Connection between Mixed Integer Non-Linear Programming (MINLP) and Sparse Optimization

Samah Mohammed Osman Mohammed Alamin (samaha@aims.ac.za)

سماح محمد عثمان محمد الأمين

African Institute for Mathematical Sciences (AIMS)

Supervised by: Prof. Dr. Ekaterina A. Kostina

Heidelberg University, Institute of Applied Mathematics, Germany

23 May 2019

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

In optimization problems, several types of mixed-integer non-linear programming (MINLP) have been well studied in literature due to the computational efficiency of the algorithms in finding the optimal solution of the MINLP problem. Generally, in optimization problems, the sparse regression of the objective and constraint functions provides useful information that will help us to reach the optimization goal. It has, however, many formulations such as the LASSO regression problem. To solve such kind of regression problems, the sparse optimization can be reformulated as an MINLP problem. Also, the sparse regression method can significantly improve MINLP algorithms. In this research, we show a new insight for the L_0-norm optimization from the perspective of the mixed-integer quadratic optimization (MIQP). Moreover, we show relaxations of the MIQP problems and analyze a connection between MIQP and sparse optimization problems.

Keywords

Sparse regression, LASSO regression, Mixed-Integer Quadratic Optimization, Perspective Relaxation, Best Subset Selection Problem, Penalizations.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Samah Mohamed Osman Abd-Alrhaman, 23 May 2019

Contents

1. Introduction

In general, the sparse optimization problem has a significant number of applications in different fields of science, such as compressed sensing, image processing, statistical and machine learning. In machine learning, the sparse optimization is called sparse learning (Pham, 2016). There is a vast number of papers where sparse learning was used in models problems such as optimal control, dynamical systems, signal and image processing. The high volume of research in recent years has caused great advancements in the field. Accordingly, many extremely efficient algorithms have been developed for sparse machine learning. Despite sparse learning problems having higher computational costs than their non-sparse counterparts, it might be shortly be agreed on that their sparsity constraints or penalties may help in reducing the computational costs of learning.

In this work, we are interested in one of the sparse optimization problems, namely the sparse linear regression model (i.e. least square regression), which has the following form:

$$\min_{\beta} \frac{1}{2} \|X\beta - y\|_2^2 + \sum_i \rho(\beta_i; \lambda_i),$$

where $\rho(\cdot\,;\cdot)$ is a penalty function that produces sparsity of an optimal solution $\hat{\beta}$, and $\{\lambda_i\}$ are some other penalty parameters used to modify the shape of each of these functions. Setting different values for λ's gives rise to different problems in optimization. In this work we focus on LASSO regression (Dong et al., 2015) and best subset selection problem (Miller, 2002).

Bertsimas et al. (2016) noticed that algorithmic advances in integer optimization combined with hardware improvements resulted in a MIQP problem being solved at a much faster rate. As a result, these authors were able to develop a MIQP solver that could find optimal solutions which gave more accurate solutions than LASSO solvers. Since some sparse optimization problems can be formulated as Mixed Integer Quadratic Programs (MIQP), then we can solve these problems using the efficient methods of mixed integer optimization. Using similar analysis, we can reformulate LASSO regression to a MIQP, and use a MIQP solver to get an optimal solution for this regression problem. One the other side, one can analyse relaxations of MIQP. One of these relaxations, so-called perspective relaxations, can be interpreted as LASSO problems with different penalty functions.

In this essay, we first present a general description of the sparse optimization problem, and specifically expand on least square regressions. Also, in Chapter 2, some background on Mixed Integer Quadratic Optimization Problem (MIQP) will be presented. In Chapter 3, we explore MIQP in more detail with Best subset regression. Then we investigate the connection between LASSO and MIQP in Chapter 4. Finally, we conclude in Chapter 5.

Notation

Notation	definition
y	vector in $\mathbb{R}^n$
X	Matrix in $\mathbb{R}^{n\times p}$
β	vector $\mathbb{R}^p$
$\|\beta\|_1$	$\sum_{i=1}^{p} \|\beta_i\|$
$\|\beta\|_2$	$\sqrt{\sum_{i=1}^{p} \beta_i^2}$
$\|\beta\|_\infty$	$\max_{i=1}^{p} \|\beta_i\|$
$\|\beta\|_l$	$\sum_{i=1}^{p} (\|\beta_i\|^l)^{\frac{1}{l}}$
$\mathbf{0}_{n\times n}$	n-by-n Zero Matrix
$\mathbb{R}_+$	The Set of Non-Negative Real Numbers
$\varnothing$	empty set
$\langle v, w\rangle$	The Standard Euclidean Inner Product Of v and w
conv(X)	The Closure of the Convex Hull of X
$g(\beta)$	$\frac{1}{2}\|y - X\beta\|_2^2$
$\nabla g(\beta)$	$-X^T(y - X\beta)$
$\mathtt{diag}(\delta)$	p-by-p Diagonal Matrix where the main diagonal vector is equal to $\delta \in \mathbb{R}^p$

2. Least Squares Regression and Mixed-Integer Optimization Formulations

In this chapter, we will introduce the least squares regression and the mixed-integer problem in order to draw the connection between the mixed-integer quadratic problems and the LASSO least squares regression later on.

2.1 Least Squares Regression

In this section, we focus on the sparse linear regression model, more specifically on the least squares regression. The problems are known to induce sparsity to the solutions.

2.1.1 Problem formulation. Let p be the number of predictor variables and n the number of observations. Given a model matrix $X \in \mathbb{R}^{n\times p}$ of n input variables in $\mathbb{R}^p$, that is $X = [x_{ij}]$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$, and a vector response variable $y \in \mathbb{R}^n$, we therefore have a prediction problem of n cases. The least squares regression formula is defined as follows to get the minimum constraint $\beta = \{\beta_j\}_{j=1,\cdots,p}$ on the regression (Atsmtürk and Gómea, 2019; Tibshirani, 1995)

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \quad & \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 + \mu \|\beta\|_1 \\ \text{s.t} \quad & \|\beta\|_0 \leq k, \end{aligned} \tag{2.1}$$

where the non-negative regularization parameters λ and μ are meant to control the amount of penalization in the model complexity, and the desired sparsity (or tuning parameter) $k \in \mathbb{Z}^+$ is used to control the number of regression coefficients required to explain the observations from the explanatory variables in X with the L_0-norm constraint on the regression β (fus; Tibshirani, 1995; Atsmtürk and Gómea, 2019).

There exists a wide range of regression models under the least squares regression problem. They differ on the values of the parameters (λ, μ). Below are some special cases (Atsmtürk and Gómea, 2019):

- **Ridge regression** (E. Hoerl and W. Kennard, 1970) for $\lambda > 0,\ \mu = 0,\ k \geq p$. Ridge regression, in statistical literature, is known as weight decay in machine learning, but it is also called Tikhonov regularization by Andrey Tikhonov who modified Equation (2.1) using the shrinkage term which consists of the penalty λ and the L_2-norm. Then, for the ridge regression problem, Equation (2.1) is the same as

  $$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \quad & \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2, \\ \text{s.t} \quad & \|\beta\|_0 \leq k. \end{aligned}$$

  More details are given in Section 2.1.2.

- **LASSO regression** (Tibshirani, 1995) for $\lambda = 0,\ \mu > 0$, and $k \geq p$.

  Least Absolute Shrinkage and Selection Operator or LASSO, was proposed by Leo Breiman. This problem is very similar to the ridge regression, except that it is a regression with the L_1-norm

penalty, that is

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \|y - X\beta\|_2^2 + \mu \|\beta\|_1 , \\ \text{s.t} \quad & \|\beta\|_0 \leq k. \end{aligned}$$

Moreover, there are a lot of methods that generalize the LASSO regression problem. A few examples of them are described in the table below.

