International Journal of Analysis and Applications
Volume 19, Number 2 (2021), 264-279
URL: https://doi.org/10.28924/2291-8639
DOI: 10.28924/2291-8639-19-2021-264

SMOOTHING APPROXIMATIONS FOR LEAST SQUARES MINIMIZATION WITH L_1-NORM REGULARIZATION FUNCTIONAL

HENRIETTA NKANSAH*, FRANCIS BENYAH, HENRY AMANKWAH

Department of Mathematics, University of Cape Coast, Cape Coast, Ghana

**Corresponding author:* hnkansah@ucc.edu.gh

Abstract. The paper considers the problem of least squares minimization with L_1-norm regularization functional. It investigates various smoothing approximations for the L_1-norm functional. It considers Quadratic, Sigmoid and Cubic Hermite functionals. A Tikhonov regularization is then applied to each of the resulting smooth least squares minimization problem. Results of numerical simulations for each smoothing approximation are presented. The results indicate that our regularization method is as good as any other non-smoothing method used in developed solvers.

1. Introduction

We consider the problem

$$\min_{\alpha} g(\alpha) = f(\alpha) + \lambda j(\alpha) \tag{1.1}$$

where $f(\alpha)$ is smooth, $j(\alpha)$ is non-smooth and $\lambda > 0$ is the regularization parameter. In particular, we examine $f(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2$ and $j(\alpha) = \|\alpha\|_1$.

Therefore, the problem becomes

$$\min_{\alpha} g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \|\mathbf{L}\alpha\|_1 , \tag{1.2}$$

and $\mathbf{L}$ is the $p \times n$ discrete approximation of the $(n-p)$-th derivative operator.

Received October 8th, 2019; accepted November 4th, 2019; published March 5th, 2021.

2010 *Mathematics Subject Classification.* 68W25.

Key words and phrases. least squares minimization; regularization; smoothing approximations; over-determined systems.

In this paper, we focus on over-determined linear model of the form

$$\mathbf{X}\alpha = \mathbf{y},$$

where $\alpha \in \Re^p$ is the vector of unknowns, $\mathbf{y} \in \Re^m$ is the vector of observations and $\mathbf{X} \in \Re^{m \times p}$ is the decision matrix. When $m \geq p$ and the columns of $\mathbf{X}$ are linearly independent, we can determine α by solving the least squares problem of minimizing the quadratic loss $\|\mathbf{X}\alpha - \mathbf{y}\|_2^2$, where $\|\mathbf{v}\|_2 = (\sum_i \mathbf{v}_i^2)^{\frac{1}{2}}$ denotes the L_2-norm of $\mathbf{v}$ and $\|\alpha\|_1 = \sum |\alpha_i|$. When m is not large enough compared to p, a simple least-squares problem leads to over-regularization.

1.1. **Background.** The L_1-norm minimization has attracted attention in data fitting as an effective technique for solving over-determined systems of linear equations and has also been proposed for regularization ([7], [10], [11]). The L_1- norm is a matrix norm that penalizes the sum of maximum absolute values of each row. This regularizer encourages row sparsity, that is, it encourages the entire rows of the matrix to have zero elements. In essence, this type of regularization aims at extending the L_1 framework for learning sparse models to a setting where the goal is to learn a set of sparse models. Learning algorithms based on L_1 regularized loss functions have had a relatively long history in machine learning, covering a wide range of applications such as sparse sensing ([5], [6]), L_1-logistic regression [8] and structure learning of Markov networks [9]. A well known property of L_1 regularized models is their ability to recover sparse solutions. Because of this, they are suitable for applications where discovering significant features is of value and where computing features is expensive. In addition, it has been shown that in some cases, L_1 regularization can lead to sample complexity bounds that are logarithmic in the number of input dimensions, making it suitable for learning in high dimensional spaces [9]. [11] developed an interior point algorithm for optimizing a twice differentiable objective regularized problem with an L_1- norm. One of the limitations of this approach is that it requires the exact computation of the Hessian of the objective function. This might be computationally expensive for some applications both in terms of memory and time. An alternative approach was proposed by [10], who combined a gradient-descent method with independent L_∞ projections. For the special case of a linear objective function, the regularization problem can be expressed as a linear programme [12]. While this is feasible for small problems, it does not scale to problems with large number of variables. ([13], [16]) also proposed an L_1 projection algorithm which is a special case of the algorithm where $m = 1$. The derivation of the general case for L_1 regularization is significantly more involving, as it requires reducing a set of L_∞ regularization problems tied together through a common L_1 -norm to a problem that can be solved efficiently. Similar to the L_1- norm, the L_2 -norm has also been proposed for sparse approximation. This norm penalizes the sum of the L_2- norms of each row ([14], [15],[17]).

