# On the MSE Performance and Optimization of Regularized Problems

Thesis by

Ayed Mofareh Alrashdi

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

November, 2016

# EXAMINATION COMMITTEE PAGE

The thesis of Ayed Mofareh Alrashdi is approved by the examination committee.

Committee Chairperson: Prof. Tareq Al-Naffouri
Committee Members: Prof. Mohamed-Slim Alouini, Prof. Taous-Meriem LALEG-KIRATI, Dr. Tarig Ballal

# ABSTRACT


## On the MSE Performance and Optimization of Regularized Problems

Ayed Mofareh Alrashdi

The amount of data that has been measured, transmitted/received, and stored in the recent years has dramatically increased. So, today, we are in the world of big data. Fortunately, in many applications, we can take advantages of possible structures and patterns in the data to overcome the curse of dimensionality. The most well known structures include sparsity, low-rankness, block sparsity. This includes a wide range of applications such as machine learning, medical imaging, signal processing, social networks and computer vision. This also led to a specific interest in recovering signals from noisy compressed measurements (Compressed Sensing (CS) problem). Such problems are generally ill-posed unless the signal is structured. The structure can be captured by a regularizer function. This gives rise to a potential interest in regularized inverse problems, where the process of reconstructing the structured signal can be modeled as a regularized problem. This thesis particularly focuses on finding the optimal regularization parameter for such problems, such as ridge regression, LASSO, square-root LASSO and low-rank Generalized LASSO. Our goal is to optimally tune the regularizer to minimize the mean-squared error (MSE) of the solution when the noise variance or structure parameters are unknown. The analysis is based on the framework of the Convex Gaussian Min-max Theorem (CGMT) that has been used recently to precisely predict performance errors.

# ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my advisor Prof. Tareq Al-Naffouri. Your help and guidance are invaluable. Thanks for being patient with me and for encouraging me. I also would like to thank Dr. Tarig Ballal.

My special thanks to my family for their patience and their understanding of me being away from them. Thank you my father Mofareh for everything you did for me, my words cannot express my feelings towards you. Thank you my mother Fahadah for being a great mom and for your everyday wishes. Thank you my brothers and sisters for your always support and encouragement.

I would like also to thank all of my friends at KAUST, especially Raid AlRowais and Ismail Ben Atitallah. Special thanks to my old teachers Yasser Al-Refei and Khalid Al-Rmali for supporting me in my very beginning days and for making me love math. I also thank my friend Abdullah Eid Alrashdi for his great support and encouragement.

Finally, I would like to thank the University of Hail for sponsoring me and giving me the opportunity to be in the world of KAUST.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF ACRONYMS

| | |
|---|---|
| AO | Auxiliary Optimization |
| CGMT | Convex Gaussian Min-max Theorem |
| COPRA | Constrained Perturbation Regularization Approach |
| CS | Compressed Sensing |
| DCT | Discrete Cosine Transform |
| GCV | Generalized Cross Validation |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| MSE | Mean Squared Error |
| NSE | Normalized Squared Error |
| OLS | Ordinary Least-Squares |
| PO | Primary Optimization |
| RLS | Regularized Least-Squares |
| SVD | Singular Value Decomposition |

# Chapter 1

# Introduction

## 1.1 Overview

Nowadays, our world is full of technology and our lives become more and more surrounded and involved in this technology. From smartphones to our laptops and smart watches we generate more and more data. In fact, the amount of data that has been generated, measured, and stored in the recent years has tremondously increased leading to the revolutionary era of big data. The commencement of such era is associated with many challenges pertaining the way in which the large volumes of data are handled and calls for the development of new methods for data processing that can cope with the high dimentionality. Tranditional methods based on classical frameworks fall short as the sizes of data sets grow. However, there are always ways around. Exploiting structures that reside in the data is one of these ways. In many applications, we can take advantages of certain structures and patterns in the data to overcome the curse of dimensionality [1]. Such structures include sparsity, low-rankness, and block sparsity. These structures appear in a wide range of applications such as machine learning, medical imaging, signal processing, social networks, statisctics, sensor networks and computer vision. The structure can be captured by regularizer function. This gives rise to a potential interest in regularized inverse problems, where the process of reconstructing the structured signal can be modeled as a regularized problem, and can be solved generally by non-smooth convex optimization algorithms [2]. One of the most challenging problems is how to optimally tune the regularization param-

eter involved to obtain the best reconstruction of the signal of interest. This thesis focuses on developing methods that leverage the structure how to optimally select the regularization parameter for a set of regularized problems including ridge regression, LASSO, square-root LASSO and low-rank generalized LASSO, when the structure parameters are unknown. The proposed techniques are based on the new framework of the CGMT that has been shown to give precise predictions of the estimation error given in the form of the mean-squared error (MSE) and the normalized-squared error (NSE) [3],[2],[4].

## 1.2 Related Work

Estimating a signal from noisy linear measurements can be traced back to Gauss, where the ordinary least square (OLS) estimate appears. Due to the poor performance of the OLS, the regularized least squares (RLS) approach was proposed. There exist a few methods in the literature that tackles the problem of optimally and automatically selecting the regularization parameter for the regularized problems. In fact, in many areas people tune this parameter manually or based on some heuristics. Some of the methods used for the regularization parameter tuning include Generalized Cross-Validation (GCV) [5],[6], L-curve [7],[8], quasi-optimal [9],[10] and COPRA [11]. The performance of these methods vary based on the nature of the problem. These methods are generally agonistic to any structures that are present in the data. For a general inverse problem and where the signal is structured, only in the recent years there have been contributions to this area. In the recent work of [2], [3], [4] and [12] the authors proposed a new method to optimally tune the regularizer for the generalized M-estimators. However, this method assumes that all the structure parameters (for example the sparsity, or noise variance) of the problem are known. However, such knowledge is not usually available in practice. This is the major focus of this thesis. In other words, we attempt to optimally tune the regularization

parameter for a number of estimation problems when certain structures are present in the data, but the value of the parameters associated with these structures are not known.

## 1.3 Contributions

The contributions of this thesis can be summarized as follows:

- We propose a new method that is based on the CGMT [13] framework, to find the optimal regularization parameter for the regularized problem of estimating a structured signal from noisy measurements. We focus on the case where the structure parameters are unknown.
- We derive an analytical expression for the mean squared error (MSE) performance of the ridge regression problem that is also based on the CGMT. We also derive an analytical expression for the optimal cost function.
- We present a simple smoothing spline method that can be used to fit a noisy data. We take advantage of this method to reduce the number of simulation iterations required to estimate the structure parameters to only one iteration.

## 1.4 Organization

We start by presenting a constrained penalized smoothing splines method in Chapter 2. Then, we develop the analytical MSE performance and then use it to optimally select the regularization parameter for the Ridge Regression problem in Chapter 3. After that, we move to the case of structured signals and we discuss the proposed approach for optimal regularizer tuning, when the structure parameters are unknown in Chapter 4. Finally, we conclude this thesis in Chapter 5 and give some possible future directions.

## 1.5 Notation

$\mathbb{E}[\cdot]$ denotes the expectation of a random variable, $\mathbb{P}(\cdot)$ denotes the probability of an event $(\cdot)$, $\mathbf{x}_i$ represents the $i^{th}$ component of a vector $\mathbf{x}$. $f^{'}, f^{''}$ represent the first and second derivative of a function $f$ respectively.

# Chapter 2

# Constrained Penalized Smoothing Splines

## 2.1 Overview

In this chapter, we will discuss the problem of fitting noisy data with spline functions. In section 2.3, we will introduce the idea of using spline functions to fit the data for the noiseless case. On the other hand, section 2.4 is dedicated to describe our method of using cubic smoothing splines in the presence of noise in the data. It is worth to mention that this chapter is not the main component of this thesis. It describes a tool that will be used in the later chapters to smooth the cost function of some regularized problems in order to obtain their optimal tuning parameter.

## 2.2 Interpolation

In signal and image processing and also in statistics, smoothing is required to reduce the experimental noise while trying to keep the most important imprints of the data. Let us assume that our data points are coming from the model:

$$y_i = f(x_i) + \epsilon_i, i = 1, 2, ..., n \tag{2.1}$$

where $f(x_i)$ is a smooth but unknown function and $\epsilon_i$ are error or noise terms that are assumed to be independent and identically distributed (i.i.d.).
Smoothing $y$ relies upon finding a spline function $\tilde{f}$ that best estimate the original data. Smoothing is generally done by means of parametric or non-parametric regres-

sion. Parametric regression assumes a priori knowledge of a predetermined analytical function that represent the data. However, most of the data values cannot be expressed in terms of a single function, so non-parametric regression is usually used for data smoothing. The most important approaches to nonparametric regression include kernel smoothing [14], penalized least-squares regression which was first introduced by Whittaker [15] and has been studied a lot since the work of Wahba [5] and the work of Schoenberg on smoothing splines. M. Unser, et al. produced some splendid references on interpolation and splines that are dedicated to the use in signal and biomedical image processing [16], [17],[18]. Weinert, et al. presented a straightforward approach to express the roughness term using a second divided difference and used different transforms and decompositions [19],[20],[21]. D. Garcia also presented a robust smoothing algorithm where the Discrete Cosine Transform (DCT) had been used [22].

A spline function is a function constructed from polynomial segments that are joined together and subject to conditions or continuity at their knots [23],[24]. The cubic spline functions are the mostly used functions. In this chapter, we will present the algorithm of the cubic smoothing splines based on the simple mathematical model of natural cubic spline interpolation that is penalized to obtain the required smoothness in noisy data scenarios. The goal here is to estimate the best spline coefficients based on a Regularized Least Squares (RLS) problem formulation, and then use them to fit noisy data in a way that balances between the fit of the data and the desired smoothness.

## 2.3 Interpolating Splines

Given a set of data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where $x_1 < x_2 < ... < x_n$ ($x_1, ..., x_n$ are also called the knots). We want to find cubic polynomials (also known as spline functions) $S_i(x), i = 1, 2, ..., n-1$ that connect between each pair of adjacent

points $(x_i, y_i), (x_{i+1}, y_{i+1})$. A cubic spline is defined by

$$S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3 \tag{2.2}$$

where $x \in [x_i, x_{i+1}]$ and $a_i, b_i, c_i$ and $d_i$ are the polynomial coefficients. For a given $n$ data points, we need $n-1$ spline functions. These spline functions are joined together to form a continuous curve with continuous first and second derivatives, i.e., $S(x), S^{'}(x), S^{''}(x)$ must all be continuous and $S(x)$ must interpolate the data points at the given knots [23]. To interpolate between the data points, we have $4(n-1)$ unknown spline coefficients to be determined (i.e., $a_i, b_i, c_i$, and $d_i$), so we need $4(n-1)$ simultaneous equations. The first and second derivatives of the spline function are

$$S^{'}(x) = b_i + 2c_i x + 3d_i x^2, \tag{2.3}$$

and

$$S^{''}(x) = 2c_i + 6d_i x \tag{2.4}$$

Each spline function should interpolate all of the data points, which means:

$$S_i(x_i) = y_i, \forall i = 1, 2, ..., n-1,$$

or equivalently,

$$a_i + b_i x_i + c_i x_i^2 + d_i x_i^3 = y_i. \tag{2.5}$$

Also, the splines must be equal at the inner knots, which leads to the following continuity condition:

$$S_i(x_i) = S_{i+1}(x_i), \forall i = 2, 3, ..., n-2,$$

or equivalently,

$$(a_i - a_i) + (b_i - b_{i+1})x_i + (c_i - c_{i+1})x_i^2 + (d_i - d_{i+1})x_i^3 = 0. \tag{2.6}$$

The condition that the first derivatives should be equal at the inner knots can be expressed as

$$S_i^{'}(x_{i+1}) = S_{i+1}^{'}(x_{i+1}), \forall i = 1, 2, ..., n-2$$

or equivalently,

$$(b_i - b_{i+1}) + 2(c_i - ci + 1)x_{i+1} + 3(d_i - d_{i+1})x_{i+1}^2 = 0. \tag{2.7}$$

The condition that the second derivatives should be equal at the inner knots gives

$$S_i^{''}(x_{i+1}) = S_{i+1}^{''}(x_{i+1}), \forall i = 1, 2, ..., n-2$$

or equivalently,

$$2(c_i - ci + 1) + 6(d_i - d_{i+1})x_{i+1} = 0. \tag{2.8}$$

For the end point conditions, we will use the natural cubic spline (NCS) conditions

$$S_1^{''}(x_1) = 0,$$

and

$$S_{n-1}^{''}(x_n) = 0,$$

or equivalently,

$$2c_1 + 6d_1x_1 = 0, \tag{2.9}$$

and

$$2c_{n-1} + 6d_{n-1}x_n = 0. \tag{2.10}$$

To make the interpolating cubic splines model more convenient, we will use the matrix-vector form to represent the system of equations for the spline conditions given above.

Let $\mathbf{u_i} = \begin{bmatrix} a_i & b_i & c_i & d_i \end{bmatrix}^T$, where $i = 1, 2, ..., n-1$, be a $(4 \times 1)$ vector of the coefficients of the $i^{\text{th}}$ spline function $S_i(x)$.

Also, let $\mathbf{p_i} = \begin{bmatrix} 1 & x_i & x_i^2 & x_i^3 \end{bmatrix}^T$, where $i = 1, 2, ..., n$, be a $(4 \times 1)$ vector of spline basis.

Define $\mathbf{s} = \begin{bmatrix} \mathbf{u_1}^T & \mathbf{u_2}^T & \mathbf{u_3}^T & \cdots & \mathbf{u_{n-1}}^T \end{bmatrix}^T$, be a $4(n-1) \times 1$ vector of the concatenation of all the spline coefficients of the model.

The interpolation conditions (equation (2.5)) can be represented by the following equation:

$$\mathbf{A}_d \, \mathbf{s} = \mathbf{y}, \tag{2.11}$$

where

$$\mathbf{A_d} = \begin{bmatrix} \mathbf{p_1}^T & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{p_2}^T & \mathbf{0}^T & \cdots & \mathbf{0}^T \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \mathbf{0}^T & \mathbf{p_n}^T \end{bmatrix} \in \mathbb{R}^{n \times 4(n-1)},$$

where $\mathbf{0}$ in the above matrix is a $4 \times 1$ vector of all zeros.

