The Graduate University for Advanced Studies,
SOKENDAI

Doctoral Thesis

Sparse Regression for Correlated Variables

Author:
Masaaki Takada

Supervisor:
Dr. Hironori Fujisawa

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Statistical Science
School of Multidisciplinary Sciences

September, 2020

The Graduate University for Advanced Studies, SOKENDAI

Abstract

School of Multidisciplinary Sciences

Department of Statistical Science

Doctor of Philosophy

Sparse Regression for Correlated Variables

by Masaaki TAKADA

Sparse regularization, such as ℓ_1 regularization, is a quite effective technique for high dimensional learning problems. Its effectiveness has been supported empirically and theoretically. However, one of the most significant issues in ℓ_1 regularization is that its performance is quite sensitive to correlations among variables. Typical theoretical supports are based on a kind of small correlation assumptions including the incoherence condition and the restricted eigenvalue condition. Besides, ℓ_1 regularization empirically incurs many false-positive variables, resulting in large estimation errors and low interpretability. In this thesis, we propose a new regularization method, "Independently Interpretable Lasso" (IILasso). We mitigate small correlation assumptions among variables; instead, we impose small correlation assumptions among true active variables. Our proposed regularizer incurs a large penalty for selecting correlated variables, hence the output models have low correlations among active variables. We can intuitively interpret regression coefficients in our model because each active variable affects the response independently. In our theoretical analyses, we show that our method achieves correct sign recovery under the generalized incoherence condition, which is milder than the ordinary incoherence condition. Additionally, we show that our method is beneficial for its estimation error for correlated design because our generalized restricted eigenvalue condition is milder than the ordinary restricted eigenvalue condition. Furthermore, we extend our method and its theoretical results to generalized linear models. Synthetic and real data analyses indicate the effectiveness of our method.

Acknowledgements

I could not have done this work without help from many people.

First and foremost, I would like to say thanks to my supervisor Dr. Hironori Fujisawa. He has understood and motivated me well. His guidance helped me in all the time of research and writing of this thesis. I appreciate all his contributions.

Besides my advisor, I would like to offer my special thanks to Dr. Taiji Suzuki. He was a great collaborator in my research. I learned a lot from his insightful feedback.

I would like to thank my sub-supervisors, Dr. Kenji Fukumizu and Dr. Ryo Yoshida, for their supports. I would also like to thank my examiners, Dr. Yoshiyuki Ninomiya, Dr. Hideitsu Hino, Dr. Shuichi Kawano, and Dr. Masaaki Imaizumi, for their helpful comments.

I would like to thank the members of Fujisawa Lab and SOKENDAI for their stimulation. I would also like to thank my colleagues in my company for their kind consideration.

Lastly, I would like to thank my family for all their love and encouragement. My wife has supported me emotionally and gave me time to study. My children have given me happy and refreshing moments. My parents have given me a lot of support up to this point.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sparse Regression: Strengths and Weaknesses

High dimensional data appear in many fields such as biology, economy, and industry. A common approach for high dimensional regression is a sparse regularization strategy such as the Lasso (Least absolute shrinkage and selection operator) (Tibshirani, 1996). Since the sparse regularization performs both parameter estimation and variable selection simultaneously,

1. it offers *interpretable* results by identifying informative variables,
2. and it can effectively *avoid overfitting* by discarding redundant variables.

Because of these properties, the sparse regularization has had huge success in a wide range of data analysis in science and engineering. In addition, several theoretical studies have been developed to support the effectiveness of sparse regularization, and several optimization methods also have been proposed so that sparse learning is efficiently executed.

One of the significant issues, however, is its performance in the presence of correlated variables. Its performance is theoretically guaranteed only under “small correlation” assumptions that variables are not much correlated with each other. Actually, typical theoretical supports are based on a kind of small correlation assumptions including the incoherence condition (Fuchs, 2005; Tropp, 2006; Wainwright, 2009) and the restricted eigenvalue condition (Bickel, Ritov, Tsybakov, et al., 2009; Bühlmann and Van De Geer, 2011; Hastie, Tibshirani, and Wainwright, 2015). Besides, the Lasso empirically incurs many false-positive variables in the presence of correlated variables, resulting in large estimation errors.

Correlated variables also hinder the interpretation of models. A standard interpretation of a (generalized) linear model comes from *regression coefficients* and *effects* (products of regression coefficients and standard deviations). A single coefficient represents the degree to which an increase of its variable by one unit increases the prediction “when all other active variables remain fixed”. When an output model contains correlated variables, we must consider the influence of all other correlated active variables simultaneously, which can be intractable. A single effect represents the degree to which its variable affects the prediction “on average”. However, the value of the effect may become misleading in the presence of strong correlations because it is unlikely to increase a variable by one standard deviation but not changing other correlated variables. Similar properties were described as disadvantages of linear models in Section 4.1 in Molnar et al. (2018). This kind of interpretability was also referred to as “decomposability” in Lipton (2018), which means the ability of whether we can decompose a model into some parts and interpret each component independently. In this sense, correlated variables degrade the decomposability of the model. Therefore, it is beneficial to construct a model with uncorrelated variables for both estimation error and interpretability.

1.2 Research for Correlated Variabes

Several methods have been proposed to resolve the problem induced by correlations among variables.

One line of research taking correlations among variables into account is based on a strategy in which correlated variables are either all selected or not selected at all. Examples of this line are the Elastic Net (Zou and Hastie, 2005), Pairwise Elastic Net (Lorbert, Eis, Kostina, Blei, and Ramadge, 2010), and Trace Lasso (Grave, Obozinski, and Bach, 2011). These methods select not only important variables but also unimportant variables that are strongly correlated with important variables. Although these methods often give stable generalization error, this strategy makes it hard to interpret the model. This is because the output model incorporates many correlated variables, and their coefficients and effects receive large influence from the number of correlated variables and the degree of correlations.

Another line of research, including ours, is proposed based on the strategy in which uncorrelated variables are selected. The Uncorrelated Lasso (Chen, Ding, Luo, and Xie, 2013) intends to construct a model with uncorrelated variables. However, it still tends to select “negatively” correlated variables, and hence the correlation problem is not resolved. The Exclusive Group Lasso (Kong, Fujimaki, Liu, Nie, and Ding, 2014) is also in this line, but it is necessary to group correlated variables beforehand. They suggest grouping variables whose correlations are higher than a certain threshold. However, determination of the threshold is not a trivial problem, and practically it causes unstable results. Moreover, these methods are lack of theoretical support, including sign recovery and estimation errors.

1.3 Our Contribution

We address the issue that the performance of sparse regression is degraded when the correlation among variables is high. Existing theoretical support is based on "small correlation" assumptions such as incoherence condition and restricted eigenvalue condition. There are many false-positive variables empirically, and it results in large estimation errors in the presence of correlated variables.

We theoretically mitigate "small correlation" assumptions among variables; instead, we impose "small correlation" assumptions among true active variables. Under this assumption, we propose a new regularization method, "Independently Interpretable Lasso" (IILasso). Our proposed regularizer incurs a large penalty for selecting correlated variables; hence the output models have low correlations among active variables. Each active variable affects the response independently in our model so that we can interpret regression coefficients intuitively. Our method offers efficient variable selection, which does not select negatively correlated variables and is free from a specific pre-processing such as grouping.

To support the effectiveness of our proposal, we give the following contributions:

- We show a necessary and sufficient condition for the sign consistency of variables selection. We show that our method achieves the sign consistency under a milder condition than the Lasso for correlated design.
- The convergence rate of the estimation error is analyzed. We show that our estimation error achieves the almost minimax optimal rate and has an advantage over the Lasso.
- We propose a coordinate descent algorithm to find a local optimum of the objective function, which is guaranteed to converge to a stationary point. Additionally, we show that every local optimal solution achieves the same statistical error rate as the global optimal solution and thus is almost minimax optimal.

- We extend linear models to generalized linear models and derive its convergence rate.

1.4 Outline

The rest of the thesis is organized as follows: In Chapter 2, we review sparse regression methods. In Chapter 3, we propose a new regularization method for linear models and introduce its optimization method and theoretical results. In Chapter 4, we extend the IILasso to generalized linear models. In Chapter 5, both synthetic and real-world data experiments, including ten microarray datasets, are illustrated. In Chapter 6, we summarize our thesis.

1.5 Notations

Let $v \in \mathbb{R}^p$. Let $\mathrm{Diag}(v) \in \mathbb{R}^{p \times p}$ be the diagonal matrix whose j-th diagonal element is v_j. Let $|v|$ be the element-wise absolute vector whose j-th element is $|v_j|$. Let $\mathrm{sgn}(v)$ be the sign vector whose elements are 1 for $v_j > 0$, -1 for $v_j < 0$, and 0 for $v_j = 0$. Let $\mathrm{supp}(v)$ be the support set of v, i.e., $\{j \in \{1, \cdots, p\} | v_j \neq 0\}$. Let $\|v\|_q$ be the ℓ_q-norm, i.e., $\|v\|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}$.

Let $M \in \mathbb{R}^{n \times p}$. We use subscripts for the columns of M, i.e., M_j denotes the j-th column. Let $\|M\|_q$ be the operator norm (induced norm), i.e., $\|M\|_q = \sup_{v \in \mathbb{R}^p} \|Mv\|_q / \|v\|_q$. Specifically, $\|M\|_2 = \sup_{v \in \mathbb{R}^p} \|Mv\|_2 / \|v\|_2$ is the spectral norm (the largest singular value of M), and $\|M\|_\infty = \max_i \sum_j |M_{ij}|$ is the maximum absolute column sum norm. Let $\|M\|_{\max}$ be the max norm, i.e., $\|M\|_{\max} = \max_{ij} |M_{ij}|$.

Let $M \in \mathbb{R}^{p \times p}$. Let $M \succeq O$ and $M \succ O$ denote positive semi-definite matrix and positive definite matrix, i.e., $v^\top M v \geq 0$ for $\forall v \in \mathbb{R}^p$ and $v^\top M v > 0$ for $\forall v \in \mathbb{R}^p$ and $v \neq 0$, respectively.

Let S be a subset of $\{1, \cdots, p\}$. Let $|S|$ be the number of the elements in S. Let S^c be the complement subset of S, i.e., $S^c = \{1, \cdots, p\} \backslash S$. Let v_S be the vector v restricted to the index set S. Let $M_{S_1 S_2}$ be the matrix whose row indexes are restricted to S_1 and column indexes are restricted to S_2.

Chapter 2

Sparse Regression

2.1 Problem Setting

Consider a problem of predicting a response $y \in \mathbb{R}^n$, given a design matrix $X \in \mathbb{R}^{n \times p}$, assuming a linear model

$$y = X\beta + \varepsilon, \tag{2.1}$$

where $\beta \in \mathbb{R}^p$ is a regression coefficient, and $\varepsilon \in \mathbb{R}^n$ is a noise. We assume that the variables are standardized such that $\Sigma_{i=1}^n X_{ij} = 0$, $\Sigma_{i=1}^n X_{ij}^2 / n = 1$ and $\Sigma_{i=1}^n y_i = 0$.

The usual least-squares estimator is based on minimizing the squared-error loss

$$\min_\beta \frac{1}{2n} \|y - X\beta\|_2^2.$$

If $p < n$ and X is full column rank, the unique solution is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

The estimation error of the squared-error estimator is known to $O(p/n)$.

On the other hand, if $p > n$, the least-squares estimators are not unique, and there is an infinite set of solutions.

2.2 The Lasso

The Lasso (Tibshirani, 1996) is a standard method for high-dimensional $(n < p)$ data to estimate a sparse model. The Lasso optimizes

$$\min_\beta \; \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \tag{2.2}$$

where $\lambda = \lambda_n$ is a regularization parameter, which is typically determined by cross-validation. We also write $\lambda = \lambda_n$ as we explicitly express its dependence of n. The formulation (2.2) has several attractive properties as follows:

Convexity. The objective function is convex. Convexity also holds for the penalty $\|\beta\|_q^q$ where $q \geq 1$. Simple greedy algorithms converge to a global optimum.

Sparsity. The solution is sparse, that is, a large λ causes some of the coefficients to be exactly zero. Sparsity also holds for the penalty $\|\beta\|_q^q$ where $q \leq 1$.

Interpretability. Sparse (generalized) linear models are easy to interpret.

Algorithm Efficiency. Efficient optimization algorithms are available such as the coordinate descent algorithm.

Statistical Property. Several theoretical analyses support the Lasso. The incoherence condition guarantees sign recovery. The ℓ_2 estimation error is $O(s \log(p)/n)$, where s is the number of true active variables, under the restricted eigenvalue condition. It is almost minimax-optimal.

We explain the above properties.

2.2.1 Convexity

Another form for the Lasso (2.2) is

$$\begin{aligned} &\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2, \\ &\text{s.t. } \|\beta\|_1 \leq t. \end{aligned} \tag{2.3}$$

The constraint region is a convex set, and the objective function is a strongly convex function, so the Lasso is a convex problem. Because the convex problem has no local minima, simple greedy algorithms converge to a global optimum.

The penalty $\|\beta\|_q^q$ for $q > 1$ is also convex. Ridge regression is an example using squared-ℓ_2 penalty:

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where λ is a regularization parameter. The solution is easily obtained as

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y.$$

Ridge regression shrinks the estimate to zero, but does not induce sparsity.

The penalty $\|\beta\|_q^q$ for $q < 1$ is nonconvex. It is hard to optimize nonconvex problems because there exists a local minimum. Best subset selection can be seen as a limit case of $q \to 0$ so that it is called ℓ_0 penalty. This optimization is challenging because of its nonconvexity and discontinuity.

2.2.2 Sparsity

The Lasso solution is sparse. To see this, we consider the ℓ_1 geometry. The constraint region of the Lasso (2.3) is a square, and the contours of the objective function is an ellipse. If their contours hit the corner of the square, the corresponding component of the solution equals to zero, which indicates the sparse solution.

The penalty $\|\beta\|_q^q$ for $q < 1$ also induces sparsity, because its constraint region is sharp on each component. On the other hand, the penalty $\|\beta\|_q^q$ for $q > 1$ does not induce sparsity, because their constraint regions are not sharp for any directions.

The importance of sparsity relies on estimation error and interpretability. Under the assumption that the true signal is sparse, the Lasso can recover the true signal. Besides, it can estimates coefficients as if we know true active variables in advance. Additionally, the output model is easy to interpret because the Lasso eliminates uninformative variables and selects only a few informative variables.

2.2.3 Interpretability

Interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019). It has received much attention from the machine learning community in recent years (Doshi-Velez and Kim, 2017; Miller, 2019; Molnar, 2020). There are many reasons why interpretability matters. One representative reason is that problem formulation using a single metric, such as a risk function, is not enough to describe the original real-world problem. Thus, we need to understand the behavior of the model for cofirming safety, detecting bias, social acceptance, debugging, and auditing for example.

Methods for interpretability can be classified as intrinsic methods and post hoc methods. Intrinsic methods refer to exploiting interpretable models due to their simple structure, while post hoc interpretation methods refer to the application of interpretation methods to complicated trained models. Intrinsic methods are effective especially for scientific knowledge discovery because there is almost no gap between the model and interpretation (scientific discovery). Intrinsic methods includes linear regression models, logistic regression models, generalized linear models, general additive models, decision trees, k-nearest neighbors, rule based learners, and Bayesian models.

The Lasso is an intrinsically interpretable model and has some superiorities in its model transparency. First, the Lasso has high simulatability, that is, it is easy to simulate a model by a human. This is because the model is a linear function and it includes only a small number of variables. Second, it has decomposability, that is, each of the parts of a model can be interpreted. This is because its coefficients represents the association between variables and responses. However, this is not the case under high correlations between active variables, as described in Chapter 3.2 and our proposed method mitigates this drawback. Third, it also has algorithmic transparency, that is, the optimization algorithms are intuitive and easy to interpret because it converges to a unique solution.

2.2.4 Optimization

Several optimization algorithms can solve the Lasso problem. In this subsection, we introduce four algorithms: proximal gradient descent, coordinate descent, LAR, and ADMM.

Proximal Gradient Descent

Proximal gradient descent is a basic algorithm for nondifferentiable function optimization. Let an objective function can be decomposed as $f = g + h$ where g is convex and differentiable, and h is convex but nondifferentiable. Then, proximal gradient descent update forms

$$\beta^{t+1} \leftarrow \mathrm{argmin}_{\beta} \left\{ g(\beta^t) + \langle \nabla g(\beta^t), \beta - \beta^t \rangle + \frac{1}{2s^t} \|\beta - \beta^t\|_2^2 + h(\beta) \right\},$$

where s^t is a stepsize. The stepsize can be either taken any small fixed constant or chosen by backtracking line search. Taking $g = \|y - X\beta\|_2^2/2n$ and $h = \lambda\|\beta\|_1$, we have the proximal descent update for the Lasso (Daubechies, Defrise, and De Mol, 2004) as

$$\beta_j^{t+1} \leftarrow \mathcal{S}\left(\beta_j^t + s^t \frac{1}{n} X_j^\top (y - X\beta^t), s^t\lambda \right),$$

for $j = 1, \ldots, p$, where $\mathcal{S}(z, \gamma)$ is a soft thresholding function

$$\begin{aligned} \mathcal{S}(z,\gamma) :&= \operatorname{sgn}(z)(|z| - \gamma)_+ \\ &= \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z|, \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z|, \\ 0 & \text{if } |z| \leq \gamma. \end{cases} \end{aligned} \tag{2.4}$$

Furthermore, Nesterov's acceleration scheme (Nesterov, 2013) can be used as

$$\begin{aligned} \beta^{t+1} &\leftarrow \mathcal{S}\left(\theta^t + s^t \frac{1}{n} X^\top (y - X\theta^t), s^t \lambda\right), \\ \theta^{t+1} &\leftarrow \beta^{t+1} + \frac{t}{t+3}(\beta^{t+1} - \beta^t). \end{aligned}$$

This is essentially equivalent to FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) (Beck and Teboulle, 2009).

Coordinate Descent

For high-dimensional data, simultaneous update for all components of coefficients needs high computation cost. The coordinate descent algorithm (Friedman, Hastie, Höfling, Tibshirani, et al., 2007; Friedman, Hastie, and Tibshirani, 2010) updates a single coordinate at a single iteration. Specifically, the cyclic coordinate descent algorithm forms for $j = 1, \ldots, p, 1, \ldots, p, \ldots,$

$$\begin{aligned} \beta_j^{t+1} &\leftarrow \operatorname{argmin}_{\beta_j} f(\beta_1^t, \ldots, \beta_{j-1}^t, \beta_j, \beta_{j+1}^t, \ldots, \beta_p^t), \\ \beta_k^{t+1} &\leftarrow \beta_k^t \quad \text{for} \quad k \neq j. \end{aligned}$$

For the Lasso, differentiating the objective function (2.2) with respect to β_j yields

$$-\frac{1}{n} X_j^\top (y - X_{-j}\beta_{-j}) + \beta_j + \lambda \operatorname{sgn}(\beta_j),$$

where β_{-j} denotes β without the j-th component, and X_{-j} denotes X without j-th column. Hence, we obtain the update rule as

$$\beta_j \leftarrow \mathcal{S}\left(\frac{1}{n} X_j^\top (y - X_{-j}\beta_{-j}), \lambda\right),$$

where $\mathcal{S}(z, \gamma)$ is a soft thresholding function (2.4). Several implementation techniques are effective for the computational cost, including covariance updating, warm-start, active-set convergence, strong-set convergence, and safe screening. See (Hastie, Tibshirani, and Wainwright, 2015) for example.

LAR

Least angle regression (LAR) (Efron, Hastie, Johnstone, Tibshirani, et al., 2004) is a specific algorithm for the Lasso with squared-error loss. It can provide exact entire solution paths for regularization parameters. At first step, it finds the variable most correlated with the response. Then, it moves its coefficient continuously to the least-squares estimate. When another variable catches up in terms of correlation with the residual, it enters the active set, and move together to the least-squares estimate keeping their correlations equally. This process is continued until it ends at the full least-squares estimate.

ADMM

The alternating direction method of multipliers (ADMM) (Boyd, Parikh, Chu, Peleato, Eckstein, et al., 2011) is an augmented Lagrangian based approach. It is applicable to a wide range of optimization problems, including the Lasso. Consider an optimization problem such that

$$\min_{\beta,\theta} \; f(\beta) + g(\theta), \quad \text{s.t.} \quad A\beta + B\theta = c,$$

where $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are convex functions, and $A \in \mathbb{R}^{n\times d}, B \in \mathbb{R}^{n\times d}$, and $c \in \mathbb{R}^d$. Then, the augmented Lagrangian is

$$L_\rho(\beta, \theta, \mu) := f(\beta) + g(\theta) + \langle \mu, A\beta + B\theta - c\rangle + \frac{\rho}{2}\|A\beta + B\theta - c\|_2^2.$$

The ADMM algorithm forms

$$\begin{aligned}
\beta^{t+1} &\leftarrow \text{argmin}_\beta L_\rho(\beta, \theta^t, \mu^t) \\
\theta^{t+1} &\leftarrow \text{argmin}_\theta L_\rho(\beta^{t+1}, \theta, \mu^t) \\
\mu^{t+1} &\leftarrow \mu^t + \rho(A\beta^{t+1} + B\theta^{t+1} - c).
\end{aligned}$$

Taking $f = \|y - X\beta\|_2^2/2n$ and $g = \lambda\|\beta\|_1$, we have the ADMM algorithm for the Lasso as

$$\begin{aligned} \beta^{t+1} &\leftarrow (X^\top X + \rho I)^{-1}(X^\top y + \rho\theta^t - \mu^t) \\ \theta_j^{t+1} &\leftarrow \mathcal{S}\left(\beta_j^{t+1} + \frac{\mu^t}{\rho}, \frac{\lambda}{\rho}\right) \quad \text{for} \quad j = 1, \dots, p \\ \mu^{t+1} &\leftarrow \mu^t + \rho(\beta^{t+1} - \theta^{t+1}). \end{aligned}$$

2.2.5 Theoretical Properties

The Lasso has a large amount of theoretical support. In this subsection, we review two primary theoretical results; sign recovery and estimation error. Consider a linear model (2.1) and the Lasso (2.2).

Sign Recovery

Sign recovery refers to whether estimators can estimate correct signs (positive/zero/negative) for all coefficients. The Lasso has a simple condition for correct sign recovery. This result follows straightforwardly from optimality conditions for convex programs (Wainwright, 2009).

Theorem 1. *Assume $X_S^\top X_S/n$ is invertible. Then, there exists a solution $\hat{\beta}$ of (2.2) with correct sign recovery $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)$ if and only if the following two conditions hold:*

$$\begin{aligned} \operatorname{sgn}\left(\beta_S^* - \left(\frac{1}{n}X_S^\top X_S\right)^{-1}\left(\lambda \operatorname{sgn}(\beta_S^*) - \frac{1}{n}X_S^\top \varepsilon\right)\right) &= \operatorname{sgn}(\beta_S^*), \\ \left|\frac{1}{n}X_{S^c}^\top X_S\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\left(\lambda \operatorname{sgn}(\beta_S^*) - \frac{1}{n}X_S^\top \varepsilon\right) + \frac{1}{n}X_{S^c}^\top \varepsilon\right| &\leq \lambda, \end{aligned}$$

where both of these vector inequalities are taken elementwise.

Furthermore, the Lasso yields correct sign recovery with high probability under the assumption of sub-Gaussian noise (Assumption 1 in Section 3.5). This result is the same as (Wainwright, 2009), although the parameterizations of the following are different from the original one. Our theoretical results include the result as a particular case.

