Convex and Non-Convex Approaches for Statistical Inference with Class-Conditional Noisy Labels

Hyebin Song HPS5320@PSU.EDU
Department of Statistics
The Pennsylvania State University
State College, PA 16801, USA

Ran Dai RAN.DAI@UNMC.EDU
Department of Biostatistics
University of Nebraska Medical Center
Omaha, NE 68198, USA

Garvesh Raskutti RASKUTTI@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Rina Foygel Barber RINA@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Abstract

We study the problem of estimation and testing in logistic regression with class-conditional noise in the observed labels, which has an important implication in the Positive-Unlabeled (PU) learning setting. With the key observation that the label noise problem belongs to a special sub-class of generalized linear models (GLM), we discuss convex and non-convex approaches that address this problem. A non-convex approach based on the maximum likelihood estimation produces an estimator with several optimal properties, but a convex approach has an obvious advantage in optimization. We demonstrate that in the low-dimensional setting, both estimators are consistent and asymptotically normal, where the asymptotic variance of the non-convex estimator is smaller than the convex counterpart. We also quantify the efficiency gap which provides insight into when the two methods are comparable. In the high-dimensional setting, we show that both estimation procedures achieve ℓ_2-consistency at the minimax optimal $\sqrt{s \log p/n}$ rates under mild conditions. Finally, we propose an inference procedure using a de-biasing approach. We validate our theoretical findings through simulations and a real-data example.

Keywords: generalized linear model, non-convexity, class-conditional label noise, PU-learning, regularization

1. Introduction

Label noise is a common phenomenon in a number of classification applications. For example, label noise occurs when humans are involved in labeling due to inattention or subjectivity (Ipeirotis et al., 2010; Smyth et al., 1995). Label noise can also come from bad data entry (Sculley and Cormack, 2008) or is sometimes intentionally introduced to protect

the privacy of a respondent (van den Hout and van der Heijden, 2002). Consequently, it is important to investigate how to carry out valid statistical inference in the presence of label noise.

An important example of the label noise problem includes the Positive-Unlabeled (PU) learning problem, where labeled samples are known to be positive, but unlabeled samples may be either positive or negative. Positive-Unlabeled learning arises in many applications, where obtaining negative responses is more expensive or intractable. One concrete example arises from deep mutational scanning (DMS) data sets in biochemistry (Fowler and Fields, 2014), where a data set consists of functional (positive) variants of a protein, together with unknown functionality (unlabeled) variants from an initial library. Numerous other applications of PU-learning arise (see e.g., Liu et al., 2003; Yang et al., 2014; Elkan and Noto, 2008).

This article addresses the estimation and testing problems of a binary logistic regression model where noise is present in responses. We assume a standard logistic model between the true binary responses and the known features, and a contamination process of the true labels. In particular, we assume that labels are corrupted with asymmetric probabilities based on their values, but those probabilities are not affected by the features. The goal is to estimate, and perform inferences on, the parameter in the logistic model, which parametrizes the relationship between the features and the true labels.

1.1 Related Work

There is a substantial literature on the subject of learning with label noise data. Since the random classification noise model was first proposed in Angluin and Laird (1988), extensive studies have been conducted to develop algorithms for building a classifier that effectively separates true positive and negative samples from data with label noise, and to establish theoretical guarantees for the proposed classifiers (Natarajan et al., 2018; Li and Bradic, 2018; see also Frénay and Verleysen, 2014 for a comprehensive survey).

Parameter estimation problems using probabilistic approaches have also been thoroughly studied by a number of authors, where the likelihood-based method with a latent variable model has been the primary approach. Both settings where the noise rates are known (Magder and Hughes, 1997; Hausman et al., 1998) and unknown (Pepe, 1992; Bollinger and David, 1997; Lyles et al., 2011) have been considered. In the latter case, an additional validation set, consisting of both true and noisy labels, is assumed to be available in addition to the main data set (Bollinger and David, 1997; Lyles et al., 2011; Pepe, 1992), or a semi-parametric approach was used to model the function containing noise rates nonparametrically (Hausman et al., 1998). Additionally, a general treatment of the noisy response problem as a variable with measurement errors is found in Chapter 13 of Carroll et al. (2006). In particular Carroll et al. (2006) uses a quasi-likelihood method which in our case is equivalent to the likelihood approach we propose later.

On the other hand, Ward et al. (2009) considered modeling of presence and absence of species, assuming that the prevalence of the positive examples in the unlabeled set was known a priori. Treating an indicator of true positive and absence of species as a latent variable, the latent variable model was fitted via the EM algorithm. The non-identification problem of the prevalence of the positives in the unlabeled set without any parametric

model assumption has also been pointed out in the same paper (Ward et al., 2009), and in the follow-up paper about the estimation of this prevalence parameter (Hastie and Fithian, 2013).

In all these aforementioned works in the parametric framework, either convergence to a local optimum was established without theoretical guarantees for the obtained estimators being provided, or the maximum likelihood estimator was considered in the theoretical analysis without discussion of the feasibility of obtaining such global optimum in a non-convex problem, all in low-dimensional settings. In contrast, one of the contributions of our paper is that we demonstrate achieving the global maximum is possible with high probability.

In high dimensions, Song and Raskutti (2018) studied the estimation problem in PU-learning and a case-control scheme where the prevalence of positives in an unlabeled set was assumed to be known a priori. They proposed an estimation based on the ℓ_1 penalized likelihood and devised an algorithm for which the estimator converges to a stationary point of the objective function, where the feasible space was constrained to be an intersection of ℓ_1 and ℓ_2 balls. They provided a theoretical mean-squared error guarantee for any stationary point of the objective. In this work, we consider a more general non-convex noisy labels problem which includes the PU problem in Song and Raskutti (2018) as a special case. Compared to the results in Song and Raskutti (2018), where the mean-squared error guarantees can only apply to the PU problem within a case-control scheme, our results can be applied to provide mean-squared error guarantees for noisy labels models in both prospective and case-control schemes, while removing the ℓ_1 constraint in their optimization problem. We also study in this paper an estimator based on a convex objective that can serve as a good starting point for the initialization of the non-convex method. Also, compared to Song and Raskutti (2018), where only an estimation problem was studied, both estimation and testing problems have been addressed in this paper in both low and high dimensions.

1.2 Our Contributions

In this paper, we study the parametric estimation and testing problem given observations where labels are observed with noise. One of the consequences of the label noise is that the maximum likelihood objective yields a non-convex minimization problem (Magder and Hughes, 1997; Bootkrajang and Kabán, 2012; Song and Raskutti, 2018). On the other hand, the surrogate loss based on an unbiased estimate of the original loss function leads to a convex minimization problem (Chaganty and Liang, 2014; Natarajan et al., 2018; Du Plessis et al., 2015). We propose and compare these two approaches in the classical regime and the high-dimensional regime, where the number of features p is fixed or grows with n, potentially at a faster speed. In this paper, we make the following contributions:

- Theoretical guarantees for parameter estimates for both non-convex likelihood-based and convex surrogate approaches in the classical regime (Proposition 2 and 5). Our guarantee is for *any* local minimum, by establishing that the likelihood function has actually at most one stationary point with high probability (Proposition 6). In contrast, prior work either proves convergence to a local minimum or proves theory for the global minimizer without any guarantee of finding this point.

- Quantification of the efficiency gap of the two estimators based on the conditions of the design matrix $\mathbf{X}$, which provides an insight into the performance of the convex versus non-convex estimators (Corollary 3).

- Mean-squared error guarantees and valid testing procedures in high dimensions, for the two estimators based on non-convex and convex approaches (Theorem 11 and 13). The error bounds match with the optimal $s \log p/n$ rates known as minimax optimal in the sparse regression literature (Raskutti et al., 2011). The testing procedure in high dimensions is based on de-biasing a penalized estimator and to the best of our knowledge, the first such theoretical analysis of testing procedures.

- A simulation study and a real data analysis to empirically support our theoretical findings. Our simulation study also indicates a potential advantage of using the convex surrogate of the likelihood in a very sparse regime, in contrast with the classical regime where the likelihood-based approach is provably optimal.

Now we outline the remainder of the paper. We begin by discussing the set-up of the work in Section 2. In Section 3, we discuss how our noisy logistic regression model can be represented as a generalized linear model, and introduce the convex and non-convex approaches for parameter estimation. We establish point estimation guarantees and hypothesis testing in both low dimensions (Section 4) and high dimensions (Section 5). In Section 6 and 7, we apply convex and non-convex methods to synthetic and real data and compare the performance of the two estimators. Finally, we conclude the paper with remarks in Section 8.

