THEORETICAL RESULTS FOR THE LASSO AND A NEW ROOT-LOG CONCAVE REGULARISER IN HIGH-DIMENSIONAL LINEAR REGRESSION

Tony Wang

Supervisor: Dr. Zdravko Botev

School of Mathematics and Statistics
UNSW Sydney

August 2022

Submitted in partial fulfilment of the requirements of the degree of
Master of Mathematics

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed:  Date: 04/08/2022

Acknowledgements

It has been a long two years, and am I ever so glad to be finishing up the Masters of Mathematics and putting this period of my life behind me.

I would like to give thanks to a number of key supporters who have held me up through this period of my life.

Firstly, to my housemates, who remember me despite my insistence on disappearing into my room at the first hint of stress, and who suffered and laughed with me through our most unique housing situation in 2022.

Next, to my parents, who have always been the most ardent of my supporters, and who have always welcomed me back home any time.

Finally, to my supervisor Zdravko Botev, without whom this project would not have been completed, let alone started. His technical expertise and ability to illuminate complex ideas, particularly in the proofs of Chapter 3, has been a wonder to behold. It has been a pleasure discussing ideas, proofs and mathematics with him in our weekly meetings, and building up our understanding of the root-log penalty from the ground up. I can only hope that this thesis represents well the work that we did together.

Abstract

Under the high dimensional assumption where $p >> n$ and n is allowed to diverge to ∞, standard linear regression techniques such as the ordinary least squares method and stepwise regression perform poorly. Model selection is essential to reduce the dimensionality of the problem, improve interpretability of the model and reduce overfitting on unseen data.

Regression models which naturally perform model selection include the well-known lasso and a large class of concave penalised least squares estimators, where the L^1-penalty of the lasso is replaced by a concave penalty function satisfying certain regularity conditions. Under the high-dimensional assumption, an important question for such regression models is as follows: given that there exists an underlying sparse structure, under what conditions can the regression technique recover this structure? This is termed the model selection consistency property, and it is desirable for this property to hold under easily satisfiable conditions.

The end goal of this thesis is to study the model selection properties of a penalised least squares estimator, where the penalty function is the newly-introduced "root-log penalty", a symmetric, concave function that is non-differentiable at 0. The root-log penalty is of particular interest due to the special form of its soft-thresholding one-dimensional solution; furthermore, due to its concavity and behaviour under computer simulation, it is believed to achieve model selection consistency under less restrictive conditions than the lasso.

To understand how these facts may be shown to hold, we first study the theoretical properties of the lasso. We give proofs of existence and uniqueness of solutions, find asymptotic bounds on the parameter error and mean squared error, and present the proof of model selection consistency under the strong irrepresentability condition. We then delve into the general theory of concave penalised least squares estimators by first considering properties of their solutions under the orthogonal ($p < n$) assumption, before studying in detail the behaviour of local minimisers of the root-log penalised problem under the high dimensional assumption. In particular, this thesis presents two main results concerning their model selection properties: under sufficient conditions that are satisfied with high probability (in particular, without requiring the strong irrepresentability condition), our first result states that there exists a local minimiser that attains the same support set as the true regression parameter, and our second result states that path-finding algorithms starting from $\mathbf{0}$ will find this model selection consistent solution with high probability. The proof technique for these two theorems is derived from similar methods for the lasso and select papers in the concave penalised least squares estimator literature, including recent work by Feng & Zhang on sorted concave regression models.

Contents

Chapter 1

High Dimensional Regression Theory

Suppose we make n observations $\{(\mathbf{X}_i, Y_i)\}_{1 \leq i \leq n}$, with each $\mathbf{X}_i \in \mathbb{R}^p$ containing p variables of interest and $Y_i \in \mathbb{R}$ the associated response variable. For simplicity, we preclude discrete/categorical variables and assume that each covariate is measured on a continuous scale in $\mathbb{R}$. Collect up the n observations $\mathbf{X}_i$ as rows in a design matrix $\mathbf{X}$, with columns labelled $\boldsymbol{v}_j, 1 \leq j \leq p$, and the response variables in a vector $\boldsymbol{Y} \in \mathbb{R}^n$. In linear regression, we are interested in modelling the linear relationship between the covariates $\mathbf{X}$ and the response variables $\boldsymbol{Y}$

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{1.0.1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is an as-of-yet unknown true regression parameter that we wish to estimate, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(0, \sigma^2 \boldsymbol{I})$ is a vector of independent homoscedastic normal error terms, each with the same variance σ^2.

Classical regression analysis assumes we have more data n points than variables p. In this thesis, we will be exploring regression techniques that perform well under the *high-dimensional assumption*: that is, we have $p = p_n >> n$ and in general we allow $p \to \infty$ as $n \to \infty$. In this situation, standard regression techniques perform poorly, if at all. For example, the design matrix $\mathbf{X}$ cannot be of full rank, so that $\mathbf{X}^\top \mathbf{X}$ is not invertible, and the ordinary least squares estimate is not well-defined. Even replacing the inverse with a matrix pseudo-inverse, an infinite number of solutions exist that minimise the residual sum of squares, making interpretation of the model difficult. Stepwise regression methods are computationally infeasible due to the exponentially increasing number of models that need to be considered.

In order to make any progress, we must enforce a number of *regularity conditions* on the sparsity of the true model $\boldsymbol{\beta}^*$ and on the design matrix $\mathbf{X}$. In particular, we will work under the hard-sparse assumption, where the true regression vector $\boldsymbol{\beta}^*$ is supported solely on a set $\mathcal{S} \subsetneq \{1, 2 \ldots, p\}$, and the goal is to attempt to recover $\boldsymbol{\beta}^*$ with an estimator $\hat{\boldsymbol{\beta}}$ that has the same support $\mathcal{S}$. An estimator $\hat{\boldsymbol{\beta}}$ (and the regression model associated with it) that is able to recover the true support $\mathcal{S}$ is called *model selection consistent*.

The *bet on sparsity* principle in statistical learning (as described in p.2 [7]) justifies why the hard-sparsity assumption is of theoretical interest. Indeed, if there *were* a true sparse structure in $\boldsymbol{\beta}^*$, methods designed to reveal this structure (such as the lasso) would perform well. On the other hand, in the absence of such a structure, *no models perform well*, owing to the lack of data compared to the number of parameters that must be estimated. Hence, it makes sense to default to a sparse regression model in the high-dimensional case.

There are also a myriad of real-world justifications for working under the hard sparse assumption. Well-chosen sparse models are computationally efficient, more interpretable and generally exhibit better generalisation to unseen data compared to overfit, large parameter models. There may be rigorous, scientific reasons to assume that only a small subset of the parameters is of interest. In the statistical analysis of genomics, for example, out of a large number of possible interacting genes, we may reasonably expect that only a small number influence the expression of a certain trait.

As part of our study of regression techniques which naturally induce sparsity, we give a brief description of the following regression models:

(i) Lasso.
(ii) General concave penalised least squares estimators (concave PLSE), including SCAD and MCP.
(iii) Root-log penalty.

The lasso, SCAD and MCP were first studied in the literature years prior, while the root-log penalty is a new penalty function that is introduced for the first time in this thesis. Interest in the root-log penalty was motivated by good performance in model selection under computer simulation in previous work by the supervisor of this thesis. In Section 1.1.4 (and later on, as a focus in Chapter 3), we will describe its properties and why it is of particular interest among a number of explicit examples of concave regularised penalty functions.

We note that our work in Chapter 1 mainly serves as references, and we aim for the tone to be explanatory; we will not give proofs, rigorous definitions of terminology or state theorems, but instead point to them in the later chapters while giving an intuitive description of what these results mean. Our aim is to produce a coherent narrative in the study of sparse penalised regression models, from the lasso to concave regularised models to the root-log penalty.

1.1 Regression Models in Sparse Statistical Learning

1.1.1 The Lasso

In 1996, the seminal paper [9] was published, introducing the Least Absolute Shrinkage and Selection Operator (LASSO, or lasso) to the statistics literature. In lasso regression, we penalise the residual sum of squares by the L^1-norm of the parameters $\boldsymbol{\beta}$: that is, the lasso solution arises from the minimisation problem

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{1.1.1}$$

The Lagrangian dual formulation of (1.1.1) gives some more insight into the nature of the lasso solution: it is given by

$$\begin{gathered} \hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\}, \\ \text{subject to } \|\boldsymbol{\beta}\|_1 \leq \lambda. \end{gathered} \tag{1.1.2}$$

From (1.1.2), we see that the lasso solution is the minimiser of the residual sum of squares when $\boldsymbol{\beta}$ is contained within the λ-sphere in the L^1-norm. Since the lasso penalty (1.1.1) is convex in $\boldsymbol{\beta}$, we can apply the subgradient calculus (see Section A.2 and Example A.2.1) to obtain necessary and sufficient conditions for solutions of (1.1.1):

$$\boldsymbol{\gamma} = \frac{1}{n\lambda}\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{1.1.3}$$

where

$$\gamma_i \in \begin{cases} \{\text{sign}(\hat{\beta}_i)\} & \text{if } \hat{\beta} \neq 0, \\ [-1, 1] & \text{if } \hat{\beta}_i = 0. \end{cases}$$

is the subgradient of the L^1-norm $\|\cdot\|_1$.

In the case where $\hat{\beta}_i = 0$, the subgradient calculus allows us to take γ_i to be the "most favourable" value (in an inequality, for example), or otherwise taken as any arbitrary value in $[-1, 1]$. The subgradient equation (1.1.3) can also be interpreted as the Karush-Kuhn-Tucker (KKT) conditions for the minimisation problem (1.1.1); such first-order conditions for local minimality play a center role in the analysis of model selection consistency of the lasso (see, for example, Section 2.3), and also of concave penalised estimators (see Section 3.2).

In the case where $\mathbf{X}$ is orthogonal (necessarily, $p < n$), the minimisation problem (1.1.1) decomposes into coordinate-wise minimisation problems; these can be solved with simple calculus (Proposition 2.2.1) to obtain

$$\hat{\beta}_i^{\text{LASSO}} = \text{sign}(\beta_i^{\text{LS}})(|\beta_i^{\text{LS}}| - \lambda)_+.$$

From the above equation, we see that the lasso shifts the least squares estimate towards 0 by λ, setting all coordinates that change sign after this shifting to 0, automatically giving us sparse models when λ is large enough. From the orthogonal analysis, we see that the lasso is biased, tending to indiscriminately underestimate the size of parameters by approximately λ, even when the parameters themselves are large and hence unlikely to be zero in the true model.

Under the high-dimensional assumption, the orthogonal analysis does not hold anymore, and it is unclear whether solutions for (1.1.1) are even unique, as they were in the orthogonal case. It can be shown, however, that under the assumption that the entries of $\mathbf{X}$ are drawn from a continuous distribution, the solutions of (1.1.1) are unique with probability 1 (Corollary 2.4.1, which was derived from work in [10]), and under further regularity assumptions on $\mathbf{X}$ and a lower bound on the magnitude of λ, we can find asymptotic bounds on the L^2-parameter error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ and the mean square error

$\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}^*\|_2^2}{n}$, of the form

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2 \in \mathcal{O}\left(\frac{\log p}{n}\right),$$
$$\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}^*\|_2^2}{n} \in \mathcal{O}\left(\frac{\log p}{n}\right).$$

The proof of these asymptotic upper bounds is the subject of Theorem 2.6.1, Theorem 2.7.1 and Corollary 2.7.1.

Of course, simply knowing bounds on the parameter error gives us no information on the *structure* of the model found by $\hat{\boldsymbol{\beta}}$. The question of interest is as follows: given that the true model $\boldsymbol{\beta}^*$ is supported solely on a set $\mathcal{S}$, can we guarantee that the lasso solution is model selection consistent, i.e., supported on $\mathcal{S}$ as well? As it turns out, the answer is "yes", but not without some further assumptions. Under the *strong irrepresentability condition* on $\mathbf{X}$ (Definition 2.8.1) and a *minimal signal strength condition* (Definition 2.8.2) on $\boldsymbol{\beta}^*$, the unique lasso solution can be shown to be model selection consistent, the proof of which is the subject of Theorem 2.8.1. Now, while the minimal signal strength condition (a lower bound on the magnitude of the active variables in $\boldsymbol{\beta}^*$) is, intuitively, a realistic condition to enforce (signals that are too weak would be masked by the noise $\boldsymbol{\varepsilon}$ and hence be impossible to distinguish), the strong irrepresentability condition does not scale well as p and n increases due to spurious correlations occurring in high dimensional datasets, hindering the lasso's model selection consistency. One can interpret this as the overly high biased estimation of the lasso reduces its ability to detect the true model when spurious correlations between noise and true variables occurs at a sufficiently high level.

We may wonder if the strong irrepresentability condition is necessary if we are able to reduce this bias somehow. This leads us to consider properties of estimators when the inflexible L^1-norm is replaced by a suitable concave penalty function. We discuss this class of estimators in the next section.

1.1.2 Concave Penalised Least Squares Estimators

As we noted in the previous section, the lasso is biased, with bias increasing to the order of λ. This is an important issue, since to get good sparse models, we need the regularisation parameter λ to be "large enough" to produce sparse solutions that are consistent with the true model $\boldsymbol{\beta}^*$ (a lower bound is given in, for example, Theorem 2.6.1), yet also small enough for solutions to be "close" to the true $\boldsymbol{\beta}^*$. The fault lies with the inflexible, constant penalisation of the L^1-norm (see Figure 1.1), even when the parameters $\hat{\boldsymbol{\beta}}$ are large and hence are unlikely to be zero in the true model $\boldsymbol{\beta}^*$.

If we replace the L^1-penalty function with a concave penaliser ρ_λ with gradient decaying to zero for large parameters, we would naturally expect a smaller bias compared to the lasso, and possibly better theoretical properties for estimators generated from such penalisers. Formally, we are interested in (local) minimisers of the optimisation problem

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\sum_{i=1}^{p}\rho_\lambda(\beta_i)\right\}. \tag{1.1.4}$$

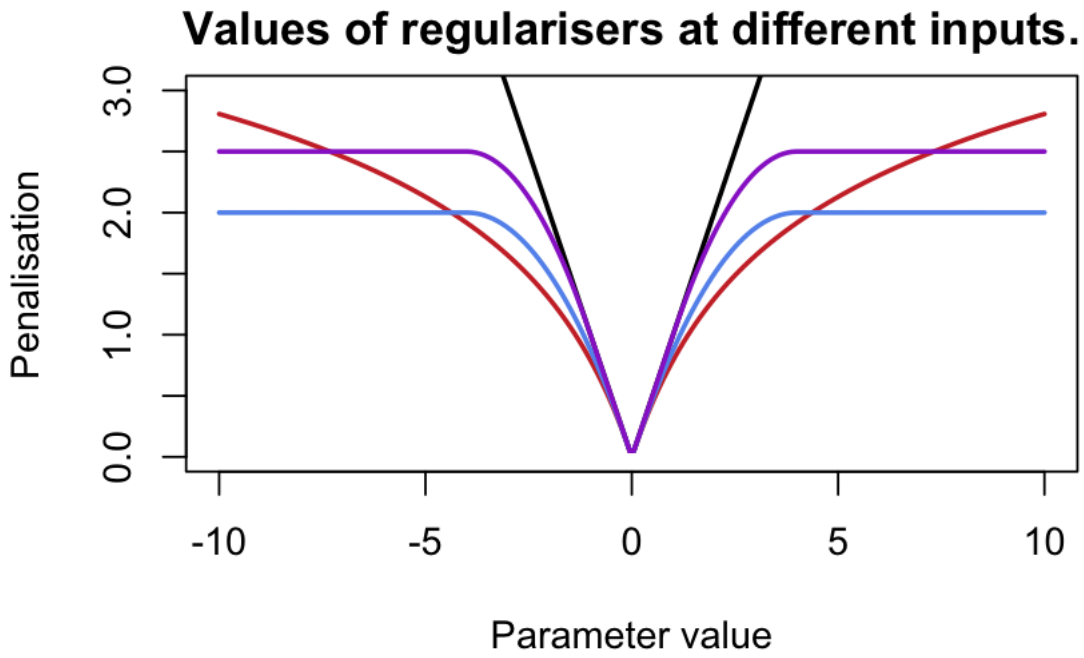

Figure 1.1: Plots of the lasso (black), MCP (blue), SCAD (purple) and root-log regulariser (orange) at different parameter values with $\gamma = 4$ and $\lambda = 1$. Aside from the lasso, the other regularisers are folded concave regularisers with a cusp at 0. Both MCP and SCAD are constant after the critical value $|t| = \gamma = 4$, while the root-log regulariser has an everywhere non-zero derivative that vanishes in the limit $|t| \to \infty$.

where $\rho_\lambda(\cdot) = \lambda^2\rho(\cdot/\lambda)$ for a symmetric, concave function ρ (for a full list of properties, see Chapter 3). Penalty functions of this type are called "folded concave penalisers", and solutions of the minimisation problem (1.1.4) are called "(folded) concave penalised least squares estimators" (concave PLSE), in reference to the shape of the one-dimensional penalty ρ. Plots of such functions in one-dimension with $\lambda = 1$ can be found in Figure 1.1, including the lasso, SCAD, MCP and the root-log penalty, the last three of which we will describe later.

We draw particular attention to the singular derivatives of each regulariser at $0 \in \mathbb{R}$ and the asymptotic gradients of each regulariser in the limit $|t| \to \infty$. As we will present in Theorem 3.1.1, the singular derivative on the regulariser is essential for inducing sparsity in the final model, and is what differentiates regression models such as ridge regression (which does not produce sparse solutions) with the lasso. Furthermore, the asymptotically zero gradient is what differentiates the sparse solutions produced by the lasso and those produced by concave PLSE: the constant gradient of the L^1-norm induces inflexible penalisation and hence difficulty in uncovering the true sparse $\boldsymbol{\beta}^*$, if there exists such a structure. Surprisingly, this difference is what allows us to prove model selection consistency for the concave PLSE without the restrictive strong irrepresentability condition, the proof which is the goal of Chapter 3.

To complete our review, we introduce SCAD and MCP as examples of concave penalty functions that have been studied in the literature, before introducing a novel concave penalty function which we dub the "root-log regulariser" or "root-log penalty".

1.1.3 SCAD & MCP

The Smoothly Clipped Absolute Deviation penalty (SCAD) was first introduced in [4] as an explicit example of a concave regulariser with a number of good properties. It is defined as follows. If $\gamma > 2$, then the one-dimensional penalty is

$$\rho^{\mathrm{SCAD}}(t) := \int_0^{|t|} \left(1 - \frac{x}{\gamma}\right)_+ dx,$$

and the regularised penalty is $\rho_\lambda^{\mathrm{SCAD}}(t) = \lambda^2 \rho_\lambda^{\mathrm{SCAD}}(t/\lambda)$. From p.134, [3], we obtain explicit formulas for the regularised penalty function, the derivative and the one-dimensional minimisation solution. Firstly, the regularised penalty function is given by

$$\rho_\lambda^{\mathrm{SCAD}}(t) = \begin{cases} \lambda|t| & \text{if } |t| \leq \lambda, \\ \frac{2\gamma\lambda|t| - t^2 - \lambda^2}{2(\gamma-1)} & \text{if } \lambda < |t| \leq \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2} & \text{otherwise.} \end{cases}$$

with derivative

$$\rho_\lambda^{\mathrm{SCAD}\prime}(t) = \lambda \left[1_{\{t \leq \lambda\}} + \frac{(\gamma\lambda - t)_+}{\lambda(\gamma-1)} 1_{\{t > \lambda\}}\right].$$

The SCAD penalty function corresponds to a quadratic spline with knots at λ and $\gamma\lambda$. Indeed, when $|t| \leq \lambda$, the SCAD penalty is the regularised lasso; when $\lambda < |t| \leq \gamma\lambda$, the SCAD penalty is a concave quadratic polynomial interpolating to $\frac{\lambda^2(\gamma+1)}{2}$, which the penalty function stays constant on when $|t| > \gamma\lambda$.

In one dimension, we can explicitly write out the solution of the minimisation problem

$$\operatorname*{argmin}_{\beta \in \mathbb{R}} \left\{ \frac{1}{2}(z - \beta)^2 + \rho_\lambda^{\mathrm{SCAD}}(\beta) \right\}.$$

It is given by

$$\hat{\beta}^{\mathrm{SCAD}}(z) = \begin{cases} (|z| - \lambda)_+ & \text{if } |z| \leq 2\lambda, \\ \frac{(\gamma-1)z - \gamma\lambda}{\gamma-2} & \text{if } 2\lambda < |z| \leq \gamma\lambda, \\ z & \text{if } |t| > \gamma\lambda. \end{cases}$$

which is the lasso solution when $|z| \leq 2\lambda$, the unchanged solution z when $|t| > \gamma\lambda$ and a linearly interpolated value between $|z| - \lambda$ and z when $2\lambda < |z| \leq \gamma\lambda$.

