Penalized and Constrained Regression

Gareth M. James, Courtney Paulson and Paat Rusmevichientong [*]

December 15, 2013

Abstract

Motivated by applications in areas as diverse as finance, image reconstruction, and curve estimation, we consider the constrained high-dimensional generalized linear model (GLM) problem, where the underlying parameters satisfy a collection of linear constraints. We develop the *Penalized and Constrained* regression method (PAC) for computing the penalized coefficient paths on high-dimensional GLM fits, subject to a set of linear constraints. PAC is an extremely general method, and we show that it encompasses many statistical approaches, such as the fused lasso, monotone curve estimation and the generalized lasso. Computing the PAC coefficient path poses some technical challenges but we develop an efficient algorithm for fitting the path over a grid of tuning parameters. Non-asymptotic error bounds are presented, which suggest that PAC should outperform unconstrained penalized GLM methods in situations where the true parameters satisfy the underlying constraints. Extensive numerical experiments show that our method performs well, both computationally and statistically. Finally, we apply PAC to a real dataset to estimate the demand curve for a particular type of auto loan as a function of interest rate, and demonstrate that it outperforms more standard approaches.

Keywords: PAC; Linear constraints; Penalized regression; Demand function; Generalized Linear Models

[*] Marshall School of Business, University of Southern California.

1. Introduction

In a generalized linear model with a canonical link function the response variable Y_i and the predictors $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip}) \in \mathbb{R}^p$ are related through an exponential family distribution defined by:

$$\Pr\left\{Y_i = y \;\middle|\; \mathbf{X}_i = \boldsymbol{x}\right\} = \exp\left(\frac{[y \cdot (\boldsymbol{x}^T\boldsymbol{\beta})] - \psi\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)}{a(\sigma)} + c(y, \sigma)\right), \tag{1}$$

where $\mu_i = \mathsf{E}[Y_i \mid \mathbf{X}_i] = \psi'\left(\sum_{j=1}^p X_{ij}\beta_j\right)$. The *link function* $g(\cdot)$ connects the response and predictors through: $g(\mu_i) = \sum_{j=1}^p X_{ij}\beta_j$. Common link functions include the identity link used for normal response data and the logistic link used for binary response data. For notational simplicity we have assumed in (1) that g is the canonical link, although all the ideas generalize naturally to other link functions.

A great deal of attention has recently been focused on the problem of fitting GLMs in situations where the number of predictors, p, is large relative to the sample size, n. In this setting there are many approaches that outperform standard maximum likelihood methods (Frank and Friedman, 1993). Early work on these “large p” problems concentrated on the linear regression model, with penalized regression methods commonly adopted. A small sampling of these approaches include the lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive lasso (Zou, 2006), CAP (Zhao *et al.*, 2009), the Dantzig selector (Candes and Tao, 2007), the relaxed lasso (Meinshausen, 2007), and VISA (Radchenko and James, 2008).

However, more recently there has been growing interest in other GLM settings. For example, Park and Hastie (2007) discuss a natural GLM extension of the lasso where, for a fixed λ, they choose $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ to minimize a penalized version of the log likelihood,

$$\arg\min_{\boldsymbol{\beta}} \left\{ -\sum_{i=1}^{n} \mathsf{LogLik}_i(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{2}$$

where $\mathsf{LogLik}_i(\boldsymbol{\beta})$ denotes the log likelihood of the i^{th} observation. Friedman *et al.* (2010) further extend this work by developing an extremely efficient coordinate descent algorithm for solving (2). Of course the ℓ_1 penalty is just one possible option. We may wish to solve the more general penalized regression problem

$$\arg\min_{\boldsymbol{\beta}} \left\{ -\sum_{i=1}^{n} \mathsf{LogLik}_i(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \rho(|\beta_j|) \right\}, \tag{3}$$

where $\rho(\cdot)$ is a pre-specified penalty function. The lasso can be seen as a special case of (3) where the response follows a Gaussian distribution and $\rho(t) = t$. However, (3) can also be used to fit other response distributions, such as Bernoulli or Poisson, and alternative penalty functions, such as the elastic net or SCAD penalties.

In this article we are interested in solving a constrained version of (3)

$$\arg\min_{\boldsymbol{\beta}} \left\{ -\sum_{i=1}^{n} \mathsf{LogLik}_i(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \rho(|\beta_j|) \right\} \quad \text{subject to} \quad \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \tag{4}$$

where $\mathbf{C} \in \mathbb{R}^{m \times p}$ and $\mathbf{b} \in \mathbb{R}^m$ are predefined matrices and vectors. Equation (4) places a set of m linear constraints on the coefficients, so we call this the *Penalized and Constrained* (PAC) regression method. Equations (3) and (4) appear similar to each other so why is the constrained problem interesting? We address this question in three parts.

Broad Range of Applications: First, we demonstrate in Section 2 that many commonly applied statistical problems, such as the fused lasso, generalized lasso, monotone curve estimation, and many others, can be expressed as special cases of (4). Hence, an efficient algorithm for solving the constrained problem can be used to fit a large class of statistical methodologies.

New Algorithmic Development: Second, it turns out that solving (4) is a non-trivial problem. In the special case of linear regression with an ℓ_1 penalty (4) reduces to a constrained version of the lasso

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}. \tag{5}$$

There has been some work on solving (5) using path algorithms in the spirit of LARS (Efron *et al.*, 2004; Tibshirani and Taylor, 2011; Zhou and Lange, 2013). However, path algorithms rely heavily on the piecewise linear nature of the lasso paths. This property does not hold for general likelihood or penalty functions, in which case path algorithms become far less appealing.

Alternatively one could adopt a coordinate descent method (Wu and Lange, 2008). Friedman *et al.* (2010) demonstrate that these methods have clear efficiency and stability advantages for fitting penalized GLM problems of the form in (2). Unfortunately, a naive implementation of coordinate descent for (4) generally results in a solution that fails to reach the global optimum because the solution may get stuck at the constraint boundaries. However, in Section 3, we propose a modified coordinate descent algorithm for (4) that avoids these convergence issues. In addition, our PAC algorithm can be implemented using standard lasso software packages, which greatly improves the robustness and computational speed of PAC. This is an important point because other algorithms generally require more complicated fitting methods. For example, Zhou and Wu (2013) suggest the EPSODE algorithm which is a very general approach for optimizing constrained criterion. The method is elegant and works well on the smaller scale problems that they consider. However, EPSODE must solve an ordinary differential equation for each value of the tuning parameter, which may prove prohibitive in high dimensional problems. By contrast we demonstrate that PAC can run efficiently on problems involving $p = 500$ or more predictors.

New Non-Asymptotic Error Bounds: Third, in Section 4, we extend the theoretical bounds on the errors in the unconstrained fits given by (2) to those for the PAC coefficient estimates. The unconstrained and PAC bounds have similar forms, but our bounds clearly demonstrate the potential improvements in accuracy that can be derived from adding the additional constraints.

The paper is structured as follows. In Section 2 we provide a number of motivating examples which illustrate the wide range of situations where PAC is applicable. Our modified coordinate descent algorithm is presented in Section 3. We first consider the situation where (4) involves equality constraints and then show how the algorithm can be extended to the more general inequality setting through the use of slack variables. In Section 4, we develop non-asymptotic error bounds for PAC under the assumption that the constraints hold for the true coefficients. Our error bounds clearly show that the PAC estimates will have smaller errors than unconstrained ones. There are many settings where we know that certain constraints must hold on the parameters we are estimating. For instance, in demand curve estimation, we want the curve to be non-increasing in price, and in portfolio optimization, the total weights allocated to all assets must add up to one. Our theoretical results show that, in these settings, directly incorporating these constraints will produce superior estimates. We compare PAC to a variety of unconstrained methods on a series of simulated data sets in Section 5. PAC is also implemented on a real data set consisting of loan applications for a large number of potential customers. The aim here is to estimate the demand curve as a function of interest rate. We conclude with a discussion of future extensions of this work in Section 6.

2. Motivating Examples

In this section we briefly describe a few of the situations in which PAC can be applied.

2.1 Monotone Curve Estimation

Consider the problem of fitting a smooth function, $h(x)$, to a set of observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, subject to the constraint that h must be monotone e.g. non-increasing. There are many examples of such a setting. For instance, when estimating demand as a function of price we need to ensure that the curve is non-increasing. Alternatively, when estimating a cumulative distribution function (CDF) we need to ensure that the function is non-decreasing. Similarly, when performing curve registration to align a set of curves, the estimated warping functions must be non-decreasing.

If h is modeled as a parametric function the monotonicity constraint is not hard to impose. However, one often wishes to produce a more flexible fit using a non-parametric approach. For example, we could model $h(x) = \mathbf{b}(x)^T\boldsymbol{\beta}$ where $\mathbf{b}(x)$ is some p-dimensional basis function. Then one natural approach to produce a very flexible estimate for $h(x)$ would be to choose a large value for p and then to smooth the estimate by minimizing (3) subject to $g(\mu_i) = h(x_i) = \mathbf{b}(x_i)^T\boldsymbol{\beta}$. However, minimizing (3) does not ensure a monotone curve. Hence, we need to constrain the coefficients such that

$$\mathbf{C}\boldsymbol{\beta} \leq \mathbf{0}, \tag{6}$$

where the l^{th} row of $\mathbf{C}$ is the derivative $\mathbf{b}'(u_l)$ of the basis functions evaluated at u_ℓ, where $u_1, \ldots, u_m$ are a fine grid of points over the range of x. Enforcing (6) ensures that the derivative of h is non-positive so h will be monotone decreasing. Of course minimizing (3) subject to (6) is a special case of PAC.

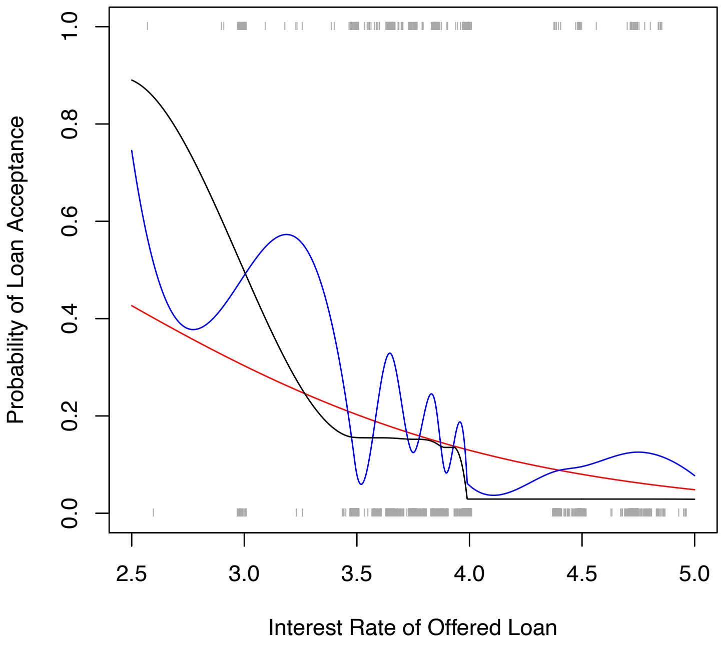

Figure 1: Auto loan data demand curves as a function of interest rate for the PAC (black), the logistic regression (red), and the lasso (blue).

Figure 1 provides an example of this approach applied to the auto loan data considered in Section 5.3. We have used the lasso (blue), standard logistic regression (red) and PAC (black) to estimate demand for loans as a function of interest rate. PAC and the lasso were both fit using a spline basis function which produces a flexible representation for the demand curve. The logistic regression curve is constrained in the shape it can model while the lasso fit is not sensible because it lacks a monotone shape. However, PAC produces a flexible, but still monotone, curve estimate.

2.2 Portfolio Optimization

Suppose we have p assets indexed by $1, 2, \ldots, p$ whose covariance matrix is denoted by $\boldsymbol{\Sigma}$. Markowitz (1952, 1959) developed the seminal framework for mean-variance analysis. In particular his approach involved choosing asset weights, $\boldsymbol{w}$, to minimize the portfolio risk

$$\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \tag{7}$$

subject to $\boldsymbol{w}^T\mathbf{1} = 1$. One may also choose to impose additional constraints on $\boldsymbol{w}$ to control the expected return of the portfolio, the allocations among sectors or industries, or the risk exposures to certain known risk factors.

In practice $\boldsymbol{\Sigma}$ is unobserved so must be estimated using the sample covariance matrix, $\hat{\boldsymbol{\Sigma}}$. However, it has been well documented in the finance literature that when p is large, which is the norm in real world applications, minimizing (7) using the sample covariance matrix gives poor estimates for $\boldsymbol{w}$. One possible solution involves regularizing $\hat{\boldsymbol{\Sigma}}$, but more recently attention has focused on directly penalizing or constraining the weights, an analogous approach to penalizing the coefficients in a regression setting. Fan *et al.* (2012) recently adopted this framework using a portfolio optimization problem with a gross-exposure parameter c defined by:

$$\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{w}, \quad \|\boldsymbol{w}\|_1 \leq c \tag{8}$$

subject to $\boldsymbol{w}^T\mathbf{1} = 1$, where c is a tuning parameter. Fan *et al.* (2012) reformulated (8) as a penalized linear regression problem and used LARS to estimate the weights. However, as they point out, their regression formulation only approximately solves (8). It is not hard to verify that (8) can be expressed in the form (5), where $\mathbf{C}$ has at least one row (to constrain $\boldsymbol{w}$ to sum to one) but may also have additional rows if we place constraints on the expected return, industry weightings, etc. Hence, implementing PAC with an identity link function allows us to exactly solve the gross-exposure portfolio optimization problem.

2.3 Generalized Lasso and Related Methods

Tibshirani and Taylor (2011) introduce the generalized lasso problem:

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_1, \tag{9}$$

where $\boldsymbol{D} \in \mathbb{R}^{r\times p}$. When $\mathsf{rank}(\boldsymbol{D}) = r$, and thus $r \leq p$, Tibshirani and Taylor (2011) show that the generalized lasso can be converted to the classical lasso problem. However, if $r > p$ then such a reformulation is not possible.

Lemma 1 shows that when $r > p$ and $\boldsymbol{D}$ is full column rank, then there is an interesting connection between the generalized lasso and the constrained lasso, which is a special case of (4).

Lemma 1 (Generalized Lasso is a Special Case of PAC)**.** *If $r > p$ and $\mathsf{rank}(\boldsymbol{D}) = p$ then there exist matrices $\mathbf{A}, \mathbf{C}$ and $\mathbf{X}$ such that, for all values of λ, the solution to (9) is equal to $\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is given by:*

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \textit{subject to} \quad \mathbf{C}\boldsymbol{\beta} = \mathbf{0}.$$

The proof of Lemma 1 is provided in Appendix A. Hence, any problem that falls into the generalized lasso paradigm can be solved as a constrained lasso problem with $\mathbf{b} = \mathbf{0}$.