Definition 1 (Incoherence Condition and beta-min Condition)**.** *We say that the incoherence condition holds if there exists some incoherence parameter $\kappa \in (0, 1]$*

such that

$$\left\| \frac{1}{n} X_j^\top X_S \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \right\|_1 \leq 1 - \kappa,$$

for $\forall j \in S^c$. *We say that the beta-min condition holds if*

$$\beta^*_{\min} := \min_{j \in S} |\beta^*_j| > \lambda_n \left(\left\| \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \right\|_\infty + \frac{4\sigma}{\sqrt{\varphi}} \right).$$

Theorem 2. *Suppose that Assumptions 1 (sub-Gaussian noise) and Definition 1 (the incoherence condition and the beta-min condition) with a constant* $\kappa \in (0, 1]$ *are satisfied. Suppose that there exists a constant* $\varphi > 0$ *such that* $\frac{1}{n} X_S^\top X_S \succeq \varphi I$. *Suppose that the regularization parameter satisfies*

$$\lambda_n \geq \max \left\{ \frac{1}{4}, \frac{2\sigma}{\kappa} \right\} \sqrt{\frac{2 \log(2p/\delta)}{n}}.$$

Then, there exists a solution $\hat{\beta}$ *of* (2.2) *with correct sign recovery* $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)$ *with probability at least* $1 - 2\delta$.

Estimation Error

Estimation errors refer to the distances between estimates and true coefficients. There are various conditions for estimation errors. The following is one of the most straightforward conditions.

Definition 2 (Restricted Eigenvalue Condition). *Let a set of vectors* $\mathcal{B}(S, C)$ *be*

$$\mathcal{B}(S, C) := \Big\{ v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leq C \|v_S\|_1 \Big\}.$$

We say that the restricted eigenvalue condition holds if we have $\phi_{\mathrm{RE}} > 0$ *where*

$$\phi_{\mathrm{RE}} = \phi_{\mathrm{RE}}(S, C) := \inf_{v \in \mathcal{B}(S, C)} \frac{v^\top \frac{1}{n} X^\top X v}{\|v\|_2^2}.$$

ℓ_2 error of the Lasso is bounded by parameters s, λ_n, and ϕ_{RE} with high probability. This result is essentially the same as the existing results (Bickel, Ritov, Tsybakov, et al., 2009; Bühlmann and Van De Geer, 2011). Our theoretical results include the result as a particular case.

Theorem 3. *Suppose that Assumption 1 is satisfied. Suppose that the regularization parameters satisfy*

$$\lambda_n \geq 3\sigma\sqrt{\frac{2\log(2p/\delta)}{n}},$$

and that Assumption RE$(S, 2)$ (Definition 2) is satisfied. Then, it holds that

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{64 s\lambda_n^2}{9\phi_{\mathrm{RE}}(S,2)^2},$$

with probability at least $1 - \delta$.

Different constants give different statements. For example, we can obtain

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{9 s\lambda_n^2}{\phi_{\mathrm{RE}}(S,3)^2},$$

with probability at least $1 - \delta$, assuming

$$\lambda_n \geq 2\sigma\sqrt{\frac{2\log(2p/\delta)}{n}},$$

and Assumption RE$(S, 3)$ (Definition 2).

We note that the convergence rate is roughly evaluated as

$$\|\hat{\beta} - \beta^*\|_2^2 = O_p\left(\frac{s\log(p)}{n}\right),$$

by taking $\lambda_n = O_p(\sqrt{\log p/n})$, which is almost the minimax optimal rate (Raskutti, Wainwright, and Yu, 2011).

2.3 Beyond the Lasso

These theoretical analyses, as well as empirical results, show the strengths and weaknesses of the Lasso. We point out three major issues of the Lasso.

1. The Lasso has estimation bias induced by the ℓ_1-norm. Some methods have been proposed to aim at low bias estimators. The examples include SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010), Adaptive Lasso (Zou, 2006), relaxed Lasso (Meinshausen, 2007; Hastie, Tibshirani, and Tibshirani, 2017), and others.

2. The Lasso is unstable in the presence of correlations among features. Several stable estimators have been proposed, such as Elastic Net (Zou and

Hastie, 2005), Pairwise Elastic Net (Lorbert, Eis, Kostina, Blei, and Ramadge, 2010), and Trace Lasso (Grave, Obozinski, and Bach, 2011).

3. The Lasso typically includes many active variables, which results in many false positives. Some methods take correlations among variables into account to obtain more sparse solutions. These include Exclusive Group Lasso (Kong, Fujimaki, Liu, Nie, and Ding, 2014), Uncorrelated Lasso (Chen, Ding, Luo, and Xie, 2013), and our proposed method in this paper.

Various methods for high-dimensional data have been motivated by these issues. We introduce them in the following subsections.

2.3.1 Low Bias Estimators

The Lasso has estimation bias because it imposes ℓ_1 norm as a penalty of the objective function. An ideal alternative is ℓ_0 norm; that is, the penalty proportional to the number of active variables. This is also called best subset selection. In general, ℓ_0 minimization problems are known to be NP-hard (Natarajan, 1995), so that the computational cost is problematic. There are several heuristic approaches, such as forward (Efroymson and Ray, 1966; Draper and Smith, 1966), or backward selection, but they do not converge to a global optimum. Recently, best subset selection via a mixed integer optimization problem has been proposed (Bertsimas, King, and Mazumder, 2016). This method can solve a global optimum efficiently, but it is still too slow for thousands or more variables (Hastie, Tibshirani, and Tibshirani, 2017).

An intermediate formulation between ℓ_0 and ℓ_1 norm is ℓ_q penalty with $0 < q < 1$:

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q.$$

Although this is nonconvex, some greedy algorithms have been proposed such as Iteratively Reweighted ℓ_1 minimization (Gasso, Rakotomamonjy, and Canu, 2009; Zou and Li, 2008), Iteratively Reweighted Least Squares (Rao and Kreutz-Delgado, 1999; Gorodnitsky and Rao, 1993; Gorodnitsky and Rao, 1997), and Iteratively Thresholding Method (She et al., 2009).

Some clipped penalties are proposed to reduce estimation bias. SCAD (Fan and Li, 2001) uses a smoothly clipped penalty with a linear, quadratic, and constant term as

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \sum_{j=1}^{p} \rho(\beta_j; \lambda, \gamma),$$

$$\rho(\theta; \lambda, \gamma) = \begin{cases} \lambda\theta & \text{if } \theta \leq \lambda \\ \dfrac{2\lambda\gamma\theta - \theta^2 - \lambda^2}{2(\gamma - 1)} & \text{if } \lambda \leq \theta \leq \gamma\lambda \\ \dfrac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)} & \text{if } \theta \geq \gamma\lambda. \end{cases}$$

Similarly, MCP (Zhang et al., 2010) uses a quadratic and constant term as

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \sum_{j=1}^{p} \rho(\beta_j; \lambda, \gamma),$$

$$\rho(\theta; \lambda, \gamma) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2\gamma} & \text{if } \beta \leq \gamma\lambda \\ \dfrac{1}{2}\gamma\lambda^2 & \text{if } \beta \geq \gamma\lambda. \end{cases}$$

SCAD and MCP are also nonconvex, but some greedy algorithms have been proposed such as (Breheny and Huang, 2011; Loh and Wainwright, 2015). Moreover, estimation errors for every local minimum are theoretically quantified (Loh and Wainwright, 2015).

There are many other low bias estimator such as Relaxed Lasso (Meinshausen, 2007; Hastie, Tibshirani, and Tibshirani, 2017), capped-ℓ_1 (Zhang, Zhang, et al., 2012), and Adaptive Lasso (Zou, 2006).

2.3.2 Stable Estimators

If there are strong correlations among variables, the Lasso solutions tend to be unstable. In particular, identical variables lose the uniqueness of the Lasso solution; if $\hat{\beta}$ is a Lasso solution and $X_j = X_k$, then $\hat{\beta}'$ is another solution of the Lasso objective function, where

$$\hat{\beta}' = \begin{cases} \hat{\beta}_l & \text{if } l \neq j \text{ and } l \neq k \\ (\hat{\beta}_j + \beta_k)u & \text{if } l = j \\ (\hat{\beta}_j + \beta_k)(1 - u) & \text{if } l = k, \end{cases}$$

for any $u \in [0, 1]$. This fact implies that high correlations flatten the Lasso objective function around minimizer, and the optimal solution can vary drastically with a small noise.

Elastic Net (Zou and Hastie, 2005) is a representative method that yields both sparse and stable solutions. The objective function is constructed by squared ℓ_2 penalty in addition to ℓ_1 penalty:

$$\min_{\beta} \ \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2.$$

Elastic Net is stable because it has a grouping effect. Formally, it is shown that the estimate of Elastic Net $\hat{\beta}$ satisfies

$$|\hat{\beta}_j - \hat{\beta}_k| \leq \frac{\|y\|_1}{\lambda_2}\sqrt{2(1 - X_j^\top X_k/n)}.$$

Hence, if the variables X_j and X_k are strongly correlated tending to be 1, then estimates of β_j and β_k get closer. This effect contributes to stable solutions.

There is some research in this direction. Pairwise Elastic Net (Lorbert, Eis, Kostina, Blei, and Ramadge, 2010) is

$$\min_{\beta} \ \|y - X\beta\|_2^2 + \lambda\left(\|\beta\|_2^2 + (1-\gamma)\|\beta\|_1^2 - (1-\gamma)|\beta|^\top R|\beta|\right), \ R_{jk} = \frac{1}{n}\left|X_j X_k\right|.$$

An appropreate choice of γ yields the convexity of the problem. Trace Lasso (Grave, Obozinski, and Bach, 2011) is

$$\min_{\beta} \ \|y - X\beta\|_2^2 + \lambda\left\|X\mathrm{Diag}(\beta)\right\|_*,$$

where M_* denotes the trace norm, that is, the sum of the singular values of the matrix M. This formulation is also convex. Pairwise Elastic Net and Trace Lasso take advantage of the correlations among variables to add strong convexity exactly in the directions where needed, while Elastic Net blindly adds squared ℓ_2 norm in every direction.

One of the problems, when we use these kinds of stable methods, is that they tend to include many correlated variables, and each coefficient no longer indicates an independent variable contribution to the response. As a result, it is hard to understand which variables are truly active and how variables affect the objective variable.

2.3.3 Low Correlation Estimators

The Lasso typically includes many active variables for correlated data, which results in many false positives. Some methods take correlations among variables into account to exclude redundant variables.

The Uncorrelated Lasso (ULasso) (Chen, Ding, Luo, and Xie, 2013) aims to reduce correlations among active variables. It optimizes the following objective function:

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\beta^\top R\beta, \tag{2.5}$$

where $R \in \mathbb{R}^{p\times p}$ with each element $R_{jk} = (\frac{1}{n}X_j^\top X_k)^2$. Although they intended to exclude correlated variables, we found that the ULasso does not necessarily select uncorrelated variables. For example, consider the case $X = [X_1, X_2]$. The last term of (2.5) is $\lambda_2(\beta_1^2 + \beta_2^2 + 2R_{12}\beta_1\beta_2)$. If $R_{12} \neq 0$, then the term $R_{12}\beta_1\beta_2$ encourages $|\beta_1\beta_2|$ larger with $\beta_1\beta_2 < 0$. This example implies that the ULasso tends to select correlated variables and set coefficients to the opposite sign. In particular, X_1 and X_2 are strongly correlated, then it reduces $\lambda_2(\beta_1 + \beta_2)^2$, which induces $\beta_1 = -\beta_2$. It is not a significant problem when X_1 and X_2 are positively correlated but is a significant problem when X_1 and X_2 are negatively correlated.

The Exclusive Group Lasso (EGLasso) (Kong, Fujimaki, Liu, Nie, and Ding, 2014) is also the same direction of exclusive selection. It optimizes the following objective function:

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\sum_{k=1}^{K}\|\beta^{(k)}\|_1^2, \tag{2.6}$$

where $\beta^{(k)}$ consists of the variables of β within a group of predictors $g_k \subset \{1, \cdots, p\}$ and K is the number of groups. The last ℓ_1/ℓ_2 penalty term acts on exclusive variable selection. For example, when $p = 3, g_1 = \{1, 2\}$ and $g_2 = \{3\}$, then the last term becomes $\lambda_2((|\beta_1| + |\beta_2|)^2 + |\beta_3|^2)$. This penalty enforces sparsity over each intra-group. They suggest putting highly correlated variables into the same group to select uncorrelated variables. They use $|r_{ij}| > \theta$ with $\theta \approx 0.90$ as a threshold.

The idea of uncorrelated variable selection is widely used in a feature selection framework. For example, mRMR (Ding and Peng, 2005) selects variables by maximizing relevance between X_j and y and minimizing redundancy among X_j's. They used mutual information instead of correlation, but the idea is quite similar.

We note that the authors above did not refer to the aspect of interpretability and theoretical properties. From the viewpoint of interpretability, we claim that we can easily interpret the model with an uncorrelated model. From the theoretical viewpoint, we can relax other assumptions, including the incoherence condition and the restricted eigenvalue condition, in exchange for the additional assumption that the true active variables are uncorrelated.

Chapter 3

Independently Interpretable Lasso

3.1 Proposed Method

One of the significant issues of sparse regularization is its performance and interpretability in the presence of correlated variables, as described in the introduction. To overcome this problem, we propose a new regularization formulation as follows:

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right) =: \mathcal{L}(\beta), \tag{3.1}$$

where $\alpha > 0$ is a regularization parameter for the new regularization term, and $R \in \mathbb{R}^{p\times p}$ is a symmetric matrix whose component $R_{jk} \geq 0$ is a monotonically increasing function of the absolute correlation $r_{jk} = \frac{1}{n}|X_j^\top X_k|$ for $j \neq k$. Some concrete definitions of R are described later. The last term $(\lambda\alpha/2)|\beta|^\top R|\beta| = (\lambda\alpha/2)\sum_{j=1}^p \sum_{k=1}^p R_{jk}|\beta_j||\beta_k|$ is an additional term to the Lasso. Since R_{jk} represents the similarity between X_j and X_k, correlated variables are hard to be selected simultaneously in our formulation. In particular, when X_j and X_k are strongly correlated, the squared error does not change under the condition that $\beta_j + \beta_k$ is constant, but the penalty $R_{jk}|\beta_j||\beta_k|$ strongly induces either $\beta_j = 0$ or $\beta_k = 0$. On the contrary, if X_j and X_k are uncorrelated, i.e., R_{jk} is small, then the penalty of selecting both β_j and β_k is negligible, and it reduces to the ordinary Lasso formulation. Hence, our formulation has exclusive effect only on correlated variables.

We can consider some definition variations of the similarity matrix R. One of the natural choices is $R_{jk} = r_{jk}^2$. R is positive semidefinite in this case because the Hadamard product of positive semidefinite matrices is also positive semidefinite. Hence, the problem (3.1) turns to be convex and easy to solve the global optimal solution. However, it may not reduce correlations enough. Yet another choice is $R_{jk} = |r_{jk}|$, which reduces correlations more strongly. Another

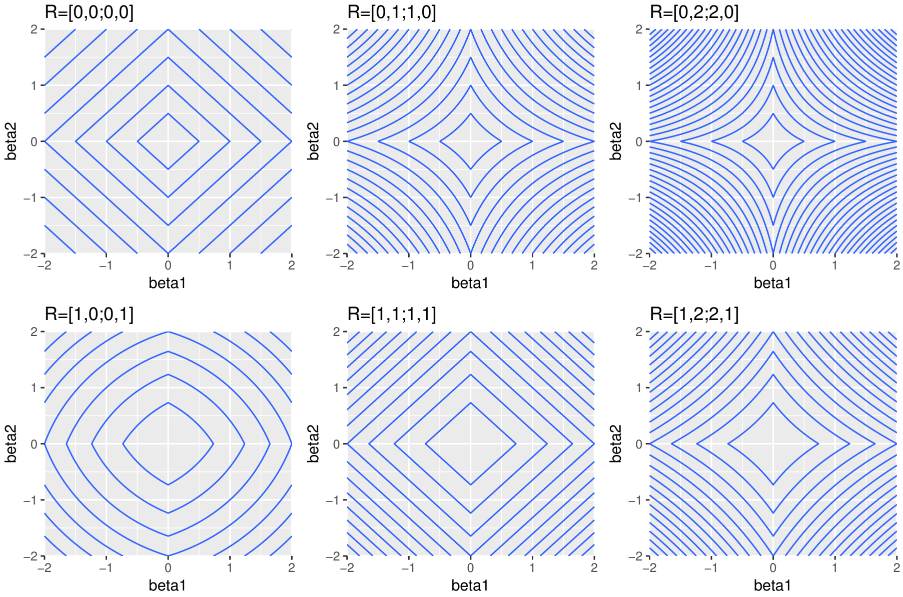

Figure 3.1: Contours of IILasso regularization terms.

effective choice is $R_{jk} = |r_{jk}|/(1-|r_{jk}|)$ for $j \neq k$, and $R_{jk} = 0$ for $j = k$. In this case, if a correlation between two certain variables becomes higher, i.e., $r_{jk} \to 1$, then the penalty term diverges infinitely, and the IILasso cannot simultaneously select both of them. We use the last one in our numerical experiments, because it is favorable from theoretical studies.

Constraint regions corresponding to our regularization term indicate how our method incurs exclusive effect. Figure 3.1 illustrates the constraint regions of $\|\beta\|_1 + |\beta|^T R|\beta|/2$ for the case $p = 2$. As diagonal elements of R increases (from the top to the bottom panel), the contours become smooth at the axes of coordinates. Because of this, the solution tends to select both variables if two variables are strongly correlated. This is the grouping effect of the Elastic Net, as we will describe later. On the other hand, as off-diagonal elements of R increase (from the left to the right panel), the contours become pointed at the axes of coordinates, and the solution tends to be sparser. This is the exclusive effect for correlated variables. The ℓ_q $(0 < q < 1)$ penalty, SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010), and other methods also have the exclusive nature among variables to obtain sparse solutions, as their contours are pointed at the axes; however, the contours of the IILasso is adaptive for correlations among variables, so that our penalty achieves both sparse and stable solutions.

We show contours of various regularization terms in Figure 3.2 including the SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010), ℓ_q-norm, log penalty (Candes, Wakin, and Boyd, 2008), and ordered weighted ℓ_1-norm (Bogdan, Van Den

Berg, Sabatti, Su, and Candès, 2015; Figueiredo and Nowak, 2016; Zeng and Figueiredo, 2014). We compare them with the IILasso (Figure 3.1) and summarize their properties as follows:

- The shapes of the SCAD and MCP are similar to each other; they behave like the Lasso for large γ or small magnitude of β, while the contours are pointed at the axes and parallel to the axes for small γ and large magnitude of β. Their shapes are adaptive for the magnitude of β depending on their hyper-parameter γ, but they are not adaptive for correlations among variables in contrast to the IILasso.
- The shapes of ℓ_q-norm for $0 < q < 1$ are pointed at the axes so that they resemble that of the IILasso for correlated variables.
- The log penalty looks similar to ℓ_q-norm for $0 < q < 1$, as well as the IILasso for correlated variables.
- The ordered weighted ℓ_1-norm is similar to the Elastic Net, but its contours are pointed at $|\beta_1| = |\beta_2|$, resulting in grouping effect.

The exclusive effect among correlated variables offers great advantages. First, it produces uncorrelated models that are easy to interpret, as described in the next section. Second, it is favorable for sign recovery and estimation error under the assumption that the truth is "well interpretable" as described in our theoretical and numerical analyses.

3.2 Interpretability

A linear model looks as if it could be perfectly interpreted, but it is not always the case. Here, we discuss the difficulties of linear model interpretation and the advantages of our proposed method. It is noted that similar difficulties were referred to in (Lipton, 2018) and Section 4.1 in (Molnar et al., 2018), but they did not offer any workaround.

One usually understands a linear model through its regression coefficients and effects. A single coefficient represents how strongly a unit change of a variable affects the prediction under the condition that "when all other active variables remain fixed". A single effect is calculated by a regression coefficient times a standard deviation of each variable so that it represents the degree to which a variable affects the prediction "on average." Coefficients and effects coincide if variables are standardized in advance.

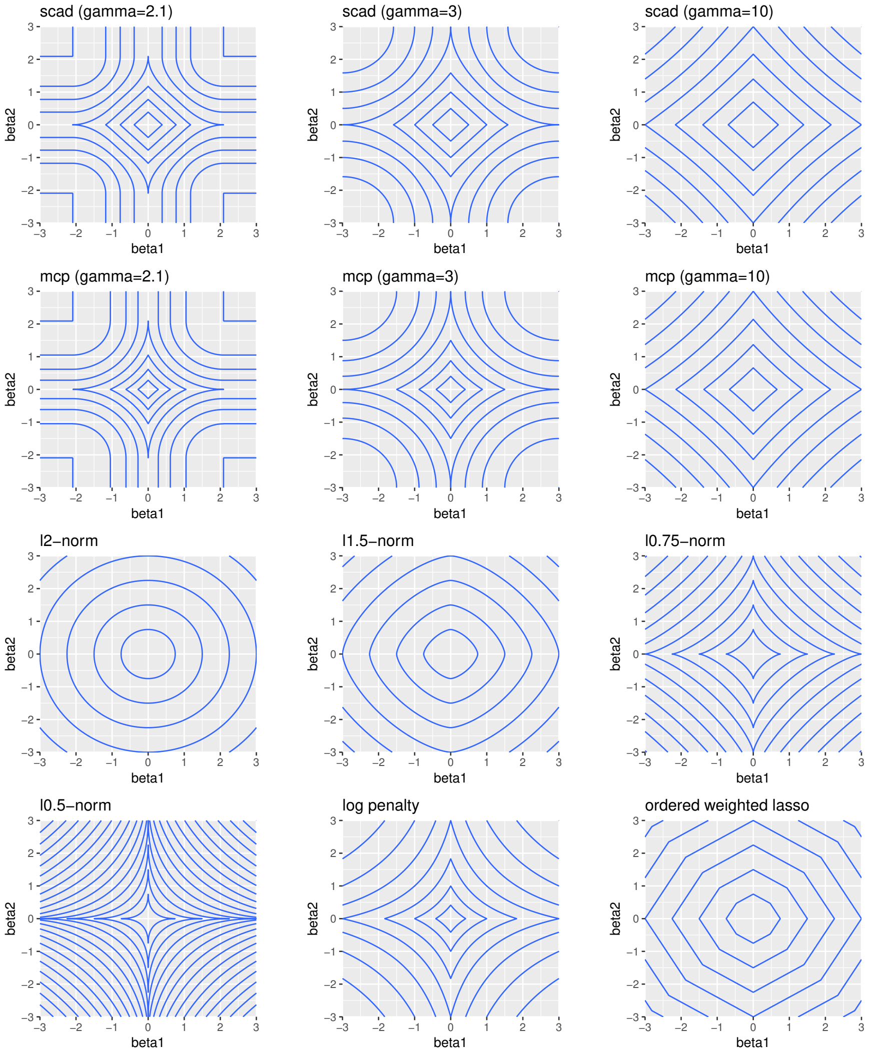

FIGURE 3.2: Contours of various regularization terms.

Correlated variables in an output model make interpretation of coefficients and effects harder. We must simultaneously consider the influence of all other correlated active variables when the model contains correlated variables. Neglecting other active variables may lead to misinterpretation because increasing a variable by one standard deviation but not changing others can be counterfactual. On the other hand, when the model does not contain correlated variables, the footnote that "when all other active variables remain fixed" is a reasonable assumption.