2. Problem Setup

First we define the problem and introduce the major notation. We assume access to samples $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ where $(\mathbf{z}_i)_{i=1}^n$ are observed labels and $\mathbf{x}_i \in \mathbb{R}^p$ is a p-dimensional feature vector such that $\mathbf{x}_i = [1, \mathbf{x}_{2i}, \ldots, \mathbf{x}_{pi}]$. Each observed label $\mathbf{z}_i$ is a corrupted version of a latent binary outcome $\mathbf{y}_i$, where $\mathbf{y}_i \sim p_{\beta_0}(y_i|\mathbf{x}_i)$ is a true response with p.d.f is given by a logistic model,

$$p_\beta(y|\mathbf{x}) = \exp(y\mathbf{x}^\top \beta - \log(1 + \exp(\mathbf{x}^\top \beta))), \tag{1}$$

and $\mathbf{z}_i$ is generated by flipping the value of $\mathbf{y}_i$ randomly based on known noise rates ρ_0 and ρ_1, with

$$\rho_0 := \mathbb{P}(\mathbf{z} = 1|\mathbf{y} = 0) \quad \text{and} \quad \rho_1 := \mathbb{P}(\mathbf{z} = 0|\mathbf{y} = 1),$$

for $\rho_0 + \rho_1 < 1$. We assume that $\mathbf{z}_i$ and $\mathbf{x}_i$ are conditionally independent given the true response $\mathbf{y}_i$. The goal is to estimate and perform inference on β_0.

We note that in the case of the data sets from deep mutational scanning (DMS) experiments, which we discuss as a concrete example of the noisy labels problem in Section 7, the conditional independence assumption between $\mathbf{z}$ and $\mathbf{x}$ given $\mathbf{y}$ is satisfied and the noise rates are known from the previous experiments. In other applications, however, the conditional independence or the known noise rates assumptions can be limiting. The noise rates

(a)

	$x=0$	$x=1$
$\mathbb{P}(\mathbf{y}=0\|\mathbf{x}=x)$	0.5	0.25
$\mathbb{P}(\mathbf{y}=1\|\mathbf{x}=x)$	0.5	0.75

(b)

	$x=0$	$x=1$
$\mathbb{P}(\mathbf{y}=0\|\mathbf{x}=x)$	0.5	0.75
$\mathbb{P}(\mathbf{y}=1\|\mathbf{x}=x)$	0.5	0.25

→ (c)

	$x=0$	$x=1$
$\mathbb{P}(\mathbf{z}=0\|\mathbf{x}=x)$	0.5	0.5
$\mathbb{P}(\mathbf{z}=1\|\mathbf{x}=x)$	0.5	0.5

Table 1: A simple illustration that two different $(\mathbf{x}, \mathbf{y})$ distributions (table a,b) can result in the same observed $\mathbf{x}$ and $\mathbf{z}$ distribution (table c) if the noise rates are allowed to be dependent on both $\mathbf{x}$ and $\mathbf{y}$. In this example, the observed label is contaminated only when $\mathbf{x}=1$ ($\rho_y(0)=0, \forall y$). The noise rates are ($\rho_0(1)=0, \rho_0(1)=\frac{1}{3}$) for (a) and ($\rho_0(1)=\frac{1}{3}, \rho_1(1)=0$) for (b).

may depend on both the true class labels $\mathbf{y}_i$ and the features $\mathbf{x}_i$ (violation of conditional independence assumption), or the noise rates may not be known a priori.

Unfortunately, in the case where the conditional independence assumption is violated, not much can be said about the results of the estimation unless the functional dependence of the noise rates on $\mathbf{x}$ is known or can be estimated from external sources, e.g., using a validation set containing both $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)_{i=1}^{nV}$ (Neuhaus, 1999; Lyles et al., 2011). Otherwise, we cannot identify the true mean function because different pairs $(\mathbb{E}[\mathbf{y}|\mathbf{x}=x], \rho_y(x) := \mathbb{P}(\mathbf{z} \neq y|\mathbf{y}=y, \mathbf{x}=x))$ can lead to the same observed mean function $\mathbb{E}[\mathbf{z}|\mathbf{x}=x]$, resulting in a lack of identifiability of the true mean function $\mathbb{E}[\mathbf{y}|\mathbf{x}=x]$ and contamination mechanism $\rho_y(x)$ (Table 1).

Similarly, in the case where the noise rates are unknown but the noise is conditionally independent of the features given the class y, we may consider jointly estimating (β, ρ_0, ρ_1) where (ρ_0, ρ_1) are unknown nuisance parameters. However, this parametric approach can be problematic unless there are additional sources of information available for the estimation of (ρ_0, ρ_1). Although (ρ_0, ρ_1) are identifiable under the logistic model assumption (1), (ρ_0, ρ_1) are not identifiable without (1) (Hausman et al., 1998; Magder and Hughes, 1997; Ward et al., 2009). The identifiability of the noise parameters depends entirely on the assumed parametric form, and slight deviations from the assumed parametric model can produce very different estimation results for (ρ_0, ρ_1) (e.g., Hastie and Fithian, 2013). Hence we focus on the setting where the conditional independence assumption is satisfied and the noise rates are known.

The relationship between the conditional mean of $\mathbf{y}$ and the conditional mean of $\mathbf{z}$ can be obtained under the conditional independence assumption. By the factorization theorem, we have

$$\begin{aligned}\mathbb{E}[\mathbf{z}|\mathbf{x}] &= \mathbb{P}(\mathbf{z}=1|\mathbf{y}=1)\mathbb{E}[\mathbf{y}|\mathbf{x}] + \mathbb{P}(\mathbf{z}=1|\mathbf{y}=0)(1-\mathbb{E}[\mathbf{y}|\mathbf{x}]) \\ &= (1-\rho_1)\mathbb{E}[\mathbf{y}|\mathbf{x}] + \rho_0(1-\mathbb{E}[\mathbf{y}|\mathbf{x}]) \\ &= (1-\rho_1-\rho_0)\mathbb{E}[\mathbf{y}|\mathbf{x}] + \rho_0. \end{aligned} \tag{2}$$

For the remainder of the paper, we let $\mathbb{P}_\beta$ be the distribution of the data when $\mathbf{y}|\mathbf{x} \sim p_\beta(\cdot|\mathbf{x})$ in (1). We will sometimes write $\mathbb{P}(\cdot)$ for the probability distribution evaluated at the true parameter β_0, i.e., $\mathbb{P}(\cdot) = \mathbb{P}_{\beta_0}(\cdot)$, where β_0 is the unique minimizer of the population loss function, i.e., $\beta_0 := \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}[-\log p_\beta(\mathbf{y}|\mathbf{x})]$.

Connection to Positive-Unlabeled Learning

The set-up in the previous section has an important implication in Positive-Unlabeled (PU) learning. In PU learning, we learn a model with two sets of samples, where the first set consists of *labeled and positive* subjects and the second set consists of *unlabeled* subjects whose associated responses are unknown.

Two schemes are considered for PU-learning: the first scheme is a single training set scheme (Elkan and Noto, 2008) whose complete observations $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)_{i=1}^n$ are from a single distribution and only $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ are recorded. The second scheme is where observations in the positive and unlabeled set are drawn separately, with the unlabeled set drawn from the general population (Ward et al., 2009; Song and Raskutti, 2018). A subtle but important difference between the two schemes is that a sample from the first scheme has the same distribution as the joint distribution of the population but a sample from the second scheme does not. In the second scheme, positive subjects are over-represented in the data set, since the distribution of the unlabeled sample is the same as the population distribution and the labeled set consists of only positive subjects. Therefore, a case-control sampling model (McCullagh and Nelder, 1989) is necessary in the second scheme, where different inclusion probabilities are allowed based on the value of the true responses.

We demonstrate how both PU schemes fit into the set-up of our label noise problem and also show how the error rates ρ_1 and ρ_0 are related with the number of labeled (n_ℓ) and unlabeled samples (n_u), and the proportion of positives in the unlabeled set $\pi := \mathbb{P}(\mathbf{y} = 1|\mathbf{z} = 0)$. We assume a parametric logistic model between $(\mathbf{x}, \mathbf{y})$ as in (1). In both schemes, flipping probabilities from $\mathbf{y}$ to $\mathbf{z}$ do not depend on $\mathbf{x}$. Also, $\rho_0 = \mathbb{P}(\mathbf{z} = 1|\mathbf{y} = 0) = 0$ since all labeled elements ($\mathbf{z} = 1$) are positive ($\mathbf{y} = 1$) by the set-up of the PU problem. On the other hand, by Bayes' theorem we have

$$\begin{aligned}
\rho_1 &= \frac{\mathbb{P}(\mathbf{y} = 1|\mathbf{z} = 0)\mathbb{P}(\mathbf{z} = 0)}{\mathbb{P}(\mathbf{y} = 1|\mathbf{z} = 1)\mathbb{P}(\mathbf{z} = 1) + \mathbb{P}(\mathbf{y} = 1|\mathbf{z} = 0)\mathbb{P}(\mathbf{z} = 0)} \\
&= \frac{\pi\mathbb{P}(\mathbf{z} = 0)}{\mathbb{P}(\mathbf{z} = 1) + \pi\mathbb{P}(\mathbf{z} = 0)} \\
&\approx \frac{\pi n_u}{n_\ell + \pi n_u},
\end{aligned}$$

where we use the definition $\pi := \mathbb{P}(\mathbf{y} = 1|\mathbf{z} = 0)$ and $\mathbb{P}(\mathbf{z} = 0)/\mathbb{P}(\mathbf{z} = 1) \approx n_u/n_\ell$. Thus, the knowledge of π practically amounts to knowing error rates (ρ_0, ρ_1) in PU-learning.