Method	Reference	Formulation
Grouped LASSO	(Yuan and Lin, 2007b)	$\sum_g \|\beta_g\|_2$
Elastic net	(Zou and Hastie, 2005)	$\lambda_1 \sum_i \|\beta_i\| + \lambda_2 \sum_i \beta_i^2$
Fused LASSO	(fus)	$\lambda \sum_i \|\beta_{i+1} - \beta_i\|$
Adaptive LASSO	(Zou, 2006)	$\lambda_1 \sum_i w_i \|\beta_i\|$
Dantzig selector	(Emmanuel and Terence, 2007)	$\min_\beta \left\|X^T(y - X\beta)\right\|_\infty , \quad \|\beta\|_1 < t$
Near isotonic regularization	(Ryan J. et al., 2011)	$\sum_i (\beta_i - \beta_{i+1})_+$
Matrix completion	(Candès and Terence; Rahul et al., 2010)	$\left\|X - \hat{X}\right\|^2 + \lambda \left\|\hat{X}\right\|_*$
Compressive sensing	(David, 2004; Emmanuel J, 2006)	$\min_\beta(\| \beta \|_1)$ subject to $y = X\beta$
Multivariate methods	(Jolliffe et al., 2003; Witten et al.)	Sparse principal component analysis, linear discriminant analysis and canonical correlation analysis

For more details on the LASSO regression, you may refer to Section 2.1.2.

- **Elastic net** (Zou and Hastie, 2005) for $\lambda,\ \mu > 0$, and $k \geq p$.

  It combines the penalties of the ridge and LASSO regressions, that means it uses the two L_1-norm and L_2-norm penalties. It is also a generalization of LASSO, where in LASSO $p > n$ but for the Elastic net $p < n$. Therefore, the elastic net formulation is the same as in Equation (2.1) when $p < n$.

- **Best subset selection (normal least square or linear regression)** (Miller, 2002) for $\lambda = \mu = 0$, and $k < p$.

  This problem is similar of the ordinary linear regression since it does not contain any penalty term, that is

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \|y - X\beta\|_2^2 , \\ \text{s.t.} \quad & \|\beta\|_0 \leq k. \end{aligned} \tag{2.2}$$

- For solving the high-dimensional regression problem (Bertsimas and Parys, 2017), when $\lambda > 0,\ \mu = 0$ and $k < p$.
  Bertsimas and Parys (2017) presented a novel binary convex reformulation of the sparse regression problem where they presented this method to solve the high-dimensional sparse regression problem

for any sample size n and any number of regressors (or regression parameters) $p > n$. The computational cost in time was about 100000 seconds. On the other hand, the difficulty of a problem increases with the sample size n but that is now solved by Bertsimas and Parys (2017) using this method when $\lambda > 0,\ \mu = 0$ and $k < p$. In fact, their method solves the problem extremely fast (even faster than LASSO).

- For solving the problems with Signal to Noise Ratios (SNR) (Mazumder et al., 2017), when $\lambda = 0,\ \mu > 0$ and $k < p$.
  Mazumder et al. (2017) proposed a new method inspired from the best subset selection regression and the shrinkage methods (LASSO and Ridge regressions). The best subset selection regression works extremely well when the Signal to Noise Ratio (SNR) is high. However, its work performance degrades when the Signal to Noise Ratio (SNR) is low. In the latter case, the shrinkage methods beats the best subset selection regression which leads Mazumder et al. (2017) to adopt this new method by combining the best subset selection regression and the shrinkage methods for any case of SNR.

There are many applications for least squares regression of all kinds. For example in financial time series analysis, macroeconomics, biology and medical sciences. In the medical sciences, the least squares regression is particularly useful as it can be used for the cancer cells study, the prediction of the chemotherapeutic response in patients, in the genetic expression of tumours, and for drug sensitivity databases.

The following subsection will discuss the regularization techniques, that are used in special cases of regression.

2.1.2 Regularization techniques. The regularization techniques of least squares regression vary depending on the cases. When $k \geq p$, the fundamental constraint with the L_0-norm is excessive and the least squares regression formula (2.1) is easily solvable because it is a convex optimization problem. On the other hand, when $k < p$, we have a non-convex optimization problem then the problem is NP-hard and finding an optimal solution requires a great computational effort.

As earlier mentioned we have different types of regression. The **best subset selection regression** (Miller, 2002) is the regression formula in (2.1) without bias and shrinkage that is when $\lambda = \mu = 0$ and $k < p$ in the usual ordinary least squares such that

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \|y - X\beta\|_2^2, \\ \text{s.t} \quad & \|\beta\|_0 \leq k. \end{aligned} \tag{2.3}$$

A common modification in the ordinary least squares is to add the regularization terms with non-negative λ and μ in the objective function (2.3) to avoid overfitting by adding a penalty in the model. That is often included as a factor of $\|\beta\|_1$ or $\|\beta\|_2$ or as a linear combination of both norms which is the same as having the last two terms in the objective function of Equation (2.1). For example, we mentioned in the previous section for the **LASSO** and **ridge** regressions that when we add a penalty term to the $\|\beta\|_2^2$ part then we come up with a ridge regression which gives result to shrinkage and bias which can be desirable, but the result is sparse.

When we solve the ridge regression for the minimization problem, the vector of ridge regression estimates $\hat{\beta}_{ridge} \in \mathbb{R}^n$ is expressed as

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1}(X^T y), \quad \text{with } X \in \mathbb{R}^{n \times p},\ y \in \mathbb{R}^n,\ \lambda > 0,$$

where I denotes the identity matrix in $\mathbb{R}^{p\times p}$, and λ denotes the penalty term. One can see from this equation that when λ is equal to zero, $\hat{\beta}_{ridge}$ is the same as β_{OLR} from the linear regression in the case of Ordinary Linear Regression. On the other hand, as λ becomes larger (i.e. when the penalty is high enough), $\hat{\beta}_{ridge}$ approaches zero (i.e the coefficient of λ shrinks towards zero), and that is the goal for the ridge regression since we would like to find a minimal constraint β as the optimal solution (E. Hoerl and W. Kennard, 1970). Furthermore, the bias and the variance of the ridge regression estimates are, respectively, given by (E. Hoerl and W. Kennard, 1970)

$$\begin{aligned}\text{Bias}(\hat{\beta}_{ridge}) &= -\lambda(X^TX+\lambda I)^{-1}\beta,\\ \text{Var}(\hat{\beta}_{ridge}) &= \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1},\end{aligned}$$

where X is a model matrix in $\mathbb{R}^{n\times p}$, σ^2 is the variance of X and β is the regression parameter in $\mathbb{R}^p$. When λ tends to infinity, then, for the ridge regression estimates, the bias increases as opposed to the variance which decreases.

As a second example, when we add a penalty term μ to $\|\beta\|_1$, then we have the Least Absolute Selection and Shrinkage Operator or **LASSO** regression. This method provides a good performance and result in shrinkage. However, we cannot control the shrinkage and sparsity, which leads to a poor performance in some cases (Zhang and Zhang, 2012; Miller, 2002). Moreover, LASSO has some disadvantages due to these restrictions (Zou and Hastie, 2005). We can cite a few of them.

- If $n < p$, LASSO selects n features before it saturates.
- LASSO tends to select the variables one by one instead of considering all of them at once, and that without looking at the pairwise correlations between them.