A standard technique to prevent over-regularization is Tikhonov regularization or the L_2-norm [3] given by

$$\min_{\alpha} \quad g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \|\alpha\|_2^2 . \tag{1.3}$$

The L_2- regularized least squares problem (LSP) has an analytic solution of

$$\alpha_{L_2} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

Solutions to the LSP can either be computed by direct methods or by nondirect method, that is, applying iterative methods to the linear system of equations $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\alpha = \mathbf{X}^T\mathbf{y}$. Iterative methods are efficient especially when there are fast algorithms for the matrix-vector multiplications with the decision matrix $\mathbf{X}$ and its transpose $\mathbf{X}^T$.

Singular Value Decomposition of $\mathbf{X}$ in component form is

$$\alpha_{reg} = \sum_{i=1}^{p} f_i \frac{\mathbf{U}_i^T\mathbf{y}}{\sigma_i}\mathbf{V}_i \tag{1.4}$$

where f_i are the filter factors and is given by

$$f_i = \frac{\sigma_i^2}{\lambda + \sigma_i^2}.$$

If $\lambda << \sigma_i^2,\ f_i \approx 1,$ which indicates that the filter factors has no effect on the solution. Thus, Equation (1.4) is without regularization.

However, if

$\lambda >> \sigma_i^2, \quad f_i \approx \frac{\sigma_i^2}{\lambda}$, and $\frac{\sigma_i^2}{\lambda} \to 0$. This indicates that the filter factors has effect on the solution. In this case, it reduces the effect of the smaller singular values. Thus, Equation (1.4) gives the solution with regularization.

Another technique is the L_1-Regularized Least Squares in which we substitute a sum of absolute values for the sum of squares used in the L_2-norm regularization, to obtain Equation (1.2). This problem in Equation (1.2) always has a solution but it needs not be unique.

The first part of $g(\alpha)$ is smooth but the second part is non-smooth. In this paper, we explore three smoothing approximations that can be used to replace the L_1-norm regularized term thereby enabling us to apply the Tikhonov regularization method. These approximations are the Quadratic Approximation of a function [2], Sigmoid Function Approximation [1] and the Cubic Hermite Approximation. These three approximations are used to obtain a regularized solution to the least-squares problem in the case where $\mathbf{L} = \mathbf{I}_p$, which is the Tikhonov regularization of order zero. In each case, we will compare the solution from our regularization method with that of the Modified Newton's Method, which is mostly used in the literature. We begin by implementing the Modified Newton's Method used by Lee et al.(2006) for solving an unconstrained optimization problem in order to ascertain the challenges associated with the method.

2. Smoothing Approximations

2.1. Quadratic Approximation. Lee et al. (2006), proposed a method for transforming the non-differentiable L_1-norm function into a differentiable function by replacing it with a differentiable approximation. For a one dimensional case, the approximation to the absolute value function is given by

$$|x| \approx \sqrt{x^2 + \epsilon}.$$

To determine the best approximate solution, we first examine the nature of the plot for various values of ϵ. Approximation of the absolute value function for different values of ϵ, is given in Figure 1.

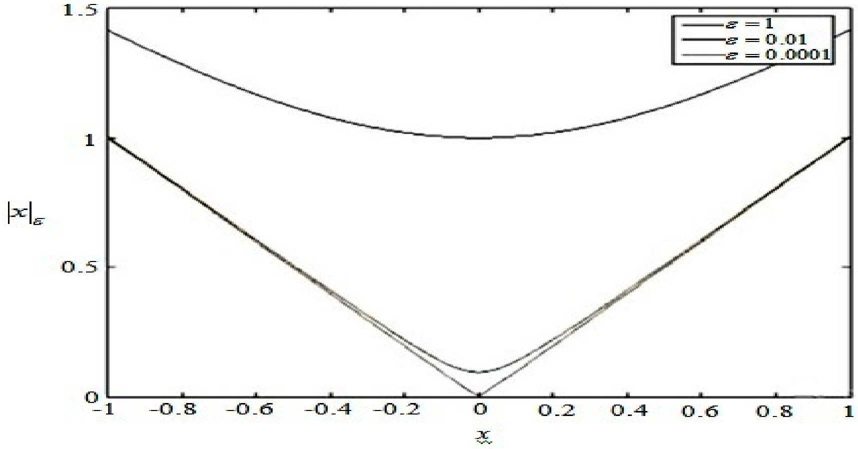


Figure 1. Quadratic Approximation of $|x|_\epsilon$ for various values of the approximation parameter, ϵ.