Tibshirani and Taylor (2011) demonstrate that a variety of common statistical methods can be formulated as special cases of the generalized lasso. One example is the 1d fused lasso (Tibshirani *et al.*, 2005) which is defined as the solution to

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

The fused lasso encourages blocks of adjacent estimated coefficients to all have the same value. This type of structure often makes sense in situations where there is a natural ordering in the coefficients. If instead the data have a two-dimensional ordering, such as for an image reconstruction, this idea can be extended to the 2d fused lasso

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j,j'} |\beta_{j,j'}| + \lambda_2 \sum_{j \neq j'} |\beta_{j,j'} - \beta_{j,j'-1}| + \lambda_3 \sum_{j \neq j'} |\beta_{j,j'} - \beta_{j-1,j'}|,$$

where $\beta_{j,j'}$ is the coefficient at location j, j'.

Other examples of statistical methodologies that fall into the generalized (and hence constrained) lasso setting include; polynomial trend filtering where one penalizes discrete differences to produce smooth piecewise polynomial curves, Wavelet smoothing, and the FLiRTI method (James *et al.*, 2009). See Tibshirani and Taylor (2011) for more details on these methods and further examples of the generalized lasso paradigm. She (2010) also considers a similar criterion to (9) and discusses special cases such as the "clustered lasso". Since the constrained lasso is a special case of PAC with an identity link function and ℓ_1 penalty all of these various approaches can be solved using the PAC algorithm.

3. A Constrained GLM Fitting Algorithm

In the standard generalized linear models setting, it is common to maximize the likelihood by computing a quadratic approximation using the current parameter estimate, maximizing the quadratic, recomputing the approximation, and iterating; the so called *iterative reweighted least squares algorithm.* We take a similar approach to solve the PAC criterion (4). Let $g(\cdot)$ denote the canonical link function, and $V(\cdot)$ the variance function. Analogous to the iterative reweighted least squares approach, given the current parameter estimates $\bar{\boldsymbol{\beta}}$, a second-order Taylor approximation to the optimization problem in (4), up to irrelevant constants, is given by

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n} w_i\left(z_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}\gamma_j|\boldsymbol{\beta}_j| \quad \text{such that} \quad \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \tag{10}$$

where $\gamma_j = \rho'(|\bar{\beta}_j|)$, $z_i = \bar{\beta}_0 + \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}} + \frac{y_i - \bar{\mu}_i}{w_i}$, $w_i = V(\bar{\mu}_i)$, and $\bar{\mu}_i = g^{-1}\left(\bar{\beta}_0 + \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}}\right)$.

For example, consider a logistic regression involving binary responses. In this case, $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, $V(\mu) = \mu(1-\mu)$, and for $i = 1, \ldots, n$,

$$\mathsf{LogLik}_i(\boldsymbol{\beta}) = y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\left(1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}\right).$$

Thus, given the current parameter estimates

$$z_i = \bar{\beta}_0 + \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}} + \frac{y_i - \bar{\mu}_i}{w_i}, \quad w_i = \bar{\mu}_i(1-\bar{\mu}_i), \quad \text{and} \quad \bar{\mu}_i = \frac{e^{\beta_0 + \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}}}}{1 + e^{\beta_0 + \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}}}}.$$

Hence, to solve (4) we need to iteratively compute the solution to (10). In the next section, we develop an algorithm for solving a special case of (10), where we have equality constraints. Then, in Section 3.2, we show how to extend our proposed method to incorporate inequality constraints.

3.1 Equality Constraints

A common approach to deal with inequality constraints involves augmenting the set of parameters to include *slack variables*, $\boldsymbol{\delta}$, and hence reparameterizing the problem using equality constraints. We show in Section 3.2 that this augmentation approach also works to solve the PAC criterion. Throughout this section, we let $\mathbf{C}_{\mathcal{A}}$ represent the m columns of a matrix $\mathbf{C}$ associated with an index set $\mathcal{A}$, which is always of size m, corresponding to the number of equality constraints. Similarly let $\mathbf{C}_{\bar{\mathcal{A}}}$ correspond to the remaining columns of $\mathbf{C}$. To reduce notation, we assume without loss of generality that $\mathcal{A}$ corresponds to the first m columns of $\mathbf{C}$ so $\mathbf{C} = [\mathbf{C}_{\mathcal{A}}\ \ \mathbf{C}_{\bar{\mathcal{A}}}]$.

We begin by considering the equality constrained version of (10), which is given by

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{W}^{1/2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})\|_2^2 + \lambda\|\boldsymbol{\Gamma}\boldsymbol{\beta}\|_1 \quad \text{such that} \quad \mathbf{C}\boldsymbol{\beta}=\mathbf{b}, \tag{11}$$

where $\mathbf{Z}$ is an n-dimensional vector with ith entry z_i, $\mathbf{W}$ is an n-dimensional diagonal matrix corresponding to w_i and $\boldsymbol{\Gamma}$ is a p-dimensional diagonal matrix formed from γ_j. Unfortunately, the constraint on the coefficients makes it difficult to directly solve (11). However, Lemma 2 provides an approach for reformulating (11) as an unconstrained optimization problem.

Lemma 2 (Removing Constraints)**.** *For any $\lambda > 0$ and index set $\mathcal{A}$, define $\boldsymbol{\theta}_{\bar{\mathcal{A}}} \in \mathbb{R}^{p-m}$ by:*

$$\boldsymbol{\theta}_{\bar{\mathcal{A}}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{p-m}} \frac{1}{2}\|\mathbf{Z}^* - \mathbf{X}^*\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\|_1 + \lambda\|\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b}-\mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\right)\|_1 \ , \tag{12}$$

where $\mathbf{Z}^* = \mathbf{W}^{1/2}(\mathbf{Z}-\mathbf{X}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{b})$, $\mathbf{X}^* = \mathbf{W}^{1/2}(\mathbf{X}_{\bar{\mathcal{A}}}-\mathbf{X}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}})$. If $\mathbf{C}_{\mathcal{A}}$ is non-singular, then the solution to (11) *is given by*

$$\boldsymbol{\beta} = \begin{bmatrix} \mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b}-\mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}}}\right) \\ \boldsymbol{\theta}_{\bar{\mathcal{A}}} \end{bmatrix}.$$

Again, to reduce notation, we assume without loss of generality that the elements of $\boldsymbol{\beta}$ are ordered so that the first m correspond to $\mathcal{A}$. The optimization problem in (12) has the advantage that *there is no constraint on the parameters*. However, unfortunately, the second ℓ_1 term in (12) can not be separated into additive functions of θ_j which makes the criterion difficult to optimize directly. Fortunately, we can make use of Lemma 3 below which shows that an alternative, more tractable, criterion can be used to compute $\boldsymbol{\theta}_{\bar{\mathcal{A}}}$.

Lemma 3 (Transforming to a Separable Objective)**.** *For any $\lambda > 0$, an index set $\mathcal{A}$, and a vector $\mathbf{s} \in \{-1,+1\}^m$, define $\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}} \in \mathbb{R}^{p-m}$ by:*

$$\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{p-m}} \frac{1}{2}\|\tilde{\mathbf{Z}} - \mathbf{X}^*\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\|_1 \tag{13}$$

where $\tilde{\mathbf{Z}} = \mathbf{Z}^* + \lambda\mathbf{X}^-\left(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{s}$, and $\mathbf{X}^-$ is a matrix such that $\mathbf{X}^{*T}\mathbf{X}^- = \mathbf{I}$.

Then, it will be the case that $\boldsymbol{\theta}_{\bar{\mathcal{A}}} = \boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}$ provided that

$$\mathbf{s} = \mathit{sign}\left(\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}\right), \tag{Sign-Consistency}$$

where $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}} \equiv \mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b}-\mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right) \in \mathbb{R}^m$ and sign$(\mathbf{a})$ is a vector of the same dimension as $\mathbf{a}$ with ith element equal to 1 or -1 depending on the sign of a_i.

The proofs of Lemmas 2 and 3 are provided in Appendix B. There is a simple intuition behind Lemma 3. The difficulty in computing (12) lies in the non-differentiability and non-separability of the second ℓ_1 penalty. However, if $\mathbf{s}$ satisfies the Sign-Consistency condition, then

$$\|\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b}-\mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right)\|_1 = \mathbf{s}^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b}-\mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right)$$

and we can replace the ℓ_1 penalty by a differentiable term which no longer needs to be separable.

Choosing $\mathcal{A}$ and $\mathbf{s}$: Of course, for this approach to work we must select $\mathcal{A}$ and $\mathbf{s}$ such that the Sign-Consistency condition holds. The key idea here is that we fit PAC over a fine grid of values for λ, and at each point on the grid, we select $\mathcal{A}$ and $\mathbf{s}$ using the coefficient estimate associated with the previous value of λ. In particular let $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\lambda)$ represent the most recent coefficient estimate on the solution path. Then for our new grid point, $\tilde{\lambda}$, we pick $\mathcal{A}$ to be the m largest elements in absolute value of $\hat{\boldsymbol{\beta}}$ and $\mathbf{s} = \text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$. In order for this strategy to work, we only require that the m largest elements of $\hat{\boldsymbol{\beta}}$ satisfy the Sign-Consistency condition at $\tilde{\lambda}$; that is, they have the same sign as the corresponding elements in the final solution to (11). Lemma 4 assures us that the Sign-Consistency condition will be satisfied for all $\tilde{\lambda}$ close to λ ensuring that our proposed strategy will produce the entire solution path provided we use a fine enough grid of values.

Lemma 4 (Continuity of the Signs)**.** *For any $\lambda > 0$, index set $\mathcal{A}$, and $\mathbf{s} \in \{-1, 1\}^m$, if the* Sign-Consistency *condition holds for $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}$ at λ and every coordinate of $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}$ is bounded away from zero, then there exists $\epsilon > 0$ such that the* Sign-Consistency *condition also holds for all $\tilde{\lambda} \in (\lambda - \epsilon, \lambda + \epsilon)$.*

The proof of Lemma 4 is given in Appendix C. We emphasize that once we compute the coefficient estimate $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\tilde{\lambda})$ at the new grid point $\tilde{\lambda}$, the m largest elements in absolute value under $\tilde{\boldsymbol{\beta}}$ can be different from those of $\hat{\boldsymbol{\beta}}$, and thus, the set $\mathcal{A}$ chosen under $\tilde{\boldsymbol{\beta}}$ may change. This can happen because with $m < p$, there are many solutions satisfying the constraint $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$, and the Sign-Consistency condition does not impose any requirement on the magnitudes of the coefficients. So, as we move through the grid of λ, both $\mathcal{A}$ and $\mathbf{s}$ will change, depending on the solution at each grid point.

Lemmas 3 and 4 provide an attractive means of computing the PAC solution because the optimization problem (13) is a standard lasso criterion so can be solved using any one of the many highly efficient lasso optimization approaches, such as the LARS algorithm or coordinate descent. These results suggest Algorithm 1 for solving (11) over a grid of λ.

Algorithm 1 : Solution Path for the GLM with Equality Constraints

1. Initialize $\boldsymbol{\beta}_0$ by solving the GLM problem with equality constraints at the initial $\lambda_0 = \lambda_{\max}$. Set $k = 1$.

2. At step k, select $\mathcal{A}_k$ and $\mathbf{s}_k$ using the largest m elements of $|\boldsymbol{\beta}(\lambda_{k-1})|$ and set

   $$\lambda_k \leftarrow 10^{-\alpha}\lambda_{k-1},$$

   where $\alpha > 0$ controls the step size.

3. Set $\boldsymbol{\beta}_k = \begin{bmatrix} \boldsymbol{\theta}_{\mathcal{A}_k,\mathbf{s}_k} \\ \boldsymbol{\theta}_{\bar{\mathcal{A}}_k,\mathbf{s}_k} \end{bmatrix}$ where $\boldsymbol{\theta}_{\bar{\mathcal{A}}_k,\mathbf{s}_k}$ is a solution to the optimization problem in (13) with the index set $\mathcal{A}_k$ and sign vector $\mathbf{s}_k$, and $\boldsymbol{\theta}_{\mathcal{A}_k,\mathbf{s}_k} = \mathbf{C}_{\mathcal{A}_k}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}_k}\boldsymbol{\theta}_{\bar{\mathcal{A}}_k,\mathbf{s}_k}\right)$.

4. Use the latest parameter estimates to update $\mathcal{A}_k, \mathbf{s}_k, \mathbf{Z}, \mathbf{W}$ and $\boldsymbol{\Gamma}$. Repeat Steps 3 and 4 until convergence. This produces a solution to the constrained GLM problem for λ_k.

5. $k \leftarrow k + 1$ and return to Step 2.

6. Iterate until $\lambda_k < \lambda_{\min}$. The algorithm outputs the solution path $\{\boldsymbol{\beta}_k : k = 0, 1, 2, \ldots\}$.

In Step 3 of Algorithm 1, we must also check that the Sign-Consistency condition holds. Lemma 4 guarantees that the condition holds provided the change in λ is small enough. Hence, if the condition is violated, then we halve the step size and repeat 3. In practice, unless m is very large or α is set too large, the Sign-Consistency condition generally holds without needing to decrease the step size. Step 3 is the main computational component of this algorithm but $\boldsymbol{\theta}_{\bar{\mathcal{A}}_k,\mathbf{s}_k} \in \mathbb{R}^{p-m}$ is easy to compute because (13) is just a standard lasso criterion so we can use any one of a number of optimization tools.

Initialization: The initial solution can be computed by noting that as $\lambda \to \infty$ the solution to (11) will be

$$\arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Gamma}\boldsymbol{\beta}\|_1 \quad \text{such that} \quad \mathbf{C}\boldsymbol{\beta} = \mathbf{b}\ , \tag{14}$$

which is a linear programming problem that can be efficiently solved using standard algorithms. We also implement a reversed version of this algorithm where we first set $\lambda_0 = \lambda_{\min}$, compute $\boldsymbol{\beta}_0$ as the solution to a quadratic programming problem and then increase λ at each step until $\lambda_k > \lambda_{\max}$. We discuss some additional implementation details in Appendix D.

PAC Solution Paths and Why Path-Following Algorithms Do Not Work? Figure 2 compares the unconstrained GLM solution with the PAC coefficients for a Gaussian simulated data set (top row) and a logistic response data set (bottom row), plotted as a function of the ℓ_1 norm of the coefficients. In both cases the coefficients satisfy a set of m linear constraints. The left-hand plots correspond to standard unconstrained lasso or GLM fits. The right-hand plots represent the same set of coefficients computed using Algorithm 1, after incorporating the m equality constraints. Where λ is small (right-hand side of each plot), the two plots are similar. However, for large values of λ the unconstrained and PAC coefficient estimates differ significantly from each other because the linear constraints do not allow the coefficients to be set to zero.

