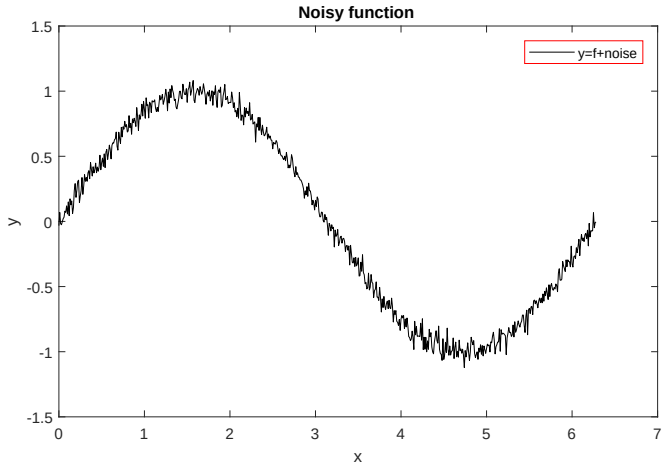
The variable "η" is AWGN noise (vector) which is generated using "randn" a MATLAB® syntax.



(a)



(b) Additive white Gaussian noise



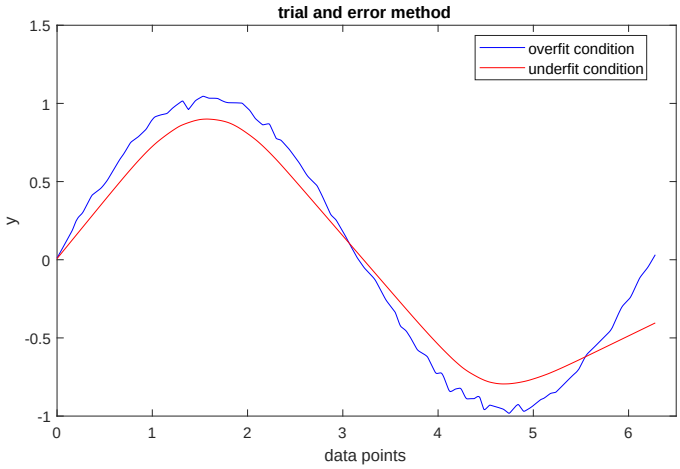
(c) Noisy function

Figure 5.1: Graphs depicting the respective information provided in table 5.1

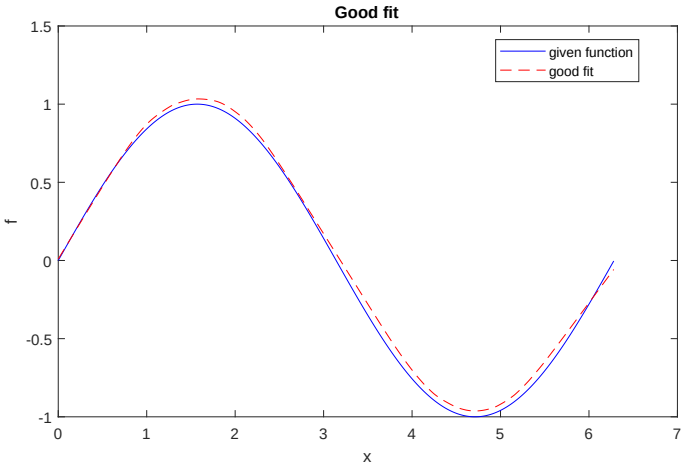
5.1.1.1 Eye-balling method

It is also known as trial-and-error method. The user needs to start with an initial value and check if the solution is either overfit or an underfit condition. This suggest whether one must increase or decrease the choice of regularization parameter. This process continues until the a certain value that provides a good fit is selected. This does not ensure that the selected parameter is optimal.

This method proves to be tedious for a user when dealing with large scale data and also an unfeasible technique. The results of a few parameters are shown below,



(a)



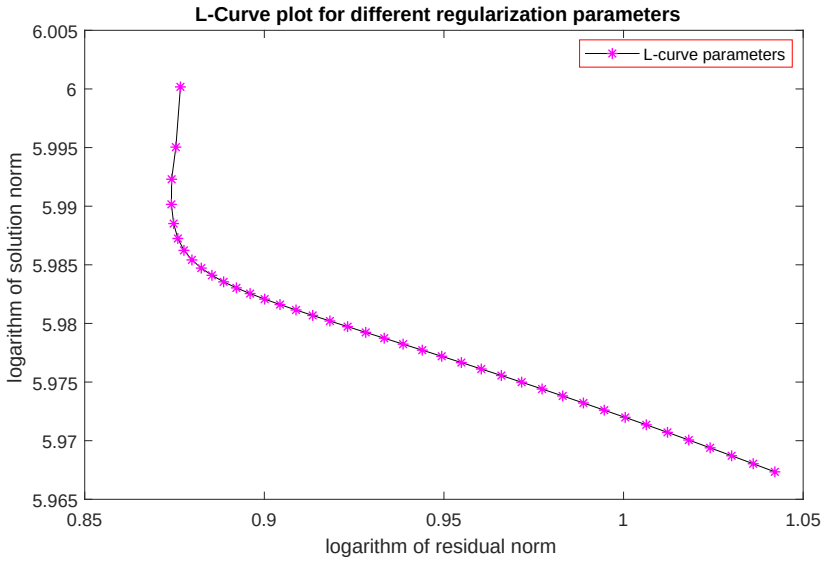
s

(b)

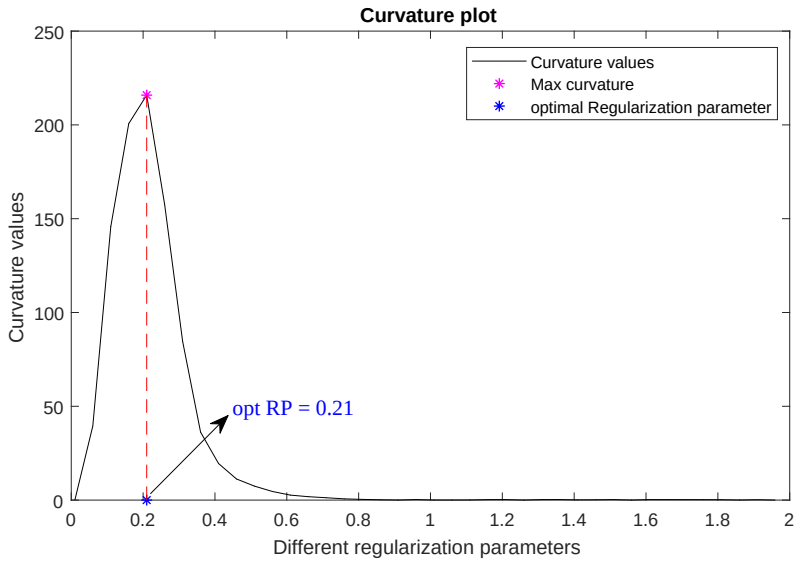
Figure 5.2: Graphs depicting the overfit and underfit condition (a) and good fit (b) for test function 1

5.1.1.2 L-curve method

We now implement the L-curve method to determine the optimal regularization parameter that helps to balance the under-fit and over-fit condition.The results are shown below,



(a) A graph representing the L-curve graph



(b) A graph depicting curvature graph

Figure 5.3: L-curve results for test function 1

5.1.1.3 NCP method

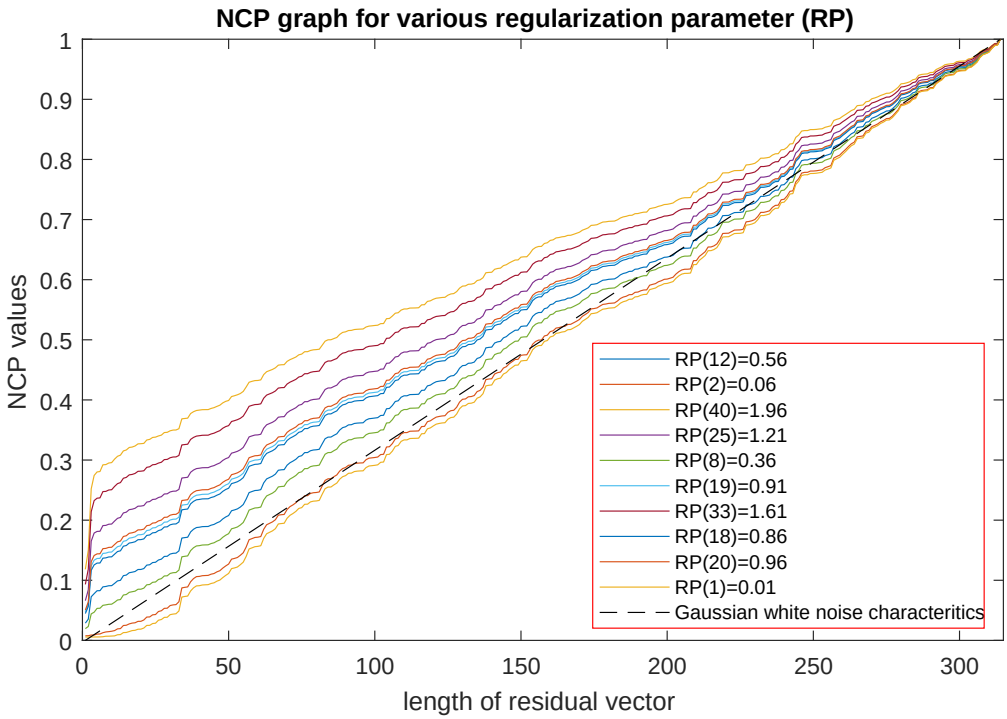


Figure 5.4: A graph showing the behavior of NCP values for ten different regularization parameter

The figure 5.4 makes our inference regarding the NCP a bit tedious.Hence we isolate the optimal regularization parameter from our entire set of regularization parameter and this is shown in figure 5.5.