LASSO is similar to the ridge regression in terms of definition as well as in solving the regression problem. But the LASSO regression uses the absolute values of the coefficients of β where the ridge regression uses the squares of the coefficients of β. Also, one of the major differences between them resides in the regularization techniques. For the ridge regression, when the penalty increases, most of the parameters tend to zero without vanishing while, for LASSO, we have many parameters vanish. That is one advantage that the LASSO regression has over the ridge regression. Unlike its counterpart, LASSO automatically selects the relevant features and deselects others (Niz et al., 2016), whereas the ridge regression selects any feature.

To avoid those restrictions on LASSO, the **elastic net** $(\lambda, \mu > 0, k \geq p)$ (Zou and Hastie, 2005) combines penalties of the ridge and LASSO regressions.

In the next section, we will talk about mixed-integer optimization to find almost optimal solutions of Problem (2.1) by using mixed-integer optimization solvers, we will also focus more on the properties of the mixed-integer quadratic programs.

2.2 Mixed-Integer Quadratic Programming (MIQP) Formulation

In this section, we present the (MIQP) formulation for the ordinary regression (or the best subset selection regression) as well as an approach to solving the MIO problem, together with a reformulation of the mixed-integer quadratic program (MIQP) to the LASSO regression problem by using some properties in MIQP.

2.2.1 Brief Background on MIQP. In general, the mixed integer quadratic programming (MIQP) problems are optimization problems in which a quadratic objective function is minimized over a set of linear constraints with bounded variables, it can be written as

$$\begin{aligned} \min \ & \alpha^T Q\alpha + \alpha^T a, \\ \text{s.t.} \ & A\alpha \leq b, \\ & \alpha_i \in \{0,1\}, \ \forall i = 1, \cdots, n, \\ & \alpha_j \in \mathbb{R}_+, \forall j \notin \mathcal{I}; \ \mathcal{I} \subset \{1, \cdots, n\}, \end{aligned} \tag{2.4}$$

where $a \in \mathbb{R}^n$, $A \in \mathbb{R}^{k\times n}$, and $Q \in \mathbb{R}^{n\times n}$ is positive semi-definite and, where $\alpha \in \mathbb{R}^n$ contains both discrete $(\alpha_i, \ i \in \mathcal{I})$ and continuous $(\alpha_i, \ i \notin \mathcal{I})$ variables. Furthermore, if $\mathcal{I} = \varnothing$, the problem becomes a convex quadratic optimization problem; if $Q = \mathbf{0}_{n\times n}$, it is a mixed-integer problem; and for $(\mathcal{I} = \varnothing, \ Q = \mathbf{0}_{n\times n})$, we have a linear optimization problem. These three classes are all subclasses of Mixed-Integer optimization.

2.2.2 MIQP Formulations for the Best Subset Selection Problem. In this Subsection, we present a simple reformulation of Problem (2.2) as a MIQP problem (Bertsimas et al., 2016).

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \ & \frac{1}{2}\|y - X\beta\|_2^2 \\ \text{s.t} \ & -Mz_i \leq \beta_i \leq Mz_i, \quad i = 1, \cdots, p, \\ & \sum_i z_i \leq k, \\ & \beta \in \mathbb{R}^p, z = \{z_i\}_{i=1}^p \in \{0,1\}^p, \end{aligned} \tag{2.5}$$

where M is a constant, z is a binary variable where $z_i = 1$ means that $|\beta_i| \leq M$, otherwise, $\beta_i = 0$. Since $\hat{\beta}$ is the minimizer of this problem, which means that M should be large enough such that $M \geq \left\|\hat{\beta}\right\|_\infty$. In this case, $\hat{\beta}$ is an optimal solution of Problem (2.5) and (2.2) because they are equivalent.

When we use the convex hull of the constraints in Problem (2.5)

$$\begin{aligned} \text{Conv}\left(\left\{\beta : |\beta_i| \leq Mz_i, z_i \in 0,1, i = 1, \cdots, p, \sum_{i=1}^p z_i \leq k\right\}\right) \\ = \{\beta : \|\beta\|_\infty \leq M, \|\beta\|_1 \leq Mk\} \subseteq \{\beta : \|\beta\|_1 \leq Mk\}. \end{aligned}$$

We can make interesting conclusions, e.g. the solution to Problem (2.5) is lower-bounded by the optimum objective value of the convex optimization problem, that is

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2, \quad \text{s.t.} \ \|\beta\|_\infty \leq M, \|\beta\|_1 \leq Mk, \tag{2.6}$$

and

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2, \quad \text{s.t.} \ \|\beta\|_1 \leq Mk. \tag{2.7}$$

Problem (2.7) is similar to LASSO problem, where it is weaker than Problem (2.6). Additionally, the constraint of the LASSO problem in (2.7) is weaker than the root node relaxation, and MIO improves the quality of the root node solution and toward the optimal solution. Hence, we may use MIQP to solve the reformulation of LASSO.

2.2.3 Formulations By Specially Ordered Sets (SOS). Any optimal solution (feasible solution) of Problem (2.5) satisfies the condition $(1-z_i)\beta_i = 0,\ \forall i \in \{1,\cdots,p\}$ which is called integer optimization with the use of Specially Ordered Sets Type 1 (SOS-1). Note that SOS-1 means that at least one variable β_i is non-zero where $\beta = \{\beta_i\}$. Let us write the previous condition as follows

$$(1-z_i)\beta_i = 0 \Leftrightarrow (\beta_i, 1-z_i) : SOS-1; \forall i = 1,\cdots,p\ .$$

Then, we can use this condition in Equation (2.2)

$$\begin{aligned} \min_{\beta,z} \quad & \frac{1}{2}\left\|y - X\beta\right\|_2^2, \\ \text{s.t} \quad & (\beta_i, 1-z_i) : SOS-1, i = 1,\cdots,p, \\ & z_i \in \{0,1\}, i = 1,\cdots,p, \\ & \sum_{i=1}^{p} z_i \leq k. \end{aligned} \tag{2.8}$$

We see from Problem (2.8) that we can get a global solution without knowing the parameter M from Problem (2.2) to (2.7), and we can use the Problem (2.8) to obtain global solutions for these problems (2.2) to (2.7) without knowing the parameter M. In addition, we can see in the objective function of the Problem (2.8) that there is a convex quadratic function in the continuous variable β, which is a representation of Problem (2.8) formulated explicitly by the following

$$\begin{aligned} \min_{\beta,z} \quad & \frac{1}{2}\beta^T X^T X\beta - \langle X^{'}y, \beta\rangle + \frac{1}{2}\left\|y\right\|_2^2 \\ \text{s.t} \quad & (\beta_i, 1-z_i) : SOS-1, i = 1,\cdots,p \\ & z_i \in 0,1, i = 1,\cdots,p, \\ & \sum_{i=1}^{p} z_i \leq k, \\ & -M \leq \beta_i \leq M, i = 1,\cdots,p \\ & \left\|\beta\right\|_1 \leq M, \end{aligned} \tag{2.9}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean inner product. Furthermore, this problem can be reformulated as

$$\begin{aligned} \min_{\beta,z,\zeta} \quad & \frac{1}{2}\zeta^T\zeta - \langle X^{'}y, \beta\rangle + \frac{1}{2}\left\|y\right\|_2^2 \\ \text{s.t} \quad & \zeta = X\beta \\ & (\beta_i, 1-z_i) : SOS-1, i = 1,\cdots,p \\ & z_i \in 0,1, i = 1,\cdots,p, \\ & \sum_{i=1}^{p} z_i \leq k, \\ & -M \leq \beta_i \leq M, i = 1,\cdots,p \\ & \left\|\beta\right\|_1 \leq M, \\ & -M \leq \zeta_i \leq M, i = 1,\cdots,n\ , \\ & \left\|\zeta\right\|_1 \leq M, \end{aligned} \tag{2.10}$$

where the variables (optimization variables) are $\beta \in \mathbb{R}^p, \zeta \in \mathbb{R}^n, z \in \{0,1\}^p$ and $M \in [0,\infty]$. The Problem (2.10) is a quadratic problem with n variables (i.e $y = \{y_j\}_{j=1}^n$) with n a linear functions X

and p variables ($X = \{x_i\}_{j,i=1}^{n,p}$).
Problem (2.10) is equivalent to Problem (2.2) which is the best subset problem

$$\begin{aligned} \min_{\beta} \quad & \|y - X\beta\|_2^2 \\ \text{s.t} \quad & \|\beta\|_0 \leq k, \\ & \|\beta\|_\infty \leq M, \|\beta\|_1 \leq M, \\ & \|X\beta\|_\infty \leq M, \|X\beta\|_1 \leq M. \end{aligned} \tag{2.11}$$

Problems (2.10) and (2.11) differ in the size of the quadratic forms, where Problem (2.11) has more variables and is then useful for high-dimensional problems ($p > n$). Furthermore, the constraints on ζ and β enhance the strength of MIQP.