The top row of Figure 2 shows that the lasso paths are piecewise linear (Efron *et al.*, 2004). However, the logistic regression coefficient paths in the bottom row are *not* piecewise linear. Thus, it would be difficult to compute the paths of the PAC coefficients using traditional path-following algorithms, such as LARS, that rely heavily on piecewise linear paths. Algorithm 1 does not require linearity of the paths.

3.2 Inequality Constraints

We now consider the more general setting with inequality constraints. One might imagine that a reasonable approach would be to initialize with $\boldsymbol{\beta}$ such that

$$\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b} \tag{15}$$

and then apply a coordinate descent algorithm subject to ensuring that at each update (15) is not violated. Unfortunately, this approach typically gets stuck at a constraint boundary point where no improvement is possible by changing a single coordinate. In this setting the criterion can often be improved by moving along the constraint boundary but such a move requires adjusting multiple coefficients simultaneously which is not possible using coordinate descent because it only updates one coefficient at a time.

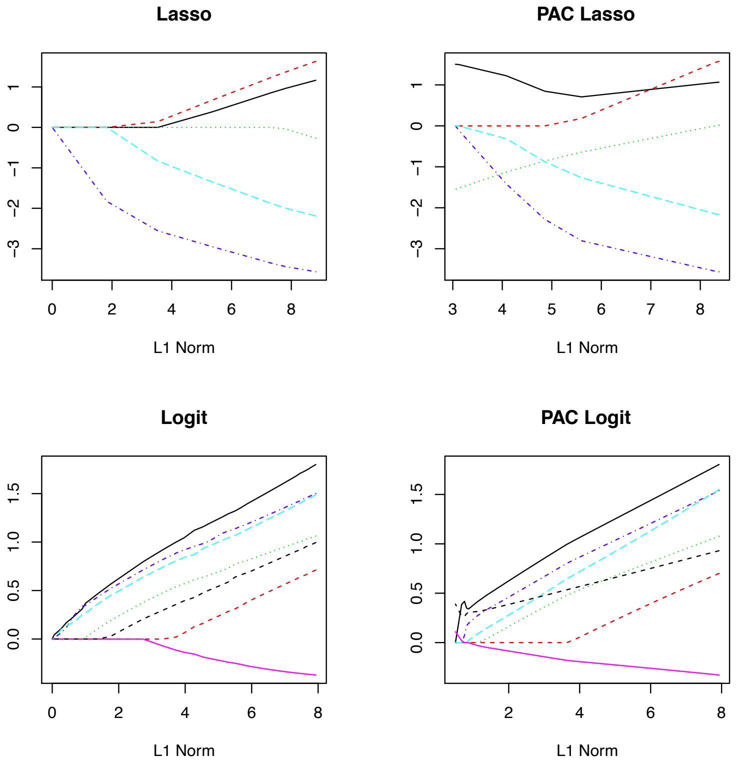

Figure 2: *Top Left:* Unconstrained lasso coefficient paths for a simulated data set with $n = 50$ observations and $p = 5$ predictors. *Top Right:* PAC coefficient paths computed using Algorithm 1, after incorporating $m = 2$ linear constraints. *Bottom Left:* Unconstrained GLM logistic coefficient paths for a simulated data set with $n = 50$ observations and $p = 7$ predictors. *Bottom Right:* PAC coefficient paths computed after incorporating $m = 5$ linear constraints.

Instead we introduce a set of m slack variables, $\boldsymbol{\delta}$, so that (15) can be equivalently expressed as

$$\mathbf{C}\boldsymbol{\beta} + \boldsymbol{\delta} = \mathbf{b}, \boldsymbol{\delta} \geq \mathbf{0} \quad \text{or} \quad \tilde{\mathbf{C}}\tilde{\boldsymbol{\beta}} = \mathbf{b}, \boldsymbol{\delta} \geq \mathbf{0}, \tag{16}$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \boldsymbol{\delta})$ is a $p + m$-dimensional vector, $\tilde{\mathbf{C}} = [\mathbf{C} \;\; \mathbf{I}] \in \mathbb{R}^{m \times (p+m)}$ and $\mathbf{I}$ is an m-dimensional identity matrix. Let $\mathbf{e}_{\boldsymbol{\delta}}(\mathbf{a})$ be a function which selects out the elements of $\mathbf{a}$ that correspond to $\boldsymbol{\delta}$. For example, $\mathbf{e}_{\boldsymbol{\delta}}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\delta}$ while $\mathbf{e}_{\bar{\boldsymbol{\delta}}}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Then, the inclusion of the slack variables in (16) allows us to reexpress the Taylor approximation of (4) as

$$\arg \min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+m}} \frac{1}{2}\|\mathbf{W}^{1/2}(\mathbf{Z} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}})\|_2^2 + \lambda\|\boldsymbol{\Gamma}\mathbf{e}_{\bar{\boldsymbol{\delta}}}(\tilde{\boldsymbol{\beta}})\|_1 \quad \text{such that} \quad \tilde{\mathbf{C}}\tilde{\boldsymbol{\beta}} = \mathbf{b},\ \mathbf{e}_{\boldsymbol{\delta}}(\tilde{\boldsymbol{\beta}}) \geq \mathbf{0}, \tag{17}$$

where $\tilde{\mathbf{X}} = [\mathbf{X} \ \ \mathbf{0}] \in \mathbb{R}^{n \times (p+m)}$ and $\mathbf{0}$ is a n-by-m matrix of zero elements. The criterion in (17) is very similar to the equality constrained version (11). The only differences are that the components of $\tilde{\boldsymbol{\beta}}$ corresponding to $\boldsymbol{\delta}$ do not appear in the penalty term and are required to be non-negative.

Even with these minor differences, the same basic approach from Section 3.1 can still be adopted for fitting (17). In particular Lemma 5 provides a set of conditions under which (17) can be solved.

Lemma 5. *For a given index set $\mathcal{A}$, and vector $\mathbf{s} \in \{-1, +1\}^m$ such that $\mathbf{e}_{\boldsymbol{\delta}}(\mathbf{s}) = 0$, define $\tilde{\boldsymbol{\theta}}_{\bar{\mathcal{A}},\mathbf{s}}$ by:*

$$\tilde{\boldsymbol{\theta}}_{\bar{\mathcal{A}},\mathbf{s}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2}\|\tilde{\mathbf{Z}} - \mathbf{X}^* \boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\Gamma} \mathbf{e}_{\bar{\boldsymbol{\delta}}}(\boldsymbol{\theta})\|_1 \ \ \textit{such that} \ \ \mathbf{e}_{\boldsymbol{\delta}}(\boldsymbol{\theta}) \geq 0, \tag{18}$$

where $\tilde{\mathbf{Z}} = \mathbf{W}^{1/2}(\mathbf{Z} - \tilde{\mathbf{X}}_{\mathcal{A}} \tilde{\mathbf{C}}_{\mathcal{A}}^{-1} \mathbf{b}) + \lambda \mathbf{X}^{-} \left(\tilde{\mathbf{C}}_{\mathcal{A}}^{-1} \tilde{\mathbf{C}}_{\bar{\mathcal{A}}}\right)^T \boldsymbol{\Gamma}_{\mathcal{A}} \mathbf{s}$, $\mathbf{X}^* = \mathbf{W}^{1/2}(\tilde{\mathbf{X}}_{\bar{\mathcal{A}}} - \tilde{\mathbf{X}}_{\mathcal{A}} \tilde{\mathbf{C}}_{\mathcal{A}}^{-1} \tilde{\mathbf{C}}_{\mathcal{A}})$, and $\mathbf{X}^{-}$ is a matrix such that $\mathbf{X}^{*T} \mathbf{X}^{-} = \mathbf{I}$. Suppose

$$\begin{aligned} \mathbf{e}_{\bar{\boldsymbol{\delta}}}(\mathbf{s}) &= \textit{sign}\left(\mathbf{e}_{\bar{\boldsymbol{\delta}}}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{A},\mathbf{s}}\right)\right), & (19) \\ \mathbf{e}_{\boldsymbol{\delta}}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{A},\mathbf{s}}\right) &\geq 0, & (20) \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{\mathcal{A},\mathbf{s}} = \tilde{\mathbf{C}}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \tilde{\mathbf{C}}_{\bar{\mathcal{A}}} \tilde{\boldsymbol{\theta}}_{\bar{\mathcal{A}},\mathbf{s}}\right)$. Then, the solution to (17) is given by $\mathbf{e}_{\bar{\delta}}(\tilde{\boldsymbol{\beta}})$ where,

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \tilde{\boldsymbol{\theta}}_{\mathcal{A},\mathbf{s}} \\ \tilde{\boldsymbol{\theta}}_{\bar{\mathcal{A}},\mathbf{s}} \end{bmatrix}.$$

The proof of this result is similar to that for Lemma 3 so we omit it here. Lemma 5 shows that, provided an appropriate $\mathcal{A}$ and $\mathbf{s}$ are chosen, the solution to PAC can still be computed by solving a lasso type criterion, (18). However, we must now ensure that both (19) and (20) hold. Condition 19 is equivalent to the Sign-Consistency condition in the equality constraint setting, while (20) along with the constraint in (18) ensure that $\boldsymbol{\delta} \geq \mathbf{0}$.

Hence, we can use almost exactly the same approach as in the equality constraint setting. In particular the only change that is required in Algorithm 1 is to compute $\tilde{\boldsymbol{\theta}}_{\mathcal{A}_k,\mathbf{s}_k}$ by solving (18) instead of (13). Solving (18) poses little additional complication over fitting the standard lasso. The only differences are that the elements of $\boldsymbol{\theta}$ that correspond to $\boldsymbol{\delta}$, i.e. $\mathbf{e}_{\boldsymbol{\delta}}(\boldsymbol{\theta})$, have zero penalty and must be non-negative. However, these changes are simple to incorporate into a coordinate descent algorithm. For any θ_j that is an element of $\mathbf{e}_{\boldsymbol{\delta}}(\boldsymbol{\theta})$ we first compute the unshrunk least squares estimate, $\hat{\theta}_j$, and then set

$$\tilde{\theta}_j = \left[\hat{\theta}_j\right]_+ . \tag{21}$$

It is not hard to show that (21) enforces the non-negative constraint on $\boldsymbol{\delta}$ while also ensuring that no penalty term is applied to the slack variables. The update step for the original $\boldsymbol{\beta}$ coefficients, $\mathbf{e}_{\bar{\boldsymbol{\delta}}}(\boldsymbol{\theta})$ (those that do not involve $\boldsymbol{\delta}$), is identical to that for the standard lasso.

We must also ensure that (19) and (20) hold at each step. However, Lemma 4 guarantees that the conditions hold provided the change in λ is small enough. Hence, if either (19) or (20) are violated we halve the step size and repeat Step 3. The initial solution, $\tilde{\boldsymbol{\beta}}_0$, can still be computed by solving a standard linear programming problem, subject to inequality constraints.

4. Statistical Properties: Error Bounds

In this section we show that, assuming the true coefficients correspond to a given set of constraints, one can obtain sharper error bounds for the constrained estimate relative to the traditional unconstrained GLM estimate. To describe the results, let us introduce the statistical model for our problem. We assume that the data are generated so that $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)$ are independent and identically distributed, with $(\mathbf{X}_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ for all i. Also,

$$\Pr\left\{Y = y \;\middle|\; \mathbf{X} = \boldsymbol{x}\right\} \;=\; \exp\left(\frac{[y\cdot(\boldsymbol{x}^T\boldsymbol{\beta}^*)] - \psi\left(\boldsymbol{x}^T\boldsymbol{\beta}^*\right)}{a(\sigma)} + c(y,\sigma)\right),$$

where $a(\cdot)$ is a scaling function with $\sigma > 0$ a *fixed and known* scale parameter, whereas $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is fixed but *unknown.* Our goal is to estimate $\boldsymbol{\beta}^*$. It is a standard result that $\psi(\cdot)$ is infinitely differentiable, and its second derivative $\psi''(\cdot)$ is strictly positive on $\mathbb{R}$. The above model encompasses the standard examples in GLM, including linear, logistic, and Poisson regression models.

Let $\widehat{\boldsymbol{\beta}}(\lambda)$ denote a solution of the constrained GLM problem defined by:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \mathcal{L}\left(\boldsymbol{\beta}; \mathbf{X}, \boldsymbol{Y}\right) + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \mathbf{C}\boldsymbol{\beta} = \mathbf{b}, \tag{22}$$

where

$$\mathcal{L}\left(\boldsymbol{\beta}; \mathbf{X}, \boldsymbol{Y}\right) = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\boldsymbol{\beta}^T\mathbf{X}_i\right) - \left[Y_i \cdot \left(\boldsymbol{\beta}^T\mathbf{X}_i\right)\right]$$

We make the following assumptions about the underlying parameter $\boldsymbol{\beta}^*$, the constraint matrix $\mathbf{C}$, the covariates $\mathbf{X}$, and the link function ψ.

Assumption 1 (Sparsity)**.** *Let* $\mathsf{supp}(\boldsymbol{\beta}^*) = \left\{j : \beta_j^* \neq 0\right\}$*. The parameter* $\boldsymbol{\beta}^*$ *is* s*-sparse with* $|\mathsf{supp}(\boldsymbol{\beta}^*)| = s$*.*

Assumption 2 (Constraint Matrix)**.** *There exists a matrix* $\mathbf{C} \in \mathbb{R}^{m \times p}$ *with* $\mathsf{rank}(\mathbf{C}) = m \leq s$ *such that* $\mathbf{C}\boldsymbol{\beta}^* = \mathbf{b}$*. Moreover, the matrix* $\mathbf{C}_{\mathcal{A}}$ *is non-singular for some* $\mathcal{A} \subseteq \mathsf{supp}(\boldsymbol{\beta}^*)$ *such that* $|\mathcal{A}| = m$*.*

Assumption 3 (Covariates)**.** *The random vectors* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *are i.i.d with zero mean and covariance matrix* $\boldsymbol{\Sigma}$ *such that the smallest eigenvalue of* $\boldsymbol{\Sigma}$ *is bounded below by* $\kappa_\ell > 0$*, and for any vector* $\mathbf{v} \in \mathbb{R}^p$*, the random variable* $\mathbf{v}^T\mathbf{X}_i$ *is sub-Gaussian with a parameter at most* $\kappa_u \left\|\mathbf{v}\right\|_2^2$*.*

Assumption 4 (Regularity of the Link Functions)**.** [1] *There exists a constant* $M_\psi > 0$ *such that* $\sup_{u \in \mathbb{R}} \psi''(u) \leq M_\psi$*.*

Assumption 1 is a standard assumption on the true parameter. By Assumption 2, we can write $\boldsymbol{\beta}^* = [\boldsymbol{\beta}^*_{\mathcal{A}}, \boldsymbol{\beta}^*_{\bar{\mathcal{A}}}]$. Let $S = \mathsf{supp}(\boldsymbol{\beta}^*_{\bar{\mathcal{A}}})$ denote the support of the vector $\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}$. Note that Assumptions 1 and 2 imply that $|S| = s - m$. Assumption 3 ensures that the covariates are "non-degenerate" with minimum eigenvalue bounded away from zero. Finally, it is easy to verify that Assumption 4 is satisfied for the linear and logistic with $M_\psi = 1$.