Figure 1 indicates that

$$\lim_{\epsilon \to 0} |x|_\epsilon = |x|.$$

Thus, we choose $\epsilon = 0.0001$ for the subsequent implementation. The gradient, $\nabla(|x|_\epsilon)$, and the Hessian, $\nabla^2(|x|_\epsilon)$, of the smoothing approximation of the absolute value function given in single variable form are derived as follows:

$$\nabla(|x|_\epsilon) = \frac{x}{\sqrt{x^2 + \epsilon}} \quad \text{and} \quad \nabla^2(|x|_\epsilon) = \frac{\epsilon}{\left(\sqrt{x^2 + \epsilon}\right)^3}.$$

For $\mathbf{x} \in \Re^p$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| \approx \sum_{i=1}^{p} |x_i|_\epsilon.$$

The loss function given in Equation (1.2) therefore becomes

$$g(\alpha) \approx \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \mu \sum_i^p \sqrt{\alpha_i^2 + \epsilon}. \tag{2.1}$$

The regularized solution of (2.1)is given as

$$\alpha_\mu = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mu = \frac{1}{2}\lambda\epsilon^{-\frac{1}{2}}$. The regularized solution α_μ written in terms of singular value decomposition (SVD) is given in component form as

$$\alpha_\mu = \sum_{i=1}^p \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i,$$

where σ_i are the singular values of the matrix $\mathbf{X}$.

2.1.1. *Derivation of Analytic Solution using Lee-Quadratic Approximation.* Let

$$k(\mathbf{x}) = \sum_i^p \sqrt{x_i^2 + \epsilon} = \|\mathbf{x}\|_\epsilon \approx \|\mathbf{x}\|_1 .$$

Since $k(\mathbf{x})$ is differentiable at $x = 0$, a Taylor's expansion about $\mathbf{x} = 0$ is given as

$$\begin{aligned} k(\mathbf{x}) &\approx k(\mathbf{x}_0) + \nabla k(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T\nabla^2 k(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \cdots \\ &\approx k(0) + \nabla k(0)^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\nabla^2 k(0)\mathbf{x} + \cdots , \end{aligned}$$

where

$$k(0) = p\epsilon^{\frac{1}{2}}, \quad \nabla k(0) = \nabla k(\mathbf{x}) = \left.\frac{x_i}{\sqrt{x_i^2 + \epsilon}}\right|_{x=0} = 0,$$

and

$$\nabla^2 k(0) = \nabla^2 k(\mathbf{x}) = \left.\frac{\epsilon}{\left(\sqrt{x_i^2 + \epsilon}\right)^3}\right|_{x=0} = \begin{pmatrix} \epsilon^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \epsilon^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \epsilon^{-\frac{1}{2}} \end{pmatrix} = \epsilon^{-\frac{1}{2}}\mathbf{I}_p.$$

Therefore,

$$k(\mathbf{x}) = p\epsilon^{\frac{1}{2}} + \frac{1}{2}\mathbf{x}^T\epsilon^{-\frac{1}{2}}\mathbf{x}\mathbf{I}_p.$$

Thus, $g(\alpha)$ in (2.1) becomes

$$g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda\Big(p\epsilon^{\frac{1}{2}} + \frac{1}{2}\alpha^T\epsilon^{-\frac{1}{2}}\alpha\Big) \tag{2.2}$$

which is now differentiable. Finding the gradient of $g(\alpha)$ and equating to zero gives

$$\begin{aligned} 2\mathbf{X}^T\mathbf{X}\alpha - 2\mathbf{X}^T\mathbf{y} + \lambda\alpha\epsilon^{-\frac{1}{2}} &= 0 \\ \mathbf{X}^T\mathbf{X}\alpha + \frac{1}{2}\lambda\alpha\epsilon^{-\frac{1}{2}} &= \mathbf{X}^T\mathbf{y} \\ \Big(\mathbf{X}^T\mathbf{X} + \frac{1}{2}\lambda\epsilon^{-\frac{1}{2}}\mathbf{I}\Big)\alpha &= \mathbf{X}^T\mathbf{y} \end{aligned}$$

$$\alpha_\mu = \Big(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}\Big)^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.3}$$

where $\mu = \frac{1}{2}\lambda\epsilon^{-\frac{1}{2}}$.

The method usually considered in the literature after obtaining a smoothing approximation to replace the L_1-norm functional is an unconstrained optimization method known as the Modified Newton's Method. To compare our results using regularization with the Modified Newton's Method, we now implement this method.

The algorithm based on the implementation of Modified Newton's Method is formulated as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k\mathbf{H}(\mathbf{x}_k)^{-1}\nabla g(\mathbf{x}_k),$$

where $\mathbf{x}_{k+1}$ is the next iterate, $\mathbf{x}_k$ is the current iterate, $\nabla g(\mathbf{x}_k)$ is the gradient at the current iterate $\mathbf{x}_k$, $\beta_k > 0$ is the step size and $\mathbf{H}(\mathbf{x}_k)$ is the Hessian at the current iterate.
From (2.1), the gradient of $g(\alpha)$ is given as

$$\nabla g(\alpha) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda G(\alpha),$$

where

$$G(\alpha) = \Big[\alpha_1(\alpha_1^2+\epsilon)^{-\frac{1}{2}},\ \alpha_2(\alpha_2^2+\epsilon)^{-\frac{1}{2}},\ \cdots,\ \alpha_p(\alpha_p^2+\epsilon)^{-\frac{1}{2}}\Big]^T,$$

and the Hessian is also given as

$$\mathbf{H}(\alpha) = 2\mathbf{X}^T\mathbf{X} + \epsilon\lambda\mathbf{h}(\alpha),$$

where

$$\mathbf{h}(\alpha) = \operatorname{diag}\Big[(\alpha_1^2+\epsilon)^{-\frac{3}{2}},\ (\alpha_2^2+\epsilon)^{-\frac{3}{2}},\ \cdots,(\alpha_p^2+\epsilon)^{-\frac{3}{2}}\Big].$$

We now consider the implementation of the algorithm.