As an example, Figure 3.3 shows scatter plots of uncorrelated and correlated variables in output models. Two variables X_1 and X_2 are mean 0 and standard deviation 1 in both cases, but correlations among them differ from each other. In the uncorrelated case, increasing X_1 by 1 (standard deviation) but not changing X_2 is reasonable, because fixing X_2 does not affect the range of X_1.

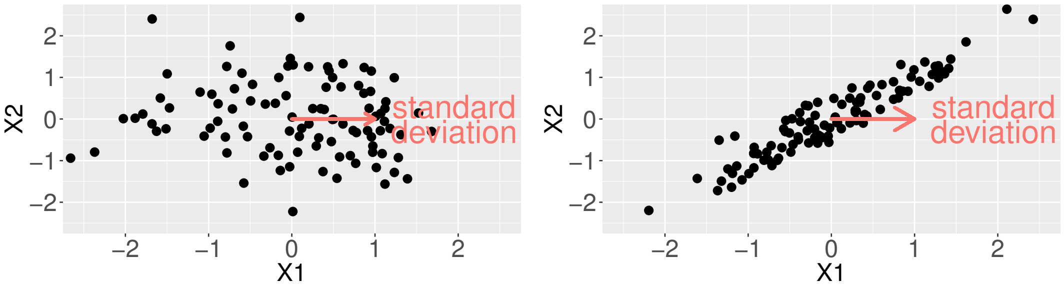

Figure 3.3: Standard deviations for uncorrelated variables (left) and correlated variables (right).

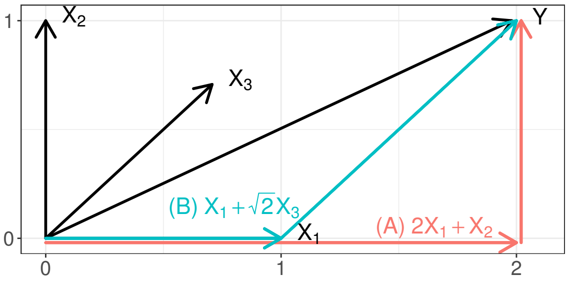

Figure 3.4: Uncorrelated model (A) versus correlated model (B)

On the other hand, in the correlated case, fixing X_2 strongly affects the range of X_1. Thus, increasing X_1 by 1 but not changing X_2 falls into an unrealistic (counterfactual) setting, that is, it gets out of the support of the distribution. For this reason, the effect no longer represents the average influence on the response fixing others in the correlated case. This example implies that correlations among active variables hinder the interpretation of linear models and leads to misinterpretation. These difficulties are mitigated in our proposed method by inducing low correlations.

Moreover, even when we have a moderately interpretable correlated model, it might be possible to obtain an uncorrelated model with refined interpretability. Let us consider a simple example with $p = 3$. Specifically, suppose $X = [X_1, X_2, X_3] \in \mathbb{R}^{n\times 3}$ is standardized, X_1 and X_2 are orthogonal, and $X_3 = (X_1 + X_2)/\sqrt{2}$. Consider two models: (A) $y = 2X_1 + X_2$ and (B) $y = X_1 + \sqrt{2}X_3$. Figure 3.4 illustrates X_j's and y. Both models output the same prediction. However, it seems that the model (A) is more interpretable than (B). This is because, in (A), active variables (X_1 and X_2) are uncorrelated, hence we can decompose the model ($2X_1 + X_2$) into each component ($2X_1$ and X_2) and interpret each coefficient as independent variable contribution from each variable to the response. On the other hand, in (B), active variables (X_1 and X_3) are correlated; hence each coefficient is no longer independent variable contribution. For example, imagine an example of predicting traffic congestion using three variables 'X_1 : Saturday', 'X_2 : Sunday', and 'X_3 : Weekend.' We can interpret the model with 'Saturday' and 'Sunday' more intuitively than that of 'Saturday' and 'Weekend' because 'Weekend' includes 'Saturday.' Of course, we can easily interpret the 'Saturday' and 'Weekend' model since their relationship is clear. However, in general, it is not easy to unravel complicated interactions among correlated variables since they might imply confounders, mediators, or cause-effect relationships. The Lasso selects (B) because ℓ_1 norm of its coefficients is small; while the IILasso tends to select (A) because our regularization term excludes correlations. This perfectly collinear example is not merely an extreme case, because it was reported that there were many good solutions (solutions which have almost the same error) around the Lasso solution in real applications (Hara and Maehara, 2017) so that severe collinearity often occurs in high-dimensional data.

3.3 Optimization

We introduce Coordinate Descent Algorithm (CDA) to solve the IILasso problem (3.1), which was originally proposed for the Lasso ($\alpha = 0$ for the IILasso) (Friedman, Hastie, Höfling, Tibshirani, et al., 2007; Friedman, Hastie, and Tibshirani, 2010). It is a simple and efficient algorithm, particularly for high dimensional data. CDA follows simply: For each $j \in \{1, \cdots, p\}$, we optimize the objective function with respect to β_j with the remaining elements of β fixed at their most recently updated values.

To derive the update equation, when $\beta_j \neq 0$, differentiating $\mathcal{L}(\beta)$ with respect to β_j yields

$$\partial_{\beta_j}\mathcal{L}(\beta) = -\frac{1}{n}X_j^\top \left(y - X_{-j}\beta_{-j}\right) + \left(1 + \lambda\alpha R_{jj}\right)\beta_j + \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\mathrm{sgn}(\beta_j),$$

where β_{-j} denotes β without the j-th component, X_{-j} denotes X without j-th column and $R_{j,-j}$ denotes the j-th row vector without j-th column of R. Solving $\partial_{\beta_j}\mathcal{L}(\beta) = 0$, we obtain the update rule as

$$\beta_j \leftarrow \frac{1}{1 + \lambda\alpha R_{jj}}\mathcal{S}\left(\frac{1}{n}X_j^\top \left(y - X_{-j}\beta_{-j}\right), \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\right), \tag{3.2}$$

Algorithm 1 CDA for the IILasso

```
for $\lambda = \lambda_{\max}, \cdots, \lambda_{\min}$ do
    initialize $\beta$
    while until convergence do
        for $j = 1, \cdots, p$ do
            $\beta_j \leftarrow \frac{1}{1+\lambda\alpha R_{jj}} \mathcal{S}\left(\frac{1}{n} X_j^\top (y - X_{-j}\beta_{-j}), \lambda(1 + \alpha R_{j,-j}|\beta_{-j}|)\right)$
        end for
    end while
end for
```

where $\mathcal{S}(z, \gamma)$ is a soft thresholding function

$$\begin{aligned} \mathcal{S}(z, \gamma) &:= \operatorname{sgn}(z)(|z| - \gamma)_+ \\ &= \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z|, \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z|, \\ 0 & \text{if } |z| \leq \gamma. \end{cases} \end{aligned}$$

The whole algorithm for solving the IILasso is described in Algorithm 1. We search several λ from $\lambda_{\max}$ to $\lambda_{\min}$. β is initialized at each λ in some ways such as (i) zeros for all elements, (ii) the solution of previous λ, or (iii) the solution of the ordinary Lasso.

In Algorithm 1, the objective function monotonically decreases at each update, and the estimate converges a stationary point.

Proposition 4. *Let $\{\beta^t\}_{t=0,1,\cdots}$ be a sequence of β in Algorithm 1. Then, every cluster point of $\{\beta^t\}_{t\equiv(p-1)\mathrm{mod}p}$ is a stationary point.*

Proof. The proof is based on Theorem 4.1 in (Tseng, 2001). First, we can see that the level set $\{\beta | \mathcal{L}(\beta) \leq \mathcal{L}(\beta^0)\}$ is compact and $\mathcal{L}(\beta)$ is continuous. Moreover, $\mathcal{L}(\beta)$ has a unique minimum with (3.2) in terms of β_j. Therefore, every cluster point of $\{\beta^t\}_{t\equiv(p-1)\mathrm{mod}p}$ is a coordinatewise minimum point. In addition, since $\mathcal{L}(\beta)$ can be seen as a locally quadratic function in any directions, $\mathcal{L}(\beta)$ is *regular* at the cluster point. Hence, Theorem 4.1 (c) in (Tseng, 2001) concludes the assertion. □

3.4 Related Work

IILasso looks similar to exclusive selection methods such as Uncorrelated Lasso and Exclusive Group Lasso.

The Uncorrelated Lasso (ULasso) (Chen, Ding, Luo, and Xie, 2013) aims to reduce correlations among active variables. Although the ULasso quite resembles our formulation, there exists a critical difference that they use β instead of $|\beta|$ in the objective function (2.5). We found that the ULasso does not necessarily select uncorrelated variables. For example, consider the case $X = [X_1, X_2]$. The last term of (2.5) is $\lambda_2(\beta_1^2 + \beta_2^2 + 2R_{12}\beta_1\beta_2)$. Minimizing $R_{12}\beta_1\beta_2$ makes $\beta_1\beta_2 < 0$ and $|\beta_1\beta_2|$ larger if $R_{12} \neq 0$. This implies that the ULasso tends to select correlated variables and set coefficients to the opposite sign. In particular, X_1 and X_2 are strongly correlated, then it reduces $\lambda_2(\beta_1 + \beta_2)^2$, which induces $\beta_1 = -\beta_2$. It is not a major problem when X_1 and X_2 are positively correlated, but is a significant problem when X_1 and X_2 are negatively correlated. This problem is overcome in our method. Therefore, the difference between their ULasso and our IILasso is essential and crucial.

The Exclusive Group Lasso (EGLasso) (Kong, Fujimaki, Liu, Nie, and Ding, 2014) is also the same direction of exclusive selection. EGLasso can be seen as a particular case of the IILasso. Let R be a group indicator matrix such as $R_{jk} = 1$ if X_j and X_k belong to the same group and $R_{jk} = 0$ otherwise. Then the IILasso is reduced to EGLasso. For the above example, if we define similarity matrix $R = [1, 1, 0; 1, 1, 0; 0, 0, 1]$, then the last term of the IILasso objective function (3.1) becomes $\lambda\,(\beta_1^2 + 2|\beta_1||\beta_2| + \beta_2^2 + \beta_3^2)$, which is the same as the last term of (2.6). As we see, EGLasso needs to determine the threshold θ and group variables beforehand, which causes severely unstable estimation.

3.5 Theoretical Properties

In this section, we show the sign recovery condition and the estimation error bound of the IILasso for linear models. These results explain the effectiveness of the IILasso in terms of interpretability and estimation error. Moreover, we show the property of a local minimum, which implies that every local optimal solution achieves the same statistical error rate as the global optimal solution. In this chapter, let β^* denote the true parameter. Let S denote the true active sets, i.e., $S = \mathrm{supp}(\beta^*) = \{j \in \{1, \cdots, p\} | \beta_j^* \neq 0\}$ and $s = |S|$.

3.5.1 Sign Recovery

First, we give a necessary and sufficient condition of sign recovery. We define

$$\begin{aligned}U :=&\frac{1}{n}X_S^\top X_S + \lambda\alpha \mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*)),\\ w :=&\mathrm{sgn}(\beta_S^*) + \alpha \mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))\beta_S^*.\end{aligned}$$

Theorem 5. *Assume U is invertible. Then, there exists a critical point $\hat{\beta}$ of (3.1) with correct sign recovery $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$ if and only if the following two conditions hold:*

$$\mathrm{sgn}\left(\beta_S^* - U^{-1}\left(\lambda w - \frac{1}{n}X_S^\top \varepsilon\right)\right) = \mathrm{sgn}(\beta_S^*), \tag{3.3}$$

$$\begin{aligned}&\left|\frac{1}{n}X_{S^c}^\top X_S U^{-1}\left(\lambda w - \frac{1}{n}X_S^\top \varepsilon\right) + \frac{1}{n}X_{S^c}^\top \varepsilon\right|\\ &\leq \lambda\left(1 + \alpha R_{S^cS}\left|\beta_S^* - U^{-1}\left(\lambda w - \frac{1}{n}X_S^\top \varepsilon\right)\right|\right),\end{aligned} \tag{3.4}$$

where both of these vector inequalities are taken elementwise.

The proof is given in 3.6.1. The sign recovery condition is derived from the standard conditions for optimality. We note that $\alpha = 0$ reduces the condition into the ordinary Lasso condition in (Wainwright, 2009). The invertible assumption of U is not restrictive because it is true for almost all λ if $X_S^\top X_S$ is invertible, which is the same assumption as standard analysis of the Lasso.

According to Theorem 5, the IILasso has the advantage in sign recovery for correlated design compared to the Lasso, as long as the truth is “well interpretable”, that is, the true non-zero components are independent. This is because, when R_{SS} is small enough, (3.3) is the same as the Lasso and (3.4) is easier to be satisfied unless $\alpha R_{S^cS} = 0$. Besides, the condition gets milder as R_{S^cS} gets large.

We note that Theorem 5 is *not* the condition of a global optimal solution. However, if global optimal solutions are finite, they must be a critical point. Hence, there exists a global optimal solution with correct sign recovery only if (3.3) and (3.4) hold.

Next, we give a sufficient condition for sign recovery. The following theorem clarifies the sign recovery conditions on the design matrix X and the true parameter β^* since it does not depends on the realization of the noise ε. We prepare some assumptions and definitions.

Assumption 1 (Sub-Gaussian). *The noise sequence* $\{\varepsilon_i\}_{i=1}^n$ *is an i.i.d. sub-Gaussian sequence with parameter* $\sigma > 0$*, i.e.,* $\mathrm{E}[\exp(t\varepsilon_i)] \leq \exp(\sigma^2 t^2/2)$ *for* $\forall t \in \mathbb{R}$*.*

We give sub-Gaussian properties in section 3.6.2 for our theoretical analyses.

Assumption 2. *There exists a constant* $D > 0$*,* $\varphi > 0$*, and* ψ *such that* $1 + \lambda_n \alpha \psi > 0$*,*

$$\|R_{SS}\|_{\max} \leq D, \quad \frac{1}{n} X_S^\top X_S \succeq \varphi I,$$

$$\left(\frac{1}{n} X_S^\top X_S\right)^{-1/2} \mathrm{Diag}(\mathrm{sgn}(\beta_S^*)) R_{SS} \mathrm{Diag}(\mathrm{sgn}(\beta_S^*)) \left(\frac{1}{n} X_S^\top X_S\right)^{-1/2} \succeq \psi I. \tag{3.5}$$

We note that ψ is not necessarily positive, but is larger than $\psi > -1/\lambda_n\alpha$. The condition (3.5) is satisfied if $R_{SS} \succeq \psi_R I$, $X_S^\top X_S/n \preceq \varphi_{\max} I$, $\psi_R/\varphi \geq \psi$, and $\psi_R/\varphi_{\max} \geq \psi$.

Definition 3 (Generalized Incoherence Condition). *We say that the generalized incoherence condition holds if there exists some incoherence parameter* $\kappa \in (0, 1]$ *such that*

$$\left\|\frac{1}{n} X_j^\top X_S U^{-1}\right\|_1 \leq \left(\frac{1 + \alpha \|R_{Sj}\|_1 \Delta_{\min}}{1 + \alpha D \|\beta_S^*\|_1}\right)(1 - \kappa), \tag{3.6}$$

for $\forall j \in S^c$*, where* $\beta_{\min}^* := \min_{j \in S} |\beta_j^*|$ *and*

$$\Delta_{\min} := \beta_{\min}^* - \lambda_n \left((1 + \alpha D \|\beta_S^*\|_1) \|U^{-1}\|_\infty + \frac{4\sigma}{\sqrt{\varphi}(1 + \lambda_n \alpha \psi)}\right) > 0.$$

The generalized incoherence condition (Definition 3) is a generalized notion of the incoherence condition (Fuchs, 2005; Tropp, 2006; Wainwright, 2009). The generalized incoherence condition reduces to the ordinary incoherence condition when $\alpha = 0$. The ordinary incoherence condition is quite restrictive and is much stronger than the restricted eigenvalue condition. For example, if we have $X_S^\top X_S/n = I$, the ordinary incoherence condition requires that $\max_{j \in S^c} \sum_{k \in S} |\sum_i X_{ij} X_{ik}/n| < 1$. This condition is hard to be satisfied if there are correlations between informative and uninformative variables.

On the other hand, the generalized incoherence condition offers a great advantage because the right-hand side of (3.6) is *not* upper bounded by 1 when $\alpha \neq 0$. Specifically, consider a case where the true model is "well interpretable", i.e., $R_{SS} = O$. The generalized incoherence condition reduces to

$$\left\| \frac{1}{n} X_j^\top X_S \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \right\|_1 \leq (1 + \alpha \| R_{Sj} \|_1 \Delta_{\min}) \, (1 - \kappa), \tag{3.7}$$

for $\forall j \in S^c$, where

$$\Delta_{\min} := \beta^*_{\min} - \lambda_n \left(\left\| \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \right\|_\infty + \frac{4\sigma}{\sqrt{\varphi}(1 + \lambda_n \alpha \psi)} \right) > 0.$$

This condition is milder than the ordinary incoherence condition since the right-hand side of (3.7) is larger than $1 - \kappa$.

Now, we fix any $0 < \delta < 1$ and define γ_n as

$$\gamma_n = \gamma_n(\delta) := \sigma \sqrt{\frac{2 \log(2p/\delta)}{n}}. \tag{3.8}$$

Then, we obtain a sufficient condition for sign recovery.

Theorem 6. *Suppose that Assumptions 1 and 2 and Definition 3 (generalized incoherence condition) with a constant* $\kappa \in (0, 1]$ *are satisfied. Suppose that the regularization parameter satisfies*

$$\lambda_n \geq \max \left\{ \frac{1}{4\sigma}, \frac{2}{\kappa}, \left(\frac{\lambda_n \alpha \psi}{1 + \lambda_n \alpha \psi} \right)^2 \frac{2}{\kappa} \right\} \gamma_n, \tag{3.9}$$

with γ_n *in* (3.8)*. Then, there exists a critical point* $\hat{\beta}$ *of* (3.1) *with correct sign recovery* $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)$ *with probability at least* $1 - 2\delta$*.*

The proof is given in section 3.6.3. We note that $\alpha = 0$ reduces to the result of ordinary Lasso (Wainwright, 2009). When $\alpha \neq 0$, our method has much advantage for sign recovery as long as the true model is "well interpretable", i.e., $R_{SS} = O$. This is because (i) the generalized incoherence condition (Definition 3) for the IILasso is milder than that for the ordinary Lasso as described above, and (ii) Assumption 2 and (3.9) reduces to the ordinary Lasso condition. Various classes of X satisfy the incoherence condition (Meinshausen, Bühlmann, et al., 2006; Zhao and Yu, 2006). This implies that the generalized incoherence condition also holds in various situations as long as R_{SS} is small enough.

3.5.2 Estimation Error

We give an estimation error bound of approximately global minimum solutions of our method. Before we give the statement, the assumption and definition are prepared.

Definition 4 (Generalized Restricted Eigenvalue Condition for Linear Models (GRE(S, C, C')))**.** *Let a set of vectors $\mathcal{B}(S, C, C')$ be*

$$\mathcal{B}(S, C, C') := \Big\{ v \in \mathbb{R}^p : \|v_{S^c}\|_1 + \frac{C'\alpha}{2}|v_{S^c}|^\top R_{S^cS^c}|v_{S^c}| \\ + C'\alpha|v_{S^c}|^\top R_{S^cS}|v_S + \beta^*_S| \leq C\|v_S\|_1 \Big\}.$$

We say that the genelarized restricted eigenvalue condition holds if we have $\phi_{\mathrm{GRE}} > 0$ where

$$\phi_{\mathrm{GRE}} = \phi_{\mathrm{GRE}}(S, C, C') := \inf_{v \in \mathcal{B}(S,C,C')} \frac{v^\top \frac{1}{n} X^\top X v}{\|v\|_2^2}.$$

The generalized restricted eigenvalue (GRE) condition (Definition 4) is a generalized notion of the *restricted eigenvalue (RE) condition* (Bickel, Ritov, Tsybakov, et al., 2009; Bühlmann and Van De Geer, 2011) tailored for our regularization. One can see that, if $\alpha = 0$ or $C' = 0$, then $\phi_{\mathrm{GRE}}(S, C, C')$ is reduced to the ordinary restricted eigenvalue (Bickel, Ritov, Tsybakov, et al., 2009; Raskutti, Wainwright, and Yu, 2010) for the analysis of the Lasso.

The GRE condition is not restrictive. Since there are additional terms related to $|\beta|^\top R|\beta|$, the set $\mathcal{B}(S, C, C')$ is smaller than that for the ordinary restricted eigenvalue if the same C is used. In particular, the term $|\beta_{S^c}|^\top R_{S^cS}|\beta_S + \beta^*_S|$ strongly restricts the amplitude of coefficients for unimportant variables (especially the variables with large R_{S^cS}). Hence, Assumption GRE(S, C, C') is milder than Assumption RE(S, C). Because the RE(S, C) condition is satisfied in general class of Gaussian design (Raskutti, Wainwright, and Yu, 2010), the GRE condition also holds in general class of Gaussian design.

Then, we obtain the convergence rate of approximately global minimum solution of the IILasso as follows.

Theorem 7. *Suppose that Assumption 1 is satisfied. Suppose R satisfy*

$$\|R_{SS}\|_{\max} \leq D, \tag{3.10}$$

for some positive constant D, and the estimator $\hat{\beta}$ of (3.1) is approximately minimizing the objective function so that

$$\mathcal{L}(\hat{\beta}) \leq \mathcal{L}(\beta^*). \tag{3.11}$$

Suppose that the regularization parameters satisfy

$$3\gamma_n \leq \lambda_n \text{ and } \alpha \leq \frac{1}{4D\|\beta^*_S\|_1}, \tag{3.12}$$

with γ_n in (3.8). Suppose that Assumption GRE$(S, 3, 3/2)$ (Definition 4) is satisfied. Then, it holds that

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{16 s \lambda_n^2}{\phi_{\mathrm{GRE}}^2},$$

with probability at least $1 - \delta$.

The proof is given in 3.6.4. The obtained convergence rate is roughly evaluated as

$$\|\hat{\beta} - \beta^*\|_2^2 = O_p\left(\frac{s \log(p)}{n}\right),$$

by taking $\lambda_n = O_p(\sqrt{\log p/n})$, which is almost the minimax optimal rate (Raskutti, Wainwright, and Yu, 2011).

We compare the convergence rate of the Lasso and IILasso. For comparison, we have a little bit stricter bound

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{\left(\frac{8}{3} + 5\alpha D\|\beta^*_S\|_1 + \frac{3}{4}(\alpha D\|\beta^*_S\|_1)^2\right)^2 s\lambda_n^2}{\phi_{\mathrm{GRE}}^2},$$

under Assumption GRE$(S, C, 3/2)$ (Definition 4) where $C = 2 + 15\alpha D\|\beta^*_S\|_1/4 + 9(\alpha D\|\beta^*_S\|_1)^2/16$, with high probability. The proof is given in 3.6.5. We can easily see that when $\alpha = 0$, then the convergence rate analysis is reduced to the standard one for the ordinary Lasso (Bickel, Ritov, Tsybakov, et al., 2009; Bühlmann and Van De Geer, 2011). We note that our theorem includes the additional assumption (3.10) compared to the ordinary Lasso theorem, and instead relaxes the restricted eigenvalue condition. This indicates that our theorem holds under the milder condition in terms of correlations among all variables, in exchange for the additional condition of the true active variables. Under "well interpretable" cases where the true non-zero components S are independent, i.e., $R_{SS} = O$, the error bounds for the Lasso and IILasso are the same except for the term ϕ_{GRE}. Since R_{S^cS} and $R_{S^cS^c}$ shrink the set of vectors $\mathcal{B}(S, C, C')$ in Definition 4, ϕ_{GRE} of the IILasso is larger than that of the Lasso. Therefore, our approximately global minimum solution has a better error bound than the ordinary ℓ_1 regularization in this situation. Besides, our method has more advantageous when the variables are correlated between informative and non-informative variables.