In the case-control sampling model, only *selected* subjects $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_i = 1)_{i=1}^n$ are available in the data set where $\mathbf{s}_i \in \{0, 1\}$ represents whether the ith subject is selected or not. It is a well-known result (e.g., McCullagh and Nelder, 1989) that case-control probabilities $\mathbb{P}(\mathbf{y} = 1|\mathbf{x}, \mathbf{s} = 1)$ differ from $\mathbb{P}(\mathbf{y} = 1|\mathbf{x})$ by the intercept, whose adjustment term is given by the log ratio of the different selection probabilities. More concretely, $p_\beta(y|\mathbf{x}, \mathbf{s} = 1)$ can

be written as

$$p_\beta(y|\mathbf{x}, \mathbf{s} = 1) = \exp\{y(\mathbf{x}^\top \beta^\gamma) - \log(1 + \exp(\mathbf{x}^\top \beta^\gamma))\},$$

for $\beta^\gamma \in \mathbb{R}^p$ such that $\beta_1^\gamma = \beta_1 + \gamma$ and $\beta_j^\gamma = \beta_j$, $\forall j \geq 2$, and where $\gamma := \log(\mathbb{P}(\mathbf{s} = 1|\mathbf{y} = 1)/\mathbb{P}(\mathbf{s} = 1|\mathbf{y} = 0))$ is the log ratio of the different selection probabilities. The log ratio γ can also be expressed as functions of n_ℓ, n_u, and π. Specifically, $\gamma = \log(1 + n_\ell/\pi n_u)$ was derived in Ward et al. (2009).

We note that in both PU schemes the conditional distribution of $\mathbf{y}$ follows a logistic model, with the parameter β_0 in the first scheme and β_0^γ in the second scheme. Since our target of interest is the coefficients of the model and $\beta_{0j} = \beta_{0j}^\gamma, \forall j \geq 2$, from this point on we will treat both sampling models the same. Specifically, we will omit conditioning on $\mathbf{s}$ and dependence of γ in β_0^γ, and we assume $y|\mathbf{x} \sim p_{\beta_0}(y|\mathbf{x}) = \exp\{y(\mathbf{x}^\top \beta_0) - \log(1 + \exp(\mathbf{x}^\top \beta_0))\}$ in both PU schemes.

3. Convex and Non-Convex Approaches for Inference

In this section, we briefly review generalized linear models (McCullagh and Nelder, 1989) and discuss how all models discussed above can be fitted into the generalized linear model (GLM) framework. Then we introduce two approaches to estimate the true parameter β_0, i.e., the parameter from which the data is generated. The first approach is to use a negative log-likelihood loss function, which is a non-convex function of β. In the second approach, we discuss how we can construct a convex surrogate function.

3.1 Generalized Linear Models (GLMs)

Generalized linear models (McCullagh and Nelder, 1989) are an extension of linear models, where a response $\mathbf{z} \in \mathcal{Z}$ has a p.d.f of the form

$$p_\theta(z) = c(z) \exp(z\theta - A(\theta)), \tag{3}$$

for $\theta \in \mathbb{R}$ which can depend on $\mathbf{x}$, and $A(\theta) = \log \int_{\mathcal{Z}} c(z) \exp(z\theta) dz$. The mean and variance of $\mathbf{z}$ can be derived from (3):

$$\mathbb{E}_\theta(\mathbf{z}|\mathbf{x}) = \mu = A'(\theta) \quad \text{and} \quad \text{Var}_\theta(\mathbf{z}|\mathbf{x}) = A''(\theta) = \mathcal{V}(\mu), \tag{4}$$

where the variance function $\mathcal{V}$ is defined as $\mathcal{V} := A'' \circ (A')^{-1}$ so that it is a function of μ.

Another important component of the GLM is the link function g, which relates the linear predictor $\mathbf{x}^\top \beta$ to the mean of the response μ by $g(\mu) = \mathbf{x}^\top \beta$. By definition of g and $\mu = A'(\theta)$, $\theta = (g \circ A')^{-1}(\mathbf{x}^\top \beta)$, and we can rewrite (3) in terms of the linear predictor $\mathbf{x}^\top \beta$ and the link function g. Therefore, the assumed distribution and the link function are two defining components of the GLM. We define the following:

Definition 1 (GLM) *We say a sample $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ is from a* **(GLM)** *with parameters (A, g) if the p.d.f of $\mathbf{z} \in \mathcal{Z}$ has the form*

$$p_\beta(z|\mathbf{x}) = c(z) \exp(zh(\mathbf{x}^\top \beta) - A(h(\mathbf{x}^\top \beta))), \tag{5}$$

for some c only depending on z and $h := (A')^{-1} \circ g^{-1}$.

We require g to be strictly increasing so that responses are positively related with linear predictors. A GLM is called canonical if $g = (A')^{-1}$ which implies $h(\cdot) = I(\cdot)$, an identity function. Suppose a random variable $\mathbf{y}$ is from a canonical GLM $(A, (A')^{-1})$. Then we have

$$p_\beta(y|\mathbf{x}) = c(y)\exp(y\mathbf{x}^\top\beta - A(\mathbf{x}^\top\beta)).$$

For example, the logistic model (1) is an example of a canonical GLM.

As we will discuss shortly in more detail, the statistical models for noisy labels belong to a special class of non-canonical GLMs whose mean is linearly related to the mean A' of a canonical GLM. In this type of case, the link function g is determined by such linear relationship since the link function is the inverse of the mean, i.e., $\mathbb{E}_\beta[\mathbf{z}|\mathbf{x}] = g^{-1}(\mathbf{x}^\top\beta)$. More concretely, suppose we have the following linear relationship

$$\mathbb{E}_\beta[\mathbf{z}|\mathbf{x}] = aA'(\mathbf{x}^\top\beta) + b, \tag{6}$$

for some $a > 0$, and $b \geq 0$. Then g has to satisfy the equation $g(aA'(\mathbf{x}^\top\beta) + b) = \mathbf{x}^\top\beta$, i.e.,

$$g(t) = (A')^{-1}\left(\frac{t-b}{a}\right). \tag{7}$$

Conversely, if g is taken to be as in (7), the linear relationship (6) is satisfied. We refer to this sub-class of GLM, where the link function g follows the form in (7), as **(GLM-L)** with parameters (A, a, b).

3.2 Statistical Models for Noisy Labels and GLMs

Now we relate the statistical models for noisy labels with the GLM framework. Since we have $\mathbf{z} \in \{0, 1\}$,

$$\begin{aligned} p_\beta(z|\mathbf{x}) &= (\mathbb{E}_\beta(\mathbf{z}|\mathbf{x}))^z(1 - \mathbb{E}_\beta(\mathbf{z}|\mathbf{x}))^{1-z} \\ &= \exp\Big(z\theta - \log(1 + e^\theta)\Big) \end{aligned}$$

for $\theta = \log\left(\frac{\mathbb{E}_\beta(\mathbf{z}|\mathbf{x})}{1-\mathbb{E}_\beta(\mathbf{z}|\mathbf{x})}\right)$, and thus $p_\beta(z|\mathbf{x})$ belongs to a GLM with $A(t) = \log(1 + e^t)$. Also by (1) and (2), we have the representation

$$\begin{aligned} \mathbb{E}_\beta[\mathbf{z}|\mathbf{x}] &= (1 - \rho_1 - \rho_0)\mathbb{E}_\beta[\mathbf{y}|\mathbf{x}] + \rho_0 \\ &= (1 - \rho_1 - \rho_0)\frac{e^{\mathbf{x}^\top\beta}}{1 + e^{\mathbf{x}^\top\beta}} + \rho_0. \end{aligned} \tag{8}$$

From (8), we obtain the link function g_{LN} (label-noise) by solving $g_{LN}(\mathbb{E}_\beta[\mathbf{z}|\mathbf{x}]) = \mathbf{x}^\top\beta$ for $\mathbf{x}^\top\beta$:

$$g_{LN}(t) = \text{logit}\left(\frac{t - \rho_0}{1 - \rho_1 - \rho_0}\right). \tag{9}$$

Therefore $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ belongs to **(GLM-L)** with $(\log(1 + \exp(\cdot)), (1 - \rho_1 - \rho_0), \rho_0)$.