The interpolation function continuity, first derivative continuity, and second derivative continuity conditions (equations (2.6)-(2.8)) can be expressed in a matrix-vector form as follows:

$$\mathbf{C} \, \mathbf{s} = \mathbf{0}_{3n-2}, \tag{2.12}$$

where $\mathbf{C} \in \mathbb{R}^{3(n-2) \times 4(n-1)}$ is given by

$$\mathbf{C} = \begin{bmatrix} \mathbf{p_2}^T & -\mathbf{p_2}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \mathbf{q_2}^T & -\mathbf{q_2}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \mathbf{z_2}^T & -\mathbf{z_2}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{p_3}^T & -\mathbf{p_3}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{q_3}^T & -\mathbf{q_3}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{z_3}^T & -\mathbf{z_3}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \mathbf{0}^T \\ \vdots & & & & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \mathbf{0}^T & \mathbf{p_{n-1}}^T & -\mathbf{p_{n-1}}^T \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \mathbf{0}^T & \mathbf{q_{n-1}}^T & -\mathbf{q_{n-1}}^T \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \cdots & \cdots & \mathbf{0}^T & \mathbf{z_{n-1}}^T & -\mathbf{z_{n-1}}^T \end{bmatrix},$$

where $\mathbf{q_i} = \begin{bmatrix} 0 & 1 & 2x_i & 3x_i^2 \end{bmatrix}^T$, $\mathbf{z_i} = \begin{bmatrix} 0 & 0 & 2 & 6x_i \end{bmatrix}^T$ and $\mathbf{0}$ is a $4 \times 1$ vector of all zeros.

Finally, for the end point conditions (NCS):

$$\tilde{\mathbf{Z}}\ \mathbf{s} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{2.13}$$

where $\tilde{\mathbf{Z}} \in \mathbb{R}^{2 \times 4(n-1)}$ is given by

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{z_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{z_n} \end{bmatrix}.$$

Let $\mathbf{A_c} = \begin{bmatrix} \mathbf{C}^T & \tilde{\mathbf{Z}}^T \end{bmatrix}^T$ be the matrix of the concatenation of the conditions imposed on the spline model, such that:

$$\mathbf{A}_c\ \mathbf{s} = \mathbf{0}_{3n-4}. \tag{2.14}$$

To form the final natural cubic spline model, we will concatenate all of the conditions matrices into a single matrix $\mathbf{A}$, so that our NCS model becomes:

$$\mathbf{A}\ \mathbf{s} = \tilde{\mathbf{y}}, \tag{2.15}$$

where $\mathbf{A} = \begin{bmatrix} \mathbf{A_d}^T & \mathbf{A_c}^T \end{bmatrix}^T$, and $\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}^T & \mathbf{0^T}_{3n-4} \end{bmatrix}^T$.

For the system of equations in 2.15, a simple least-squares estimation approach can be used to solve for the optimal coefficients $\hat{\mathbf{s}}$. Once these coefficients are estimated, one can use the interpolation equation ($\mathbf{y} = \mathbf{A}_d\ \hat{\mathbf{s}}$) to find the final fit of the data $\mathbf{y}$.

Figure 2.1 illustrates the NCS fitting of a non-noisy data generated from the function $y = \frac{1}{1+25x^2}$. This kind of interpolation is simple due to the absence of noise. The more challenging case is when the data is corrupted by noise, which will be discussed in the next section.

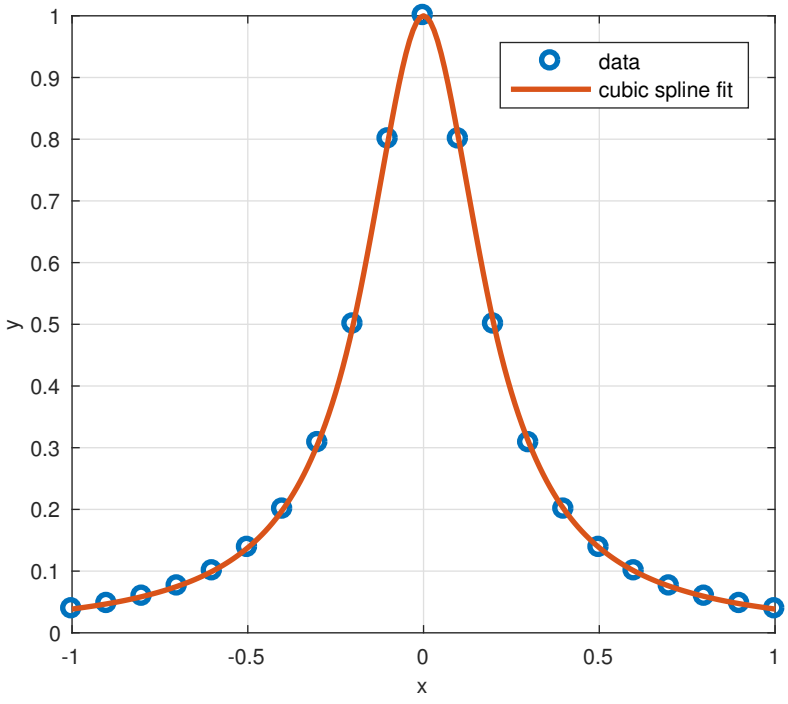


Figure 2.1: Interpolation of noiseless data, generated from $y = \frac{1}{1+25x^2}$

## 2.4 The Proposed Constrained Penalized Smoothing Spline Method

The interpolating splines are useful in representing a smooth function $f(x)$ only when the data points are exactly on the path of the function or very close to it. However, if the data is corrupted by noise and scattered randomly around the function path, then an interpolating polynomial will follow the same fluctuations and cannot give a reasonable approximation of the underlying function [23]. Therefore, for the sake of smoothness, we may want to allow the spline function to depart from the given noisy points by means of regularization of the interpolation model.

In this case, we can try to construct the function $f(x)$ by using a cubic spline function $S(x)$ with coefficients $\mathbf{s}$ that solves the following minimization problem:

$$\begin{aligned} \hat{\mathbf{s}} &= \underset{\mathbf{s}}{\operatorname{argmin}} \; ||\mathbf{A_d s} - \mathbf{y}|| + \gamma ||\mathbf{s}|| \\ \text{subject to :} \quad & \mathbf{A}_c \mathbf{s} = \mathbf{0}. \end{aligned} \tag{2.16}$$

where $\mathbf{A}_d$, $\mathbf{A}_c$, and $\mathbf{s}$ are the same as defined before in the interpolating splines section, and $\mathbf{A}_c$ is the matrix including the conditions of the splines. The reason here for making the conditions as a separate constraint is that we want them to be exact since they are not affected by the noise.

The singular value decomposition (SVD) of the condition matrix $\mathbf{A}_c$ is defined as:

$$\mathbf{A}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{2.17}$$

Using the SVD of $\mathbf{A}_c$, the constraint in (2.16) becomes

$$\begin{aligned} \mathbf{U}\boldsymbol{\Sigma}\underbrace{\mathbf{V}^T\mathbf{s}}_{\mathbf{w}} &= \mathbf{0} \\ \implies \boldsymbol{\Sigma}\mathbf{w} &= \mathbf{0} \\ \implies \begin{bmatrix} \boldsymbol{\Sigma}_1 & \underbrace{\boldsymbol{\Sigma}_2}_{\mathbf{0}} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} &= \mathbf{0} \\ \implies \boldsymbol{\Sigma}_1\mathbf{w}_1 &= \mathbf{0}, \\ \implies \mathbf{w}_1 &= \mathbf{0}. \end{aligned}$$

The matrix $\boldsymbol{\Sigma}_2$ is set to zero due to the specific structure of the matrix $\mathbf{A}_c$.

Based on the above results, we can transform the constrained minimization problem in 2.16 into the following unconstrained problem

$$\min_{\mathbf{w}} ||\mathbf{A}_d\mathbf{V}\mathbf{w} - \mathbf{y}|| + \gamma||\mathbf{V}\mathbf{w}|| \tag{2.18}$$

Taking $\mathbf{V} = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}$, and recalling that $\mathbf{w}_1 = \mathbf{0}$, the minimization problem in 2.18 simplifies to:

$$\min_{\mathbf{w}_2{}'} ||\mathbf{A}_d\ \mathbf{w}_2{}' - \mathbf{y}|| + \gamma||\mathbf{w}_2{}'||, \tag{2.19}$$

where, $\mathbf{w}_2{}' = \mathbf{V}_2\ \mathbf{w}_2$. Solving this reduced optimization problem in 2.19 and recalling that $\mathbf{s} = \mathbf{V}\mathbf{w}$ will give the final estimate of the optimal spline coefficients $\hat{\mathbf{s}}$. The final step is to fit the noisy data. This is done by using Equation (2.11).

$\gamma$ is the smoothing (regularization) parameter that balances between the closeness to the data and the smoothness. It is chosen automatically based on any existing method such as GCV.

When $\gamma = 0$, the solution of 2.19 boils down to the NCS solution that passes exactly through the data points. The proposed smoothed spline method, besides its benefit for different applications, will be used later in this thesis to estimate some of the signal

parameters (such as sparsity, noise variance,...) which is in turn used to select the optimal regularization parameter in a novel way. We can see the performance of the proposed method in the presence of noise in the data in figure 2.2, where we assume no priori information. From this figure, we can see that even when the data is noisy the regularization helped getting closer results to the original data. The performance is compared with a set of benchmarks including the algorithms proposed in [20], [21] and [22].

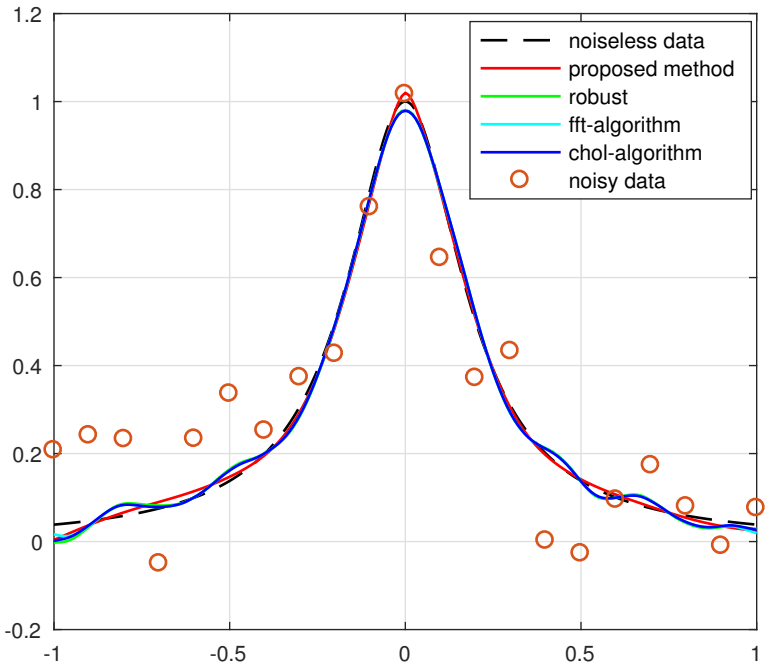


Figure 2.2: smoothing a noisy data, generated from $y = \frac{1}{1+25x^2}$ with white noise

# Chapter 3

# Optimal Regularization Parameter Selection for Ridge Regression

## 3.1 Introduction

In this thesis, we consider the problem of estimating an unknown signal $\mathbf{x_0} \in \mathbb{R}^n$ from a noisy linear measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x_0} + \mathbf{z} \in \mathbb{R}^m, \tag{3.1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the measurement matrix that has independent and identically distributed (i.i.d.) Gaussian entries $\mathcal{N}(0, \frac{1}{n})$, and $\mathbf{z} \in \mathbb{R}^m$ is the noise vector.
A commonly used approach to obtain an estimate $\hat{\mathbf{x}}$ of the signal $\mathbf{x_0}$ is by solving the following non-smooth convex optimization problem:

$$\hat{\mathbf{x}} := \arg \min_{\mathbf{x}} \quad \mathcal{L}(\mathbf{y} - \mathbf{A}\mathbf{x}) + \lambda f(\mathbf{x}), \tag{3.2}$$

where $\mathcal{L} : \mathbb{R}^m \to \mathbb{R}$ is the loss function that measures the mismatch between the observations $\mathbf{y}$ and $\mathbf{A}\hat{\mathbf{x}}$, the regularizer $f : \mathbb{R}^n \to \mathbb{R}$ as discussed before promote the structure of $\mathbf{x}_0$, and the regularization parameter $\lambda$ balances between the two.
Different choices of $\mathcal{L}$ and $f$ lead to different popular problems [25]:

- Ordinary Least Squares(OLS): $\mathcal{L} = \frac{1}{2}||\cdot||_2^2$, $f(\cdot) = 0$.
- $\ell_2^2$-LASSO: $\mathcal{L} = \frac{1}{2}||\cdot||_2^2$, $f(\cdot) = ||\cdot||_1$, that is a popular for sparse recovery.

- $\ell_2$- (Square-root) LASSO: $\mathcal{L} = ||\cdot||_2$, $f(\cdot) = ||\cdot||_1$.
- Generalized LASSO: $\mathcal{L} = \frac{1}{2}||\cdot||_2^2$ or $\mathcal{L} = ||\cdot||_2$, which is a generalization of the LASSO to arbitrary convex regularizer $f(\cdot)$.
- Ridge Regression or Regularized Least Squares (RLS): $\mathcal{L} = \frac{1}{2}||\cdot||_2^2$, $f(\cdot) = ||\cdot||_2^2$.
- Regularized Least Absolute Deviation (LAD): $\mathcal{L} = ||\cdot||_1$. It is used for example in the case of sparse noise.

There are several other choices that exist in the literature such as Support Vector Machines, Hubber loss, etc.