The selection of the optimal value can be performed by determining the L2-norm between Gaussian white noise characteristic line (dotted black line in figure 5.5) and various NCP values for respective regularization parameter. One with the least L2-norm can be selected as optimal because this particular line (blue line in figure 5.5) lies closest to the desired.

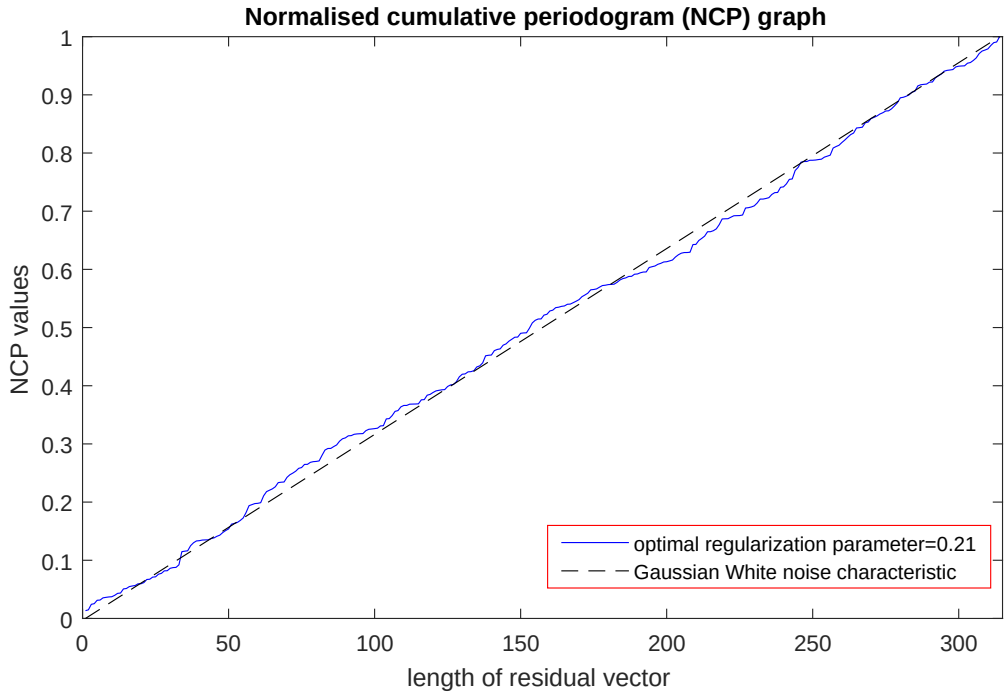


Figure 5.5: A graph showing the behavior of NCP values for the optimal regularization parameter

5.1.1.4 GCV method

As seen in figure 5.6, selection of optimal value is comparatively easier as the aim of GCV method is to minimize the GCV curve. Therefore the regularization parameter corresponding to minimum GCV value is selected as our optimal value (blue marker in figure 5.6).

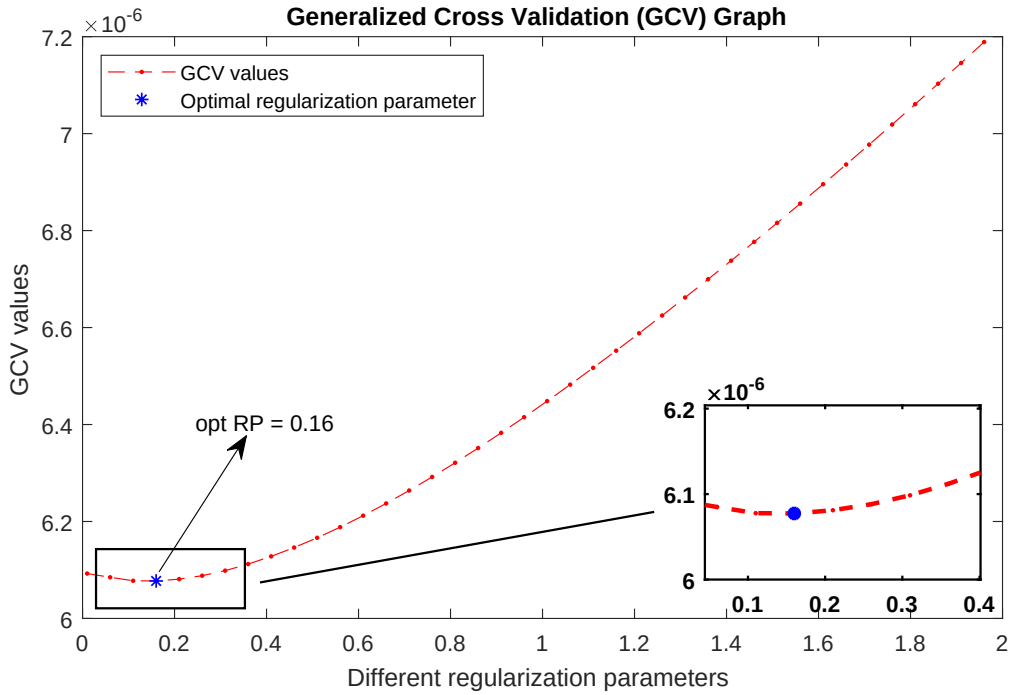


Figure 5.6: A graph showing the behavior of GCV values for various regularization parameter

Note: Magnification of the graph was performed using a MATLAB® code [16].

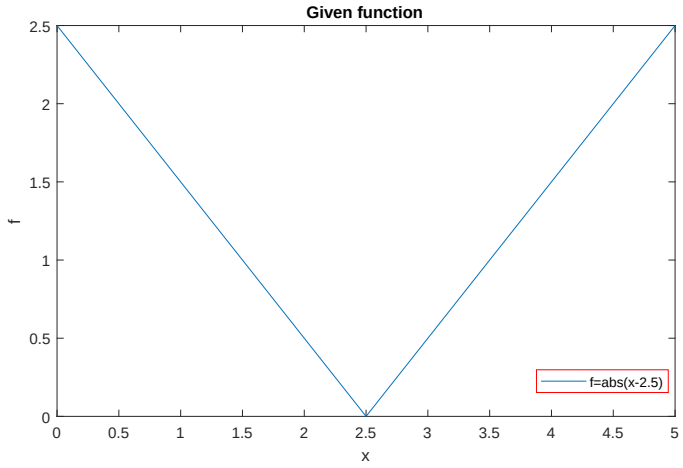
5.1.2 Test function 2

Data points	$x = 0 : 0.01 : 5$
Given function	$\lvert x - 2.5 \rvert$
Standard deviation of AWGN	0.05
Noisy function	$y = f + \eta$

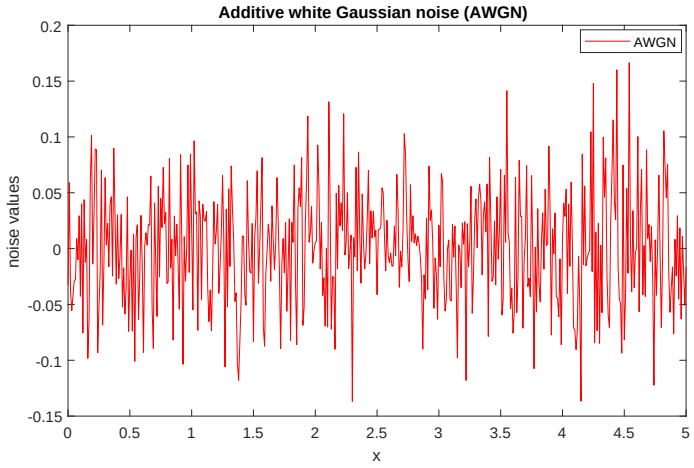
Table 5.2: Given information for test function 2

Note:

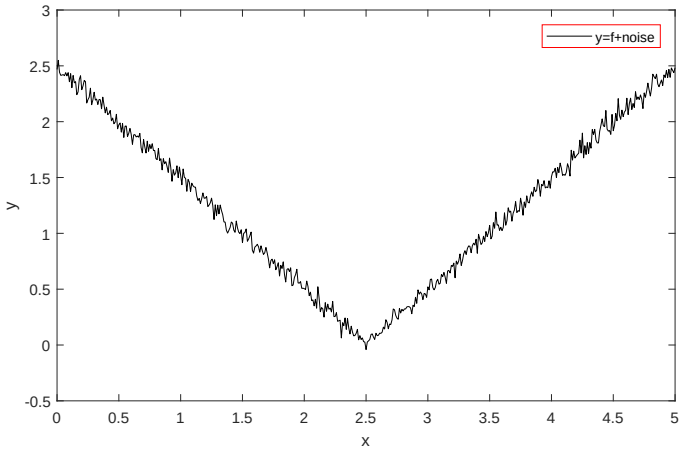
The variable "η" is AWGN noise (vector) which is generated using "randn" a MATLAB® syntax.