2.1.2. *Numerical Experiment.* To illustrate our results, we make use of the 12 × 7 Hilbert submatrix of the 12 × 12 Hilbert matrix which constitute an overdetermined system. Hilbert matrices are known to be very ill-conditioned because the coefficient matrix $\mathbf{X}^T\mathbf{X}$ is almost near zero. $\mathbf{y}$ is chosen such that the true solution is $\alpha = [1,\ 1,\ 1,\ 1,\ 1, 1,\ 1]^T$. We want to find $\alpha \in \Re^p$ such that $\mathbf{X}\alpha = \mathbf{y}$.

Numerical simulations are performed to obtain an optimal regularization parameter μ which will hopefully give a solution close to the true solution. We define $\mu = 10^{-16}$ for the Modified Newton's Method and $\mu = 10^{-30}$ for the Regularization Method after several iterations.

Script-files are created in OCTAVE 4.0 to compute the solutions at various iterations. In the implementations, we define $\epsilon = 0.0001$, $\beta_k = \beta = 2$ and set initial guess $\mathbf{x}_0 = 0.25 * \text{ones}(7, 1)$. We also initialize $\mathbf{x} := \text{ones}(7, 1)$. Numerical simulations are performed with the best approximate solution occurring at the 81st iteration. The result of the implementation of the algorithm based on Lee's approximation is given in Table 1 along with its regularized solution.

Table 1 shows the solutions corresponding to the optimal regularization parameter of the two methods.

Table 1: Modified Newton's Method(MNM) vrs Regularization Method(RM) with Quadratic Approximation

Method	μ	$\hat{\alpha}$	$\|\|\alpha_{exact} - \hat{\alpha}\|\|$
MNM	10^{-16}	1.000000680934881	$1.31254062182729e - 003$
		0.999975996291894	
		1.000210550708668	
		0.999239773222684	
		1.001312540621827	
		0.998920937855809	
		1.000339707272387	
RM	10^{-30}	0.999999999998873	$2.07948647190648e - 009$
		1.000000000038512	
		0.999999999666076	
		1.000000001202600	
		0.999999997920514	
		1.000000001715320	
		0.999999999457774	

From Table 1, by increasing the value of μ from 10^{-16}, the solution seems to deteriorate. The iterations show that as we move away from the 81st iterate, there is not much difference between the solutions from the 82nd to the 100th iteration.

The accuracy in the computed solution of MNM corresponding to $\mu = 10^{-16}$ is just about 3 digits. The loss in the accuracy of the solution is due to the fact that the coefficient matrix $\mathbf{X}^T\mathbf{X}$ of the normal equations is ill-conditioned, with a condition number $\kappa \approx 2.31648078701200e + 015$.

These results verify the findings of Lee et al. (2006) that gives a slow convergence which is due to the ill-conditioning of $\mathbf{X}^T\mathbf{X}$. In order to overcome the undesirable effects of ill-conditioning, we make use of regularization method. It is found that the best approximate solution for the MNM occurred at $\mu = 10^{-16}$, with the step size $\beta = 2$, and at the 81st iteration. For the RM, the best approximate solution occurred at $\mu = 10^{-30}$ with nine digit accuracy.

2.2. **Sigmoid Function Approximation.** In this section, we consider the Sigmoid Function approximation to the L_1-norm functional. It takes advantage of the non-negative projection operators

$$(x)_+ = \max(x, 0) \quad \text{and} \quad (-x)_+ = \max(-x, 0).$$

This projection function can be smoothly approximated by the integral of a sigmoid function [1] given as

$$(x)_+ \approx p(x, \kappa) = x + \frac{1}{\kappa}\log(1 + e^{-\kappa x})$$

and

$$(-x)_+ \approx p(-x, \kappa) = -x + \frac{1}{\kappa}\log(1 + e^{\kappa x}).$$

The functions $p(x, \kappa)$ and $p(-x, \kappa)$ are members of a class of smoothing functions presented in [1]. These smoothing approximations of the projections have been used to transform the standard L_1-norm formulation into an efficiently solved unconstrained problem [2]. By combining $p(x, \kappa)$ and $p(-x, \kappa)$, we obtain the identity

$$|x| = (x)_+ + (-x)_+$$

where $(x)_+ + (-x)_+$ are the left and right parts of $\text{abs}(x) = |x|$.