This chapter focuses on the Ridge regression or the so called regularized least squares (RLS) solution of the problem and finding its optimal regularization parameter especially when the noise variance is not known. Other optimization problems will be treated in the subsequent chapter.

## 3.2 Background

The OLS method attempts to find an estimate $\hat{\mathbf{x}}$ for $\mathbf{x}$ by minimizing the $ell_2$-norm of the residual. That is

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \ ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2. \tag{3.3}$$

The solution to (3.3) is given by [26]

$$\hat{\mathbf{x}} \ = \ (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y}, \tag{3.4}$$

where $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is the singular value decomposition (SVD) of $\mathbf{A}$, $\mathbf{u}_i$ and $\mathbf{v}_i$ are the left and the right orthogonal singular vectors, while $\sigma_i$ is $i^{th}$ singular value that satisfies $\sigma_1 \geq \sigma_1 \geq \cdots \geq \sigma_n$ [27]. The OLS solution is very sensitive to perturbations in the data such as the one that results from the noise.

This means any small change in the data results in a huge change in the solution. To overcome this issue, regularization techniques are frequently used. The most common form of regularization is Tikhonov regularization [28], which is also known as Ridge Regression in machine learning and statistics contexts. The Tikhonov regularization problem is given by:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \ ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2 + \lambda||\mathbf{x}||^2, \tag{3.5}$$

where $\lambda \geq 0$ is the regularization parameter that balances between the fit to the data (the residual norm) and the magnitude of the coefficients of the solution (for example to avoid over fitting). It has been proved that the solution to (3.5) is given by the RLS estimator

$$\hat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{y}. \tag{3.6}$$

The optimal regularizer for (3.6) is given by the linear minimum mean square error (LMMSE) estimator [26]. For any signal $\mathbf{x}$ and noise $\mathbf{z}$, the LMMSE estimator is given by

$$\hat{\mathbf{x}}_{\text{LMMSE}} = \left(\mathbf{A}^T\mathbf{R}_{\mathbf{z}}^{-1}\mathbf{A} + \mathbf{R}_{\mathbf{x}}^{-1}\right)^{-1}\mathbf{A}^T\mathbf{R}_{\mathbf{z}}^{-1}\mathbf{y}, \tag{3.7}$$

where $\mathbf{R}_{\mathbf{x}} \triangleq \mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right)$ is the covariance matrix of $\mathbf{x}$, and $\mathbf{R}_{\mathbf{z}} \triangleq \mathbb{E}\left(\mathbf{z}\mathbf{z}^T\right)$ is the noise covariance matrix. The main issue with the LMMSE estimator is that it requires a knowledge about the signal covarince matrix $\mathbf{R}_{\mathbf{x}}$ and the noise variance $\sigma_{\mathbf{z}}^2$. Such knowledge is usually not available. When $\lambda = 0$, (3.6) reduces to the OLS estimator (3.4). One important problems is to find a regularizer that gives the best estimate of the signal (in terms of the mean-squared error (MSE) for example). Several regularization parameter selection methods exist in the literature, including Generalized Cross-Validation (GCV) [5],[6], L-curve [7],[8], quasi-optimal [9],[10] and Constrained Perturbation Regularization Approach (COPRA) [11]. The performance of these methods varies significantly depending on the nature of the problem. These regularization algorithms can be summarized as follow

- GCV: The basic idea in GCV is that if any data point $y_i$ is omitted and we compute the solution $x_i$ of the new smaller problem, the estimate of $y_i$ from $x_i$ should be a good estimate. The regularizer of GCV is obtained as the minimizer of the GCV function which suffers from the shortcoming that it may have a very flat minimum, with makes it very challenging to be located numerically [6].

- L-curve: The L-curve is a graphical tool to obtain the regularization parameter by plotting the norm function of the regularized solution $||\hat{\mathbf{x}}||_2$ versus that of the residual $||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2$. This plot (in log scale) appears as an L-shape and that the solution is located exactly in the corner of the shape. The high computational complexity is a main drawback of this method.

- Quasi-Optimal: The quasi-optimality criterion chooses the regularization parameter without taking into account the noise level. This method has a very poor performance for some noise levels.

- COPRA: It is one of the recent methods, where tools of random matrix theory have been used to solve the problem in the case of random measurement matrix. However, one of the limitations of COPRA is that it requires the system to be overdetermined [11] and cannot handle underdetermined systems.

In the next section, we will derive an analytical expression of the MSE for the Ridge Regression problem that is based on the Convex Gaussian Min-max Theorem (CGMT) framework [3], [13]. Then, in later sections, we will use this to find the optimal regularization parameter of the RLS problem. We will also derive an analytical prediction of the optimal cost function.

## 3.3 Performance Analysis

In this section, we will use the CGMT framework to study the performance of the RLS estimator in terms of the MSE. Also, we use the same framework to derive a precise

analytical expression of the optimal cost of the RLS at the optimal solution. The analysis is performed when the system dimensions ($m$ and $n$) grow simultaneously large at a fixed ratio $\delta = \lim_{n\to\infty} \frac{m}{n}$. However, we need to state the CGMT first.

### 3.3.1 Convex Gaussian Min-max Theorem (CGMT)

The CGMT theorem associates with a primary optimization (PO) problem a simplified auxiliary optimization (AO) problem from which we can tightly infer properties of the original (PO), such as the optimal cost, the norm of the optimal solution. The (AO) problem is often easier to analyze because it does not involve big random matrices but only random vectors, in contrast to the (PO) which depends on the random design matrix $\mathbf{A}$ [2], [29].

Specifically, the (PO) and (AO) optimizations are given as follows:

$$\Phi(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_w} \max_{\mathbf{u}\in\mathcal{S}_u} \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{3.8a}$$

$$\phi(\mathbf{g},\mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_w} \max_{\mathbf{u}\in\mathcal{S}_u} ||\mathbf{w}||\mathbf{g}^T\mathbf{u} - ||\mathbf{u}||\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{3.8b}$$

where $\mathbf{G} \in \mathbb{R}^{m\times n}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n, \mathcal{S}_w \subset \mathbb{R}^n, \mathcal{S}_u \subset \mathbb{R}^m$ and $\psi : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}$. Denote by $\mathbf{w}_\Phi := \mathbf{w}_\Phi(\mathbf{G})$ and $\mathbf{w}_\phi := \mathbf{w}_\phi(\mathbf{g},\mathbf{h})$ any optimal minimizers of 3.7a and 3.7b respectively. Then, the CGMT can be stated as follows.

Theorem 1 (CGMT [13]). Let $\mathcal{S}_w, \mathcal{S}_u$ be convex, compact sets, and $\psi(\mathbf{w},\mathbf{u})$ be convex-concave continuous. For any open subset $\mathcal{S} \subset \mathcal{S}_w$:

$$\lim_{n\to\infty} \mathbb{P}(\hat{\mathbf{w}}_{(\mathrm{AO})} \in \mathcal{S}) = 1 \implies \lim_{n\to\infty} \mathbb{P}(\hat{\mathbf{w}}_{(\mathrm{PO})} \in \mathcal{S}) = 1$$

The result of applying the CGMT to the RLS problem is summarized in the following theorem.

Theorem 2 (RLS error). Let $\hat{\mathbf{x}}$ be a minimizer of the RLS problem, and let $\lambda > 0$ and $\delta > 0$, then

$$\lim_{n\to\infty} \frac{1}{n}||\hat{\mathbf{x}} - \mathbf{x}_0||^2 = \alpha_*,$$

where

$$\alpha_* = \frac{\sigma_z^2 + \delta\sigma_x^2\Upsilon^2(\lambda,\delta)}{\delta(1+\Upsilon(\lambda,\delta))^2 - 1}, \tag{3.9}$$

and

$$\Upsilon(\lambda,\delta) = \frac{1-\delta+\lambda+\sqrt{(1-\delta+\lambda)^2+4\lambda\delta}}{2\delta}$$

It follows the same direction as in .

*Proof.* By following the same directions as in [30], let us start by transforming the model in 3.1 to the following model:

$$\mathbf{y} = \sigma_x^2(\mathbf{A}\tilde{\mathbf{x}_0} + \tilde{\mathbf{z}}). \tag{3.10}$$

Let $\tilde{\sigma_x}^2 = \text{var}[\tilde{\mathbf{x}_0}] = 1$, and $\sigma^2 = \text{var}[\tilde{\mathbf{z}}] = \frac{\sigma_z^2}{\sigma_x^2}$. Let us call $\tilde{\mathbf{y}} = \mathbf{A}\tilde{\mathbf{x}_0} + \tilde{\mathbf{z}}$.
We will prove for $\sigma_x^2 = 1$ case, then we generalize it to arbitrary $\sigma_x^2$. The MSE for $\sigma_x^2 = 1$ is given by

$$\lim_{n\to\infty} \frac{1}{n}||\hat{\mathbf{x}} - \tilde{\mathbf{x}_0}||^2 = \zeta_*. \tag{3.11}$$

where

$$\zeta_* = \frac{\sigma^2 + \delta\Upsilon^2(\lambda,\delta)}{\delta(1+\Upsilon(\lambda,\delta))^2 - 1}, \tag{3.12}$$

Once we proved this, it is easy to show that

$$\alpha_* = \sigma_x^2\ \zeta_* \tag{3.13}$$

For convenience, we consider the error vector $\mathbf{w} = \mathbf{x} - \tilde{\mathbf{x}_0}$, and recall $\tilde{\mathbf{y}} = \mathbf{A}\tilde{\mathbf{x}_0} + \mathbf{z}$.

Then, the RLS problem in 3.5 can be reformulated as:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \; ||\mathbf{z} - \mathbf{A}\mathbf{w}||^2 + \lambda||\mathbf{w} + \mathbf{x}_0||^2. \tag{3.14}$$

The problem in 3.14 is still not a form of a (PO) of the CGMT, so first we need to write it in form that suits the CGMT. To do so, first express the loss function of 3.14 in its dual form through the Fenchel conjugate to be

$$||\mathbf{z} - \mathbf{A}\mathbf{w}||^2 = \max_{\mathbf{u}} \sqrt{n}\mathbf{u}^T(\mathbf{z} - \mathbf{A}\mathbf{w}) - \frac{n}{4}||\mathbf{u}||^2$$

Then, 3.14 can be rewritten as:

$$\min_{\mathbf{w}} \max_{\mathbf{u}} \sqrt{n}\mathbf{u}^T\mathbf{A}\mathbf{w} - \sqrt{n}\mathbf{u}^T\mathbf{z} - \frac{n}{4}||\mathbf{u}||^2 + \lambda||\mathbf{w} + \mathbf{x}_0||^2. \tag{3.15}$$

Now, we can see that 3.15 is in the form of a (PO) problem of the CGMT. Therefore, we can determine its corresponding (AO) problem as :

$$\min_{\mathbf{w}} \max_{\mathbf{u}} \; -||\mathbf{w}||\mathbf{g}^T\mathbf{u} - ||\mathbf{u}||\mathbf{h}^T\mathbf{w} - \sqrt{n}\mathbf{u}^T\mathbf{z} - \frac{n}{4}||\mathbf{u}||^2 + \lambda||\mathbf{w} + \mathbf{x}_0||^2. \tag{3.16}$$

Computing the MSE via the (AO): for any $\epsilon > 0$, we define the set

$$\mathcal{S} = \left\{ \mathbf{v} : \left| \frac{1}{n}||\mathbf{v}||^2 - \zeta_* \right| < \epsilon \right\},$$

where $\zeta_*$ is as defined in equation 3.12. Denote the optimal solution of the (AO) problem by $\tilde{\mathbf{w}}$. We will prove that $\tilde{\mathbf{w}} \in \mathcal{S}$ with probability one. Then by applying the CGMT to the set $\mathcal{S}$ we can conclude that $\hat{\mathbf{w}} \in \mathcal{S}$ with probability one. This establishes the asymptotic expression of the MSE.

Simplifying the (AO): The vectors $\mathbf{g}$ and $\mathbf{h}$ are independent, so:

$-||\mathbf{w}||\mathbf{g}^T\mathbf{u} - \sqrt{n}\mathbf{u}^T\mathbf{z} \stackrel{d}{=} \sqrt{||\mathbf{w}||^2 + n\sigma^2}\mathbf{g}^T\mathbf{u}$. Therefore, 3.16 is equivalent to

$$\min_{\mathbf{w}} \max_{\mathbf{u}} \sqrt{||\mathbf{w}||^2 + n\sigma^2}\mathbf{g}^T\mathbf{u} - ||\mathbf{u}||\mathbf{h}^T\mathbf{w} - \frac{n}{4}||\mathbf{u}||^2 + \lambda||\mathbf{w} + \mathbf{x}_0||^2. \tag{3.17}$$

Fixing the magnitude of $\mathbf{u}$ to be $||\mathbf{u}|| = \beta$, and switching the order of min-max problem in 3.17, the optimization over the direction of $\mathbf{w}$ can be done easily by aligning it with $\mathbf{g}$. Hence, the (AO) can be simplified to:

$$\max_{\beta \geq 0} \min_{\mathbf{w}} \sqrt{n}\beta(\sqrt{\frac{||\mathbf{w}||^2}{n} + \sigma^2}||\mathbf{g}|| - \frac{\mathbf{h}^T\mathbf{w}}{\sqrt{n}}) - \frac{n\beta^2}{4} + \lambda||\mathbf{w} + \mathbf{x}_0||^2. \tag{3.18}$$

Switching the order of the optimization back, and in order to make the cost function separable, we can apply the fact that for any real $r \geq 0$, $\sqrt{r} = \min_{\tau > 0} \frac{\tau}{2} + \frac{r}{2\tau}$, this yields:

$$\min_{\tau>0} \max_{\beta \geq 0} \frac{\sqrt{n}\beta\tau||\mathbf{g}||}{2} + \frac{\sqrt{n}\beta\sigma^2||\mathbf{g}||}{2\tau} - \frac{n\beta^2}{4} + ||\mathbf{x}_0||^2\lambda + \sum_{i=1}^{n} \min_{\mathbf{w}_i} (\frac{\beta||\mathbf{g}||}{2\tau\sqrt{n}} + \lambda)\mathbf{w}_i^2 - (\beta\mathbf{h}_i - 2\lambda\mathbf{x}_{0,i})\mathbf{w}_i. \tag{3.19}$$

Let $\tau_n$ and $\beta_n$ be the optimal solutions to the optimization in 3.19. Then, if $\beta_n > 0$, the optimal $\tilde{\mathbf{w}_i}$ is given by:

$$\tilde{\mathbf{w}_i} = \frac{\beta_n\mathbf{h}_i - 2\lambda\mathbf{x}_{0,i}}{\frac{\beta_n||\mathbf{g}||}{\tau_n\sqrt{n}} + 2\lambda}. \tag{3.20}$$

Therefore, $\tau_n$ and $\beta_n$ are solutions to:

$$\min_{\tau>0} \max_{\beta>0} \frac{\sqrt{n}\beta}{2}(\tau||\mathbf{g}|| + \frac{\sigma^2||\mathbf{g}||}{\tau}) - \frac{n\beta^2}{4} + \sum_{i=1}^{n} v(\tau, \beta), \tag{3.21}$$

where

$$v(\tau, \beta) = -\frac{(\beta\mathbf{h}_i - 2\lambda\mathbf{x}_{0,i})^2}{\frac{2\beta||\mathbf{g}||}{\tau\sqrt{n}} + 4\lambda}. \tag{3.22}$$

Convergence of the (AO): After getting the simplified (AO) in 3.21, we can now

analyze its asymptotic behavior. Before beginning, we need to normalize the (AO) cost function by dividing it by $n$, and redefine $\tau = \frac{\tau}{\sqrt{\delta}}$.