2.2.4 Mixed-Integer Quadratic Optimization Reformulation to LASSO regression . Several techniques exist to solve the problem (2.1). Bixby (2012) applied the mixed-integer optimization technique to come up with a new method and uses its benefit to get a solution to the regression problem in (2.1). In fact, he observed MIO as a more computationally efficient method.

Bixby (2012) introduced the indicator variable $z = \{z_i\}_{i=1}^p \in \{0,1\}^p$ which is a binary variable, for any i, where, if $z_i = 1$ then $\beta_i \in [-M, M]$ for some constant M produced by the big-M method (Bertsimas et al., 2016), otherwise $\beta_i = 0$. Moreover, the big-M constraints are used to capture the non-convexity and to confirm the equivalence between Problem (2.1) and MIO (Bertsimas and Parys, 2017). Then, we can present a simple reformulation of Problem (2.1) as the mixed-integral quadratic program (MIQP) similar to (2.5)

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 + \mu \|\beta\|_1 \\ \text{s.t} \quad & -Mz_i \leq \beta_i \leq Mz_i, \quad i = 1, \cdots, p, \\ & \sum_i z_i \leq k, \\ & \beta \in \mathbb{R}^p, z = \{z_i\}_{i=1}^p \in \{0,1\}^p. \end{aligned} \tag{2.12}$$

Furthermore, we can replace the big-M constraints $-Mz_i \leq \beta_i \leq Mz_i$ in (2.12) by the conditions $\beta_i(1 - z_i) = 0$ to solve (2.1) with no need to specify the value of the parameter M. Remark that in the regression problem, we cannot find any constant M that will bound the variables β_i and that does not lead to finding the β_i's solutions (Atsmtürk and Gómea, 2019). We can then reformulate the problems (2.1) and (2.12) as (Atsmtürk and Gómea, 2019)

$$\begin{aligned} \min \quad & y^T y - 2y^T X\beta + \beta^T (X^T X + \lambda I)\beta + \mu \sum_{i=1}^p u_i \\ \text{s.t} \quad & \sum_{i=1}^p z_i \leq k, \\ & \beta_i \leq u_i, \quad -\beta_i \leq u_i, \quad i = 1, \cdots, p, \\ & \beta_i(1 - z_i) = 0, \quad i = 1, \cdots, p, \\ & \beta \in \mathbb{R}^p, \ z \in \{0,1\}^p, \ u \in \mathbb{R}_+^p. \end{aligned} \tag{2.13}$$

There exist two different algorithms for solving the Mixed-integer problem (2.12) which are the branch-and-bound algorithm and the cutting-plane algorithm. These are the core techniques to solve the

mixed-integer problem but we will only focus on the first one in this section, especially on how the branch-and-bound algorithm works. The Branch-and-bound algorithm works in the MIO problem, moreover, we find that it uses strong convex relaxations of any optimization problem to reduce the size of the solutions set and the number of sub-problems we deal with. Therefore, we can apply this algorithm to solve Equation (2.13) in a much faster way when the value of big-M is available because it finds the convex relaxations that approximate the non-convex problem; otherwise, without use of the idea in the branch-and-bound algorithm, it takes a large amount of time to compute a solution of (2.13).

We will now study the properties of the mixed-integral quadratic program (2.12) and will focus more on the LASSO regression.

2.2.5 Properties of the mixed-integer quadratic program reformulation of LASSO.

- When we choose M sufficiently large such that

$$M > \max_{z\in\{0,1\}^p} \|\beta_z^*\|_\infty , \tag{2.14}$$

in which case the LASSO regression becomes

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \quad & \frac{1}{2}\|y - X\beta\|_2^2 + \bar{\lambda}\sum_i z_i \\ \text{s.t} \quad & |\beta_i| \le M z_i \\ & z_i \in \{0,1\}, \quad i = 1, \cdots, p. \end{aligned} \tag{2.15}$$

Then, Equation (2.15) can be considered as a special continuous relaxation of MIQP that is equivalent to the LASSO (the classical L_1 approximation) regression problem with penalty parameter $\frac{\lambda}{M}$ where M is chosen to be sufficiently large as in (2.14) and the condition on $z_i \in \{0,1\}$ relaxes to $z_i \in [0, +\infty]$. Moreover, from Equation (2.15), we must have $z_i \le \frac{|\beta_i|}{M}$ to get the optimal solution of MIQP. Therefore, applying this continuous relaxation in Equation (2.15), we have the following equivalent LASSO (a convex approximation to L_0) problem

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \bar{\lambda}\sum_i |\beta_i|, \qquad \text{(LASSO)}.$$

Then, when we find the optimal solution to this linear regression which is given by

$$\beta_s^* \in \arg\min\Big\{\frac{1}{2}\|x\beta - y\|_2^2 ; \beta_i = 0, \forall i \notin s\Big\}.$$

This is an optimal solution for MIQP and also for LASSO with penalty parameter $\bar{\lambda} = \frac{\lambda}{M}$ in LASSO.

- We can use convex relaxation techniques for MIQP$_{\lambda,M}$. The parameters λ, M are independent on lifting and conic optimization for bringing new ideas to develop some methods for variable selection which getting optimal solution.

3. Discrete Optimization Method to Solve LASSO Problem and Applications

In this chapter, we will describe the first order discrete optimization methods to solve the mixed-integer quadratic Programs (MIQP) in an optimal way. After developing extensions of these methods in convex optimization, the approximate optimal solution to Problem (2.2) will be deduced.

3.1 Finding Stationary Solutions for Minimizing Smooth Convex Functions With Cardinality Constraints

3.1.1 Related work and contributions. Proposed iterative hard threshold algorithms in the literature on signal processing (Blumensath and E. Davies, 2008, 2009) were generalized to the context of L_0-norm regularization of the least squares problems. The convergence properties of the algorithms using the coherence of the model matrix X (Blumensath and E. Davies, 2008) or its restricted isometry property (Blumensath and E. Davies, 2009). The method we are about to show can be applied to a larger class of constrained optimization problems of the form (3.1). Moreover, the idea of the algorithm is derived from gradient descent methods in first-order convex optimization problems (Nesterov, 2004), which are generalized in the discrete optimization problem (3.1). We consider the optimization problem as follows

$$\min_{\beta} g(\beta) \quad \text{s.t} \ \ \|\beta\|_0 \leq k, \tag{3.1}$$

where g is a convex function and also non-negative which means that it has a Lipschitz continuous gradient, that is there exists $\ell > 0$ such that for any vector $\beta, \tilde{\beta}$, we have

$$\left\|\nabla g(\beta) - \nabla g(\tilde{\beta})\right\| \leq \ell \left\|\beta - \tilde{\beta}\right\|. \tag{3.2}$$

If $g(\beta) = \|\beta - c\|_2^2$ for some constant vector c, Problem (3.1) has a closed-form solution.