(a)



(b) Additive white Gaussian noise

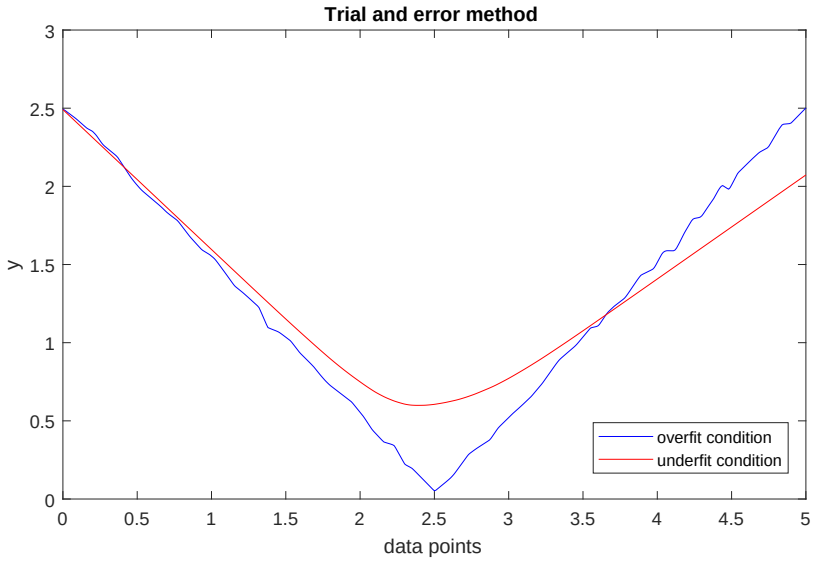


(c) Noisy function

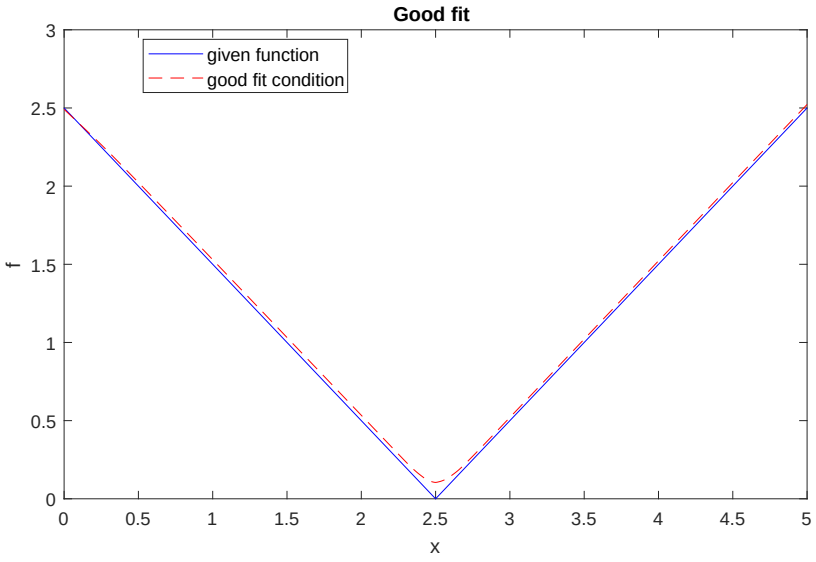
Figure 5.7: Graphs depicting the respective information provided in table 5.2

5.1.2.1 Eye-balling method

Similar to 5.1.1.1, this method is computationally and economically not viable.Some of the results are shown below,



(a)



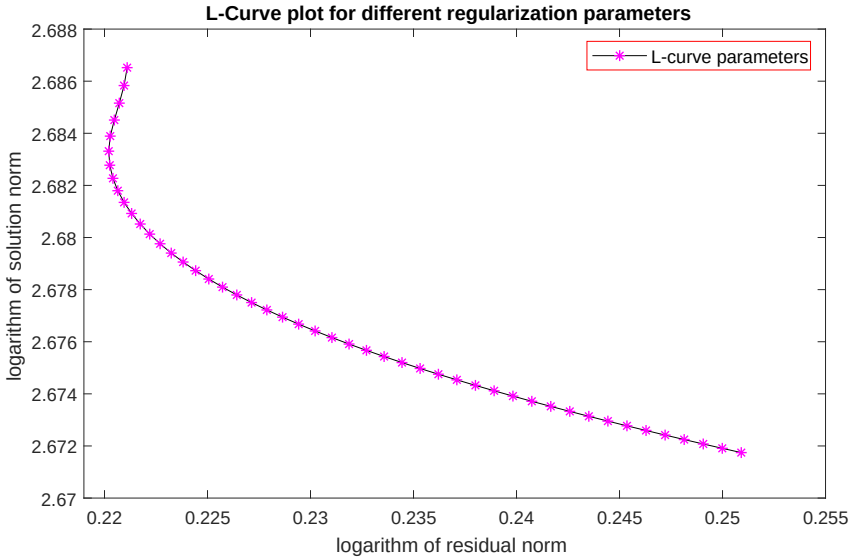
s

(b)

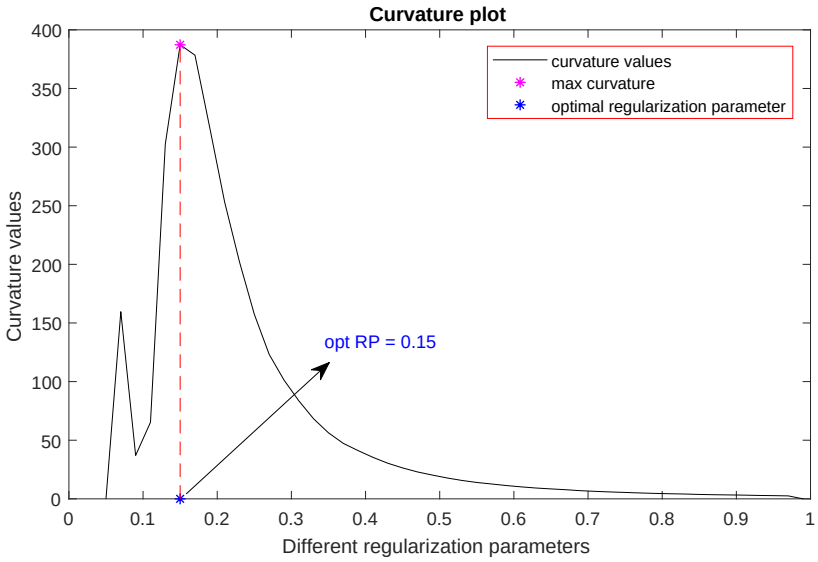
Figure 5.8: Graphs depicting the overfit and underfit condition (a) and good fit (b) for test function 2

5.1.2.2 L-curve method

Similar to 5.1.1.2,we present below the results for L-curve for test function 2.



(a) A graph representing the L-curve graph



(b) A graph depicting curvature graph

Figure 5.9: L-curve results for test function 2

5.1.2.3 NCP method

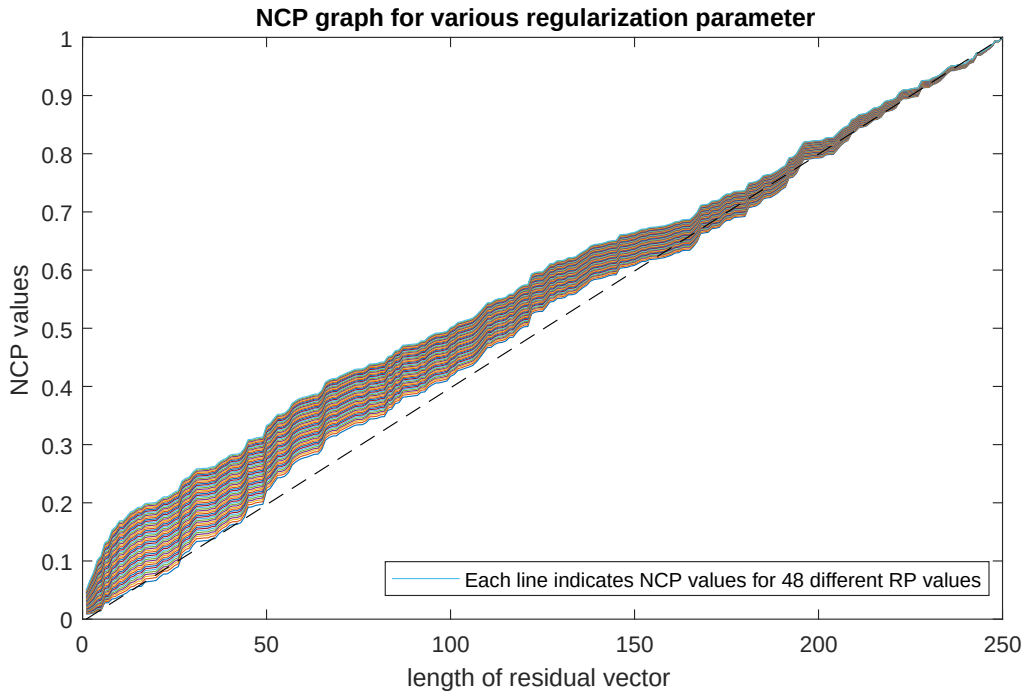


Figure 5.10: A graph showing the behavior of NCP values for ten different regularization parameter

As seen in figure 5.10,selection of optimal value is complex as the NCP values for closest to the Gaussian white noise characteristic (dotted black line in figure 5.11) yields a very small value in this case.This is shown below,