Using the weak law of large numbers (WLLN), $\frac{||\mathbf{g}||}{\sqrt{n}} \xrightarrow{P} \sqrt{\delta}$, and for all $\tau > 0$ and $\beta > 0, \frac{1}{n}v(\tau,\beta) \xrightarrow{P} -\frac{\beta^2+4\lambda^2}{\frac{2\beta}{\tau}+4\lambda}$. Hence, the cost function in 3.21 converges (point-wise convergence) to:

$$D(\tau,\beta) = \frac{\beta\delta\tau}{2} + \frac{\beta\sigma^2}{2\tau} - \frac{\beta^2}{4} - \frac{\beta^2+4\lambda^2}{2\frac{\beta}{\tau}+4\lambda}. \tag{3.23}$$

Therefore, the (AO) problem can be written as:

$$\min_{\tau>0}\max_{\beta>0} D(\tau,\beta) = \beta\delta\tau + \frac{\beta\sigma^2}{\tau} - \frac{\beta^2}{2} - \frac{\beta^2+4\lambda^2}{\frac{\beta}{\tau}+2\lambda}. \tag{3.24}$$

Furthermore, it is possible to show that for $\lambda \neq 0, \beta^*$ with probability one:

The functions $\tau \mapsto \max_{\beta>0} \frac{\sqrt{n}\beta}{2}(\tau||\mathbf{g}||+\frac{\sigma^2||\mathbf{g}||}{\tau}) - \frac{n\beta^2}{4} + \sum_{i=1}^{n} v(\tau,\beta)$, and $\tau \mapsto \max_{\beta>0} D(\tau,\beta)$ are convex. Hence, one can show using theorem 2.7 in [31] that $\tau_n \xrightarrow{P} \tau^*$, where $\tau^*$ and $\beta^*$ are the optimal solution of 3.24, which can be found by solving the following system of first order optimality conditions:

$$\delta - \frac{\sigma^2}{\tau^2} - \frac{\beta^2+4\lambda^2}{(\beta+2\lambda\tau)^2} = 0 \tag{3.25}$$

$$\delta\tau + \frac{\sigma^2}{\tau} - \beta - \tau\frac{\beta^2+4\beta\lambda\tau-4\lambda^2}{(\beta+2\lambda\tau)^2} = 0. \tag{3.26}$$

Solving the above system of equations yields the following expressions for the optimal solutions:

$$\tau^* = \sqrt{\left(\frac{\left(\frac{\Upsilon(\lambda,\delta)}{1+\Upsilon(\lambda,\delta)}\right)^2+\sigma^2}{\delta - \frac{1}{(1+\Upsilon(\lambda,\delta))^2}}\right)}, \quad \text{and} \quad \beta^* = 2\lambda\Upsilon^{-1}(\lambda,\delta)\tau^*,$$

where $\Upsilon(\lambda,\delta)$ is as defined before in Theorem 2.

Proving $\tilde{\mathbf{w}} \in \mathcal{S}$:

Using the expression in 3.20, we can write: $||\tilde{\mathbf{w}}||^2 = \left|\left| \frac{\beta^* \mathbf{h} - 2\lambda}{\frac{\beta^* ||\mathbf{g}||}{\tau^* \sqrt{n}} + 2\lambda} \right|\right|^2$, then by the WLLN:

$$\frac{1}{n}||\tilde{\mathbf{w}}||^2 \xrightarrow{P} \frac{\beta^{*2} + 4\lambda^2}{(\frac{\beta^*}{\tau^*} + 2\lambda)^2}. \tag{3.27}$$

Recalling $||\mathbf{g}||/\sqrt{n} \xrightarrow{P} \sqrt{\delta}$, $\tau_n \xrightarrow{P} \tau^*$, and plugging values of $\tau^*$, and $\beta^*$ into 3.27 and after some algebraic manipulations, we can show that:

$$\frac{1}{n}||\tilde{\mathbf{w}}||^2 \xrightarrow{P} \zeta * . \tag{3.28}$$

Hence, $\tilde{\mathbf{w}} \in \mathcal{S}$, and then by the CGMT, $\hat{\mathbf{w}} \in \mathcal{S}$. Therefore,

$$\frac{1}{n}||\hat{\mathbf{w}}||^2 \xrightarrow{P} \zeta_*. \tag{3.29}$$

Finally, $\alpha_* = \sigma_x^2 \, \zeta_*$ which completes the proof of theorem 2.

□

Note: evaluation of 3.24 at optimal solutions will give us $D^*$ which we will call the optimal cost function of ridge regression problem.

Simulation results show that our analytical predictions are very accurate. Figure 3.1 illustrates both the analytical and empirical curves for the MSE of the ridge regression problem. From figure 3.2 we can see that the analytical cost function and the simulation are very close.

The theory of the approach in this section requires the problem dimensions to grow large to infinity, but simulation results show that these predictions are also accurate for dimensions ranging to few hundreds. See figure 3.3.

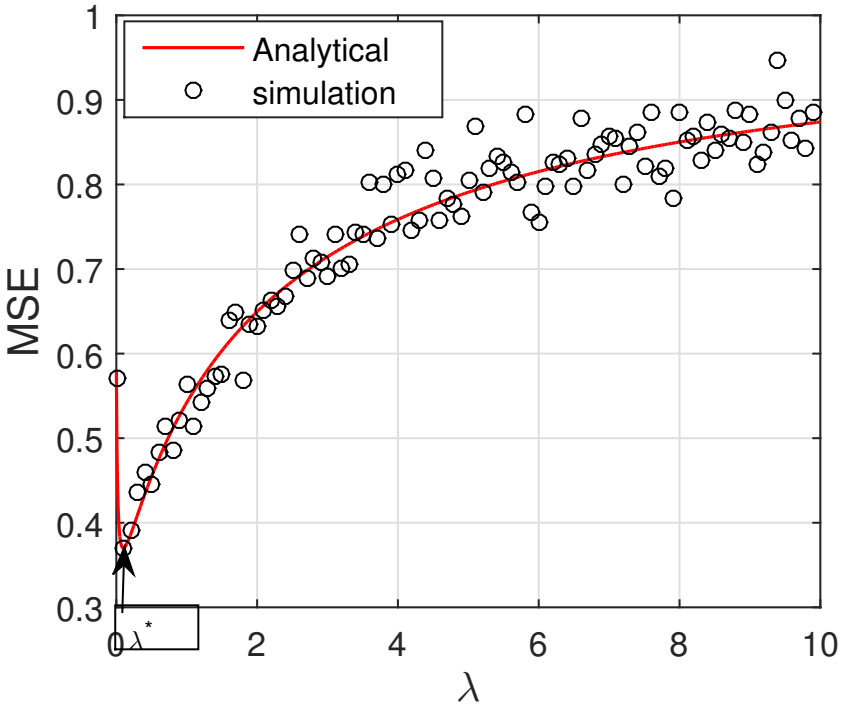


Figure 3.1: MSE, $m = 1200, n = 1500, \sigma^2 = 0.1$

## 3.4 Optimal Selection of Regularization Parameter–Known Noise Variance

From figure 3.1, we can see that there is an optimal regularizer $\lambda^*$ that gives the best MSE. So, if all the parameters are available, one can use this approach to obtain the optimal regularization parameter $\lambda^*$. This method outperforms other existing regularization methods such as GVC, COPRA and L-curve. Such comparison is demonstrated in figure 3.4.

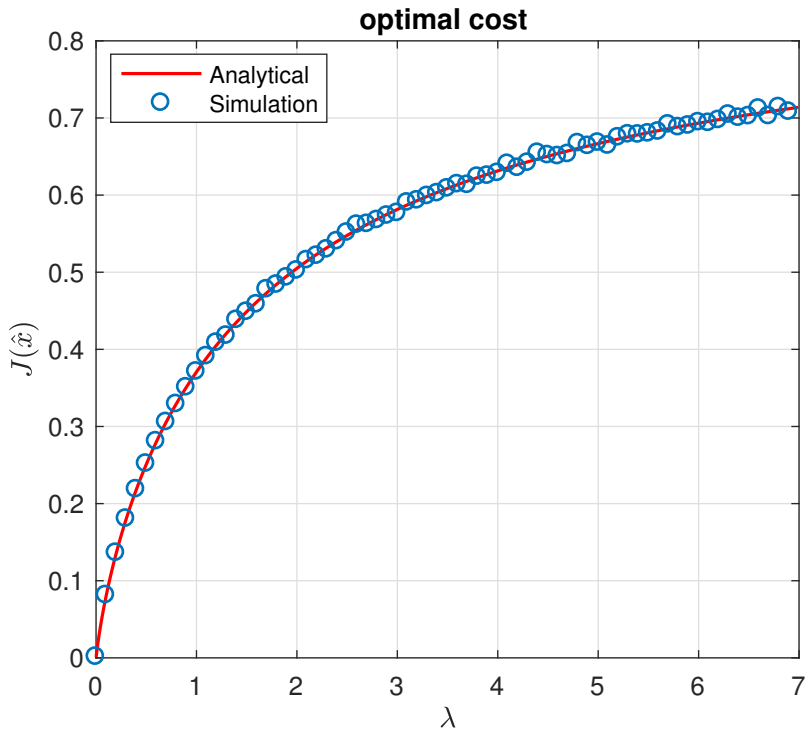


Figure 3.2: optimal cost function, $m = 1200, n = 1500, \sigma^2 = 0.1$

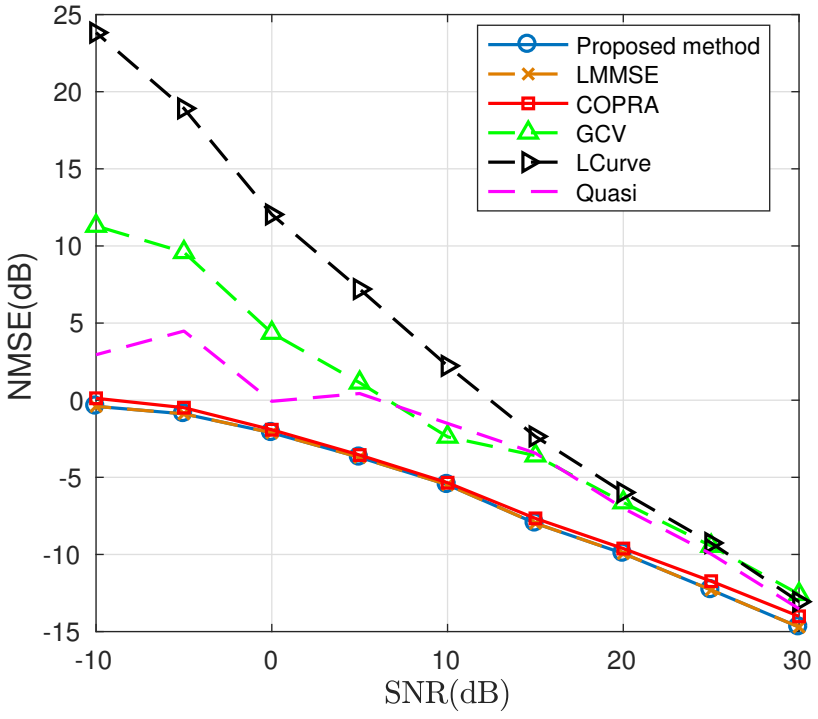


Figure 3.4: $m = 100, n = 100$, and $\mathbf{x}$ has i.i.d. Gaussian entries with $\mathcal{N}(0, 1)$, exact SNR

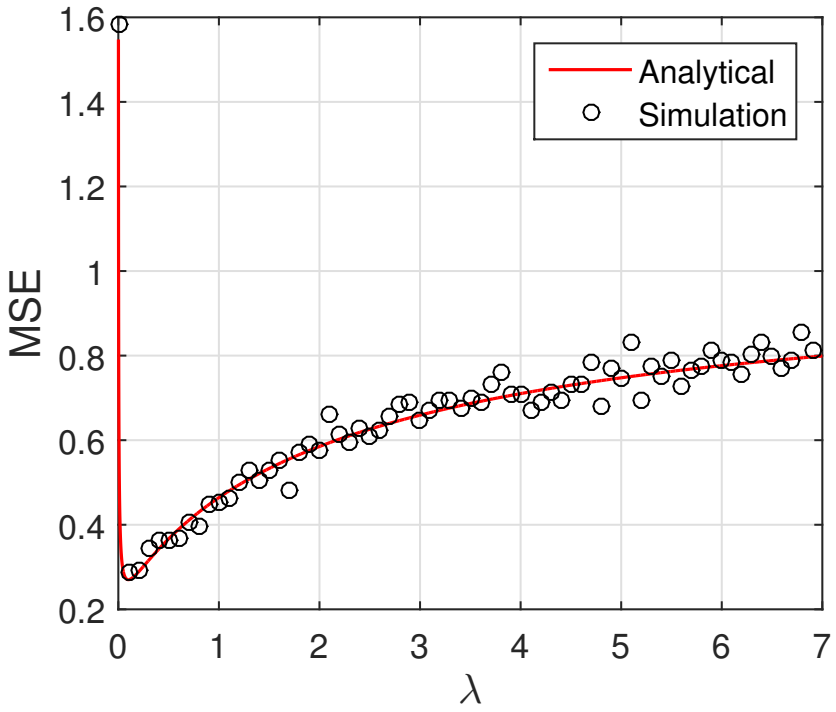


Figure 3.3: MSE, $m = 100, n = 100, \sigma_z^2 = 0.2$

## 3.5 Optimal Selection of Regularization Parameter–Unknown Noise Variance

In Section 3.4, we discussed how to optimally tune the regularization parameter if all the parameters of the model are available. However, such availability might not exist in a real scenario. Hence, one can try estimate some of these parameters.