The main result of this write-up is stated in the following theorem, which provides a deterministic error bound for each instance of $\mathbf{X}$ and $\boldsymbol{Y}$. The proof of Theorem 1 is given in Appendix E. Let us introduce the following notation. For any vector $\boldsymbol{u}$, let $\Pi_S(\boldsymbol{u})$ and $\Pi_{S^\perp}(\boldsymbol{u})$ be defined by:

$$[\Pi_S(\boldsymbol{u})]_j = \begin{cases} u_j & \text{if } j \in S, \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad [\Pi_{S^\perp}(\boldsymbol{u})]_j = \begin{cases} u_j & \text{if } j \notin S, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\boldsymbol{u} = \Pi_S(\boldsymbol{u}) + \Pi_{S^\perp}(\boldsymbol{u})$. Also, for any $J > 1$, let a cone $\mathbb{C}_J \in \mathbb{R}^p$ be defined by:

$$\mathbb{C}_J \;\equiv\; \left\{ \boldsymbol{\Delta} \in \mathbb{R}^p \;:\; \mathbf{C}\boldsymbol{\Delta} = \mathbf{0} \quad \text{and} \quad \|\Pi_{S^\perp}(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 \leq \frac{J+1}{J-1}\Big(\|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \Big) \right\} \quad .$$

It is easy to verify that $\mathbb{C}_J$ is a cone because if $\boldsymbol{x} \in \mathbb{C}_J$, then $t\boldsymbol{x} \in \mathbb{C}_J$ for all $t \geq 0$.

[1] Assumption 4 can be extended to allow for Poisson regression by requiring that $|X_{ij}| \leq 1$ for all i and j and $\mathsf{E}\left\{\left[\max_{|u| \leq 1} \psi''(\boldsymbol{X}^T\boldsymbol{\beta}^* + u)\right]^\alpha\right\} \leq M_\psi$ for some $\alpha \geq 2$. The proof is similar and we omit the details.

Theorem 1 (Deterministic Bounds)**.** *Under Assumptions 1 and 2, suppose that for some* $J > 1$*, the loss function* $\mathcal{L}(\cdot\,;\mathbf{X},\boldsymbol{Y})$ *satisfies the* **restricted strong convexity** *on* $\mathbb{C}_J \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$ *with parameter* $\kappa_{\mathcal{L}} > 0$*; that is, for all* $\boldsymbol{\Delta} \in \mathbb{C}_J \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$,

$$\mathcal{L}\left(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \mathbf{X}, \boldsymbol{Y}\right) - \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right) - \nabla\mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)^T \boldsymbol{\Delta} \geq \kappa_{\mathcal{L}} \left\|\boldsymbol{\Delta}\right\|_2^2 \ .$$

Moreover, suppose that

$$J \left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)\right\|_\infty \quad \leq \quad \lambda \quad < \quad \frac{J\,\kappa_{\mathcal{L}}}{(J+1)\sqrt{2}\,\max\{\sqrt{s-m}, \sqrt{m}\,\}} \ .$$

Then,

$$\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_2 \leq \sqrt{\max\{\, s-m \ , \ m\,\}} \cdot \sqrt{2}\left(1 + \frac{1}{J}\right) \cdot \frac{\lambda}{\kappa_{\mathcal{L}}}$$

and

$$\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_1 \leq \max\{\, s-m \ , \ m\,\} \cdot 4\left(1 + \frac{2}{J-1}\right) \cdot \frac{\lambda}{\kappa_{\mathcal{L}}}.$$

Once we incorporate the probabilistic information about the covariates $\mathbf{X}$ and the structure of the link function, we can obtain probabilistic error bounds, which are given in the following theorem. The proof is given in Appendix F.

Theorem 2 (High Probability Error Bounds for PAC)**.** *Suppose Assumptions 1 - 4 hold. Then, there exist positive constants* (c_0, c_1, c_2, c_3) *that depend only on* κ_ℓ*,* κ_u*, and* M_ψ *such that if* $n \ > \ c_0 \max\{s-m, m\}\cdot \log p$ *and* $\lambda = c_1\sqrt{\frac{\log p}{n}}$*, then any solution to the constrained GLM in* (22) *satisfies the following bound*

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_2 \ &\leq \ c_2\sqrt{\max\{\, s-m \ , \ m\,\}} \cdot \sqrt{\frac{\log p}{n}} \\ \left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_1 \ &\leq \ c_2 \ \max\{\, s-m \ , \ m\,\} \ \cdot \sqrt{\frac{\log p}{n}}\end{aligned}$$

with probability at least $1 - (c_3/n)$*.*

We note that the error bounds have the same flavor as the more standard unconstrained bound in Negahban *et al.* (2012), but our error bound scales with $\max\{s-m, m\}$, instead of just s. Since $\max\{s-m, m\} \leq s$ the above error bound demonstrates the benefits of PAC, especially for m close to $s/2$. For $m \leq s/2$, our ℓ_2 bound is of order $O(\sqrt{s-m}\,)$. This rate follows from the fact that Assumption 2 implies

$$\boldsymbol{\beta}^*_{\mathcal{A}} = \mathbf{C}_{\mathcal{A}}{}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}\right). \tag{23}$$

Hence, the m coordinates of $\boldsymbol{\beta}^*_{\mathcal{A}}$ are completely determined by the remaining $p-m$ coordinates of $\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}$. Since $\left|\mathsf{supp}\left(\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}\right)\right| = s-m$, the vector $\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}$ is $(s-m)$-sparse, which can be interpreted as the "adjusted sparsity" level. Why then do the bounds in Theorem 2 also depend on m? In fact, this term reflects the error due to model selection. To see this, note that when $m = s$, it follows from Assumptions 1 and 2 and Equation (23) that we can exactly recover $\boldsymbol{\beta}^*$, but only if we know the locations of the non-zero entries $\mathcal{A}$. There are $\binom{p}{s} \sim p^s$ possible locations of the non-zero entries. Intuitively, this corresponds to $O(p^s)$ possible hypotheses, and information theoretic arguments show that even if we have $m = s$ constraints, we still need n to be of order $\log\binom{p}{s} \approx s\log p$ to identify the correct hypothesis, and hence ensure that our error $\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_2$ is small.

We note that the bounds in Theorem 2 only depend on s and m, and not on the constraint matrix $\mathbf{C}$. It is possible to develop an error bound that depends on $\mathbf{C}$, in which case the bound scales with $\sqrt{s-m} + \sqrt{m}\,\nu_{\max}(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}})$, where $\nu_{\max}(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}})$ is the largest singular value of $\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}$. This error bound is useful, for example, when the magnitude of the elements of $\mathbf{C}_{\bar{\mathcal{A}}}$ are much smaller than those of $\mathbf{C}_{\mathcal{A}}$. For example, in the extreme case where the constraints only involve the m largest coefficients then $\mathbf{C}_{\bar{\mathcal{A}}} = \mathbf{0}$ and $\nu_{\max} = 0$ so the bound scales with $\sqrt{s-m}$. The proof of this bound is similar to that of Theorem 2, so we omit the details

Sharper Error Bounds for Constrained Lasso: In the linear model where $\boldsymbol{Y}$ is normally distributed, Corollary 1 provides explicit constants for the error bounds, which are competitive with the results in the existing literature. The result follows from Theorem 1, and uses a similar proof technique as Theorem 2, so we omit the details.

Corollary 1 (Probabilistic Error Bounds for Linear Case)**.** *Under Assumptions 1 and 2, suppose the matrix $\boldsymbol{X}$ satisfies the* **restricted eigenvalue** *condition on $\mathbb{C}$ with $\frac{\|\boldsymbol{X}\boldsymbol{u}\|_2^2}{n} \geq 2\kappa_{\mathcal{L}}\|\boldsymbol{u}\|_2^2$ for all $\boldsymbol{u} \in \mathbb{C}$. If $\lambda = (4\sqrt{2}-2)\sigma\sqrt{\frac{\log p}{n}}$ and $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ are i.i.d. mean-zero sub-Gaussian random variables with variance σ^2, then with probability at least $1 - 2/p$,*

$$\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_2 \leq \sqrt{\max\{\, s-m\ ,\ m\,\}} \cdot \frac{8\sigma}{\kappa_L}\sqrt{\frac{\log p}{n}} \tag{24}$$

By comparison Negahban *et al.* (2012) show that, for $\lambda = 4\sigma\sqrt{\frac{\log p}{n}}$, the standard unconstrained lasso estimate $\widehat{\boldsymbol{\beta}}_L(\lambda)$ satisfies the following error bound:

$$\left\| \widehat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}^* \right\|_2 \leq \sqrt{s} \cdot \frac{8\sigma}{\kappa_{\mathcal{L}}}\sqrt{\frac{\log p}{n}} \quad ,$$

with probability at least $1 - 2/p$. Hence, the lasso and PAC bounds differ only in the first term, with the PAC bound strictly better than that of the lasso for $1 \leq m < s$. It is worth noting that the constant in (24) can be lowered further by choosing a slightly lower value for λ but we have not done so because our goal here is to highlight the difference in the dependence on s between the PAC and lasso. A bound of the same form as (24) can also be produced for the ℓ_1 error, although a somewhat larger value for λ is optimal in that case.

5. Empirical Results

In this section, we present empirical results for our proposed PAC formulation and algorithm. In Section 5.1, we compare PAC with unconstrained GLM for binomial and Poisson responses, and with unconstrained lasso for linear responses. When the true underlying parameters satisfy equality constraints, our PAC formulation can yield significant improvements in the prediction accuracy over unconstrained methods. In addition, Section 5.2 shows that our approach is robust, so that even when the true parameters violate some of the constraints, PAC still yields superior estimates. In Section 5.3, we apply our method to real data from an online auto lending company, to estimate the demand curve. Finally, in Section 5.4, we also test our method on simulated demand data. The numerical experiments show that PAC yields a more realistic and accurate estimate of the demand curve than existing approaches in the literature.

5.1 Equality Constraints

In this section we examine the performance of PAC with binomial and Poisson responses, as well as with a Gaussian linear relationship between the predictors and responses. In all cases we use a standard ℓ_1 penalty. We provide comparison results between the binomial PAC and the unconstrained logistic generalized linear model, the Poisson PAC and the unconstrained Poisson generalized linear model, and finally the Gaussian PAC and the unconstrained lasso (Tibshirani, 1996). While the GLM methods are unconstrained they still contain an ℓ_1 penalty and are fit using the glmnet function in R (Friedman *et al.*, 2010)

For each method, we considered six simulation settings, three different combinations of n observations and p predictors (incorporating both traditional and high-dimensional problems) and those same combinations with an imposed correlation structure. For each setting we randomly generated a training set, fit each method to the data, and computed the error over a test set of $N = 10000$ observations. The metric used to calculate error necessarily varied between the types of data. For the binomial responses, the error metric was the percentage of incorrect binomial predictions; for the Gaussian and Poisson cases, the error computations were done using root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - E(Y_i|X_i)\right)^2}.$$

This process was repeated 100 times for each of the six settings. The training data sets were produced using a random design matrix generated from a normal distribution. For the binomial and Gaussian results, this was a standard normal distribution, but due to the potential for very large extreme values with randomly-generated Poisson responses, the normal distribution used for the Poisson data was limited to $N(0, 0.25)$.

Two different correlation structures were investigated as well: $\rho_{ij} = 0$ and $\rho_{ij} = 0.5^{|i-j|}$ (where ρ_{ij} is the correlation between the ith and jth variables). In all cases, the m-by-p constraint matrix $\mathbf{C}$ and the constraint vector $\mathbf{b}$ were randomly generated from the same normal distribution as the data. The true coefficient vector $\boldsymbol{\beta}^*$ was produced by first generating $\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}$ using 5 non-zero random uniform components and $p - m - 5$ zero entries, and then computing $\boldsymbol{\beta}^*_{\mathcal{A}}$ from (23). Note that this process resulted in $\boldsymbol{\beta}^*$ having at most $m + 5$ non-zero entries, and ensured that the constraints held for the true coefficient vector. For each set of simulations, the optimal value of λ was chosen by minimizing error on a separate validation set, which was independently generated using the same parameters as for the given simulation setting.

For each method we explored three combinations of n, p, and m: a low-dimensional, traditional problem with few constraints, a more complex, higher-dimensional problem with few constraints, and a high-dimensional problem with more constraints. Tables 1 through 3 respectively provide the test error values for the binomial, Poisson and Gaussian responses. For each method, we compared results from four different approaches: the standard penalized GLM fit (Friedman *et al.*, 2010), PAC, the relaxed GLM, and the relaxed PAC. The latter

	ρ	Bayes	GLM	PAC	Relaxed GLM	Relaxed PAC
$n = 100, p = 50$	0	12.27(0.11)	19.36(0.23)	18.56(0.24)	19.33(0.45)	17.68(0.31)
$m = 5$	$0.5^{\|i-j\|}$	9.30(0.10)	14.60(0.18)	14.08(0.19)	14.51(0.20)	13.34(0.25)
$n = 1000, p = 500$	0	11.02(0.19)	12.33(0.22)	12.00(0.26)	12.14(0.26)	11.68(0.28)
$m = 10$	$0.5^{\|i-j\|}$	8.60(0.19)	10.15(0.28)	9.76(0.31)	10.01(0.33)	9.44(0.27)
$n = 50, p = 100$	0	8.20(0.06)	43.17(1.07)	36.06(0.85)	41.60(1.01)	31.23(0.63)
$m = 30$	$0.5^{\|i-j\|}$	7.26(0.10)	37.73(1.58)	28.03(0.78)	35.67(1.47)	24.54(0.70)

Table 1: Average misspecification error (in percentages) over 100 training data sets for four binomial methods tested in three different simulation settings and two different correlation structures. The Bayes error rate is given for comparison; it represents the minimum error rate. The numbers in parentheses are standard errors, also in percentages.