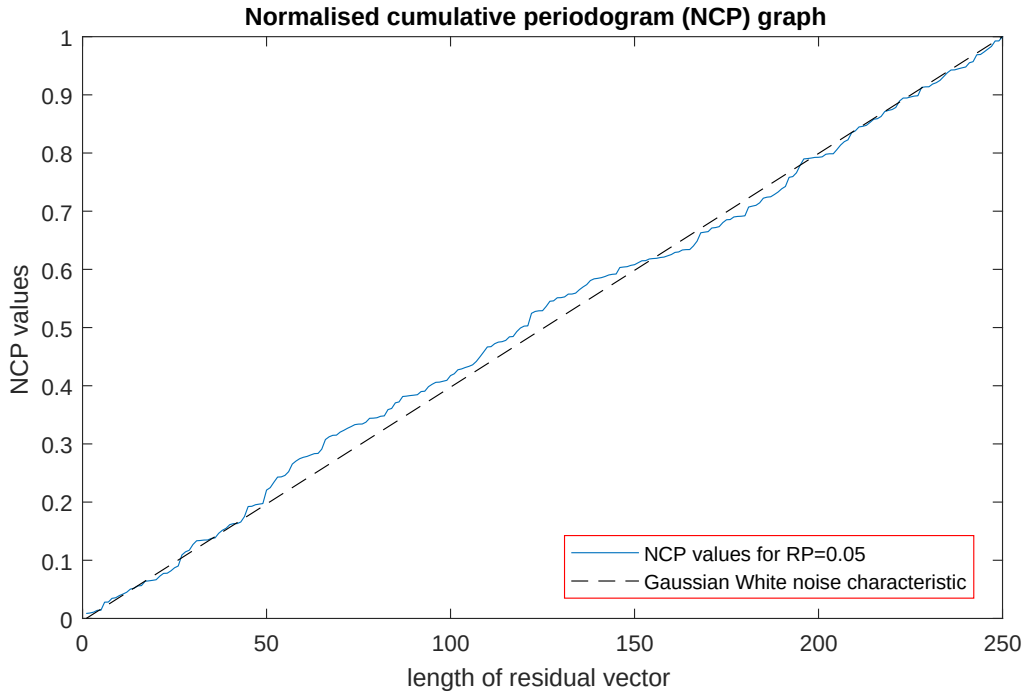


Figure 5.11: A graph showing the behavior of NCP values for the optimal regularization parameter

Hence we select the optimal values from the other methods discussed for test function 2.

5.1.2.4 GCV method

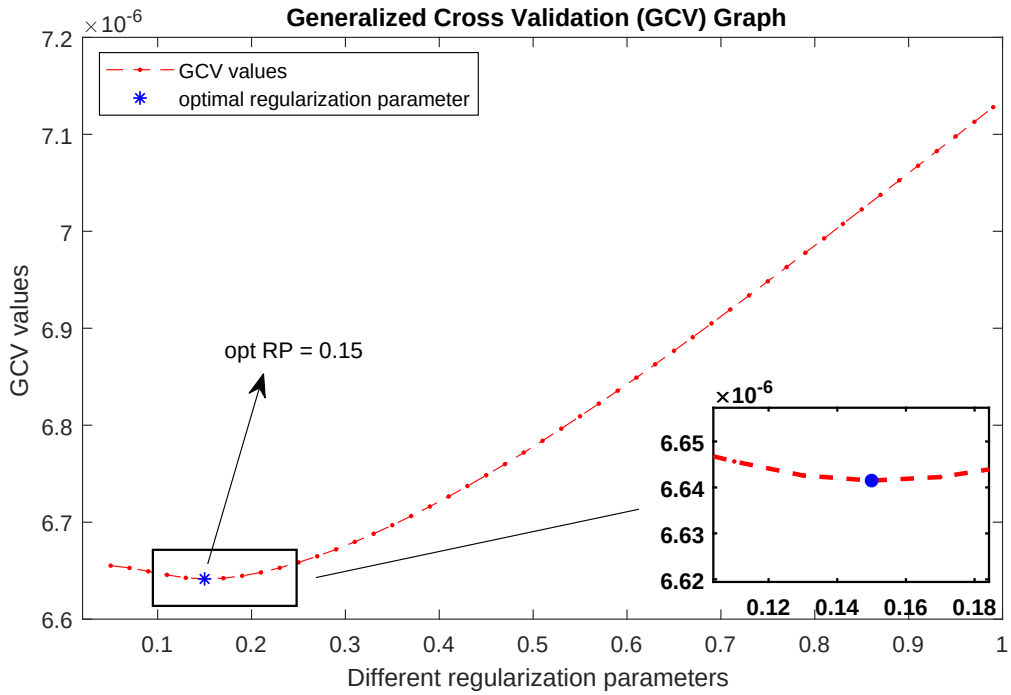


Figure 5.12: A graph showing the behavior of GCV values for various regularization parameter

Finally, we employ the optimal value from L-curve method in TV regularization to obtain numerically determined function. This is shown in the figure below,

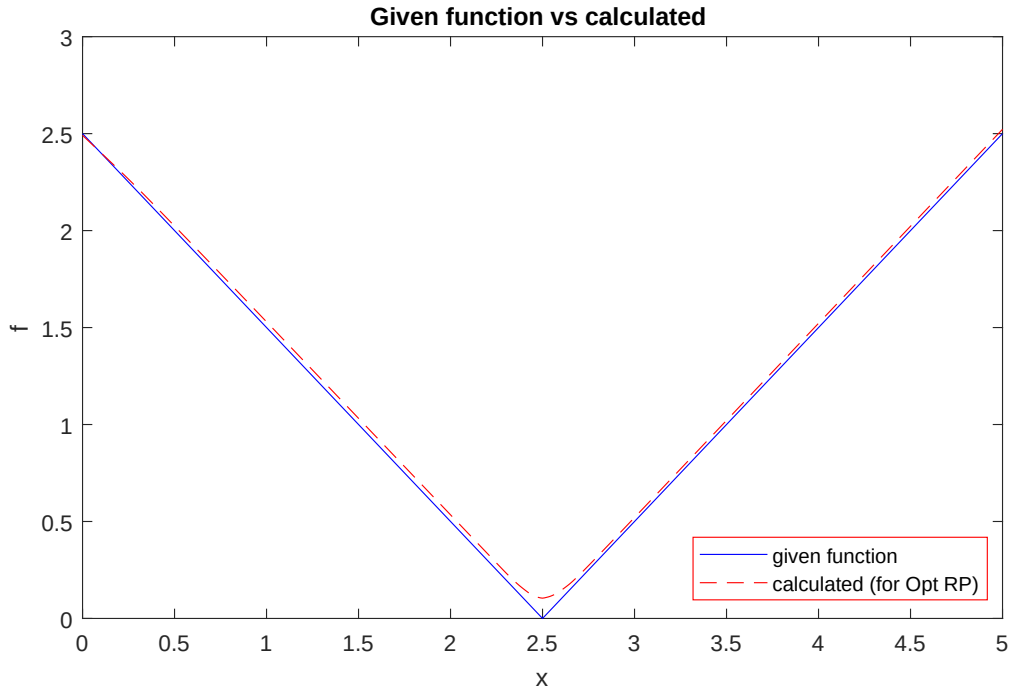


Figure 5.13: A graph showing the given function and numerically determined function

As both the given and determined functions are identical from figure 5.13, we can draw the inference that the selected optimal value is satisfactory.

5.1.3 Test function 3

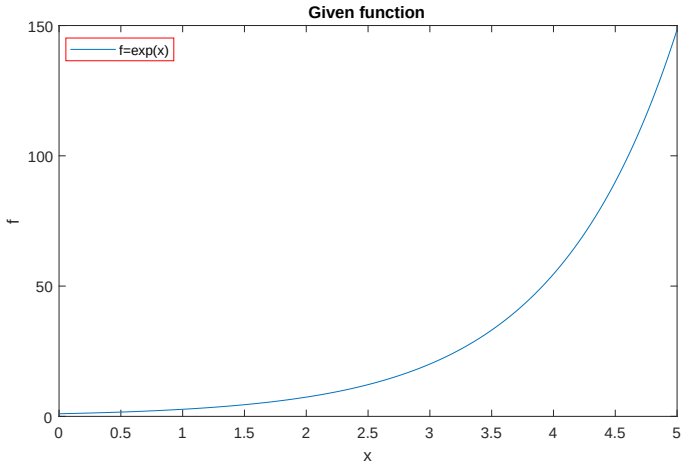
Data points	$x = 0 : 0.01 : 5$
Given function	$\exp(x)$
Standard deviation of AWGN	0.5
Noisy function	$y = f + \eta$

Table 5.3: Given information for test function 3

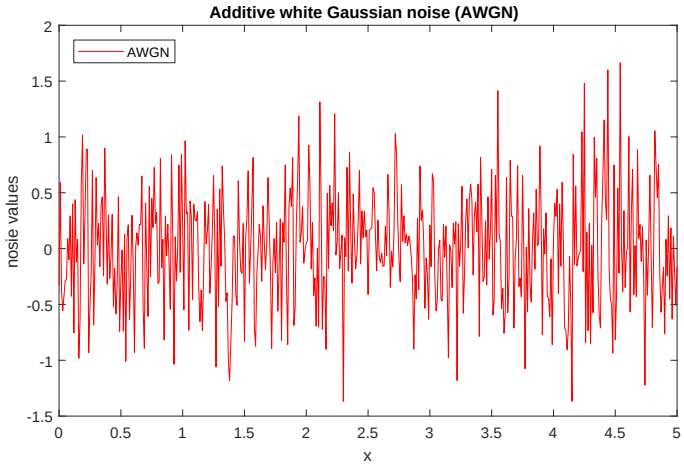
Note:

(a) The variable "η" is AWGN noise (vector) which is generated using "randn" a MATLAB® syntax.