In this section, we will show how to estimate the noise level $\sigma_z^2$ and then we use the same approach in 3.4 to get $\lambda^*$. In this approach, we will make use of the smoothing spline method introduced earlier in Chapter 2. This method is used to smooth a curve of the MSE or cost function that is obtained by a single iteration in the simulation. From figure 3.5, we can see that the resultant smoothed curve is very close to the analytical one (obtained by formulas in Section (3.2)).

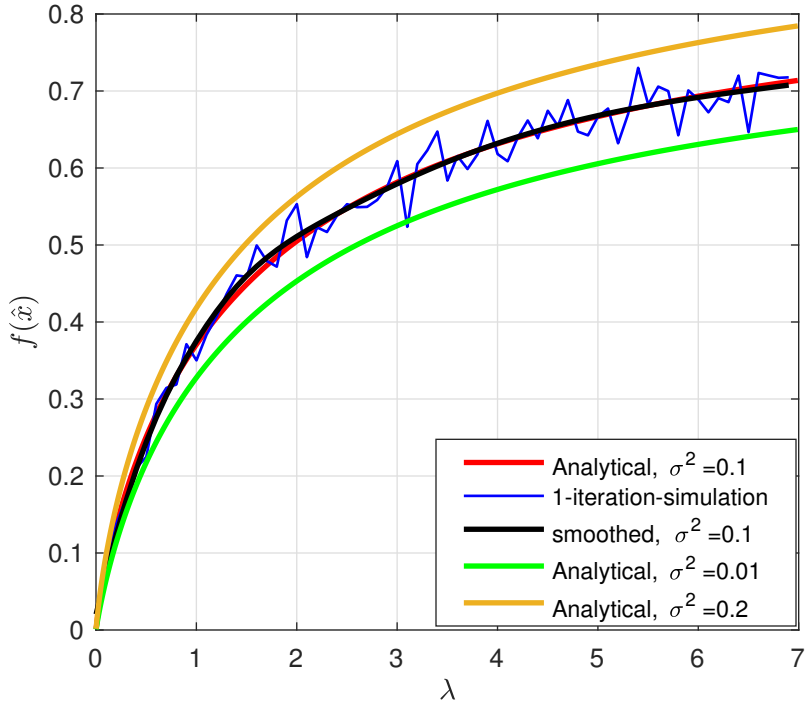


Figure 3.5: Cost function,$m = 100, n = 100, \mathbf{x}$ has i.i.d entries $\mathcal{N}(0, 1)$

This observation motivates us to use a smoothed curve of the cost function to estimate the unknown noise level $\sigma_z^2$ by matching it with the closest analytical curve (whose $\sigma_z^2$) is known. Algorithm 1 below summarizes the idea of estimating $\sigma_z^2$.

Once $\sigma_z^2$ is estimated, one can use the same approach in section 4.3 to find the optimal regularizer $\lambda^*$. That is simply, evaluate the analytical MSE expression for different $\lambda$ values and then choose $\lambda^*$ that minimizes the MSE.

From figure 3.6, we can see that the performance of the estimated $\sigma_z^2$ is very close to the exact $\sigma_z^2$. In addition, Figure 3.7 illustrates that the performance of the proposed method here outperforms other methods such as GCV, L-curve, quasi-optimal and COPRA methods with COPRA being the closest.

**Algorithm 1** Estimation of noise variance $\sigma_z^2$

1: Take a sufficient range of $\lambda$ values, save it in vector $\boldsymbol{\lambda}$.
2: Find the numerical values of optimal cost in (3.5) for the given $\lambda$ range.
3: Save the empirical results in vector $\mathbf{r}_{emp}^{'}$.
4: Use smoothing splines to smooth $(\boldsymbol{\lambda}, \mathbf{r}_{emp}^{'})$ to get smoothed $\mathbf{r}_{emp}$ vector.
5: Take two initial guesses $\sigma_{z,1}^2, \sigma_{z,2}^2$, evaluate the analytical cost function 3.24 for the $\lambda$ range for each of them
6: Evaluate the analytical cost function $\sigma\eta$ in Lemma 1 for same range of $\lambda$ as in step 1,
7: Save the results in vector $\mathbf{r}_{Th_1}, \mathbf{r}_{Th_2}$.
8: Find the closest to $\mathbf{r}_{emp}$, say $\mathbf{r}_{Th_1}$.
9: Update $\mathbf{r}_{Th_1} = \mathbf{r}_{Th_1} \pm \Delta$, where $\Delta$ is step size, if $\mathbf{r}_{Th_1} > \mathbf{r}_{emp}$ decrease, else increase
10: Choose $\sigma^2$ such that

$$||\mathbf{r}_{emp} - \mathbf{r}_{Th}||^2 \leq \rho,$$

where $\rho$ is stopping condition

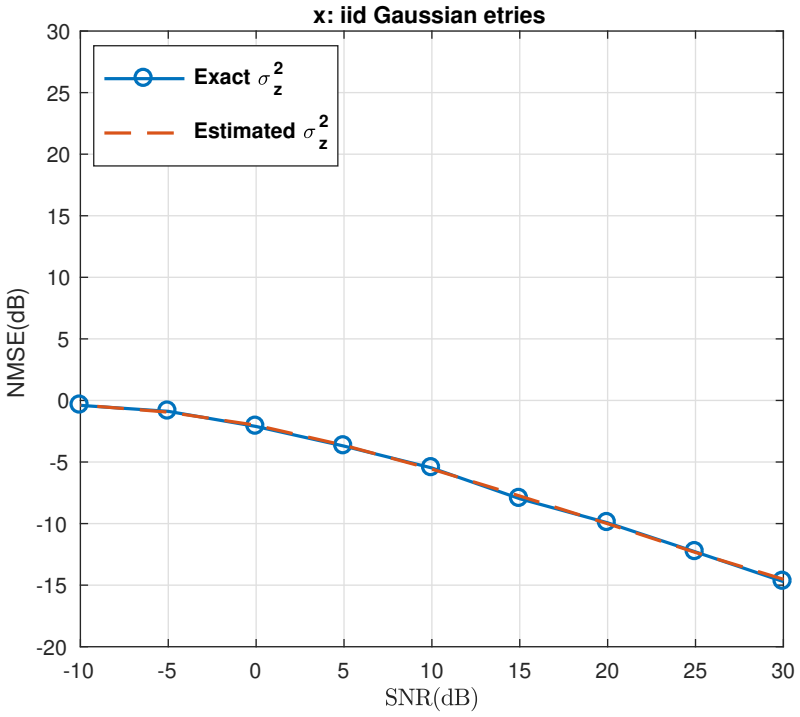


Figure 3.6: $m = 100, n = 100, \sigma_z^2 = 0.1, \mathbf{x}$ has i.i.d entries $\mathcal{N}(0, 1)$

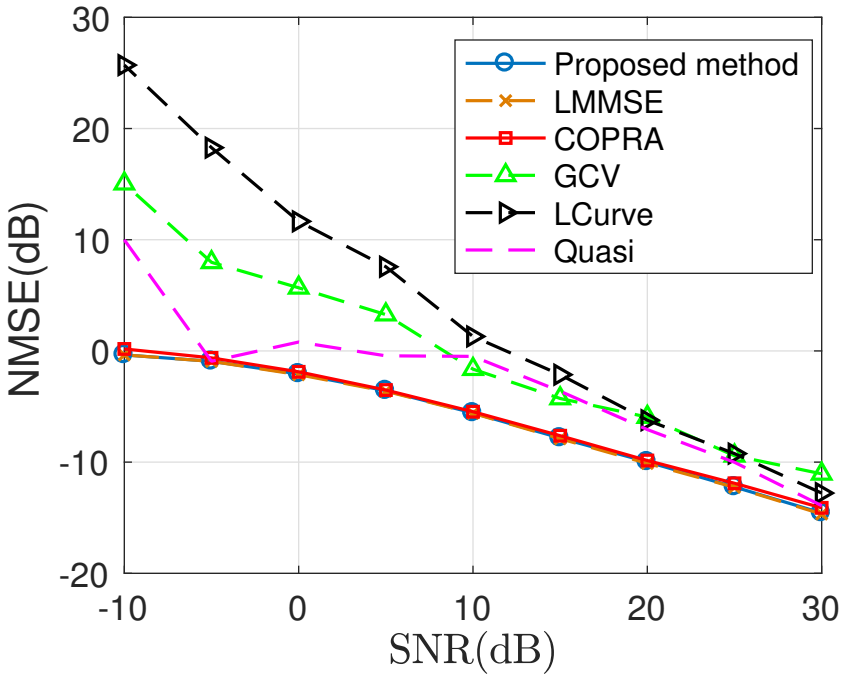


Figure 3.7: $m = 100, n = 100$, and $\mathbf{x}$ has i.i.d Gaussian entries $\mathcal{N}(0, 1)$, estimated SNR

For the case, when $m < n$, the COPRA algorithm fails, figure 3.8 illustrates this case using $\mathbf{x}$ vector that is drawn from a uniform distribution.

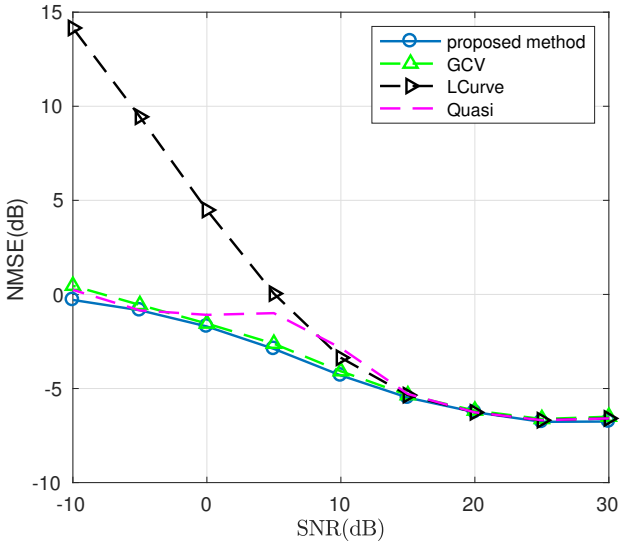


Figure 3.8: Estimated $\sigma_z^2, m = 1200, n = 1500$, $\mathbf{x}$ has i.i.d. entries uniformally drawn from $[-0.5, 0.5]$

## 3.6 Summary

In this chapter, we proposed a new asymptotic exact formulas for both the MSE and the optimal cost function of the ridge regression problem. The approach is based on the CGMT and as have been seen it is more natural. Then, we used these formulas to find the optimal regularization parameter of the estimation problem when the noise variance is unknown. The proposed method of tuning the regularization parameter is shown to outperform a set of benchmarks.

Finally, we can summarize the main procedure for finding the RLS estimate $\hat{\mathbf{x}}$ as follows:

- Estimate the noise variance $\sigma_z^2$ based on Algorithm 1.
- Now, since we have all the parameters, use the analytical formula of the MSE in Equation (3.9) to find the optimal regularization parameter $\lambda^*$.
- Solve the RLS problem by the closed form in (3.6), to obtain $\hat{\mathbf{x}}$.

# Chapter 4

# Optimal Regularization Parameter Selection for Recovery of Noisy Structured Signals

## 4.1 Introduction

In this chapter, we consider the problem of finding the optimal regularization parameter $\lambda$ for the problem of estimating an unknown signal $\mathbf{x_0} \in \mathbb{R}^n$ from a noisy linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x_0} + \mathbf{z} \in \mathbb{R}^m$. Where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the measurement matrix and $\mathbf{z} \in \mathbb{R}^m$ is the noise vector. In today's world of big data, the ambient dimension of the signal of interest $\mathbf{x_0}$ is very large. There are many applications where this is the case, such as machine learning, wireless communication, image processing, sensor networks, social networks, massive MIMO, DNA microarrays, etc [1]. However, in many applications, the desired properties of the signal may lie in a manifold of much lower dimension than the original ambient space. Such signals are called structured signals. Examples of well known structures include: sparsity, low rankness, block-sparsity, etc. In many cases of today's big data world, we are interested in the scenario of compressed measurements where $m < n$, such inverse problems are generally ill-posed unless $\mathbf{x_0}$ is structured [25]. The structure of the signal $\mathbf{x_0}$ can be captured by a convex function $f(\cdot)$. We will call $f(\cdot)$ the structure inducing regularizer function. For sparse signals, the associated structure inducing function is the $\ell_1$ norm. In the case of low-rank matrix signals, $f$ will be the nuclear norm (sum of the singular values) and when the signal is block-sparse $f$ is taken to be the $\ell_{1,2}$ norm.