In the subsequent analysis the variances of a clean label $\mathbf{y}$ and a noisy response $\mathbf{z}$ will play an important role. First we define mean functions μ and μ_z as $\mu(t) := A'(t)$ and

$\mu_z(t) := A'(h_{LN}(t))$, for $A(\cdot) = \log(1 + \exp(\cdot))$ and $h_{LN} := (A')^{-1} \circ g_{LN}^{-1}$. In particular, we have $\mathbb{E}_\beta[\mathbf{z}|\mathbf{x}] = \mu_z(\mathbf{x}^\top\beta)$ and $\mathbb{E}_\beta[\mathbf{y}|\mathbf{x}] = \mu(\mathbf{x}^\top\beta)$. By the definition of $\mathcal{V}$ in (4), we have

$$\begin{aligned}\mathrm{Var}_\beta(\mathbf{z}|\mathbf{x}) &= A''(h_{LN}(\mathbf{x}^\top\beta)) \ = \mathcal{V}(\mu_z(\mathbf{x}^\top\beta)) \\ \mathrm{Var}_\beta(\mathbf{y}|\mathbf{x}) &= A''(\mathbf{x}^\top\beta) = \mathcal{V}(\mu(\mathbf{x}^\top\beta))\end{aligned}$$

where the last equality uses the fact that $(A')^{-1} \circ \mu = I$.

3.3 Non-Convex Approach Using a Negative Log-likelihood Loss

Given a sample $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ from **(GLM)** with (A, g), a natural approach for the estimation of β_0 is to take a likelihood-based approach due to the several optimality properties of a likelihood function. A negative log-likelihood loss can be obtained directly from (5) as

$$\mathcal{L}_n^\ell(\beta) := \frac{1}{n}\sum_{i=1}^n A(h(\mathbf{x}_i^\top\beta)) - \mathbf{z}_i h(\mathbf{x}_i^\top\beta) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i^\top\beta, \mathbf{z}_i), \tag{10}$$

where we define $\ell(\mathbf{x}^\top\beta, \mathbf{z}) := A(h(\mathbf{x}^\top\beta)) - \mathbf{z}h(\mathbf{x}^\top\beta)$. In general, the likelihood becomes a non-convex function of β unless $g = (A')^{-1}$ i.e., g is canonical and h is an identity function.

The first and second derivatives of the likelihood function are

$$\nabla\mathcal{L}_n^\ell(\beta) = \frac{1}{n}\sum_{i=1}^n \ell'(\mathbf{x}_i^\top\beta, \mathbf{z}_i)\mathbf{x}_i, \quad \nabla^2\mathcal{L}_n^\ell(\beta) = \frac{1}{n}\sum_{i=1}^n \ell''(\mathbf{x}_i^\top\beta, \mathbf{z}_i)\mathbf{x}_i\mathbf{x}_i^\top,$$

where we write

$$\ell'(t, z) = \big(A'(h(t)) - z\big)\, h'(t) \tag{11}$$

$$\begin{aligned}\ell''(t, z) &= A''(h(t))h'(t)^2 + (A'(h(t)) - z)h''(t) \\ &:= \rho_I(t) + \rho_R(t, z)\end{aligned} \tag{12}$$

for $\rho_I(t) := A''(h(t))h'(t)^2$ and $\rho_R(t, z) := (A'(h(t)) - z)h''(t)$. Although $\rho_I \geq 0$, the sign of ρ_R is arbitrary, and thus $\nabla^2\mathcal{L}_n^\ell(\beta)$ is not necessarily a positive semi-definite matrix.

3.4 Construction of a Convex Surrogate Loss

Next, we discuss an alternative approach involving a convex surrogate function when a sample is from a **(GLM-L)** model with parameters (A, a, b). Essentially, we construct an unbiased estimator of a convex loss function with the same minimizer, which is a well-known idea in stochastic optimization (Nemirovski et al., 2009) and has also been investigated in the latent variable model literature (Loh and Wainwright, 2012; Chaganty and Liang, 2014; Natarajan et al., 2018). More concretely, if the responses $(\mathbf{y}_i)_{i=1}^n$ from a canonical GLM are available, we can minimize a convex loss $\mathcal{L}_n^c(\beta)$ which we define as

$$\mathcal{L}_n^c(\beta) := \frac{1}{n}\sum_{i=1}^n A(\mathbf{x}_i^\top\beta) - \mathbf{y}_i(\mathbf{x}_i^\top\beta). \tag{13}$$

For example, we can take this convex approach if labels are not contaminated. Since $(\mathbf{y}_i)_{i=1}^n$ are not available, we construct a surrogate function $\mathcal{L}_n^s(\beta)$ by replacing $\mathbf{z}$ with a function output $T(\mathbf{z})$ while keeping $h(\cdot) = I(\cdot)$:

$$\mathcal{L}_n^s(\beta) := \frac{1}{n}\sum_{i=1}^n A(\mathbf{x}_i^\top \beta) - T(\mathbf{z}_i)(\mathbf{x}_i^\top \beta). \tag{14}$$

To obtain a consistent estimator, the function T needs to satisfy $\mathbb{E}_\beta[T(\mathbf{z})|\mathbf{x}] = A'(\mathbf{x}^\top\beta) = \mathbb{E}_\beta[\mathbf{y}|\mathbf{x}]$. Such a function T is available by the **(GLM-L)** model class assumption. Specifically, we let T be $T(t) := (t-b)/a$ so that $\mathbb{E}_\beta[T(\mathbf{z})|\mathbf{x}] = \mathbb{E}_\beta[(\mathbf{z}-b)/a] = A'(\mathbf{x}^\top\beta)$ by (6). For a future reference we define

$$\ell_s(\mathbf{x}^\top\beta, \mathbf{z}) := A(\mathbf{x}^\top\beta) - T(\mathbf{z})(\mathbf{x}^\top\beta) \tag{15}$$

so that $\mathcal{L}_n^s(\beta) = n^{-1}\sum_{i=1}^n \ell_s(\mathbf{x}_i^\top\beta, \mathbf{z}_i)$.

At any fixed parameter β, the surrogate loss (14) is an unbiased estimate of the loss (13). We note

$$\begin{aligned}
\mathbb{E}_{\beta_0}[\mathcal{L}_n^s(\beta)] &= \mathbb{E}_{\beta_0}\left[\frac{1}{n}\sum_{i=1}^n A(\mathbf{x}_i^\top\beta) - \mathbb{E}_{\beta_0}[T(\mathbf{z}_i)|\mathbf{x}_i](\mathbf{x}_i^\top\beta)\right] \\
&= \mathbb{E}_{\beta_0}\left[\frac{1}{n}\sum_{i=1}^n A(\mathbf{x}_i^\top\beta) - A'(\mathbf{x}_i^\top\beta_0)(\mathbf{x}_i^\top\beta)\right] \\
&= \mathbb{E}_{\beta_0}[\mathcal{L}_n^c(\beta)],
\end{aligned}$$

where we use the law of iterative expectation and $\mathbb{E}_{\beta_0}[\mathbf{y}|\mathbf{x}] = A'(\mathbf{x}^\top\beta_0)$.

3.5 Notation

Before proceeding, we pause to define some notation that will be useful in presenting our theoretical results. For $v \in \mathbb{R}^p$, we denote the ℓ_1, ℓ_2, and ℓ_∞ norms as $\|v\|_1 := \sum_{i=1}^p |v_i|$, $\|v\|_2 := \sqrt{v^\top v}$, and $\|v\|_\infty := \sup_{1\le j\le p}|v_j|$. Similarly, for a function f, we define $\|f\|_p := (\int |f(x)|^p dx)^{1/p}$ and $\|f\|_\infty := \sup_x |f(x)|$. In the case of matrix norm, for $A \in \mathbb{R}^{m\times n}$, we denote a Frobenius norm as $\|A\|_F := \sqrt{\sum_{i,j}|A_{ij}|^2}$, an operator norm as $\|A\|_2 := \sigma_{\max}(A)$, and an element-wise max norm as $\|A\|_{\max} := \max_{i,j}|A_{ij}|$. We define a condition number of A as $\kappa(A) := \sigma_{\max}(A)/\sigma_{\min}(A)$.

For a set S, we use $|S|$ to denote the cardinality of S. For $v \in \mathbb{R}^p$ and any subset $S \subseteq \{1,\dots,p\}$, $v_S \in \mathbb{R}^{|S|}$ denotes the sub-vector of the vector v by selecting the components with indices in S. Likewise for any matrix $A \in \mathbb{R}^{m\times n}$, $A_S \in \mathbb{R}^{m\times|S|}$ denotes a sub-matrix having columns in S. For matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{m\times n}$, we say $A \succeq B$ if $A - B$ is a positive semi-definite matrix and $A \succ B$ if $A - B$ is positive definite. Also we write $\mathcal{C}(A)$ to refer to a column space of A. Also we use $\mathbb{B}_q(r; v)$ to denote a ball with radius r in the ℓ_q norm centered at $v \in \mathbb{R}^p$. If $v = 0$, we simply use $\mathbb{B}_q(r)$ to denote the ball.