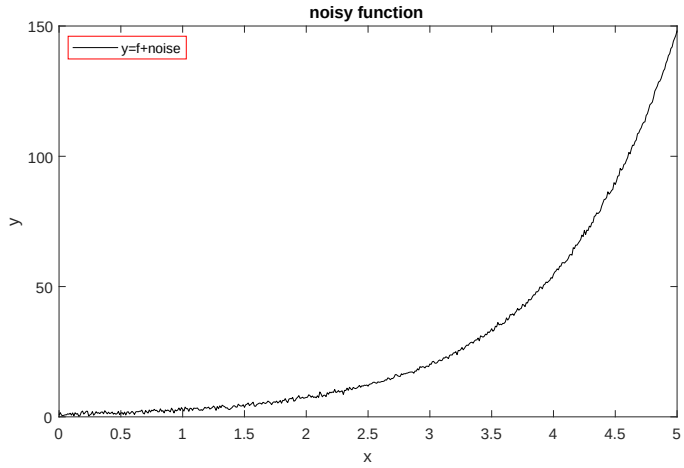
(b) Since we are dealing with exponential function, the standard deviation should be higher for AWGN in order to affect the test function significantly. Hence we choose a much higher standard deviation for test function 3 in comparison to test function 1 & 2.



(a)



(b) Additive white Gaussian noise

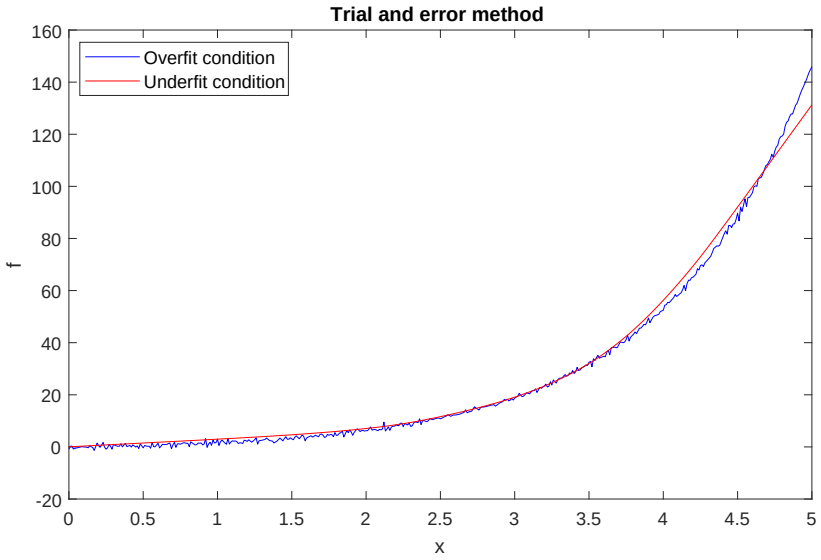


(c) Noisy function

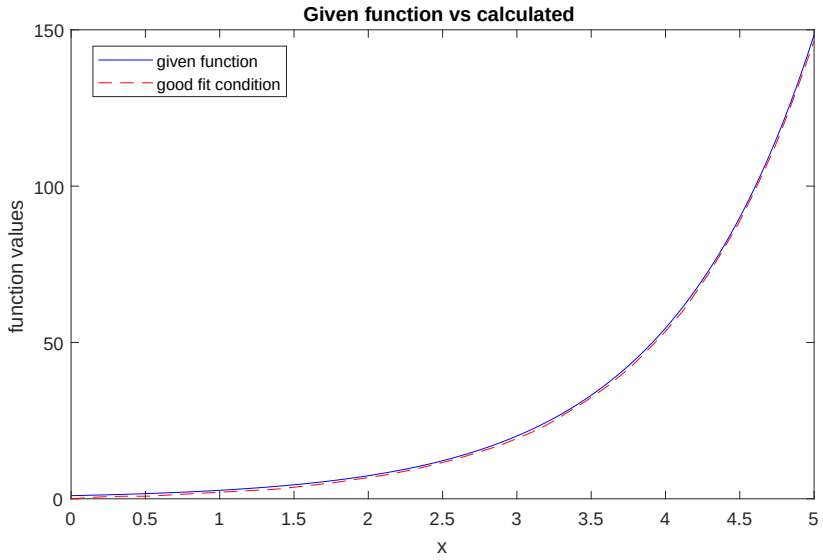
Figure 5.14: Graphs depicting the respective information provided in table 5.3

5.1.3.1 Eye-balling method

Similar to the explanation in 5.1.1.1, this method is inefficient in nature. Few results are shown below,



(a)



s

(b)

Figure 5.15: Graphs depicting the overfit and underfit condition (a) and good fit (b) for test function 3

5.1.3.2 L-curve method

The results for L-curve and its curvature plot are shown in figure 5.16 and 5.17 respectively.

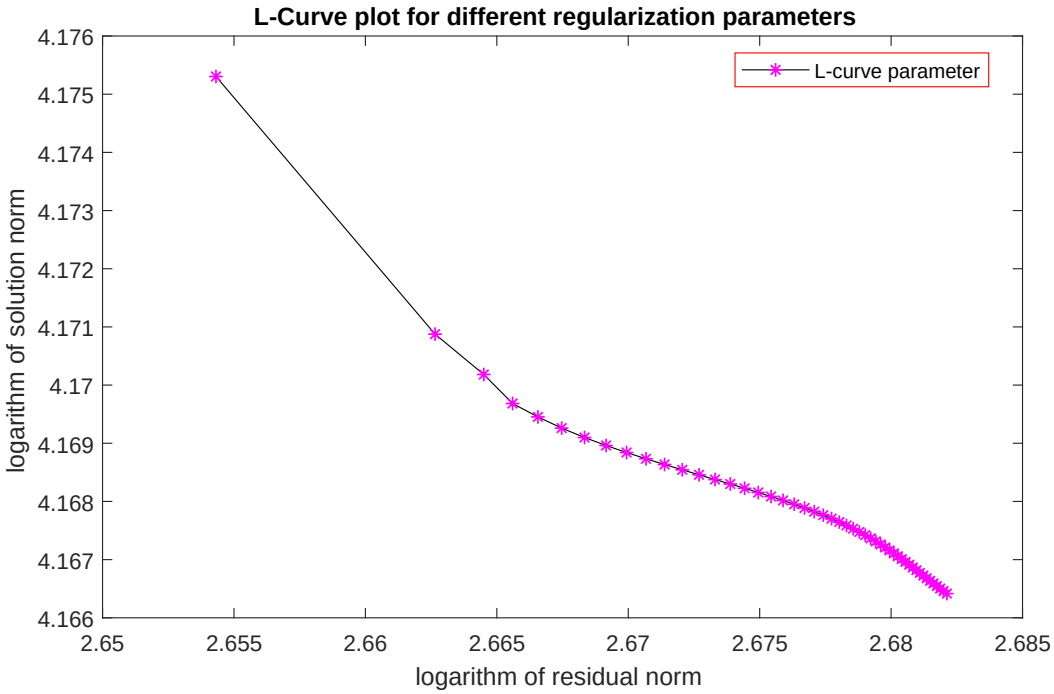


Figure 5.16: A graph representing L-curve

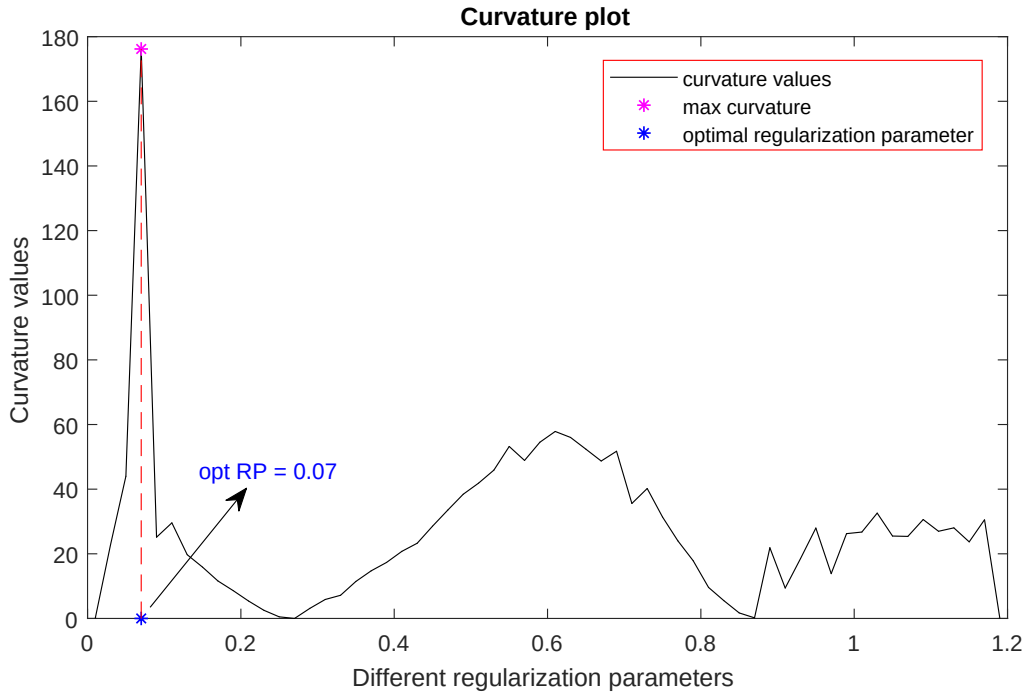


Figure 5.17: A graph showing curvature plot

5.1.3.3 NCP method

The below clearly depicts the nature of NCP values for various regularization parameter.