We arrive at a smoothing approximation for the absolute value function that consists of the sum of the integral of two sigmoid functions given by

$$\begin{aligned} |x| &\approx (x)_+ + (-x)_+ = p(x, \kappa) + p(-x, \kappa) \\ &= \frac{1}{\kappa}[\log(1 + e^{-\kappa x}) + \log(1 + e^{\kappa x})] \\ &\overset{def}{=} |x|_\kappa \end{aligned}$$

A graph of different values of the parameter κ in approximating the absolute value function is given in Figure 2.

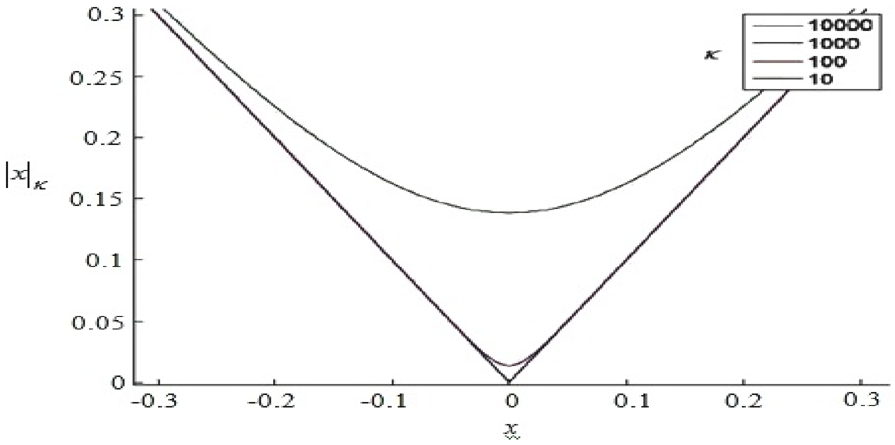


FIGURE 2. Sigmoid Approximations of $|x|_\kappa$ for various values of κ.

Figure 2 indicates that $|x|_\kappa \to |x|$ as $\kappa \to \infty$. That is,

$$\lim_{\kappa\to\infty} |x|_\kappa = |x|\,.$$

Thus, it would be suitable to choose $\kappa = 1000$ for the subsequent implementation.

Given the smoothing approximation, the gradient $\nabla(|x|_\kappa)$ and the Hessian $\nabla^2(|x|_\kappa)$ in single variable form is derived as follows:

$$\begin{aligned}
|x|_\kappa &= \frac{1}{\kappa}\Big[\log(1+e^{-\kappa x}) + \log(1+e^{\kappa x})\Big] \\
\nabla(|x|_\kappa) &= \frac{1}{\kappa}\Big[\frac{-\kappa e^{-\kappa x}}{1+e^{-\kappa x}} + \frac{\kappa e^{\kappa x}}{1+e^{\kappa x}}\Big] \\
&= \frac{(1+e^{\kappa x}) - (1+e^{-\kappa x})}{(1+e^{-\kappa x})(1+e^{\kappa x})} \\
&= \frac{(1+e^{\kappa x})}{(1+e^{-\kappa x})(1+e^{\kappa x})} - \frac{(1+e^{-\kappa x})}{(1+e^{-\kappa x})(1+e^{\kappa x})} \\
&= \frac{1}{1+e^{-\kappa x}} - \frac{1}{1+e^{\kappa x}}.
\end{aligned}$$

Therefore,

$$\nabla(|x|_\kappa) = (1+e^{-\kappa x})^{-1} - (1+e^{\kappa x})^{-1},$$

and

$$\begin{aligned}
\nabla^2(|x|_\kappa) &= (1+e^{-\kappa x})^{-2}\kappa e^{-\kappa x} + (1+e^{\kappa x})^{-2}\kappa e^{\kappa x} \\
&= \frac{\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}\Big[\frac{(1+e^{\kappa})^2}{(1+e^{-\kappa x})^2}(e^{-\kappa x})^2 + 1\Big] \\
&= \frac{\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}\Big[\frac{(e^{-\kappa x}(1+e^{\kappa x})^2)}{(1+e^{-\kappa x})^2} + 1\Big]
\end{aligned}$$

$$\begin{aligned} &= \frac{\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}\Big[\frac{(e^{-\kappa x}+1)^2}{(1+e^{-\kappa x})^2}+1\Big] \\ &= \frac{2\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}. \end{aligned}$$

Therefore,

$$\nabla^2(|x|_\kappa) = \frac{2\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}.$$

For $\mathbf{x} \in \Re^p$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| \approx \sum_{i=1}^{p} |x_i|_\kappa .$$

The loss function in (1.2) therefore becomes

$$g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \sum_{i}^{p} \frac{1}{\kappa}\Big[\log(1+e^{-\kappa\alpha_i}) + \log(1+e^{\kappa\alpha_i})\Big]. \tag{2.4}$$

which is now differentiable. The gradient of $g(\alpha)$ is given as

$$\nabla g(\alpha,\kappa) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda\sum_{i=1}^{p}(1+e^{-\kappa\alpha_i})^{-1} - (1+e^{\kappa\alpha_i})^{-1}.$$