The Minimax Concave Penalty (MCP) is a concave penalty function that was introduced in [11]. It is defined as follows: for $\gamma > 1$, the one-dimensional penalty is

$$\rho^{\mathrm{MCP}}(t) = \int_0^{|t|} \left(1 - \frac{x}{\gamma}\right)_+ dx,$$

and $\rho_\lambda^{\mathrm{MCP}}(t) = \lambda^2 \rho_\lambda^{\mathrm{MCP}}(t)$. Explicitly, we can write

$$\rho_\lambda^{\mathrm{MCP}}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\gamma} & \text{if } |t| \le \gamma\lambda \\ \frac{\gamma\lambda^2}{2} & \text{otherwise.} \end{cases}$$

with derivative

$$\rho_\lambda^{\mathrm{MCP}\prime}(t) = \begin{cases} \operatorname{sign}(t)\left(\lambda - \frac{|t|}{\gamma}\right) & \text{if } |t| \le \gamma\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

The MCP is of particular interest since it has the property of minimising the maximum concavity κ_ρ (defined in Definition 3.1.1). In the study of asymptotic bounds and model selection consistency, this allows a smaller lower bound in certain inequalities (as an example, see our work in Section 3.2).

The curves of $\rho_1^{\mathrm{SCAD}}(t)$ and $\rho_1^{\mathrm{MCP}}(t)$ are plotted in purple and blue respectively with $\gamma = 4$ for both curves in Figure 1.1. Of particular note is the fact that SCAD and MCP have zero gradient after a finite point, while the gradient of the root-log regulariser (to-be-introduced) *decays* to zero in the limit $t \to \infty$.

1.1.4 Root-Log Regulariser

Finally, we introduce the special regularising function that motivated our initial interest in concave penalised least squares estimators and the work in this thesis. It is defined as follows. First, define the one-dimensional penalty ρ^{RL} as

$$\rho^{\mathrm{RL}}(t) = \frac{\sqrt{t^4 + 4t^2} - t^2}{4} + \frac{1}{2}\ln\left(1 + \frac{\sqrt{t^4 + 4t^2} + t^2}{2}\right). \tag{1.1.5}$$

Then the regularised root-log penalty[1] as $\rho_\lambda^{\mathrm{RL}}(t) = \lambda^2 \rho^{\mathrm{RL}}(t/\lambda)$. The *root-log estimator* is found as follows (if it exists):

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^p \rho_\lambda^{\mathrm{RL}}(\beta_i) \right\}. \tag{1.1.6}$$

It is differentiable everywhere except 0, with the derivative

$$\rho'(t) = \frac{\sqrt{t^2\,(t^2+4)} - t^2}{2t} = \frac{\operatorname{sign}(t)\sqrt{t^2+4} - t}{2} \tag{1.1.7}$$

[1]The naming of the root-log penaliser is not quite so inspired: it merely speaks to the presence of the logarithm and the square roots in (1.1.5).

and regularised derivative

$$\frac{d}{dt}\rho_\lambda(t) = \lambda^2 \frac{d}{dt}\rho(t/\lambda) = \lambda\rho'(t/\lambda).$$

Despite the intimidating nature of (1.1.5), the root-log regulariser is of particular interest among the folded concave regularisers due to its computational properties. Indeed, as we prove in Proposition 3.1.1, it is known that the one-dimensional minimiser of (1.1.6) is of the form

$$\hat{\beta}(z) = z\left(1 - \frac{\lambda^2}{z^2}\right)_+. \tag{1.1.8}$$

Hence, a simple algorithm to try to compute minimisers of (1.1.6) involves cyclical coordinate gradient descent using (1.1.8) as the coordinate updates, similar to how gradient descent is used for the lasso problem. However, one should note that since (1.1.6) is non-convex in $\boldsymbol{\beta}$, there are no theoretical guarantees (in general) that any solution found by such numerical methods is a global minimum. Hence, our work in Chapter 3 focuses on applying the first-order KKT conditions to divine properties of *local minimisers* of (1.1.6). Such an issue does not exist for the lasso due to the convexity of the objective function.

With the basic review complete, we turn to the formal study of theoretical properties of the lasso in Chapter 2.

Chapter 2

Lasso Regression

2.1 Introduction

Throughout this chapter, fix a vector of observations $\boldsymbol{Y} \in \mathbb{R}^p$ and a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, with columns $\boldsymbol{v}_i$, $1 \leq i \leq p$ representing a measured variable and rows $\mathbf{X}_i$, $1 \leq i \leq n$ representing a data point. We consider the linear model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{2.1.1}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ is assumed to be a vector of independent normal errors each with variance σ^2, and $\boldsymbol{\beta}^*$ is an unknown vector of coefficients representing the *true* regression parameters. The lasso regression problem solves the following minimisation problem

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.1.2}$$

where $\lambda > 0$ is some regularisation parameter.

2.2 Orthogonal Analysis

In order to understand the behaviour of lasso solutions, we first consider the simple case where $\mathbf{X}$ is a normalised orthogonal matrix in the low-dimensional $p < n$ setting. Although this analysis cannot be used in the general setting $p >> n$ that we are interested in, the general tendency of the lasso to select sparse models is made clear in the following proposition.

Proposition 2.2.1. Suppose the design matrix $\mathbf{X}$ is normalised so that $\mathbf{X}^\top \mathbf{X} = n\boldsymbol{I}$ (necessarily, this implies $p < n$). Then the lasso problem (2.1.2) has solution $\hat{\boldsymbol{\beta}}$ given by

$$\begin{aligned} \hat{\beta}_i &= \operatorname{sign}(\hat{\beta}_i^{\text{LS}})(|\hat{\beta}_i^{\text{LS}}| - \lambda)_+ \\ &= \begin{cases} \operatorname{sign}(\hat{\beta}_i^{\text{LS}})(|\hat{\beta}_i^{\text{LS}}| - \lambda) & \text{if } |\hat{\beta}_i^{\text{LS}}| > \lambda, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.2.1}$$

That is, the solution to the lasso is simply the ordinary least squares estimate shifted down by the regularisation parameter, thresholded to equal zero if the estimate changes sign.

Proof. From the basic theory, the ordinary least squares (OLS) estimate $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$ is given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{Y} = \frac{1}{n}\mathbf{X}^\top \boldsymbol{Y}. \tag{2.2.2}$$

By expanding the function in (2.1.2), we get the equivalent minimisation problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} &= \min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n}\boldsymbol{Y}^\top \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \right\} \\ &= \min_{\boldsymbol{\beta}} \left\{ -(\hat{\boldsymbol{\beta}}^{\mathrm{LS}})^\top \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \right\}. \end{aligned} \tag{2.2.3}$$

Now, noticing that the objective function is separable into its individual components, it suffices to minimise

$$\mathcal{L}_i(\beta_i) = -\beta_i^{\mathrm{LS}}\beta_i + \frac{1}{2}\beta_i^2 + \lambda|\beta_i|$$

for each fixed $1 \leq i \leq p$. Now, we note that $\mathrm{sign}(\beta_i^{\mathrm{LS}}) = \mathrm{sign}(\beta_i)$. Indeed, the remaining terms β_i^2 and $|\beta_i|$ are non-negative, and so any minimiser β_i of $\mathcal{L}_i$ must make the term $-\beta_i^{\mathrm{LS}}\beta_i$ negative. Hence, we can deal with minimising $\mathcal{L}_i$ in cases depending on the sign of β_i^{LS}:

(i): If $\beta_i^{\mathrm{LS}} = 0$, then clearly the minimiser of $\mathcal{L}_i$ is $\hat{\beta}_i = 0$.

(ii): If $\beta_i^{\mathrm{LS}} > 0$, then we must have $\beta_i > 0$. Setting the derivative of $\mathcal{L}_i$ equal to zero, we have

$$\beta_i = \beta_i^{\mathrm{LS}} - \lambda.$$

Now, if $\beta_i^{\mathrm{LS}} < \lambda$, then we should set $\hat{\beta}_i = 0$, since we can write $\mathcal{L}_i(\beta_i) = \frac{1}{2}\beta_i^2 + (\lambda - |\beta_i^{\mathrm{LS}}|)|\beta_i|$; hence, we conclude that

$$\hat{\beta}_i = (\beta_i^{\mathrm{LS}} - \lambda)_+ = \mathrm{sign}(\beta_i^{\mathrm{LS}})(|\beta_i^{\mathrm{LS}}| - \lambda)_+.$$

(iii): If $\beta_i^{\mathrm{LS}} < 0$, then similarly to the positive case, we must have $\beta_i < 0$, and repeating the derivation, we obtain

$$\hat{\beta}_i = -(-\beta_i^{\mathrm{LS}} - \lambda)_+ = \mathrm{sign}(\beta_i^{\mathrm{LS}})(|\beta_i^{LS}| - \lambda)_+.$$

Hence, the solution $\hat{\boldsymbol{\beta}}$ is as claimed in (2.2.1). □

Remark 2.2.1. This result shows why the lasso is important in *sparse statistical learning*. In ordinary linear regression, hypothesis tests on the parameters $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$ provide a rigorous way of concluding whether the optimal regression parameters are (statistically) close enough to zero to have arisen from random chance. This allows one to make inferences on the presence and direction of effect in a model. Variable selection, overall, is an important part of the model building process in any statistical problem. On the other hand, we have shown rigorously (at least in the orthogonal case) that the lasso *automates* the model selection process: it naturally produces sparse solutions, in the sense that if any of the

optimal parameters in $\hat{\boldsymbol{\beta}}^{\text{LASSO}}$ are small enough, the lasso sets it to zero. Indeed, trivially, if we pick $\lambda > \|\hat{\boldsymbol{\beta}}^{\text{LS}}\|_\infty$, then we must have $\hat{\boldsymbol{\beta}} = \mathbf{0}$. In real world datasets, the optimal λ would be picked via a combination of prior knowledge and cross-validation on training data.

This is only the beginning of the theory on the lasso problem. By imposing full-rank assumptions on $\mathbf{X}$, elementary methods allow us to obtain a closed-form solution of the lasso. Such a nice closed-form solution for $\hat{\boldsymbol{\beta}}$ does not hold under the high-dimensional assumption $p >> n$: indeed, it is not immediately obvious that a (global) solution must exist at all, and if it does, that such a solution is unique. We tackle such issues in the next section.

2.3 Existence of Lasso Solutions

In this section, we follow the work of [10] to show that under continuous distribution assumptions on the columns of $\mathbf{X}$, we can guarantee both existence and uniqueness for the lasso minimisation problem (2.1.2). We are mainly interested in the case $p > n$ (when $p \leq n$, existence and uniqueness), although we will explicitly state if we enforce this assumption in our proofs.

Theorem 2.3.1 (Basic Facts)**.** Consider the problem (2.1.2). For any $\boldsymbol{Y}, \mathbf{X}$ and $\lambda \geq 0$, the following are satisfied:

(i) (Alternatives): There either exists a unique solution $\hat{\boldsymbol{\beta}}$ or uncountably many solutions to the lasso minimisation problem (2.1.2).

(ii) (Prediction & Parameter Equality): Any two solutions $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$ must give the same fitted values, and the parameter L^1-penalties must be equal. That is,

$$\mathbf{X}\hat{\boldsymbol{\beta}}^{(1)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(2)} \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}^{(1)}\|_1 = \|\hat{\boldsymbol{\beta}}^{(2)}\|_1.$$

Proof. From the convexity of the minimisation problem (2.1.2), we automatically obtain the existence of a minimiser. Since the minimisation problem (2.1.2) is convex, the solution set

$$\left\{ \hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \{ \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \} \right\}$$

is a convex set. Hence, if there exist distinct $\hat{\boldsymbol{\beta}}^{(1)} \neq \hat{\boldsymbol{\beta}}^{(2)}$ that are solutions to (2.1.2), then the uncountably many convex combinations $\alpha\hat{\boldsymbol{\beta}}^{(1)} + (1-\alpha)\hat{\boldsymbol{\beta}}^{(2)}$ are also solutions to (2.1.2) for $\alpha \in [0, 1]$. For the second part, suppose by contradiction the existence of lasso solutions $\hat{\boldsymbol{\beta}}^{(1)} \neq \hat{\boldsymbol{\beta}}^{(2)}$ such that $\mathbf{X}\hat{\boldsymbol{\beta}}^{(1)} \neq \mathbf{X}\hat{\boldsymbol{\beta}}^{(2)}$. Then for any $\alpha \in (0, 1)$, the solution $\alpha\hat{\boldsymbol{\beta}}^{(1)} + (1-\alpha)\hat{\boldsymbol{\beta}}^{(2)}$ achieves a lower value of the objective function than $\hat{\boldsymbol{\beta}}^{(1)}$ or $\hat{\boldsymbol{\beta}}^{(2)}$. To

see this, if we set c as the common minimum value of the lasso's objective function, we get

$$\begin{aligned}
&\|\boldsymbol{Y} - \mathbf{X}(\alpha\hat{\boldsymbol{\beta}}^{(1)} + (1-\alpha)\hat{\boldsymbol{\beta}}^{(2)})\|_2^2 + \lambda\|\alpha\hat{\boldsymbol{\beta}}^{(1)} + (1-\alpha)\hat{\boldsymbol{\beta}}^{(2)}\|_1 \\
&< \alpha\|\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(1)}\|_2^2 + (1-\alpha)\|\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(2)}\|_2^2 + \alpha\lambda\|\hat{\boldsymbol{\beta}}^{(1)}\|_1 + (1-\alpha)\lambda\|\hat{\boldsymbol{\beta}}_1^{(2)}\| \\
&= \alpha\left(\|\boldsymbol{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}}^{(1)}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}^{(1)}\|_1\right) + (1-\alpha)\left(\|\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(2)}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}^{(2)}\|_1\right) = c,
\end{aligned}$$

where the hard inequality results from the strict convexity of $\|\cdot\|_2^2$ (this follows immediately since the Hessian $\nabla^2\|\cdot\|_2^2 = 2I > 0$ is positive-definite).

Knowing that any two solutions must give the same fitted values automatically gives us that the L^1-penalties must be equal. Indeed, the value of the function $\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ must be equal for any two solutions, and we know that the fitted values $\boldsymbol{X}\,\hat{\boldsymbol{\beta}}$ must be equal as well, which forces the L^1-penalties $\|\hat{\boldsymbol{\beta}}\|_1$ to be equal as well. □

In order to obtain more powerful results, we must employ the first order KKT conditions that minimisers of (2.1.2) must satisfy. The lasso KKT conditions are as follows:

$$\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = n\lambda\boldsymbol{\gamma}, \tag{2.3.1}$$

where $\boldsymbol{\gamma}$ is the subgradient of $\|\cdot\|_1$:

$$\boldsymbol{\gamma} \in \partial\|\hat{\boldsymbol{\beta}}\|_1 \implies \gamma_i \in \begin{cases} \{\operatorname{sign}\hat{\beta}_i\}, & \text{if } \hat{\beta}_i \neq 0, \\ [-1, 1], & \text{if } \hat{\beta}_i = 0. \end{cases} \tag{2.3.2}$$

If interested, we direct readers to the Appendix for theory on the subgradient calculus and how (2.3.1) is derived.

2.4 Uniqueness of Lasso Solutions

In this section, we give some useful characterisations for when we can expect the solution of the lasso problem to be unique. From now on, we assume $\lambda > 0$ (note that the case $\lambda = 0$ corresponds to the ordinary least squares problem).

Definition 2.4.1. The *equicorrelation set* is given by

$$\mathcal{S} = \{i \in \{1, 2, \ldots, p\} : |\boldsymbol{v}_i^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})| = \lambda\}. \tag{2.4.1}$$

We may also refer to these as the *equicorrelation indices*, and the corresponding columns $\boldsymbol{v}_i$ ($i \in \mathcal{S}$) to be the *equicorrelation variables*. The *equicorrelation sign* is given by

$$\boldsymbol{s} = \operatorname{sign}(\mathbf{X}_\mathcal{S}^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})) \in \mathbb{R}^{|\mathcal{S}|}. \tag{2.4.2}$$

The set $\mathcal{S}$ corresponds exactly to the columns of $\mathbf{X}$ which attain the largest possible correlation with the residuals $\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. By the KKT conditions (2.3.1), the set $\mathcal{S}^c = \{1, 2, \ldots, p\} \setminus \mathcal{S}$ corresponds to a *subset* of the indices of $\hat{\boldsymbol{\beta}}$ which must be zero. Note the emphasis on subset: the equicorrelation indices may not be *exactly* the indices not supported by $\hat{\boldsymbol{\beta}}$, since the subdifferential $\boldsymbol{\gamma}$ could still take values -1 or 1 when $\hat{\boldsymbol{\beta}}_i = 0$.

Hence, we need to make distinct the idea of the *active* subset $\mathcal{A} := \operatorname{supp}(\hat{\boldsymbol{\beta}}) \subset \mathcal{S}$ of the lasso solution versus the equicorrelation set. The equicorrelation sign $\boldsymbol{s}$ gives the direction of the correlations of the active constraint variables with the residuals, and is equal to the subgradient $\boldsymbol{\gamma}$ on the indices i where $|\gamma_i| = 1$.

In the following theorem, we shall show that uniqueness of the lasso solution is determined, in some sense, by the kernel of the design matrix $\mathbf{X}$ on the equicorrelation set $\mathcal{S}$. When the kernel is trivial, the estimated solution is exactly zero on the non-equicorrelation indices; otherwise, it is given by a solution very similar to the least-squares solution shifted towards 0 by λ:

Theorem 2.4.1 (Equicorrelation Condition for Uniqueness)**.** If $\ker(\mathbf{X}_{\mathcal{S}}) = 0$, then for any $\boldsymbol{Y}, \mathbf{X}$ and $\lambda > 0$, the solution to the lasso problem (2.1.2) is unique and is given by

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0, \quad \hat{\boldsymbol{\beta}}_{\mathcal{S}} = (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1}(\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{Y} - \lambda \boldsymbol{s}) \tag{2.4.3}$$

Proof. Since $\hat{\beta}_i = 0$ for $i \in \mathcal{S}^c$, we can rewrite the linear equations (2.3.1) to those on $\mathcal{S}$ only:

$$\mathbf{X}_{\mathcal{S}}^\top (\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}} \hat{\boldsymbol{\beta}}_{\mathcal{S}}) = \lambda \boldsymbol{s}. \tag{2.4.4}$$

From this equation, we see that $\lambda s \in \operatorname{Row}(\mathbf{X}_{\mathcal{S}}) = \operatorname{Range}(\mathbf{X}_{\mathcal{S}}^\top)$, and so by Proposition A.1.2, we can write $\lambda \boldsymbol{s} = \mathbf{X}_{\mathcal{S}}^\top (\mathbf{X}_{\mathcal{S}}^\top)^+ \lambda \boldsymbol{s}$, where $(\mathbf{X}_{\mathcal{S}}^\top)^+$ is the (unique) matrix pseudo-inverse of $\mathbf{X}_{\mathcal{S}}^\top$. Substituting this expression into (2.4.4), we obtain

$$\mathbf{X}_{\mathcal{S}}^\top (\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}} \hat{\boldsymbol{\beta}}_{\mathcal{S}}) = \lambda \boldsymbol{s} = \mathbf{X}_{\mathcal{S}}^\top (\mathbf{X}_{\mathcal{S}}^\top)^+ \lambda \boldsymbol{s},$$

and after some rearranging, we obtain

$$\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} \hat{\boldsymbol{\beta}}_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}}^\top (\boldsymbol{Y} - (\mathbf{X}_{\mathcal{S}}^\top)^+ \lambda \boldsymbol{s}), \tag{2.4.5}$$

which implies that the lasso solution $\mathbf{X}\hat{\boldsymbol{\beta}}$ is given by

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}_{\mathcal{S}} \hat{\boldsymbol{\beta}}_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}} (\mathbf{X}_{\mathcal{S}})^+ (\boldsymbol{Y} - (\mathbf{X}_{\mathcal{S}}^\top)^+ \lambda \boldsymbol{s}).$$

From here, we see that any lasso solution $\hat{\boldsymbol{\beta}}$ satisfies

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0 \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{S}} = (\mathbf{X}_{\mathcal{S}})^+ (\boldsymbol{Y} - (\mathbf{X}_{\mathcal{S}}^\top)^+ \lambda \boldsymbol{s}) + \boldsymbol{b} \tag{2.4.6}$$

for some $\boldsymbol{b} \in \ker(\mathbf{X}_{\mathcal{S}})$. In general, there could be infinitely many possible vectors $\boldsymbol{b}$ such that $\hat{\boldsymbol{\beta}}$ is a solution to the lasso minimisation problem(2.1.2): by the KKT conditions (2.3.1), we only need to make sure $\boldsymbol{b}$ is chosen so that

$$\operatorname{sign}(\hat{\beta}_i) = s_i = \gamma_i \quad \text{for } i \in \mathcal{S}. \tag{2.4.7}$$

Equivalently, we can write

$$s_i \cdot \operatorname{sign}(\hat{\beta}_i) \geq 0 \quad \text{for } i \in \mathcal{S}. \tag{2.4.8}$$

Of course, if $\ker(\mathbf{X}_\mathcal{S}) = 0$, then the solution $\hat{\boldsymbol{\beta}}$ in (2.4.6) is unique. The form of the solution in (2.4.3) is obtained by rearranging (2.4.6) using the generalised inverse identity $\boldsymbol{A}^+ = (\boldsymbol{A}^\top \boldsymbol{A})^+ \boldsymbol{A}^\top$. The proof is complete. □

However, we note that requiring $\ker(\mathbf{X}_\mathcal{S}) = 0$ is not very natural. Indeed, determining the set $\mathcal{S}$ requires us to find an optimal solution $\hat{\boldsymbol{\beta}}$ first and such a solution depends on $\boldsymbol{Y}$, $\mathbf{X}$ and λ. What we would like to obtain is a condition that is based strictly on the observed data. As it turns out, some natural conditions exist which implies that $\ker(\mathbf{X}_\mathcal{S}) = 0$, depending solely on the design matrix $\mathbf{X}$:

Definition 2.4.2 (Antipodal General Position)**.** A matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is in *antipodal general position* if, for all $k < \min\{n, p\}$, no k-dimensional subspace contains more than $k + 1$ elements of $\{\pm \boldsymbol{v}_1, \pm \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p\}$, where $\boldsymbol{v}_i$ is the ith column of $\mathbf{X}$.