For functions f and g, we write $f(n) = O(g(n))$ if there exists a constant $C > 0$ such that $f(n) \le Cg(n)$, $\forall n$, and $f(n) \asymp g(n)$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$. Also

for a random variable X_n, we write $X_n = O_p(a_n)$ if X_n/a_n is bounded in probability and $X_n = o_p(a_n)$ if X_n/a_n converges to 0 in probability. Also for simplicity, we sometimes use $\mathbf{x}_1^n$ to refer to the collection of random variables $(\mathbf{x}_i)_{i=1}^n$. We write a.s. to denote 'almost surely', i.e., an event that occurs with probability 1. Also, for a sequence of events $(\mathcal{E}_n)_{n\geq 1}$, we say $\mathcal{E}_n$ holds with high probability (w.h.p) if $\mathbb{P}(\mathcal{E}_n) \xrightarrow{n} 1$.

4. Estimation and Testing in the Classical Regime

In this section, we discuss the statistical properties of two estimators from convex and non-convex approaches in the classical regime where the number of features p is fixed. In particular, we demonstrate that both approaches yield consistent estimators, but the estimator based on the non-convex approach has better efficiency than the convex counterpart in the large n limit. Also, we quantify the efficiency gap between the two approaches and discuss when two approaches can be comparable.

4.1 Consistency and Relative Asymptotic Efficiency

We define a global minimizer of $\mathcal{L}_n^\ell(\beta)$ and $\mathcal{L}_n^s(\beta)$ as

$$\widehat{\beta}_\ell \in \underset{\beta\in\mathbb{B}_2(r)}{\arg\min}\, \mathcal{L}_n^\ell(\beta) \quad \text{and} \quad \widehat{\beta}_s \in \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \mathcal{L}_n^s(\beta). \tag{16}$$

By definition of $\mathcal{L}_n^\ell$ and $\mathcal{L}_n^s$, $\widehat{\beta}_\ell$ is the solution of a non-convex optimization problem, whereas $\widehat{\beta}_s$ is based on the convex problem. In the case of the non-convex optimization problem, we limit the search space to a compact region $\mathbb{B}_2(r)$, where r is some large number such that $r \geq \|\beta_0\|_2$. The use of a compact search space ensures that the gradient of the non-convex loss function is uniformly bounded away from zero for values of β not near the true parameter.

Clearly, it is not obvious whether it is feasible to obtain $\widehat{\beta}_\ell$ in practice, since finding a global minimizer of a non-convex function is in general a challenging problem. However, obtaining a stationary point of $\mathcal{L}_n^\ell(\beta)$ is in fact enough when n is sufficiently large, as we will demonstrate in Proposition 6 that in the classical regime, with high probability, $\mathcal{L}_n^\ell(\beta)$ has a unique stationary point (i.e. the global minimizer).

In the following Proposition 2, we show that both estimators are consistent for β_0 and also quantify their asymptotic efficiency. We first state the following minimum eigenvalue condition, which is a standard assumption in the classical regime with the fixed design (e.g., Fahrmeir and Tutz, 2001; Shao, 2003).

A1. *There exist* $C_\lambda > 0$ *and* $C_X < \infty$ *such that* $\lambda_{\min}(n^{-1}\sum_{1\leq i\leq n}\mathbf{x}_i\mathbf{x}_i^\top) \geq C_\lambda$ *and* $\sup_{1\leq i\leq n}\|\mathbf{x}_i\|_\infty \leq C_X, \forall n$.

Proposition 2 *(Fixed design) Suppose a sample* $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ *is from a* **(GLM-L)** *with* $(\log(1+\exp(\cdot)), (1-\rho_1-\rho_0), \rho_0)$ *and* $\mathbf{z}_i \in \{0,1\}$. *Assume* **A1** *and the classical regime where the number of features* p *is fixed and the sample size* $n \to \infty$. *Then,*

$$\sqrt{n}\mathcal{I}_n^\ell(\beta_0)^{1/2}(\widehat{\beta}_\ell - \beta_0) \xrightarrow{d} \mathcal{N}(0, I_p)$$
$$\sqrt{n}\mathcal{I}_n^s(\beta_0)^{1/2}(\widehat{\beta}_s - \beta_0) \xrightarrow{d} \mathcal{N}(0, I_p),$$

for positive definite matrices $\mathcal{I}_n^\ell(\beta), \mathcal{I}_n^s(\beta)$ *defined as*

$$\mathcal{I}_n^\ell(\beta) := (1-\rho_1-\rho_0)^2 \cdot \frac{1}{n}\sum_{i=1}^n \frac{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))^2}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))}\mathbf{x}_i\mathbf{x}_i^\top,$$
$$\mathcal{I}_n^s(\beta) := (1-\rho_1-\rho_0)^2 \cdot \left(\frac{1}{n}\sum_{i=1}^n \mathcal{V}(\mu(\mathbf{x}_i^\top\beta))\mathbf{x}_i\mathbf{x}_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^n \mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{V}(\mu(\mathbf{x}_i^\top\beta))\mathbf{x}_i\mathbf{x}_i^\top\right).$$

The proof essentially uses classical likelihood and generalized estimating equations theory and is provided in the Appendix A.1. One point that deserves special attention is the similarity between $\mathcal{I}_n^\ell(\beta_0)$ and $\mathcal{I}_n^s(\beta_0)$ in Proposition 2. In particular, if $\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta)) \approx \mathcal{V}(\mu(\mathbf{x}_i^\top\beta))$ for all i, the two information matrices will turn out to be very similar.

The following Corollary shows that $\mathcal{I}_n^\ell(\beta_0) \succeq \mathcal{I}_n^s(\beta_0)$ and quantifies the discrepancy between the two information matrices. First we define two weight matrices $W_y(\beta)$ and $W_z(\beta)$ as

$$W_y(\beta) := \mathrm{diag}(\{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))\}_{i=1}^n) \quad \text{and} \quad W_z(\beta) := \mathrm{diag}(\{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))\}_{i=1}^n),$$

whose diagonal entries consist of the conditional variances of $\mathbf{y}_i$ and $\mathbf{z}_i$ given $\mathbf{x}_i$, respectively. We suppress the dependence on β if $\beta = \beta_0$ and let $W_y := W_y(\beta_0)$ and $W_z := W_z(\beta_0)$ for ease of notation. Also, we define the gap $\widehat{\delta}(\mathcal{M},\mathcal{N})$ between two vector subspaces $\mathcal{M},\mathcal{N}$ as (e.g., Kato, 2013)

$$\widehat{\delta}(\mathcal{M},\mathcal{N}) := \max\{\delta(\mathcal{M},\mathcal{N}), \delta(\mathcal{N},\mathcal{M})\}, \text{ for } \delta(\mathcal{M},\mathcal{N}) := \sup_{u\in\mathcal{M},\|u\|_2=1}\inf_{v\in\mathcal{N}}\|u-v\|_2. \quad (17)$$

The gap measures the distance between two subspaces, with $\widehat{\delta}(\mathcal{M},\mathcal{N}) = 0$ if and only if $\mathcal{M} = \mathcal{N}$. Now we present the following Corollary.

Corollary 3 *Assume the conditions as in Proposition 2. We have* $\mathcal{I}_n^\ell(\beta_0) \succeq \mathcal{I}_n^s(\beta_0)$ *and*

$$\|I_p - \mathcal{I}_n^\ell(\beta_0)^{-1/2}\mathcal{I}_n^s(\beta_0)\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2 \le c_n\widehat{\delta}^2(\mathcal{C}(W_z^{-1}W_y\mathbf{X}),\mathcal{C}(\mathbf{X})) \quad (18)$$

where $c_n := \kappa(\mathbf{X}^\top\mathbf{X}/n)\kappa(W_y^2)\kappa(W_z^2)$ *and* $c_n = O(1)$*. In particular,* $\widehat{\beta}_s$ *achieves asymptotic efficiency if* $\mathcal{C}(W_z^{-1}W_y\mathbf{X}) = \mathcal{C}(\mathbf{X})$*.*

The proof is provided in Appendix A.2. We note if $p = 1$, the relative ℓ_2 difference equals to

$$\|I_p - \mathcal{I}_n^\ell(\beta_0)^{-1/2}\mathcal{I}_n^s(\beta_0)\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2 = \left|1 - \frac{\mathcal{I}_n^\ell(\beta_0)^{-1}}{\mathcal{I}_n^s(\beta_0)^{-1}}\right| = 1 - \mathrm{ARE}(\widehat{\beta}_s,\widehat{\beta}_\ell;\beta_0)$$

where $\mathrm{ARE}(\widehat{\beta}_s,\widehat{\beta}_\ell;\beta_0)$ denotes the asymptotic relative efficiency of $\widehat{\beta}_s$ with respect to $\widehat{\beta}_\ell$. In general, we can find the direction u such that $\|u\|_2 = 1$ and

$$\|I_p - \mathcal{I}_n^\ell(\beta_0)^{-1/2}\mathcal{I}_n^s(\beta_0)\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2 \ge 1 - \mathrm{ARE}(u^\top\widehat{\beta}_s, u^\top\widehat{\beta}_\ell;\beta_0).$$