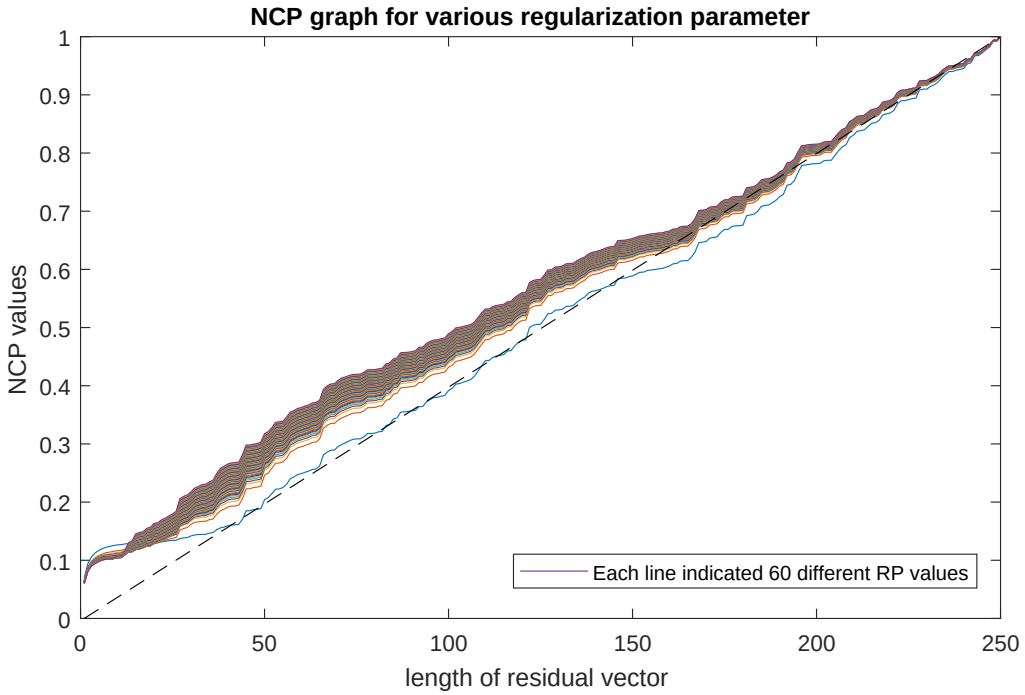


Figure 5.18: A graph showing curvature plot

Figure 5.18 indicates that the NCP values for the first regularization parameter (blue line) lies closer to the Gaussian white noise characteristic line (dotted black line). Hence we isolate those particular values and its shown in figure 5.19.

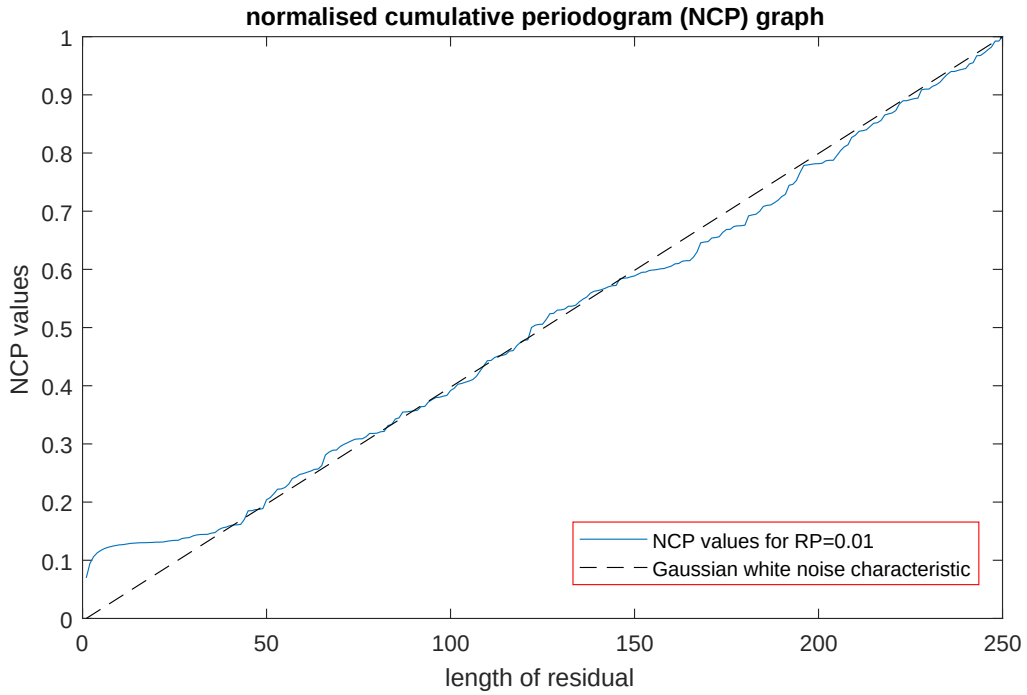


Figure 5.19: A graph NCP values for regularization parameter=0.01

IN the case of test function 3, the optimal values from either L-curve or NCP provides satisfactory numerical obtained solution.As seen in figure 5.20 the chosen optimal values proves to be satisfactory as the given function is identical to the numerical one.

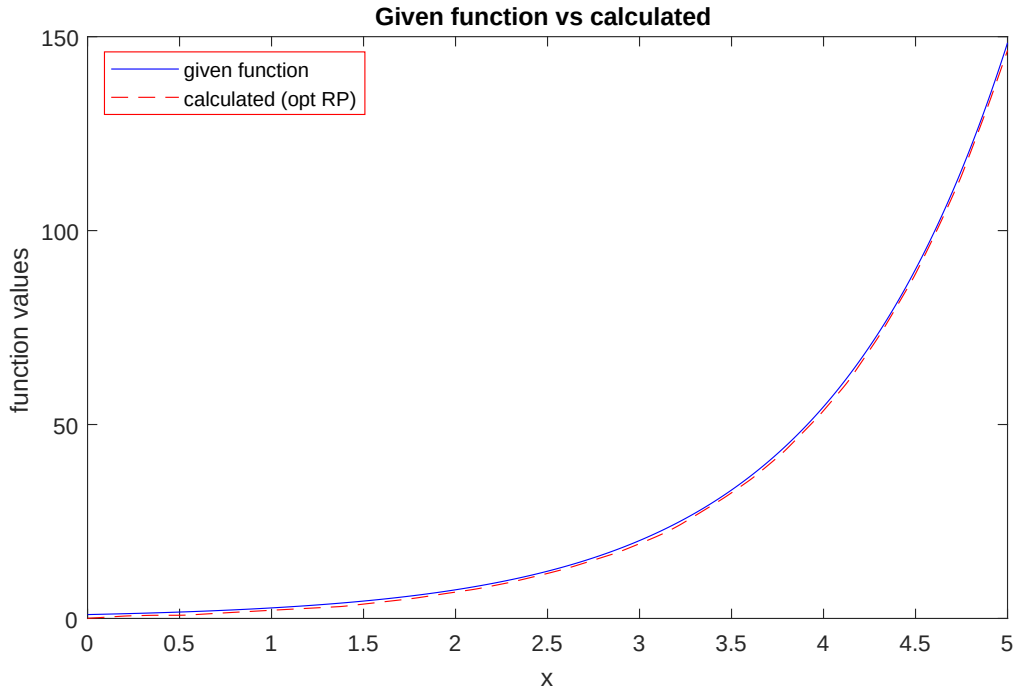


Figure 5.20: A graph showing given function and numerically obtained function for test function 3

5.2 Effect of different noise levels (standard deviation) on regularization parameter

In section 5.1, we provide the results using different test functions to determine the optimal values but for only one noise level (standard deviation). This section shines light on the effect of standard deviation of noise (AWGN) on regularization parameter.

We consider the test function 2 and change the standard deviation as shown in the table below,

Standard deviation 1	0.02
Standard deviation 2	0.05
Standard deviation 3	0.1
Standard deviation 4	0.2

Table 5.4: Different standard deviation

As explained in 5.1.2.2, we perform the analysis for various standard deviation in table 5.4 and the appropriate results were collected. The graph below shown the relationship between standard deviation and regularization parameter.

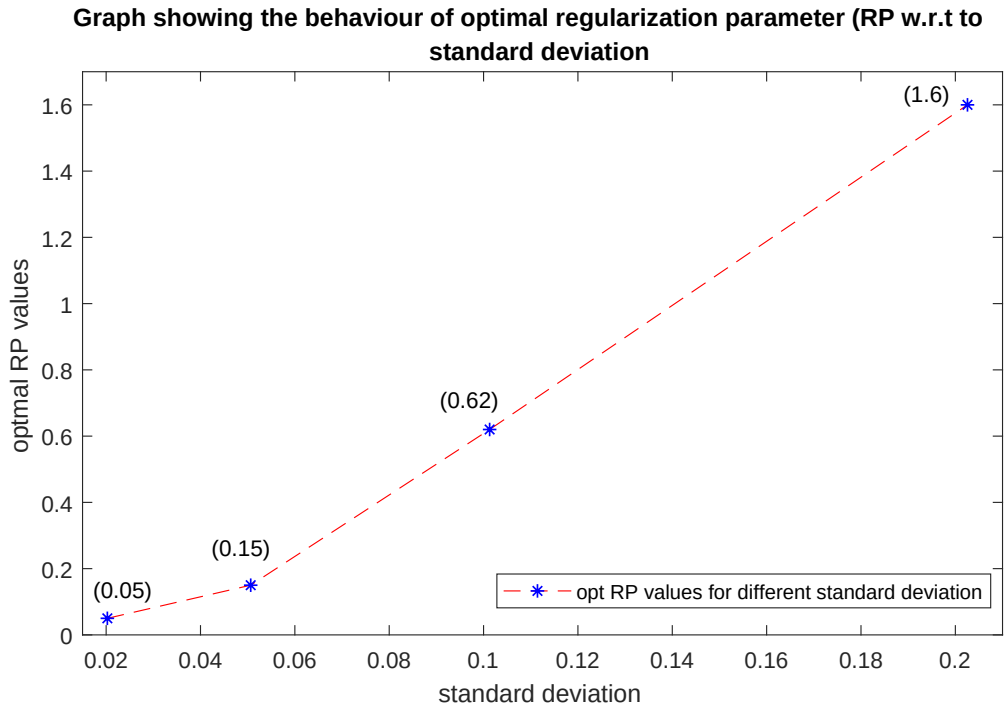


Figure 5.21: A graph showing the behavior of optimal regularization value w.r.t different standard deviation for test function 2

Note:

The marked values in figure 5.21 are the respective optimal values for the standard deviations mentioned in table 5.4.