In this chapter, we will study the so called Generalized LASSO algorithm, that finds an estimate $\hat{\mathbf{x}}$ of the unknown signal $\mathbf{x}_0$ by solving the following optimization problem:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{A}\mathbf{x}|| + \lambda f(\mathbf{x}), \tag{4.1}$$

where $f(\cdot)$ is the structured inducing function discussed above, and the measurement matrix $\mathbf{A}$ is assumed to have i.i.d standard normal entries (i.e. $\mathbf{A}_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1)$). The noise vector $\mathbf{z}$ is assumed to have i.i.d zero-mean normal entries with variance $\sigma^2$ (i.e. $\mathbf{z}_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$). It is important to characterize how good the estimate $\hat{\mathbf{x}}$ is. A commonly used measure of performance is the normalized squared error (NSE), which is defined as [4]:

$$\text{NSE} = \frac{||\mathbf{x}_0 - \hat{\mathbf{x}}||^2}{\sigma^2}.$$

We will focus on the problem of the selection of the regularization parameter for specific instances of the Generalized LASSO, namely $\ell_1$-square-root LASSO, $\ell_1$-LASSO and low-rank estimation. Our interest is in the case where the structure parameter values are unknown, for example, when sparsity level $k$ is not available.

## 4.2 Estimation Performance Analysis

In this section, we will use a slightly different framework that has been proposed by Thrampoulidis et al. in [2], [3] and [4] to derive a precise analysis of the Generalized LASSO in the high SNR regime ( $\sigma \to 0$). We will give a brief discussion of the general framework proposed by Theampoulidis etal. in [4] for the precise analysis of the normalized squared error (NSE) and the optimal cost of the Generalized LASSO, then we will use some results from this framework to optimize the selection of regularization parameter later in this chapter.

### 4.2.1 Preliminaries

This section introduces the background required to the statements on the NSE of the Generalized LASSO algorithm, which includes concepts from convex geometry such as: subdifferentials, Gaussian squared distance, Gaussian squared distance to the scaled subdifferential.

Subdifferential

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $\mathbf{x}_0 \in \mathbb{R}^n$ be an arbitrary vector which is not minimizer of $f$. The subdifferential of $f$ at $\mathbf{x}_0$ is defined to be the set of vectors:

$$\partial f(\mathbf{x}_0) = \big\{\mathbf{s} \in \mathbb{R}^n \mid f(\mathbf{x}_0 + \mathbf{w}) \geq f(\mathbf{x}_0) + \mathbf{s}^T\mathbf{w}, \forall \mathbf{w} \in \mathbb{R}^n\big\}.$$

This set is closed and convex [4],[32]. It, also, does not contain the origin since $\mathbf{x}_0$ is assumed not to be a minimizer of $f$.

Gaussian Squared Distance

For any vector $\mathbf{u} \in \mathbb{R}^n$, write its projection and its distance to a closed convex set $\mathcal{C} \subset \mathbb{R}^n$ as

$$\text{Proj}(\mathbf{u}, \mathcal{C}) = \underset{\mathbf{s}\in\mathcal{C}}{\text{argmin}} \; ||\mathbf{u} - \mathbf{s}||, \quad \text{and} \quad \text{dist} = ||\mathbf{u} - \text{Proj}(\mathbf{u}, \mathcal{C})||.$$

Definition 1 (Gaussian squared distance). Let $\mathbf{h} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries. The Gaussian squared distance of a set $\mathcal{C} \subset \mathbb{R}^n$ is defined as

$$\mathbf{D}(\mathcal{C}) = \mathbb{E}\big[\text{dist}^2(\mathbf{h}, \mathcal{C})\big].$$

One important quantity to us is the Gaussian squared distance to the scaled subdifferential

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) = \mathbb{E}\big[\text{dist}^2(\mathbf{h}, \lambda\partial f(\mathbf{x}_0))\big].$$

$\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is well studied in the compressed sensing literature [33]. It can be calculated in a closed form form the most important structure inducing functions such as $\ell_1-$norm, nuclear norm, etc. (see Appendix A).

### 4.2.2 First-Order Approximation

The $\ell_2$-LASSO is written as:

$$\hat{\mathbf{x}}_{\ell_2} = \underset{\mathbf{x}}{\operatorname{argmin}} \ ||\mathbf{y} - \mathbf{A}\mathbf{x}|| + \lambda f(\mathbf{x}). \tag{4.2}$$

The key idea used in [4], and [3] is to use the first-order approximation of the structure inducing function $f(\cdot)$ in 4.2 (which we will denote by $\hat{f}(\cdot)$) instead of $f(\cdot)$. Defining the error vector as $\mathbf{w} := \mathbf{x} - \mathbf{x_0}$, and recalling $\mathbf{y} = \mathbf{A}\mathbf{x_0} + \mathbf{z} = \mathbf{A}\mathbf{x_0} + \sigma\mathbf{v}$ for $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_m})$, we can get the corresponding "Approximated LASSO" problem:

$$\hat{\mathbf{w}}_{\ell_2} = \left\{ \underset{\mathbf{w}}{\operatorname{argmin}} \ ||\mathbf{A}\mathbf{w} - \sigma\mathbf{v}|| + \sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x_0})} \mathbf{s}^{\mathbf{T}}\mathbf{w} \right\}. \tag{4.3}$$

The approximated problem in 4.3 is simpler than the original one in 4.2 but still hard to analyze. Gordon's Lemma is further applied to simplify the approximated problem. Corollary 4.1 below summarizes the result of applying Gordon's Lemma to the approximated LASSO problem.

Corollary 1 (Lower Key Optimization [4]). Let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_m})$, $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_n})$and $h \sim \mathcal{N}(0, 1)$ be independent of each other. Define the following optimization problem:

$$\mathscr{L}(\mathbf{g}, \mathbf{h}) = \min_{\mathbf{w}} \ \left\{ \sqrt{||\mathbf{w}||^2 + \sigma^2}||\mathbf{g}|| - \mathbf{h}^{\mathbf{T}}\mathbf{w} + \sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x_0})} \mathbf{s}^{\mathbf{T}}\mathbf{w} \right\}. \tag{4.4}$$

Then, for any $c \in \mathbb{R}$:

$$\mathbb{P}(\mathscr{F}_{\ell_2}(\mathbf{A}, \mathbf{v}) \geq c) \geq 2\mathbb{P}(\mathscr{L}(\mathbf{g}, \mathbf{h}) - h\sigma \geq c) - 1.$$

Where $\mathscr{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ is the optimal cost function of the approximated LASSO in 4.3.

Analysis of the Lower Key Optimization

-Scalarization: $\mathscr{L}(\mathbf{g}, \mathbf{h})$ can be reduced to a scalar optimization by fixing the norm of $\mathbf{w}$ (i.e.,$||\mathbf{w}|| = \alpha$), then 4.4 becomes equivalent to:

$$\mathscr{L}(\mathbf{g}, \mathbf{h}) = \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + \sigma^2} ||\mathbf{g}|| - \alpha \underbrace{\min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} ||\mathbf{h} - \mathbf{s}||}_{\text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))} \right\}.$$

-Deterministic Result: Let $\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})$ be an optimal solution of 4.4. if $||\mathbf{g}|| \geq \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))$, then,

$$\mathscr{L}(\mathbf{g}, \mathbf{h}) = \sigma \sqrt{||\mathbf{g}||^2 - \text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))},$$

and,

$$||\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})||^2 = \sigma^2 \frac{\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))}{||\mathbf{g}||^2 - \text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))}.$$

-Probabilistic Analysis: The $\ell_2$-norm and the distance to the scaled subdifferential are 1-Lipschitz functions. Applying Lemma 5 in [4], shows that $||\mathbf{g}||^2$ and $\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))$ concentrate around their means $\mathbb{E}\big[||\mathbf{g}||^2\big] = m$, and $\mathbb{E}\big[\text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))\big] = \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Combining the deterministic results above with this results, we can conclude with the following Lemma 1.

Lemma 1 (Probablistic Result[4]). Assume that $(1 - \epsilon_L)m \geq \mathbf{D}(\lambda \partial f(\mathbf{x_0}))\epsilon_L m$ for some constant $\epsilon_L > 0$. Define,

$$\eta = \sqrt{m - \mathbf{D}(\lambda \partial f(\mathbf{x_0}))} \quad \text{and} \quad \gamma = \frac{\mathbf{D}(\lambda \partial f(\mathbf{x_0}))}{m - \mathbf{D}(\lambda \partial f(\mathbf{x_0}))}.$$

Then, for any $\epsilon > 0$, there exists a constant $c > 0$ such that, for sufficiently large $m$,

with probability $1 - exp(-cm)$,

$$\left|\mathscr{L}(\mathbf{g}, \mathbf{h}) - \sigma\eta\right| \leq \epsilon\sigma\eta, \quad \text{and} \quad \left|\frac{||\tilde{\mathbf{w}}(\mathbf{g}, \mathbf{h})||^2}{\sigma^2} - \gamma\right| \leq \epsilon\gamma.$$

In summary, the authors in [4], and [3] proved that following are true:

- Similar to $\mathscr{L}(\mathbf{g}, \mathbf{h})$, the optimal cost $\mathscr{F}_{\ell_2}(\mathbf{A}, \mathbf{v})$ of the approximated $\ell_2$-LASSO concentrates around $\sigma\eta$.
- Similar to $\frac{||\tilde{\mathbf{w}}(\mathbf{g},\mathbf{h})||^2}{\sigma^2}$, the NSE of the approximated $\ell_2$-LASSO $\frac{||\hat{\mathbf{w}}(\mathbf{A},\mathbf{v})||^2}{\sigma^2}$ concentrates around $\gamma$.
- The final step is to translate these bounds on the optimal cost and NSE of the approximated problem into bounds on the original LASSO problem 4.2. This can be done by choosing $\sigma$ small enough such that $\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$, then conclude that the NSE of the original problem concentrates around $\gamma$ with high probability.

We refer the reader to [3], and [4] for the complete analysis of the framework.

## 4.3 Selection of the Optimal Regularizer Parameter - Known Structure Parameter Values

From Lemma 1 above, one can see the implicit dependence of both the NSE and the optimal cost expressions on $\lambda, m, \partial f(\mathbf{x}_0)$. In addition, from Appendix A, we can see that the closed form expressions of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, and hence the NSE expression depend only on the structure of $\mathbf{x}_0$ and not on the unknown signal $\mathbf{x}_0$ itself.

In this section, we will show how accurate the analytical expressions of Lemma 1 compared to the empirical results by simulation. This is done for different types of structured signals (sparse, low-rank, and block sparse) for the generalized LASSO problem in 4.1. Also, we will see how one can easily tune the regularization parameter

$\lambda$ of the problem optimally in the case where all the model parameters are available. Finally, we will use results from [3] , and [4] to get a corresponding optimal regularizer for the so called $\ell_2^2$-LASSO problem 4.2.

### 4.3.1 Sparse Signal Recovery ($\ell_1$ Minimization)

One of the most celebrated structures in the literature (compressed sensing, machine learning,etc.) is sparsity. Let $\mathbf{x}_0 \in \mathbb{R}^2$ be a $k$-sparse vector, which means only k of its entries are non-zero ( usually $k << n$). The regularizer function here as discussed before will be $f(\cdot) = ||\cdot||_1$.

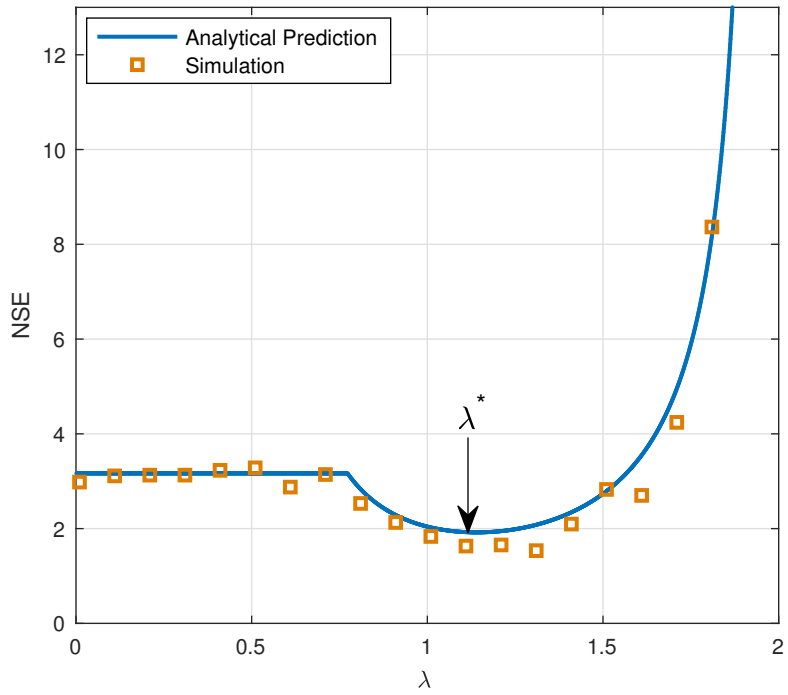


Figure 4.1: NSE for sparse signal, $n = 1500, m = 750, k = 150$

For the simulation, we used $n = 1500, m = 750, k = 150$ and fixed the noise variance to $\sigma^2 = 10^{-5}$ (for high SNR regime). The analytical predictions are obtained based on formulas from Appendix A. The results are shown in figure 4.1 and 4.2. It can be seen that the analytical predictions match the simulation. This also will be

verified for the nuclear norm and block sparse estimation in the following subsections.