3.5.3 Local Optimality

The objective function of the IILasso is *not* necessarily convex in exchange for better statistical properties, as observed above. Our next theoretical interest is about the local optimality of our optimization algorithm (Algorithm 1). Since our optimization method is greedy, there is no confirmation that it achieves the global optimum. However, as we see in this chapter, the local solution achieves almost the same estimation error as the global optimum satisfying (3.11). For theoretical simplicity, we assume the following a little stronger condition.

Assumption 3. *There exists $\phi > 0$ and $q_n \leq p$ such that, for all $V \subset \{1, \ldots, p\}$ satisfying $|V| \leq q_n$ and $V \cap S = \emptyset$, it holds that*

$$\frac{1}{n} X_{S \cup V}^{\top} X_{S \cup V} \succ \phi \mathrm{I}.$$

Moreover, there exists $\bar{D}$ such that the maximum absolute value of the eigenvalue of R is bounded as

$$\sup_{u \in \mathbb{R}^{S \cup V}} u^{\top} (R_{S \cup V, S \cup V}) u \leq \bar{D} \|u\|_2^2.$$

Then, we obtain the convergence rate of local minimum solutions of IILasso as follows.

Theorem 8. *Suppose Assumptions 1 and 3 are satisfied. Suppose that $\hat{\beta}$ is a local optimal solution of (3.1) satisfying $|\operatorname{supp}(\hat{\beta})| \leq |S| + q_n$. Let the regularization parameters satisfy*

$$\gamma_n < \lambda_n \text{ and } \alpha < \min \left\{ \frac{\sqrt{s}}{2\bar{D}\|\beta^*\|_2}, \frac{\phi}{2\bar{D}\lambda_n} \right\},$$

with γ_n in (3.8). Then, $\hat{\beta}$ should satisfy

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{25 s \lambda_n^2}{\phi^2},$$

with probability at least $1-\delta$.

The proof is given in 3.6.6. Theorem 8 indicates that every local optimal solution achieves the same convergence rate with the ideal optimal solution. In other words, there is *no local optimal solution* with sparsity level $|S|+q_n$ far from the true vector β^*.

3.6 Proofs

3.6.1 Proof of Theorem 5

Proof. By standard conditions for optimality, $\hat{\beta}$ is a critical point if and only if there exists a subgradient $\hat{z} \in \partial\|\hat{\beta}\|_1 := \{\hat{z} \in \mathbb{R}^p | \hat{z}_j = \mathrm{sgn}(\hat{\beta}_j) \text{ for } \hat{\beta}_j \neq 0,\ |\hat{z}_j| \leq 1 \text{ otherwise}\}$ such that $\partial_{\hat{\beta}}\mathcal{L}(\beta) = 0$. Because $\partial_\beta \frac{1}{2}|\beta|^\top R|\beta| = \mathrm{Diag}(R|\beta|)z$, the condition $\partial_{\hat{\beta}}\mathcal{L}(\beta) = 0$ yields

$$-\frac{1}{n}X^\top(y - X\hat{\beta}) + \lambda\hat{z} + \lambda\alpha\mathrm{Diag}\left(R|\hat{\beta}|\right)\hat{z} = 0. \tag{3.13}$$

Substituting $y = X\beta^* + \varepsilon$ in (3.13), we have

$$-\frac{1}{n}X^\top(X(\beta^* - \hat{\beta}) + \varepsilon) + \lambda\hat{z} + \lambda\alpha\mathrm{Diag}\left(R|\hat{\beta}|\right)\hat{z} = 0. \tag{3.14}$$

Let the true active set $S = \{1, \cdots, s\}$ and inactive set $S^c = \{s+1, \cdots, p\}$ without loss of generality, then (3.14) is turned into

$$\frac{1}{n}X_S^\top X_S\left(\hat{\beta}_S - \beta_S^*\right) + \frac{1}{n}X_S^\top X_{S^c}\hat{\beta}_{S^c} - \frac{1}{n}X_S^\top\varepsilon + \lambda\hat{z}_S + \lambda\alpha\mathrm{Diag}\left(R_{SS}|\hat{\beta}_S|\right)\hat{z}_S = 0, \tag{3.15}$$

$$\frac{1}{n}X_{S^c}^\top X_S\left(\hat{\beta}_S - \beta_S^*\right) + \frac{1}{n}X_{S^c}^\top X_{S^c}\hat{\beta}_{S^c} - \frac{1}{n}X_{S^c}^\top\varepsilon + \lambda\hat{z}_{S^c} + \lambda\alpha\mathrm{Diag}\left(R_{S^cS}|\hat{\beta}_S|\right)\hat{z}_{S^c} = 0. \tag{3.16}$$

Hence, there exists a critical point with correct sign recovery if and only if there exists $\hat{\beta}$ and $\hat{z}$ such that (3.15), (3.16), $\hat{z} \in \partial\|\hat{\beta}\|_1$ and $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$. The latter two conditions can be written as

$$\hat{z}_S = \mathrm{sgn}(\beta_S^*), \tag{3.17}$$

$$|\hat{z}_{S^c}| \leq 1, \tag{3.18}$$

$$\mathrm{sgn}(\hat{\beta}_S) = \mathrm{sgn}(\beta_S^*), \tag{3.19}$$

$$\hat{\beta}_{S^c} = 0. \tag{3.20}$$

The condition (3.17) and (3.20) yield

$$\begin{aligned}\frac{1}{n}X_S^\top X_S\left(\hat{\beta}_S-\beta_S^*\right)-\frac{1}{n}X_S^\top\varepsilon+\lambda\,\mathrm{sgn}(\beta_S^*)&\\+\lambda\alpha\mathrm{Diag}\left(R_{S\cdot}|\hat{\beta}|\right)\mathrm{sgn}(\beta_S^*)=0,&\end{aligned}\tag{3.21}$$

$$\frac{1}{n}X_{S^c}^\top X_S\left(\hat{\beta}_S-\beta_S^*\right)-\frac{1}{n}X_{S^c}^\top\varepsilon+\lambda\hat{z}_{S^c}+\lambda\alpha\mathrm{Diag}\left(R_{S^c\cdot}|\hat{\beta}|\right)\hat{z}_{S^c}=0.\tag{3.22}$$

Since

$$\begin{aligned}\mathrm{Diag}(R_{SS}|\hat{\beta}_S|)\,\mathrm{sgn}(\beta_S^*)=&\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}|\hat{\beta}_S|\\=&\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))\hat{\beta}_S,\end{aligned}$$

(3.21) can be rewritten as

$$U(\hat{\beta}_S-\beta_S^*)+w-\frac{1}{n}X_S^\top\varepsilon=0,$$

where

$$\begin{aligned}U&:=\frac{1}{n}X_S^\top X_S+\lambda\alpha\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*)),\\w&:=\lambda\,\mathrm{sgn}(\beta_S^*)+\lambda\alpha\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))\beta_S^*.\end{aligned}$$

If we assume U is invertible, we obtain

$$\hat{\beta}_S=\beta_S^*-U^{-1}\left(w-\frac{1}{n}X_S^\top\varepsilon\right).\tag{3.23}$$

Substituting this in (3.22), we have

$$\begin{aligned}&\frac{1}{n}X_{S^c}^\top X_S\left(-U^{-1}\left(w-\frac{1}{n}X_S^\top\varepsilon\right)\right)-\frac{1}{n}X_{S^c}^\top\varepsilon+\lambda\hat{z}_{S^c}\\&+\lambda\alpha\mathrm{Diag}\left(R_{S^cS}\left|\beta_S^*-U^{-1}\left(w-\frac{1}{n}X_S^\top\varepsilon\right)\right|\right)\hat{z}_{S^c}=0,\end{aligned}$$

that is,

$$\begin{aligned}&\left(1+\alpha\mathrm{Diag}\left(R_{S^cS}\left|\beta_S^*-U^{-1}\left(w-\frac{1}{n}X_S^\top\varepsilon\right)\right|\right)\right)\lambda\hat{z}_{S^c}\\=&\frac{1}{n}X_{S^c}^\top X_SU^{-1}\left(w-\frac{1}{n}X_S^\top\varepsilon\right)+\frac{1}{n}X_{S^c}^\top\varepsilon.\end{aligned}\tag{3.24}$$

Combining (3.18), (3.19), (3.23) and (3.24), we concludes the assertion.

□

3.6.2 Sub-Gaussian Tail Bounds

We briefly summarize the definition and properties of sub-Gaussian since it plays a key role in our non-asymptotic analyses. See (Wainwright, 2009; Rigollet and Hütter, 2015) for details for example. Let ε be a zero-mean random variable. We say that ε is a sub-Gaussian variable with parameter $\sigma > 0$ if it holds for $\forall t \in \mathbb{R}$

$$\mathrm{E}[\exp(t\varepsilon)] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right). \tag{3.25}$$

By applying the Chernoff bound to (3.25), we have a sub-Gaussian tail bound for $\forall z > 0$

$$P\left(|\varepsilon| > z\right) \leq 2\exp\left(-\frac{z^2}{2\sigma^2}\right). \tag{3.26}$$

We obtain the following lemma for a sequence of sub-Gaussian variables. This is useful for our theoretical analyses of sign recovery.

Lemma 9. *Let $\{\varepsilon_i\}_{i=1}^n$ be i.i.d. zero-mean sub-Gaussian variables with a parameter σ. Then, we have for $\forall a \in \mathbb{R}^n$ and $\forall z > 0$,*

$$P\left(\left|\sum_{i=1}^n a_i \varepsilon_i\right| > z\right) \leq 2\exp\left(-\frac{z^2}{2\|a\|_2^2\sigma^2}\right).$$

Proof. From the definition of sub-Gaussian, we have

$$\begin{aligned}\mathrm{E}\left[\exp\left(t\sum_{i=1}^n a_i\varepsilon_i\right)\right] &= \prod_{i=1}^n \mathrm{E}\left[\exp\left(ta_i\varepsilon_i\right)\right] \\ &\leq \prod_{i=1}^n \exp\left(\frac{a_i^2\sigma^2t^2}{2}\right) = \exp\left(\frac{\|a\|_2^2\sigma^2t^2}{2}\right).\end{aligned}$$

Therefore, the sub-Gaussian tail bound (3.26) concludes the assertion. □

In addition, we prepare the following collorary using Lemma 9. This is useful for our theoretical analyses of estimation error and local optimality.

Corollary 10. Suppose that Assumption 1 and $\sum_{i=1}^{n} X_{ij}^2/n \leq 1$ for $\forall j = 1, \dots, p$ are satisfied. For $\forall \delta > 0$, define $\gamma_n := \gamma_n(\delta)$ as (3.8). Then, we have

$$P\left(\left\|\frac{1}{n}X^\top \varepsilon\right\|_\infty \geq \gamma_n\right) \leq \delta.$$

Proof. Notice that

$$\begin{aligned}
P\left(\left\|\frac{1}{n}X^\top \varepsilon\right\|_\infty \geq \gamma_n\right) &= P\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i\right| \geq \gamma_n\right) \\
&= P\left(\bigcup_{1\leq j\leq p}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i\right| \geq \gamma_n\right\}\right) \\
&\leq \sum_{j=1}^{p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i\right| \geq \gamma_n\right) \\
&\leq p \max_{1\leq j\leq p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i\right| \geq \gamma_n\right) \\
&\leq 2p \max_{1\leq j\leq p} \exp\left(-\frac{n^2\gamma_n^2}{2\sigma^2\|X_j\|_2^2}\right) \\
&\leq \exp\left(-\frac{n\gamma_n^2}{2\sigma^2} + \log(2p)\right),
\end{aligned}$$

where we used Lemma 9 in the fifth line. Since we set $\delta = \exp\left(-n\gamma_n^2/2\sigma^2 + \log(2p)\right)$, we concludes the assertion. □

3.6.3 Proof of Theorem 6

Proof. We derive sufficient conditions for (3.3) and (3.4) in Theorem 5.

In terms of (3.3), it is sufficient if

$$\beta^*_{\min} > \left\|U^{-1}\left(\lambda_n w - \frac{1}{n}X_S^\top \varepsilon\right)\right\|_\infty .$$

By the triangular inequality, we have

$$\left\|U^{-1}\left(\lambda_n w - \frac{1}{n}X_S^\top \varepsilon\right)\right\|_\infty \leq \lambda_n \left\|U^{-1} w\right\|_\infty + \left\|U^{-1}\frac{1}{n}X_S^\top \varepsilon\right\|_\infty . \tag{3.27}$$

The first term on the right-hand side of (3.27) is bounded as

$$\lambda_n \left\|U^{-1} w\right\|_\infty \leq \lambda_n \left(1 + \alpha D \|\beta^*_S\|_1\right) \|U^{-1}\|_\infty .$$

Consider the j-th element of random variable of the second term on the right-hand side of (3.27),

$$T_j := e_j^\top U^{-1}\frac{1}{n}X_S^\top \varepsilon,$$

where $e_j \in \mathbb{R}^s$ represents a unit vector with 1 for the j-th element and 0 for others. From Lemma 9, we have for $\forall t > 0$,

$$P(|T_j| > t) \le 2\exp\left(-\frac{t^2 n^2}{2\sigma^2 \left\|X_S U^{-1} e_j\right\|_2^2}\right).$$

Using Assumption 2, we have

$$\begin{aligned}
&\left(\frac{1}{n}X_S^\top X_S\right)^{-1/2} U \left(\frac{1}{n}X_S^\top X_S\right)^{-1/2} \\
&= I + \lambda_n\alpha\left(\frac{1}{n}X_S^\top X_S\right)^{-1/2} \mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))\left(\frac{1}{n}X_S^\top X_S\right)^{-1/2} \\
&\succeq (1+\lambda_n\alpha\psi)I \\
\Rightarrow & U\left(\frac{1}{n}X_S^\top X_S\right)^{-1} U \\
&= \left(\frac{1}{n}X_S^\top X_S\right)^{1/2}\left(\left(\frac{1}{n}X_S^\top X_S\right)^{-1/2} U \left(\frac{1}{n}X_S^\top X_S\right)^{-1/2}\right)^2 \left(\frac{1}{n}X_S^\top X_S\right)^{1/2} \\
&\succeq \varphi(1+\lambda_n\alpha\psi)^2 I \\
\Rightarrow & \|X_S U^{-1} e_j\|_2^2 = e_j^\top U^{-1} X_S^\top X_S U^{-1} e_j \le n/\varphi(1+\lambda_n\alpha\psi)^2
\end{aligned}$$

Hence, we obtain

$$P\left(\max_{j\in S}|T_j| > t\right) \le 2s\exp\left(-\frac{t^2 n\varphi(1+\lambda_n\alpha\psi)^2}{2\sigma^2}\right).$$

Setting $t = 4\lambda_n\sigma/\sqrt{\varphi}(1+\lambda_n\alpha\psi)$, we have

$$P\left(\max_{j\in S}|T_j| > \frac{4\lambda_n\sigma}{\sqrt{\varphi}(1+\lambda_n\alpha\psi)}\right) \le \exp\left(-8\lambda_n^2 n + \log(2s)\right).$$

Therefore, if it holds $\Delta_{\min} > 0$ where

$$\Delta_{\min} := \beta_{\min}^* - \lambda_n\left((1+\alpha D\|\beta_S^*\|_1)\,\|U^{-1}\|_\infty + \frac{4\sigma}{\sqrt{\varphi}(1+\lambda_n\alpha\psi)}\right),$$

then (3.3) is satisfied with probability at least $1 - \exp\left(-8\lambda_n^2 n + \log(2s)\right)$.

In terms of (3.4), it is sufficient if for $\forall j \in S^c$,

$$\begin{aligned} &\left| \frac{\lambda_n}{n} X_j^\top X_S U^{-1} w \right| + \left| \frac{1}{n} X_j^\top \left(I - \frac{1}{n} X_S U^{-1} X_S^\top \right) \varepsilon \right| \\ &\leq \lambda_n \left(1 + \alpha \| R_{Sj} \|_1 \Delta_{\min} \right), \end{aligned} \tag{3.28}$$

since we have now

$$\left| \beta_S^* - U^{-1} \left(\lambda w - \frac{1}{n} X_S^\top \varepsilon \right) \right| > \Delta_{\min} > 0.$$

The first term on the left-hand side of (3.28) is bounded as

$$\left| \frac{\lambda_n}{n} X_j^\top X_S U^{-1} w \right| \leq \frac{\lambda_n}{n} \left(1 + \alpha D \| \beta_S^* \|_1 \right) \left\| X_j^\top X_S U^{-1} \right\|_1 .$$

Consider the random variable

$$Z_j := \frac{1}{n} X_j^\top \left(I - \frac{1}{n} X_S U^{-1} X_S^\top \right) \varepsilon,$$

in the second term on the left-hand side of (3.28). From Lemma 9, we have for $\forall z_j > 0$

$$P(|Z_j| > z_j) \leq 2 \exp \left(- \frac{n^2 z_j^2}{2\sigma^2 \left\| X_j^\top \left(I - \frac{1}{n} X_S U^{-1} X_S^\top \right) \right\|_2^2} \right).$$

Since $U \succeq \varphi(1 + \lambda_n \alpha \psi) I \succeq O$, we have

$$I - \frac{1}{n} X_S U^{-1} X_S^\top \preceq I. \tag{3.29}$$

In addition, we have

$$\begin{aligned} & I - \frac{1}{n} X_S U^{-1} X_S^\top \\ =& I - X_S \left(X_S^\top X_S \right)^{-1/2} \left(\left(\frac{1}{n} X_S^\top X_S \right)^{-1/2} U \left(\frac{1}{n} X_S^\top X_S \right)^{-1/2} \right)^{-1} \left(X_S^\top X_S \right)^{-1/2} X_S^\top \\ \succeq& I - \frac{1}{1 + \lambda_n \alpha \psi} X_S (X_S^\top X_S)^{-1} X_S^\top \\ \succeq& \left(1 - \frac{1}{1 + \lambda_n \alpha \psi} \right) I, \end{aligned} \tag{3.30}$$

where we used the fact that $X_S(X_S^\top X_S)^{-1}X_S^\top$ is the projection matrix to the image of $X_S^\top$ in the last line. (3.29) and (3.30) give

$$\left\| I - \frac{1}{n} X_S U^{-1} X_S^\top \right\|_2^2 \leq \max\left\{1, \left(\frac{\lambda_n \alpha \psi}{1+\lambda_n \alpha\psi}\right)^2\right\} =: \nu^2$$

Hence, we obtain

$$P\left(\bigcup_{j\in S^c} \{|Z_j| > z_j\}\right) \leq \sum_{j\in S^c} 2\exp\left(-\frac{n z_j^2}{2\sigma^2\nu^2}\right).$$

Setting $z_j = \lambda_n(1+\alpha\|R_{Sj}\|_1\Delta_{\min})\kappa/2$, we have

$$\begin{aligned} & P\left(\bigcup_{j\in S^c} \{|Z_j| > \lambda_n(1+\alpha\|R_{Sj}\|_1\Delta_{\min})\kappa/2\}\right) \\ \leq & 2\sum_{j\in S^c} \exp\left(-\frac{n\lambda_n^2\kappa^2(1+\alpha\|R_{Sj}\|_1\Delta_{\min})^2}{8\sigma^2\nu^2}\right) \\ \leq & 2(p-s)\exp\left(-\frac{n\lambda_n^2\kappa^2}{8\sigma^2\nu^2}\right) \\ = & \exp\left(-\frac{n\lambda_n^2\kappa^2}{8\sigma^2\nu^2} + \log(2(p-s))\right). \end{aligned}$$

Therefore, the generalized incoherence condition (Definition 3) yields the condition (3.4) with probability at least $1-\exp(-n\lambda_n^2\kappa^2/8\sigma^2\nu^2 + \log(2(p-s)))$.

Overall, the conditions (3.3) and (3.4) hold with probability at least $1-\exp\left(-8\lambda_n^2 n + \log(2s)\right) - \exp(-n\lambda_n^2\kappa^2/8\sigma^2\nu^2 + \log(2(p-s)))$. Since we set λ_n and δ as $\lambda_n \geq \max\{1/4\sigma, 2\nu/\kappa\}\gamma_n$ and $\delta = \exp(-n\gamma_n^2/2\sigma^2 + \log(2p))$, the probability is bounded by $1-2\delta$.

□

3.6.4 Proof of Theorem 7

Proof. By $\mathcal{L}(\hat\beta) \leq \mathcal{L}(\beta^*)$ and $y = X\beta^* + \varepsilon$, it holds that

$$\begin{aligned} & \frac{1}{2n}\|X(\hat\beta-\beta^*)\|_2^2 + \lambda_n\left(\|\hat\beta\|_1 + \frac{\alpha}{2}|\hat\beta|^\top R|\hat\beta|\right) \\ \leq & \frac{1}{n}\varepsilon^\top X(\hat\beta-\beta^*) + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right). \end{aligned} \tag{3.31}$$

By Corollary 10, it holds that

$$P\left(\left\|\frac{1}{n}X^\top\varepsilon\right\|_\infty > \gamma_n\right) \leq \delta.$$

Hereafter, we assume that the event $\{\left\|\frac{1}{n}X^\top\varepsilon\right\|_\infty \leq \gamma_n\}$ is happening.