	ρ	GLM	PAC	Relaxed GLM	Relaxed PAC
$n = 100, p = 50$	0	1.09(0.05)	0.89(0.02)	0.96(0.02)	0.80(0.02)
$m = 5$	$0.5^{\|i-j\|}$	1.42(0.06)	1.22(0.03)	1.29(0.03)	1.08(0.02)
$n = 1000, p = 500$	0	0.62(0.03)	0.58(0.03)	0.58(0.02)	0.53(0.02)
$m = 10$	$0.5^{\|i-j\|}$	0.83(0.08)	0.68(0.07)	0.69(0.04)	0.58(0.04)
$n = 50, p = 100$	0	6.87(0.46)	5.79(0.24)	6.16(0.26)	4.94(0.17)
$m = 30$	$0.5^{\|i-j\|}$	8.99(0.51)	7.42(0.52)	7.78(0.45)	6.37(0.39)

Table 2: Average RMSE over 100 training data sets, for four Poisson methods tested in three different simulation settings and two different correlation structures. The numbers in parentheses are standard errors.

two methods use a two-step approach in an attempt to reduce the over shrinkage problem commonly exhibited by the ℓ_1 penalty. In the first step, the given method is used to select a candidate set of predictors. In the second step, the final model is produced using an unshrunk ordinary least squares fit on the variables selected in the first step.

The first set of results in each table correspond to a setting with $n = 100, p = 50$ and $m = 5$. In the binomial case, PAC shows a moderate but not large improvement over the standard GLM fit, which one might expect given this is a relatively low dimensional problem with only a small number of constraints. However, even with this low value for m, in the Poisson and Gaussian cases PAC shows highly statistically significant improvements over the unconstrained methods. Both relaxed methods display lower error rates than their unrelaxed counterparts, and the higher correlations in the $\rho = 0.5^{|i-j|}$ design structure do not change the relative rankings of the four approaches.

	ρ	Lasso	PAC	Relaxed Lasso	Relaxed PAC
$n = 100, p = 50$	0	0.57(0.01)	0.50(0.01)	0.43(0.01)	0.29(0.01)
$m = 5$	$0.5^{\|i-j\|}$	0.59(0.01)	0.46(0.01)	0.50(0.02)	0.32(0.01)
$n = 50, p = 500$	0	3.39(0.08)	3.28(0.09)	3.29(0.09)	3.07(0.10)
$m = 10$	$0.5^{\|i-j\|}$	2.61(0.08)	2.39(0.09)	2.50(0.09)	2.23(0.10)
$n = 50, p = 100$	0	3.90(0.08)	1.57(0.05)	3.81(0.07)	1.47(0.04)
$m = 30$	$0.5^{\|i-j\|}$	3.15(0.06)	1.45(0.04)	3.00(0.05)	1.34(0.03)

Table 3: Average RMSE over 100 training data sets, for four lasso methods tested in three different simulation settings and two different correlation structures. The numbers in parentheses are standard errors.

The second set of results in the tables represents a more complex setting with $n = 50, p = 500$, and $m = 10$ for the Gaussian. We use a larger sample size ($n = 1000$) in the binomial and Poisson settings because these distributions provide less information for estimating the regression coefficients. As one would expect, given the low ratio of m relative to p, PAC only shows small improvements over its unconstrained counterparts. However, our main purpose in examining this setting was to prove that the PAC algorithm is still efficient enough to optimize the constrained criterion even for large data sets and very high dimensional data.

The final setting in all three cases examined data with $n = 50, p = 100$ and a larger number of constraints, $m = 30$. This setting is more statistically favorable for PAC because it has the potential to produce significantly more accurate regression coefficients by correctly incorporating the larger number of constraints. However, this is also a computationally difficult setting for PAC because a large value of m causes the coefficient paths to be highly variable. Nevertheless, the large improvements in accuracy for both PAC and relaxed PAC in Tables 1, 2, and 3 demonstrate that our algorithm is quite capable of dealing with this added complexity.

Figure 3 provides a better feeling for the general improvement in accuracy from incorporating the constraints into our fit. Here we display boxplots of the RMSE for the Gaussian setting for each of our four comparison methods, over the 100 training data sets on two of the six simulation settings. The improvement ranges from moderate in the first setting when m is low, to more dramatic in the second setting where m is higher.

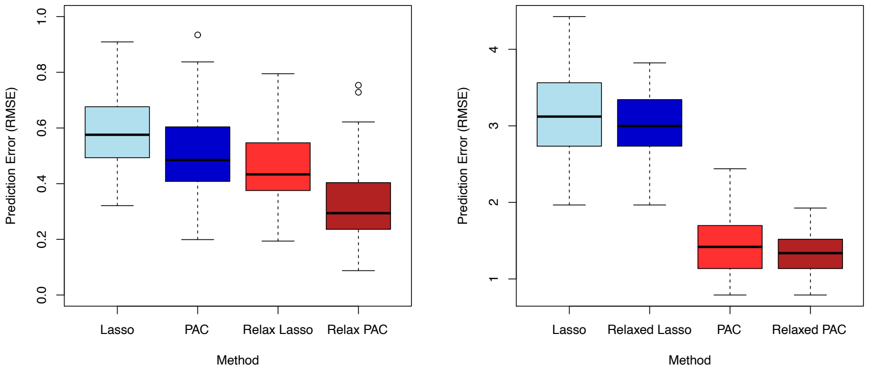

Figure 3: Boxplots of the 100 simulation RMSE values in the Gaussian setting using the $\rho = 0.5^{|i-j|}$ correlation structure. *Left*: The $n = 100, p = 50, m = 5$ setting. *Right*: The $n = 50, p = 100, m = 30$ setting.

5.2 Violations of Constraints

The results presented in the previous section all correspond to an ideal situation where the true regression coefficients exactly match the equality constraints. We also investigated the sensitivity of PAC to deviations of the regression coefficients from the assumed constraints. In particular we generated the true regression coefficients according to

$$\mathbf{C}\boldsymbol{\beta}^* = (\mathbf{1} + \mathbf{u}) \cdot \mathbf{b}, \tag{25}$$

where $\mathbf{u} = (u_1, \ldots, u_m)$, $u_l \sim Unif(0, a)$ for $l = 1, \ldots m$ and the vector product is taken pointwise. The PAC and relaxed PAC were then fit using the usual constraint, $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$.

Table 4 reports the new RMSE values for three Gaussian settings under the $\rho = 0$ correlation structure; the first two settings correspond exactly to the first two settings considered in Section 5.1, while the last is a setting with a very large number of constraints to demonstrate robustness even when $n < m$.

We tested four values for a: $0.25, 0.50, 0.75$ and 1.00. The largest value of a corresponds to a 50% average error in the constraint. The results suggest that PAC and relaxed PAC are surprisingly robust to random violations in the constraints. While both methods deteriorated slightly as a increased, they were still both superior to their unconstrained counterparts for all values of a and all settings.

	a	Lasso	PAC	Relaxed Lasso	Relaxed PAC
$n = 100$	0.25	0.57(0.01)	0.51(0.01)	0.43(0.01)	0.30(0.01)
$p = 50$	0.50	0.57(0.01)	0.51(0.01)	0.42(0.01)	0.33(0.01)
$m = 5$	0.75	0.57(0.01)	0.52(0.01)	0.42(0.01)	0.35(0.01)
	1.00	0.57(0.01)	0.53(0.01)	0.42(0.01)	0.38(0.01)
$n = 50$	0.25	3.34(0.08)	3.22(0.09)	3.24(0.09)	2.98(0.10)
$p = 500$	0.50	3.34(0.08)	3.22(0.09)	3.23(0.09)	2.99(0.10)
$m = 10$	0.75	3.37(0.08)	3.28(0.09)	3.26(0.09)	3.07(0.10)
	1.00	3.36(0.08)	3.28(0.09)	3.23(0.09)	3.06(0.10)
$n = 50$	0.25	6.66(0.07)	1.17(0.03)	6.80(0.08)	0.95(0.03)
$p = 100$	0.50	6.66(0.07)	1.18(0.03)	6.81(0.08)	0.99(0.03)
$m = 60$	0.75	6.65(0.07)	1.21(0.03)	6.79(0.08)	1.03(0.03)
	1.00	6.68(0.07)	1.27(0.04)	6.86(0.08)	1.08(0.04)

Table 4: Average RMSE over 100 training data sets in three different simulation settings using the $\rho = 0$ correlation structure. The numbers in parentheses are standard errors. The true regression coefficients were generated according to (25)

5.3 On-line Auto Loan Data

We now consider an application of PAC to estimate a non-parametric demand function on the "On-line Auto Lending" data set, courtesy of *The Center for Pricing and Revenue Management* of the Columbia Business School. As the name suggests, the data contains observations on all United States auto loans approved by an online lender. Each record corresponds to a specific customer who was approved for a loan, then either chose to accept or decline the offer. The data also includes characteristics of the customer and offered loan, such as FICO score and loan duration. In particular, the data is segmented by three variables: credit tier (Tier I-VII, with Tier I being the highest), length of loan (0-36 months, 37-48 months, 49-60 months, or over 60 months), and type of car loan (new, used, or refinanced). As with previous analyses on the auto lending data set (Besbes *et al.*, 2010), the analysis here is performed on a segment of the overall data: Tier I loans for new cars with a duration of 0 to 36 months. This subset contained observations on 8698 customers, which we randomly divided into one-half training data, one-quarter validation data, and one-quarter test data.

We used the offered rate as the predictor and the loan's funding status, 1 if the customer accepted the loan and 0 otherwise, as the response.

A cubic spline basis, $\mathbf{b}(t)$, was used to model the demand curve as a non-linear function of the interest rate. To ensure that the demand function was monotonically decreasing as a function of the interest rate we computed the first derivative of the spline over a fine grid of m interest rates, between the minimum and the maximum observed rate. We then fit PAC (using the logistic link function) over the training data with the ith row of $\mathbf{X}$ derived from $\mathbf{b}(x_i)$, where x_i is the ith interest rate, and the first derivative constrained to be less than zero at each grid point i.e. $\mathbf{b}'(u_l)^T\boldsymbol{\beta} \leq 0$ for $l = 1, \ldots, m$. We compared the PAC fit to an unconstrained lasso with the same spline design matrix. The degrees of freedom for $\mathbf{b}(t)$ and λ for the lasso and PAC were both chosen by minimizing RMSE on the validation data. To provide a comparison with a traditional parametric approach we also fit a logistic regression curve to the data.

Figure 1 plots the three resulting demand functions, with PAC in black, the traditional logistic in red, and the standard lasso in blue. Jittered versions of the test data are also provided for context. As expected the linear logistic curve provides a smooth monotone decreasing fit to the data. However, the constrained parametric form of the logistic means that it is not capable of both modeling the high acceptance rate near 2.5% and the much lower acceptance rate near 5%. By comparison the lasso fit using the spline basis is far more flexible, modeling both the high acceptance rate near 2.5% and the much lower acceptance rate near 5%. However, because the curve is unconstrained it provides an unrealistic shape which actually suggests demand is increasing and decreasing in wild oscillations. Of course if the validation data had suggested a smaller degrees of freedom the lasso would produce a smoother curve, but even in that case the resulting fit will generally not be monotone.

By comparison the PAC fit combines both the flexibility of the lasso with the intuitively reasonable monotone decreasing shape of the logistic curve. The PAC suggests a sharp decline in demand between interest rates of 2.5% and 3.5%, followed by a relatively constant demand between 3.5% and 4.0%, and then a further decline for rates above 4%. It is interesting that both 3.5% and 4.0% seem to represent distinct change points in demand, possibly corresponding to psychological effects from mentally rounding to the nearest half percent.

γ	Lasso	Logit	PAC
0	1.64(0.08)	0.45(0.04)	1.22(0.08)
1/3	1.69(0.08)	0.87(0.04)	1.25(0.09)
2/3	2.05(0.10)	2.00(0.05)	1.56(0.10)
1	2.16(0.11)	3.81(0.07)	1.70(0.10)

Table 5: A comparison of the prediction accuracies of the lasso, logistic, and PAC methods. The values are the mean integrated squared differences over 200 simulation runs (multiplied by a factor of 10^3 for comparison purposes, as are the standard errors in parentheses).

5.4 Demand Function Simulation

We ran a simulation study to test out the accuracy of PAC in a setting such as the on-line auto loan demand estimation problem. To mimic the real life setting we generated Bernoulli responses with

$$P(Y_i = 1|X_i) = (1-\gamma)L(X_i) + \gamma P(X_i), \quad 0 \le \gamma \le 1, \tag{26}$$

where $L(X)$ and $P(X)$ respectively corresponded to the logistic and PAC fits to the auto loan data. With $\gamma = 0$ the data was generated from a smooth underlying demand function, while for larger values of γ the true demand function had more flexibility.

We tested $\gamma = 0, 1/3, 2/3$ and 1, and for each value generated 200 data sets according to (26). The simulated interest rates were generated along an equally spaced grid from 2.5% to 5.0% in step sizes of 0.002% for a total of 1251 observations. Each data set was then randomly divided into one-half training data, one-quarter validation data, and one-quarter testing data. The PAC, lasso and logistic regression were then fit in the same fashion as in Section 5.3. In particular PAC and the lasso were each fit over a range of degrees of freedom (df) and λ values, then the combination of df and λ which yielded the minimum $SSE = \sum_i (Y_i - \hat{Y}_i)^2$ value on the validation set was chosen as the optimal model for a particular simulation. That model was then used to fit the testing data.

Table 5 provides a comparison of the prediction accuracies over the testing data for the three methods on each of the four choices of γ. The values in the table are the estimated mean integrated squared differences (ISD)

$$\int \Big(P(Y=1|X) - \hat{Y}(x)\Big)^2 dx$$

computed by averaging the test observations and the 200 simulation runs, where $\hat{Y}(x)$ is the predicted demand function at x for a given method. As one would expect, the logistic regression provides superior results for low values of γ. However, as the true demand function becomes more flexible PAC starts to dominate. At $\gamma = 1$ PAC has an ISD that is less than half that of the logistic regression fit. While the lasso also improves relative to the logistic fit for larger values of γ, it is clearly inferior to PAC in all settings.

6. Conclusion

In this paper we have illustrated a few of the wide range of statistical applications for the PAC formulation, and developed computationally efficient path algorithms for computing its solutions. Our theoretical results demonstrate that, for correctly chosen constraints, the PAC solution provides the potential for improvements in prediction accuracy. Furthermore, our simulation results show that the PAC estimates generally outperform the unconstrained estimates, not only when the constraints hold exactly, but also when there is some reasonable error in the assumption on the constraints.