A linear approximation to

$$\nabla(|\alpha|_\kappa) = (1+e^{-\kappa\alpha})^{-1} - (1+e^{\kappa\alpha})^{-1}$$

is obtained as

$$k(\alpha) = \frac{1}{2}\Big(\kappa\alpha + \frac{(\kappa\alpha)^3}{3!} + \cdots\Big)$$

after some expansion and simplification. By ignoring terms of higher order, we obtain the linear approximation

$$k(\alpha) = \frac{1}{2}\kappa\alpha.$$

Using this linear approximation and equating $\nabla g(\alpha)$ to zero, we solve the equation to obtain

$$\mathbf{X}^T\mathbf{X}\alpha + \frac{1}{2}\lambda k(\alpha) = \mathbf{X}^T\mathbf{y},$$

which gives

$$\alpha_\mu = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

where $\mu = \frac{1}{4}\lambda\kappa$.

Using the linear approximation for $k(\alpha)$, we obtain the minimization of $g(\alpha)$ as

$$\mathbf{X}^T\mathbf{X}\alpha + \frac{1}{4}\lambda\kappa\alpha = \mathbf{X}^T\mathbf{y}.$$

Thus,

$$\alpha_\mu = \Big(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}\Big)^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.5}$$

where $\mu = \dfrac{1}{4}\lambda\kappa$.

Therefore, the singular value decomposition solution in component form is given as

$$\alpha_\mu = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \mu} (\mathbf{U}_i^T \mathbf{y}) \mathbf{V}_i.$$

To implement the Modified Newton's Method for sigmoid approximation, we want to find $\alpha \in \Re^p$ such that $\mathbf{X}\alpha = \mathbf{y}$. From (2.4), the gradient of $g(\alpha)$ is given by

$$\nabla g(\alpha) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda G(\alpha)$$

where $G(\alpha) = \Big[(1+e^{-\kappa\alpha_1})^{-1} - (1+e^{\kappa\alpha_1})^{-1},\ \cdots,\ (1+e^{-\kappa\alpha_p})^{-1} - (1+e^{\kappa\alpha_p})^{-1}\Big]^T$
and Hessian

$$\nabla^2(g(\alpha)) = 2\mathbf{X}^T\mathbf{X} + 2\kappa\lambda\ \mathbf{h}(\alpha),$$

where $\mathbf{h}(\alpha) = \text{diag}\left[\dfrac{e^{\kappa\alpha_1}}{(1+e^{\kappa\alpha_1})^2}, \dfrac{e^{\kappa\alpha_2}}{(1+e^{\kappa\alpha_2})^2}, \cdots, \dfrac{e^{\kappa\alpha_p}}{(1+e^{\kappa\alpha_p})^2}\right]$.

In the implementation, we define $\kappa = 300$, $\beta_k = \beta = 3$ and set initial guess $\alpha_0 = 0.25 * \text{ones}(7,1)$. We also initialize $\alpha := \text{ones}(7,1)$. Numerical simulations are performed with the best approximate solution occurring at the 84th iterate.

The results of the implementation based on sigmoid approximation is given in Table 2 for the Modified Newton's Method and the Regularization Method.

Table 2: Modified Newton's Method(MNM) vrs Regularization Method(RM) with Sigmoid Approximation

Method	μ	$\hat{\alpha}$	$\lVert\alpha_{exact} - \hat{\alpha}\rVert$
MNM	10^{-16}	1.000005092022318 0.999820688361910 1.001571753611182 0.994327784550849 1.009789408764376 0.991954381867874 1.002532280742229	$9.78940876437595e-003$
RM	10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$

From Table 2, the accuracy in the computed solution of MNM corresponding to $\mu = 10^{-16}$ is just about 3 digits. The accuracy in that of the RM is up to about 9 digits.

2.3. **Cubic Hermite Approximation.** The Cubic Hermite approximation is a spline where each piece is a third-degree polynomial specified in Hermite form: that is, by its values and first derivatives at the end points of the corresponding domain interval. The Hermite form of a cubic polynomial defines the polynomial $p(x)$ by specifying two distinct points $[-\gamma,\ \gamma]$, and providing values for the following four equations

$$\begin{bmatrix} 0 & 1 & 2\gamma & 3\gamma^2 \\ 0 & 1 & -2\gamma & 3\gamma^2 \\ 1 & \gamma & \gamma^2 & \gamma^3 \\ 1 & -\gamma & \gamma^2 & -\gamma^3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ \gamma \\ \gamma \end{bmatrix}. \tag{2.6}$$