Then, if $\gamma_n \leq \lambda_n/3$, by (3.31),

$$\begin{aligned}
&\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_n\left(\|\hat{\beta}\|_1 + \frac{\alpha}{2}|\hat{\beta}|^\top R|\hat{\beta}|\right)\\
\leq&\frac{1}{n}\|\varepsilon^\top X\|_\infty\|\beta^*-\hat{\beta}\|_1 + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right)\\
\leq&\gamma_n\|\beta^*-\hat{\beta}\|_1 + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right)\\
\leq&\frac{1}{3}\lambda_n\|\beta^*-\hat{\beta}\|_1 + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right).
\end{aligned} \tag{3.32}$$

Since

$$\|\hat{\beta}-\beta^*\|_1 = \|\hat{\beta}_S-\beta^*_S\|_1 + \|\hat{\beta}_{S^c}-\beta^*_{S^c}\|_1 = \|\hat{\beta}_S-\beta^*_S\|_1 + \|\hat{\beta}_{S^c}\|_1,$$

and

$$\begin{aligned}
&|\beta^*_S|^\top R_{SS}|\beta^*_S| - |\hat{\beta}_S|^\top R_{SS}|\hat{\beta}_S|\\
\leq&\sum_{(j,k)\in S\times S} R_{jk}|\beta^*_j\beta^*_k - \hat{\beta}_j\hat{\beta}_k|\\
=&\sum_{(j,k)\in S\times S} R_{jk}|\beta^*_j\beta^*_k - (\hat{\beta}_j-\beta^*_j+\beta^*_j)(\hat{\beta}_k-\beta^*_k+\beta^*_k)|\\
\leq&2\sum_{(j,k)\in S\times S} R_{jk}|\beta^*_j(\beta^*_k-\hat{\beta}_k)| + \sum_{(j,k)\in S\times S} R_{jk}|(\beta^*_j-\hat{\beta}_j)(\beta^*_k-\hat{\beta}_k)|\\
=&2|\beta^*_S|^\top R_{SS}|\beta^*_S-\hat{\beta}_S| + |\beta^*_S-\hat{\beta}_S|^\top R_{SS}|\beta^*_S-\hat{\beta}_S|\\
\leq&2\|R_{SS}|\beta^*_S|\|_\infty\|\beta^*_S-\hat{\beta}_S\|_1 + D\|\beta^*_S-\hat{\beta}_S\|_1^2,
\end{aligned}$$

we obtain that

$$\begin{aligned}
&\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2+\lambda_n\left(\|\hat{\beta}_S\|_1+\|\hat{\beta}_{S^c}\|_1+\frac{\alpha}{2}|\hat{\beta}_S|^\top R_{SS}|\hat{\beta}_S|+\frac{\alpha}{2}\sum_{(j,k)\notin S\times S}R_{jk}|\hat{\beta}_j\hat{\beta}_k|\right)\\
&\le\frac{1}{3}\lambda_n(\|\hat{\beta}_S-\beta_S^*\|_1+\|\hat{\beta}_{S^c}\|_1)+\lambda_n\left(\|\beta_S^*\|_1+\frac{\alpha}{2}|\beta_S^*|^\top R_{SS}|\beta_S^*|\right)\\
\Rightarrow&\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2+\lambda_n\left(\frac{2}{3}\|\hat{\beta}_{S^c}\|_1+\frac{\alpha}{2}\sum_{(j,k)\notin S\times S}R_{jk}|\hat{\beta}_j\hat{\beta}_k|\right)\\
&\le\frac{1}{3}\lambda_n\|\hat{\beta}_S-\beta_S^*\|_1+\lambda_n\left(\|\beta_S^*\|_1-\|\hat{\beta}_S\|_1+\alpha\|R_{SS}|\beta_S^*|\|_\infty\|\beta_S^*-\hat{\beta}_S\|_1+\frac{\alpha D}{2}\|\beta_S^*-\hat{\beta}_S\|_1^2\right)\\
\Rightarrow&\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2+\lambda_n\left(\frac{2}{3}\|\hat{\beta}_{S^c}\|_1+\frac{\alpha}{2}\sum_{(j,k)\notin S\times S}R_{jk}|\hat{\beta}_j\hat{\beta}_k|\right)\\
&\le\lambda_n\left(\frac{4}{3}\|\hat{\beta}_S-\beta_S^*\|_1+\alpha\|R_{SS}|\beta_S^*|\|_\infty\|\beta_S^*-\hat{\beta}_S\|_1+\frac{\alpha D}{2}\|\beta_S^*-\hat{\beta}_S\|_1^2\right). \qquad (3.33)
\end{aligned}$$

On the other hand, (3.32) also gives

$$\begin{aligned}
&\|\hat{\beta}_S\|_1+\|\hat{\beta}_{S^c}\|_1\le\frac{1}{3}(\|\hat{\beta}_S-\beta_S^*\|_1+\|\hat{\beta}_{S^c}\|_1)+\|\beta_S^*\|_1+\frac{\alpha}{2}|\beta_S^*|^\top R_{SS}|\beta_S^*|\\
\Rightarrow\quad&\frac{2}{3}\|\hat{\beta}_S-\beta_S^*\|_1+\frac{2}{3}\|\hat{\beta}_{S^c}\|_1\le 2\|\beta_S^*\|_1+\frac{\alpha}{2}|\beta_S^*|^\top R_{SS}|\beta_S^*|\\
\Rightarrow\quad&\|\hat{\beta}_S-\beta_S^*\|_1\le 3\|\beta_S^*\|_1+\frac{3}{4}\alpha|\beta_S^*|^\top R_{SS}|\beta_S^*|\\
\Rightarrow\quad&\|\hat{\beta}_S-\beta_S^*\|_1\le\left(3+\frac{3}{4}\alpha\|R_{SS}|\beta_S^*|\|_\infty\right)\|\beta_S^*\|_1.
\end{aligned}$$

Therefore, (3.33) gives

$$\begin{aligned}
&\frac{2}{3}\|\hat{\beta}_{S^c}\|_1+\frac{\alpha}{2}\sum_{(j,k)\notin S\times S}R_{jk}|\hat{\beta}_j\hat{\beta}_k|\\
&\le\left(\frac{4}{3}+\alpha\|R_{SS}|\beta_S^*|\|_\infty+\frac{3}{2}\alpha D\|\beta_S^*\|_1\left(1+\frac{\alpha}{4}\|R_{SS}|\beta_S^*|\|_\infty\right)\right)\|\hat{\beta}_S-\beta_S^*\|_1.
\end{aligned} \qquad (3.34)$$

The second term of the left side is evaluated as

$$\begin{aligned}
\sum_{(j,k)\notin S\times S}R_{jk}|\hat{\beta}_j\hat{\beta}_k|=&\sum_{j\in S^c,k\in S^c}R_{jk}|\hat{\beta}_j\hat{\beta}_k|+2\sum_{j\in S,k\in S^c}R_{jk}|(\hat{\beta}_j-\beta_S^*+\beta_S^*)\hat{\beta}_k|\\
=&|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}|+2|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S-\beta_S^*+\beta_S^*|.
\end{aligned}$$

Hence, (3.34) gives

$$\begin{aligned}
&\frac{2}{3}\|\hat{\beta}_{S^c}\|_1+\frac{\alpha}{2}|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}|+\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S-\beta_S^*+\beta_S^*|\\
&\leq\left(\frac{4}{3}+\alpha\|R_{SS}|\beta_S^*|\|_\infty+\frac{3}{2}\alpha D\|\beta_S^*\|_1\left(1+\frac{\alpha}{4}\|R_{SS}|\beta_S^*|\|_\infty\right)\right)\|\hat{\beta}_S-\beta_S^*\|_1\\
\Rightarrow\quad&\|\hat{\beta}_{S^c}\|_1+\frac{3}{4}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}|+\frac{3}{2}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S-\beta_S^*+\beta_S^*|\\
&\leq\left(2+\frac{15}{4}\alpha D\|\beta_S^*\|_1+\frac{9}{16}(\alpha D\|\beta_S^*\|_1)^2\right)\|\hat{\beta}_S-\beta_S^*\|_1. \qquad (3.35)
\end{aligned}$$

If $\alpha\leq\frac{1}{4D\|\beta_S^*\|_1}$, we have

$$\|\hat{\beta}_{S^c}\|_1+\frac{3}{4}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}|+\frac{3}{2}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S-\beta_S^*+\beta_S^*|\leq 3\|\hat{\beta}_S-\beta_S^*\|_1.$$

Therefore, we can see that

$$\hat{v}\in\mathcal{B}(S,C,C'),$$

where $\hat{v}=\hat{\beta}-\beta^*$, $C=3$ and $C'=\frac{3}{2}$. By applying the definition of ϕ_{GRE} to (3.33), it holds that

$$\frac{\phi_{\mathrm{GRE}}}{2}\|\hat{\beta}-\beta^*\|_2^2\leq\lambda_n\left(\frac{4}{3}+\frac{5}{2}\alpha D\|\beta_S^*\|_1+\frac{3}{8}(\alpha D\|\beta_S^*\|_1)^2\right)\|\hat{\beta}_S-\beta_S^*\|_1.$$

Because $\|\hat{\beta}_S-\beta_S^*\|_1^2\leq s\|\hat{\beta}_S-\beta_S^*\|_2^2$, we have

$$\begin{aligned}
&\|\hat{\beta}-\beta^*\|_2\leq\frac{\left(\frac{8}{3}+5\alpha D\|\beta_S^*\|_1+\frac{3}{4}(\alpha D\|\beta_S^*\|_1)^2\right)\sqrt{s}\lambda_n}{\phi_{\mathrm{GRE}}}\\
\Rightarrow\quad&\|\hat{\beta}-\beta^*\|_2^2\leq\frac{\left(\frac{8}{3}+5\alpha D\|\beta_S^*\|_1+\frac{3}{4}(\alpha D\|\beta_S^*\|_1)^2\right)^2 s\lambda_n^2}{\phi_{\mathrm{GRE}}^2}\leq\frac{16s\lambda_n^2}{\phi_{\mathrm{GRE}}^2}. \qquad (3.36)
\end{aligned}$$

This concludes the assertion. □

3.6.5 Corollary of Theorem 7

For comparison with the IILasso and Lasso, we use the following a little bit stricter bound.

Corollary 11. *Suppose the same assumption of Theorem 7 except for Assumption GRE*$(S,3,\frac{3}{2})$. *Instead, suppose that Assumption GRE*$(S,C,\frac{3}{2})$ *(Definition*

4) where $C = 2 + \frac{15}{4}\alpha D\|\beta^*_S\|_1 + \frac{9}{16}(\alpha D\|\beta^*_S\|_1)^2$ is satisfied. Then, it holds that

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{\left(\frac{8}{3} + 5\alpha D\|\beta^*_S\|_1 + \frac{3}{4}(\alpha D\|\beta^*_S\|_1)^2\right)^2 s\lambda_n^2}{\phi^2_{\mathrm{GRE}}},$$

with probability at least $1 - \delta$.

Proof. This is derived basically in the same way as Theorem 7. From (3.35), we can directly see that

$$\hat{v} \in \mathcal{B}(S, C, C'),$$

where $\hat{v} = \hat{\beta} - \beta^*$, $C = 2 + \frac{15}{4}\alpha D\|\beta^*_S\|_1 + \frac{9}{16}(\alpha D\|\beta^*_S\|_1)^2$, and $C' = \frac{3}{2}$. This and (3.36) concludes the assertion. □

From this corollary, we compare the IILasso with $R_{SS} = O$ and Lasso.

- If $\alpha = 0$, we have

  $$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{64 s\lambda_n^2}{9\phi^2_{\mathrm{GRE}}},$$

  with $\mathcal{B}(S, C, C')$ where $C = 2$ and $C' = 0$. This is a standard Lasso result.

- If $D = 0$, we have

  $$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{64 s\lambda_n^2}{9\phi^2_{\mathrm{GRE}}},$$

  with $\mathcal{B}(S, C, C')$ where $C = 2$ and $C' = \frac{3}{2}$. Since ϕ_{GRE} is the minimum eigenvalue restricted by $\mathcal{B}(S, C, C')$, ϕ_{GRE} of the IILasso is larger than that of the Lasso.

3.6.6 Proof of Theorem 8

Proof. Let

$$\check{\beta} := \mathop{\arg\min}_{\beta \in \mathbb{R}^p : \beta_{S^c} = 0} \|y - X\beta\|_2^2.$$

That is, $\check{\beta}$ is the least squares estimator with the true non-zero coefficients. Let $\tilde{\beta}$ be a local optimal solution. For $0 < h < 1$, letting $\beta(h) := \tilde{\beta} + h(\check{\beta} - \tilde{\beta})$, then it holds that

$$\begin{aligned}\mathcal{L}(\beta(h)) - \mathcal{L}(\tilde{\beta}) =& \frac{h^2-2h}{2n}\|X(\tilde{\beta}-\breve{\beta})\|_2^2 - \frac{h}{n}(X\breve{\beta}-y)^\top X(\tilde{\beta}-\breve{\beta}) \\ &+ \lambda_n(\|\beta(h)\|_1 - \|\tilde{\beta}\|_1) + \frac{\lambda_n\alpha}{2}(|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}|).\end{aligned} \tag{3.37}$$

First we evaluate the term $\frac{1}{n}(X\breve{\beta}-y)^\top X(\tilde{\beta}-\breve{\beta}) = \frac{1}{n}(X\breve{\beta}-y)^\top X_S(\tilde{\beta}_S-\breve{\beta}_S) + \frac{1}{n}(X\breve{\beta}-y)^\top X_{S^c}(\tilde{\beta}_{S^c}-\breve{\beta}_{S^c})$ as follows:
(1) Since $\breve{\beta}$ is the least squares estimator and $\frac{1}{n}X_S^\top X_S$ is invertible by the assumption, we have

$$\breve{\beta}_S = (X_S^\top X_S)^{-1}X_S^\top y, \quad \breve{\beta}_{S^c} = 0.$$

Therefore,

$$\frac{1}{n}X_S^\top(X\breve{\beta}-y) = \frac{1}{n}X_S^\top(X_S(X_S^\top X_S)^{-1}X_S^\top - I)y.$$

Here, $I - X_S(X_S^\top X_S)^{-1}X_S^\top$ is the projection matrix to the orthogonal complement of the image of $X_S^\top$. Hence, $\frac{1}{n}(X\breve{\beta}-y)^\top X_S(\tilde{\beta}_S-\breve{\beta}_S) = 0$.
(2) Noticing that

$$\begin{aligned}\frac{1}{n}X_{S^c}^\top(X\breve{\beta}-y) &= -\frac{1}{n}X_{S^c}^\top(I - X_S(X_S^\top X_S)^{-1}X_S^\top)y \\ &= -\frac{1}{n}X_{S^c}^\top(I - X_S(X_S^\top X_S)^{-1}X_S^\top)(X_S\beta_S^* + \varepsilon) \\ &= -\frac{1}{n}X_{S^c}^\top(I - X_S(X_S^\top X_S)^{-1}X_S^\top)\varepsilon,\end{aligned}$$

where we used $(I - X_S(X_S^\top X_S)^{-1}X_S^\top)X_{S^c} = 0$ in the last line. Because $(I - X_S(X_S^\top X_S)^\top X_S^\top)$ is a projection matrix, we have $\|(I-X_S(X_S^\top X_S)^{-1}X_S^\top)X_j\|_2^2 \leq \|X_j\|_2^2$. This and Corollary 10 gives

$$\left\|\frac{1}{n}X_{S^c}^\top(X\breve{\beta}-y)\right\|_\infty \leq \gamma_n,$$

with probability $1-\delta$. Hence, let $V := \mathrm{supp}(\tilde{\beta})\backslash S$, then we have

$$\left|\frac{1}{n}(\tilde{\beta}_{S^c}-\breve{\beta}_{S^c})^\top X_{S^c}^\top(X\breve{\beta}-y)\right| \leq \gamma_n\|\tilde{\beta}_{S^c}-\breve{\beta}_{S^c}\|_1 = \gamma_n\|\tilde{\beta}_V\|_1.$$

where we used the assumption $V \subseteq S^c$ and $\breve{\beta}_V = 0$.

Combining these inequalities, we have that

$$\left|\frac{1}{n}(X\check{\beta}-y)^\top X(\tilde{\beta}-\check{\beta})\right| \leq \gamma_n\|\tilde{\beta}_V\|_1. \tag{3.38}$$

As for the regularization term, we evaluate each term of $\lambda_n(\|\beta(h)\|_1 - \|\tilde{\beta}\|_1) + \frac{\lambda_n}{2}(|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}|)$ in the following.

(i) Evaluation of $\|\beta(h)\|_1 - \|\tilde{\beta}\|_1$. Because of the definition of $\beta(h)$, it holds that

$$\begin{aligned}\|\beta(h)\|_1 - \|\tilde{\beta}\|_1 &= \|\tilde{\beta} + h(\check{\beta}-\tilde{\beta})\|_1 - \|\tilde{\beta}\|_1 \\ &= \|\tilde{\beta}_S + h(\check{\beta}_S - \tilde{\beta}_S)\|_1 - \|\tilde{\beta}_S\|_1 + \|\tilde{\beta}_V + h(\check{\beta}_V - \tilde{\beta}_V)\|_1 - \|\tilde{\beta}_V\|_1 \\ &= \|\tilde{\beta}_S + h(\check{\beta}_S - \tilde{\beta}_S)\|_1 - \|\tilde{\beta}_S\|_1 + (1-h)\|\tilde{\beta}_V\|_1 - \|\tilde{\beta}_V\|_1 \\ &\leq h\|\check{\beta}_S - \tilde{\beta}_S\|_1 - h\|\tilde{\beta}_V\|_1. \end{aligned} \tag{3.39}$$

(ii) Evaluation of $|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}|$. Note that

$$\begin{aligned}&|\beta(h)_j|R_{jk}|\beta(h)_k| - |\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| \\ &= |(1-h)\tilde{\beta}_j + h\check{\beta}_j|R_{jk}|(1-h)\tilde{\beta}_k + h\check{\beta}_k| - |\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| \\ &\leq (1-h)^2|\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| + h(1-h)(|\check{\beta}_j|R_{jk}|\tilde{\beta}_k| + |\tilde{\beta}_j|R_{jk}|\check{\beta}_k|) \\ &\quad + h^2|\check{\beta}_j|R_{jk}|\check{\beta}_k| - |\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| \\ &= -2h|\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| + h(|\check{\beta}_j|R_{jk}|\tilde{\beta}_k| + |\tilde{\beta}_j|R_{jk}|\check{\beta}_k|) + O(h^2) \\ &= h[(|\check{\beta}_j| - |\tilde{\beta}_j|)R_{jk}|\tilde{\beta}_k| + |\tilde{\beta}_j|R_{jk}(|\check{\beta}_k| - |\tilde{\beta}_k|)] + O(h^2). \end{aligned} \tag{3.40}$$

If $j, k \in S$, then the right hand side of Eq. (3.40) is bounded by

$$\begin{aligned}&h(|\check{\beta}_j - \tilde{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k| + |\check{\beta}_j - \tilde{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k|) \\ &\quad + h(|\check{\beta}_j - \tilde{\beta}_j|R_{jk}|\check{\beta}_k| + |\check{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k|) + O(h^2).\end{aligned}$$

If $j \in V$ and $k \in S$, then the right hand side of Eq. (3.40) is bounded by

$$h|\tilde{\beta}_j|R_{jk}(|\check{\beta}_k| - |\tilde{\beta}_k|) + O(h^2) \leq h|\tilde{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k| + O(h^2).$$

If $j \in V$ and $k \in V$, then the right hand side of Eq. (3.40) is bounded by

$$0 + O(h^2) = O(h^2).$$

Based on these evaluations, we have

$$\begin{aligned}
&|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}| \\
\le& 2h\left(|\check{\beta}_S - \tilde{\beta}_S|^\top R_{SS}|\check{\beta}_S - \tilde{\beta}_S| + |\check{\beta}_S - \tilde{\beta}_S|^\top R_{SS}|\check{\beta}_S| + |\tilde{\beta}_V|^\top R_{VS}|\check{\beta}_S - \tilde{\beta}_S|\right) + O(h^2) \\
\le& 2h\left(|\check{\beta} - \tilde{\beta}|^\top R|\check{\beta} - \tilde{\beta}| + |\check{\beta}_S - \tilde{\beta}_S|^\top R_{SS}|\check{\beta}_S|\right) + O(h^2) \\
\le& 2h\bar{D}(\|\check{\beta} - \tilde{\beta}\|_2^2 + \|\check{\beta}\|_2\|\check{\beta}_S - \tilde{\beta}_S\|_2) + O(h^2).
\end{aligned}$$

Here, we will show later in Eq. (3.42) that $\|\check{\beta} - \beta^*\|_2 \le \sqrt{s}\lambda_n/\phi$, and thus it follows that

$$\|\check{\beta}\|_2 \le \|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi.$$

Therefore, we obtain that

$$\begin{aligned}
&|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}| \\
\le& 2h\bar{D}\left(\|\check{\beta} - \tilde{\beta}\|_2^2 + (\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2\right) + O(h^2).
\end{aligned} \tag{3.41}$$

Applying the inequalities (3.38), (3.39) and (3.41) to (3.37) yields that

$$\begin{aligned}
&\mathcal{L}(\beta(h)) - \mathcal{L}(\tilde{\beta}) \\
\le& h\Big\{ - \frac{1}{n}\|X(\check{\beta} - \tilde{\beta})\|_2^2 + \lambda_n\|\tilde{\beta}_S - \check{\beta}_S\|_1 - (\lambda_n - \gamma_n)\|\tilde{\beta}_V\|_1 \\
&+ \lambda_n\alpha\bar{D}[\|\check{\beta} - \tilde{\beta}\|_2^2 + (\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2]\Big\} + O(h^2) \\
\le& h\Big\{ - \phi\|\check{\beta} - \tilde{\beta}\|_2^2 + \lambda_n\|\tilde{\beta}_S - \check{\beta}_S\|_1 \\
&+ \lambda_n\alpha\bar{D}[\|\check{\beta} - \tilde{\beta}\|_2^2 + (\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2]\Big\} + O(h^2) \\
\le& h\left\{ \left(-\phi + \lambda_n\alpha\bar{D}\right)\|\check{\beta} - \tilde{\beta}\|_2^2 \right. \\
&\left. + \lambda_n\left(\|\tilde{\beta}_S - \check{\beta}_S\|_1 + \alpha\bar{D}(\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2\right) \right\} + O(h^2),
\end{aligned}$$

where we used the assumption $\lambda_n > \gamma_n$ in the second inequality.

Since we have assumed $\alpha < \min\left\{\frac{\sqrt{s}}{2\bar{D}\|\beta^*\|_2}, \frac{\phi}{2\bar{D}\lambda_n}\right\}$, the right hand side is further bounded by

$$h\left\{-\frac{\phi}{2}\|\check{\beta} - \tilde{\beta}\|_2^2 + 2\lambda_n\sqrt{s}\|\check{\beta}_S - \tilde{\beta}_S\|_2\right\} + O(h^2).$$

Because of this, if $\|\check{\beta} - \tilde{\beta}\|_2 > \frac{4\sqrt{s}\lambda_n}{\phi}$, then the first term becomes negative, and we conclude that, for sufficiently small $\eta > 0$, it holds that

$$\mathcal{L}(\beta(h)) < \mathcal{L}(\tilde{\beta}),$$

for all $0 < h < \eta$. In other word, $\tilde{\beta}$ is not a local optimal solution. Therefore, we must have

$$\|\check{\beta} - \tilde{\beta}\|_2 \leq \frac{4\sqrt{s}\lambda_n}{\phi}$$

Finally, notice that $\|\tilde{\beta} - \beta^*\|_2^2 \leq (\|\tilde{\beta} - \check{\beta}\|_2 + \|\beta^* - \check{\beta}\|_2)^2$ and

$$\begin{aligned}\|\check{\beta} - \beta^*\|_2^2 &= \|(X_S^\top X_S)^{-1} X_S^\top y - \beta_S^*\|_2^2 = \|(X_S^\top X_S)^{-1} X_S^\top (X_S \beta_S^* + \varepsilon) - \beta_S^*\|_2^2 \\ &= \|(X_S^\top X_S)^{-1} X_S^\top \varepsilon\|_2^2 \leq \phi^{-2} \|\frac{1}{n} X_S^\top \varepsilon\|_2^2 \leq \phi^{-2} s \gamma_n^2 \leq \phi^{-2} s \lambda_n^2, \quad (3.42)\end{aligned}$$

which concludes the assertion. □

Chapter 4

Extensions to Generalized Linear Models

4.1 Method for GLMs

The idea of the IILasso is not restricted in linear models with quadratic loss. In this chapter, we extend the method and theory to a more general setting, especially for generalized linear models (GLMs) with general loss without assuming the truth is included in GLMs.