The bound (18) shows that the relative ℓ_2 difference between $\mathcal{I}_n^\ell(\beta_0)$ and $\mathcal{I}_n^s(\beta_0)$ depends on how dissimilar $W_z^{-1}W_y$ is from the identity matrix. We observe that $W_z^{-1}W_y$ is a diagonal matrix where the diagonal entries are ratios of the variances of $\mathbf{z}$ and $\mathbf{y}$, i.e.,

$$\begin{aligned}(W_z^{-1}W_y)_{ii} &= \frac{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta_0))}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta_0))} \\ &= \frac{\mu(\mathbf{x}_i^\top\beta_0)(1-\mu(\mathbf{x}_i^\top\beta_0))}{\{(1-\rho_1-\rho_0)\mu(\mathbf{x}_i^\top\beta_0)+\rho_0\}\{(1-\rho_1-\rho_0)(1-\mu(\mathbf{x}_i^\top\beta_0))+\rho_1\}},\end{aligned} \tag{19}$$

noting $\mathcal{V}(\mu) = \mu(1-\mu)$. In light of these observations, the inefficiency of a surrogate convex loss function can be understood as the result of sub-optimal weighting of the observations due to the mis-specification of the variance matrix for $\mathbf{z}$. In fact, in the special case of the intercept-only model, no covariate information is available for the optimal weighting of the observations. In this case, we have $W_y = w_1 I_n$, $W_z = w_2 I_n$ for some $w_1, w_2 > 0$, and thus $\mathcal{C}(W_z^{-1}W_y\mathbf{X}) = \mathcal{C}(\mathbf{X})$ and the inequality (18) is sharp.

We also note that the variance ratios are also functions of the noise rates. Each diagonal entry $(W_z^{-1}W_y)_{ii}$ is a point on the variance ratio curve

$$r(t) = \frac{\mu(t)(1-\mu(t))}{\{(1-\rho_1-\rho_0)\mu(t)+\rho_0\}\{(1-\rho_1-\rho_0)(1-\mu(t))+\rho_1\}}$$

at $t = \mathbf{x}_i^\top\beta_0$, where the curve $r(t)$ is a function of the noise rates ρ_1 and ρ_0. When there is no noise in the labels, i.e., $\rho_0 = \rho_1 = 0$, $r(t) \equiv 1$, all diagonal entries are 1. In this case, we again have $\mathcal{C}(W_z^{-1}W_y\mathbf{X}) = \mathcal{C}(\mathbf{X})$. With positive noise rates, higher noise rates tend to be associated with larger amounts of perturbation to the column space of $\mathbf{X}$, with the caveat that the locations of $\{\mathbf{x}_i^\top\beta_0\}_{i=1}^n$ also play a role in determining the amount of perturbation given the noise rates. We provide more discussion on the relationship between the amounts of noise in labels and the relative efficiency of the two estimators in Section 6.3.

So far, we have considered the fixed design setting. Now we present the result equivalent to Proposition 2 in the random design. We assume that rows in the random design matrix satisfy a sub-gaussian tail condition. We define this sub-gaussian tail condition as follows:

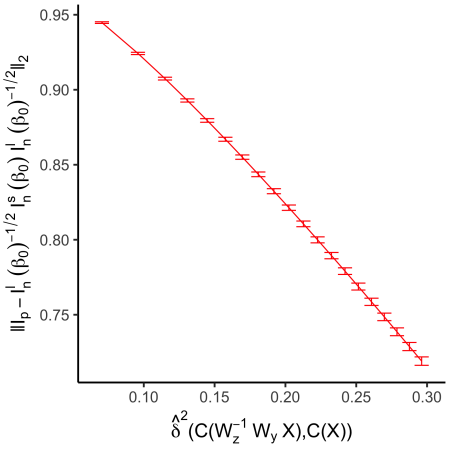


Figure 1: The plot of the relative ℓ_2 difference between $\mathcal{I}_n^\ell(\beta_0)$ and $\mathcal{I}_n^s(\beta_0)$ as a function of $gap^2 = \widehat{\delta}^2(\mathcal{C}(W_z^{-1}W_y\mathbf{X}), \mathcal{C}(\mathbf{X}))$. To generate design matrix $\mathbf{X}$ with various gap values, each $\mathbf{x}_i$ is sampled from an equal mixture of multivariate Gaussian distribution with different centers; see Section 6 for details. The results were averaged over 10000 repetitions at each center, and the bars denote one standard error.

Definition 4 (sub-gaussian tail condition) *We say a random vector $\mathbf{x} \in \mathbb{R}^p$ satisfies the sub-gaussian tail condition with parameter K if*

$$\sup_{u \in \mathbb{R}^p; \|u\|_2 = 1} \mathbb{E}[\exp(u^T \mathbf{x})^2 / K^2] \leq 2. \tag{20}$$

For example, a random vector $\mathbf{x} \in \mathbb{R}^p$ with $\mu_X = \|\mathbb{E}[\mathbf{x}]\|_2$ satisfies (20) with $K = c_1 \mu_X + c_2 \sigma_X$ for some absolute constants $c_1, c_2 > 0$ if the centered vector $\mathbf{x} - \mathbb{E}[\mathbf{x}]$ is sub-gaussian with parameter σ_X, i.e.,

$$\sup_{u \in \mathbb{R}^p; \|u\|_2 = 1} \mathbb{E}[\exp(t(u^T \mathbf{x} - \mathbb{E}[u^T \mathbf{x}]))] \leq \exp(t^2 \sigma_X^2 / 2), \forall t \in \mathbb{R}.$$

We replace **A1** with the following assumption:
A1'. *(Random design) For a random feature vector $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x}$ satisfies the sub-gaussian tail condition with parameter K_X for a positive constant $K_X < \infty$. Also, there exist $C_\lambda > 0$ and $C_X < \infty$ such that $\lambda_{\min}(\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top]) \geq C_\lambda$ and $\sup_{1 \leq i \leq n} \|\mathbf{x}_i\|_\infty \leq C_X, \forall n$.*

Proposition 5 *(Random design) Assume the conditions of Proposition 2 where* **A1** *is replaced with* **A1'**. *Then,*

$$\sqrt{n}(\widehat{\beta}_\ell - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^\ell(\beta_0)^{-1})$$
$$\sqrt{n}(\widehat{\beta}_s - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^s(\beta_0)^{-1}),$$

for $\mathcal{I}^\ell(\beta), \mathcal{I}^s(\beta)$ defined as

$$\mathcal{I}^\ell(\beta) := (1 - \rho_1 - \rho_0)^2 \mathbb{E}_\beta \left(\frac{\mathcal{V}(\mu(\mathbf{x}^\top \beta))^2}{\mathcal{V}(\mu_z(\mathbf{x}^\top \beta))} \mathbf{x}\mathbf{x}^\top \right),$$
$$\mathcal{I}^s(\beta) := (1 - \rho_1 - \rho_0)^2 \mathbb{E}_\beta \left(\mathcal{V}(\mu(\mathbf{x}^\top \beta)) \mathbf{x}\mathbf{x}^\top \right) \mathbb{E}_\beta \left(\mathcal{V}(\mu_z(\mathbf{x}^\top \beta)) \mathbf{x}\mathbf{x}^\top \right)^{-1} \mathbb{E}_\beta \left(\mathcal{V}(\mu(\mathbf{x}^\top \beta)) \mathbf{x}\mathbf{x}^\top \right).$$

Also, $\mathcal{I}^\ell(\beta_0) \succeq \mathcal{I}^s(\beta_0)$.

The result follows from classical M-estimation theory (see e.g., van der Vaart, 1998). $\mathcal{I}^\ell(\beta_0) \succeq \mathcal{I}^s(\beta_0)$ follows from Theorem 1 in Morton (1981).

The final result that we will present in this section is about the comparability between the global and local minimizer in the low-dimensional setting. So far, we have only considered the global minimizer of the empirical risk function $\mathcal{L}_n^\ell(\beta)$ which is the MLE. However, since $\mathcal{L}_n^\ell(\beta)$ is non-convex, obtaining the global minimizer $\widehat{\beta}_\ell$ is in general computationally intractable, and algorithms on the optimization of non-convex functions focus on finding a stationary point of the objective function.

The population risk, albeit non-convex, can be shown to be unimodal and also strongly convex around β_0 in some GLMs. Often, fast probability tail decay of $\mathbf{x}$ and $\mathbf{z}$, and boundedness of the derivatives of the loss function allow enough concentration of the empirical risk function around the population counterpart that the empirical risk function has a unique stationary point, which in fact is the global minimum (Mei et al., 2018). We make the following assumption **A2** about the smoothness of ℓ, for ℓ defined in (10). In Corollary 7, we

show that Assumption **A2** is satisfied for the label noise model, **(GLM)** with parameters $(\log(1+\exp(\cdot)), g_{LN})$.