We note that this definition may differ from typical definitions of general position in linear algebra. As we will see from the next lemma, the "antipodal general position" condition naturally falls out as a way to ensure the kernel $\ker(\mathbf{X}_\mathcal{S})$ is trivial for *any* possible equicorrelation set $\mathcal{S}$.

Lemma 2.4.1 (General Position implies Uniqueness)**.** If the columns of $\mathbf{X}$ are in antipodal general position, then for all $\boldsymbol{Y}$ and $\lambda > 0$, the lasso solution $\hat{\boldsymbol{\beta}}$ is unique.

Proof. It suffices to show that the columns of $\mathbf{X}$ being in general position implies that $\ker(\mathbf{X}_\mathcal{S}) = \{0\}$. By contraposition, suppose that $\ker(\mathbf{X}_\mathcal{S}) \neq \{0\}$, so there exists some $\boldsymbol{b} \neq 0$ such that $\mathbf{X}_\mathcal{S} \boldsymbol{b} = \{0\}$. Since $\boldsymbol{b} \neq 0$, we can rearrange this linear equation to get

$$\boldsymbol{v}_i = \sum_{j \in \mathcal{S} \setminus \{i\}} c_j \boldsymbol{v}_j$$

for some $i \in \mathcal{S}$ and some coefficients c_j, $j \in \mathcal{S} \setminus \{i\}$. Multiplying by s_i and noting that $s_j^2 = 1$ for all j, we get

$$s_i \boldsymbol{v}_i = \sum_{j \in \mathcal{S} \setminus \{i\}} (s_i s_j c_j) \cdot (s_j \boldsymbol{v}_j). \tag{2.4.9}$$

By definition of the equicorrelation sign $\boldsymbol{s}$ in (2.4.2) and the KKT conditions (2.3.1), we see that for all $j \in \mathcal{S}$, we have

$$\boldsymbol{v}_j^\top (\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = s_j \lambda,$$

and so by taking the dot product of (2.4.9) with $\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, we get

$$\begin{aligned}(\boldsymbol{s}_i \boldsymbol{v}_i)^\top (\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \sum_{j \in \mathcal{S} \setminus \{i\}} (s_i s_j c_j) \cdot (s_j \boldsymbol{v}_j)^\top (\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \sum_{j \in \mathcal{S} \setminus \{i\}} s_i s_j c_j \lambda.\end{aligned}$$

The left-hand side of the above equation simplifies to λ, since $(s_i \boldsymbol{v}_i)^\top (\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (s_i)^2 \lambda = \lambda$. Dividing by λ, we get

$$1 = \sum_{j \in \mathcal{S} \setminus \{i\}} s_i s_j c_j,$$

which shows that

$$s_i \boldsymbol{v}_i = \sum_{j \in \mathcal{S} \setminus \{i\}} (s_i s_j c_j) \cdot (s_j \boldsymbol{v}_j) \tag{2.4.10}$$

is in the affine span of the vectors $s_j \boldsymbol{v}_j$ for $j \in \mathcal{S} \setminus \{i\}$ (recall Definition A.1.2). Hence, if we can ensure that the affine span never contains elements of the form (2.4.10), no matter the set $\mathcal{S}$, we can be assured that $\ker(\mathbf{X}_\mathcal{S}) = 0$ always holds. This is exactly what is implied when $\mathbf{X}$ is in antipodal general position from Definition 2.4.1, finishing the proof. □

From here, we get a strong sufficient condition for uniqueness of the lasso solution (with probability 1):

Corollary 2.4.1 (Continuous Distributions)**.** If the entries of $\mathbf{X}$ are drawn from a continuous probability distribution, then the lasso solution $\hat{\boldsymbol{\beta}}$ is unique with probability 1.

Proof. It suffices to prove that if the entries of $\mathbf{X}$ are drawn from a continuous probability distribution, then the columns of $\mathbf{X}$ will, with probability 1, be in antipodal general position. Suppose $k < \min\{n, p\}$ and that we have drawn entries $x_1, x_2, \ldots, x_k$ from some continuous probability distribution (continuous with respect to the Lebesgue measure on $\mathbb{R}^{np}$). The affine span $\mathrm{Aff}\{x_1, x_2, \ldots, x_{k+1}\}$ has Lebesgue measure zero in $\mathbb{R}^n$, and so $P(x_{k+2} \in \mathrm{Aff}\{x_1, x_2, \ldots, x_{k+1}\}) = 0$; taking finite unions over all $k < \min\{n, p\}$ and all possible sign changes $\pm x_i$, we conclude that the columns of $\mathbf{X}$ are in antipodal general position with probability 1. □

It can be shown that the triviality of $\ker(\mathbf{X}_\mathcal{S})$ *characterises* the uniqueness of solutions for the lasso problem. The statement is given in the following lemma:

Lemma 2.4.2 (Necessary Condition for Uniqueness)**.** Suppose we have a fixed design matrix $\mathbf{X}$ and regularisation parameter $\lambda > 0$. For almost every response vector $\boldsymbol{Y}$, if there exists a unique solution $\hat{\boldsymbol{\beta}}$ to (2.1.2), then $\ker(\mathbf{X}_\mathcal{S}) = 0$.

The proof of this lemma relies on the fact that a modified version of the Least Angle Regression (LARS) algorithm converges to a true solution of the lasso, even in the case $\mathrm{rank}(\mathbf{X}) < p$ (which must occur when $p > n$). Indeed, as noted in p.8 [10], the basic LARS algorithm is not necessarily correct due to the possible non-uniqueness of the found minimiser, although due to Corollary 2.4.1, such an issue went unnoticed in the literature since many simulation studies and real-world applications assume draws from a continuous distribution. Due to time and space constraints, we do not expand upon the proof of correctness (Appendix A.1, p.21 [10]) or the proof of Lemma 2.4.2 (which can be found as Lemma 16, p.20, [10]).

2.5 Sign Consistency of Non-Unique Lasso Solutions

The use of lasso regression as a form of model selection means that the estimated lasso solution $\hat{\boldsymbol{\beta}}$ needs to be *interpretable*. In particular, we are usually interested in the *signs* of the parameters in $\hat{\boldsymbol{\beta}}$, since this suggests the direction of correlation. As we have shown in the above section, uniqueness of lasso solutions can be characterised by a continuous distribution assumption on the columns of $\mathbf{X}$. However, even when such conditions do not hold, the following result gives us confidence in the interpretability of lasso solutions:

Proposition 2.5.1 (Sign Consistency)**.** Given two distinct solutions $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$, we must have

$$\hat{\beta}_i^{(1)} \cdot \hat{\beta}_i^{(2)} \geq 0 \quad \text{for all } i \in I. \tag{2.5.1}$$

Proof. The proof follows almost immediately from the sufficient and necessary KKT conditions. Indeed, suppose we have two distinct solutions $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$. By rearranging the KKT conditions (2.3.1), we get that the subgradient $\boldsymbol{\gamma}$ satisfies

$$\boldsymbol{\gamma} = \frac{1}{\lambda}\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{2.5.2}$$

for any lasso solution $\hat{\boldsymbol{\beta}}$. Of course, since the right-hand side of (2.5.2) is deterministic, the subgradient $\boldsymbol{\gamma}$ must be unique. In particular, $\boldsymbol{\gamma}$ is the subgradient for both $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$, and so we have

$$\hat{\beta}_i^{(1)} \cdot \hat{\beta}_i^{(2)} \geq 0 \quad \text{for all } i \in I.$$

We must note that this doesn't mean their support sets must match up. With this argument, we can guarantee at most that the estimated parameters either have the same sign or one is zero when compared with other distinct lasso solutions in the non-uniqueness case. □

2.6 Parameter Error

We now proceed to heavier analysis on the error rates associated with lasso solutions. In this first section, we will focus on bounding the L^2-parameter error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ asymptotically as $p, n \to \infty$. We will assume we are in a situation where the existence and uniqueness of lasso solutions is not of concern: for example, the entries of $\mathbf{X}$ are drawn from a continuous distribution (Corollary 2.4.1), so that the estimated solution $\hat{\boldsymbol{\beta}}$ is the unambiguously the unique solution of (2.1.2). In any case, the solution $\hat{\boldsymbol{\beta}}$ is assumed to minimise the lasso criterion (2.1.2) and also satisfy the lasso KKT conditions (2.3.1).

The proofs from the rest of the chapter (parameter error, prediction error and model selection consistency for the lasso) are sourced from [7]. As will be the basis for the rest of this thesis except for the orthogonal analyses sections, we assume, in general, that $p >> n$, that p and n are allowed to diverge as $n \to \infty$ and that the true parameter vector $\boldsymbol{\beta}^*$ has support $\text{supp}\,(\boldsymbol{\beta}^*) = \mathcal{S} \subset \{1, 2, \ldots, p\}$. We may also write $\lambda = \lambda$ to make the dependence of λ on n explicit, and $|\mathcal{S}| = k < p$.

Unfortunately, without some regularity assumptions on $\mathbf{X}$, the error analysis for $\hat{\boldsymbol{\beta}}$ would be impossible. We begin with some basic, important definitions.

Definition 2.6.1 (Strong Convexity)**.** Let $f : \mathbb{R}^p \to \mathbb{R}^p$ be a differentiable function. Then we say f is *strongly convex* with parameter $\gamma > 0$ at $\boldsymbol{\theta} \in \mathbb{R}^p$ if it satisfies

$$f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) \geq \nabla f(\boldsymbol{\theta})^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2. \tag{2.6.1}$$

Note that the case $\gamma = 0$ corresponds to ordinary convexity.

When $f \in C^2(\mathbb{R}^p)$, there is an alternative characterisation of strong convexity in terms of the Hessian of f:

Lemma 2.6.1 (Characterisation of Strong Convexity in C^2)**.** Suppose $f \in C^2(\mathbb{R}^p)$ is a twice continuously-differentiable function. Then f is strongly convex with parameter γ around $\boldsymbol{\beta}^* \in \mathbb{R}^p$ if and only if the minimum eigenvalue of the Hessian $\nabla^2 f(\boldsymbol{\beta})$ is at least γ for all vectors $\boldsymbol{\beta}$ in a neighbourhood of $\boldsymbol{\beta}^*$.

Example 2.6.1. A short calculation with

$$f_n(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

shows that

$$\nabla^2 f_n(\boldsymbol{\beta}) = \frac{\mathbf{X}^\top \mathbf{X}}{n},$$

and so f_n is strongly convex if and only if the eigenvalues of $\mathbf{X}^\top \mathbf{X}$ are bounded away from 0 uniformly, but this is not possible in the case $p > n$: indeed, $\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{p \times p}$ is of rank at most $\min\{n, p\} < p$, and so $\mathbf{X}^\top \mathbf{X}$ must have at least one zero eigenvalue. Fortunately, for our analysis, we do not require such a strong condition, only that strong convexity holds on some suitably defined set:

Definition 2.6.2 (Restricted Strong Convexity)**.** A function $f \in C^2(\mathbb{R}^p; \mathbb{R})$ satisfies *restricted strong convexity* at $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with respect to some set $\mathcal{C}$ if there is a constant $\gamma > 0$ such that

$$\frac{\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{\beta}) \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geq \gamma \quad \text{for all nonzero } \boldsymbol{v} \in \mathcal{C}, \tag{2.6.2}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ in a neighbourhood of $\boldsymbol{\beta}^*$.

In the linear regression case, where $f(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, this is equivalent to requiring

$$\frac{\frac{1}{n}\boldsymbol{v}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geq \gamma \quad \text{for all nonzero } \boldsymbol{v} \in \mathcal{C}, \tag{2.6.3}$$

for some set $\mathcal{C}$. Matrices $\mathbf{X}$ satisfying (2.6.3) are said to satisfy the γ-*restricted eigenvalue condition* (γ-RE).

What sets $\mathcal{C}$ could we use? It turns out that a useful set to define is the *cone set*:

Definition 2.6.3 (Cone Set)**.** Let $\mathcal{S} \subset \{1, 2, \ldots, p\}$ and $\alpha \geq 1$. The cone set $\mathcal{C}(\mathcal{S}; \alpha)$ is defined to be

$$\mathcal{C}(\mathcal{S}; \alpha) := \{\boldsymbol{v} \in \mathbb{R}^p \mid \|\boldsymbol{v}_{\mathcal{S}^c}\|_1 \leq \alpha \|\boldsymbol{v}_{\mathcal{S}}\|_1\}, \tag{2.6.4}$$

where $\alpha \geq 1$.

Essentially, the L^1-norm of the inactive coordinates of elements of $\mathcal{C}(\mathcal{S}; \alpha)$ are constrained by the L^1-norm of the coordinates on the active set $\mathcal{S}$. In fact, we will show that the solutions of the lasso naturally satisfy the cone inequality (2.6.4) for $\alpha = 3$ and $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*)$ in the process of proving Theorem 2.6.1, whenever the λ is large enough.

Theorem 2.6.1 (Parameter Error Bounds)**.** Suppose the design matrix $\mathbf{X}$ satisfies the restricted eigenvalue bound (2.6.3) with parameter $\gamma > 0$ and restricted set $\mathcal{C}(\mathcal{S}; 3)$. Then given a regularisation parameter $\lambda \geq 2\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty / n > 0$, any solution $\hat{\boldsymbol{\beta}}$ of the lasso (2.1.2) satisfies

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{3}{\gamma}\sqrt{k}\lambda. \tag{2.6.5}$$

Before we give the proof of the parameter error theorem, we prove a technical inequality under the hard-sparse assumption. As we will see, this inequality will be very useful in finding upper bounds on both the parameter error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ and the mean square error $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 / n$.

Lemma 2.6.2 (Basic Inequality)**.** Write $\boldsymbol{v} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Then

$$0 \leq \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} \leq \frac{\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty \|\boldsymbol{v}\|_1}{n} + \lambda(\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1). \tag{2.6.6}$$

and

$$0 \leq \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} \leq \left(\frac{\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n} - \lambda\right) \|\boldsymbol{v}\|_1 + 2\lambda \|\boldsymbol{\beta}^*\|_1. \tag{2.6.7}$$

Proof. First, define the function

$$G(\boldsymbol{v}) := \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}(\boldsymbol{\beta}^* + \boldsymbol{v})\|_2^2 + \lambda \|\boldsymbol{\beta}^* + \boldsymbol{v}\|_1. \tag{2.6.8}$$

By construction, $\boldsymbol{v} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ minimises G since $\hat{\boldsymbol{\beta}}$ is the minimiser of the lasso problem (2.1.2). We also note that $G(\boldsymbol{v}) \leq G(\mathbf{0})$. From this, we obtain

$$\frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1$$

Rearranging and making the substitution $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ again, we get the inequality

$$\frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} \leq \boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{v} + \lambda(\|\boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\beta}^* + \boldsymbol{v}\|_1). \tag{2.6.9}$$

Using the fact that $\boldsymbol{\beta}^*_{\mathcal{S}^c} = \mathbf{0}$, the additive decomposition $\|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + \|\boldsymbol{\beta}^*\|_{\mathcal{S}^c}$ and the reverse triangle inequality $\|\boldsymbol{\beta}^* + \boldsymbol{v}\|_1 \geq \|\boldsymbol{v}\|_1 - \|\boldsymbol{\beta}^*\|_1$, we get

$$\|\boldsymbol{\beta}^* + \boldsymbol{v}\|_1 = \|\boldsymbol{\beta}^*_{\mathcal{S}} + \boldsymbol{v}_{\mathcal{S}}\|_1 + \|\boldsymbol{v}_{\mathcal{S}^c}\|_1 \geq \|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}}\|_1 + \|\boldsymbol{v}_{\mathcal{S}^c}\|_1,$$

and substituting this into (2.6.9) gives

$$\begin{aligned}\frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} &\leq \frac{\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{v}}{n} + \lambda(\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1) \\ &\leq \frac{\|\mathbf{X}^\top\boldsymbol{\varepsilon}\|_\infty\|\boldsymbol{v}\|_1}{n} + \lambda(\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1).\end{aligned} \tag{2.6.10}$$

This is exactly the first inequality we wanted. As for the second inequality (2.6.7), we return to (2.6.9) and instead apply the second reverse triangle inequality $\|\boldsymbol{\beta}^* + \boldsymbol{v}\|_1 \geq \|\boldsymbol{\beta}^*\|_1 - \|\boldsymbol{v}\|_1$ along with Hölder's inequality to get

$$\frac{\|X\boldsymbol{v}\|_2^2}{2n} \leq \left(\|\mathbf{X}^\top\boldsymbol{\varepsilon}\|_\infty - \lambda\right)\|\boldsymbol{v}\|_1 + 2\lambda\|\boldsymbol{\beta}^*\|_1,$$

which is exactly (2.6.7). This ends the proof. □

We now turn to the proof of the parameter error theorem.

Proof. Suppose $\hat{\boldsymbol{\beta}}$ is the optimal solution found by the constrained optimisation problem and $\boldsymbol{\beta}^*$ is the fixed true regression vector.

Note the appearance of the term $\|\mathbf{X}^\top\boldsymbol{\varepsilon}\|_\infty/n$ in the right-hand side of (2.6.6). By the assumptions of the theorem, this norm is bounded from above by $\lambda/2$; the upper bound of Lemma 2.6.2 then becomes

$$\begin{aligned}\frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} &\leq \frac{\lambda}{2}\|\boldsymbol{v}\|_1 + \lambda(\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1) \\ &= \frac{\lambda}{2}(\|\boldsymbol{v}_{\mathcal{S}}\|_1 + \|\boldsymbol{v}_{\mathcal{S}^c}\|_1) + \lambda(\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1) \\ &= \frac{3\lambda}{2}\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \frac{\lambda}{2}\|\boldsymbol{v}_{\mathcal{S}^c}\|_1 \\ &\leq \frac{3\lambda}{2}\|\boldsymbol{v}_{\mathcal{S}}\|_1 \leq \frac{3}{2}\sqrt{k}\lambda\|\boldsymbol{v}\|_2,\end{aligned} \tag{2.6.11}$$

where in the last step we used Hölder's inequality to get $\|\boldsymbol{v}_{\mathcal{S}}\|_1 \leq \sqrt{k}\|\boldsymbol{v}_{\mathcal{S}}\|_2 \leq \sqrt{k}\|\boldsymbol{v}\|_2$.

It remains now to relate the term $\|\mathbf{X}\boldsymbol{v}\|_2^2$ with $\|\boldsymbol{v}\|_2^2$. To do so, we wish to apply the γ-RE condition, which requires us to know that $\boldsymbol{v}$ is contained within a cone set. This is not difficult to show. Indeed, from the second line of (2.6.11), we have

$$0 \leq \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} \leq \frac{\lambda}{2}(\|\boldsymbol{v}_{\mathcal{S}}\|_1 + \|\boldsymbol{v}_{\mathcal{S}^c}\|_1) + \lambda(\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1),$$

which, after rearranging, gives us $\|\boldsymbol{v}_{\mathcal{S}}\|_1 \leq 3\|\boldsymbol{v}_{\mathcal{S}^c}\|_1$, so by definition $\boldsymbol{v} \in \mathcal{C}(\mathcal{S}; 3)$. This means we can now apply the γ-RE condition $\|\mathbf{X}\boldsymbol{v}\|_2^2/n \geq \gamma\|\boldsymbol{v}\|_2^2$. Combining this with (2.6.11), we finally obtain

$$\begin{aligned} \gamma\frac{\|\boldsymbol{v}\|_2^2}{2} \leq \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} &\leq \frac{3}{2}\sqrt{k}\lambda\|\boldsymbol{v}\|_2, \\ \implies \|\boldsymbol{v}\|_2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &\leq \frac{3\sqrt{k}\lambda}{\gamma}, \end{aligned}$$

completing the proof. □

2.7 Prediction Error

In the previous section, we were interested in finding an upper bound on the parameter error of the lasso. In this section, we turn to finding upper bounds on the *prediction error* of the lasso. While the ability to approximate $\boldsymbol{\beta}^*$ well with $\hat{\boldsymbol{\beta}}$ guarantees a small prediction error as well, this is not a necessity for good prediction accuracy. In the theorem below, we give upper bounds for $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/n$ in cases where there is an upper bound on $\|\boldsymbol{\beta}^*\|_1$ (L^1-sparsity) and $\boldsymbol{\beta}^*$ is supported on $\mathcal{S}$ (hard-sparsity):

Theorem 2.7.1. Suppose $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top\boldsymbol{\varepsilon}\|_\infty$.