Consider independent and identically distributed data $\{(X_i, Y_i)\}_{i=1}^n$ with $(X_i, Y_i) \in \mathscr{X} \times \mathscr{Y}$ where X_i is a fixed covariable in some space $\mathscr{X}$ and Y_i is a response variable in some space $\mathscr{Y} \subset \mathbb{R}$. Note that we suppose X_i's are deterministic variables. Let $\boldsymbol{F}$ be the space of the whole measurable functions. Let $\rho_f : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ be a general loss function for a model $f \in \boldsymbol{F}$. We also denote $\rho_f(x,y) := \rho(f(x), y)$. Typical examples of loss functions include quadratic loss $\rho_f(x,y) = \rho(f(x),y) = (y - f(x))^2$ for $y \in \mathbb{R}$, and logistic loss $\rho_f(x,y) = \rho(f(x),y) = -yf(x) + \log(1 + \exp(f(x)))$ for $y \in \{0,1\}$. Consider a generalized linear model subspace

$$\mathscr{F} := \left\{ f_\beta(x) := \sum_{j=1}^p \beta_j \psi_j(x) : \beta \in \mathbb{R}^p, x \in \mathscr{X} \right\} \subset \boldsymbol{F}.$$

Here, the map $x \mapsto (\psi_1(x), \ldots, \psi_p(x))^\top \in \mathbb{R}^p$ is an arbitrary measurable feature map from the space $\mathscr{X}$ to the space $\mathbb{R}^p$. We define the population mean risk

$$P\rho_f := \frac{1}{n} \sum_{i=1}^n \mathrm{E}_{Y_i|X_i} \left[\rho_f(X_i, Y_i) | X_i \right],$$

and the empirical average risk

$$P_n\rho_f := \frac{1}{n}\sum_{i=1}^{n}\rho_f(X_i, Y_i).$$

We further define the target as the minimizer of the theoretical mean risk

$$f^0 := \arg\min_{f\in\boldsymbol{F}} P\rho_f. \tag{4.1}$$

Note that we allow model misspecification, in other words, it may hold $f^0 \notin \mathscr{F}$. In this setting, our interest is to estimate GLMs close to the target (4.1). IILasso for GLMs forms

$$\hat{\beta} = \arg\min_{\beta} P_n\rho_{f_\beta} + \lambda\left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right) =: \mathcal{L}(\beta),$$

where $\lambda > 0$ and $\alpha > 0$ are regularization parameters, and $R \in \mathbb{R}^{p\times p}$ is a symmmetric matrix whose component $R_{jk} \geq 0$ is typically a monotonically increasing function of the absolute correlation $r_{jk} = |\sum_i \psi_j(X_i)^\top\psi_k(X_i)|/n$ for $j \neq k$ and $r_{jk} = 0$ for $j = k$.

4.2 Theoretical Properties for GLMs

We introduce some additional notations. Let $\|f\|_\infty$ be L_∞-norm, i.e., $\|f\|_\infty = \sup_z |f(z)|$. Let $\|f\|$ be $L_2(Q_n)$-norm with Q_n the empirical measure of $\{z_i\}_{i=1}^n$, i.e., $\|f\| := \|f\|_n := \sqrt{\sum_{i=1}^n f^2(z_i)/n}$. We denote S as an arbitrary variable index set and $S_\beta := \mathrm{supp}(\beta)$. Note that we do *not* use S as the true active variables in this chapter, because it may hold $f^0 \notin \mathscr{F}$. We define the excess risk as

$$\mathcal{E}(f) := P(\rho_f - \rho_{f^0}).$$

Next, we prepare some key assumptions and definitions.

Assumption 4. *We assume the loss function is a Lipschitz loss with L, i.e.,*

$$|\rho(a, y) - \rho(a', y)| \leq L\,|a - a'|\,,\ \forall a, a', y.$$

Definition 5 (Quadratic Margin Condition)**.** *We say that the quadratic margin condition holds with* κ *for* $\boldsymbol{F}_\eta := \{f \in \boldsymbol{F} : \|f - f^0\|_\infty \leq \eta\}$*, if we have*

$$\mathcal{E}(f) \geq \kappa \left\|f - f^0\right\|^2, \ \forall f \in \boldsymbol{F}_\eta.$$

Definition 6 (Generalized Restricted Eigenvalue Condition for GLMs (GRE (S, C, C')))**.** *Let a set of vectors* $\mathcal{B}(S, C, C')$ *be*

$$\mathcal{B}(S, C, C') := \Big\{ v \in \mathbb{R}^p : \|v_{S^c}\|_1 + \frac{C'\alpha}{2} |v_{S^c}|^\top R_{S^c S^c} |v_{S^c}| \\ + \alpha C' |v_{S^c}|^\top R_{S^c S} |v_S + \beta^*_S(S)| \leq C \|v_S\|_1 \Big\}.$$

where $\beta^*(S) := \arg\min_{\beta : S_\beta = S} \mathcal{E}(f_\beta)$*. We say that the generalized restricted eigenvalue condition for GLMs holds with a set of vectors* $\mathcal{B}(S, C, C')$*, if we have* $\phi_{\mathrm{GRE}} > 0$ *where*

$$\phi_{\mathrm{GRE}} = \phi_{\mathrm{GRE}}(S, C, C') := \inf_{v \in \mathcal{B}(S, C, C')} \frac{\|f_v\|^2}{\|v\|_2^2}.$$

Definition 7 (Oracle)**.** *Let* $\mathscr{S}$ *be a collection of variable index sets. We define the oracle* β^* *as*

$$\beta^* = \beta^*(\mathscr{S}) := \arg \min_{\beta : S_\beta \in \mathscr{S}} \left\{ 3\mathcal{E}(f_\beta) + \frac{9\lambda_n^2 |S_\beta|}{\kappa \phi^*_{\mathrm{GRE}}} \right\}, \tag{4.2}$$

where $\phi^*_{\mathrm{GRE}} = \phi_{\mathrm{GRE}}(\mathscr{S}) := \inf_{S \in \mathscr{S}} \phi_{\mathrm{GRE}}(S, C, C')$.

These assumptions and definitions are based on (Bühlmann and Van De Geer, 2011) and tailored for our proposed method. The quadratic margin condition (Definition 5) means that the excess risk for every model f near the target f^0 is bounded below by a quadratic function of $\|f - f^0\|$. The GRE condition for GLMs (Definition 6) is a natural extension of the GRE condition for linear models. The oracle β^* (Definition 7) represents an ideal solution because it minimizes the sum of the excess risk and the penalty in proportion to the number of selected variables.

Now, we derive an estimation error bound for GLMs.

Theorem 12. *Let Assumption 4 with a constant* L *be satisfied. Suppose that* $\|\psi_j\| \leq 1$ *for* $\forall j$*. Let* $\mathscr{S}$ *be an arbitrary collection of variable index sets. For every variable index set* $S \in \mathscr{S}$*, suppose that assumption GRE*$(S, 3, 1)$ *(Definition 6) is satisfied. Define the oracle* β^* *as* (4.2) *(Definition 7) for* $\mathscr{S}$ *and some* *constant* $\kappa > 0$. *Fix any* $0 < \delta < 1$, *define*

$$\gamma_n = \gamma_n(\delta) := 2L\left(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right), \tag{4.3}$$

$$M^* := \frac{1}{\gamma_n}\left(3\mathcal{E}(f_{\beta^*}) + \frac{9\lambda_n^2 |S_{\beta^*}|}{\kappa\phi^*_{\mathrm{GRE}}}\right).$$

Suppose the quadratic margin condition (Definition 5) holds with κ *for* $\boldsymbol{F}_\eta$. *In addition, assume* $f_\beta \in \boldsymbol{F}_\eta$ *for all* $\|\beta - \beta^*\|_1 \le M^*$, *as well as* $f_{\beta^*} \in \boldsymbol{F}_\eta$. *Let* $\hat\beta$ *be an approximately minimizer of the IILasso such that* $\mathcal{L}(\hat\beta) \le \mathcal{L}(\beta^*)$ *and* $\mathcal{L}(t\hat\beta + (1-t)\beta^*) \le \mathcal{L}(\beta^*)$ *for* $t = M^*/(M^* + \|\hat\beta - \beta^*\|_1)$. *Let the regularization parameters satisfy*

$$\alpha \le \frac{1}{12D\|\beta^*\|_1} \text{ and } 20\gamma_n \le 3\lambda_n,$$

where $D := \|R_{S_*S_*}\|_{\max}$ *and* $S_* := \mathrm{supp}(\beta^*)$. *Then, we have*

$$\|\hat\beta - \beta^*\|_2^2 \le \frac{22\mathcal{E}(f_{\beta^*})}{\kappa\phi^*_{\mathrm{GRE}}} + \frac{60\lambda_n^2|S_*|}{\kappa^2\phi^{*2}_{\mathrm{GRE}}},$$

with probability at least $1-\delta$.

The proof is given in 4.4.1. The obtained covergence rate in Theorem 12 is roughly evaluated as

$$\|\hat\beta - \beta^*\|_2^2 = O_P\left(\frac{|S_*|\log(p)}{n}\right),$$

under $\mathcal{E}(f_{\beta^*}) = o_p(1)$, $\phi^*_{\mathrm{GRE}} = O_p(1)$, and $\lambda_n = O_p(\sqrt{\log p/n})$, which is almost the minimax optimal rate (Raskutti, Wainwright, and Yu, 2011). If $D = 0$, then ϕ^*_{GRE} gets larger as $R_{S_*^cS_*}$ and $R_{S_*^cS_*^c}$ become large. Hence, the IILasso for general loss with quadratic margin is preferable when true active variables are not correlated and others are strongly correlated.

The collection of variable index sets $\mathscr{S}$ is arbitrary. Taking $\mathscr{S}$ as specific definitions yields the following corollaries.

Corollary 13. *Define the best linear approximation*

$$\beta^0_{\mathrm{GLM}} := \arg\min_\beta P\rho_{f_\beta}.$$

Take $\mathscr{S} = \{S^0_{\mathrm{GLM}}\} := \{\mathrm{supp}(\beta^0_{\mathrm{GLM},j})\}$*. Then, under the same assumptions as in Theorem 12, we have* $\beta^* = \beta^0_{\mathrm{GLM}}$ *and*

$$\|\hat{\beta} - \beta^0_{\mathrm{GLM}}\|_2^2 \leq \frac{22\mathcal{E}(f_{\beta^0_{\mathrm{GLM}}})}{\kappa\phi_{\mathrm{GRE}}(S^0_{\mathrm{GLM}}, 3, 1)} + \frac{60\lambda_n^2|S^0_{\mathrm{GLM}}|}{\kappa^2\phi_{\mathrm{GRE}}(S^0_{\mathrm{GLM}}, 3, 1)^2}.$$

Proof.

$$\beta^* = \arg\min_{\beta: S_\beta = S^0_{\mathrm{GLM}}} \left\{3\mathcal{E}(f_\beta) + \frac{9\lambda_n^2|S^0_{\mathrm{GLM}}|}{\kappa\phi^*_{\mathrm{GRE}}}\right\} = \beta^0_{\mathrm{GLM}},$$

and $\phi^*_{\mathrm{GRE}} = \phi_{\mathrm{GRE}}(S^0_{\mathrm{GLM}}, 3, 1)$ concludes the assertion. □

Corollary 14. *Suppose the true model is linear, i.e.,*

$$f^0 = f_{\beta^0} \text{ and } \mathcal{E}(f_{\beta^0}) = 0.$$

Take $\mathscr{S} = \{S^0\} := \{\mathrm{supp}(\beta^0)\}$*. Then, under the same assumptions as in Theorem 12, we have* $\beta^* = \beta^0$ *and*

$$\|\hat{\beta} - \beta^0\|_2^2 \leq \frac{60\lambda_n^2|S^0|}{\kappa^2\phi_{\mathrm{GRE}}(S^0, 3, 1)^2}.$$

Proof.

$$\beta^* = \arg\min_{\beta: S_\beta = S^0} \left\{3\mathcal{E}(f_\beta) + \frac{9\lambda_n^2|S^0|}{\kappa\phi^*_{\mathrm{GRE}}}\right\} = \beta^0,$$

and $\phi^*_{\mathrm{GRE}} = \phi_{\mathrm{GRE}}(S^0, 3, 1)$ concludes the assertion. □

Corollary 13 shows that the error between the IILasso estimate and the best linear approximation is bounded by the excess risk of the best linear approximation, the number of active variables in the best linear approximation, and some constants depending on n and p. Also, Corollary 14 shows that the error between the IILasso estimate and the true linear model (if exists) is bounded by the number of true active variables and some constants depending on n and p.

4.3 Extension to Logistic Models

We examine logistic regression models as an illustration. Consider data $\{(X_i, Y_i)\}_{i=1}^n$ with $(X_i, Y_i) \in \mathbb{R}^p \times \{0, 1\}$. Define $\pi(x) := P(Y = 1|X = x)$. The logistic model is

$$\log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = f(x) = \mu + \sum_{j=1}^{p} \beta_j x_j,$$

and the logistic loss is

$$\rho_f(x, y) = \rho(f(x), y) := -yf(x) + \log(1 + \exp(f(x))).$$

We take a target as

$$f^0(x) := \log\left(\frac{\pi(x)}{1-\pi(x)}\right),$$

because it minimizes $\mathrm{E}_{Y|X}[\rho_f(X, Y)|X = x]$.

We derive coordinate descent algorithm of the IILasso for logistic regression. The objective function is

$$\mathcal{L}(\beta) = -\frac{1}{n}\sum_i \left(y_i X^i\beta - \log(1 + \exp(X^i\beta))\right) + \lambda\left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right),$$

where X^i is the i-th row of $X = [1, X_1, \cdots, X_p]$ and $\beta = [\beta_0, \beta_1, \cdots, \beta_p]$. Forming a quadratic approximation with the current estimate $\bar{\beta}$, we have

$$\bar{L}(\beta) = -\frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - X^i\beta)^2 + C(\bar{\beta}) + \lambda\left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right),$$

where

$$\begin{aligned} z_i &= X^i\bar{\beta} + \frac{y_i - \bar{p}(X^i)}{\bar{p}(X^i)(1-\bar{p}(X^i))}, \\ w_i &= \bar{p}(X^i)(1-\bar{p}(X^i)), \\ \bar{p}(X^i) &= \frac{1}{1+\exp(-X^i\bar{\beta})}. \end{aligned}$$

Algorithm 2 CDA for Logistic IILasso

```
for $\lambda = \lambda_{\max}, \cdots, \lambda_{\min}$ do
  initialize $\beta$
  while until convergence do
    update the quadratic approximation using the current parameters $\bar{\beta}$
    while until convergence do
      for $j = 1, \cdots, p$ do
        $\beta_j \leftarrow \frac{1}{\frac{1}{n}\sum_{i=1}^{n} w_i X_{ij}^2 + \lambda\alpha R_{jj}} S\left(\frac{1}{n}\sum_{i=1}^{n} w_i \left(z_i - X_{i,-j}\beta_{-j}\right) X_{ij},\ \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\right)$
      end for
    end while
  end while
end for
```

To derive the update equation, when $\beta_j \neq 0$, differentiating the quadratic objective function with respect to β_j yields

$$
\begin{aligned}
\partial_{\beta_j} \bar{L}(\beta) &= -\frac{1}{n}\sum_{i=1}^{n} w_i (z_i - X^i\beta) X_{ij} + \lambda\left(\mathrm{sgn}(\beta_j) + \alpha R_j^\top |\beta| \,\mathrm{sgn}(\beta_j)\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} w_i \left(z_i - X_{i,-j}\beta_{-j}\right) X_{ij} + \left(\frac{1}{n}\sum_{i=1}^{n} w_i X_{ij}^2 + \lambda R_{jj}\right)\beta_j \\
&\quad + \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\mathrm{sgn}(\beta_j).
\end{aligned}
$$

This yields

$$
\beta_j \leftarrow \frac{1}{\frac{1}{n}\sum_{i=1}^{n} w_i X_{ij}^2 + \lambda\alpha R_{jj}} \mathcal{S}\left(\frac{1}{n}\sum_{i=1}^{n} w_i \left(z_i - X_{i,-j}\beta_{-j}\right) X_{ij},\ \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\right).
$$

These procedures amount to a sequence of nested loops. The whole algorithm is described in Algorithm 2.

We obtain the following estimation error bound.

Corollary 15. *Suppose that for some constant* $0 < \varepsilon_0 < 1$,

$$
\varepsilon_0 < \pi(x) < 1 - \varepsilon_0,
$$

for $\forall x \in \mathcal{X}$. *Suppose that* $\|X_j\| \leq 1$ *for* $\forall j$. *Suppose that assumption GRE*$(S, 3, 1)$ *(Definition 6) is satisfied for every variable index set* $S \in \mathscr{S}$. *Define the oracle* β^* *as* (4.2) *(Definition 7) with* $\kappa = 1/(\exp(\eta)/\varepsilon_0 + 1)^2$. *Fix any* $0 < \delta < 1$ *and*

$\eta > 0$, *define*

$$\gamma_n := \gamma_n(\delta) :=2 \left(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \right),$$
$$M^* :=\frac{1}{\gamma_n} \left(3\mathcal{E}(f_{\beta^*}) + \frac{9(\exp(\eta)+\varepsilon_0)^2\lambda_n^2|S_{\beta^*}|}{\varepsilon_0^2\phi_{\mathrm{GRE}}^*} \right).$$

Let the regularization parameters satisfy

$$\alpha \leq \frac{1}{12D\|\beta^*\|_1} \text{ and } \frac{20\gamma_n}{3} \leq \lambda_n \leq T\gamma_n$$

for some constant T. In addition, suppose that

$$\frac{T^2(\exp(\eta)+\varepsilon_0)^2\gamma_n|S_{\beta^*}|}{\varepsilon_0^2\phi_{\mathrm{GRE}}^*} \leq \frac{\eta}{27}, \ \|f^* - f^0\|_\infty \leq \frac{\eta}{3}, \text{ and } \frac{\mathcal{E}(f^*)}{\gamma_n} \leq \frac{\eta}{9}.$$

Let $\hat{\beta}$ be an approximately minimizer of the IILasso such that $\mathcal{L}(\hat{\beta}) \leq \mathcal{L}(\beta^*)$ and $\mathcal{L}(t\hat{\beta}+(1-t)\beta^*) \leq \mathcal{L}(\beta^*)$ for $t = M^*/(M^* + \|\hat{\beta}-\beta^*\|_1)$. Then, we have

$$\|\hat{\beta}-\beta^*\|_2^2 \leq \frac{22(\exp(\eta)+\varepsilon_0)^2\mathcal{E}(f_{\beta^*})}{\varepsilon_0^2\phi_{\mathrm{GRE}}^*} + \frac{60(\exp(\eta)+\varepsilon_0)^4\lambda_n^2|S_{\beta^*}|}{\varepsilon_0^4{\phi_{\mathrm{GRE}}^*}^2},$$

with probability at least $1-\delta$.

The proof is given in 4.4.2.

4.4 Proofs

4.4.1 Proof of Theorem 12

Proof. In the proof, we use the short-hand notation $S_* := S_{\beta^*}$. Let $c_1, c_2, \ldots$ be some constants, which are defined concretely at the end of the proof. Define the empirical process as

$$v_n(\beta) := (P_n - P)\rho_{f_\beta},$$

and let

$$Z_M := \sup_{\|\beta-\beta^*\|_1 \leq M} |v_n(\beta) - v_n(\beta^*)|.$$

First, we prepare the following lemma.

Lemma 16. *Let Assumption 4 hold. Suppose $\|\psi_j\| \leq 1$ for $\forall j$. Then, we have $P(\{Z_M \leq \gamma_n M\}) \geq 1 - \delta$ where γ_n is defined as (4.3).*

Proof. See example 14.2 in section 14.8 (Bühlmann and Van De Geer, 2011) in details. □

Following the lemma, we have $Z_{M^*} \leq \gamma_n M^*$ with high probability, where

$$M^* := \frac{1}{\gamma_n}\left(c_1\mathcal{E}(f_{\beta^*}) + \frac{c_2\lambda_n^2|S_*|}{\kappa\phi^*_{\mathrm{GRE}}}\right),$$

with some positive constants c_1 and c_2. Hereafter, we assume this holds. If $\|\hat{\beta} - \beta^*\|_1 \leq M^*$, which we will show later, then we have

$$\begin{aligned}
&\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\left(\|\hat{\beta}\|_1 + \frac{\alpha}{2}|\hat{\beta}|^\top R|\hat{\beta}|\right)\\
\leq& -\left(v_n(\hat{\beta}) - v_n(\beta^*)\right) + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right)\\
\leq& Z_{M^*} + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right)\\
\leq& \gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(\|\beta^*\|_1 + \frac{\alpha}{2}|\beta^*|^\top R|\beta^*|\right).
\end{aligned}$$

Substituting $\beta = \beta_{S_*} + \beta_{S_*^c}$, we have

$$\begin{aligned}
&\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\Big(\|\hat{\beta}_{S_*}\|_1 + \|\hat{\beta}_{S_*^c}\|_1\\
&+ \frac{\alpha}{2}|\hat{\beta}_{S_*}|^\top R_{S_*S_*}|\hat{\beta}_{S_*}| + \alpha|\hat{\beta}_{S_*^c}|^\top R_{S_*^cS_*}|\hat{\beta}_{S_*}| + \frac{\alpha}{2}|\hat{\beta}_{S_*^c}|^\top R_{S_*^cS_*^c}|\hat{\beta}_{S_*^c}|\Big)\\
\leq& \gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2}|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}|\right).
\end{aligned} \tag{4.4}$$

Using this inequality, we can evaluate $\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\|\hat{\beta} - \beta^*\|_1$ as

$$\begin{aligned}
&\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\|\hat{\beta} - \beta^*\|_1\\
=& \mathcal{E}(f_{\hat{\beta}}) + \lambda_n\|\hat{\beta}_{S_*} - \beta^*_{S_*}\|_1 + \lambda_n\|\hat{\beta}_{S_*^c}\|_1\\
\leq& \gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(2\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2}|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}|\right).
\end{aligned} \tag{4.5}$$

To obtain a tighter bound, we need to use quadratic terms of $\hat{\beta}$. Hereafter, we reparameterize $\hat{v} := \hat{\beta} - \beta^*$. Then, we have, from (4.4),

$$\begin{aligned}
&\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\left(\|\hat{v}_{S_*^c}\|_1 + \alpha|\hat{v}_{S_*^c}|^\top R_{S_*^cS_*}|\hat{v}_{S_*} + \beta^*_{S_*}| + \frac{\alpha}{2}|\hat{v}_{S_*^c}|^\top R_{S_*^cS_*^c}|\hat{v}_{S_*^c}|\right)\\
\leq& \gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(\|\hat{v}_{S_*}\|_1 + \frac{\alpha}{2}\left(|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}| - |\hat{\beta}_{S_*}|^\top R_{S_*S_*}|\hat{\beta}_{S_*}|\right)\right)
\end{aligned}$$

We can evaluate the last two terms as

$$\begin{aligned}
&|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}| - |\hat\beta_{S_*}|^\top R_{S_*S_*}|\hat\beta_{S_*}| \\
=&|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}| - |\beta^*_{S_*}|^\top R_{S_*S_*}|\hat\beta_{S_*}| + |\beta^*_{S_*}|^\top R_{S_*S_*}|\hat\beta_{S_*}| - |\hat\beta_{S_*}|^\top R_{S_*S_*}|\hat\beta_{S_*}| \\
\le&|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*} - \hat\beta_{S_*}| + |\hat\beta_{S_*} - \beta^*_{S_*}|^\top R_{S_*S_*}|\hat\beta_{S_*}| \\
\le&|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*} - \hat\beta_{S_*}| + |\hat\beta_{S_*} - \beta^*_{S_*}|^\top R_{S_*S_*}(|\beta^*_{S_*}| + |\beta^*_{S_*} - \hat\beta_{S_*}|) \\
\le&(2|\beta^*_{S_*}| + |\beta^*_{S_*} - \hat\beta_{S_*}|)^\top R_{S_*S_*}|\beta^*_{S_*} - \hat\beta_{S_*}| \\
\le&2\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty \|\hat v_{S_*}\|_1 + D\|\hat v_{S_*}\|_1^2
\end{aligned}$$

where $D := \|R_{S_*S_*}\|_{\max}$. Hence, we obtain

$$\begin{aligned}
&\mathcal{E}(f_{\hat\beta}) + \lambda_n\left(\|\hat v_{S_*^c}\|_1 + \alpha|\hat v_{S_*^c}|^\top R_{S_*^cS_*}|\hat v_{S_*} + \beta^*_{S_*}| + \frac{\alpha}{2}|\hat v_{S_*^c}|^\top R_{S_*^cS_*^c}|\hat v_{S_*^c}|\right) \\
\le&\gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n\left(\|\hat v_{S_*}\|_1 + \alpha\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\|\hat v_{S_*}\|_1 + \frac{\alpha D}{2}\|\hat v_{S_*}\|_1^2\right).
\end{aligned} \tag{4.6}$$

We further characterize this bound deviding into 2 cases.