A2. *ℓ'' is Lipschitz w.r.t its first argument, i.e., $|\ell''(a,t) - \ell''(a',t)| \leq L_\ell |a-a'|, \forall t$. Furthermore, there exists $C_\ell < \infty$ such that $\max\{\|\ell'\|_\infty, \|\rho_I\|_\infty, \|\rho_R\|_\infty\} \leq C_\ell$.*

Proposition 6 *Suppose a sample $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ is from a* **(GLM)** *with parameters (A, g) and assume* **A1'** *and* **A2**. *Moreover, assume $\mathbf{z}_i|\mathbf{x}_i$ satisfies the sub-gaussian tail condition with parameter K_Z for a positive constant $K_Z < \infty$, i.e.,*

$$\mathbb{E}[e^{t(\mathbf{z}_i - \mathbb{E}[\mathbf{z}_i|\mathbf{x}_i])}|\mathbf{x}_i] \leq e^{t^2 K_Z^2/2} \ \text{a.s.}, \ \forall t \in \mathbb{R}.$$

Then, for any given $\epsilon > 0$, there exists a unique stationary point of $\mathcal{L}_n^\ell(\beta)$ in $\mathbb{B}_2(r)$ with probability at least $1-\epsilon$, given a sufficiently large $n \geq C\log(1/\epsilon)p\log n$ where the constant C depends only on the model parameters in our assumptions. The unique stationary point is the global minimum of $\mathcal{L}_n^\ell(\beta)$.

Corollary 7 *Under Assumption* **A1'**, *the log-likelihood for the label noise GLM has a unique stationary point in $\mathbb{B}_2(r)$, which is the MLE, with high probability.*

Proofs of Proposition 6 and Corollary 7 are provided in Appendix A.3 and A.4.

5. Estimation and Testing in the High-Dimensional Regime

5.1 ℓ_1 and ℓ_2 Consistency

In many modern data sets, the number of the features p may be comparable to sample size n, or may even be substantially larger ($p \gg n$). In this section, we discuss the estimation of β_0 in the high-dimensional regime. For the non-convex optimization problem, as in the previous section, we restrict the search space to be $\mathbb{B}_2(r)$ for some large enough r that the true parameter β_0 is an interior point of the parameter space. We propose two estimators $\widehat{\beta}_\ell^H$, $\widehat{\beta}_s^H$ as solutions of the following optimization problems

$$\widehat{\beta}_\ell^H \in \underset{\beta \in \mathbb{B}_2(r)}{\arg\min}\, \mathcal{L}_n^\ell(\beta) + \lambda_\ell \|\beta\|_1 \quad \text{and} \quad \widehat{\beta}_s^H \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathcal{L}_n^s(\beta) + \lambda_s \|\beta\|_1, \tag{21}$$

where $\mathcal{L}_n^\ell(\beta)$ is a non-convex negative log-likelihood loss and $\mathcal{L}_n^s(\beta)$ is a convex surrogate loss. Here λ_ℓ and λ_s are tuning parameters which need to be chosen appropriately, and we will discuss their choices shortly. Finally, we note that in many cases, it is common to leave a finite number of coordinates unpenalized. An important special example is when the model includes an intercept feature. The theory that we develop in this section has a straightforward extension when the ℓ_1 penalty is modified to exclude a finite subset of features. In the low-dimensional setting, we established that the global minimizer can be obtained with high probability, but it is hard for a similar result to hold in the high-dimensional regime. Therefore, instead of $\widehat{\beta}_\ell^H$ we make use of a stationary point and define $\widetilde{\beta}_\ell^H$ to be a stationary point of the first optimization problem in (21).

We now study the statistical guarantees of the two estimators in the high-dimensional regime. First, we impose the standard sparsity assumption on β_0, $s_0 := \|\beta_0\|_0$. The core

condition which needs to be established is the restricted strong convexity (RSC) condition, the notion of which was first proposed by Negahban et al. (2012) for convex loss functions and extended for non-convex functions by Loh (2017) and Loh and Wainwright (2015). Similarly as in the definition in Loh (2017), we define the RSC condition as follows.

Definition 8 (RSC condition) *We say $\mathcal{L}_n$ satisfies a restricted strong convexity (RSC) condition with respect to β_0 with curvature α, a tolerance function $\tau_{n,p}(\cdot) : \mathbb{R} \to \mathbb{R}$, and a region $\Omega \subseteq \mathbb{R}^p$ if there exist $\alpha > 0$, $\tau_{n,p}(\cdot)$ such that*

$$\langle \nabla \mathcal{L}_n(\beta) - \nabla \mathcal{L}_n(\beta_0), \beta - \beta_0 \rangle \geq \alpha \|\beta - \beta_0\|_2^2 - \tau_{n,p}(\|\beta - \beta_0\|_1), \quad \forall \beta \in \Omega. \tag{22}$$

For example, the RSC condition in Loh (2017) corresponds to the choices $\tau_{n,p}(t) = \tau(\log p/n)t^2$ for a constant $\tau \geq 0$ and $\Omega = \mathbb{B}_2(\delta; \beta_0)$ with a radius $\delta > 0$. The main idea behind the definition of RSC is that it is the relaxed version of the strong convexity; when $\alpha > 0$, $\tau_{n,p} \equiv 0$ and the inequality (22) holds for all β and $\beta_0 \in \mathbb{R}^p$. Even if $\mathcal{L}_n$ is convex, $\mathcal{L}_n$ cannot be strongly convex in the high-dimensional regime due to the rank deficiency, which causes the curvature to vanish in some directions. The RSC condition guarantees that gradient information can still be exploited to direct the algorithm to the optimal point β_0 in the lack of strong convexity.

We will establish the RSC condition (22) with the choices $\tau_{n,p}(t) = \tau_\ell \sqrt{\log p/n} t$ and $\tau_{n,p}(t) = \tau_s(\log p/n)t^2$ for $\mathcal{L}_n^\ell(\beta)$ and $\mathcal{L}_n^s(\beta)$, respectively, for some $\tau_\ell, \tau_s > 0$. First, we discuss some additional conditions needed to establish the RSC condition for the negative log-likelihood loss $\mathcal{L}_n^\ell(\beta)$.

A3. *There exist $C_\rho > 0$ such that $\sup_t |\ell''(t, z)t| \leq C_\rho$, for all $z \in \mathcal{Z}$.*

A4. *There exist $C_d, C_b < \infty$ such that $\max_{1 \leq i \leq n} |\mathbf{x}_i^\top (\beta_0/\|\beta_0\|_2)| \leq C_d$, a.s. and $\|\beta_0\|_2 \leq C_b$.*

Assumption **A3** is a technical assumption which ensures that $\ell''(t.z)$ decays at least on the order of $1/t$ as $t \to \pm\infty$. We will show in Corollary 12 that Assumption **A3** is satisfied for the noisy labels model, where $\ell''(t, z) = A''(h_{LN}(t))h'_{LN}(t)^2 + (A'(h_{LN}(t)) - z)h''_{LN}(t)$ for $A(t) = \log(1 + \exp(t))$ and $h_{LN}(\cdot)$ defined in Section 3.2. Assumption **A4** concerns the boundedness of the signal. In particular, we assume that the size of $\mathbf{x}$ projected onto β_0 as well as $\|\beta_0\|_2$ are bounded.

Now we present two propositions to establish the RSC conditions with high probability for $\mathcal{L}_n^\ell(\beta)$ and $\mathcal{L}_n^s(\beta)$.

Proposition 9 *Suppose a sample is from a* **(GLM)** *with (A, g) which satisfies the random design condition* **A1'**. *We assume a high-dimensional regime where $p \gg n$ and $\log p/n = o(1)$. Also, we assume that ℓ is smooth and has a fast decaying tail (***A2-A3***), and that the linear signal is bounded (***A4***). Then for any given $\epsilon > 0$, there exist positive constants α_ℓ and τ_ℓ such that the following event*

$$\left(\nabla \mathcal{L}_n^\ell(\beta) - \nabla \mathcal{L}_n^\ell(\beta_0)\right)^\top (\beta - \beta_0) \geq \alpha_\ell \|\beta - \beta_0\|_2^2 - \tau_\ell \sqrt{\frac{\log p}{n}} \|\beta - \beta_0\|_1, \quad \forall \beta \in \mathbb{B}_2(r) \tag{23}$$

holds with probability at least $1 - \epsilon$, given a sufficiently large sample size $n \geq C(1/\epsilon)^{1/7}$ for a constant C depending only on the model parameters.

The proof is deferred to the Appendix B.1. We also present an equivalent result for the convex surrogate loss when a sample is from a **(GLM-L)** model. We recall that the convex

approach discussed in the previous section is available when a sample is from **(GLM-L)** model.