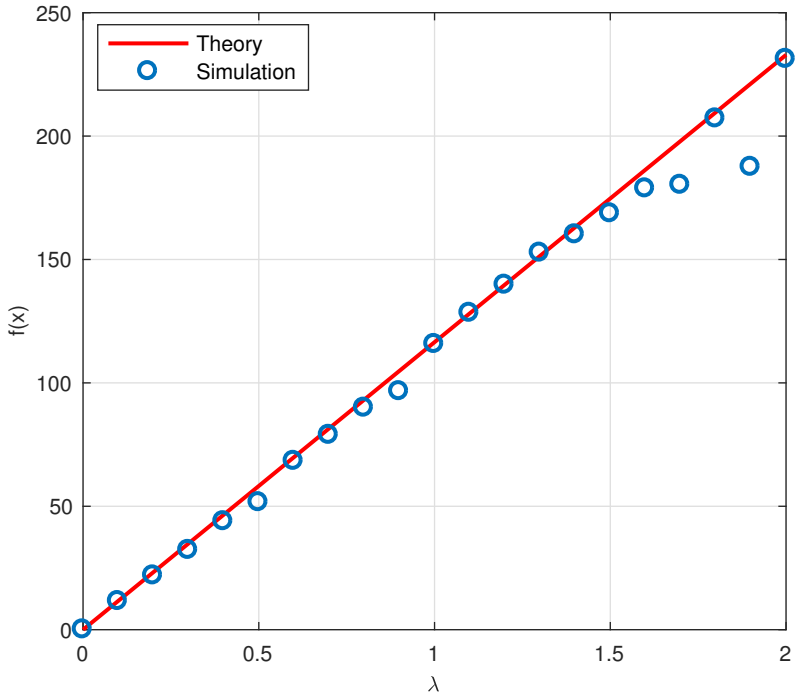


Figure 4.2: cost function

### 4.3.2 Low-Rank Matrix Recovery (Nuclear Norm Minimization)

Low-rank estimation is appropriate for many applications, including recommender systems, information retrieval, system identification, control, etc[34]. Let $\mathbf{H} \in \mathbb{R}^{m1 \times m2}$ be a low-rank matrix, then this means that it only has at most $r << \min\{m1, m2\}$ non-zero singular values. To promote this low-rank structure, the nuclear norm (also called the trace norm) is used (i.e. $f(\cdot) = ||\cdot||_*$). (How?) By definition, the nuclear norm is the sum of the singular values of a matrix. So, using a nuclear norm regularizer function encourages sparsity in the singular values vector, or equivalently for the matrix to be low-rank.

For the purposes of this chapter, consider a low-rank matrix $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$, then its vector representation is $\mathbf{x}_0 = \text{vec}(\mathbf{X}_0) \in \mathbb{R}^n, n = d^2$. Let $\mathbf{y} = \mathbf{A}.\text{vec}(\mathbf{X}_0) + \mathbf{z}$, then for the $\ell_2$ LASSO, we solve:

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times d}} ||\mathbf{y} - \mathbf{A}.\text{vec}(\mathbf{X})|| + \lambda||\mathbf{X}||_*.$$

For simulation, fix $d = 45, r = 6, m = 0.6d^2, \sigma^2 = 10^{-5}$, and for the analytical predictions, formulas from Appendix A has been used. Figure 4.3 summarizes the results.

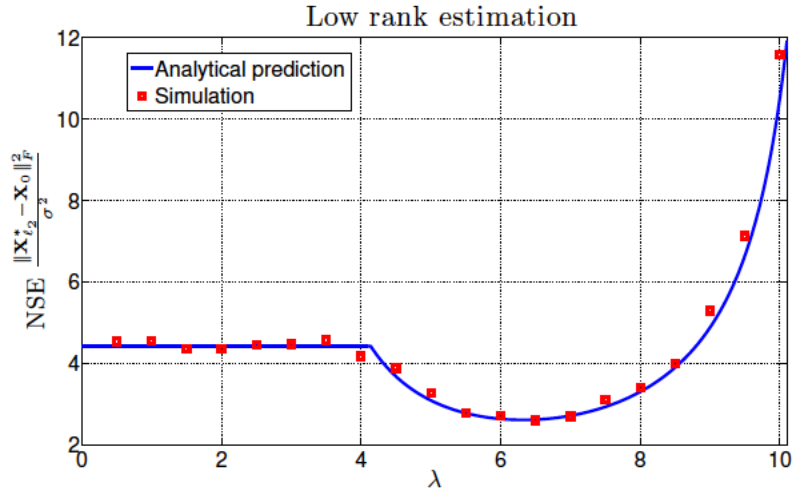


Figure 4.3: NSE for a low-rank matrix, $d = 45, m = 0.6d^2, r = 6$

### 4.3.3 Optimal Tuning of the Regularization Parameter

Both of figures 4.1, and 4.3, show that there is a value of the regularization parameter $\lambda$ for which the NSE is the best. We marked this value as $\lambda^*$ in the previous plots. This suggests that if all the model parameters are available, one can use the analytical NSE expression (i.e. $\gamma$) that was presented in Lemma 1 to find the optimal regularization parameter $\lambda^*$ that minimizes the NSE. In summary, the optimal regularizer of of the Generalized $\ell_2$(square-root)-LASSO can be selected as:

$$\lambda^* = \underset{\lambda \geq 0}{\operatorname{argmin}} \ \gamma,$$

or equivalently,

$$\lambda^* = \underset{\lambda \geq 0}{\operatorname{argmin}} \ \mathbf{D}(\lambda \partial f(\mathbf{x}_0)).$$

## 4.4 Selection of the Optimal Regularization Parameter - Unknown Structure Parameter Values

So far, we considered the problem of the optimal tuning of the regularization parameter $\lambda$ when the structure of $\mathbf{x}_0$ is known. However, in many real scenarios this knowledge is not usually available. In this section, we will develop a method for finding the optimal regularization parameter when the structure of the signal $\mathbf{x}_0$ is not available. This method takes advantage of the dependence of both the NSE $\gamma$ and optimal cost expressions (in Lemma 1) on $\lambda$ and on the structure of the signal (implicitly through $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$). The structure could be the sparsity $k$, low-rankness $r$,etc.. Our method is based on first estimating this structure of the signal, then one can use it back in the NSE formula to find the optimal regularizer.

In figure 4.4, given a received signal $\mathbf{y}$ with fixed noise variance $\sigma^2$, and unknown structure (for example the sparsity level $k$). We generated a curve of the optimal cost function for a range of $\lambda$ values by doing only one iteration of simulation, then the curve has been smoothed using the smoothing spline method introduced in chapter 2. After that, we generated curves for different sparsity levels $k$ based on the analytical expression of the optimal cost $\sigma\eta$. As we can see, there will be a $k$ value in which the empirical and analytical curves are very close. This value will be our estimate of the sparsity $k^*$. Once this value is determined, we can use the the same methodology as in the previous section to optimally tune the regularizer $\lambda$, by simply using this $k$

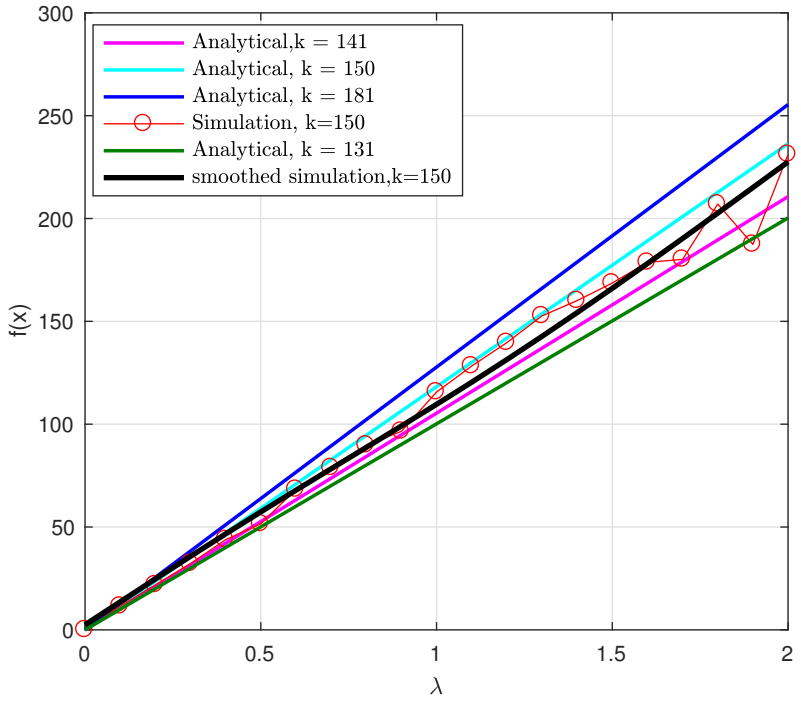


Figure 4.4: Optimal cost function of the LASSO for sparse signal with,$m = 750, n = 1500, \sigma^2 = 10^{-4}$

value in the analytical expression of the NSE to get $\lambda^*$ that corresponds to the best NSE. Algorithm 2 summarizes the idea of estimating the parameter in the case of spare signal.

For other types of structures (such as low-rankness, block sparsity,etc.), similar algorithms can be used where the sparsity $k$ is replaced by the appropriate structure ($r$: for low rank, etc.).

For the case of sparse signal, figure 4.6 shows the values of the NSE for both actual and estimated $k$ values. It can be seen that the performance using the estimated $k$ value is very close to the performance of the actual one.

**Algorithm 2** Estimation of sparsity level $k$

1: Take a sufficient range of $\lambda$ values, save it in vector $\boldsymbol{\lambda}$.
2: Find numerical values of optimal cost of 4.1 for $\lambda$ range.
3: Save the empirical results in vector $\mathbf{r}_{emp}^{'}$.
4: Use smoothing splines to smooth $(\boldsymbol{\lambda}, \mathbf{r}_{emp}^{'})$ to get smoothed $\mathbf{r}_{emp}$ vector.
5: Take sufficient $k^{'}$ integer value.
6: **for** $k = 1 : k^{'}$ **do**
7: Evaluate the analytical cost function $\sigma\eta$ in Lemma 1 for same range of $\lambda$ as in 1,
8: save the results in a vector $\mathbf{r}_{Th}^{k}$.
9: **end for**
10: Find the estimate of sparsity $k^*$ such that:

$$k^* = \underset{k}{\operatorname{argmin}} \; ||\mathbf{r}_{emp} - \mathbf{r}_{Th}^{k}||^2.$$

11: For this $k^*$ value, find $\lambda^*$ that minimizes analytical NSE in Lemma 1.

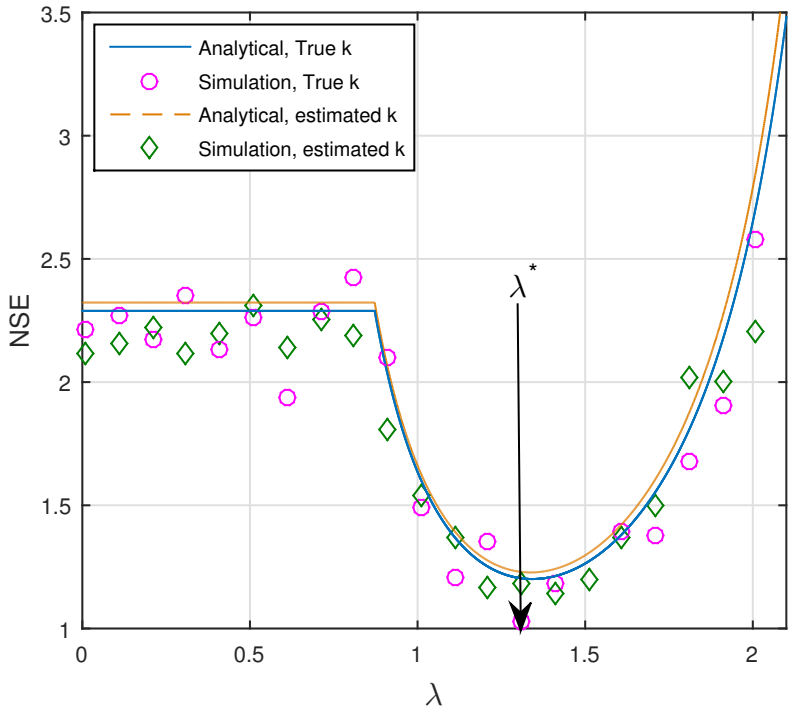


Figure 4.5: NSE for sparse signal, $m = 100, n = 500, k = 10, \sigma^2 = 10^{-4}$