(i) (Prediction error for L^1-sparse $\boldsymbol{\beta}^*$): If $\|\boldsymbol{\beta}^*\|_1 \leq R_1$ for some $R_1 > 0$, then any optimal solution $\hat{\boldsymbol{\beta}}$ satisfies

$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leq 12R_1\lambda.$$

(ii) (Prediction error bound for k-sparse $\boldsymbol{\beta}^*$): If the optimal parameter vector $\boldsymbol{\beta}^*$ is supported on $\mathcal{S}$ with $|\mathcal{S}| = k < p$ and $\mathbf{X}$ satisfies the restricted eigenvalue bound for $\gamma > 0$ over $\mathcal{C}(S; 3)$, then any optimal lasso $\hat{\boldsymbol{\beta}}$ satisfies

$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leq \frac{9k\lambda^2}{\gamma}.$$

Note: in general, $\boldsymbol{\varepsilon}$ is a random vector, so that the lower bound on λ cannot be known precisely. However, it can be shown that $\lambda = c\sigma\sqrt{\frac{\log p}{n}}$ satisfies the lower bound $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top\boldsymbol{\varepsilon}\|_\infty$ with high probability. We formalise this fact towards the end of this chapter assuming that $\boldsymbol{\varepsilon}$ is Gaussian.

Proof. (i): Applying the assumption $\lambda \geq \frac{2\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n}$ to Lemma 2.6.2 (2.6.6), we obtain

$$\begin{aligned} 0 &\leq \left(\frac{\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n} - \lambda \right) \|\boldsymbol{v}\|_1 + 2\lambda \|\boldsymbol{\beta}^*\|_1 \\ &\leq \frac{1}{2}\lambda \left(-\|\boldsymbol{v}\|_1 + 4\|\boldsymbol{\beta}^*\|_1 \right) \end{aligned}$$

After rearranging, this gives us $\|\boldsymbol{v}\|_1 \leq 4\|\boldsymbol{\beta}^*\|_1 \leq 4R_1$. Using Lemma 2.6.2 (Equation (2.6.6)) again but with the reverse triangle inequality $\|\boldsymbol{v}_{\mathcal{S}}\|_1 - \|\boldsymbol{v}_{\mathcal{S}^c}\|_1 \leq \|\boldsymbol{v}\|_1$), we also obtain

$$\begin{aligned} \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} &\leq \left(\frac{\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n} + \lambda \right) \|\boldsymbol{v}\|_1 \\ &\leq 4R_1 \left(\frac{3\lambda}{2} \right) = 6\lambda R_1. \end{aligned}$$

This is exactly (i).

Proof of ii): Recall that we are assuming hard-sparsity on $\boldsymbol{\beta}^*$, supported on $\mathcal{S}$, and the restricted eigenvalue bound with $\gamma > 0$ on $\mathcal{S}(\mathcal{S}; 3)$. From (2.6.6) and applying Hölder's inequality, we have

$$\begin{aligned} \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{2n} &\leq \frac{3\lambda}{2}\|\boldsymbol{v}\|_1 \\ \implies \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{n} &\leq 3\lambda\sqrt{k}\|\boldsymbol{v}_{\mathcal{S}}\|_2. \end{aligned} \tag{2.7.1}$$

From the proof of Theorem 2.6.1, we know that $\boldsymbol{v} \in \mathcal{C}(\mathcal{S}; 3)$, and so the restricted eigenvalue bound (2.6.3) applies, giving us $\|\boldsymbol{v}\|_2^2 \leq \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{n\gamma}$. Taking square roots and substituting this into (2.7.1) gives us

$$\frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{n} \leq 3\lambda\sqrt{k}\frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n\gamma}},$$

and so

$$\frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{n} = \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leq \frac{9k\lambda^2}{\gamma},$$

as we wanted. □

In both Theorem 2.6.1 and Theorem 2.7.1, we required a lower bound on the regularisation parameter $\lambda \geq \frac{2\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n}$. Of course, since $\boldsymbol{\varepsilon}$ is a Gaussian random variable, this lower bound can only hold up to some probability. We show, in the next corollary, that there exists a so-called *universal threshold level* $\lambda = 2\sigma\sqrt{\frac{\log p}{n}}$ such that we can guarantee the lower bound condition with high probability while simultaneously giving us agreeable asymptotic bounds on both the parameter error and prediction error.

Corollary 2.7.1. If $\lambda = 2\sigma\sqrt{\frac{\log p}{n}}$ and σ, γ and k are constant in p and n, then

$$\mathbb{P}\left(\frac{\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n} \leq \lambda\right) \geq 1 - \frac{1}{p \log p} \tag{2.7.2}$$

and

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \leq\in \mathcal{O}\left(\frac{\log p}{n}\right). \tag{2.7.3}$$

Furthermore, under the L^1-sparse assumption $\|\boldsymbol{\beta}^*\|_1 \leq R_1$ for some $R_1 > 0$, we have the asymptotic bound

$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leq 24 R_1 \sigma \sqrt{\frac{\log p}{n}} \in \mathcal{O}\left(\frac{\log p}{n}\right). \tag{2.7.4}$$

and under the k-hard sparse assumption $|\mathcal{S}| = k$, we have

$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} = \frac{36\sigma^2}{\gamma}\frac{\log p}{n} \in \mathcal{O}\left(\frac{\log p}{n}\right). \tag{2.7.5}$$

Proof. Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, we have $\mathbf{X}^\top \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{X}^\top \mathbf{X})$. Since we normalised $\boldsymbol{v}_i^\top \boldsymbol{v}_i = n$ for all $i \in \{1, 2, \ldots, p\}$, we have $\boldsymbol{e}_i^\top \mathbf{X}^\top \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 n)$, and so $\frac{\boldsymbol{e}_i^\top \mathbf{X}^\top \boldsymbol{\varepsilon}}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1)$ (where $\boldsymbol{e}_i$ is the ith unit vector in $\mathbb{R}^n$). By the Gaussian tail bound (Proposition A.1.1), we have

$$\mathbb{P}\left(\frac{|\boldsymbol{e}_i^\top \mathbf{X}^\top \boldsymbol{\varepsilon}|}{\sigma\sqrt{n}} > z\right) \leq \frac{2\exp(-z^2/2)}{z},$$

and so by the union bound over p variables,

$$\mathbb{P}\left(\frac{\|\boldsymbol{e}_i^\top \mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{\sqrt{n}\sigma} > z\right) \leq \frac{2p\exp(-z^2/2)}{z}.$$

By rearranging and taking $z = \frac{\sqrt{n}\lambda}{\sigma}$ with $\lambda = 2\sigma\sqrt{\frac{\log p}{n}}$, we get

$$\begin{aligned}
\mathbb{P}\left(\frac{\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty}{n} > \lambda\right) &\leq \frac{2p\exp(-z^2/2)}{z} \\
&= \frac{2p\sigma}{\sqrt{n}\lambda}\exp\left(-\frac{n\lambda^2}{2\sigma^2}\right) \\
&= \frac{p}{\sqrt{\log p}} \cdot \frac{1}{p^2} \\
&= \frac{1}{p \log p}.
\end{aligned}$$

This gives us (2.7.2). As for the asymptotic bounds (2.7.3), (2.7.4) and (2.7.5), replace λ with $2\sigma\sqrt{\frac{\log p}{n}}$ in (2.6.5) and Theorem 2.7.1 (i) and (ii) respectively. If we assume σ, γ

and k are constant in n and $p = p_n$, then we get the $\mathcal{O}(\frac{\log p}{n})$ bounds as claimed. This completes the proof. □

2.8 Model Selection Consistency of the LASSO

In the previous two sections, we obtained asymptotic estimates for the parameter error and the prediction error. These error rates do not tell us about the *structure* of the model recovered by the lasso. We are now interested in the following (more important) question: under what conditions is the lasso model selection consistent when the true model $\boldsymbol{\beta}^*$ is sparse?

Formally, suppose that $\text{supp}\,(\boldsymbol{\beta}^*) = \mathcal{S}$ with $|\mathcal{S}| = k < p$. An estimate $\hat{\boldsymbol{\beta}}$ produced from a minimisation problem of the form (2.1.2) is said to be *model selection consistent* if $\text{supp}\,(\hat{\boldsymbol{\beta}}) = \mathcal{S}$. The first condition we will need to prove model selection consistency is the following:

Definition 2.8.1 (Strong Irrepresentability Condition)**.** Let $\mathcal{S} \subset \{1, 2, \ldots, p\}$. A matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ satisfies the *strong irrepresentability condition*[1] if there exists $\gamma > 0$ such that

$$\max_{j \in \mathcal{S}^c} \|(\mathbf{X}_\mathcal{S}\mathbf{X}_\mathcal{S})^{-1}\mathbf{X}_\mathcal{S}^\top \boldsymbol{v}_j\|_1 \leq 1 - \gamma \tag{2.8.1}$$

for columns $\boldsymbol{v}_j$ of $\mathbf{X}_{\mathcal{S}^c}$.

Essentially, the strong irrepresentability condition states that the "correlation" of the active variables of $\mathbf{X}$ with the inactive variables should not be too large. As we will see in Theorem 2.8.1, this additional requirement, on top of the γ-RE condition, is essential in proving model selection consistency. We can interpret the irrepresentability condition as follows: in order for the lasso to divine the true underlying model, the maximum linear correlation between the non-active variables and the active variables cannot be too high (indeed, the vector in (2.8.1) is the least squares predictor trained on the active variables applied to the inactive columns of $\mathbf{X}$). When explained this way, requiring (2.8.1) makes intuitive sense. However, we emphasise that the strong irrepresentability condition does not scale well as p increases. Indeed, of the numerical experiments conducted with Gaussian data and varying levels of sparsity in the true model (Figure 11.5, p.304 in [7]), none satisfied the strong irrepresentability condition with $p = 2000$ and $n = 1000$, yet we may generally expect larger p and n (and larger ratios p/n) in high dimensional applications. This is also not difficult to see from the form of (2.8.1): requiring a constant L^∞-bound over $|\mathcal{S}^c| = p - n >> 0$ values is difficult when large spurious correlations can be generated through stochastic fluctuations despite there being no true underlying relationship.

Finally, we shall also need a condition on the minimal size of the signal of interest:

Definition 2.8.2 (Minimal Signal Strength)**.** The true regression vector $\boldsymbol{\beta}^*$ satisfies the *minimal signal strength* condition if there exists some $c > 0$ such that

$$\min_{j \in \mathcal{S}} |\beta_j^*| > c. \tag{2.8.2}$$

[1] The strong irrepresentability condition, named as such in [11] is also known as the *mutual incoherence property*, as described on p.302 [7].

Theorem 2.8.1 (Lasso Model Selection Consistency)**.** Suppose the data matrix $\mathbf{X}$ satisfies the strong irrepresentability property and that the matrix $\mathbf{X}$ is normalised so that $\mathrm{diag}(\mathbf{X}^\top\mathbf{X}) = n\mathbf{1}_p$. Suppose further that there exists a lower bound on the minimum eigenvalue:

$$\lambda_{\min}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}{n}\right) \geq C_{\min}, \tag{2.8.3}$$

for some $C_{\min} > 0$, and that

$$\lambda \geq \frac{\sigma}{\gamma}\sqrt{\frac{\log p}{n}}.$$

Then with probability greater than $1 - c_1\exp(-c_2 n\lambda^2)$ for some constants $c_1, c_2 > 0$, the following results hold:

(i) Uniqueness: The optimal lasso solution $\hat{\boldsymbol{\beta}}$ of (2.1.2) is unique, even without the assumption that the entries of $\mathbf{X}$ are drawn from a continuous distribution.

(ii) No False Inclusion: The unique estimate $\hat{\boldsymbol{\beta}}$ has support $\mathrm{supp}\,(\hat{\boldsymbol{\beta}}) \subset \mathrm{supp}\,(\boldsymbol{\beta}^*)$.

(iii) Uniform Bound: The error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ satisfies the L^∞-bound

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_\infty \leq \lambda\left[\frac{\gamma}{\sqrt{C_{\min}}} + \left\|\frac{(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}}{n}\right\|_\infty\right] := B(\lambda, \gamma; \mathbf{X}). \tag{2.8.4}$$

(iv) Model Selection Consistency: Under the minimal signal strength condition with $c = B(\lambda, \sigma; \mathbf{X})$, the lasso estimate $\hat{\boldsymbol{\beta}}$ has support $\mathcal{S} = \mathrm{supp}\,(\boldsymbol{\beta}^*)$.

Proof. The proof relies on the "primal-dual-witness" (PDW) construction, which we describe now. Supposing we knew in advance the true support set $\mathcal{S}$,

(1) Set $\hat{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0$.

(2) Solve the *oracle sub-problem*

$$\hat{\boldsymbol{\beta}}_S \in \underset{\boldsymbol{\beta}_S \in \mathbb{R}^k}{\mathrm{argmin}}\left\{\frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}_S\boldsymbol{\beta}_S\|_2^2 + \lambda\|\boldsymbol{\beta}_S\|_1\right\} \tag{2.8.5}$$

and obtain the estimate $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$. Set $\hat{\boldsymbol{z}}_{\mathcal{S}} = \mathrm{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{S}})$, with the value set to anything in $[-1, 1]$ if a coordinate of $\hat{\boldsymbol{\beta}}$ is 0 (this is allowable under the subgradient calculus). Then $\hat{\boldsymbol{z}}_{\mathcal{S}}$ is a subgradient of $\partial\|\hat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1$, satisfying the $\mathcal{S}$-KKT conditions

$$\frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top(\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}}\hat{\boldsymbol{\beta}}_{\mathcal{S}}) + \lambda\hat{\boldsymbol{z}}_{\mathcal{S}} = 0. \tag{2.8.6}$$

(3) In general, the subgradient $\hat{\boldsymbol{z}}$ should satisfy the lasso KKT conditions

$$\frac{1}{n}\mathbf{X}^\top\left(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \lambda\hat{\boldsymbol{z}} = 0, \tag{2.8.7}$$

not just on $\mathcal{S}$. Solve for $\hat{\boldsymbol{z}}_{\mathcal{S}^c}$ in (2.8.7) and check whether $\|\hat{\boldsymbol{z}}_{\mathcal{S}^c}\|_\infty < 1$ holds (the *strict dual feasibility* condition).

(4) Show that the oracle estimator $\hat{\boldsymbol{\beta}}$ constructed in the presence of $\hat{\boldsymbol{z}}$ is the unique optimal solution of the lasso minimisation problem (2.1.2).

As we discussed in Section 2.3, Equation (2.8.5) always has a solution, and by the subgradient calculus, (2.8.6) is satisfied by $\hat{\boldsymbol{z}}_{\mathcal{S}}$. Hence, to complete the primal-dual-witness construction and the proof of Theorem 2.8.1, we prove that we can solve for $\hat{\boldsymbol{z}}_{\mathcal{S}^c}$ in (2.8.7) such that $\|\hat{\boldsymbol{z}}_{\mathcal{S}^c}\|_\infty < 1$ (Lemma 2.8.1), that the solution $\hat{\boldsymbol{\beta}}$ is unique and has the no false inclusion property (Lemma 2.8.2), before finally proving the uniform bound (2.8.4) and model selection consistency in Lemma 2.8.3.

Lemma 2.8.1 (Strict Dual Feasibility)**.** The primal-dual-witness method constructs a subgradient vector $\hat{\boldsymbol{z}}$ such that $\|\hat{\boldsymbol{z}}_{\mathcal{S}}\|_\infty < 1$.

Proof. By writing $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and $\mathbf{X} = [\mathbf{X}_{\mathcal{S}}\, \mathbf{X}_{\mathcal{S}^c}]$ and

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{X}_{\mathcal{S}}^\top \\ \mathbf{X}_{\mathcal{S}^c}^\top \end{bmatrix} [\mathbf{X}_{\mathcal{S}} \mathbf{X}_{\mathcal{S}^c}] = \begin{bmatrix} \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} & \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}^c} \\ \mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}} & \mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}^c}, \end{bmatrix}$$

we can rewrite the zero subgradient condition (2.8.6) to get

$$\begin{aligned} \frac{1}{n}\mathbf{X}^\top(\mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}) + \lambda \hat{\boldsymbol{z}} = 0, \\ \frac{1}{n}\begin{bmatrix} \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} & \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}^c} \\ \mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}} & \mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}^c} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}} \\ 0 \end{bmatrix} + \frac{1}{n}\begin{bmatrix} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon} \\ \mathbf{X}_{\mathcal{S}^c}^\top \boldsymbol{\varepsilon} \end{bmatrix} + \lambda \begin{bmatrix} \hat{\boldsymbol{z}}_{\mathcal{S}} \\ \hat{\boldsymbol{z}}_{\mathcal{S}^c} \end{bmatrix} = 0, \end{aligned} \tag{2.8.8}$$

which gives us two linear equations

$$\begin{aligned} \frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}}) + \frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon} + \lambda \hat{\boldsymbol{z}}_{\mathcal{S}} = 0, \\ \frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}}) + \frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^\top \boldsymbol{\varepsilon} + \lambda \hat{\boldsymbol{z}}_{\mathcal{S}^c} = 0. \end{aligned} \tag{2.8.9}$$

From the first equation of (2.8.9), we can rearrange to get

$$\begin{aligned} \frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}}) = -\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon} + \lambda \hat{\boldsymbol{z}}_{\mathcal{S}}\right), \text{ so} \\ \boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}} = -\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} + \lambda \hat{\boldsymbol{z}}_{\mathcal{S}}\right), \end{aligned} \tag{2.8.10}$$

where we have used the assumed invertibility of $\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}$ (this follows from the eigenvalue condition (2.8.3)). We can substitute this expression for the error $\boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}}$ into the second equation of (2.8.9) to get

$$\frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}}\left[-\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} + \lambda \hat{\boldsymbol{z}}_{\mathcal{S}}\right)\right] + \frac{1}{n}\mathbf{X}_{\mathcal{S}^c}^\top \boldsymbol{\varepsilon} + \lambda \hat{\boldsymbol{z}}_{\mathcal{S}^c} = 0,$$

which, after some expansion and factorisation, simplifies to

$$\frac{\mathbf{X}_{\mathcal{S}^c}^\top}{n}\big(\boldsymbol{I} - \mathbf{X}_{\mathcal{S}^c}^\top\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top\big)\boldsymbol{\varepsilon} + \lambda\hat{\boldsymbol{z}}_{\mathcal{S}^c} - \lambda\mathbf{X}_{\mathcal{S}^c}^\top\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}^\top}\mathbf{X}_{\mathcal{S}})^{-1}\hat{\boldsymbol{z}}_{\mathcal{S}} = 0,$$

and rearranging for $\hat{\boldsymbol{z}}_{\mathcal{S}^c}$ gives

$$\hat{\boldsymbol{z}}_{\mathcal{S}^c} = \underbrace{\mathbf{X}_{\mathcal{S}^c}^\top\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}}^*)}_{\boldsymbol{\mu}} - \underbrace{\mathbf{X}_{\mathcal{S}^c}^\top\big(\boldsymbol{I} - \mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top\big)\frac{\boldsymbol{\varepsilon}}{\lambda n}}_{\boldsymbol{V}_{\mathcal{S}^c}}. \tag{2.8.11}$$

Note that we have replaced $\boldsymbol{z}_{\mathcal{S}}$ by $\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}}^*)$, since $\boldsymbol{z}_{\mathcal{S}} \in \partial\|\boldsymbol{\beta}^*\|_1$. Using the triangle inequality with $\|\cdot\|_\infty$, we get that

$$\|\hat{\boldsymbol{z}}_{\mathcal{S}^c}\|_\infty \leq \|\boldsymbol{\mu}\|_\infty + \|\boldsymbol{V}_{\mathcal{S}^c}\|_\infty. \tag{2.8.12}$$

We now bound both $\boldsymbol{\mu}$ and $\boldsymbol{V}_{\mathcal{S}^c}$ on the right-hand side of (2.8.12). By the strong irrepresentability condition (2.8.1), we can bound $\|\boldsymbol{\mu}\|_\infty$: this is because

$$\begin{aligned}
\|\boldsymbol{\mu}\|_\infty &= \|\mathbf{X}_{\mathcal{S}^c}^\top\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}}^*)\|_\infty \\
&= \|\mathbf{X}_{\mathcal{S}^c}^\top\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{1}_k\|_\infty \\
&= \|\mathbf{1}_k^\top(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}^c}\|_\infty \\
&= \max_{j\in\mathcal{S}^c}\|(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top\boldsymbol{v}_j\|_1, \\
&\leq 1-\gamma
\end{aligned}$$

for some $\gamma > 0$. We now need to bound the random quantity $\|\boldsymbol{V}_{\mathcal{S}^c}\|_\infty$. Let us write $\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top$ as $\boldsymbol{P}$, the projection onto the column space of $\mathbf{X}_{\mathcal{S}}$, satisfying $\boldsymbol{P}^2 = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{P}$. Then $\boldsymbol{I} - \boldsymbol{P}$ is the projection orthogonal to $\operatorname{Col}(\mathbf{X}_{\mathcal{S}})$, and satisfies $(\boldsymbol{I} - \boldsymbol{P})^2 = \boldsymbol{I} - \boldsymbol{P}$. We consider $\boldsymbol{V}_{\mathcal{S}^c}$ coordinate-wise: for $j \in \mathcal{S}^c$, define

$$V_j := \boldsymbol{v}_j^\top\left[\boldsymbol{I} - \mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top\right]\frac{\boldsymbol{\varepsilon}}{\lambda n} = \boldsymbol{v}_j(\boldsymbol{I} - \boldsymbol{P})\frac{\boldsymbol{\varepsilon}}{\lambda n}.$$