(i) If $\lambda_n\left(\|\hat v_{S_*}\|_1 + \alpha\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\|\hat v_{S_*}\|_1 + \frac{\alpha D}{2}\|\hat v_{S_*}\|_1^2\right) \le \gamma_n M^*$, then the inequality (4.6) reduces to

$$\begin{aligned}
&\mathcal{E}(f_{\hat\beta}) + \lambda_n\left(\|\hat v_{S_*^c}\|_1 + \alpha|\hat v_{S_*^c}|^\top R_{S_*^cS_*}|\hat v_{S_*} + \beta^*_{S_*}| + \frac{\alpha}{2}|\hat v_{S_*^c}|R_{S_*^cS_*^c}|\hat v_{S_*^c}|\right) \\
\le&2\gamma_n M^* + \mathcal{E}(f_{\beta^*}).
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\mathcal{E}(f_{\hat\beta}) + \lambda_n\|\hat v\|_1 &= \mathcal{E}(f_{\hat\beta}) + \lambda_n\|\hat v_{S_*^c}\|_1 + \lambda_n\|\hat v_{S_*}\|_1 \\
&\le 3\gamma_n M^* + \mathcal{E}(f_{\beta^*}) \\
&= (3c_1+1)\mathcal{E}(f_{\beta^*}) + \frac{3c_2\lambda_n^2|S_*|}{\kappa\phi^*_{\mathrm{GRE}}},
\end{aligned}$$

where we use $\lambda_n\|\hat v_{S_*}\|_1 \le \gamma_n M^*$ in the second line.

(ii) If $\lambda_n\left(\|\hat v_{S_*}\|_1 + \alpha\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\|\hat v_{S_*}\|_1 + \frac{\alpha D}{2}\|\hat v_{S_*}\|_1^2\right) \ge \gamma_n M^*$, we need the compatibility condition and the margin condition. In this case, the inequality

(4.6) reduces to

$$\begin{aligned}
&\mathcal{E}(f_{\hat\beta}) + \lambda_n \left(\|\hat v_{S_*^c}\|_1 + \alpha |\hat v_{S_*^c}|^\top R_{S_*^c S_*} |\hat v_{S_*} + \beta^*_{S_*}| + \frac{\alpha}{2} |\hat v_{S_*^c}| R_{S_*^c S_*^c} |\hat v_{S_*^c}| \right) \\
\leq& \gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n \left(\|\hat v_{S_*}\|_1 + \alpha \|R_{S_* S_*} |\beta^*_{S_*}|\|_\infty \|\hat v_{S_*}\|_1 + \frac{\alpha D}{2} \|\hat v_{S_*}\|_1^2 \right) \\
\leq& \left(1 + \frac{1}{c_1}\right) \gamma_n M^* + \lambda_n \left(\|\hat v_{S_*}\|_1 + \alpha \|R_{S_* S_*} |\beta^*_{S_*}|\|_\infty \|\hat v_{S_*}\|_1 + \frac{\alpha D}{2} \|\hat v_{S_*}\|_1^2 \right) \\
\leq& \left(2 + \frac{1}{c_1}\right) \lambda_n \left(\|\hat v_{S_*}\|_1 + \alpha \|R_{S_* S_*} |\beta^*_{S_*}|\|_\infty \|\hat v_{S_*}\|_1 + \frac{\alpha D}{2} \|\hat v_{S_*}\|_1^2 \right) \\
\leq& \left(2 + \frac{1}{c_1}\right) \lambda_n \left(1 + \alpha \|R_{S_* S_*} |\beta^*_{S_*}|\|_\infty + \frac{\alpha D}{2} \|\hat v_{S_*}\|_1 \right) \|\hat v_{S_*}\|_1. \qquad (4.7)
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
&\mathcal{E}(f_{\hat\beta}) + \lambda_n \|\hat v\|_1 \\
=& \mathcal{E}(f_{\hat\beta}) + \lambda_n \|\hat v_{S_*^c}\|_1 + \lambda_n \|\hat v_{S_*}\|_1 \\
\leq& \lambda_n \left(\left(3 + \frac{1}{c_1}\right) + \left(2 + \frac{1}{c_1}\right) \alpha \|R_{S_* S_*} |\beta^*_{S_*}|\|_\infty + \left(2 + \frac{1}{c_1}\right) \frac{\alpha D}{2} \|\hat v_{S_*}\|_1 \right) \|\hat v_{S_*}\|_1. \qquad (4.8)
\end{aligned}$$

To characterize $\|\hat v_{S_*}\|_1$ in parentheses, we can see from (4.5)

$$\begin{aligned}
&\lambda_n (\|\hat v_{S_*}\|_1 + \|\hat v_{S_*^c}\|_1) \\
\leq& \gamma_n M^* + \mathcal{E}(f_{\beta^*}) + \lambda_n \left(2\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2} |\beta^*_{S_*}|^\top R_{S_* S_*} |\beta^*_{S_*}| \right) \\
\leq& \left(1 + \frac{1}{c_1}\right) \gamma_n M^* + \lambda_n \left(2\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2} |\beta^*_{S_*}|^\top R_{S_* S_*} |\beta^*_{S_*}| \right). \qquad (4.9)
\end{aligned}$$

We further devide into 2 cases.

(ii-a) If $\lambda_n \left(2\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2} |\beta^*_{S_*}|^\top R_{S_* S_*} |\beta^*_{S_*}| \right) \leq 2\gamma_n M^*$, then (4.5) reduces to

$$\begin{aligned}
\mathcal{E}(f_{\hat\beta}) + \lambda_n \|\hat v\|_1 &\leq 3\gamma_n M^* + \mathcal{E}(f_{\beta^*}) \\
&\leq (3c_1 + 1)\, \mathcal{E}(f_{\beta^*}) + \frac{3 c_2 \lambda_n^2 |S_*|}{\kappa \phi^*_{\mathrm{GRE}}}.
\end{aligned}$$

(ii-b) If $\lambda_n \left(2\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2}|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}|\right) \geq 2\gamma_n M^*$, then the inequality (4.9) reduces to

$$\begin{aligned}\lambda_n(\|\hat{v}_{S_*}\|_1 + \|\hat{v}_{S_*^c}\|_1) &\leq \frac{1}{2}\left(3+\frac{1}{c_1}\right)\lambda_n\left(2\|\beta^*_{S_*}\|_1 + \frac{\alpha}{2}|\beta^*_{S_*}|^\top R_{S_*S_*}|\beta^*_{S_*}|\right)\\ &\leq \frac{1}{2}\left(3+\frac{1}{c_1}\right)\lambda_n\left(2+\frac{\alpha}{2}\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\right)\|\beta^*_{S_*}\|_1,\end{aligned}$$

which indicates

$$\|\hat{v}_{S_*}\|_1 \leq \frac{1}{2}\left(3+\frac{1}{c_1}\right)\left(2+\frac{\alpha}{2}\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\right)\|\beta^*_{S_*}\|_1. \tag{4.10}$$

Hence, incorporating (4.8) and (4.10) yields,

$$\begin{aligned}&\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\|\hat{v}\|_1\\ \leq&\lambda_n\left(\left(3+\frac{1}{c_1}\right)+\left(2+\frac{1}{c_1}\right)\alpha\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\right.\\ &\left.+\frac{1}{2}\left(2+\frac{1}{c_1}\right)\left(3+\frac{1}{c_1}\right)\frac{\alpha D}{2}\left(2+\frac{\alpha}{2}\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\right)\|\beta^*_{S_*}\|_1\right)\|\hat{v}_{S_*}\|_1\\ \leq&\lambda_n\left(\left(3+\frac{1}{c_1}\right)+\frac{1}{2}\left(2+\frac{1}{c_1}\right)\left(5+\frac{1}{c_1}\right)\alpha D\|\beta^*\|_1\right.\\ &\left.+\frac{1}{8}\left(2+\frac{1}{c_1}\right)\left(3+\frac{1}{c_1}\right)(\alpha D\|\beta^*\|_1)^2\right)\|\hat{v}_{S_*}\|_1.\end{aligned}$$

If we take α' such that $\alpha \leq \frac{\alpha'}{D\|\beta^*\|_1}$ and define

$$c_3 := \left(3+\frac{1}{c_1}\right)+\frac{1}{2}\left(2+\frac{1}{c_1}\right)\left(5+\frac{1}{c_1}\right)\alpha' + \frac{1}{8}\left(2+\frac{1}{c_1}\right)\left(3+\frac{1}{c_1}\right)\alpha'^2,$$

we have

$$\mathcal{E}(f_{\hat{\beta}}) + \lambda_n\|\hat{v}\|_1 \leq c_3\lambda_n\|\hat{v}_{S_*}\|_1. \tag{4.11}$$

On the other hand, we can restrict the feasible region for v by (4.7) and (4.10) as

$$\begin{aligned}
&\|\hat{v}_{S_*^c}\|_1 + \alpha|\hat{v}_{S_*^c}|^\top R_{S_*^c S_*}|\hat{v}_{S_*} + \beta^*_{S_*}| + \frac{\alpha}{2}|\hat{v}_{S_*^c}|^\top R_{S_*^c S_*^c}|\hat{v}_{S_*^c}| \\
\leq &\left(2 + \frac{1}{c_1}\right)\Bigg(1 + \alpha\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty \\
&+ \frac{\alpha D}{2}\frac{1}{2}\left(3 + \frac{1}{c_1}\right)\left(2 + \frac{\alpha}{2}\|R_{S_*S_*}|\beta^*_{S_*}|\|_\infty\right)\|\beta^*_{S_*}\|_1\Bigg)\|\hat{v}_{S_*}\|_1 \\
\leq &\left(2 + \frac{1}{c_1}\right)\left(1 + \frac{1}{2}\left(5 + \frac{1}{c_1}\right)\alpha D\|\beta^*_{S_*}\|_1 + \frac{1}{8}\left(3 + \frac{1}{c_1}\right)\left(\alpha D\|\beta^*_{S_*}\|_1\right)^2\right)\|\hat{v}_{S_*}\|_1 \\
\leq &\left(2 + \frac{1}{c_1}\right)\left(1 + \frac{1}{2}\left(5 + \frac{1}{c_1}\right)\alpha' + \frac{1}{8}\left(3 + \frac{1}{c_1}\right)\alpha'^2\right)\|\hat{v}_{S_*}\|_1.
\end{aligned}$$

If we take c and α' satisfying

$$\left(2 + \frac{1}{c_1}\right)\left(1 + \frac{1}{2}\left(5 + \frac{1}{c_1}\right)\alpha' + \frac{1}{8}\left(3 + \frac{1}{c_1}\right)\alpha'^2\right) \leq 3,$$

then we have $\hat{v} \in \mathcal{B}(S_*, 3, 1)$ where

$$\begin{aligned}
&\mathcal{B}(S_*, 3, 1) \\
=&\left\{v : \|v_{S_*^c}\|_1 + \alpha|v_{S_*^c}|^\top R_{S_*^c S_*}|v_{S_*} + \beta^*_{S_*}| + \frac{\alpha}{2}|v_{S_*^c}|^\top R_{S_*^c S_*^c}|v_{S_*^c}| \leq 3\|v_{S_*}\|_1\right\}.
\end{aligned}$$

Hence, we have only to impose the restricted eigenvalue condition for the set $\mathcal{B}(S_*, 3, 1)$.

According to the restricted eigenvalue condition, we have

$$\|v\|_2^2 \leq \frac{\|f_v\|^2}{\phi^*_{\mathrm{GRE}}}, \ \forall v \in \mathcal{B}(S_*, 3, 1).$$

Incorporating $\|\hat{v}_{S_*}\|_1 \leq \sqrt{|S_*|}\|\hat{v}_{S_*}\|_2 \leq \sqrt{|S_*|}\|\hat{v}\|_2$, we have

$$\|\hat{v}_{S_*}\|_1 \leq \frac{\sqrt{|S_*|}\|f_{\hat{v}}\|}{\sqrt{\phi^*_{\mathrm{GRE}}}}.$$

Hence, from (4.11), we have

$$\begin{aligned}
\mathcal{E}(f_{\hat\beta}) + \lambda_n\|\hat v\|_1 &\le c_3\lambda_n \frac{\sqrt{|S_*|}\|f_{\hat\beta} - f_{\beta^*}\|}{\sqrt{\phi^*_{\text{GRE}}}} \\
&\le \frac{4c_3\lambda_n^2|S_*|}{\kappa\phi^*_{\text{GRE}}} + \frac{c_3\kappa}{16}\|f_{\hat\beta} - f_{\beta^*}\|^2 \\
&\le \frac{4c_3\lambda_n^2|S_*|}{\kappa\phi^*_{\text{GRE}}} + \frac{c_3\kappa}{16}\Big(\|f_{\hat\beta} - f^0\| + \|f_{\beta^*} - f^0\|\Big)^2 \\
&\le \frac{4c_3\lambda_n^2|S_*|}{\kappa\phi^*_{\text{GRE}}} + \frac{c_3\kappa}{8}\|f_{\hat\beta} - f^0\|^2 + \frac{c_3\kappa}{8}\|f_{\beta^*} - f^0\|^2 \\
&\le \frac{4c_3\lambda_n^2|S_*|}{\kappa\phi^*_{\text{GRE}}} + \frac{c_3}{8}\mathcal{E}(f_{\hat\beta}) + \frac{c_3}{8}\mathcal{E}(f_{\beta^*}),
\end{aligned}$$

where we use the restricted eigenvalue condition in the first line, $uv \le 4u^2 + v^2/16$ in the second line, the triangular inequality in the third line, $(u+v)^2 \le 2(u^2+v^2)$ in the fourth line, and the margin condition with $f_{\hat\beta} \in \boldsymbol{F}_\eta$ and $f_{\beta^*} \in \boldsymbol{F}_\eta$ in the last line. This implies

$$\left(1 - \frac{c_3}{8}\right)\mathcal{E}(f_{\hat\beta}) + \lambda_n\|\hat v\|_1 \le \frac{c_3}{8}\mathcal{E}(f_{\beta^*}) + \frac{4c_3\lambda_n^2|S_*|}{\kappa\phi^*_{\text{GRE}}},$$

hence we obtain

$$\mathcal{E}(f_{\hat\beta}) + \lambda_n\|\hat v\|_1 \le \frac{c_3}{8 - c_3}\mathcal{E}(f_{\beta^*}) + \frac{32c_3\lambda_n^2|S_*|}{(8 - c_3)\kappa\phi^*_{\text{GRE}}}.$$

where we assume $c_3 < 8$, which is satisfied by taking appropreate c_1 and α'.

Incorporating the case (i), (ii-a) and (ii-b), we have shown that

$$\mathcal{E}(f_{\hat\beta}) + \lambda_n\|\hat v\|_1 \le c_4\gamma_n M^*, \tag{4.12}$$

where

$$c_4 := \max\left\{3 + \frac{1}{c_1}, \frac{c_3}{c_1(8 - c_3)}, \frac{32c_3}{c_2(8 - c_3)}\right\}.$$

On the other hand, we have

$$\begin{aligned}
\mathcal{E}(f_{\hat\beta}) &\ge \kappa\|f_{\hat\beta} - f^0\|^2 \\
&\ge \frac{1}{2}\kappa\|f_{\hat\beta} - f_{\beta^*}\|^2 - \kappa\|f_{\beta^*} - f^0\|^2 \\
&\ge \frac{1}{2}\kappa\phi^*_{\text{GRE}}\|\hat\beta - \beta^*\|_2^2 - \mathcal{E}(f_{\beta^*}),
\end{aligned}$$

where we used $(u+v)^2 \leq 2(u^2+v^2)$ in the second line, and the generalized restricted and margin conditions in the last line. Therefore, we have

$$\frac{1}{2}\kappa\phi^*_{\mathrm{GRE}}\|\hat{\beta}-\beta^*\|_2^2 \leq (c_1c_4+1)\mathcal{E}(f_{\beta^*}) + \frac{c_2c_4\lambda_n^2|S_*|}{\kappa\phi^*_{\mathrm{GRE}}},$$

that is,

$$\|\hat{\beta}-\beta^*\|_2^2 \leq \frac{2(c_1c_4+1)}{\kappa\phi^*_{\mathrm{GRE}}}\mathcal{E}(f_{\beta^*}) + \frac{2c_2c_4\lambda_n^2|S_*|}{\kappa^2{\phi^*_{\mathrm{GRE}}}^2},$$

This concludes the assertion.

Now, we show $\|\hat{\beta}-\beta^*\|_1 \leq M^*$. Define

$$\tilde{\beta} := \beta^* + t(\hat{\beta}-\beta^*),\ t := \frac{M^*}{M^* + \|\hat{\beta}-\beta^*\|_1}.$$

Then it holds $\|\tilde{\beta}-\beta^*\|_1 \leq M^*$. Because $\mathcal{L}(\tilde{\beta}) \leq \mathcal{L}(\beta^*)$ by the assumption, all the above inequalities hold substuting $\hat{\beta}$ into $\tilde{\beta}$. In particular, (4.12) implies

$$\|\tilde{\beta}-\beta^*\|_1 \leq c_4\frac{\gamma_n}{\lambda_n}M^* \leq \frac{M^*}{2},$$

for which we assume $c_4\frac{\gamma_n}{\lambda_n} \leq \frac{1}{2}$. Because $\|\tilde{\beta}-\beta^*\|_1 = t\|\hat{\beta}-\beta^*\|_1$, we have

$$\frac{M^*}{M^* + \|\hat{\beta}-\beta^*\|_1}\|\hat{\beta}-\beta^*\|_1 \leq \frac{M^*}{2},$$

hence we obtain $\|\hat{\beta}-\beta^*\|_1 \leq M^*$.

Finally, we take $c_1 = 3,\ c_2 = 9,$ and $\alpha' = 1/12$, which satisfy all the aforementioned assumptions.

□

4.4.2 Proof of Corrolary 15

Proof. We can see that the logistic loss is Lipschitz with a Lipschitz constant 1, because it holds that

$$\left|\frac{\partial}{\partial f}l(f,y)\right| = \left|-y + \frac{\exp(f)}{1+\exp(f)}\right| \leq 1.$$

Next, we show that the logistic regression satisfies the quadratic margin condition. For $f \in \boldsymbol{F}_\eta$,

$$
\begin{aligned}
\frac{\partial^2}{\partial f^2} l(f, \cdot) &= \frac{\exp(f)}{(1+\exp(f))^2} \\
&\geq \frac{\exp(|f^0| + \eta)}{(1+\exp(|f^0|+\eta))^2} \\
&\geq \frac{1}{(1+\exp(|f^0|+\eta))^2} \\
&= \frac{1}{(1+\exp(\eta)\max\left\{\frac{\pi}{1-\pi}, \frac{1-\pi}{\pi}\right\})^2} \\
&\geq \frac{1}{(1+\exp(\eta)\frac{1-\varepsilon_0}{\varepsilon_0})^2} \\
&\geq \frac{1}{(\exp(\eta)/\varepsilon_0 + 1)^2}.
\end{aligned}
$$

Hence, the quadratic margin condition holds with $\kappa := (\exp(\eta)/\varepsilon_0 + 1)^{-2}$.

On the other hand, we have

$$
\|f_\beta - f^0\|_\infty \leq \|f_\beta - f_{\beta^*}\|_\infty + \|f_{\beta^*} - f^0\|_\infty.
$$

For $\|\beta - \beta^*\|_1 \leq M^*$, it holds that

$$
\begin{aligned}
\|f_\beta - f_{\beta^*}\|_\infty &= \|(\beta - \beta^*)X\|_\infty \\
&\leq \|\beta - \beta^*\|_1 \|X\|_{\max} \\
&\leq M^* \\
&= \frac{1}{\gamma_n}\left(3\mathcal{E}(f_{\beta^*}) + \frac{9(\exp(\eta)/\varepsilon_0 + 1)^2 \lambda_n^2 |S_{\beta^*}|}{\phi^*_{\mathrm{GRE}}}\right).
\end{aligned}
$$

Since we assume

$$
\|f_{\beta^*} - f^0\|_\infty \leq \frac{\eta}{3}, \ \mathcal{E}(f_{\beta^*})/\gamma_n \leq \frac{\eta}{9}, \text{ and } \frac{\mathrm{T}^2(\exp(\eta)/\varepsilon_0 + 1)^2 \gamma_{\mathrm{n}} |\mathrm{S}_{\beta^*}|}{\phi^*_{\mathrm{GRE}}} \leq \frac{\eta}{27},
$$

with $\lambda_n \leq T\gamma_n$, we obtain

$$
\|f_\beta - f^0\|_\infty \leq \eta.
$$

From the above, all of the assumptions in Theorem 12 have been verified. □

Chapter 5

Numerical Experiments

5.1 Synthetic Data Experiments for Linear Models

First, we validated the effectiveness of the IILasso for linear models. We considered the case in which the true active variables are uncorrelated and many inactive variables are strongly correlated with the active variable. If all of the active and inactive variables are uncorrelated, it is easy to estimate which is active or inactive. On the other hand, if the inactive variables are strongly correlated with the active variables, it is hard to distinguish which one is active. We simulate such a situation.

We generated a design matrix $X \in \mathbb{R}^{n \times p}$ from the Gaussian distribution of $\mathcal{N}(0, \Sigma)$ where $\Sigma = \mathrm{Diag}(\Sigma^{(1)}, \cdots, \Sigma^{(b)})$ was a block diagonal matrix whose element $\Sigma^{(l)} \in \mathbb{R}^{q \times q}$ was $\Sigma^{(l)}_{jk} = 0.95$ for $j \neq k$ and $\Sigma^{(l)}_{jk} = 1$ for $j = k$. We set $n = 50$, $p = 100$, $b = 10$ and $q = 10$. Thus, there were 10 groups containing 10 strongly correlated variables. Next, we generated a response y by the true active variables $X_1, X_{11}, X_{21}, \cdots, X_{91}$, such that $y = 10X_1 - 9X_{11} + 8X_{21} - 7X_{31} + \cdots + 2X_{81} - X_{91} + \varepsilon$, with a standard Gaussian noise ε. Each group included one active variable. We generated three datasets for training, validation, and test as above procedures independently.