In our analysis of the statistical error bounds, we have concentrated on results for the equality constrained PAC. Results in the inequality setting are more complicated. Intuitively, one can see that if $\boldsymbol{\beta}^*$ lies well inside the region $\mathbf{C}\boldsymbol{\beta}^* \leq \mathbf{b}$, then both PAC and unconstrained GLM should give similar answers because the constraints will play little role in the fit. However, if $\boldsymbol{\beta}^*$ lies close to the constraint boundary, then PAC should offer significant improvements relative to the unconstrained estimation. As shown in Section 3, one can reformulate the GLM with inequality constraints using equality constraints with the addition of a set of m slack variables $\boldsymbol{\delta}$. If $\boldsymbol{\beta}^*$ lies close to the boundary then these slack variables will be close to, but not exactly, zero. Hence, to develop bounds for the inequality setting one must adapt Assumption 1 to one where the augmented set of coefficients is not necessarily sparse, but can be well approximated by a sparse vector. This is a promising area for future research.

A. Proof of Lemma 1

Since $\boldsymbol{D}$ has full column rank, by reordering the rows if necessary, we can write $\boldsymbol{D}$ as

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \end{bmatrix}$$

where $\boldsymbol{D}_1 \in \mathbb{R}^{p \times p}$ is an invertible matrix and $\boldsymbol{D}_2 \in \mathbb{R}^{r-p \times p}$. Then,

$$\begin{aligned} \frac{1}{2}\left\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_1 &= \frac{1}{2}\left\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{D}_1^{-1}\boldsymbol{D}_1\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{D}_1\boldsymbol{\theta}\right\|_1 + \lambda\left\|\boldsymbol{D}_2\boldsymbol{\theta}\right\|_1 \\ &= \frac{1}{2}\left\|\boldsymbol{Y} - (\tilde{\mathbf{X}}\boldsymbol{D}_1^{-1})\boldsymbol{D}_1\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{D}_1\boldsymbol{\theta}\right\|_1 + \lambda\left\|\boldsymbol{D}_2\boldsymbol{D}_1^{-1}\boldsymbol{D}_1\boldsymbol{\theta}\right\|_1 \end{aligned}$$

Using the change of variables

$$\boldsymbol{\beta}_1 = \boldsymbol{D}_1\boldsymbol{\theta}, \quad \boldsymbol{\beta}_2 = \boldsymbol{D}_2\boldsymbol{D}_1^{-1}\boldsymbol{D}_1\boldsymbol{\theta} = \boldsymbol{D}_2\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_1, \quad \text{and} \quad \boldsymbol{\beta} \;=\; \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix},$$

we can rewrite the generalized lasso problem as follows:

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_1 &= \min_{\boldsymbol{\beta}\in\mathbb{R}^r} \left\{ \frac{1}{2}\left\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_1\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1 \;\middle|\; \boldsymbol{D}_2\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \mathbf{0} \right\}, \\ &= \min_{\boldsymbol{\beta}\in\mathbb{R}^r} \left\{ \frac{1}{2}\left\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1 \;\middle|\; \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \right\}, \end{aligned}$$

where $\mathbf{X} = \begin{bmatrix}\tilde{\mathbf{X}}\boldsymbol{D}_1^{-1} & \mathbf{0}\end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix}\boldsymbol{D}_2\boldsymbol{D}_1^{-1} & -\boldsymbol{I}\end{bmatrix}$. Note that $\boldsymbol{\theta} = [\boldsymbol{D}^{-1} \;\; \mathbf{0}] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$, and thus, the generalized lasso is a special case of the constrained lasso.

B. Proofs of Lemmas 2 and 3

Consider any index set $\mathcal{A}$ such that $\mathbf{C}_{\mathcal{A}}$ is non-singular. The constraint $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$ can be written as

$$\mathbf{C}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}} = \mathbf{b} \quad \Leftrightarrow \quad \boldsymbol{\beta}_{\mathcal{A}} = \mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right),$$

and thus, we can determine $\boldsymbol{\beta}_{\mathcal{A}}$ from $\boldsymbol{\beta}_{\bar{\mathcal{A}}}$. Then, for any $\boldsymbol{\beta}$ such that $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$,

$$\begin{aligned}
&\frac{1}{2}\left\|\mathbf{W}^{1/2}\left(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}\right)\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}\boldsymbol{\beta}\right\|_1 \\
&= \frac{1}{2}\left\|\mathbf{W}^{1/2}\left(\mathbf{Z} - \mathbf{X}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}\right)\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right\|_1 + \lambda\left\|\boldsymbol{\Gamma}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}\right\|_1 \\
&= \frac{1}{2}\left\|\mathbf{W}^{1/2}\left(\mathbf{Z} - \mathbf{X}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}} - \mathbf{X}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right)\right)\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right\|_1 + \lambda\left\|\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right)\right\|_1 \\
&= \frac{1}{2}\left\|\mathbf{W}^{1/2}\left[\mathbf{Z} - \mathbf{X}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{b}\right] - \mathbf{W}^{1/2}\left(\mathbf{X}_{\bar{\mathcal{A}}} - \mathbf{X}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right\|_2^2 \\
&\qquad + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right\|_1 + \lambda\left\|\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}\right)\right\|_1 .
\end{aligned}$$

By using the change of variable $\boldsymbol{\theta} = \boldsymbol{\beta}_{\bar{\mathcal{A}}}$, the constrained GLM problem is equivalent to the following unconstrained optimization problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{p-m}} \frac{1}{2}\left\|\mathbf{Z}^* - \mathbf{X}^*\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\right\|_1 + \lambda\left\|\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\right)\right\|_1 ,$$

and let $\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)$ denote a solution to the above optimization problem. Then, a solution to the original PAC lasso problem is given

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{\mathcal{A}} \\ \boldsymbol{\beta}_{\bar{\mathcal{A}}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}}}\right) \\ \boldsymbol{\theta}_{\bar{\mathcal{A}}} \end{bmatrix} ,$$

and this completes Lemma 2.

To prove Lemma 3, fix an arbitrary $\lambda > 0$ and consider $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}$ and $\mathbf{s}$ such that $\mathbf{s} = \operatorname{sign}\left(\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}\right)\right)$. Let $F : \mathbb{R}^{p-m} \rightarrow \mathbb{R}_+$ denote the objective function of the optimization problem in Equation (12); that is, for each $\boldsymbol{\theta} \in \mathbb{R}^{p-m}$,

$$F(\boldsymbol{\theta}) = \frac{1}{2}\left\|\mathbf{Z}^* - \mathbf{X}^*\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\right\|_1 + \lambda\left\|\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}\right)\right\|_1 .$$

By definition of $\boldsymbol{\theta}_{\bar{\mathcal{A}}}$, we have $F(\boldsymbol{\theta}_{\bar{\mathcal{A}}}) \leq F(\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}})$. To complete the proof, it suffices to show that $F(\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}) \leq F(\boldsymbol{\theta}_{\bar{\mathcal{A}}})$. Suppose, on the contrary, that $F(\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}) > F(\boldsymbol{\theta}_{\bar{\mathcal{A}}})$. For each $\alpha \in [0, 1]$, let $\boldsymbol{\theta}_\alpha \in \mathbb{R}^{p-m}$ and $g(\alpha) \in \mathbb{R}_+$ be defined by:

$$\boldsymbol{\theta}_\alpha \equiv (1-\alpha)\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}} + \alpha\boldsymbol{\theta}_{\bar{\mathcal{A}}} \qquad \text{and} \qquad g(\alpha) \equiv F(\boldsymbol{\theta}_\alpha) = F\left((1-\alpha)\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}} + \alpha\boldsymbol{\theta}_{\bar{\mathcal{A}}}\right) .$$

Note that g is a convex function on $[0, 1]$ because $F(\cdot)$ is convex. Moreover, we have $g(0) = F(\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}) > F(\boldsymbol{\theta}_{\bar{\mathcal{A}}}) = g(1)$. Thus, for all $0 < \alpha \leq 1$,

$$g(\alpha) = g(\alpha \cdot 1 + (1-\alpha) \cdot 0) \leq \alpha g(1) + (1-\alpha)g(0) < g(0) .$$

By our hypothesis, $|s_i| = 1$ for all i, and thus, every coordinate of the vector $\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right)$ is bounded away from zero. So, we can choose α_0 sufficiently small so that

$$\operatorname{sign}\left(\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right)\right) = \operatorname{sign}\left(\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right)\right) \quad .$$

Then, it follows that

$$\begin{aligned} F(\boldsymbol{\theta}_{\alpha_0}) &= \frac{1}{2}\left\|\mathbf{Z}^* - \mathbf{X}^*\boldsymbol{\theta}_{\alpha_0}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right\|_1 + \lambda\mathbf{s}^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\left(\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right) \\ &= \frac{(\mathbf{Z}^*)^T\mathbf{Z}^*}{2} - (\mathbf{Z}^*)^T\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0} - \lambda\mathbf{s}^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0} + \frac{(\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0})^T\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0}}{2} \\ &\quad + \ \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right\|_1 + \lambda\mathbf{s}^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{b} \\ &= \frac{(\mathbf{Z}^*)^T\mathbf{Z}^*}{2} - (\mathbf{Z}^*)^T\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0} - \left(\lambda\mathbf{X}^-\left(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)^T\mathbf{s}\right)^T\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0} \\ &\quad + \ \frac{(\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0})^T\mathbf{X}^*\boldsymbol{\theta}_{\alpha_0}}{2} + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right\|_1 + \lambda\mathbf{s}^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{b} \\ &= \frac{1}{2}\left\|\tilde{\mathbf{Z}} - \mathbf{X}^*\boldsymbol{\theta}_{\alpha_0}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right\|_1 + \mathbf{d} \end{aligned}$$

where the last equality follows from $\tilde{\mathbf{Z}} = \mathbf{Z}^* + \lambda\mathbf{X}^-\left(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)^T\mathbf{s}$, and $\mathbf{d}$ is a constant vector defined by

$$\mathbf{d} = -\lambda(\mathbf{Z}^*)^T\mathbf{X}^-\left(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)^T\mathbf{s} \ - \ \frac{\left[\lambda\mathbf{X}^-\left(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)^T\mathbf{s}\right]^T\left[\lambda\mathbf{X}^-\left(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}}\right)^T\mathbf{s}\right]}{2} \ + \ \lambda\mathbf{s}^T\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{b} \ .$$

Also, the third equality follows from the fact that $(\mathbf{X}^-)^T\mathbf{X}^* = \mathbf{I}$.

It follows from the same argument that

$$F(\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}) = \frac{1}{2}\left\|\tilde{\mathbf{Z}} - \mathbf{X}^*\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right\|_1 + \mathbf{d}$$

Since $g(\alpha) < g(0)$, we have that $F(\boldsymbol{\theta}_{\alpha_0}) < F(\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}})$, and this implies that

$$\frac{1}{2}\left\|\tilde{\mathbf{Z}} - \mathbf{X}^*\boldsymbol{\theta}_{\alpha_0}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\alpha_0}\right\|_1 \quad < \quad \frac{1}{2}\left\|\tilde{\mathbf{Z}} - \mathbf{X}^*\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right\|_2^2 + \lambda\left\|\boldsymbol{\Gamma}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}\right\|_1 \ ,$$

but this contradicts the optimality of $\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}$! Therefore, it must be the case that $F(\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}) \ \le \ F(\boldsymbol{\theta}_{\bar{\mathcal{A}}})$, which completes the proof.

C. Proof of Lemma 4

Proof. The optimization problem in (12) can be written as

$$\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{p-m}} \frac{1}{2}\|\mathbf{Z}^* - \mathbf{X}^*\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{h}\|_1,$$

where $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Gamma}_{\bar{\mathcal{A}}} \\ -\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{C}_{\bar{\mathcal{A}}} \end{bmatrix}$ and $\boldsymbol{h} = \boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{b}$, where we write $\boldsymbol{\theta}_{\bar{\mathcal{A}}}$ as $\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)$ to emphasize the dependence on λ. With the exception of the constant vector $\boldsymbol{h}$, the above problem is an instance of the Generalized Lasso problem. Using the same argument as in Tibshirani and Taylor (2011), we can show that the solution $\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)$ is continuous in λ. By Lemma 3, $\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}(\lambda) = \boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)$. Since every coordinate of $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}(\lambda) = \mathbf{C}_{\mathcal{A}}^{-1}\left[\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)\right]$ is bounded away from zero, the continuity of $\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)$ means that for ϵ sufficiently small,

$$\operatorname{sign}\left(\mathbf{C}_{\mathcal{A}}^{-1}\left[\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\tilde{\lambda})\right]\right) \;=\; \operatorname{sign}\left(\mathbf{C}_{\mathcal{A}}^{-1}\left[\mathbf{b} - \mathbf{C}_{\bar{\mathcal{A}}}\boldsymbol{\theta}_{\bar{\mathcal{A}}}(\lambda)\right]\right) = \mathbf{s}$$

The above result shows that $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}(\tilde{\lambda})$ have the same sign as $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}(\lambda)$, and consequently, the Sign-Consistency condition holds for $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}(\tilde{\lambda})$ as well. □

D. Algorithm Implementation Details

Implementing the PAC lasso algorithm requires making a choice for $\mathbf{X}^{-}$, which is generally not difficult. If $p \leq n + m$ then it is easy to see that $\mathbf{X}^{-} = \mathbf{U}\mathbf{D}^{-1}\mathbf{V}^{T}$ satisfies $\mathbf{X}^{*T}\mathbf{X}^{-} = \mathbf{I}$ where $\mathbf{X}^{*} = \mathbf{U}\mathbf{D}\mathbf{V}^{T}$ represents the singular value decomposition of $\mathbf{X}^{*}$. If $p > n+m$ then we use the fact that in general for Lemma 3 to hold we only require $\mathbf{X}^{-}$ to be chosen such that

$$\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}} = \mathbf{X}^{-T}\mathbf{X}^{*}\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}, \tag{27}$$

where $\boldsymbol{\theta}_{\mathcal{A},\mathbf{s}}$ is the solution to (13). But standard properties of the lasso tell us that $\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}$ can have at most n non-zero components. Hence, (27) will hold if we choose $\mathbf{X}^{-}$ to be the inverse of the columns of $\mathbf{X}^{*}$ corresponding to the (at most) n non-zero columns of $\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}$. Of course we do not know a priori with complete certainty which elements of $\boldsymbol{\theta}_{\bar{\mathcal{A}},\mathbf{s}}$ will be non-zero. However, based on the solution to the previous step in the algorithm, it is easy to compute the elements that are furthest from becoming non-zero and these can generally be safely ignored in computing $\mathbf{X}^{-}$. On the rare occasions where an incorrect set of columns is selected we simply reduce the step size in λ.