Recalling $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$, we know that V_j is a zero-mean Gaussian random variable with variance

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(V_j, V_j) &= \mathbb{C}\mathrm{ov}\Big(\boldsymbol{v}_j^\top(\boldsymbol{I} - \boldsymbol{P})\frac{\boldsymbol{\varepsilon}}{\lambda n}, \boldsymbol{v}_j^\top(\boldsymbol{I} - \boldsymbol{P})\frac{\boldsymbol{\varepsilon}}{\lambda n}\Big) \\
&= \frac{1}{\lambda n}\boldsymbol{v}_j^\top(\boldsymbol{I} - \boldsymbol{P})\,\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{v}_j\frac{1}{\lambda n} \\
&= \frac{\sigma^2}{\lambda^2 n^2}\boldsymbol{v}_j^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{v}_j \\
&= \frac{\sigma^2}{\lambda^2 n^2}\boldsymbol{v}_j^\top\boldsymbol{v}_j \\
&= \frac{\sigma^2}{\lambda^2 n},
\end{aligned}$$

where in the second-last line we used that $\boldsymbol{I} - \boldsymbol{P}$ is the orthogonal projection onto $\boldsymbol{X}_{\mathcal{S}}$ but $\boldsymbol{v}_j$ is in the column space of $\boldsymbol{X}_{\mathcal{S}^c}$, and in the last line we used the normalisation of $\mathbf{X}^\top\mathbf{X}$. Using the Gaussian tail bound Proposition A.1.1, we get

$$\mathbb{P}\left[V_j \geq \frac{\gamma}{2}\right] \leq \exp\left(-\frac{\lambda^2 n(\gamma/2)^2}{2\sigma^2}\right),$$

so that

$$\begin{aligned}\mathbb{P}\left[\|\boldsymbol{V}_{\mathcal{S}^c}\|_\infty \geq \gamma/2\right] &= \mathbb{P}\left[|V_j| \geq \gamma/2 \,\forall\, j \in \mathcal{S}^c\right] \\ &\leq 2(p-k)\exp\left(-\frac{\lambda^2 n(\gamma/2)^2}{2\sigma^2}\right),\end{aligned}$$

which decays to zero whenever the ratio $\lambda^2 n \to \infty$ as $n \to \infty$. This shows that $\|\hat{\boldsymbol{z}}_{\mathcal{S}^c}\|_\infty < 1$, so that the proof of strict dual feasibility is complete. □

Lemma 2.8.2 (PDW Generates Unique Solution)**.** Suppose $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a minimiser of the lasso problem with subgradient $\hat{\boldsymbol{z}}$ such that $\hat{\boldsymbol{z}}_{\mathcal{S}} \in \partial\|\hat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1$ and $\|\hat{\boldsymbol{z}}_{\mathcal{S}^c}\|_\infty < 1$. Then $\hat{\boldsymbol{\beta}}$ is the *unique solution* of the lasso problem (2.1.2) such that $\operatorname{supp}(\boldsymbol{\beta}) \subset \operatorname{supp}(\boldsymbol{\beta}^*)$.

Proof. For notation, let us write $F : \mathbb{R}^p \to \mathbb{R}$, $\boldsymbol{\beta} \mapsto \|\boldsymbol{\beta}\|_1$. Furthermore, note that $\hat{\boldsymbol{z}}^\top\hat{\boldsymbol{\beta}} = \|\hat{\boldsymbol{\beta}}\|_1$ since $\hat{\boldsymbol{\beta}}$ is zero on $\mathcal{S}^c$ and $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ on $\mathcal{S}$, and $\hat{\boldsymbol{z}} = \operatorname{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{S}})$ on $\mathcal{S}$.

We first show uniqueness. Suppose $\boldsymbol{\beta} \in \mathbb{R}^p$ is another optimal solution of the lasso minimisation problem. Then since $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ are both minimisers and hence satisfy the KKT conditions and recalling that $\hat{\boldsymbol{z}}$ is a subgradient for $\hat{\boldsymbol{\beta}}$, we get

$$F(\boldsymbol{\beta}) + \lambda\hat{\boldsymbol{z}}^\top\boldsymbol{\beta} = F(\boldsymbol{\beta}) + \lambda\hat{\boldsymbol{z}}^\top\hat{\boldsymbol{\beta}} = F(\boldsymbol{\beta}) + \lambda\|\hat{\boldsymbol{\beta}}\|_1.$$

Subtracting $\lambda\hat{\boldsymbol{z}}^\top\boldsymbol{\beta}$ from both sides, we get

$$F(\hat{\boldsymbol{\beta}}) - \lambda\hat{\boldsymbol{z}}^\top\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) = F(\boldsymbol{\beta}) + \lambda(\|\boldsymbol{\beta}\|_1 - \hat{\boldsymbol{z}}^\top\boldsymbol{\beta}).$$

By the zero-subgradient condition (2.8.6), we have

$$\lambda\hat{\boldsymbol{z}} = -\frac{1}{n}\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}^\top\hat{\boldsymbol{\beta}}) = -\nabla F(\hat{\boldsymbol{\beta}}),$$

so that

$$F(\hat{\boldsymbol{\beta}}) - F(\boldsymbol{\beta}) + \nabla F(\hat{\boldsymbol{\beta}})^\top\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) = \lambda(\|\boldsymbol{\beta}\|_1 - \hat{\boldsymbol{z}}^\top\boldsymbol{\beta}).$$

By an equivalent characterisation of convexity for C^1-functions, the left-hand side of the above equation is negative, forcing

$$\|\boldsymbol{\beta}\|_1 \leq \hat{\boldsymbol{z}}^\top\boldsymbol{\beta} \leq \|\hat{\boldsymbol{z}}\|_\infty\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}\|_1,$$

which is true if and only if $\hat{\boldsymbol{z}}^\top \boldsymbol{\beta} = \|\boldsymbol{\beta}\|_1$, implying that $\beta_j = 0$ for all $j \in \mathcal{S}^c$. Indeed, $\hat{\boldsymbol{z}}_{\mathcal{S}}$ is a vector of ± 1's, and the elements of $\hat{\boldsymbol{z}}_{\mathcal{S}^c}$ can be chosen to be non-zero, so that if $\beta_j \neq 0$ for $j \in \mathcal{S}^c$, $\hat{\boldsymbol{z}}^\top \boldsymbol{\beta}$ cannot be equal to $\|\boldsymbol{\beta}\|_1$, a contradiction. This shows that every possible optimal solution of the oracle subproblem (2.8.5) is supported on $\mathcal{S}$. Uniqueness of the minimiser $\hat{\boldsymbol{\beta}}$ then follows from the lower eigenvalue bound (2.8.3), since this implies the strict convexity of the minimisation problem (2.8.5) (indeed, to see this, just differentiate the objective function: the L^1-norm $\|\boldsymbol{\beta}_{\mathcal{S}}\|_1$ is differentiable since all parameter values are non-zero).

We conclude that $\hat{\boldsymbol{\beta}}$ is the unique optimal solution of the lasso problem with subgradient vector satisfying $\|\hat{\boldsymbol{z}}_{\mathcal{S}^c}\|_\infty < 1$ and $\operatorname{supp}(\boldsymbol{\beta}) \subset \operatorname{supp}(\boldsymbol{\beta}^*)$. This completes the proof.

□

Lemma 2.8.3 (L^∞-bounds)**.** The error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ satisfies the L^∞-bound

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_\infty \leq \lambda \left[\frac{4\sigma}{\sqrt{C_{\min}}} + \left\| \frac{(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1}}{n} \right\|_\infty \right] := B(\lambda, \sigma; \mathbf{X}),$$

and if $\boldsymbol{\beta}^*$ satisfies the minimal strength condition with $c = B(\lambda, \sigma; \mathbf{X})$, then $\operatorname{supp}(\hat{\boldsymbol{\beta}}) = \operatorname{supp}(\boldsymbol{\beta}^*)$.

Proof. Writing $\boldsymbol{\nu}_{\mathcal{S}} = \boldsymbol{\beta}^*_{\mathcal{S}} - \hat{\boldsymbol{\beta}}_{\mathcal{S}}$, we use the triangle inequality for $\|\cdot\|_\infty$ on (2.8.10) to get

$$\|\boldsymbol{\nu}_{\mathcal{S}}\|_\infty \leq \underbrace{\left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} \right\|_\infty}_{Z} + \underbrace{\lambda \left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \right\|_\infty}_{\eta}. \tag{2.8.13}$$

The term η is deterministic (and is already present in the inequality (2.8.4), so there is nothing to do), so it suffices to bound Z. Writing

$$Z_i := \boldsymbol{e}_i^\top \left(\frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} \right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} \quad \text{for } i = 1, \ldots, k,$$

where $\boldsymbol{e}_i$ are the standard basis vectors of $\mathbb{R}^k$, we see that Z_i are zero-mean Gaussian random variables with variance

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(Z_i, Z_i) &= \mathbb{C}\mathrm{ov}\left(\boldsymbol{e}_i^\top \left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}\right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n}, \boldsymbol{e}_i^\top \left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}\right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n}\right) \\
&= \boldsymbol{e}_i^\top \left(\frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}\right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top}{n} \frac{\mathbf{X}_{\mathcal{S}}}{n} \left(\frac{\mathbf{X}_{\mathcal{S}} \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \boldsymbol{e}_i \sigma^2 \\
&= \frac{1}{n} \boldsymbol{e}_i^\top \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \boldsymbol{e}_i \sigma^2 \\
&\leq \frac{\sigma^2}{n} \lambda_{\max}\left(\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}\right) \\
&= \frac{\sigma^2}{n} \left[\lambda_{\min}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)\right]^{-1} \\
&\leq \frac{\sigma^2}{n C_{\min}},
\end{aligned}$$

so that, by the union bound over k variables,

$$\mathbb{P}\left[Z \geq t\right] \leq 2k \exp\left(-\frac{t^2 C_{\min} n}{2\sigma^2}\right) = 2 \exp\left(-\frac{t^2 C_{\min} n}{2\sigma^2} + \log k\right).$$

Now, recall that $\lambda \geq \frac{\sigma}{\gamma}\sqrt{\frac{\log p}{n}}$ and pick $t = \frac{\gamma\lambda}{\sqrt{C_{\min}}}$. With this choice[2] of t and lower bound for λ, we have

$$\begin{aligned}
-\frac{t^2 C_{\min} n}{2\sigma^2} + \log k &= -\frac{C_{\min} n}{2\sigma^2} \frac{\gamma^2 \lambda^2}{C_{\min}} + \log k \\
&= -\frac{8\gamma^2 n \lambda^2}{\sigma^2} + \log k \\
&\leq -8 \log p + \log k \\
&< 0
\end{aligned}$$

and so, following on from (2.8.13), we conclude that

$$\|\boldsymbol{\beta}_{\mathcal{S}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{S}}\|_\infty \leq \frac{\gamma\lambda}{\sqrt{C_{\min}}} + \lambda \left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \right\|_\infty := B(\lambda, \gamma; \mathbf{X}). \tag{2.8.14}$$

with probability $1 - 2k\exp(-C_1\lambda^2)$ for a constant $C_1 > 0$.

It remains now to show model selection consistency under the minimal signal strength condition. We already know that the lasso solution has the no false inclusion property from (ii), so that $\mathrm{supp}\,(\hat{\boldsymbol{\beta}}) \subset \mathrm{supp}\,(\boldsymbol{\beta}^*)$. If $\min_{j \in \mathcal{S}} |\beta_j^*| > B(\lambda, \gamma; \mathbf{X})$, then by (2.8.14), $\hat{\beta}_j \neq 0$ for all $j \in \mathcal{S}$, since otherwise the L^∞-norm would exceed the upper bound $B(\lambda, \gamma; \mathbf{X})$, and so $\mathrm{supp}\,(\hat{\boldsymbol{\beta}})$ is exactly $\mathcal{S} = \mathrm{supp}\,(\boldsymbol{\beta}^*)$. □

[2]The value of t chosen in p.309 [7] does not give the correct bounds. We have provided a corrected version here.

Combining the strict dual feasibility result in Lemma 2.8.1, the unique solution result in Lemma 2.8.2 and the L^∞-bound in Lemma 2.8.3 finishes the proof of model selection consistency of the lasso. □

Chapter 3

Concave Penalisers and the Root-Log Regulariser

3.1 General Concave Penalised Least Squares Estimators

One problem with the lasso is the inflexible penalisation that creates overly large bias when producing estimates parameter estimates. In Chapter 2, we saw that the model selection consistency of the lasso required the strong irrepresentability condition, which is an overly restrictive condition that does not hold with high probability even under reasonably sized $p >> n$. As we discussed in Chapter 1, we can ameliorate this issue by replacing the inflexible L^1-penalty with a concave function that penalises small parameter values highly, while conversely reducing the relative penalisation for large parameter values. Formally, we are interested in minimisation problems of the form

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \rho_\lambda(\boldsymbol{\beta}) \right\}, \tag{3.1.1}$$

where $\rho_\lambda(\cdot) = \lambda^2\rho(\cdot/\lambda)$ for some (continuous) function ρ satisfying the following conditions:

(i) Zero baseline penalty: $\rho(0) = 0$.
(ii) Symmetry: $\rho(t) = \rho(-t)$.
(iii) Sparsity: $\lim_{t\to 0+} \frac{\rho(t)}{t} > 0$.
(iv) Monotonicity: $\rho(t') > \rho(t)$ if $t' > t$.
(v) One-sided differentiability: the limit

$$\rho'(t\pm) := \lim_{\varepsilon\to 0} \frac{\rho_\lambda(t \pm \varepsilon) - \rho_\lambda(t)}{\boldsymbol{\varepsilon}}$$

exists for all t.
(vi) Concavity: $\rho'(t)$ is decreasing in t for $t \in (0, \infty)$.
(vii) Approximate unbiasedness: $\lim_{t\to\infty} \rho'(t) = 0$.

See p.189 [5] for the original reference. Functions that satisfy all of the above properties are called (folded) concave penalties. We can regularise a concave penalty by defining $\rho_\lambda(t) := \lambda^2\rho(t/\lambda)$. Owing to Property (iii), the derivative $\rho'(0)$ does not exist in the conventional sense. However, notation-wise, we write $\rho'(0)$ to denote any value between $[\rho(0-), \rho(0+)]$. Similar notation is used in, for example, p.186 [5].

For our work, the motivating prototype of a folded concave penalty is the *root-log penalty*, defined in the following way:

$$\rho(t) = \frac{\sqrt{t^4 + 4t^2} - t^2}{4} + \frac{1}{2}\ln\left(1 + \frac{\sqrt{t^4 + 4t^2} + t^2}{2}\right). \tag{3.1.2}$$

Asymptotically, $\rho(\cdot)$ behaves like $|\cdot|$ when $t \to 0$ and like $\log(\cdot)$ when $t \to \infty$.

In the following sections, we first prove our results for a general concave regulariser ρ_λ satisfying the above conditions before explicitly moving to the situation where ρ_λ is the root-log regulariser (3.1.2) in Sections 3.1.2 and onwards. As per usual, we assume that the true regression parameter vector $\boldsymbol{\beta}^*$ satisfies the hard sparse assumption $\text{supp}\,(\boldsymbol{\beta}^*) = \mathcal{S}$ with $|\mathcal{S}| < p$. Aside from the orthogonal analysis, we place no restrictions on p or n, allowing the general case $p = p_n >> n$ with both allowed to diverge as $n \to \infty$.

3.1.1 Orthogonal Analysis

As we did for the lasso, we unveil the behaviour of the general concave minimisation problem (3.1.1) by studying the behaviour of solutions in the one-dimensional case (this corresponds to minimising (3.1.1) coordinate-wise when $\mathbf{X}$ is orthogonal and $p < n$). This result appears as Theorem 1, p.943 in [1].

Theorem 3.1.1. Suppose ρ_λ is a folded concave regularised penalty function such that $\beta + \rho'_\lambda(\beta)$ has (at most) a single minimum on $(0, \infty)$. Fix a data point $z \in \mathbb{R}$ and consider the one-dimensional minimisation problem

$$\min_{\beta\in\mathbb{R}}\{\ell(\beta)\} = \min_{\boldsymbol{\beta}\in\mathbb{R}}\left\{\frac{(z-\beta)^2}{2} + \rho_\lambda(|\beta|)\right\}, \tag{3.1.3}$$

Then the following properties hold:

(i) Solutions to (3.1.3) exist and are unique. The solution $\hat{\beta}(z)$ (thought of as a function of the data z) is antisymmetric, satisfying $\hat{\beta}(-z) = -\hat{\beta}(z)$.

(ii) Let $\rho_0 = \inf_{\beta\geq 0}\{\beta + \rho'_\lambda(\beta)\}$. Then

$$\hat{\beta}(z) = \begin{cases} 0 & \text{if } |z| \leq \rho_0, \\ z - \text{sign}(z)\rho'_\lambda(|\hat{\beta}(z)|) & \text{if } |z| > \rho_0. \end{cases} \tag{3.1.4}$$

Furthermore, we have $|\hat{\beta}(z)| \leq |z|$.

(iii) If $\rho'_\lambda(\cdot)$ is a non-increasing function on $(0, \infty)$, then for $|z| > \rho_0$,

$$|z| - \rho_0 \leq |\hat{\beta}(z)| \leq |z| - \rho'_\lambda(|z|). \tag{3.1.5}$$

We first comment on the meaning of these results. (i) tells us that concave PLSE (at least in the one-dimensional case) do exist. (ii) tells us that the solution is soft-thresholded, in the sense that the solution sets "small enough" parameters to zero, with the soft-threshold level given by ρ_0. In (iii), we see that the size of this reduction is bounded between the threshold level ρ_0 and the gradient $\rho'_\lambda(|z|)$. It is not difficult to

show that $\rho_0 = \lambda$ when $\rho_\lambda(t) = \lambda|t|$; since the lasso is (also) concave, parameter estimates below λ are set to zero, which concurs with our results in Proposition 2.2.1.

Proof of (i) and (ii): First, consider existence and uniqueness in the case $z = 0$. Clearly, (3.1.3) has the (unique) solution $\hat{\beta}(z) = 0$, and so we can assume $z > 0$. If $z > 0$, then the loss function satisfies $\ell(-\beta) > \ell(\beta)$ (since ρ_λ is symmetric, simply note that $z - \beta$ would be smaller than $z + \beta$). This forces $\hat{\beta}(z) > 0$, since positive values of β always improve the loss function over negative values. Now, for $\beta > 0$, we can differentiate (3.1.3) to obtain

$$\ell'(\beta) = \beta - z + \rho'_\lambda(\beta). \tag{3.1.6}$$

Recall that $\rho_0 := \inf_{\beta \geq 0}\{\beta + \rho'_\lambda(\beta)\}$. We consider cases where the data point z is compared to the size of ρ_0.

Case 1: Suppose $z < \rho_0$. By (3.1.6), this implies $\ell'(\beta) > 0$, so $\ell(\beta)$ is strictly increasing on $(0, \infty)$. Hence, the global minimiser must be obtained at $\hat{\beta}(z) = 0$. This minimiser must be unique for ℓ since, if $\ell'(\beta)$ is strictly increasing, then $\ell'(\beta)$ has at most one zero.

Case 2: Suppose $z = \rho_0$. Since $\rho'_\lambda(\beta)$ is non-decreasing, ρ_0 can only be attained at one value β_u. Hence, $\ell'(\beta)$ is zero only at β_u and is positive everywhere else, this is the unique global minimiser we wanted.

Case 3: Suppose $z > \rho_0$. We now use the assumption that $-\beta - \rho'_\lambda(\beta)$ is at most unimodal: this implies that $\beta + \rho'_\lambda(\beta)$ has at most one minimum, so that $\ell'(\theta)$ has (at most) two possible roots $\alpha_1 < \alpha_2$ on $(0, \infty)$. Due to the concavity of $\ell'(\beta)$, we have $\ell''(\alpha_1) < 0$ and $\ell''(\alpha_2) > 0$, so that $\alpha_2 = \hat{\beta}$ is the (unique) minimiser of $\ell(\beta)$.