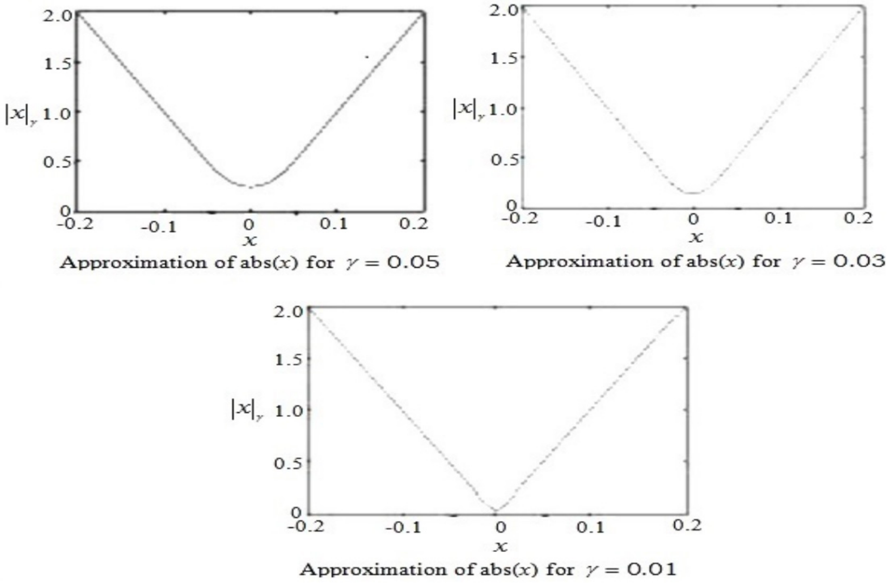
Solving for the unknown parameters in Equation (2.6), gives

$$a_0 = \frac{\gamma}{2}, \quad a_1 = 0, \quad a_2 = \frac{1}{2\gamma}, \quad a_3 = 0.$$

Therefore,

$$P(\mathbf{x}) = \frac{\gamma}{2} + \frac{1}{2\gamma}\mathbf{x}^2.$$

To determine the best approximate solution, we first examine the nature of the plot of the absolute value function for various values of γ. The graph of the abs(x) is given in Figure 3 for various values of the parameter γ.



Approximation of abs(x) for $\gamma = 0.05$

Approximation of abs(x) for $\gamma = 0.03$

Approximation of abs(x) for $\gamma = 0.01$

FIGURE 3. Cubic Hermite Approximation of $|x|_\gamma$ for various values of Approximating Parameter, γ.

Figure 3 indicates that

$$\lim_{\gamma \to 0} |x|_\gamma = |x|.$$

The implementation shows that $\gamma = 0.05$ is a more suitable choice.

Thus, the scalar Cubic Hermite approximation to the absolute value function is given as

$$|x|_\gamma \approx \frac{\gamma}{2} + \frac{1}{2\gamma}x^2.$$

The gradient $\nabla(|x|_\gamma)$ and the Hessian $\nabla^2(|x|_\gamma)$ are derived as follows:

$$\nabla(|x|_\gamma) = \frac{x}{\gamma} \quad \text{and} \quad \nabla^2(|x|_\gamma) = \frac{1}{\gamma}.$$

For $\mathbf{x} \in \Re^p$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| \approx \sum_{i=1}^{p} |x_i|_\gamma .$$

The loss function in Equation (1.2) therefore becomes,

$$g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \sum_{i}^{p} \Big(\frac{\gamma}{2} + \frac{1}{2\gamma}\alpha_i^2\Big), \tag{2.7}$$

which is now differentiable. The gradient of $g(\alpha)$ is given as

$$\nabla g(\alpha) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda G(\alpha),$$

where $G(\alpha) = (\frac{\alpha_1}{\gamma},\ \frac{\alpha_2}{\gamma},\ \cdots,\ \frac{\alpha_p}{\gamma})^T$.

Equating $\nabla g(\alpha)$ to zero and solving gives

$$\alpha_\mu = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mu = \frac{1}{2\gamma}\lambda$.

Therefore, the singular value decomposition solution in component form is given as

$$\alpha_\mu = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i.$$

To implement the Modified Newton's method for Cubic Hermite approximation, we want to find $\alpha \in \Re^p$ such that $\mathbf{X}\alpha = \mathbf{y}$.

Now, the Hessian of $g(\alpha)$ is

$$\nabla^2(g(\alpha)) = 2\mathbf{X}^T\mathbf{X} + \frac{1}{\gamma}\lambda\mathbf{I}_p.$$

In the implementation, we define $\gamma = 0.05,\ \beta_k = \beta = 3$ and set initial guess $\alpha_0 = 0.25*\text{ones}(7,1)$. We also initialize $\alpha := \text{ones}(7,1)$. Numerical simulations are performed with the best approximate solution occurring at the 87th iterate.

The result of the implementation based on Cubic Hermite approximation is given in Table 3 for the Modified Newton's Method and the Regularization Method.