Hence, with the help of figure 5.21 we can draw the inference that higher standard deviation of noise (AWGN) requires stronger regularization parameter as the process or dataset has higher variations in noise levels.

5.3 Complexities involved in derivatives of certain functions

In order to explain the derivative complexities involved in certain functions, we mainly focus on the mean squared error (MSE). We employ the method explained in the 4.2.4 to arrive at the optimal value for a given noisy function.

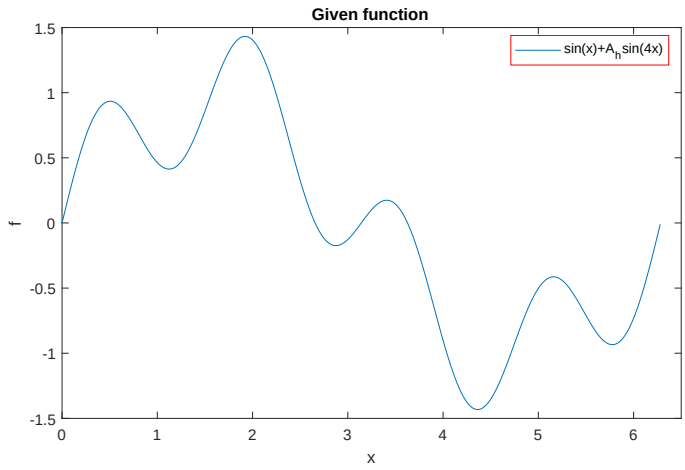
Now, similar to the earlier sections in this chapter, we select the following test function shown in the following table,

Data points	$x = 0 : 0.01 : 2\pi$
Given function	$\sin(x) + A_h \sin(4x)$
Arbitrary amplitude (A_h)	0.5
Standard deviation of AWGN	0.05
Noisy function	$y = f + \eta$

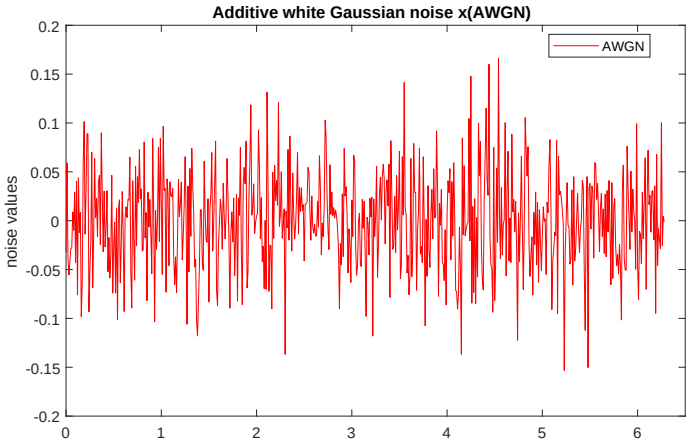
Table 5.5: Given information for test function 4

Note:

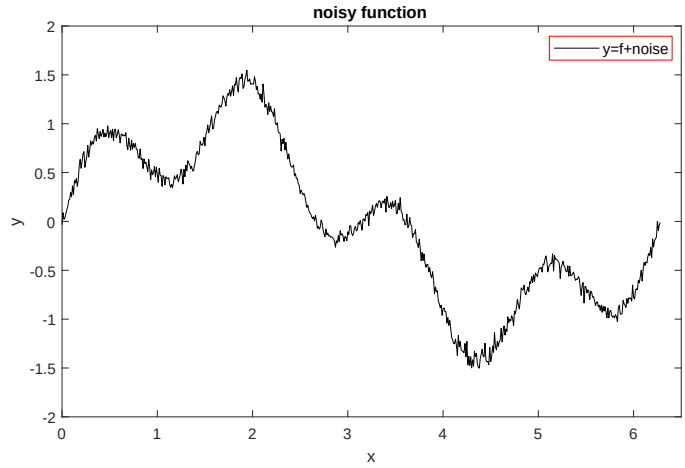
The variable "η" is AWGN noise (vector) which is generated using "randn" a MATLAB® syntax.



(a)



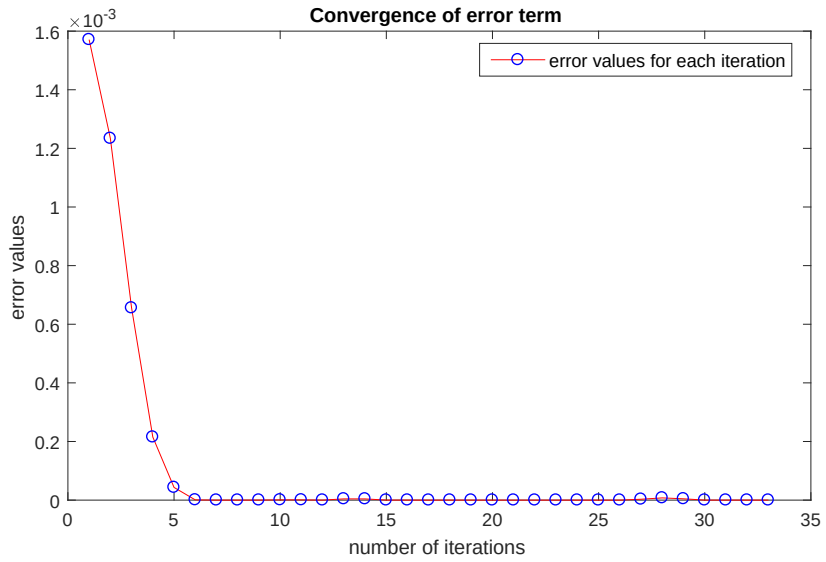
(b) Additive white Gaussian noise



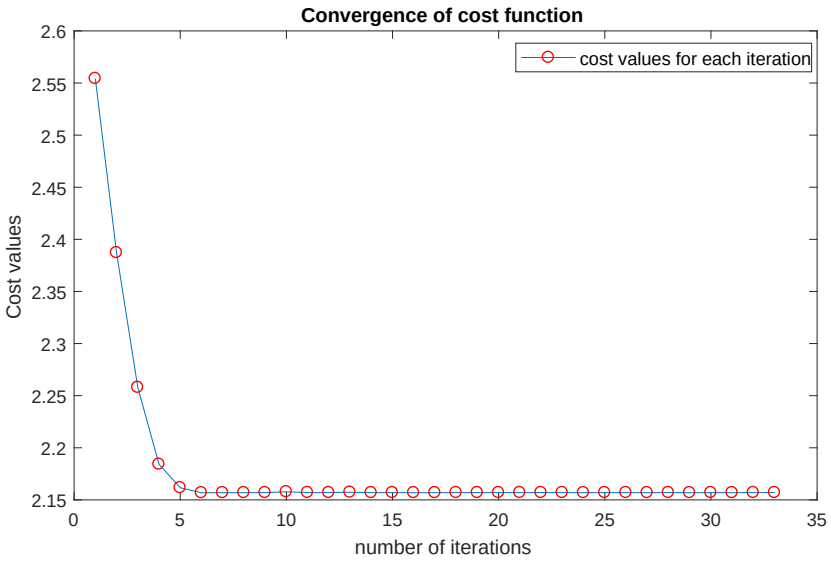
(c) Noisy function

Figure 5.22: Graphs depicting the respective information provided in table 5.5

The gradient method is performed by initializing a regularization parameter and learning curve values until the convergence criteria is met. The graphs in figure 5.23 proves our method converges.



(a)



s

(b)

Figure 5.23: Graphs depicting the error (a) and cost values (b) for respective iteration

Hence, the obtained optimal value and its MSE are shown in the table below,

Optimal regularization parameter	0.01716005
Mean squared error	0.028576623

Table 5.6: Important results of test function 4

Now, let us take a look at the how optimal value affects the derivative of test function 4.