3.1.2 Proposition. Given positive integers p and k, and a vector $c = [c_1, c_2, \cdots, c_p]^T \in \mathbb{R}^p$.
If $\hat{\beta}$ is the optimal solution to the following problem

$$\hat{\beta} \in \underset{\|\beta\|_0 \leq k}{\arg\min} \|\beta - c\|_2^2, \tag{3.3}$$

then it can be computed as follows.
First, $\hat{\beta}$ retains the k largest (in absolute value) elements of $c \in \mathbb{R}^p$ and sets the rest to zero, i.e. if $|c_{(1)}| \geq |c_{(2)}| \geq \cdots \geq |c_{(p)}|$, denotes the ordered values of the absolute values of the vector c, then

$$\hat{\beta}_i = \begin{cases} c_{(i)}, & \text{if } (i) \in \{(1), \cdots, (k)\}, \\ 0, & \text{otherwise}, \end{cases} \tag{3.4}$$

where $\hat{\beta}_i$ is the i-th component of $\hat{\beta}$. We will denote the set of solutions to Problem (3.3) by the notation $\mathbf{H}_k(c)$.

Proof. See (Bertsimas et al., 2016). □

The operator in Equation (3.4) is the hard-thresholding operator and it is also related to the optimization problem

$$\hat{\beta} \in \arg\min_{\beta} \frac{1}{2} \|\beta - c\|_2^2 + \lambda \|\beta\|_0 , \tag{3.5}$$

where λ is the penalty for the shrinkage term, $\hat{\beta}_i = c_{(i)}$ if $|c_{(i)}| > \sqrt{\lambda}$ and $\hat{\beta}_i = 0$ otherwise, for any $i = 1, \cdots, p$. Moreover, the minimum non-zero term of Problem (3.5) is greater than λ where in Problem (3.3) there is no minimum non-zero term.

3.1.3 Proposition. (Nesterov, 2004, 2007) Given a model matrix X. For a convex function g satisfying Condition (3.2) with $\ell = \lambda_{\max}(X^T X)$, for any $L \geq \ell$, we have

$$g(\eta) \leq Q_L(\eta, \beta) := g(\beta) + \frac{L}{2} \|\eta - \beta\|_2^2 + \langle \nabla g(\beta), \eta - \beta \rangle, \tag{3.6}$$

for any vectors β and η, with equality holding at $\beta = \eta$.

Applying Proposition 3.1.2 to Proposition 3.1.3 in the upper bound $Q_L(\eta, \beta)$, we have

$$\begin{aligned} \arg\min_{\|\eta\|_0 \leq k} Q_L(\eta, \beta) &= \arg\min_{\|\eta\|_0 \leq k} \left(\frac{L}{2} \left\| \eta - \left(\beta - \frac{1}{L} \nabla g(\beta) \right) \right\|_2^2 - \frac{1}{2L} \|\nabla g(\beta)\|_2^2 + g(\beta) \right), \\ &= \arg\min_{\|\eta\|_0 \leq k} \left\| \eta - \left(\beta - \frac{1}{L} \nabla g(\beta) \right) \right\|_2^2, \\ &= \mathbf{H}_k \left(\beta - \frac{1}{L} \nabla g(\beta) \right), \end{aligned} \tag{3.7}$$

to obtain the optimal solution of Equation (3.6).

Hence, we can use the algorithm below to find the stationary point of Problem (3.1).

Algorithm 1:

Input: $g, \beta_1, p, m, k, L, \epsilon$.

Output: A first order stationary solution β^*.

Algorithm:

1 Initialize with $\beta_1 \in \mathbb{R}^p$ such that $\|\beta_1\|_0 \leq k$.

2 For $m \geq 1$, apply (3.7) with $\beta = \beta_m$ to obtain β_{m+1} as:

$$\beta_{m+1} \in \mathbf{H}_k \left(\beta_m - \frac{1}{L} \nabla g(\beta_m) \right) \tag{3.8}$$

3 Repeat Step 2, until $\|\beta_{m+1} - \beta_m\|_2 \leq \epsilon$.

4 Let $\beta = \beta_m := (\beta_{m1}, \cdots, \beta_{mp})$ denote the current estimate and let $I = \text{Supp}(\beta_m) := \{i : \beta_{mi} \neq 0\}$. Solve the continuous optimization problem

$$\min_{\beta,\ \beta_i = 0,\ i \notin I} g(\beta), \tag{3.9}$$

and let β^* be the minimizer of Equation (3.9).

- Where $\ell = \lambda_{\max}(X^T X)$.

3.2 Application to Least Squares

For the constrained problem with squared error loss, for β in the support of the least squares problem, we have

$$g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 \text{ and } \nabla g(\beta) = -X^T(y - X\beta).$$

The regression coefficients in the active set can be obtained by using the least squares problem for the variables y, X_I where I is the support of the regression coefficients.

3.3 Application to Least Absolute Deviation

In this section, we apply the method in Section 3.2 to the least absolute deviation problem with the support constraints in β, that is

$$\min_{\beta} g_1(\beta) := \|Y - X\beta\|_1 \quad \text{s.t.} \quad \|\beta\|_0 \leq k. \tag{3.10}$$

As g_1 is a non-smooth function thus it cannot be applied in Algorithm 1 directly. Therefore, we have to make the function g_1 smooth by observing that

$$g_1(\beta) = \sup_{\|W\|_\infty \leq 1} \langle Y - X\beta, W\rangle.$$

Here, we used the smoothing technique of Nesterov (2005) to get the smooth approximation of g_1 as follows

$$g_1(\beta : \tau) = \sup_{\|W\|_\infty \leq 1} (\langle Y - X\beta, W\rangle - \frac{\tau}{2}\|W\|_2^2), \text{ with } \ell = \frac{\lambda_{max}(X^TX)}{\tau}.$$

Therefore, we can use $g_1(\beta : \tau)$ in Algorithm 1 with this value of ℓ. Moreover, to get a premium approximation to Problem (3.10), we can use the following strategies in practice:

- Take $\tau > 0$, initialize with $\beta_0 \in \mathbb{R}^p$ and repeat the steps (2 - 3) until convergence.
- Apply the smooth function $g_1(\beta; \tau)$ in Algorithm 1 to obtain the limiting solution β_τ^*.
- Choose γ such that $\tau\gamma \to \tau$ (e.g: $\gamma = 0.8$), and return to step(1) with $\beta_0 = \beta_\tau^*$.

  Exit from the algorithm if for some predefined tolerance TOL , the value of τ is less than TOL.

4. Connection between Mixed-Integer Quadratic Programming and LASSO

In this chapter, we would like to show the connection between Mixed-Integer Quadratic Programming (MIQP) and the LASSO regression. As a result, for some penalty functions, we will see that MIQP becomes similar to the LASSO problem. Therefore, there is a connection between LASSO and the conic relaxation MIQP.

Initially, we will give a reformulation of MIQP by using a convex relaxation of MIQP which is called perspective relaxation and is also the second order cone programming problem (SOCP) (Dong et al., 2015). The SOCP problem has a linear objective function and a second order cone constraint for a given n dimension. We will then use two penalty functions, namely the minimax concave penalty (MCP) (Zhang, 2010) and the reverse Huber penalty (Pilanci et al., 2015), in the perspective relaxation in order to get the connection between MIQP and LASSO. As a remark, Equation (2.1) is also a mixed-integer convex quadratic optimization problem (MICQP) with indicator variables.