Finally, we note that setting $\ell'(\beta) = 0$ in (3.1.6), the minimiser $\hat{\beta}(z)$ must take the form

$$\hat{\beta}(z) = \begin{cases} 0 & \text{if } |z| \leq \rho_0, \\ z - \operatorname{sign}(z)\rho'_\lambda(|\hat{\beta}(z)|) & \text{if } |z| > \rho_0. \end{cases} \tag{3.1.7}$$

Proof of (iii): We now suppose ρ'_λ is non-increasing and, without loss of generality, that $z > \rho_0$ with $z \in [0, \infty)$. Let $\beta_0 := \operatorname{argmin}_{\beta \in [0,\infty)}\{\beta + \rho'_\lambda(\beta)\}$. Then we must have $\hat{\beta}(z) > \beta_0$. Indeed, as in the proof of Case 3, we know that in such a situation, $\hat{\beta}$ is the largest root of $\ell'(\beta)$, with the minimiser of $\ell'(\beta)$ contained in a valley between the two roots of $\ell'(\beta)$. From here, we get the inequality

$$\rho'_\lambda(\hat{\beta}(z)) \leq \rho'_\lambda(\beta) \leq \beta_0 + \rho'_\lambda(\beta_0) = \rho_0, \tag{3.1.8}$$

with the first inequality of (3.1.8) following from the non-increasing property of ρ'_λ, the second inequality since $\beta_0 > 0$, and the last equality by definition of ρ_0 and β_0. From the explicit solution (3.1.7), we automatically get

$$\hat{\beta}(z) = z - \rho'_\lambda(\hat{\beta}(z)) \leq z,$$

so that $\rho'_\lambda(\hat{\beta}(z)) \geq \rho'_\lambda(z)$, again by the non-increasing property of ρ'_λ. Combining this inequality with (3.1.8) and the equality $\hat{\beta}(z) = z - \rho'_\lambda(\hat{(}\beta(z))$, we get

$$z - \rho_0 \leq \hat{\beta}(z) \leq z - \rho'_\lambda(z),$$

as we wanted. □

3.1.2 Properties of the Root-Log Regulariser

With an understanding of the general properties of concave penalised estimators in the one-dimensional case, we turn our focus to specific properties of the root-log regulariser. We begin by proving an explicit formula for the one-dimensional minimiser of the root-log regularised problem:

Proposition 3.1.1 (Root-Log Orthogonal Analysis)**.** Suppose ρ_λ is the root-log regulariser as given in (3.1.2). The solution to the one-dimensional minimisation problem

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}} \left\{ \frac{(\beta - z)^2}{2} + \rho_\lambda(\beta) \right\} \tag{3.1.9}$$

is given by

$$\hat{\beta}(z) = z \left(1 - \frac{\lambda^2}{z^2} \right)_+, \tag{3.1.10}$$

which is a continuous function of the data z. When the one-dimensional minimisation problem results from the coordinate-wise minimisation given an orthogonal data matrix $\mathbf{X}$, the solution $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\beta}_i = \hat{\beta}_i^{\mathrm{LS}} \left(1 - \frac{\lambda^2}{\hat{\beta}_i^{\mathrm{LS}^2}} \right)_+. \tag{3.1.11}$$

That is, the root-log orthogonal estimator is the least squares estimator, multiplied by a coefficient that decreases with the inverse square of the least squares estimator.

Proof. We first consider the function

$$f(t) = t + \rho_\lambda'(t) = t + \lambda^2 \frac{d}{dt} \rho(t/\lambda).$$

Through some algebra, we can simplify f to

$$f(t) = \frac{t + \operatorname{sign}(t)\sqrt{4\lambda^2 + t^2}}{2} = \frac{t + \sqrt{4\lambda^2 + t^2}}{2} \quad \text{when } t > 0.$$

Noting that f is increasing in t, and hence has (at most) one minimum on $[0, \infty)$. Furthermore, the infimum $\rho_0 := \inf_{t \geq 0} f(t)$ is given by $\lim_{t \to 0} f(t) = \lambda$. Hence, the threshold in Theorem 3.1.1 is exactly λ, and below this threshold, the optimal solution is $\hat{\beta} = 0$. Now, by differentiating the one-dimensional minimisation problem (3.1.3) and

assuming that $t > \rho_0$, we get

$$\begin{aligned}
\frac{\hat{\beta}}{2} - z + \frac{\operatorname{sign}(\hat{\beta})\sqrt{\hat{\beta}^2 + 4\lambda}}{2} &= 0 \\
\hat{\beta} + \operatorname{sign}(\hat{\beta})\sqrt{\hat{\beta}^2 + 4\lambda^2} &= 2z \\
-2z + \hat{\beta} &= -\operatorname{sign}(\hat{\beta})\sqrt{\hat{\beta}^2 + 4\lambda^2} \\
-4z\hat{\beta} + 4z^2 &= 4\lambda^2 \\
\hat{\beta} &= z\left(1 - \frac{\lambda^2}{z^2}\right).
\end{aligned}$$

This is a continuous function of the data z. As for the interpretation of z as the least squares solution $\hat{\beta}_i^{\mathrm{LS}}$, under the orthogonal design matrix assumption $\mathbf{X}^\top\mathbf{X} = n\boldsymbol{I}_n$, we have

$$\begin{aligned}
\frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{p}\rho_\lambda(\beta_i) &= \frac{1}{2n}\boldsymbol{Y}^\top\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^\top\mathbf{X}\boldsymbol{\beta} + \|\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{p}\rho_\lambda(\beta_i) \\
&= \frac{1}{2n}\boldsymbol{Y}^\top\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^{\mathrm{LS}^\top}\boldsymbol{\beta} + \|\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{p}\rho_\lambda(\beta_i).
\end{aligned} \tag{3.1.12}$$

Hence, minimising (3.1.12) is equivalent to minimising the coordinate-wise problem

$$\min_{\beta_i\in\mathbb{R}}\left\{ -\hat{\beta}_i^{\mathrm{LS}}\beta_i + \beta_i^2 + \rho_\lambda(\beta_i)\right\}$$

which is simply the minimisation problem (3.1.9) expanded with $z = \hat{\beta}_i$ (ignoring the z^2 term that is constant). This completes the proof. □

3.1.3 Concavity Analysis

An essential difference between the lasso and the root-log regulariser is the concavity of the function. One can define a numerical measure of the concavity of a penalty function, which will play an important role in analysis of error and model selection consistency in the high dimensional regime:

Definition 3.1.1 (Maximum Concavity)**.** Write

$$\overline{\kappa}(t;\rho) = \sup_{t'>t}\frac{\rho'(t') - \rho'(t)}{t - t'} \tag{3.1.13}$$

to denote the concavity of ρ at t. The maximum concavity of the penalty function ρ is defined to be

$$\overline{\kappa}(\rho) = \max_{t\geq 0,\lambda>0}\overline{\kappa}(t;\rho). \tag{3.1.14}$$

Proposition 3.1.2 (Root-Log Inequalities)**.** We have the inequalities

$$\rho'(t) \leq \frac{1}{|t|} \quad \text{and} \quad \overline{\kappa}(\rho) \leq \frac{1}{2}. \tag{3.1.15}$$

Proof. For the bound on ρ', note that, for $t > 0$,

$$\begin{aligned}\rho'(t) = \frac{-t + \operatorname{sign}(t)\sqrt{4+t^2}}{2} &= \frac{4}{2(t+\sqrt{4+t^2})} \\ &= \frac{2}{t+\sqrt{4+t^2}} \leq \frac{2}{2t} = \frac{1}{t}.\end{aligned}$$

Similar computations apply for $t < 0$. As for the bound on $\overline{\kappa}(\rho)$, note that for $t' > t > 0$,

$$\begin{aligned}\frac{\rho'(t') - \rho'(t)}{t - t'} &= \frac{1}{t-t'}\left(\frac{-t' + \sqrt{4+(t')^2}}{2} - \frac{-t+\sqrt{4+t^2}}{2}\right) \\ &= \frac{1}{2(t-t')}\left(t - t' + \sqrt{4+(t')^2} - \sqrt{4+t^2}\right) \\ &= \frac{1}{2} + \frac{\sqrt{4+(t')^2} - \sqrt{4+t^2}}{2(t-t')} \\ &= \frac{1}{2} - \frac{t+t'}{2(\sqrt{4+t^2} + \sqrt{4+(t')^2})} \leq \frac{1}{2}.\end{aligned}$$

This completes the proof. □

3.2 Candidate Solution

Now that we know some essential properties of the root-log penalty, we can work on proving our first main result of this thesis: that there exists a local minimiser of (3.1.1) which is model selection consistent.

The most prominent source for our work in this section is [6], particularly Section 2. We also acknowledge the tremendous help of the supervisor in communicating these proofs to the author of this thesis. The basic strategy is to study local minimisers of the concave minimisation problem by comparison with the so-called *candidate solution* $\tilde{\boldsymbol{\beta}}$ formed with prior knowledge of the support set $\mathcal{S}$. The goal is to define this candidate solution so that it is zero on $\mathcal{S}^c$ and satisfies the KKT conditions on $\mathcal{S}$. The relationship between the candidate solution $\tilde{\boldsymbol{\beta}}$ and the true regression vector $\boldsymbol{\beta}^*$ is then studied by first finding an upper bound on the error $\|\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}\|_\infty$, allowing a minimal signal strength condition to imply the no false exclusion property. From there, it can be shown that in fact, $\tilde{\boldsymbol{\beta}}$ is, itself, a local minimiser of the concave minimisation problem, giving us the model selection consistent local minimiser we were looking for.

If we recall the proof of model selection consistency of the lasso given in Theorem 2.8.1, then this may sound familiar. Indeed, this strategy is very much the core idea behind the primal-dual-witness method, although the proof is much simpler for the lasso due to the simplicity of the penalty function.

Before we set this plan into action, we need to define the sufficient KKT conditions for local minimisers for the concave optimisation problem (3.1.1):

Definition 3.2.1 (Local KKT Conditions)**.** A vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is said to satisfy the *local KKT conditions* at level λ on the working features $\mathcal{A} \subset \{1, 2, \ldots, p\}$ if

$$\begin{cases} \boldsymbol{\beta}_{\mathcal{A}^c} = 0, \\ |\boldsymbol{v}_i^\top (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})| \leq n\lambda \quad \text{for all } i \in \mathcal{A}^c, \\ \boldsymbol{v}_i^\top (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) = n\lambda \rho'(\beta_i/\lambda) \quad \text{for all } i \in \mathcal{A}. \end{cases} \tag{3.2.1}$$

The proof that these are indeed sufficient conditions for local minimisation is given in Theorem 4.18, p.190 [5]; these conditions also appear on p.3074 [6]. If the working features $\mathcal{A}$ is exactly the true support $\mathcal{S}$, and the local minimiser $\hat{\boldsymbol{\beta}}$ is zero on $\mathcal{S}^c$, then we have the so-called *oracle solution*:

Definition 3.2.2 (Oracle Solution)**.** A vector $\boldsymbol{\beta}^o \in \mathbb{R}^p$ is called an *oracle solution at level* λ^o if

$$\begin{cases} \min_{i \in \mathcal{S}} |\beta_i^o| > 0, \quad \boldsymbol{\beta}^o_{\mathcal{S}^c} = \mathbf{0}, \\ |\boldsymbol{v}_i^\top (\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}} \boldsymbol{\beta}^o_{\mathcal{S}})| \leq n\lambda^o, \qquad \text{for all } i \in \mathcal{S}^c, \\ |\boldsymbol{v}_i^\top (\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}} \boldsymbol{\beta}^o_{\mathcal{S}})| = n\lambda^o \rho'(\beta_i^o/\lambda^o) \quad \text{for all } i \in \mathcal{S} \end{cases} \tag{3.2.2}$$

We now define the *candidate solution* as follows:

Definition 3.2.3 (Candidate Solution)**.** A vector $\tilde{\boldsymbol{\beta}}$ is called a *candidate solution* of the concave minimisation problem (3.1.1) if

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} &= \mathbf{0}, \\ \tilde{\boldsymbol{\beta}}_{\mathcal{S}} &= (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{Y} - n\lambda (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda) \\ &= \boldsymbol{\beta}^*_{\mathcal{S}} + \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} - \lambda \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda). \end{aligned} \tag{3.2.3}$$

Remark 3.2.1. As we discussed previously, the form of (3.2.3) is specifically designed so that $\tilde{\boldsymbol{\beta}}$ satisfies the equality KKT conditions on $\mathcal{S}$ from (3.2.2). Indeed, rearranging (3.2.3) in terms of $\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)$, we obtain

$$\begin{aligned} n\lambda (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda) &= \boldsymbol{\beta}^*_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}_{\mathcal{S}} + (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon} \\ \implies n\lambda \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda) &= (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}) \left(\boldsymbol{\beta}^*_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}_{\mathcal{S}} \right) + \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}, \end{aligned}$$

which is exactly the KKT equality conditions (3.2.2) coordinate-wise.

We shall also assume a lower bound on the eigenvalue of $\mathbf{X}_{\mathcal{S}}$:

$$\left(1 - \frac{1}{C_1}\right) \min_{\|\boldsymbol{u}\|_2 \neq 0} \frac{\|\mathbf{X}_{\mathcal{S}} \boldsymbol{u}\|_2^2}{n \|\boldsymbol{u}\|_2^2} \geq \frac{1}{2} \tag{3.2.4}$$

and assume that the design matrix $\mathbf{X}$ is normalised such that $\boldsymbol{v}_i^\top \boldsymbol{v}_i = n$ for all $1 \leq i \leq p$. Our first major goal is to find an upper bound on the L^∞-error $\|\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}\|_\infty$, where $\tilde{\boldsymbol{\beta}}$ is the candidate solution and $\boldsymbol{\beta}^*$ is the true regression parameter. Before we do so, we must give a probabilistic upper bound on the L^2-error $\|\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}\|_2$:

Proposition 3.2.1 (L^2-bounds on Candidate Error)**.** Suppose $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_0 = k$ is constant in n and $\tilde{\boldsymbol{\beta}}$ is a candidate solution satisfying (3.2.3). Then for $\sqrt{n}\lambda > 1$,

$$\mathbb{P}(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 2c_1\lambda(\sqrt{5}\sigma + \|\rho'(\boldsymbol{\beta}_{\mathcal{S}}/\lambda)\|_2) \geq 1 - \exp(-n\lambda^2). \tag{3.2.5}$$

Proof. Write $\boldsymbol{\delta}_{\mathcal{S}}$ to denote the candidate error $\tilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}$. The idea of the proof is to use the candidate condition (3.2.3) to obtain an expression for the mean squared error $\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2/n$, then use the RE condition (3.2.4) to obtain a bound for $\|\boldsymbol{\delta}_{\mathcal{S}}\|_2$. To begin, we rearrange the candidate condition (3.2.3) to obtain

$$-\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}} + \mathbf{X}_{\mathcal{S}}^\top\boldsymbol{\varepsilon} = n\lambda\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda),$$

giving us the coordinate-wise equations

$$\boldsymbol{v}_i^\top\boldsymbol{\varepsilon} - \boldsymbol{v}_i^\top\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}} = n\lambda\rho'(\tilde{\beta}_i/\lambda) \quad \text{for all } i \in \mathcal{S}.$$

Multiplying by $\delta_i = \tilde{\beta}_i - \beta_i^*$ and rewriting in matrix form, we obtain

$$\begin{aligned}
&(\tilde{\beta}_i - \beta_i^*)\boldsymbol{v}_i^\top\boldsymbol{\varepsilon} - n\lambda\rho'(\tilde{\beta}_i/\lambda)(\tilde{\beta}_i - \beta_i^*) = \boldsymbol{v}_i^\top\mathbf{X}_{\mathcal{S}}\delta_{\mathcal{S}}(\tilde{\beta}_i - \beta_i^*) \\
&\implies \frac{\boldsymbol{\varepsilon}^\top\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}}{n} - \lambda\sum_{i=1}^{p}\rho'(\tilde{\beta}_i/\lambda)\delta_i = \frac{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2}{n}.
\end{aligned}$$

Adding and subtracting $\sum_{i\in\mathcal{S}}\rho'(\beta_i^*/\lambda)$, it follows that

$$\begin{aligned}
\frac{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2}{n} &= \frac{\boldsymbol{\varepsilon}^\top\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}}{n} - \lambda\sum_{i\in\mathcal{S}}\delta_i\rho'(\beta_i^*/\lambda) + \lambda\sum_{i\in\mathcal{S}}\delta_i\left[\rho'(\beta_i^*/\lambda) - \rho'(\tilde{\beta}_i/\lambda)\right] \\
&\leq^* \frac{\boldsymbol{\varepsilon}^\top\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}}{n} - \lambda\sum_{i\in\ cS}\delta_i\rho'(\beta_i^*/\lambda) + \frac{\lambda}{2}\sum_{i\in\mathcal{S}}\delta_i\frac{\beta_i^* - \tilde{\beta}_i}{\lambda} \\
&= \frac{\boldsymbol{\varepsilon}^\top\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}}{n} - \lambda\sum_{i\in\ cS}\delta_i\rho'(\beta_i^*/\lambda) + \frac{1}{2}\|\boldsymbol{\delta}_{\mathcal{S}}\|_2^2 \\
&\leq^{**} \sum_{i\in\mathcal{S}}\delta_i\left[\frac{\boldsymbol{v}_i^\top\boldsymbol{\varepsilon}}{n} - \lambda\rho'(\beta_i^*/\lambda)\right] + \left(1 - \frac{1}{C_1}\right)\frac{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2}{n},
\end{aligned}$$

where in $(*)$ we used the maximum concavity of ρ' and in $(**)$ we used the RE condition along with writing the term $\boldsymbol{\varepsilon}^\top\mathbf{X}_{\mathcal{S}}\delta_{\mathcal{S}}$ in terms of the product of columns. From here, we can rearrange the $\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2$ term and carry on:

$$\begin{aligned}
\frac{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2}{C_1 n} &\leq \sum_{i\in\mathcal{S}}\delta_i\left[\frac{\boldsymbol{v}_i^\top\boldsymbol{\varepsilon}}{n} - \lambda\rho'(\beta^*/\lambda)\right] \\
&\leq \|\boldsymbol{\delta}_{\mathcal{S}}\|_2\left\|\frac{\mathbf{X}_{\mathcal{S}}^\top\boldsymbol{\varepsilon}}{n} - \lambda\rho'(\boldsymbol{\beta}_{\mathcal{S}}/\lambda)\right\|_2.
\end{aligned}$$

From here, divide by $\|\boldsymbol{\delta}_{\mathcal{S}}\|_2$ and apply the RE condition:

$$\begin{aligned}
\frac{\|\boldsymbol{\delta}\|_{\mathcal{S}}}{2C_1} \leq \frac{\|\boldsymbol{\delta}_{\mathcal{S}}\|_2}{C_1 n} \frac{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2}{\|\boldsymbol{\delta}_{\mathcal{S}}\|_2^2} &= \frac{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}\|_2^2}{C_1 n \|\boldsymbol{\delta}_{\mathcal{S}}\|_2} \\
&\leq \left\| \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} - \lambda \rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda) \right\|_2 \\
&\leq \left\| \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} \right\|_2 + \lambda \|\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_2 \qquad (3.2.6) \\
&= \lambda \frac{K\sigma}{\lambda\sqrt{n}} + \lambda \|\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_2 \\
&= \lambda \left[\frac{K\sigma}{\lambda\sqrt{n}} + \|\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_2 \right]
\end{aligned}$$

where $K := \|\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}\|_2/(\sigma\sqrt{n})$. We can now directly apply a probabilistic bound on multivariate Gaussian random vectors (Proposition 1.1 in [8]), to obtain

$$\mathbb{P}\big(K^2 \leq 1 + 2\sqrt{\alpha_1 t} + 2\alpha_2 t\big) \geq 1 - \exp(-t),$$

with $\alpha_1 := \operatorname{tr}\left[(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})/(sn)^2\right] \leq 1$ and $\alpha_2 := \lambda_{\max}(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})/(sn) \leq 1$. With $t > 1$, we can take square roots to get

$$\mathbb{P}(K \leq \sqrt{1 + 4t}) \geq 1 - \exp(-t),$$

and so with high probability, we can bound K by $\sqrt{5}$. Multiplying by $2C_1$ and replacing K with $\sqrt{5}$ in (3.2.6) completes the proof. □

With this result, we now turn to bounding the L^∞-error $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_\infty$ from above:

Lemma 3.2.1 (L^∞-bounds on Candidate Error)**.** Write $\Sigma := \Sigma_{\mathcal{S},\mathcal{S}} := \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}/n$. If $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_0 = k$ is constant in n and $\tilde{\boldsymbol{\beta}}$ is a candidate solution satisfying (3.2.3), then as $\sqrt{n}\lambda \to \infty$, with probability at least $1 - (s+1)\exp(-n\lambda^2)$,

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leq c_2\lambda + c_3 \frac{\lambda^2}{\min_{i\in\mathcal{S}}|\beta_i^*|}, \qquad (3.2.7)$$

for $c_2 := \sigma(2 + c_1\sqrt{5\|\Sigma^{-2}\|_\infty})$ and $c_3 := \|\Sigma^{-1}\|_\infty + c_1\sqrt{\|\Sigma^{-2}\|_\infty}$.