As mentioned in Section 3.1, one can generally initialize the algorithm using the solution to (14). However, this approach could potentially fail if one of the constraints in $\mathbf{C}$ is parallel with $\|\boldsymbol{\beta}\|_1$, for example $\sum \beta_j = 1$, in which case there may not be a unique solution to (14). In this setting we use quadratic programming to initialize the algorithm, which is slightly less efficient, but does not unduly impact the computational burden because the solution only needs to be found for a single value of λ.

E. Proof of Theorem 1

The proof of Theorem 1 follows the approach pioneered by Negahban *et al.* (2012), and it makes use of the following lemma, which establishes a lower bound on the deviation of under one-norm.

Lemma 3 (Deviation Inequality)**.** *For any vectors $\boldsymbol{u}$ and $\boldsymbol{v}$,*

$$\|\boldsymbol{u}+\boldsymbol{v}\|_1 - \|\boldsymbol{u}\|_1 \geq \left\|\Pi_{\mathsf{supp}(\boldsymbol{u})^\perp}\left(\boldsymbol{v}\right)\right\|_1 - \left\|\Pi_{\mathsf{supp}(\boldsymbol{u})}\left(\boldsymbol{v}\right)\right\|_1 \ .$$

Proof. Let $B = \mathsf{supp}(\boldsymbol{u})$. Note that for any coordinate j, if $j \notin B$, then $|u_j + v_j| - |u_j| = |v_j|$. On the other hand, if $j \in B$, then $|u_j + v_j| - |u_j| \geq -\left|(u_j+v_j) - u_j\right| = -\left|v_j\right|$. So,

$$\|\boldsymbol{u}+\boldsymbol{v}\|_1 - \|\boldsymbol{u}\|_1 \;=\; \left(\sum_{j\notin B} |u_j+v_j| - |u_j|\right) + \left(\sum_{j\in B} |u_j+v_j| - |u_j|\right) = \left(\sum_{j\notin B} |v_j|\right) - \left(\sum_{j\in B} |v_j|\right) \ ,$$

which is the desired result. □

Finally, here is the proof of Theorem 1.

Proof. Fix an arbitrary λ such that

$$J \left\|\nabla \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)\right\|_\infty \quad \leq \quad \lambda \quad < \quad \frac{J\ \kappa_{\mathcal{L}}}{(J+1)\sqrt{2}\ \max\{\sqrt{s-m}, \sqrt{m}\}}$$

for some $J > 1$. Let $\widehat{\boldsymbol{\Delta}}(\lambda) \equiv \widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*$. To simplify our exposition, we will drop the reference to λ and write $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Delta}}$ to denote $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\Delta}}(\lambda)$, respectively. Let $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$ be defined by: for any $\boldsymbol{\Delta} \in \mathbb{R}^p$,

$$\mathcal{F}(\boldsymbol{\Delta}) \;\equiv\; \mathcal{L}\left(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \mathbf{X}, \boldsymbol{Y}\right) - \mathcal{L}(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}) + \lambda\left\{\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1\right\}$$

By definition of $\widehat{\boldsymbol{\beta}}$, we have that

$$\widehat{\boldsymbol{\Delta}} = \arg\min\left\{\mathcal{F}(\boldsymbol{\Delta}) \mid \mathbf{C}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) = \mathbf{b}\right\}$$

Since both $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ satisfy the constraint, we have that $\mathcal{F}(\widehat{\boldsymbol{\Delta}}) \leq \mathcal{F}(\mathbf{0}) = 0$.

Consider an arbitrary $\boldsymbol{\Delta} \in \mathbb{C}_J \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$. By the strong convexity of $\mathcal{L}(\cdot; \mathbf{X}, \boldsymbol{Y})$ on $\mathbb{C}_J \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$, we have

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}) &\geq \kappa_{\mathcal{L}} \|\boldsymbol{\Delta}\|_2^2 + \nabla \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)^T \boldsymbol{\Delta} + \lambda \left(\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1\right) \\
&\geq \kappa_{\mathcal{L}} \|\boldsymbol{\Delta}\|_2^2 - \frac{1}{J} \lambda \|\boldsymbol{\Delta}\| + \lambda \left(\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1\right) \\
&= \kappa_{\mathcal{L}} \|\boldsymbol{\Delta}\|_2^2 + \lambda \left(\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1 - \frac{1}{J} \|\boldsymbol{\Delta}\|_1\right) \ ,
\end{aligned}$$

where the inequality follows from our hypothesis and the fact that

$$\nabla \mathcal{L}(\boldsymbol{\beta}^*; \boldsymbol{Y})^T \boldsymbol{\Delta} \geq - \|\nabla \mathcal{L}(\boldsymbol{\beta}^*; \boldsymbol{Y})\|_\infty \|\boldsymbol{\Delta}\|_1 \geq -\frac{\lambda}{J} \|\boldsymbol{\Delta}\|_1 \ .$$

By definition of $\|\cdot\|_1$, we have

$$\begin{aligned}
&\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1 - \frac{1}{J} \|\boldsymbol{\Delta}\|_1 \\
&= \left(\|\boldsymbol{\beta}^*_{\bar{\mathcal{A}}} + \boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_1 - \|\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}\|_1 - \frac{1}{J} \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_1\right) + \left(\|\boldsymbol{\beta}^*_{\mathcal{A}} + \boldsymbol{\Delta}_{\mathcal{A}}\|_1 - \|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1 - \frac{1}{J} \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1\right) \\
&\geq \left(\|\Pi_{S^\perp}\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1 - \|\Pi_S\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1 - \frac{1}{J} \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_1\right) - \frac{J+1}{J} \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \\
&\geq -\frac{J+1}{J} \|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 - \frac{J+1}{J} \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \ ,
\end{aligned}$$

where the first inequality follows from the application of Lemma 3 to $\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}$ and $\boldsymbol{\Delta}_{\bar{\mathcal{A}}}$ with $S = \mathsf{supp}(\boldsymbol{\beta}^*_{\bar{\mathcal{A}}})$, and the fact that $\|\boldsymbol{\beta}^*_{\mathcal{A}} + \boldsymbol{\Delta}_{\mathcal{A}}\|_1 - \|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1 \geq - \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1$. The final inequality follows from the fact that $\|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_1 = \|\Pi_S\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\| + \|\Pi_{S^\perp}\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1$, which implies that

$$\begin{aligned}
\|\Pi_{S^\perp}\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1 - \|\Pi_S\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1 - \frac{1}{J} \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_1 &= \frac{J-1}{J} \|\Pi_{S^\perp}\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1 - \frac{J+1}{J} \|\Pi_S\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\| \\
&\geq -\frac{J+1}{J} \|\Pi_S\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)\|_1
\end{aligned}$$

Therefore,

$$\mathcal{F}(\boldsymbol{\Delta}) \ \geq \ \kappa_{\mathcal{L}} \|\boldsymbol{\Delta}\|_2^2 - \frac{J+1}{J} \lambda \left(\|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1\right) \ .$$

Since $\left|\mathsf{supp}(\boldsymbol{\beta}^*_{\bar{\mathcal{A}}})\right| = |S| = s - m$ by Assumption 1, the vector $\Pi_S\left(\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\right)$ has at most $s - m$ non-zero coordinates, which implies that

$$\|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 \ \leq \ \sqrt{s-m} \|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_2 \ \leq \ \sqrt{s-m} \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_2 \ \leq \ \max\{ \sqrt{s-m}, \sqrt{m} \} \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_2 \ ,$$

where the last inequality follows from the fact that $\Pi_S(\cdot)$ is non-expansive under $\|\cdot\|_2$. We also have

$$\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \ \leq \ \sqrt{m} \|\boldsymbol{\Delta}_{\mathcal{A}}\|_2 \ \leq \ \max\{ \sqrt{s-m}, \sqrt{m} \} \|\boldsymbol{\Delta}_{\mathcal{A}}\|_2 \ ,$$

where the first inequality follows from the fact that $\mathbf{\Delta}_{\mathcal{A}} \in \mathbb{R}^m$.

Putting everything together, we have that for all $\mathbf{\Delta} \in \mathbb{C}_J \cap \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_2 \leq 1\}$,

$$\begin{aligned} \mathcal{F}(\mathbf{\Delta}) &\geq \kappa_{\mathcal{L}} \|\mathbf{\Delta}\|_2^2 - \frac{(J+1)\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J} \left(\|\mathbf{\Delta}_{\bar{\mathcal{A}}}\|_2 + \|\mathbf{\Delta}_{\mathcal{A}}\|_2\right) \\ &\geq \kappa_{\mathcal{L}} \|\mathbf{\Delta}\|_2^2 - \frac{(J+1)\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J} \cdot \sqrt{2}\, \|\mathbf{\Delta}\|_2 \\ &= \kappa_{\mathcal{L}} \|\mathbf{\Delta}\|_2 \left(\|\mathbf{\Delta}\|_2 - \frac{(J+1)\sqrt{2}\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J\, \kappa_{\mathcal{L}}} \right) , \end{aligned} \tag{28}$$

where the second inequality follows from the fact that $\|\mathbf{\Delta}_{\bar{\mathcal{A}}}\|_2 + \|\mathbf{\Delta}_{\mathcal{A}}\|_2 \leq \sqrt{2}\, \|\mathbf{\Delta}\|_2$.

Claim: $\left\|\widehat{\mathbf{\Delta}}\right\|_2 \leq 1$

To establish this claim, suppose on the contrary that $\left\|\widehat{\mathbf{\Delta}}\right\|_2 > 1$. Let $t = 1/\left\|\widehat{\mathbf{\Delta}}\right\|_2$. Then, $0 < t < 1$ and $\left\|t\widehat{\mathbf{\Delta}}\right\|_2 = 1$. By convexity of $\mathcal{F}$, we have that

$$\mathcal{F}\left(t\widehat{\mathbf{\Delta}}\right) = \mathcal{F}\left(t\widehat{\mathbf{\Delta}} + (1-t)\mathbf{0}\right) \leq t\mathcal{F}(\widehat{\mathbf{\Delta}}) + (1-t)\mathcal{F}(\mathbf{0}) \leq 0 ,$$

where the last inequality follows from the fact that $\widehat{\mathbf{\Delta}} = \arg\min\{\mathcal{F}(\mathbf{\Delta}) \mid \mathbf{C}(\boldsymbol{\beta}^* + \mathbf{\Delta}) = \mathbf{b}\}$, and thus, $\mathcal{F}(\widehat{\mathbf{\Delta}}) \leq \mathcal{F}(\mathbf{0}) = 0$. Since $\mathbb{C}_J$ is a cone, we also have that $t\widehat{\mathbf{\Delta}} \in \mathbb{C}_J \cap \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_2 \leq 1\}$, and it follows from (28) that

$$\mathcal{F}\left(t\widehat{\mathbf{\Delta}}\right) \geq \kappa_{\mathcal{L}} \left\|t\widehat{\mathbf{\Delta}}\right\|_2 \left(\left\|t\widehat{\mathbf{\Delta}}\right\|_2 - \frac{(J+1)\sqrt{2}\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J\, \kappa_{\mathcal{L}}} \right) > 0 ,$$

because $\left\|t\widehat{\mathbf{\Delta}}\right\|_2 = 1$ and $\frac{(J+1)\sqrt{2}\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J\, \kappa_{\mathcal{L}}} < 1$ by our assumption. This is a contradiction! Therefore, it must be the case that $\left\|\widehat{\mathbf{\Delta}}\right\|_2 \leq 1$.

It follows from Lemma 4 that $\widehat{\mathbf{\Delta}} \in \mathbb{C}_J$. Therefore, $\widehat{\mathbf{\Delta}} \in \mathbb{C}_J \cap \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_2 \leq 1\}$. Since $\mathcal{F}(\widehat{\mathbf{\Delta}}) \leq \mathcal{F}(\mathbf{0}) = 0$, it follows from Equation (28) that

$$0 \geq \mathcal{F}(\widehat{\mathbf{\Delta}}) \geq \kappa_{\mathcal{L}} \left\|\widehat{\mathbf{\Delta}}\right\|_2 \left(\left\|\widehat{\mathbf{\Delta}}\right\|_2 - \frac{(J+1)\sqrt{2}\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J\, \kappa_{\mathcal{L}}} \right) ,$$

which implies that

$$\left\|\widehat{\mathbf{\Delta}}\right\|_2 \leq \frac{(J+1)\sqrt{2}\,\lambda\, \max\{\sqrt{s-m}, \sqrt{m}\}}{J\, \kappa_{\mathcal{L}}} = \max\{\sqrt{s-m}, \sqrt{m}\} \cdot \frac{\lambda}{\kappa_{\mathcal{L}}} \cdot \sqrt{2}\left(1 + \frac{1}{J}\right)$$

We will now establish an upper bound on $\left\|\widehat{\mathbf{\Delta}}\right\|_1$. Since $\widehat{\mathbf{\Delta}} \in \mathbb{C}_J$ by Lemma 4, we have

$$\left\|\Pi_{S^\perp}\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 \leq \frac{J+1}{J-1} \left\|\Pi_S\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 + \frac{J+1}{J-1} \left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_1 ,$$

and thus,

$$\begin{aligned}
\left\|\widehat{\mathbf{\Delta}}\right\|_1 &= \left\|\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right\|_1 + \left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_1 = \left\|\Pi_{S^\perp}\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 + \left\|\Pi_S\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 + \left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_1 \\
&\le \left(1 + \frac{J+1}{J-1}\right)\left(\left\|\Pi_S\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 + \left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_1 \right) \\
&\le \frac{2J}{J-1}\left(\sqrt{s-m}\left\|\Pi_S\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_2 + \sqrt{m}\left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_2 \right) \\
&\le \frac{2J}{J-1}\cdot\max\{\sqrt{s-m}, \sqrt{m}\}\cdot\left(\left\|\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right\|_2 + \left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_2 \right) \\
&\le \frac{2\sqrt{2}\,J}{J-1}\cdot\max\{\sqrt{s-m}, \sqrt{m}\}\cdot\left\|\widehat{\mathbf{\Delta}}\right\|_2 \quad ,
\end{aligned}$$

where the second inequality follows from the fact that $\Pi_S\left(\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right)$ has at most $s-m$ non-zero entries because $|S| = s-m$ and $\widehat{\mathbf{\Delta}}_{\mathcal{A}} \in \mathbb{R}^m$, and the final inequality follows from $\left\|\widehat{\mathbf{\Delta}}_{\bar{\mathcal{A}}}\right\|_2 + \left\|\widehat{\mathbf{\Delta}}_{\mathcal{A}}\right\|_2 \le \sqrt{2}\left\|\widehat{\mathbf{\Delta}}\right\|_2$. Therefore, we have that

$$\left\|\widehat{\mathbf{\Delta}}\right\|_1 \le \max\{s-m, m\}\cdot\frac{\lambda}{\kappa_{\mathcal{L}}}\cdot 4\cdot\frac{J+1}{J-1} \quad ,$$

which is the desired result. □

F. Proof of Theorem 2

The proof of Theorem 2 uses the following lemma, which shows that the error belongs to the cone $\mathbb{C}$.