We apply the perspective relaxation to a separable quadratic function of the form $\beta^T \texttt{diag}(\delta)\beta$, where, for $\delta \in \mathbb{R}^p_+$, $\texttt{diag}(\delta)$ is a non-negative diagonal matrix such that $X^TX - \texttt{diag}(\delta) \geq 0$. Then, since $X^TX = (X^TX - \texttt{diag}(\delta)) + \texttt{diag}(\delta)$, MIQP problem in (2.15) can be written as (Dong et al., 2015)

$$\begin{aligned} \min_{\beta,z} \quad & \frac{1}{2}\beta^T(X^TX - \texttt{diag}(\delta))\beta - (X^Ty)^T\beta + \frac{1}{2}\sum_i \delta_i\beta_i^2 + \lambda\sum_i z_i + \frac{1}{2}y^Ty \\ \text{s.t} \quad & -Mz_i \leq \beta_i \leq Mz_i\ , \quad i = 1, \cdots, p, \\ & z_i \in \{0,1\}, \quad i = 1, \cdots, p. \end{aligned} \tag{4.1}$$

4.0.1 Remark. Given a model matrix $X \in \mathbb{R}^{n\times p}$, we would like to study the condition $X^TX - \texttt{diag}(\delta) \geq 0$ depending on $\delta \in \mathbb{R}^p_+$. That is, if X^TX is a positive definite matrix, then $\delta \neq \mathbf{0}$. Otherwise, if X^TX has a zero eigenvalue, then $\delta = \mathbf{0}$ because only the zero-vector satisfies the above condition and we are only left with X^TX. In the last case, the perspective relaxation is not helpful in this problem. Therefore, we assume in this chapter that X^TX is positive definite.

We now apply the perspective relaxation to the separable quadratic regularization term $\lambda\|\beta\|_2^2$ in the problem (2.13). Then, we want to reformulate $\lambda\|\beta\|_2^2$ in Equation (2.13) and get the perspective relaxation equation denoted by $\zeta_{PR(\delta)}$. Furthermore, additional variables $s_i \geq 0$ can be used to reformulate $\|\beta\|_2^2$ in Equation (4.1) by using the valid perspective constraint in a convex hull for the mixed-integer quadratic function. By applying this idea, the new condition for Equation (4.1) would be $s_iz_i \geq \beta_i^2$, when $M \to +\infty$ (see Section 2.2.5, simple case). Then we get the perspective relaxation formula which is expressed as

$$\begin{aligned} \zeta_{PR(\delta)} := \min_{\beta,z} \quad & \frac{1}{2}\beta^T(X^TX - \texttt{diag}(\delta))\beta - (X^Ty)^T\beta + \frac{1}{2}\sum_i \delta_i s_i + \lambda\sum_i z_i + \frac{1}{2}y^Ty \\ \text{s.t} \quad & s_iz_i \geq \beta_i^2, s_i \geq 0, 0 \leq z_i \leq 1, \quad i = 1, \cdots, p. \end{aligned} \tag{4.2}$$

4.0.2 Remark. The perspective constraints $s_iz_i \geq \beta_i^2$ are satisfied for $\beta_i \neq 0$ only when $z_i = 1$, otherwise $\beta_i = 0$.

In the next proposition, we will show that the minimum of $\zeta_{PR(\delta)}$ always exists.

4.0.3 Proposition. (Zou and Hastie, 2005) Given a model matrix $X \in \mathbb{R}^{n\times p}$ and a penalty parameter $\lambda \geq 0$. Suppose X^TX is a positive definite matrix. If there exists $\delta \in \mathbb{R}^p_+$ such that $X^TX - \texttt{diag}(\delta) \geq 0$, then the optimal value of the perspective relaxation in Equation (4.2) is reached at some finite point.

Proof. Consider the objective function of Equation (4.2) which is given by

$$\frac{1}{2}\|X\beta - y\|_2^2 + \sum_i \left(\frac{1}{2}\delta_i(s_i - \beta_i^2) + \lambda z_i\right) \geq \frac{1}{2}\|X\beta - y\|_2^2,$$

for a given $y \in \mathbb{R}^p$ and where $\beta \in \mathbb{R}^p$, $s := \{s_i\}_i \in \mathbb{R}^p_+$, and $z := \{z_i\}_i \in \{0,1\}^p$.

The optimal solution $\hat{\beta}$ of the right-hand side equation is unique and is given by $\hat{\beta} = \arg\min_\beta \|X\beta - y\|_2^2$. Also, by choosing $\hat{s}_i = \hat{\beta}^2$ and $\hat{z}_i = 1$, $(\hat{\beta}, \hat{s}, \hat{z})$ is the solution to Equation (4.2). Moreover, using the convexity of $\|X\beta - y\|_2^2$, there exists $R > 0$ such that for any β, we have $\|\beta\|_2 \geq R$. Therefore,

$$\frac{1}{2}\|X\beta - y\|_2^2 \geq \frac{1}{2}\left\|X\hat{\beta} - y\right\|_2^2 + \lambda p = \frac{1}{2}\left\|X\hat{\beta} - y\right\|_2^2 + \sum_i \frac{1}{2}\delta_i(\hat{s}_i - \hat{\beta}_i^2) + \lambda\hat{z}_i.$$

□

In the next part, we will derive the penalization from the perspective relaxation equation (4.2) where for some values of the penalty function, for example the minimax concave penalty (MCP) (Zhang, 2010) or the reverse Huber penalty (Pilanci et al., 2015), we have an equivalent penalty to what we have in the LASSO problem. Furthermore, the next Theorem 4.0.4 shows that the perspective relaxation equation (4.2) is equivalent to the regularized regression problem which is exactly the LASSO regression problem, and that is the goal of this project to see the connection between MIQP and LASSO.

4.0.4 Theorem. *Assume $X^TX > 0, \lambda > 0$, and let $\delta \in \mathbb{R}^p_+$ and $X^TX - \texttt{diag}(\delta) \geq 0$. (PR_{δ_i})is equivalent to the following regularized regression problem*

$$\min_\beta \frac{1}{2}\|X\beta - y\|_2^2 + \sum_i \rho_{\delta_i}(\beta_i;\lambda) \qquad (PR_\delta : \ reg)$$

where

$$\rho_{\delta_i}(\beta_i;\lambda) = \begin{cases} \sqrt{2\delta_i\lambda}|\beta_i| - \frac{1}{2}\delta_i\beta_i^2 & \text{if } \delta_i\beta_i^2 \leq 2\lambda; \\ \lambda & \text{if } \delta_i\beta_i^2 > 2\lambda. \end{cases} \tag{4.3}$$

Proof. We can write Equation (4.2) (PR_δ) as follows

$$\frac{1}{2}\|X\beta - y\|_2^2 + \sum_i \frac{1}{2}\delta_i(s_i - \beta_i^2) + \lambda z_i.$$

Then can be reformulation PR_δ to regularized regression problem

$$\min_\beta \frac{1}{2}\|X\beta - y\|_2^2 + \sum_i \rho_{\delta_i}(\beta_i;\lambda) \qquad (PR_\delta : \ \text{reg}),$$

where

$$\begin{aligned} \rho_{\delta_i}(\beta_i;\lambda) = \min \quad & \frac{1}{2}\delta_i(s_i - \beta_i^2) + \lambda z_i; \\ \text{s.t} \quad & s_i z_i \geq \beta_i^2, s_i \geq 0, z_i \in [0,1]. \end{aligned} \tag{4.4}$$

When we study the optimal solution for Equation (4.4), we focus on the values of the parameters (δ_i, β_i) where, if $\delta_i = 0$ we can see that the value of $\rho_{\delta_i}(\beta_i; \lambda)$ is zero. That means we focus on the optimal solution of $\rho_{\delta_i}(\beta_i; \lambda)$ when $\delta_i > 0$, then from the other parameter β_i we have two cases. When $\beta_i = 0$, we can see that the optimal solution satisfies the condition $\rho_{\delta_i}(\beta_i; \lambda) = 0$, also $s_i = 0$ or $z_i = 0$ since $s_i z_i \geq \beta_i^2$. Otherwise, when $\beta_i \neq 0$, Equation (4.4) becomes a one-dimensional problem in s_i

$$\rho_{\delta_i}(\beta_i; \lambda) = \min_{s_i \geq \beta_i^2} \frac{1}{2}\delta_i(s_i - \beta_i^2) + \lambda \frac{\beta_i^2}{s_i}, \tag{4.5}$$

where we have the constraint $s_i z_i \geq \beta_i^2$ with $z_i \in \{0, 1\}$. Then we can take $s_i \geq \beta_i^2$, which means that $z_i = \frac{\beta_i^2}{s_i}$, to get the optimal solution.