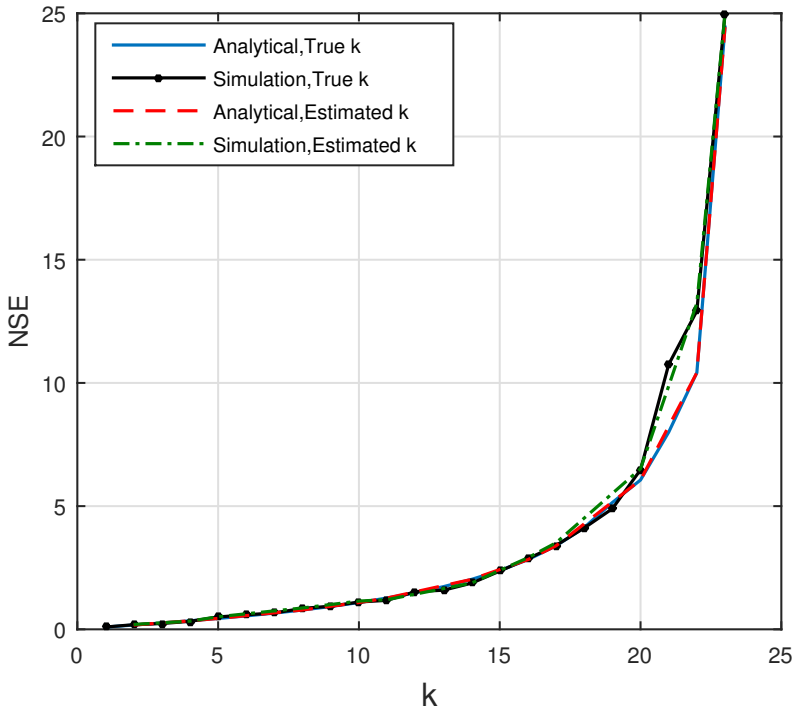


Figure 4.6: NSE for sparse signal, $m = 100, n = 500, \sigma^2 = 10^{-4}$

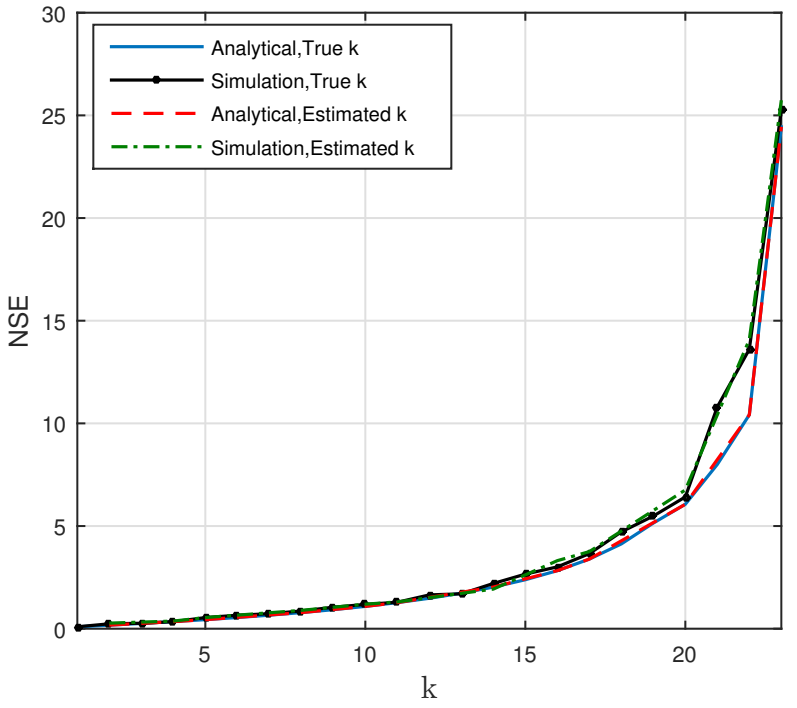


Figure 4.7: NSE for sparse signal, $m = 100, n = 500, \sigma^2 = 10^{-2}$

## 4.5 $\ell_2^2$-LASSO

Another popular version of the generalized LASSO algorithm is what is called the $\ell_2^2$-LASSO that solves:

$$\hat{\mathbf{x}}_{\ell_2^2} = \underset{\mathbf{x}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2 + \sigma\tau f(\mathbf{x}). \tag{4.5}$$

Where $f(\cdot)$ is the structure inducing function defined before. In [4], and [35], the authors provided a simple mapping between the optimal regularizers of the two problems $\lambda^*, \tau^*$ that is given by

$$\tau^* = \lambda^* \sqrt{m - \mathbf{D}(\lambda^* \partial f(\mathbf{x}_0))} \tag{4.6}$$

Therefore, to find the optimal regularizer of the $\ell_2^2$-LASSO problem, use the procedure developed for the $\ell_2$-(square-root)-LASSO and then use equation 4.6 to translate the results. Figure 4.8 illustrates the case of $k$-sparse $\mathbf{x}_0$ using the $\ell_2^2$-LASSO. It can been seen here the performance of the estimated and true sparsity are very close. Also, note that, this algorithm has the same performance as the square root-LASSO introduced before. It has been shown in [35] that the two versions of the LASSO have the same performance but the $\ell_2^2$ is more stable.

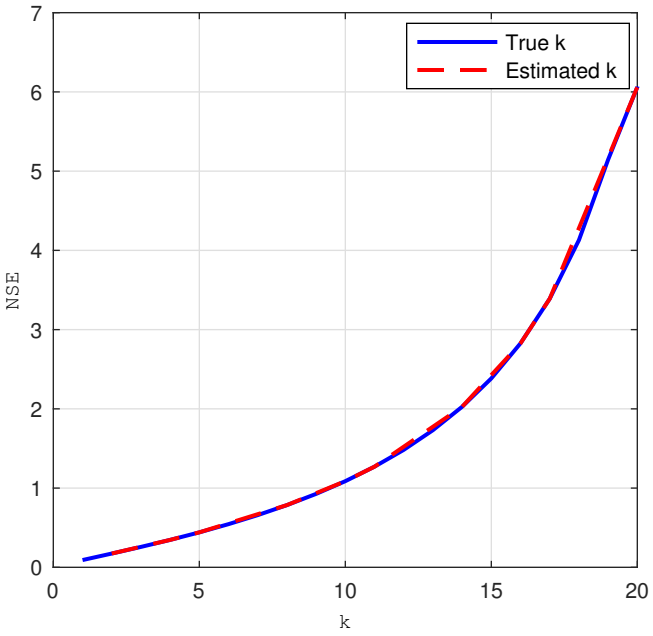


Figure 4.8: NSE for sparse signal, $m = 100, n = 500, \sigma^2 = 10^{-4}$

From figure 4.9, we can see that the simulation results support our claim that $\tau^*$ is the optimal regularizer, since using the mapping formula in 4.6 gives $\tau^* = 1.4\sqrt{750 - 493.19} \approx 18.3$ this is approximately the value of $\tau$ that minimizes the NSE in figure 4.9.

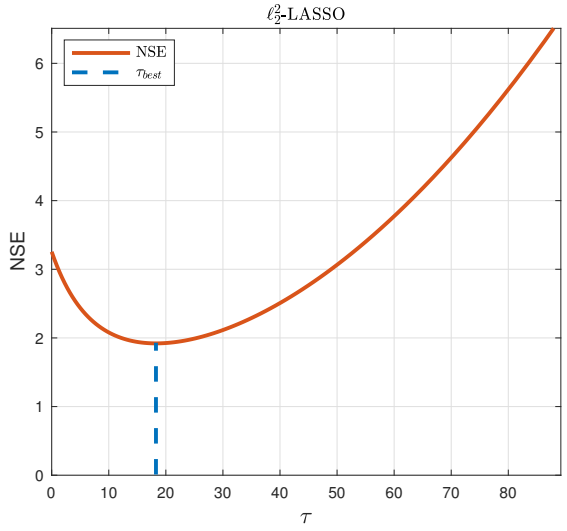


Figure 4.9: NSE for $\ell_2^2$-LASSO, of a sparse signal with $k = 150, m = 750, n = 1500, \sigma^2 = 10^{-4}$

# Chapter 5

# Concluding Remarks

## 5.1 Summary

In this thesis, we presented a new approach for optimal selection of the regularization parameter for a general regularized inverse problems. In particular, we developed new regularization algorithms for solving the Ridge regression, LASSO, square-root LASSO and low-rank generalized LASSO problems. We also presented a simple error performance analysis that is based on the CGMT framework. We focused on the case where certain model parameters are unknown and we proposed techniques that help to estimate the unknown parameters. Then, we leveraged this knowledge of the structure parameters to optimally tune the regularizer of the problem. We validated our results by numerical simulations and comparison with existing regularization methods. As a side contribution, we developed a new simple smoothing spline method that can be used to fit and interpolate noisy data.

## 5.2 Future Research Work

The work presented in this thesis can be extended in the following directions:

• Consider the case where the measurement matrix is not i.i.d. Gaussian, for example study the case of correlated matrix.

• Consider other types of structures.

• Consider other types of optimization problems.

• Analyze other cases, where $\mathbf{D}(\lambda\partial f(\mathbf{x_0}))$ is not known (in closed form) but theory

still applies. One such case is Total Variation Regularization. $\mathbf{D}$ can be obtained by simulation.

- Estimating both signal and noise variances in the case of Ridge Regression.

# REFERENCES


[1] S. Oymak, "Convex relaxation for low-dimensional representation: Phase transitions and limitations," Ph.D. dissertation, California Institute of Technology, 2015.

[2] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized m-estimators in high-dimensions," *arXiv preprint arXiv:1601.06233*, 2016.

[3] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized LASSO: A precise analysis," *CoRR*, vol. abs/1311.0830, 2013. [Online]. Available: http://arxiv.org/abs/1311.0830

[4] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Recovering structured signals in noise: least-squares meets compressed sensing, compressed sensing and its applications," *J. Vybiral*, 2015.

[5] G. Wahba, *Spline models for observational data*. Siam, 1990, vol. 59.

[6] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[7] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the l-curve," *SIAM review*, vol. 34, no. 4, pp. 561–580, 1992.

[8] P. C. Hansen and D. P. O'Leary, "The use of the l-curve in the regularization of discrete ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[9] A. Bakushinskii, "Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion," *USSR Computational Mathematics and Mathematical Physics*, vol. 24, no. 4, pp. 181–182, 1984.

[10] F. Bauer and M. Reiß, "Regularization independent of the noise level: an analysis of quasi-optimality," *Inverse Problems*, vol. 24, no. 5, p. 055009, 2008.

[11] M. Suliman, T. Ballal, A. Kammoun, and T. Y. Al-Naffouri, "Constrained perturbation regularization approach for signal estimation using random matrix theory," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1727–1731, 2016.

[12] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized lasso: A precise analysis," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*. IEEE, 2013, pp. 1002–1009.

[13] C. Thrampoulidis, S. Oymak, and B. Hassibi, "A tight version of the gaussian min-max theorem in the presence of convexity," *CoRR*, vol. abs/1408.4837, 2014. [Online]. Available: http://arxiv.org/abs/1408.4837

[14] T. Hastie and C. Loader, "Local regression: Automatic kernel carpentry," *Statistical Science*, pp. 120–129, 1993.

[15] E. T. Whittaker, "On a new method of graduation," *Proceedings of the Edinburgh Mathematical Society*, vol. 41, pp. 63–75, 1922.

[16] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal processing magazine*, vol. 16, no. 6, pp. 22–38, 1999.

[17] M. A. Unser, "Splines: a perfect fit for medical imaging," in *Medical Imaging 2002*. International Society for Optics and Photonics, 2002, pp. 225–236.

[18] S. Tirosh, D. Van De Ville, and M. Unser, "Polyharmonic smoothing splines for multi-dimensional signals with 1/∥ $\omega$∥ $\tau$-like spectra [image denoising applications]," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–297.

[19] H. L. Weinert, "A fast compact algorithm for cubic spline smoothing," *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 932–940, 2009.

[20] ——, *Fast compact algorithms and software for spline smoothing*. Springer, 2013.

[21] ——, "Efficient computation for whittaker–henderson smoothing," *Computational Statistics & Data Analysis*, vol. 52, no. 2, pp. 959–974, 2007.

[22] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," *Computational statistics & data analysis*, vol. 54, no. 4, pp. 1167–1178, 2010.

[23] D. Pollock, "Smoothing with cubic splines," 1993.

[24] C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor, *A practical guide to splines*. Springer-Verlag New York, 1978, vol. 27.

[25] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Proceedings of The 28th Conference on Learning Theory*, 2015, pp. 1683–1709.

[26] S. M. Kay, *Fundamentals of Statistical Signal Processing: Practical Algorithm Development*. Pearson Education, 2013, vol. 3.

[27] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[28] A. N. Tikhonov, A. Goncharsky, V. Stepanov, and A. G. Yagola, *Numerical methods for the solution of ill-posed problems*. Springer Science & Business Media, 2013, vol. 328.

[29] C. Thrampoulidis, "Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis," Ph.D. dissertation, California Institute of Technology, 2016.

[30] I. B. Atitallah, C. Thrampoulidis, A. Kammoun, T. Al-Naffouri, B. Hassibi, and M. S. Alouini, "Ber analysis of regularized least squares for bpsk recovery."

[31] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics*, vol. 4, pp. 2111–2245, 1994.

[32] R. T. Rockafellar, "Convex analysis. princeton landmarks in mathematics," 1997.

[33] S. Oymak and B. Hassibi, "Sharp mse bounds for proximal denoising," *Foundations of Computational Mathematics*, pp. 1–65, 2013.

[34] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *American Control Conference, 2004. Proceedings of the 2004*, vol. 4. IEEE, 2004, pp. 3273–3278.

[35] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Simple error bounds for regularized noisy linear inverse problems," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 3007–3011.

# APPENDICES

## A Explicit Formulas for Well-known Functions

In Chapter 4, we discussed the use of the Gaussian squared distance to the scaled subdifferential in the NSE characterization. Here we give a closed form expressions for this quantity for three types of structures, namely sparsity, low-rankness and group sparsity.

- **$\ell_1$ Minimization**

For a k-sparse vector $\mathbf{x}_0 \in \mathbb{R}^n$. Let $\beta = \frac{k}{n}$,then the closed form expression of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is given by [3], [4]:

$$\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{n} = (1+\lambda^2)(1-(1-\beta\text{erf}(\frac{\lambda}{\sqrt{2}}))) - \sqrt{\frac{2}{\pi}}(1-\beta)\lambda\exp(-\frac{\lambda^2}{2}) \qquad \text{(A.1)}$$

.

- **Nuclear Norm Minimization**

Assume $\mathbf{X}_0$ is $d \times d$ matrix of rank $r$ and $\mathbf{x}_0$ is its vector representation where $n = d^2$ and $\beta = \frac{r}{d}$ is fixed, then [3]:

$$\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{n} = [2\beta-\beta^2+\beta\lambda^2]+[(1-\beta)\lambda^2\Psi_0(\nu)+(1-\beta)^2\Psi_2(\nu)-2(1-\beta)^{3/2}\lambda\Psi_1(\nu)], \qquad \text{(A.2)}$$

Where,

$$\psi(x) = \begin{cases} \frac{1}{\pi}\sqrt{4-x^2}, & 0 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

$$\Psi_i(x) = \int_x^\infty x^i \psi(x) dx,$$

and

$$\nu = \frac{\lambda}{2\sqrt{1-\beta}}.$$

- **Block Sparse Minimization**

$$\mathbf{D}(\lambda \partial f(\mathbf{x}_0)) = k(b+\lambda^2) + [\Psi_1(\lambda^2) + \Psi_0(\lambda^2)\lambda^2 - 2\Psi_{\frac{1}{2}}(\lambda^2)\lambda](t-k), \tag{A.3}$$

Where,

$$\Psi_i(x) = \int_x^\infty x^i \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} \exp(-\frac{x}{2}) dx.$$