Table 3: Modified Newton's Method(MNM) vrs Regularization Method(RM)

Method	μ	$\hat{\alpha}$	$\|\alpha_{exact} - \hat{\alpha}\|$
MNM	10^{-16}	0.999983772069521	$3.18885571566037e-002$
		1.000576215743305	
		0.994919717190969	
		1.018414062557087	
		0.968111442843396	
		1.026280481466755	
		0.991709642462265	
RM	10^{-30}	0.999999999998873	$2.07948647190648e-009$
		1.000000000038512	
		0.999999999666076	
		1.000000001202600	
		0.999999997920514	
		1.000000001715320	
		0.999999999457774	

We note from Tables 1, 2 and 3 that the MNM solutions vary for the three smoothing approximations considered. However, the regularized solutions are the same for each of the three smoothing approximations at $\mu = 10^{-30}$.

We now compare the three smoothing approximations using regularization with a non-smooth method which makes use of Truncated Newton interior-point method described in [4]. In that paper, they developed a Matlab Solver for Large-Scale L_1-Regularized Least Squares Problems called l1_ls. Using our value of the parameter $\mu = 10^{-30}$ in the l1_ls, we display in Table 4 the result of all three regularization methods (RM) and that of the l1_ls.

Table 4 : Summary of methods and their solutions at $\mu = 10^{-30}$

Method	Solution corresponding to $\mu = 10^{-30}$	$\|\alpha - \hat{\alpha}\|$
RM	0.999999999998873	$2.07948647190648e-009$
	1.000000000038512	
	0.999999999666076	
	1.000000001202600	
	0.999999997920514	
	1.000000001715320	
	0.999999999457774	
l1_ls	1.000000000000292	$4.72645700355656e-010$
	0.999999999990242	
	1.000000000081881	
	0.999999999715937	
	1.000000000472646	
	0.999999999624568	
	1.000000000114463	

From Table 4, it is seen that the solutions from the three smoothing approximations by regularization method is as good as the non-smooth method, which is accurate to about 10 digits.

3. Conclusion

We have considered three different methods for approximating the absolute value function, which is non-differentiable and used in the L_1-norm minimization problem. Under each of these three methods namely, the Quadratic, Sigmoid and Cubic Hermite approximations, we have obtained optimal approximate solutions by means of the Newton's method and by regularization. It is observed that the results of the Newton's method under all three methods show visible differences and produce solutions that are accurate only to at most 3 digits. However, the regularized solution using smoothing approximations produce almost the same results that are accurate to 9 digits at the same parameter value of $\mu = 10^{-30}$. This value of μ is obviously much smaller than the usual machine representation of 10^{-16} for 0. The results obtained are as good as the non-smooth methods used in developed solvers.

Conflicts of Interest: The author(s) declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] C. Chen, O.L. Mangasarian, A class of smoothing functions for nonlinear and mixed complementarity problems, Comput. Optim. Appl. 5 (1996), 97–138.

[2] S.I. Lee, H. Lee, P. Abbeel, A.Y. Ng, Efficient L1 regularized logistic regression. https://www.aaai.org/Papers/AAAI/2006/AAAI06-064.pdf (2006).

[3] A. Neumaier, Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization, SIAM Rev. 40 (1998), 636–666.

[4] K. Koh, S.-J. Kim, S. Boyd, An interior-point method for L1-regularized logistic regression, J. Mach. Learn. Res. 8 (2007), 1519-1555.

[5] W.J. Fu, Penalized Regressions: The Bridge versus the Lasso, J. Comput. Graph. Stat. 7 (1998), 397–416.

[6] D.L. Donoho, Denoising via soft-thresholding, IEEE Trans. Info. Theory, 41 (1995), 613–627.

[7] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika. 81 (1994), 425–455.

[8] B. Efron, The Estimation of Prediction Error: Covariance Penalties and Cross-Validation, J. Amer. Stat. Assoc. 99 (2004), 619–632.

[9] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2004), 407–499.

[10] M. Schmidt, G. Fung, R. Rosales, Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches, in: J.N. Kok, J. Koronacki, R.L. de Mantaras, S. Matwin, D. Mladenič, A. Skowron (Eds.), Machine Learning: ECML 2007, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007: pp. 286–297.

[11] J.A. Tropp. Just relax: Convex programming methods for subset selection and sparse approximation. IEEE Trans. Inform. Theory, 51 (3) (2006), 1030–1051.

[12] B.A. Turlach, Shape constrained smoothing using smoothing splines, Comput. Stat. 20 (2005), 81–104.

[13] B.T. Polyak, Introduction to Optimization. Optimization Software Inc. Publication Division, New York, 1987.

[14] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Sov. Math. Dokl. 27 (2) (1983), 372-376.

[15] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An Interior-Point Method for Large-Scale l_1-Regularized Least Squares, IEEE J. Sel. Top. Signal Process. 1 (2007), 606–617.

[16] P. Zhao, B. Yu, On model selection consistency of Lasso. J. Mach. Learn. Res. 7 (2006), 2541-2563.