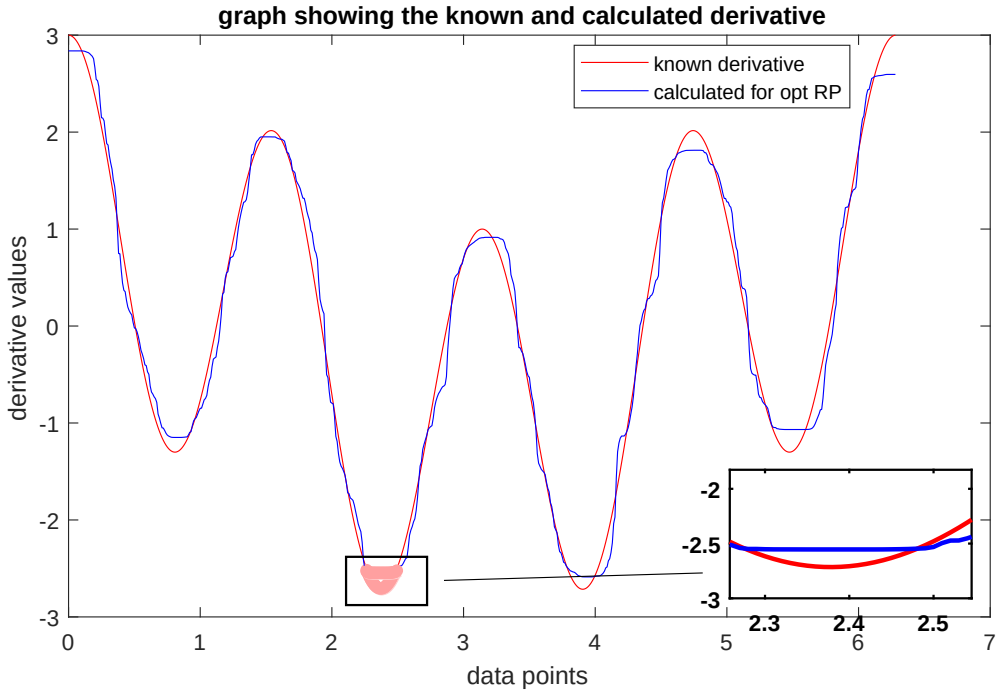


Figure 5.24: A graph showing both known derivative and calculated derivative for optimal regularization parameter

As seen from figure 5.24, the zoomed part of indicates one part of the contour of the function. As the number of data points are less the numerically obtained derivative tend to move away from the original contour. Hence it leads to a significantly higher value of MSE. The improvement of aforementioned complexities are reduced as shown,

We performed the analysis of MSE for different standard deviation of noise as well as for 2 sets of data points. The results are tabulated below,

Standard deviation	Data point 1 (629 points)	Data point 2 (6284 points)
	MSE 1	MSE 2
0.05027	0.02857	0.00946
0.20108	0.12961	0.06005
0.40217	0.41106	0.15296

Table 5.7: A comparison of MSE values for different standard deviation & data points

We can observe drastic reduction in MSE values as the number of data points are increased for the 3 cases of standard deviation. But the computational time (calculated using Intel® Core™ $i5-4210$U CPU @ 1.70GHz 2.40GHz) is approximately 51 folds more for data point 2 (6284 points) in comparison to data points 1 (629 points).

Note: Data point 1 $\implies$ 23.7053 seconds & Data point 2 $\implies$ 1208.19244 seconds

5.4 Data-driven method (sparse regression)

In this section, we perform the same procedure as mentioned in flowchart 4.5 for a test derivative function shown below,

Data points	$0:0.01:10$
Given derivative	$-0.3x+0.2x^2+0.1x^3+0.2x^4$
Standard deviation of AWGN	0.001
Noisy function	$y=f+\eta$

Table 5.8: Given information for test function 4

The test function can be obtained by using time integration scheme and AWGN is added and analysis are performed. The results obtained form the graphs prove convergence of error terms.

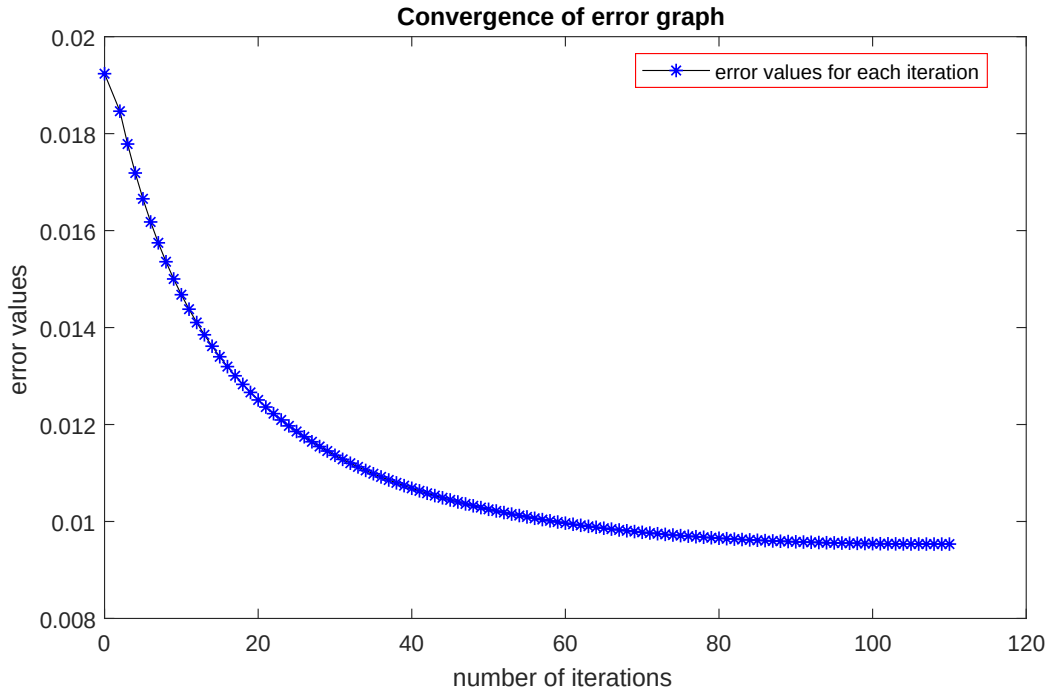


Figure 5.25: A graph showing convergence of error terms

The optimal value obtained from TV regularization is 11.8, we perform the sparse regression of and obtain the following values,

	1
1	0
2	-0.3044
3	0.2216
4	0.0713
5	0.5089
6	0

Figure 5.26: A figure showing the approximate solution

Comparing the values of coefficients of the known derivative in table 5.8 with that of the values in figure 5.26. WE can observe that the values are very close and hence the selected optimal parameter is satisfactory.

6 Conclusion

Th entire report is summaries briefly in this chapter.The phase of the thesis, which was the literature review provided us with knowledge regarding total variation regularization and its importance in the area of numerical differentiation. Hence we chose this as our basis method and performed various analysis for different test functions and noise levels.

We successfully investigate the fundamentals of large scale statistics by focusing on retrieving the information from noisy data. The method of total variation regularization helps us study thoroughly understand the concept behind regularization parameter on various test functions each at different amplitude of noise. The study behind the optimal parameter value shines light on the fact that a stronger noise level in a large scale dataset requires considerably strong optimal parameter.

As we know that, in the real life problems it very difficult to define noise from the actual measurement data because of which we also investigated the iterative process to automatically obtain regularization parameter.

The information being tracked was implemented in the process of finding differential equations by using data-driven method (sparse regression).

Bibliography

[1] 1. exploratory data analysis. pages 1892–1893.

[2] 2b.l1l2-regularization.

[3] 3 methods to deal with outliers | neural designer.

[4] Derivation of closed form lasso solution - cross validated.

[5] The l-tangent norm.

[6] least squares - how to derive the ridge regression solution? - cross validated.

[7] machine learning - parameters go to zero for lasso regularization function - cross validated.

[8] Ac waveform and ac circuit theory of sinusoids, 2013.

[9] Visible light waves, 2016.

[10] 5.1 - ridge regression | stat 897d, 5/9/2018.

[11] Murat Belge, Misha E. Kilmer, and Eric L. Miller. Efficient determination of multiple regularization parameters in a generalized l-curve framework. *Inverse Problems*, 18(4):1161–1183, 2002.

[12] Lorenzo Bruzzone, editor. SPIE Proceedings. SPIE, 2010.

[13] Rick Chartrand. Numerical differentiation of noisy, nonsmooth data. *ISRN Applied Mathematics*, 2011(1–4):1–11, 2011.

[14] Oddvar Christiansen, Tin-Man Lee, Johan Lie, Usha Sinha, and Tony F. Chan. Total variation regularization of matrix-valued images. *International journal of biomedical imaging*, 2007:27432, 2007.

[15] Dan. Microsoft powerpoint - lecture7.pptx.

[16] David Fernandez Prim. Copyright (c) 2009, david fernandez prim: on-figure-magnifier matlab.

[17] Selim Esedoglu. Total variation regularized l1 function approximation.

[18] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review*, 34(4):561–580, 1992.

[19] Per Christian Hansen and Misha E. Kilmer. A parameter-choice method that exploits residual information. *PAMM*, 7(1):1021705–1021706, 2007.

[20] Per Christian Hansen and Misha E. Kilmer. A parameter-choice method that exploits residual information. *PAMM*, 7(1):1021705–1021706, 2007.

[21] Intro to Inverse Problems. Chapter 5.

[22] Maarten Jansen, Maurits Malfait, and Adhemar Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44, 1997.

[23] Ian Knowles and Robert J. Renka. Methods for numerical differentiation of noisy data. 2014, 2014.

[24] Peng Liu and Dingsheng Liu. Selection of regularization parameter based on generalized cross-validation in total variation remote sensing image restoration. SPIE Proceedings, page 78301P. SPIE, 2010.

[25] Peter W. Macfarlane, A. van Oosterom, Olle Pahlm, Paul Kligfield, Michiel Janse, and John Camm, editors. *Comprehensive Electrocardiology*. Springer London, London, 2010.

[26] Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

[27] Bert W. Rust and Dianne P. O'Leary. Residual periodograms for choosing regularization parameters for ill-posed problems. *Inverse Problems*, 24(3):034005, 2008.

[28] Emmanuel Virot and Alexandre Ponomarenko. Popcorn: critical temperature, jump and sound. *Journal of the Royal Society, Interface*, 12(104):20141247, 2015.

[29] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996.

[30] Weisstein and Eric W. Euler-lagrange differential equation.