Observe that the objective function in Equation (4.5) is a convex function in s_i when $s_i \geq \beta_i^2$ and a non-convex function of β_i but nevertheless $(PR_\delta : \ reg)$ is a convex problem when $XX^T - \texttt{diag}(\delta)$ is a positive semi-definite matrix.

4.0.5 Remark. A function f is convex if it satisfies the following condition for any x_1, x_2

$$f(\theta x_1 + (1-\theta)(x_2)) \leq \theta f(x_1) + (1-\theta) f(x_2), \quad \theta \in [0, 1].$$

The minimum solution of the convex function is calculated by putting the gradient of this function equal to zero. Then we find the value of the variable x that satisfies that condition, i.e $\nabla f(x) = 0$. Therefore, the minimum solution of Equation (4.5) can be obtained such that

$$\frac{\partial \rho_{\delta_i}}{\partial s_i}\left(\frac{1}{2}\delta_i(s_i - \beta_i^2) + \lambda \frac{\beta_i^2}{s_i}\right) = 0,$$

that is

$$\frac{1}{2}\delta_i - \frac{\lambda \beta_i^2}{s_i^2} = 0 \Rightarrow \ s_i^2 = \frac{2\lambda\beta_i^2}{\delta_i},$$

thus the minimum solution is obtained when

$$s_i = \sqrt{\frac{2\lambda\beta_i^2}{\delta_i}}.$$

This is the case when $\sqrt{\frac{2\lambda\beta_i^2}{\delta_i}} \geq \beta_i^2$, and also $s_i = \beta_i^2$ when $\sqrt{\frac{2\lambda\beta_i^2}{\delta_i}} \leq \beta_i^2$. Therefore, when we apply this result in Equation (4.5) to get the value of $\rho_{\delta_i}(\beta_i; \lambda)$ as follows

$$\rho_{\delta_i}(\beta_i; \lambda) = \begin{cases} \sqrt{2\delta_i\lambda}|\beta_i| - \frac{1}{2}\delta_i\beta_i^2, & \text{if } \ s_i = \sqrt{\frac{2\lambda\beta_i^2}{\delta_i}}; \\ \lambda, & \text{if } \ s_i = \beta_i^2. \end{cases}$$

Remark that the two cases $\sqrt{\frac{2\lambda\beta_i^2}{\delta_i}} \geq \beta_i^2$ and $\sqrt{\frac{2\lambda\beta_i^2}{\delta_i}} \leq \beta_i^2$ can be rewritten, respectively, as $\delta_i\beta_i^2 \leq 2\lambda$ and $\delta_i\beta_i^2 > 2\lambda$. Therefore, $\rho_{\delta_i}(\beta_i; \lambda)$ can be expressed as

$$\rho_{\delta_i}(\beta_i; \lambda) = \begin{cases} \sqrt{2\delta_i\lambda}|\beta_i| - \frac{1}{2}\delta_i\beta_i^2, & \text{if } \ \delta_i\beta_i^2 \leq 2\lambda; \\ \lambda, & \text{if } \ \delta_i\beta_i^2 > 2\lambda. \end{cases}$$

Note that this formula includes the case when $\delta_i = 0$ or $\beta_i = 0$. □

For the first case of the penalty function given in (4.3), we can get

$$\begin{aligned}\rho_{\delta_i} &= \sqrt{2\delta_i\lambda}|\beta_i| - \frac{1}{2}\delta_i\beta_i^2, \\ &= \lambda - \left(\sqrt{\frac{\delta_i}{2}}|\beta_i| - \sqrt{\lambda}\right)^2 \leq \lambda.\end{aligned}$$

Therefore, that means ρ_{δ_i} has the lowest estimate λ. Also, when we look at this penalty, we can see that it is equivalent to the penalty of the LASSO regression. Then, (PR_δ) is extracted from a convex relaxation of the least squares regression formulation.

Note also that the penalty in Equation (4.3) can be rewritten as the Minimax Concave Penalty (MCP) (Zhang, 2010) with a slight change of variables expressed in the table below.

PR_δ **notation**	**MCP notation** (Zhang, 2010)
δ_i	$\frac{1}{\tilde{\gamma}}$
λ	$\frac{1}{2}\tilde{\gamma}\tilde{\lambda}^2$
$\sqrt{2\delta_i\lambda}$	$\tilde{\lambda}$

The $(PR : reg)$ is more general than the Minimax Concave Penalty (MCP) (Zhang, 2010) where MCP is to control the convexity of the penalty and only uses the parameter $\tilde{\lambda}$; where in $(PR : reg)$, δ_i is a vector of variables. Remark that if the components of δ_i are identical, then $(PR : reg)$ and MCP are equivalent. Furthermore, applying the change of variables in the previous table, MCP has the convexity condition $X^TX - (1/\tilde{\gamma})I \geq 0$ similar to the convexity condition $X^TX - \texttt{diag}(\delta) \geq 0$ in (PR_δ).

From now on, we will focus on the second example of perspective relaxation which is the reverse Huber penalty. This was proposed by Pilanci et al. (2015), where the following convex relaxation is given by

$$\min_\beta \frac{1}{2}\|X\beta - y\|_2^2 + 2\lambda B\left(\sqrt{\frac{\mu}{2\lambda}}\beta_i\right), \tag{4.6}$$

from the L_2L_0 penalized problem with $\mu > 0$, that is

$$\min_\beta \frac{1}{2}\|X\beta - y\|_2^2 + \frac{1}{2}\mu\|\beta\|_2^2 + \lambda\|\beta\|_0, \qquad (L_2L_0)$$

where B is the Huber penalty (Pilanci et al., 2015) and

$$B(t) = \begin{cases} |t|, & \text{if } |t| \leq 1; \\ \frac{t^2+1}{2}, & \text{otherwise.} \end{cases}$$

Now, we get the perspective relaxation of the (L_2L_0) problem which is equivalent to Equation (4.6). We can also reformulate the (L_2L_0) problem to the L_0 problem by using the new model matrix below

$$\min_\beta \frac{1}{2}\left\|\tilde{X}\beta - \tilde{y}\right\|_2^2 + \lambda\|\beta\|_0; \text{ where } \tilde{X} = \begin{bmatrix} X \\ \mu I_p \end{bmatrix}, \text{ and } \tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}. \tag{4.7}$$