Proof. Starting again from the candidate conditions (3.2.3) and using the definition of Σ, we have that

$$\tilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}} = \frac{\Sigma^{-1}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n} - \lambda\Sigma^{-1}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda),$$

so that

$$\frac{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty}{\lambda} = \left\|\Sigma^{-1}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}/n - \lambda\Sigma^{-1}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\right\|_\infty$$
$$\leq \underbrace{\frac{\left\|\Sigma^{-1}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}\right\|_\infty}{n\lambda}}_{(1)} + \underbrace{\frac{\left\|\Sigma^{-1}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\right\|_\infty}{n}}_{(2)} \tag{3.2.8}$$

The goal now is to bound (1) and (2) in (3.2.8). For (1), define

$$\boldsymbol{Z} = \frac{\Sigma^{-1}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{2n\sigma\lambda} = \frac{\left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}/n\right)^{-1}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{2n\sigma\lambda},$$

so that $2\sigma\|\boldsymbol{Z}\|_\infty$ is exactly (1). If $\boldsymbol{e}_i \in \mathbb{R}^p$ is the ith basis vector (viewed as a column vector), we have

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\boldsymbol{e}_i^\top\sqrt{2n}\lambda\boldsymbol{Z}, \boldsymbol{e}_i^\top\sqrt{2n}\lambda\boldsymbol{Z}) &= \boldsymbol{e}_i^\top\mathbb{C}\mathrm{ov}(\boldsymbol{Z},\boldsymbol{Z})\boldsymbol{e}_i \\
&= \frac{\boldsymbol{e}_i^\top\Sigma^{-1}\boldsymbol{e}_i}{2} \\
&= \frac{1}{2}\boldsymbol{e}_i^\top\left(\frac{\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}\boldsymbol{e}_i \\
&\leq \frac{1}{2}\lambda_{\max}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \\
&= \frac{1}{2}\lambda_{\min}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \\
&\leq 1 - \frac{1}{C_1} \leq 1.
\end{aligned}$$

Hence, each component of $\sqrt{2n}\lambda\boldsymbol{Z}$ is a zero-mean Gaussian with variance at most 1. For each component Z_i, by the Gaussian tail bound, we have

$$\mathbb{P}(\sqrt{2n}\lambda Z_i > 2n\lambda) = \mathbb{P}(Z_i > 1) \leq \exp(-n\lambda^2),$$

and so by the union bound,

$$\mathbb{P}(\|\boldsymbol{Z}\|_\infty) \leq 1) \geq 1 - k\exp(-n\lambda^2). \tag{3.2.9}$$

We now focus on bounding (2) in (3.2.8). Define $\boldsymbol{V} := \lambda\Sigma^{-1}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)$. Then we can add and subtract $\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)$ to get

$$\begin{aligned}
\|\boldsymbol{V}\|_\infty &\leq \lambda\|\Sigma^{-1}\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_\infty + \lambda\|\Sigma^{-1}(\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda) - \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\|_\infty \\
&\leq^* \lambda\|\Sigma^{-1}\|_\infty\|\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_\infty + \frac{1}{2}\|\Sigma^{-1}(\boldsymbol{\beta}^*_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}_{\mathcal{S}})\|_\infty \\
&\leq^{**} \lambda\|\Sigma^{-1}\|_\infty\|\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_\infty + \frac{1}{2}\sqrt{\|\Sigma^{-2}\|_\infty}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2,
\end{aligned}$$

where in $(*)$ we used the maximum concavity of ρ and in $(**)$ we used the Cauchy-Schwarz inequality. Noting that we can bound $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ with high probability from Proposition 3.2.1, we return to (3.2.8):

$$\begin{aligned}\frac{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty}{\lambda} &\leq 2\sigma\|\boldsymbol{Z}\|_\infty + \|\Sigma^{-1}\|_\infty\|\rho'(\boldsymbol{\beta}_{\mathcal{S}}/\lambda)\|_\infty + \frac{1}{2}\sqrt{\|\Sigma^{-2}\|_\infty}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \\ &\leq^* 2\sigma + \|\Sigma^{-1}\|_\infty\|\rho'(\boldsymbol{\beta}_{\mathcal{S}}/\lambda)\|_\infty + c_1(\sqrt{5}\sigma + \|\rho'(\boldsymbol{\beta}^*_{\mathcal{S}}/\lambda)\|_2)\sqrt{\|\Sigma^{-2}\|_\infty} \\ &= (2\sigma + c_1\sqrt{5}\sigma) + \|\rho'(\boldsymbol{\beta}_{\mathcal{S}}/\lambda)\|_\infty(\|\Sigma^{-1}\| + c_1\sqrt{\|\Sigma^{-2}\|_\infty})\end{aligned} \tag{3.2.10}$$

where, by combining our bounds on $\boldsymbol{Z}$ and the result from Proposition 3.2.1, the inequality in $(*)$ occurs with probability $1-(k+1)\exp(-n\lambda^2)$. Assuming that $\sigma, c_1, \|\Sigma^{-1}\|_\infty$ and $\sqrt{\|\sigma^{-2}\|_\infty}$ remain constant in n, and using the absolute value bound $\rho'(t) \leq |t|$ so that

$$\|\rho'(\boldsymbol{\beta}_{\mathcal{S}}/\lambda)\|_\infty \leq \frac{\lambda}{\min_{i\in\mathcal{S}}|\beta_i|},$$

we can multiply by λ on both sides in (3.2.10) to obtain

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leq c_2\lambda + c_3\frac{\lambda^2}{\min_{i\in\mathcal{S}}|\beta^*_i|},$$

with $c_2 = \sigma(2+c_1\sqrt{5\|\Sigma^{-2}\|_\infty})$ and $c_3 = \|\Sigma^{-1}\|_\infty + c_1\sqrt{\|\Sigma^{-2}\|_\infty}$, exactly as we wanted. □

Exactly as we did for the lasso, an upper bound on $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_\infty$ allows us to prove the following corollary:

Corollary 3.2.1 (No False Exclusion)**.** Under the minimal strength condition

$$\min_{i\in\mathcal{S}}|\beta^*_i| > \lambda(c_2 + \sqrt{c_3}), \tag{3.2.11}$$

the candidate solution $\tilde{\boldsymbol{\beta}}$ has the no false exclusion property: that is, $\text{supp}\,(\tilde{\boldsymbol{\beta}}) \supset \text{supp}\,(\boldsymbol{\beta}^*)$.

Proof. If $\tilde{\beta}_i = 0$ for any $i \in \mathcal{S}$ but $\min_{i\in\mathcal{S}}|\beta^*_i| > \lambda(c_2 + \sqrt{c_3})$, then the L^∞-bounds found in Lemma 3.2.1 would be violated. Hence, $\text{supp}\,(\tilde{\boldsymbol{\beta}}) \supset \text{supp}\,(\boldsymbol{\beta}^*)$, as we wanted. □

Our goal now is to show that the candidate solution $\tilde{\boldsymbol{\beta}}$ is indeed a local minimiser of (3.1.1), i.e., satisfies the local KKT conditions (3.2.1). As we discussed in Remark 3.2.1, this means showing the KKT inequality

$$\|\mathbf{X}^\top_{\mathcal{S}^c}(\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\|_\infty \leq n\lambda$$

holds. The sufficient conditions for the above inequality to hold is the subject of Lemma 3.2.2. Finally, Theorem 3.2.1 ties the results of Lemma 3.2.1 and Lemma 3.2.2 together to show that there indeed exists a model selection consistent local minimiser of (3.1.1).

Lemma 3.2.2 (Candidate Inequality)**.** Write $\boldsymbol{A} := \mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}} (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1}$ and suppose

$$\frac{2\sigma(p-k)}{\sqrt{n}\lambda} \exp\left(-\frac{n\lambda^2}{8\sigma^2}\right) \to 0 \text{ as } n \to \infty.$$

If $\mathbb{P}(\min_{i\in\mathcal{S}}|\tilde{\beta}_i| \geq 2\lambda\|\boldsymbol{A}\|_\infty) \geq 1-\alpha$ for some $\alpha \in [0,1]$, then with probability at least $1 - \alpha - \frac{2\sigma(p-k)}{\sqrt{n}\lambda} \exp\left(\frac{-n\lambda^2}{8\sigma^2}\right)$, the candidate solution $\tilde{\boldsymbol{\beta}}$ satisfies

$$\|\mathbf{X}_{\mathcal{S}^c}^\top (\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\|_\infty \leq n\lambda. \tag{3.2.12}$$

Proof. We want to bound $\|\mathbf{X}_{\mathcal{S}^c}^\top (\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\|_\infty$ by $n\lambda$. To do so, note that

$$\mathbf{X}_{\mathcal{S}^c}^\top (\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{X}_{\mathcal{S}^c}^\top (\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}}\tilde{\boldsymbol{\beta}}_{\mathcal{S}}) = \mathbf{X}_{\mathcal{S}^c}^\top \left(\mathbf{X}_{\mathcal{S}}(\boldsymbol{\beta}^*_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}_{\mathcal{S}}) + \boldsymbol{\varepsilon}\right).$$

By the candidate conditions (3.2.3), we have an expression for the candidate error:

$$\boldsymbol{\beta}^*_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}_{\mathcal{S}} = \lambda \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda) - \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \frac{\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}}{n},$$

and so

$$\begin{aligned}
\|\mathbf{X}_{\mathcal{S}^c}^\top (\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\|_\infty &= \|\mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}} \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda) + \mathbf{X}_{\mathcal{S}^c}^\top \left(\boldsymbol{I} - \mathbf{X}_{\mathcal{S}} \left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}^{-1} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}\right)\right)\|_\infty \\
&= n\lambda \|\mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}} \left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}\right)^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda) + \frac{1}{n\lambda} \mathbf{X}_{\mathcal{S}^c}^\top \left(\boldsymbol{I} - \mathbf{X}_{\mathcal{S}} \left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}^{-1} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}\right)\right)\|_\infty \\
&\leq n\lambda \left[\|\mathbf{X}_{\mathcal{S}^c}^\top \mathbf{X}_{\mathcal{S}} \left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}\right)^{-1} \rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\|_\infty + \|\frac{1}{n\lambda} \mathbf{X}_{\mathcal{S}^c}^\top \left(\boldsymbol{I} - \mathbf{X}_{\mathcal{S}} \left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}^{-1} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon}\right)\right)\|_\infty\right] \\
&= n\lambda \left[\|\boldsymbol{A}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\|_\infty + \|\sigma \boldsymbol{W}\|_\infty\right]
\end{aligned} \tag{3.2.13}$$

where $\boldsymbol{P}_{\mathcal{S}} := \mathbf{X}_{\mathcal{S}} (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^\top$ as the projection matrix onto $\text{Col}(\mathbf{X}_{\mathcal{S}})$ (and so $\boldsymbol{I} - \boldsymbol{P}_{\mathcal{S}}$ is the orthogonal projection of $\text{Col}(\mathbf{X}_{\mathcal{S}})$), and

$$\boldsymbol{W} = \frac{\mathbf{X}_{\mathcal{S}^c}^\top (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{S}}) \boldsymbol{\varepsilon}}{\sigma n \lambda}.$$

We now bound $\|\boldsymbol{A}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda\|_\infty$ and $\|\sigma \boldsymbol{W}\|_\infty$ in (3.2.13). Firstly, we have

$$\|\boldsymbol{A}\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\|_\infty \leq \|\boldsymbol{A}\|_\infty \|\rho'(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}/\lambda)\|_\infty \leq \|\boldsymbol{A}\|_\infty \frac{\lambda}{\min_{i\in\mathcal{S}}|\tilde{\beta}_i|} \tag{3.2.14}$$

using the absolute value bound $\rho'(t) \leq 1/|t|$. Next, considering $\boldsymbol{W}$ coordinate-wise,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(W_i) &= \mathbb{C}\mathrm{ov}\left(\boldsymbol{v}_i^\top \frac{(\boldsymbol{I} - \boldsymbol{P}_\mathcal{S})\boldsymbol{\varepsilon}}{n\sigma\lambda}\right) \\
&= \frac{1}{n^2\sigma^2\lambda^2}\boldsymbol{v}_i^\top \mathbb{C}\mathrm{ov}((\boldsymbol{I} - \boldsymbol{P}_\mathcal{S})\boldsymbol{\varepsilon})\boldsymbol{v}_i \\
&= \frac{1}{n^2\sigma^2\lambda^2}\boldsymbol{v}_i^\top (\boldsymbol{I} - \boldsymbol{P}_\mathcal{S})(\sigma^2\boldsymbol{I})(\boldsymbol{I} - \boldsymbol{P}_\mathcal{S})^\top \\
&=^* \frac{1}{n^2\lambda^2}(\boldsymbol{v}_i^\top (\boldsymbol{I} - \boldsymbol{P}_\mathcal{S})\boldsymbol{v}_i) \\
&= \frac{1}{n^2\lambda^2}\boldsymbol{v}_i^\top \boldsymbol{v}_i =^{**} \frac{1}{n\lambda^2},
\end{aligned}$$

so that $\sqrt{n}\lambda\boldsymbol{W}$ has coordinates that are zero-mean Gaussian random variables with variance 1. Note that in $(*)$ we used the fact that $\boldsymbol{I} - \boldsymbol{P}_\mathcal{S}$ is an orthogonal projection and $\boldsymbol{v}_i \in \mathrm{Col}(\mathbf{X}_\mathcal{S})$, and in $(**)$ we recall the normalisation $\boldsymbol{v}_i^\top \boldsymbol{v}_i = n$. Applying the Gaussian tail bound, we see that

$$\mathbb{P}\left(\sqrt{n}\lambda W_i \geq \frac{\sqrt{n}\lambda}{2\sigma}\right) \leq \frac{2\sigma}{\sqrt{n}\lambda}\exp\left(-\frac{n\lambda^2}{8\sigma^2}\right),$$

and so, by the union bound,

$$\mathbb{P}\left(\|\sigma\boldsymbol{W}\|_\infty \leq \frac{1}{2}\right) \geq 1 - \frac{2\sigma(p-k)}{\sqrt{n}\lambda}\exp\left(-\frac{n\lambda^2}{8\sigma^2}\right). \tag{3.2.15}$$

Write U_1 to denote the event $\{\min_{i\in\mathcal{S}}|\tilde{\beta}_i| \geq 2\lambda\|\boldsymbol{A}\|_\infty\}$ and U_2 to denote the event $\{\|\sigma\boldsymbol{W}\|_\infty \leq \frac{1}{2}\}$. By the hypothesis, $\mathbb{P}(U_1) \geq 1 - \alpha$ and $\mathbb{P}(U_2)$ is given in (3.2.15). On the joint event $U_1 \cap U_2$, we have

$$\min_{i\in\mathcal{S}}|\tilde{\beta}_i| \geq 2\lambda\|\boldsymbol{A}\|_\infty \geq \frac{\lambda\|\boldsymbol{A}\|_\infty}{1 - \sigma\|\boldsymbol{W}\|_\infty} \tag{3.2.16}$$

which we can see by rearranging the inequality $\|\sigma\boldsymbol{W}\|_\infty \leq \frac{1}{2}$ to obtain $2 \geq (1-\sigma\|\boldsymbol{W}\|_\infty)^{-1}$. Finally, by combining inequalities (3.2.14), (3.2.15) and (3.2.16), we see that

$$\begin{aligned}
\|\boldsymbol{A}\rho'(\tilde{\boldsymbol{\beta}}_\mathcal{S}/\lambda)\|_\infty + \|\sigma\boldsymbol{W}\|_\infty &\leq \|\sigma\boldsymbol{W}\|_\infty + \|\boldsymbol{A}\|_\infty \frac{\lambda}{\min_{i\in\mathcal{S}}|\tilde{\beta}_i|} \\
&\leq \|\sigma\boldsymbol{W}\|_\infty + \lambda\|\boldsymbol{A}\|_\infty \frac{1 - \|\sigma\boldsymbol{W}\|_\infty}{\lambda\|\boldsymbol{A}\|_\infty} = 1
\end{aligned}$$

with probability at least

$$\mathbb{P}(U_1 \cap U_2) \geq 1 - \alpha - \frac{2\sigma(p-k)}{\sqrt{n}\lambda}\exp\left(-\frac{n\lambda^2}{8\sigma^2}\right)$$

for $\alpha := (k+1)\exp(-n\lambda^2)$. With this probability, we conclude from (3.2.13) that

$$\|\mathbf{X}_{\mathcal{S}^c}^\top(\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\mathcal{S}})\|_\infty \leq n\lambda,$$

finishing the proof.

□

Theorem 3.2.1 (Candidate Solution is a Local Solution)**.** Suppose the following minimal signal strength condition is satisfied:

$$\min_{i\in\mathcal{S}}|\beta_i^*| \geq \lambda(2\|\mathbf{X}_{\mathcal{S}^c}^\top\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}})^{-1}\|_\infty + c_2 + \sqrt{c_3}) := \lambda(2\|\boldsymbol{A}\|_\infty + c_2 + \sqrt{c_3}),$$

where $c_2 = \sigma(2 + c_1\sqrt{5\|\Sigma^{-2}\|_\infty})$ and $c_3 = \|\Sigma^{-1}\|_\infty + c_1\sqrt{\|\Sigma^{-2}\|_\infty}$. Then with probability

$$1 - \frac{2\sigma(p-k)}{\sqrt{n}\lambda}\exp\left(-\frac{n\lambda^2}{8\sigma^2}\right) - (s+1)\exp(-n\lambda^2),$$

the candidate solution $\tilde{\boldsymbol{\beta}}$ satisfies the oracle KKT conditions (3.2.2) with $\lambda^o = \lambda$, and hence is a local minimiser of the root-log minimisation problem.

Proof. Let U be the event $\{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leq c_2\lambda + c_3\frac{\lambda^2}{\min_{i\in\mathcal{S}}|\beta_i^*|}\}$. Under U and the minimal signal strength condition,

$$\min_{i\in\mathcal{S}}|\tilde{\beta}_i| + \lambda c_2 + c_3\frac{\lambda^2}{\min_{i\in\mathcal{S}}|\beta_i^*|} \geq^* |\beta_i^*| \geq \lambda(2\|\boldsymbol{A}\|_\infty + c_2 + \sqrt{c_3}) \quad \text{for all } i \in \mathcal{S},$$

where $(*)$ follows from the upper bound in U. Rearranging, we see that

$$|\tilde{\beta}_i| \geq 2\lambda\|\boldsymbol{A}\|_\infty + \lambda\sqrt{c_3}\left(1 - \frac{\lambda\sqrt{c_3}}{\min_{i\in\mathcal{S}}|\beta_i^*|}\right) \geq 2\lambda\|\boldsymbol{A}\|_\infty \quad \text{for all } i \in \mathcal{S},$$

and so from the L^∞-bounds found in Lemma 3.2.1,

$$\mathbb{P}\left(\min_{i\in\mathcal{S}}|\tilde{\beta}_i| \geq 2\lambda\|\boldsymbol{A}\|_\infty\right) \geq \mathbb{P}(U) \geq 1 - (k+1)\exp(-n\lambda^2),$$

where the first inequality follows since event U guarantees the minimum bound on $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ and the second inequality is exactly the result of Lemma 3.2.1. It follows immediately from Lemma 3.2.2 that the candidate solution $\tilde{\boldsymbol{\beta}}$ satisfies

$$\|\mathbf{X}_{\mathcal{S}^c}^\top(\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\|_\infty \leq n\lambda$$

with probability

$$1 - (k+1)\exp(-n\lambda^2) - \frac{2\sigma(p-k)}{\sqrt{n}\lambda}\exp\left(-\frac{n\lambda^2}{8\sigma^2}\right)$$

which converges to 1 if $n\lambda^2 \to \infty$. Of course, by construction, $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}$ and by the no false exclusion property (Corollary 3.2.1), we have$|\tilde{\boldsymbol{\beta}}_i| > 0$ for all $i \in \mathcal{S}$. As per Remark 3.2.1, we already know that $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ satisfies the equality constraints on $\mathcal{S}$. We conclude that the candidate solution $\tilde{\boldsymbol{\beta}}$ in fact satisfies the local KKT conditions, and hence there exists a local minimiser of the concave minimisation problem that is model selection consistent. □

3.3 Analysis of Local Minimisers found by Path-Finding Algorithms

Under the minimal signal strength condition and the lower eigenvalue bound, we can guarantee there exists a model selection consistent local minimiser of the concave regularisation problem, given exactly by the candidate solution. In practice, however, an optimisation algorithm working on a concave minimisation problem will find many local minima, not necessarily the candidate solution. Can the theory guarantee that such local minima also have the model selection properties of the candidate solution? In this section, we answer in the affirmative, under some further assumptions on the algorithm that we apply.