Then, we compared the performance of the Lasso, SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010), EGLasso (Kong, Fujimaki, Liu, Nie, and Ding, 2014), and IILasso. Evaluation criteria are prediction error (mean squared error), estimation error (ℓ_2 norm between the true and estimated coefficients) and model size (the number of non-zero coefficients). The SCAD and MCP are representative methods of folded concave penalty, so their objective functions are non-convex, which are the same as our method. They have a tuning parameter γ; we set $\gamma = 2.5, 3.7, 10, 20, 100, 1000$ for the SCAD and $\gamma = 1.5, 3, 10, 20, 100, 1000$ for the MCP. The EGLasso has a parameter λ_2; we set

Table 5.1: Results of synthetic data for linear models

	prediction error	estimation error	model size
Lasso (`ncvreg`)	2.67(0.05)	4.44(0.06)	34.1(0.46)
SCAD (`ncvreg`)	1.52(0.02)	1.79(0.04)	14.6(0.23)
MCP (`ncvreg`)	1.53(0.02)	1.79(0.04)	14.6(0.24)
MCP (`sparsenet`)	2.41(0.11)	3.15(0.13)	**13.4(0.28)**
EGLasso	2.60(0.04)	4.36(0.05)	33.3(0.32)
IILasso (ours)	**1.45(0.02)**	**1.40(0.04)**	**13.5(0.23)**

$\lambda_2/\lambda_1 = 0.01, 0.1, 1, 10, 100, 1000$. For the EGLasso, we used the true group information beforehand. We used `R` packages `ncvreg` (Breheny and Huang, 2011) for the Lasso, SCAD and MCP, and `sparsenet` (Mazumder, Friedman, and Hastie, 2011) for MCP. One can solve MCP using either `ncvreg` or `sparsenet`; they differ in their optimization algorithms and ways of initialization. For the IILasso, we defined $R_{jk} = |r_{jk}|/(1 - |r_{jk}|)$ for $j \neq k$ and $R_{jk} = 0$ for $j = k$. Hence, R_{SS} takes small values if active variables are independent, and R_{SS^c} and $R_{S^cS^c}$ take large values if inactive variables are strongly correlated with other variables, which is favorable from the theoretical results. We set $\alpha = 0.01, 0.1, 1, 10, 100, 1000$. We tuned the above parameters using validation data and calculated errors using test data. We iterated this procedure 500 times and evaluated the averages and standard errors.

Table 5.1 shows the performances with their standard error in parentheses. The IILasso achieved the best prediction and estimation among all of them. This was because our penalty term excluded the correlations and avoided overfitting. Moreover, the model size of the IILasso was much less than those of the Lasso and EGLasso and comparable to MCP. As a whole, the IILasso could estimate the most accurate model with a few variables.

In our experiments, we employed cross validation for tuning the regularization parameters. We could use our theoretical results such as (3.12) to find appropreate parameters, although it requires some knowledges such as $\|\beta^*\|_1$ and D in advance.

5.2 Synthetic Data Experiments for Logistic Models

Next, we validated the performance of the IILasso for logistic regression models. We generated a response y by Bernoulli$(\pi(x))$ where log-odds of $\pi(x)$ is defined by $10X_1 - 9X_{11} + 8X_{21} - 7X_{31} + \cdots + 2X_{81} - X_{91}$. For the rest setting, we used the same parameters and procedures as Section 5.1, except that $n = 100$ instead of $n = 50$. This is just because it is difficult for all method to estimate

Table 5.2: Results of synthetic data for logistic models

	negative log-likelihood	misclassification error	estimation error	model size
Lasso (`ncvreg`)	0.484(0.006)	0.107(0.002)	16.8(0.06)	22.9(0.14)
SCAD (`ncvreg`)	0.486(0.007)	0.105(0.002)	16.00(0.10)	18.8(0.25)
MCP (`ncvreg`)	0.489(0.007)	0.106(0.002)	16.7(0.10)	18.1(0.26)
EGLasso	**0.477(0.006)**	0.104(0.002)	16.7(0.06)	25.7(0.18)
IILasso (ours)	**0.471(0.009)**	**0.096(0.002)**	**14.4(0.14)**	**12.3(0.20)**

Table 5.3: Abstract of microarray data

data	# samples	# dimensions	# positive labels
alon	62	2000	22
chiaretti	111	12625	10
gordon	181	12533	150
gravier	168	2905	111
pomeroy	60	7128	21
shipp	77	7129	58
singh	102	12600	50
subramanian	50	10100	33
tian	173	12625	137
west	49	7129	25

logistic regression models with small samples. The Lasso, SCAD, and MCP for logistic regression is supported by `ncvreg` (`sparsenet` does not support logistic regression). The EGLasso and the IILasso for logistic regression can be solved by CDA. We evaluated the negative log-likelihood, misclassification error (the rate of misclassification), estimation error, and model size.

Table 5.2 shows the performances with their standard error in parentheses. The IILasso outperformed other methods significantly in terms of misclassification error and estimation error. Besides, the IILasso presented a much smaller model size than other methods. Therefore, the IILasso could efficiently estimate the accurate logistic models with a few variables.

5.3 Real Data Experiments: Gene Expression Data

We applied our method to various gene expression data to validate its effectiveness for real applications. We used the following 10 datasets: ‘alon’ (Alon, Barkai, Notterman, Gish, Ybarra, Mack, and Levine, 1999) (colon cancer), ‘chiaretti’ (Chiaretti, Li, Gentleman, Vitale, Vignetti, Mandelli, Ritz, and Foa, 2004) (leukemia), ‘gordon’ (Gordon, Jensen, Hsiao, Gullans, Blumenstock, Ramaswamy, Richards, Sugarbaker, and Bueno, 2002) (lung cancer), ‘gravier’ (Gravier, Pierron, Vincent-Salomon, Gruel, Raynal, Savignoni,

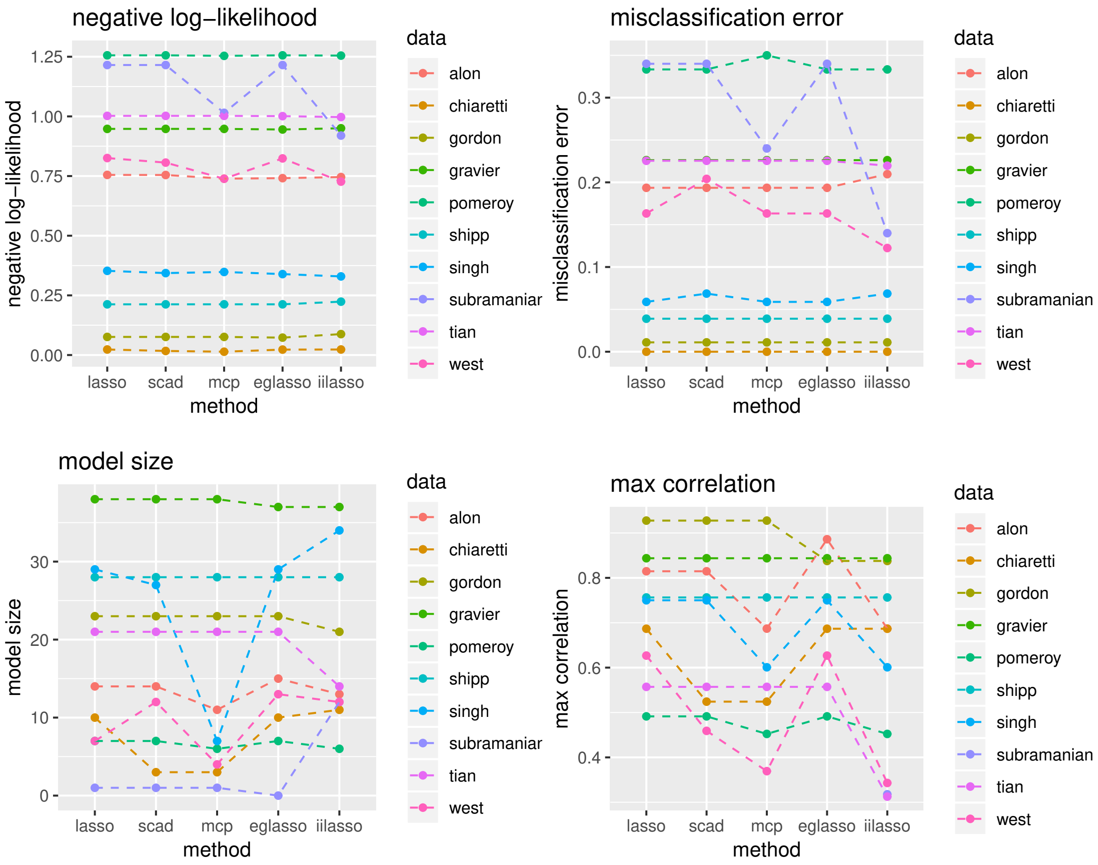

FIGURE 5.1: Results of 10 microarray datasets

De Rycke, Pierga, Lucchesi, Reyal, et al., 2010) (breast cancer), 'pomeroy' (Pomeroy, Tamayo, Gaasenbeek, Sturla, Angelo, McLaughlin, Kim, Goumnerova, Black, Lau, et al., 2002) (central nervous system disorders), 'shipp' (Shipp, Ross, Tamayo, Weng, Kutok, Aguiar, Gaasenbeek, Angelo, Reich, Pinkus, et al., 2002) (lymphoma), 'singh' (Singh, Febbo, Ross, Jackson, Manola, Ladd, Tamayo, Renshaw, D'Amico, Richie, et al., 2002) (prostate cancer), 'subramanian' (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, et al., 2005) (miscellaneous), 'tian' (Tian, Zhan, Walker, Rasmussen, Ma, Barlogie, and Shaughnessy Jr, 2003) (myeloma), 'west' (West, Blanchette, Dressman, Huang, Ishida, Spang, Zuzan, Olson, Marks, and Nevins, 2001) (breast cancer). All of these data are provided by R package `datamicroarray`. The abstract of these datasets is described in Table 5.3. All datasets are small-sample high-dimensional DNA microarray data. Since the response is binary, logistic regression was applied. We used the same settings on regularization parameters, as described in Chapter 5.1. We evaluated the negative log-likelihood, misclassification error, model size, and max correlation among active variables, using twenty-fold cross validation.

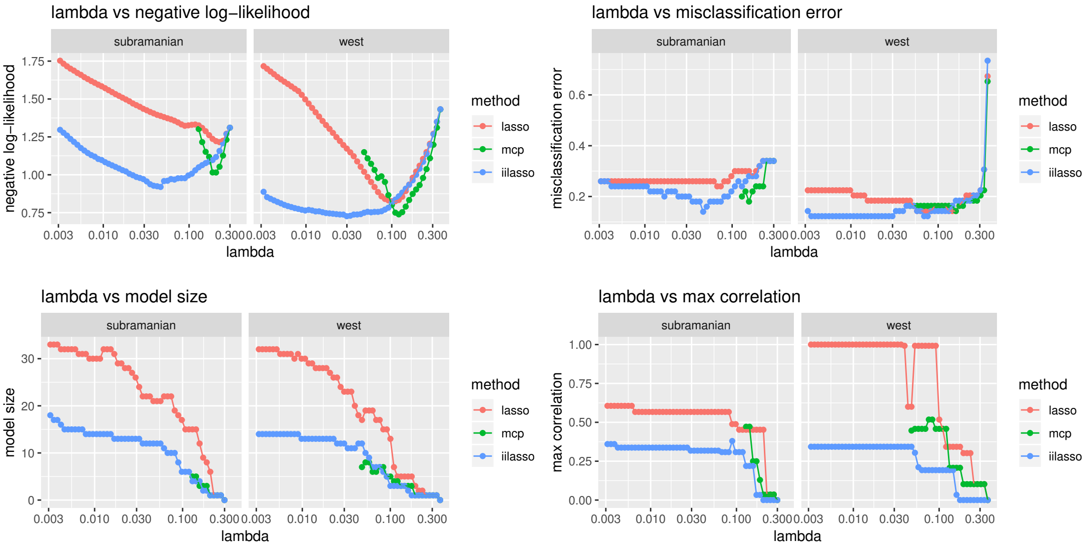

FIGURE 5.2: Results of 'subramanian' and 'west' along with the sequences of λ

The results are given in Figure 5.1. In terms of negative log-likelihood, 8 out of 10 datasets gave similar performances among all methods, and the rest two datasets 'subramanian' and 'west' showed the smallest negative log-likelihood by the IILasso. The MCP was comparable to the IILasso but fell behind the IILasso. We can see similar tendencies of misclassification errors. The IILasso won in 8 out of 10 cases, including 5 ties. Although numbers of selected variables of the IILasso were inferior to the MCP (the IILasso won in 5, including 3 ties; the MCP won in 6, including 2 ties), max correlations among active variables were superior to others (the IILasso won in 8, including 5 ties; the MCP won in 5, including 5 ties). As a whole, the IILasso could construct accurate models with small correlations.

We further investigated the influences of regularization parameters for the datasets 'subramanian' and 'west' because they showed a clear difference among methods. Figure 5.2 shows the results of the Lasso, MCP, and IILasso. Compared with the Lasso, the IILasso mitigated overfitting even when λ was small, and hence achieved low negative log-likelihood and misclassification error. Moreover, the IILasso suppressed its model size and correlations among active variables. The MCP rapidly decreased negative log-likelihood as λ decreased, and overfitting occurred early. The sequences of the MCP were broken because its algorithm iterations reached a maximum number defined in `ncvreg`.

Chapter 6

Conclusion

In this thesis, we proposed a new regularization method, "IILasso". The IILasso reduces correlations among the active variables; hence it is easy to decompose and interpret the model. We showed that the sign recovery condition of the IILasso is milder than that of the Lasso for correlated design as long as the true important variables are uncorrelated with each other. The convergence rate of the IILasso also has a better performance compared to that of the Lasso. Moreover, we extend the IILasso to the GLMs, and its convergence rate is also analyzed. Finally, we verified the effectiveness of the IILasso by synthetic and real data analyses using ten gene expression data, and we showed that the IILasso was superior in many cases on high-dimensional data.

Bibliography

Alon, Uri, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". In: *Proceedings of the National Academy of Sciences* 96.12, pp. 6745–6750.

Beck, Amir and Marc Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences* 2.1, pp. 183–202.

Bertsimas, Dimitris, Angela King, and Rahul Mazumder (2016). "Best subset selection via a modern optimization lens". In: *Annals of Statistics*, pp. 813–852.

Bickel, Peter J, Ya'acov Ritov, Alexandre B Tsybakov, et al. (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *Annals of Statistics* 37.4, pp. 1705–1732.

Bogdan, Małgorzata, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès (2015). "SLOPE: adaptive variable selection via convex optimization". In: *Annals of Applied Statistics* 9.3, p. 1103.

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends in Machine learning* 3.1, pp. 1–122.

Breheny, Patrick and Jian Huang (2011). "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection". In: *Annals of Applied Statistics* 5.1, p. 232.

Bühlmann, Peter and Sara Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Candes, Emmanuel J, Michael B Wakin, and Stephen P Boyd (2008). "Enhancing sparsity by reweighted ℓ_1 minimization". In: *Journal of Fourier Analysis and Applications* 14.5-6, pp. 877–905.

Chen, Si-Bao, Chris Ding, Bin Luo, and Ying Xie (2013). "Uncorrelated lasso". In: *Twenty-seventh AAAI conference on artificial intelligence*, pp. 166–172.

Chiaretti, Sabina, Xiaochun Li, Robert Gentleman, Antonella Vitale, Marco Vignetti, Franco Mandelli, Jerome Ritz, and Robin Foa (2004). "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival". In: *Blood* 103.7, pp. 2771–2778.

Daubechies, I., M. Defrise, and C. De Mol (2004). "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint". In: *Communications on Pure and Applied Mathematics* 57.11, pp. 1413–1457.

Ding, Chris and Hanchuan Peng (2005). "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of Bioinformatics and Computational Biology* 3.02, pp. 185–205.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.

Draper, NR and H Smith (1966). "Applied regression analysis, New York, 1966". In: *EM Pugh and GH Winslow, The Analysis of Physical Measurements, Addison-Wesley, New York*, p. 26.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. (2004). "Least angle regression". In: *Annals of Statistics* 32.2, pp. 407–499.

Efroymson, MA and TL Ray (1966). "A branch-bound algorithm for plant location". In: *Operations Research* 14.3, pp. 361–368.

Fan, Jianqing and Runze Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American Statistical Association* 96.456, pp. 1348–1360.

Figueiredo, Mario and Robert Nowak (2016). "Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects". In: *Artificial Intelligence and Statistics*, pp. 930–938.

Friedman, Jerome, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. (2007). "Pathwise coordinate optimization". In: *Annals of Applied Statistics* 1.2, pp. 302–332.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, p. 1.

Fuchs, Jean-Jacques (2005). "Recovery of exact sparse representations in the presence of bounded noise". In: *IEEE Transactions on Information Theory* 51.10, pp. 3601–3608.

Gasso, Gilles, Alain Rakotomamonjy, and Stéphane Canu (2009). "Recovering sparse signals with a certain family of nonconvex penalties and DC programming". In: *IEEE Transactions on Signal Processing* 57.12, pp. 4686–4698.

Gordon, Gavin J, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, Sridhar Ramaswamy, William G Richards, David J Sugarbaker, and Raphael Bueno (2002). "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma". In: *Cancer research* 62.17, pp. 4963–4967.

Gorodnitsky, Irina F and Bhaskar D Rao (1993). "A recursive weighted minimum norm algorithm: analysis and applications". In: *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. IEEE, pp. 456–459.

Gorodnitsky, Irina F and Bhaskar D Rao (1997). "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm". In: *IEEE Transactions on Signal Processing* 45.3, pp. 600–616.

Grave, Edouard, Guillaume R Obozinski, and Francis R Bach (2011). "Trace lasso: a trace norm regularization for correlated designs". In: *Advances in Neural Information Processing Systems*, pp. 2187–2195.

Gravier, Eléonore, Gaëlle Pierron, Anne Vincent-Salomon, Nadège Gruel, Virginie Raynal, Alexia Savignoni, Yann De Rycke, Jean-Yves Pierga, Carlo Lucchesi, Fabien Reyal, et al. (2010). "A prognostic DNA signature for T1T2 node-negative breast cancer patients". In: *Genes, chromosomes and cancer* 49.12, pp. 1125–1134.

Hara, Satoshi and Takanori Maehara (2017). "Enumerate lasso solutions for feature selection". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Hastie, Trevor, Robert Tibshirani, and Ryan J Tibshirani (2017). "Extended comparisons of best subset selection, forward stepwise selection, and the lasso". In: *arXiv preprint arXiv:1707.08692*.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Kong, Deguang, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding (2014). "Exclusive Feature Learning on Arbitrary Structures via $\ell_{1,2}$-norm". In: *Advances in Neural Information Processing Systems*, pp. 1655–1663.

Lipton, Zachary C (2018). "The mythos of model interpretability". In: *Queue* 16.3, pp. 31–57.

Loh, Po-Ling and Martin J Wainwright (2015). "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima". In: *The Journal of Machine Learning Research* 16.1, pp. 559–616.

Lorbert, Alexander, David Eis, Victoria Kostina, David Blei, and Peter Ramadge (2010). "Exploiting covariate similarity in sparse regression via the pairwise elastic net". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 477–484.

Mazumder, Rahul, Jerome H Friedman, and Trevor Hastie (2011). "Sparsenet: Coordinate descent with nonconvex penalties". In: *Journal of the American Statistical Association* 106.495, pp. 1125–1138.

Meinshausen, Nicolai (2007). "Relaxed lasso". In: *Computational Statistics & Data Analysis* 52.1, pp. 374–393.

Meinshausen, Nicolai, Peter Bühlmann, et al. (2006). "High-dimensional graphs and variable selection with the lasso". In: *Annals of Statistics* 34.3, pp. 1436–1462.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267, pp. 1–38.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Molnar, Christoph et al. (2018). "Interpretable machine learning: A guide for making black box models explainable". In: *Christoph Molnar, Leanpub*.

Natarajan, Balas Kausik (1995). "Sparse approximate solutions to linear systems". In: *SIAM journal on computing* 24.2, pp. 227–234.

Nesterov, Yu (2013). "Gradient methods for minimizing composite functions". In: *Mathematical Programming* 140.1, pp. 125–161.

Pomeroy, Scott L, Pablo Tamayo, Michelle Gaasenbeek, Lisa M Sturla, Michael Angelo, Margaret E McLaughlin, John YH Kim, Liliana C Goumnerova, Peter M Black, Ching Lau, et al. (2002). "Prediction of central nervous system embryonal tumour outcome based on gene expression". In: *Nature* 415.6870, pp. 436–442.

Rao, Bhaskar D and Kenneth Kreutz-Delgado (1999). "An affine scaling methodology for best basis selection". In: *IEEE Transactions on Signal Processing* 47.1, pp. 187–200.

Raskutti, Garvesh, Martin J Wainwright, and Bin Yu (2010). "Restricted eigenvalue properties for correlated Gaussian designs". In: *Journal of Machine Learning Research* 11.Aug, pp. 2241–2259.

Raskutti, Garvesh, Martin J Wainwright, and Bin Yu (2011). "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls". In: *IEEE Transactions on Information Theory* 57.10, pp. 6976–6994.

Rigollet, Phillippe and Jan-Christian Hütter (2015). "High dimensional statistics". In: *Lecture notes for course 18S997*.

She, Yiyuan et al. (2009). "Thresholding-based iterative selection procedures for model selection and shrinkage". In: *Electronic Journal of Statistics* 3, pp. 384–415.

Shipp, Margaret A, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, et al. (2002). "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning". In: *Nature medicine* 8.1, pp. 68–74.

Singh, Dinesh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. (2002). "Gene expression correlates of clinical prostate cancer behavior". In: *Cancer cell* 1.2, pp. 203–209.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550.

Tian, Erming, Fenghuang Zhan, Ronald Walker, Erik Rasmussen, Yupo Ma, Bart Barlogie, and John D Shaughnessy Jr (2003). "The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma". In: *New England Journal of Medicine* 349.26, pp. 2483–2494.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tropp, Joel A (2006). "Just relax: Convex programming methods for identifying sparse signals in noise". In: *IEEE Transactions on Information Theory* 52.3, pp. 1030–1051.

Tseng, Paul (2001). "Convergence of a block coordinate descent method for nondifferentiable minimization". In: *Journal of Optimization Theory and Applications* 109.3, pp. 475–494.

Wainwright, Martin J (2009). "Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE Transactions on Information Theory* 55.5, pp. 2183–2202.

West, Mike, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins (2001). "Predicting the clinical status of human breast cancer by using gene expression profiles". In: *Proceedings of the National Academy of Sciences* 98.20, pp. 11462–11467.

Zeng, Xiangrong and Mário AT Figueiredo (2014). "The Ordered Weighted ℓ_1 Norm: Atomic Formulation, Projections, and Algorithms". In: *arXiv preprint arXiv:1409.4271*.

Zhang, Cun-Hui, Tong Zhang, et al. (2012). “A general theory of concave regularization for high-dimensional sparse estimation problems”. In: *Statistical Science* 27.4, pp. 576–593.

Zhang, Cun-Hui et al. (2010). “Nearly unbiased variable selection under minimax concave penalty”. In: *Annals of Statistics* 38.2, pp. 894–942.

Zhao, Peng and Bin Yu (2006). “On model selection consistency of Lasso”. In: *Journal of Machine Learning Research* 7.Nov, pp. 2541–2563.

Zou, Hui (2006). “The adaptive lasso and its oracle properties”. In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429.

Zou, Hui and Trevor Hastie (2005). “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (statistical methodology)* 67.2, pp. 301–320.

Zou, Hui and Runze Li (2008). “One-step sparse estimates in nonconcave penalized likelihood models”. In: *Annals of Statistics* 36.4, pp. 1509–1533.