We can easily see from Equation (4.7) that $\tilde{X}^T\tilde{X} = X^TX + \mu I_p > 0$, then to derive a perspective relaxation we can set $\mu = \delta_i$ for any i. Moreover, by using Theorem 4.0.4, the perspective relaxation is given as the following by

$$\begin{aligned} &\min_{\beta} \frac{1}{2}\left\|\tilde{X}\beta - \tilde{y}\right\|_2^2 + \sum_i \rho_\mu(\beta_i;\lambda) \\ \Leftrightarrow &\min_{\beta} \frac{1}{2}\left\|X\beta - y\right\|_2^2 + \sum_i \left\{\rho_\mu(\beta_i;\lambda) + \frac{1}{2}\mu\beta_i^2\right\}. \end{aligned} \tag{4.8}$$

Obviously, Equation (4.8) is similar to Equation (4.6), where we can say that, if $\mu\beta_i^2 \leq 2\lambda$, we get $|\sqrt{\frac{\mu}{2\lambda}}\beta_i| \leq 1$ and, therefore, the penalty term which is expressed as the summation over i in Equation (4.8) is reformulated by two cases which are

$$\rho_\mu(\beta_i;\lambda) + \frac{1}{2}\mu\beta_i^2 = \sqrt{2\mu\lambda}|\beta_i| = 2\lambda\left|\sqrt{\frac{\mu}{2\lambda}}\beta_i\right| = 2\lambda B\left(\sqrt{\frac{\mu}{2\lambda}}\beta_i\right), \quad \text{if } \mu\beta_i^2 \leq 2\lambda,$$

otherwise,

$$\rho_\mu(\beta_i;\lambda) + \frac{1}{2}\mu\beta_i^2 = \lambda + \frac{1}{2}\mu\beta_i^2 = 2\lambda\frac{\left(\sqrt{\frac{\mu}{2\lambda}}\beta_i\right)^2 + 1}{2} = 2\lambda B\left(\sqrt{\frac{\mu}{2\lambda}}\beta_i\right).$$

5. Conclusion

The aim of this essay was to connect two problems in the field of optimization which are Mixed-integer Quadratic Program (MIQP) and LASSO regression. This connection could be viewed in two ways: either we apply MIQP to LASSO or we apply LASSO in MIQP.

After presenting a general description of the sparse optimization problem, we gave a background on the Mixed-Integer Quadratic Program (MIQP). Then, we presented a MIQP formulation for the best subset selection regression and an approach to solve the Mixed-Integer Optimization problem. After that, we described first order discrete optimization methods to solve the mixed-integer quadratic programs (MIQP). We were then able to make a connection between LASSO and MIQP by applying MIQP in the LASSO regression to finally get an efficient method to solve LASSO. Further, we have shown that certain relaxations of MIQP can be interpreted as LASSO formulations with special penalization functions.

Future work could further investigate the application of LASSO method to mixed integers problem in order to also derive efficient methods for solving MIQP. When such work is done, then we could use the properties of MIQP in LASSO, or vice versa to get optimal solutions for these problems.

Acknowledgements

First of all, I would like to thank Allah for giving me the strength to finish this work. Many other people contributed to the successful completion of this project:

Firstly, I am heartily thankful to my supervisor, Prof. Dr. Ekaterina A. Kostina, for the encouragement, guidance and constant supervision as well as for providing the necessary information regarding the project, God bless her.

Secondly, I am indebted to many who have supported me, the teaching staff and my colleagues from AIMS, especially, Abigail and Banan. For their kindness and loyally helping me, I am really grateful. Also, I offer my regards and blessings to my tutor Dr. Dinna Ranirina, who supported me in all aspects during the completion of this research.

Thirdly, this project would never have been possible without Miss Noluvuyo Hobana's support and guidance. I am touched by her help and the moral support she has provided me with throughout my journey at AIMS.

Lastly, I would like to show my gratitude to my family and my friends.

References

Atsmtürk, A. and Gómea, A. Rank-One Convexification for Sparse Regression. *Industrial Engineering & Operations Research*, 4:1901–1033, 2019.

Bertsimas, D. and Parys, B. V. Sparse High-Dimensional Regression: Exact Scalable Algorithms and Phase Transitions. *arXiv preprint :1709.10029v1*, 2017. URL https://arxiv.org/pdf/1709.10029.pdf.

Bertsimas, D., King, A., and Mazumder, R. Best Subset Selection via a Modern Optimization. *Annals of Statistics*, 44.2:813–852, 2016.

Bixby, R. E. A Brief History of Linear and Mixed-Integer Programming Computation. *Documenta Mathematica*, Extra Volume ISMP:63–87, 2012.

Blumensath, T. and E. Davies, M. Iterative Thresholding for Sparse Approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

Blumensath, T. and E. Davies, M. Iterative Hard Thresholding for Compressed Sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Candès, E. J. and Terence, T. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *CoRR*, abs/0903.1476. URL http://arxiv.org/abs/0903.1476.

David, D. Compressed Sensing. *Statistics School of Humanities & Sciences*, 9:4305–4065, 2004.

Dong, H., Chen, K., and Linderoth, J. Regularization vs. Relaxation: A Conic Optimization Perspective of Statistical Variable Selection. *arXiv preprint :1510.06083*, 10 2015. URL https://arxiv.org/abs/1510.06083.

E. Hoerl, A. and W. Kennard, R. Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, 12:55–67, 1970.

Emmanuel, C. and Terence, T. The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n. *Institute of Mathematical Statistics*, 35:2313–2351, 2007.

Emmanuel J, C. Compressive Sampling. *Proceedings of the International congress of mathematicians*, 3:1433–1452, 2006.

Jerome, F., Trevor, H., Holger, H., and Robert, T. Pathwise Coordinate Optimization. *Institute of Mathematical Statistics*, pages 302–332, 2007.

Jolliffe, I. T., Trendafilov, N. T., and Mudassir, U. A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, 12:531–547, 2003.

Mazumder, R., Radchenko, P., and Dedieu, A. Subset Selection With Shrinkage: Sparse Linear Modeling When the SNR Is Low. *arXiv preprint :1708.03288*, 2017. URL https://arxiv.org/pdf/1708.03288.pdf.

Miller, A. *Subset Selection in Regression*. CRC Press, 2002.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer US, 2004.

Nesterov, Y. Smooth Minimization of Non-Smooth Functions. *Mathematical Programming*, 103:127–152, 2005.

Nesterov, Y. Gradient Methods for Minimizing Composite Objective Function. *Université catholique de Louvain, Center for Operations Research and Econometrics (CORE),Technical Report number 76*, 2007.

Niz, C. D., Rahman, R., Zhao, X., and Pal, R. Algorithms for Drug Sensitivity Prediction. *Algorithms*, 9:77, 2016.

Pham, V. *Sparse optimization models with robust sketching and applications*. Phd, University of California, 2016.

Pilanci, M., Wainwright, M. J., and Ghaoui, L. E. Sparse Learning via Boolean Relaxations. *Mathematical Programming*, 151:63–87, 2015.

Rahul, M., Trevor, H., and Robert, T. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Ryan J., T., Holger, H., and Robert, T. Nearly-Isotonic Regression. *Technometrics*, 53:54–61, 2011.

Tibshirani, R. Regression Shrinkage and Selection via The LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1995.

Wikipedia. Convex Optimization. Wikipedia, the Free Encyclopedia, https://en.wikipedia.org/wiki/Convex_optimization, Accessed April 2019.

Witten, D. M., Robert, T., and Trevor, H. A Penalized Matrix Decomposition, With Applications to Sparse.

Yuan, M. and Lin, Y. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94:19–35, 2007a.

Yuan, M. and Lin, Y. Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2007b.

Zhang, C.-H. Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *Annals of Statistics*, 38:894–942, 2010.

Zhang, C.-H. and Zhang, T. A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems. *Statistical Science*, 27:576–593, 2012.

Zou, H. The Adaptive LASSO and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Zou, H. and Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301–320, 2005.