The starting point of our analysis is requiring that the candidate solutions satisfy the η-oracle KKT conditions (as is done in Equation (2.8) in [6]), where $\eta \in (0, 1)$:

$$\begin{cases} \min_{i\in\mathcal{S}}|\tilde{\beta}_i| > 0, \quad \tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}. \\ \|\mathbf{X}_{\mathcal{S}^c}^\top(\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c})\|_\infty \leq \eta\lambda n, \\ |\boldsymbol{v}_i^\top(\boldsymbol{Y} - \mathbf{X}_{\mathcal{S}}\tilde{\boldsymbol{\beta}}_{\mathcal{S}})| = n\lambda\rho'(\tilde{\beta}_i/\lambda) \quad \text{for all } i \in \mathcal{S}. \end{cases} \tag{3.3.1}$$

That is, we have strengthened the conditions on $\mathcal{S}^c$ to be a strict inequality when compared with the usual local KKT conditions (3.2.1). To find the probability of this occurring, we only need to employ Lemma 3.2.2, and change the bound on $\|\mathbf{X}_{\mathcal{S}^c}^\top(\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\mathcal{S}})\|_\infty$ to $\eta n\lambda$; since η is assumed to be constant, as long as $n\lambda^2 \to \infty$, the probability of (3.3.1) (and the rest of the KKT conditions) holding approaches 1. For the rest of this chapter, we assume that the η-KKT conditions (3.3.1) hold.
We also recall the definitions of the cone set

$$\mathcal{C}(\mathcal{S}; \xi) := \{\boldsymbol{u} \in \mathbb{R}^p \mid \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq \xi\|\boldsymbol{u}_{\mathcal{S}}\|_1\},$$

which we will apply with $\xi := \frac{1+\eta}{1-\eta}$, and we also recall the definition of the restricted eigenvalue

$$\mathrm{RE}_2(\mathcal{S}; \xi) := \inf_{\boldsymbol{u}\neq 0}\left\{\frac{\|\mathbf{X}\boldsymbol{u}\|_2^2}{n\|\boldsymbol{u}\|_2^2} \mid \boldsymbol{u} \in \mathcal{C}(\mathcal{S}; \xi)\right\},$$

which we assume is a finite positive number. Just like in the error analysis of Chapter 2, we shall assume the restricted eigenvalue bound

$$\left(1 - \frac{1}{C_1}\right)\mathrm{RE}_2(\mathcal{S}; \xi) \geq \frac{1}{2} \tag{3.3.2}$$

for some $C_1 > 1$.

Our main result (Theorem 3.3.1) concerns the accuracy of local minima $\hat{\boldsymbol{\beta}}$ satisfying the local KKT conditions such that the error $\boldsymbol{\delta} := \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$ lies in the cone $\mathcal{C}(\mathcal{S}; \xi)$, where $\tilde{\boldsymbol{\beta}}$ is the candidate solution. Why is such an assumption permissible? The answer lies in the cone-preserving properties of algorithms that compute local minimisers. In particular, a powerful result in [6] shows us that local optimisation algorithms which take "small enough steps" from a solution with error vector originally in the cone $\mathcal{C}(\mathcal{S}; \xi)$, finds local minimisers with error vectors also within the cone $\mathcal{C}(\mathcal{S}; \xi)$.

The following two propositions formalise theoretically why such an algorithm ensures that $\boldsymbol{\delta} \in \mathcal{C}(\mathcal{S}; \xi)$. We first state a paraphrased version of Proposition 2, p.3080 in [6]. Due to the notation and complexity involved, we do not give all of the details.

Proposition 3.3.1 (Cone Inequalities in Path-Finding Algorithms)**.** Suppose $\mathrm{RE}_2^2(\mathcal{S}; \xi) \geq \kappa$ for some constant $\kappa > 0$. If $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are two solutions of the local KKT conditions (3.2.1) (with potentially different regularising parameters $\hat{\lambda}$ and $\tilde{\lambda}$ respectively), then for some $\boldsymbol{\beta}^0$ and constant $a_0 > 0$,

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \in \mathcal{C}(\mathcal{S}; \xi) \quad \text{and} \quad \|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_1 \leq a_0 \tilde{\lambda} \implies \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \in \mathcal{C}(\mathcal{S}; \xi).$$

That is, if the error vector $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ is initially in the cone $\mathcal{C}(\mathcal{S}; \xi)$, then all local minimisers $\hat{\boldsymbol{\beta}}$ in an L^1-neighbourhood of $\tilde{\boldsymbol{\beta}}$ have error vectors in the cone as well.

Since we know the candidate solution $\tilde{\boldsymbol{\beta}}$ is model selection consistent, we immediately have the next proposition:

Proposition 3.3.2. Suppose $\mathrm{RE}_2^2(\mathcal{S}; \xi) \geq \kappa > 0$ and that an algorithm has computed a local minimiser $\hat{\boldsymbol{\beta}}$ of the root-log minimisation problem corresponding to a regularising parameter $\hat{\lambda}$. Then there exists a constant $a_0 > 0$ such that if $\|\hat{\boldsymbol{\beta}}\|_1 \leq a_0 \hat{\lambda}$, then $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} \in \mathcal{C}(\mathcal{S}; \xi)$, where $\tilde{\boldsymbol{\beta}}$ is the candidate solution.

Proof. In Proposition 3.3.1, take $\tilde{\boldsymbol{\beta}}$ as $\mathbf{0}$ (which is a minimiser when $\lambda = \infty$) and $\boldsymbol{\beta}^0$ as $\tilde{\boldsymbol{\beta}}$ (the candidate solution, which is a local minimiser by Theorem 3.2.1). Now, since $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}$, we naturally have

$$\mathbf{0} - \tilde{\boldsymbol{\beta}} = -\tilde{\boldsymbol{\beta}} \in \mathcal{C}(\mathcal{S}; \xi).$$

By Proposition 3.3.1, there exists $a_0 > 0$ such that

$$\|\hat{\boldsymbol{\beta}} - \mathbf{0}\|_1 = \|\hat{\boldsymbol{\beta}}\|_1 \leq a_0 \hat{\lambda},$$

which implies that $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} \in \mathcal{C}(\mathcal{S}; \xi)$. This completes the proof. □

Hence, it is valid for us to assume that $\boldsymbol{\delta} \in \mathcal{C}(\mathcal{S}; \xi)$, since all local minimisers found by such an algorithm within some L^1-ball of the origin satisfies this condition. Under this assumption, we can show that the local minimiser $\hat{\boldsymbol{\beta}}$ must actually equal the candidate solution $\tilde{\boldsymbol{\beta}}$. In essence, this result states that *most computed local minimisers of the concave minimisation problem* (3.1.1) *are model selection consistent*. We demonstrate this in the final theorem below.

Theorem 3.3.1. Suppose $\hat{\boldsymbol{\beta}}$ satisfies the local KKT conditions

$$\boldsymbol{v}_i^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = n\lambda\rho'(\hat{\beta}_i/\lambda) \quad \text{for all } i \in \{1, 2, \ldots, p\}, \tag{3.3.3}$$

such that $\boldsymbol{\delta} := \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} \in \mathcal{C}(\mathcal{S}; \xi)$, and suppose $\tilde{\boldsymbol{\beta}}$ is the candidate solution satisfying the η-KKT conditions (3.3.1). Then $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$, so that the local minimiser is in fact the model selection consistent candidate solution.

Proof. By adding and subtracting $\boldsymbol{v}_i^\top \mathbf{X}\tilde{\boldsymbol{\beta}}$ with the KKT conditions (3.3.3), we see that

$$n\lambda\rho'(\hat{\beta}_i/\lambda) = \boldsymbol{v}_i^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{v}_i^\top(\boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + \boldsymbol{v}_i^\top\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \quad \text{for all } i \in \{1, 2, \ldots, p\}.$$

Writing $\boldsymbol{r} := \boldsymbol{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$, this gives us

$$\boldsymbol{v}_i^\top\boldsymbol{r} + \boldsymbol{v}_i^\top\mathbf{X}\boldsymbol{\delta} = n\lambda\rho'(\hat{\beta}_i/\lambda) \quad \text{for all } i \in \{1, 2, \ldots, p\}.$$

Multiplying by $\hat{\beta}_i - \tilde{\beta}_i$ and summing over i and continuing, we obtain

$$\begin{aligned}
\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n} &= \sum_{i=1}^{p}(\hat{\beta}_i - \tilde{\beta}_i)\boldsymbol{v}_i^\top\mathbf{X}\boldsymbol{\delta} \\
&= \lambda\sum_{i=1}^{p}(\hat{\beta}_i - \tilde{\beta}_i)\rho'(\hat{\beta}_i/\lambda) + \sum_{i=1}^{p}(\tilde{\beta}_i - \hat{\beta}_i)\frac{\boldsymbol{v}_i^\top\boldsymbol{r}}{n} \\
&= \sum_{i=1}^{p}(\hat{\beta}_i - \tilde{\beta}_i)\left[\frac{\boldsymbol{v}_i^\top r}{n} - \lambda\rho'(\tilde{\beta}_i/\lambda)\right] + \lambda\sum_{i=1}^{p}(\tilde{\beta}_i - \hat{\beta}_i)\left[\rho'(\hat{\beta}_i/\lambda) - \rho'(\tilde{\beta}_i/\lambda)\right] \\
&=^{*} \sum_{i\in\mathcal{S}^c}\hat{\beta}_i\left[\frac{\boldsymbol{v}_i^\top\boldsymbol{r}}{n} - \lambda\rho'(0)\right] + \sum_{i\in\mathcal{S}}(b_i - \tilde{\beta}_i)\left[\frac{\boldsymbol{v}_i^\top\boldsymbol{r}}{n} - \lambda\rho'(\tilde{\beta}_i/\lambda)\right] \\
&\quad + \lambda\sum_{i=1}^{p}(\tilde{\beta}_i - \hat{\beta}_i)\left[\rho'(\hat{\beta}_i/\lambda) - \rho'(\tilde{\beta}/\lambda)\right] \\
&\leq^{**} \sum_{i\in\mathcal{S}^c}\hat{\beta}_i\left[\frac{\boldsymbol{v}_i^\top\boldsymbol{r}}{n} - 1\right] + \sum_{i=1}^{p}(\tilde{\beta}_i - \hat{\beta}_i)^2 \\
&\leq \lambda(\eta - 1)\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \frac{1}{2}\|\boldsymbol{\delta}\|_2^2.
\end{aligned} \tag{3.3.4}$$

We clarify some of the computations above. In $(*)$, we split the sum into components in $\mathcal{S}$ and $\mathcal{S}^c$ and recalled that $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}$. In $(**)$, we set $\rho'(0) = \frac{1}{\lambda}$ as the most favourable value, recalled that $\tilde{\boldsymbol{\beta}}$ satisfied the KKT conditions (3.2.1) and hence the sum over $\mathcal{S}$ is zero, and finally we used the maximum concavity of ρ for the last sum. Now, under the assumption $\boldsymbol{\delta} \in \mathcal{C}(\mathcal{S}; \xi)$, we can apply the restricted eigenvalue condition (3.3.2) to obtain

$$\frac{1}{2}\|\boldsymbol{\delta}\|_2^2 \leq \left(1 - \frac{1}{C_1}\right)\frac{\|\mathbf{X}\delta\|_2^2}{n},$$

and rearranging the result of (3.3.4), we see that

$$\frac{\mathrm{RE}_2^2(\mathcal{S};\xi)\|\boldsymbol{\delta}\|_2^2}{C_1} \leq \frac{\|\mathbf{X}\delta\|_2^2}{C_1 n} \leq \lambda(\eta-1)\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leq 0.$$

This is only possible if $\|\boldsymbol{\delta}\|_2 = 0$. We conclude that the local minimiser $\hat{\boldsymbol{\beta}}$ is in fact equal to the candidate solution $\tilde{\boldsymbol{\beta}}$, and hence has the required support $\mathcal{S}$. This completes the proof, and the final result of this thesis follows. □

Chapter 4

Conclusion

The primary objective of this thesis was to introduce the new root-log penalty function and study its theoretical model selection properties under the hard sparse assumption in high dimensional regression.

To do so, a survey of the theoretical results for the lasso and general concave penalised least squares estimators was conducted and a basic summary of their properties was described in Chapter 1. We gave the proofs of a number of properties of the lasso in Chapter 2, including the existence and uniqueness of solutions, sign consistency, asymptotic bounds on parameter and prediction error, and most importantly, the model selection consistency property. In Chapter 3, we considered properties of concave penalised least squares estimators, beginning with an orthogonal analysis of solutions to unveil their behaviour, before moving directly to proving a number of theorems on concave model selection consistency.

Our work in Chapter 2 was essential in laying the groundwork for the proofs presented in Chapter 3. The work of Feng & Zhang in [6] along with the primal-dual-witness method from [7] gave us a generalisable proof technique for showing model selection consistency in regression models. This allowed us to prove the two main theorems of this thesis concerning model selection consistency of concave PLSE, both of which hold with high probability under sufficient regularity conditions. The first theorem states that there always exists a model selection consistent local minimiser of the root-log regularised least squares problem. The second theorem states that algorithms implementing root-log regression will, with high probability, find the model selection consistent local minimiser. This is a theoretical edge over the popular lasso regression, since model selection consistency for the lasso requires the strong irrepresentability condition, which becomes less likely to hold as p and n increase, while the root-log regulariser does not. Indeed, one can view the strong irrepresentability condition as encoding the inability of the lasso to perform well given high collinearity in the data, an issue which concave penalties improve on. Naturally, then, we may expect concave least squares estimators to perform better under high collinearity.

A host of possibilities for future work is available, the most prominent of which is actually testing the proven theoretical properties in real-world applications. The first step would be to implement an algorithm to perform root-log regression. With such an algorithm, the model selection properties should be tested on both simulated and real-world data sets, and the probability that the necessary conditions for model selection consistency hold should be approximated via Monte Carlo simulation. Disparities between the model selection of the lasso and root-log regression should be noted, particularly when p and n are large and when the active and inactive variables are highly correlated.

Appendix A

Appendix

A.1 Basic Properties

Proposition A.1.1 (Gaussian Tail Bounds)**.** For a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, we have the following tail bound:

$$\mathbb{P}\left[\frac{z-\mu}{\sigma} \geq z\right] \leq \frac{1}{z}\exp\bigl(-z^2/2\bigr). \tag{A.1.1}$$

Proof. Without loss of generality, assume $Z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}\mathbb{P}(|Z| > z) = 2\,\mathbb{P}(Z > z) &= \frac{2}{\sqrt{2\pi}}\int_z^\infty \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \frac{2}{\sqrt{2\pi}}\int_z^\infty \frac{x}{z}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \frac{2}{\sqrt{2\pi}}\frac{1}{z}\int_z^\infty x\,\mathrm{e}^{-x^2/2}\mathrm{d}x = \frac{2}{\sqrt{2\pi}}\frac{1}{z}\mathrm{e}^{-z^2/2} \leq \frac{1}{z}\mathrm{e}^{-z^2/2}.\end{aligned}$$

□

Definition A.1.1 (Matrix Pseudoinverse)**.** Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. The *pseudoinverse* of $\boldsymbol{A}$ is the unique matrix $\boldsymbol{A}^+$ satisfying:

- (Pseudoinverse property)

$$\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = A \text{ and } A^+AA^+ = A^+.$$

- (Hermitian $\boldsymbol{A}^+\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^+$):

$$(AA^+)^* = (AA^+) \text{ and } (A^+A)^* = A^+A.$$

Proposition A.1.2. The matrix $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^+$ is a projection onto Range($\boldsymbol{A}$) satisfying $\boldsymbol{P}A = \boldsymbol{A}$, with the matrix $\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^+$ being the orthogonal projection onto $\ker(\boldsymbol{A})$ satisfying $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{A} = 0$.

Definition A.1.2 (Affine span)**.** The affine span of a set X in a vector space V is given by the set

$$\mathrm{Aff}(X) = \left\{ \sum_{i=1}^k \alpha_i x_i \mid k > 0, x_i \in X, \alpha_i \in \mathbb{R}, \sum_{i=1}^k \alpha_i = 1. \right\}.$$

A.2 Subgradient Calculus

Throughout this thesis, we made use of the subgradient calculus to find sufficient conditions for local minimisers of penalised least squares optimisation problems. In this section, we briefly give the main definitions and results for convenience. The source for these statements are the notes given in [2].

Definition A.2.1. A vector $\boldsymbol{z} \in \mathbb{R}^n$ is a *subgradient* of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x} \in \mathrm{dom}(f)$ if, for all $\boldsymbol{x}' \in \mathrm{dom}(f)$,

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \boldsymbol{z}^\top(\boldsymbol{x}' - \boldsymbol{x}). \tag{A.2.1}$$

The function f is *subdifferentiable at* $\boldsymbol{x} \in \mathrm{dom}(f)$ if there exists at least one subgradient at $\boldsymbol{x}$. The function f is *subdifferentiable* if f is subdifferentiable at all $\boldsymbol{x} \in \mathrm{dom}(f)$. The set of all subgradients of f at $\boldsymbol{x}$ is denoted $\partial f(\boldsymbol{x})$.

Proposition A.2.1. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. If f is differentiable at $\boldsymbol{x} \in \mathrm{dom}(f)$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. Conversely, if $\partial f(\boldsymbol{x}) = \{\boldsymbol{z}\}$ for some $\boldsymbol{z} \in \mathbb{R}^n$, then f is differentiable at $\boldsymbol{x}$ with gradient $\boldsymbol{z} = \nabla f(\boldsymbol{x})$.

Theorem A.2.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a (not necessarily convex) function. Then $\boldsymbol{x}' \in \mathrm{dom}(f)$ is a global minimiser of f if and only if f is subdifferentiable at $\boldsymbol{x}'$ and $0 \in \partial f(\boldsymbol{x}')$.

That is, subgradients determine global minimality. The issue lies in finding the form of the subgradient $\nabla f(\boldsymbol{x})$: this is the focus of the *subgradient calculus*. Even though the above theorem applies even when f is not convex, the subgradients of non-convex functions are difficult to obtain, and so in the following results, we shall describe some of the basic methods for finding subgradients of *convex functions*:

Proposition A.2.2 (Basic Properties)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $\boldsymbol{x} \in \mathrm{dom}(f)$. Then the following properties apply to the subgradient $\partial f(\boldsymbol{x})$:

(i) (Non-negative scaling): For $\alpha \geq 0$, we have $\partial(\alpha f)(\boldsymbol{x}) = \alpha \partial f(\boldsymbol{x})$.

(ii) (Sums): If $f = \sum_{i=1}^m f_i$ with f_i convex and $\boldsymbol{x} \in \mathrm{dom}(f) = \cap_{i=1}^m \mathrm{dom}(f_i)$, then

$$\partial f(\boldsymbol{x}) = \sum_{i=1}^m \partial f_i(\boldsymbol{x}).$$

(iii) (Linear chain rule): If $h(\boldsymbol{x}) := f(A\boldsymbol{x} + \boldsymbol{b})$ for a matrix $A \in \mathbb{R}^{n \times n}$, then

$$\partial h(\boldsymbol{x}) = A^\top \partial f(A\boldsymbol{x} + \boldsymbol{b})$$

(iv) (Pointwise maximum): If $f_1, \ldots, f_m$ are convex, subdifferentiable functions and

$$f(\boldsymbol{x}) := \max_{i=1,\ldots,m} f_i(\boldsymbol{x}),$$

then

$$\partial f(\boldsymbol{x}) = \mathrm{Co}\left(\cup_{i=1}^m \{\partial f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\}\right),$$

where $\mathrm{Co}(\cdot)$ denotes the convex hull of a set $\cdot$. That is, the subdifferential of a pointwise maximum of a finite set of functions is the convex hull of the union of the *active subdifferentials* (where by active we mean the functions that actually attain the maximum at $\boldsymbol{x}$).

Example A.2.1. With the last of these properties, we can find the subdifferential of the L^1-norm

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|,$$

which is essential for our analysis of the lasso in Chapter 2. Indeed, we can write

$$\|\boldsymbol{x}\|_1 = \max\left\{ \boldsymbol{s}^\top \boldsymbol{x} \mid s_i \in \{-1, 1\} \right\},$$

which is the maximum of 2^n linear functions. In order for a choice of $\boldsymbol{s} \in \{-1, 1\}^n$ to give an active function $\boldsymbol{s}^\top \boldsymbol{x}$ achieving maximality, we pick

$$s_i = \begin{cases} +1 & \text{if } x_i > 0, \\ -1 & \text{if } x_i < 0, \\ +1 \text{ or } -1 & \text{if } x_i = 0. \end{cases}$$

Then by Proposition A.2.2 above, the subdifferential is given by

$$\begin{aligned} \partial \|\boldsymbol{x}\|_1 &= \mathrm{Co}\left(\cup_{i=1}^{m} \{ \partial(\boldsymbol{s}^\top \boldsymbol{x}) \mid \boldsymbol{s}^\top \boldsymbol{x} = \|\boldsymbol{x}\|_1, \boldsymbol{s} \in \{-1\}^n \} \right) \\ &= \boldsymbol{\gamma} \in \mathbb{R}^n, \end{aligned}$$

where

$$\gamma_i \in \begin{cases} \{1\} & \text{if } x_i > 0, \\ \{-1\} & \text{if } x_i < 0, \\ [-1, 1] & \text{if } x_i = 0. \end{cases}$$

The explicit form of the subgradient above justifies why we can choose the value of the subgradient at 0 to be the most advantageous value in a given range. Finally, to study the behaviour of global minimisers of

$$\frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|,$$

we need $\mathbf{0}$ to be in its subgradient. This is attained if and only if

$$n\lambda \boldsymbol{\gamma} = \mathbf{X}^\top (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}).$$

This is exactly the lasso KKT conditions given in (2.3.1).

References

[1] Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001.

[2] Stephen Boyd, John Duchi, Mert Pilanci, and Lieven Vandenberghe. Notes for ee364b, 2022.

[3] Jianqing Fan. Comments on wavelets in statistics: A review by a. antoniadis. *Journal of the Italian Statistical Society*, 6(2):131–138, 1997.

[4] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[5] Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical foundations of data science*. Chapman and Hall/CRC, 2020.

[6] Long Feng and Cun-Hui Zhang. Sorted concave penalized regression. *The Annals of Statistics*, 47(6):3069–3098, 2019.

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143:143, 2015.

[8] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[10] Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

[11] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.