Lemma 4 (Characterization of the Error)**.** *Under Assumption 1, If* $\lambda \ge J\left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)\right\|_\infty$ *for* $J > 1$*, then the error* $\widehat{\mathbf{\Delta}}(\lambda) := \widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^* \in \mathbb{C}_J$*.*

Proof. To simplify our exposition, we will drop the reference to λ and write $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{\Delta}}$ to denote $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\mathbf{\Delta}}(\lambda)$, respectively. By Assumption 1 and our definition $\widehat{\boldsymbol{\beta}}$, we have that $\mathbf{C}\boldsymbol{\beta}^* = \mathbf{C}\widehat{\boldsymbol{\beta}} = \mathbf{b}$, and thus $\mathbf{C}\widehat{\mathbf{\Delta}} = \mathbf{0}$. Moreover, by definition of $\widehat{\boldsymbol{\beta}}$,

$$\mathcal{L}\left(\boldsymbol{\beta}^* + \widehat{\mathbf{\Delta}}; \mathbf{X}, \boldsymbol{Y}\right) + \lambda\left\|\boldsymbol{\beta}^* + \widehat{\mathbf{\Delta}}\right\|_1 \quad\le\quad \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right) + \lambda\left\|\boldsymbol{\beta}^*\right\|_1 \;,$$

which implies that

$$\begin{aligned}
0 &\ge \left\{\mathcal{L}\left(\boldsymbol{\beta}^* + \widehat{\mathbf{\Delta}}; \mathbf{X}, \boldsymbol{Y}\right) - \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)\right\} + \lambda\left\{\left\|\boldsymbol{\beta}^* + \widehat{\mathbf{\Delta}}\right\|_1 - \left\|\boldsymbol{\beta}^*\right\|_1\right\} \\
&\ge -\frac{\lambda}{J}\left\|\widehat{\mathbf{\Delta}}\right\|_1 + \lambda\left\{\left\|\boldsymbol{\beta}^* + \widehat{\mathbf{\Delta}}\right\|_1 - \left\|\boldsymbol{\beta}^*\right\|_1\right\} = \lambda\left\{\left\|\boldsymbol{\beta}^* + \widehat{\mathbf{\Delta}}\right\|_1 - \left\|\boldsymbol{\beta}^*\right\|_1 - \frac{1}{J}\left\|\widehat{\mathbf{\Delta}}\right\|_1\right\}
\end{aligned}$$

where the second inequality follows from the convexity of the loss function $\boldsymbol{\mathcal{L}}(\cdot;\boldsymbol{Y})$ because

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{\beta}^* + \widehat{\boldsymbol{\Delta}}; \mathbf{X}, \boldsymbol{Y}\right) - \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right) &\geq \nabla\mathcal{L}(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y})^T \widehat{\boldsymbol{\Delta}} \geq -\left|\nabla\mathcal{L}(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y})^T\widehat{\boldsymbol{\Delta}}\right| \\ &\geq -\|\nabla\mathcal{L}(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y})\|_\infty \left\|\widehat{\boldsymbol{\Delta}}\right\|_1 \geq -\frac{\lambda}{J}\left\|\widehat{\boldsymbol{\Delta}}\right\|_1 , \end{aligned}$$

where the third inequality follows from the Cauchy-Schwarz Inequality, and the last inequality follows from our hypothesis.

Therefore, we have that

$$\begin{aligned} 0 &\geq \left\|\boldsymbol{\beta}^* + \widehat{\boldsymbol{\Delta}}\right\|_1 - \|\boldsymbol{\beta}^*\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}\right\|_1 \\ &= \left\|\boldsymbol{\beta}^*_{\bar{\mathcal{A}}} + \widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right\|_1 - \|\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right\|_1 + \left\|\boldsymbol{\beta}^*_{\mathcal{A}} + \widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 - \|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 \\ &\geq \left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right\|_1 - \left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 \\ &= \left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right\|_1 - \frac{J+1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 \end{aligned}$$

where the second inequality follows from the application of Lemma 3 to $\boldsymbol{\beta}^*_{\bar{\mathcal{A}}}$, with $S = \mathsf{supp}(\beta^*_{\bar{\mathcal{A}}})$, and from the basic triangle inequality. Note that

$$\begin{aligned} &\left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right\|_1 - \frac{J+1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 \\ &\quad = \left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{J+1}{J}\left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{1}{J}\left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{J+1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 \\ &\quad = \frac{J-1}{J}\left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{J+1}{J}\left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{J+1}{J}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1 \\ &\quad = \frac{J-1}{J}\left\{\left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{J+1}{J-1}\left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 - \frac{J+1}{J-1}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1\right\} , \end{aligned}$$

and thus, $\left\|\Pi_{S^\perp}\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 \leq \frac{J+1}{J-1}\left\|\Pi_S\left(\widehat{\boldsymbol{\Delta}}_{\bar{\mathcal{A}}}\right)\right\|_1 + \frac{J+1}{J-1}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\right\|_1$, which implies that $\widehat{\boldsymbol{\Delta}} \in \mathbb{C}_J$. □

Here is the proof of Theorem 2.

Proof. Proposition 2 in Negahban *et al.* (2012) shows that, under Assumptions 3, there exist positive constants κ_1 and κ_2 that depend only on κ_ℓ, κ_u such that for all $\boldsymbol{\Delta} \in \mathbb{R}^p$ with $\|\boldsymbol{\Delta}\|_2 \leq 1$,

$$\mathcal{L}\left(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \mathbf{X}, \boldsymbol{Y}\right) - \mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right) - \nabla\mathcal{L}\left(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}\right)^T \boldsymbol{\Delta} \geq \kappa_1\|\boldsymbol{\Delta}\|_2\left\{\|\boldsymbol{\Delta}\|_2 - \kappa_2\sqrt{\frac{\log p}{n}}\|\boldsymbol{\Delta}\|_1\right\} ,$$

with probability at least $1 - d_1 e^{-d_2 n}$ for constants d_1 and d_2 that depend only on κ_ℓ and κ_u from Assumption 3. It follows that, with probability at least $1 - d_1 e^{-d_2 n}$, the loss function $\mathcal{L}(\cdot\ ; \mathbf{X}, \boldsymbol{Y})$ satisfies the restricted strong convexity on $\mathbb{C}_J \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$ because for each $\boldsymbol{\Delta} \in \mathbb{C}_J$,

$$
\begin{aligned}
\|\boldsymbol{\Delta}\|_1 &= \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_1 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 = \|\Pi_{S^\perp}(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 + \|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \\
&\leq \left(1 + \frac{J+1}{J-1}\right)\left(\ \|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_1 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1\ \right) \\
&\leq \frac{2J}{J-1}\left(\ \sqrt{s-m}\,\|\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})\|_2 + \sqrt{m}\,\|\boldsymbol{\Delta}_{\mathcal{A}}\|_2\ \right) \\
&\leq \frac{2J}{J-1} \cdot \max\{\ \sqrt{s-m}, \sqrt{m}\ \} \cdot \left(\ \|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_2 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_2\ \right) \\
&\leq \frac{2\sqrt{2}\,J}{J-1} \cdot \max\{\ \sqrt{s-m}, \sqrt{m}\ \} \cdot \|\boldsymbol{\Delta}\|_2 \quad ,
\end{aligned}
$$

where the second inequality follows from the fact that $\Pi_S(\boldsymbol{\Delta}_{\bar{\mathcal{A}}})$ has at most $s - m$ non-zero entries because $|S| = s - m$ and $\boldsymbol{\Delta}_{\mathcal{A}} \in \mathbb{R}^m$, and the final inequality follows from $\|\boldsymbol{\Delta}_{\bar{\mathcal{A}}}\|_2 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_2 \leq \sqrt{2}\,\|\boldsymbol{\Delta}\|_2$. Therefore, we have that on $\mathbb{C}_J \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$,

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \mathbf{X}, \boldsymbol{Y}) - \mathcal{L}(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y}) - \nabla\mathcal{L}(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{Y})^T \boldsymbol{\Delta} \\
&\geq\ \kappa_1 \left\{1 - \kappa_2 \sqrt{\frac{\log p}{n}} \cdot \frac{2\sqrt{2}\,J}{J-1} \cdot \max\{\ \sqrt{s-m}, \sqrt{m}\ \}\right\} \|\boldsymbol{\Delta}\|_2^2 \ ,
\end{aligned}
$$

If we set $J = 2$, then with high probability, $\mathcal{L}(\cdot\ ; \mathbf{X}, \boldsymbol{Y})$ satisfies the restricted strong convexity on $\mathbb{C}_2 \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leq 1\}$ with the parameter

$$
\kappa_1 \left\{1 - \kappa_2 \sqrt{\frac{\log p}{n}} \cdot 4\sqrt{2} \cdot \max\{\ \sqrt{s-m}, \sqrt{m}\ \}\right\} \geq \kappa_1 \left\{1 - \frac{4\sqrt{2}\kappa_2}{c_0}\right\} \ ,
$$

where the inequality follows from our assumption that $n > c_0 \max\{s - m, m\} \cdot \log p$. By choosing $c_0 = 8\kappa_2\sqrt{2}$, we can ensure that the loss function satisfies the restricted strong convexity with a parameter at least $\kappa_1/2$.

It remains to show that λ satisfies the requirement of Theorem with high probability. Note that, by our assumption,

$$
\lambda = c_1 \sqrt{\frac{\log p}{n}} < c_1 \sqrt{\frac{\log p}{c_0 \max\{s-m, m\} \cdot \log p}} = c_1 \sqrt{\frac{1}{8\kappa_2\sqrt{2}\max\{s-m, m\}}} \ ,
$$

where the last equality follows from our choice of $c_0 = 8\kappa_2\sqrt{2}$. For $J = 2$, to ensure that $\lambda < \frac{J\,\kappa_{\mathcal{L}}}{(J+1)\sqrt{2}\,\max\{\sqrt{s-m},\sqrt{m}\}}$, it suffice to choose and c_1 so that

$$\frac{c_1}{\sqrt{8\kappa_2\sqrt{2}}} < \frac{\kappa_1}{3\sqrt{2}} \qquad \Leftrightarrow \qquad c_1 < \frac{2^{5/4}\kappa_1\sqrt{\kappa_2}}{3}$$

So, we can pick, for example, $c_1 = \kappa_1\sqrt{\kappa_2/2}$. Now, with this choice of c_1, we will show that $J\left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y}\right)\right\|_\infty \leq \lambda$ with high probability. Lemma 6 in Negahban *et al.* (2012) shows that, under Assumptions 3 and 4, there are positive constants f_1 and f_2, depending on κ_ℓ, κ_u, and M_ψ such that for any n,

$$\Pr\left\{\left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y}\right)\right\|_\infty > f_1\sqrt{\frac{\log p}{n}}\right\} \leq \frac{f_2}{n} ,$$

which implies that for $J = 2$,

$$\begin{aligned}
\Pr\left\{2\left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y}\right)\right\|_\infty > \lambda\right\} &= \Pr\left\{\left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y}\right)\right\|_\infty > \frac{c_1}{2}\sqrt{\frac{\log p}{n}}\right\} \\
&= \Pr\left\{\left\|\nabla\mathcal{L}\left(\boldsymbol{\beta}^*;\mathbf{X},\boldsymbol{Y}\right)\right\|_\infty > f_1\sqrt{\frac{\log p}{4nf_1^2/c_1^2}}\right\} \\
&\leq \frac{f_2}{4nf_1^2/c_1^2} = \frac{f_2c_1^2/(4f_1^2)}{n} .
\end{aligned}$$

Setting $c_3 = f_2c_1^2/(4f_1^2)$, with probability at least $1-(c_3/n)$, λ satisfies the requirement of Theorem 1 and the loss function satisfies the restricted strong convexity, and thus, using $J = 2$ and $\kappa_{\mathcal{L}} \geq \kappa_1/2$, we have that

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_2 &\leq \sqrt{\max\{\,s-m\ ,\ m\,\}}\cdot\sqrt{2}\left(1+\frac{1}{J}\right)\cdot\frac{\lambda}{\kappa_{\mathcal{L}}} \\
&\leq \sqrt{\max\{\,s-m\ ,\ m\,\}}\cdot\frac{3}{\sqrt{2}}\cdot\frac{c_1}{\kappa_1/2}\cdot\sqrt{\frac{\log p}{n}} ,
\end{aligned}$$

and

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\right\|_1 &\leq \max\{\,s-m\ ,\ m\,\}\cdot 4\left(1+\frac{2}{J-1}\right)\cdot\frac{\lambda}{\kappa_{\mathcal{L}}} \\
&\leq \max\{\,s-m\ ,\ m\,\}\cdot\frac{12c_1}{\kappa_1/2}\cdot\sqrt{\frac{\log p}{n}} ,
\end{aligned}$$

and setting $c_2 = \frac{12c_1}{\kappa_1/2}$ gives the desired result. □

References

Besbes, O., Phillips, R., and Zeevi, A. (2010). Testing the validity of a demand model: an operations perspective. *Manufacturing & Service Operations Management* **12**, 1, 162–183.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Annals of Statistics* **35**, 6, 2313–2351.

Efron, B., Hastie, T., Johnston, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics* **32**, 2, 407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 456, 1348–1360.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association* **107**, 498, 592–606.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 2, 109–135.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1, 302–332.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics* **37**, 5A, 2083–2108.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* **7**, 1, 77–91.

Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York: Wiley.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis* **52**, 1, 374–393.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**, 4, 538–557.

Park, M. and Hastie, T. (2007). An L1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* **69**, 4, 659–677.

Radchenko, P. and James, G. M. (2008). Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association* **103**, 483, 1304–1315.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics* **4**, 1055–1096.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 1, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., and Zhu, J. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* **67**, 1, 91–108.

Tibshirani, R. and Taylor, J. (2011). The solution path of the generalized lasso. *Annals of Statistics* **39**, 3, 1335–1371.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2(1)**, 224–244.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics* **37**, 6A, 3468–3497.

Zhou, H. and Lange, K. (2013). A path algorithm for constrained estimation. *Journal of Computational and Graphical Statistics* **22**, 2, 261–283.

Zhou, H. and Wu, Y. (2013). A generic path algorithm for regularized statistical estimation. *Journal of the American Statistical Association (in press)* .

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 476, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 2, 301–320.