Proposition 10 *Suppose a sample is from a* **(GLM-L)** *model with* (A, a, b) *for* $a > 0$ *and* $b \geq 0$*, which satisfies the random design condition* **A1'***. Also assume the high-dimensional regime as in the Proposition 9 and* $\|\beta_0\|_2 = O(1)$*. Then for any given* $\epsilon > 0$*, there exist positive constants* α_s *and* τ_s *such that for* $n \geq C \log(1/\epsilon)$*, it holds with probability at least* $1 - \epsilon$ *that*

$$\left(\nabla \mathcal{L}_n^s(\beta) - \nabla \mathcal{L}_n^s(\beta_0)\right)^\top (\beta - \beta_0) \geq \alpha_s \|\beta - \beta_0\|_2^2 - \tau_s \frac{\log p}{n} \|\beta - \beta_0\|_1^2, \quad \forall \beta \in \mathbb{B}_2(1; \beta_0), \tag{24}$$

where the constant C *depends only on the model parameters.*

The key observation to establish the RSC result (24) is that the form of $\mathcal{L}_n^s(\beta)$ coincides with the negative log-likelihood function of a generalized linear model with the canonical link. Although $\mathbb{P}(T(\mathbf{z})|\mathbf{x})$ does not belong to the GLM family, the role of $T(\mathbf{z})$ is limited in establishing the restricted strong convexity, and the proof for the generalized linear model with the canonical link in Negahban et al. (2012) can be almost applied directly. More details are provided in Appendix B.2.

Now we state the following results regarding ℓ_1 and ℓ_2 error bounds. The first part of the theorem—for the ℓ_1 and ℓ_2 error bounds of the non-convex estimator—is a modification of Theorem 1 in Loh (2017), where Theorem 1 in Loh (2017) established error bounds for a stationary point under the RSC condition with a different tolerance function $\tau_{n,p}(t) = \tau(\log p/n)t^2$. The ℓ_1 and ℓ_2 error bounds for the convex estimator can be obtained by applying Theorem 1 in Negahban et al. (2012). To apply Theorem 1 in Negahban et al. (2012), we show that the RSC condition (24) implies the RSC condition in Negahban et al. (2012). We defer the detailed discussion to the Appendix B.3.

Theorem 11 (ℓ_1 and ℓ_2 error bound) *Assume* $\mathcal{L}_n^\ell$ *and* $\mathcal{L}_n^s$ *satisfy the RSC conditions* (23) *and* (24) *and also assume the high-dimensional regime as in the Proposition 9.*

1. *If* $\lambda_\ell \geq 4 \max\{\|\nabla \mathcal{L}_n^\ell(\beta_0)\|_\infty, \tau_\ell \sqrt{\frac{\log p}{n}}\}$*, then,*

$$\|\widetilde{\beta}_\ell^H - \beta_0\|_2 \leq c_1 \frac{\sqrt{s_0}\lambda_\ell}{\alpha_\ell} \quad \text{and} \quad \|\widetilde{\beta}_\ell^H - \beta_0\|_1 \leq 4c_1 \frac{s_0 \lambda_\ell}{\alpha_\ell}. \tag{25}$$

2. *If* $\lambda_s \geq 2\|\nabla \mathcal{L}_n^s(\beta_0)\|_\infty$ *and* $n \geq (32\tau_s/\alpha_s) s_0 \log p$*, then*

$$\|\widehat{\beta}_s^H - \beta_0\|_2 \leq c_2 \frac{\sqrt{s_0}\lambda_s}{\alpha_s} \quad \text{and} \quad \|\widehat{\beta}_s^H - \beta_0\|_1 \leq 4c_2 \frac{s_0 \lambda_s}{\alpha_s} \tag{26}$$

Here $c_1, c_2 > 0$ *are generic constants and* $s_0 := \|\beta_0\|_0$*.*

In particular, if $\|\nabla \mathcal{L}_n^\ell(\beta_0)\|_\infty, \|\nabla \mathcal{L}_n^s(\beta_0)\|_\infty = O(\sqrt{\frac{\log p}{n}})$ w.h.p, both estimators achieve the minimax-optimal error rates with the choices of $\lambda_\ell, \lambda_s \asymp \sqrt{\frac{\log p}{n}}$. In the following Corollary 12, we summarize the results about error bounds for the noisy labels model.

Corollary 12 *Suppose a sample* $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ *is from a* **(GLM)** *with* $(\log(1+\exp(\cdot)), g_{LN})$ *and* $\mathbf{z}_i \in \{0,1\}$. *Assume the high-dimensional regime as in the Proposition 9, the random design condition* **A1'**, *and the boundedness of the signal* **A4**. *For the choices of* $\lambda_\ell, \lambda_s \asymp \sqrt{\frac{\log p}{n}}$, *it holds that*

$$\|\widetilde{\beta}_\ell^H - \beta_0\|_2 \le c_1 \frac{\sqrt{s_0}\lambda_\ell}{\alpha_\ell} \quad \textit{and} \quad \|\widehat{\beta}_s^H - \beta_0\|_2 \le c_2 \frac{\sqrt{s_0}\lambda_s}{\alpha_s}$$

with probability at least $1-\epsilon$, *given a sufficiently large sample size* $n \ge C(1/\epsilon)^{1/7}$, *for a constant* C *which only depends on the model parameters.*

Notably, both estimators achieve the same optimal rates although there could still be a constant gap between the two estimators due to the different multipliers. We compare the performance of the two estimators empirically in Section 6.

5.2 Hypothesis Testing

Sparse estimators are known to have intractable limiting distributions even in the low-dimension regime (Knight and Fu, 2000). Nonetheless, it is of interest to quantify the uncertainty in the obtained estimators and test the significance of features. In this section, we discuss how we can carry out a test using the point estimates discussed in the previous section.

We take a de-biasing approach and obtain a one-step estimator whose direction is based on an estimating equation $\psi_n(\beta) := \frac{1}{n}\sum_{i=1}^n \psi(\mathbf{x}_i^\top \beta, \mathbf{z}_i)\mathbf{x}_i$. For a function $\psi : \mathbb{R} \to \mathbb{R}$, we consider ψ satisfying the following two properties:

1. ψ has an expectation of zero at $\beta = \beta_0$:

$$\mathbb{E}_{\beta_0}[\psi(\mathbf{x}^\top \beta_0, \mathbf{z})\mathbf{x}] = 0, \tag{27}$$

2. The derivative of ψ with respect to its first argument can be decomposed into the sum of ψ'_R and ψ'_I,

$$\psi'(t,z) = \psi'_I(t) + \psi'_R(t,z) \tag{28}$$

where ψ'_I and ψ'_R satisfy $\psi'_I(t) > 0, \forall t$ and $\mathbb{E}_{\beta_0}[\psi'_R(\mathbf{x}^\top \beta_0, \mathbf{z})] = 0$.

Two particular choices of ψ that we will consider subsequently will be derivatives of the log-likelihood loss and the surrogate loss,

$$\psi^\ell(\mathbf{x}^\top\beta, \mathbf{z}) := \ell'(\mathbf{x}^\top\beta, \mathbf{z}) = \{A'(h(\mathbf{x}^\top\beta)) - \mathbf{z}\}h'(\mathbf{x}^\top\beta)$$
$$\psi^s(\mathbf{x}^\top\beta, \mathbf{z}) := \ell'_s(\mathbf{x}^\top\beta, \mathbf{z}) = A'(\mathbf{x}^\top\beta) - T(\mathbf{z}),$$

where $\ell'(t,z)$ is the derivative (with respect to a linear predictor) of the log-likelihood loss defined in (10), and $\ell'_s(t,z)$ is the derivative of the surrogate loss defined in (15). Obviously, both ψ^ℓ and ψ^s satisfy (27). Also, when $\psi = \psi^\ell$, the choices of $\psi'_I(t) = \rho_I(t)$ and $\psi'_R(t,z) = \rho_R(t,z)$, for ρ_I and ρ_R are defined in (12), will satisfy (28). On the other hand, if $\psi = \psi^s$, the choices of $\psi'_I(t) = A''(t)$ and $\psi'_R \equiv 0$ will satisfy (28).

The derivative of the estimating equation plays an important role in determining the asymptotic variances of the generalized estimating equation (GEE) estimators (Godambe, 1960). We define an empirical Jacobian matrix $\psi'_{I,n}(\beta)$ of $\mathbb{E}[\psi_n(\beta)]$ and the inverse of $\mathbb{E}[\psi'_{I,n}(\beta_0)]$ as

$$\psi'_{I,n}(\beta) := \frac{1}{n}\sum_{i=1}^{n} \psi'_I(\mathbf{x}_i^\top \beta)\mathbf{x}_i\mathbf{x}_i^\top \quad \text{and} \quad \Theta(\psi) := \mathbb{E}[\psi'_I(\mathbf{x}^\top\beta_0)\mathbf{x}\mathbf{x}^\top]^{-1}, \tag{29}$$

We note that the minimum eigenvalue of $\mathbb{E}[\psi'_I(\mathbf{x}^\top\beta_0)\mathbf{x}\mathbf{x}^\top]$ can be shown to be bounded above by a positive constant under our assumptions, so that $\Theta(\psi)$ is well-defined (see Appendix B.5).

5.3 De-Biasing

For an initial estimate $\widehat{\beta}$, we define a de-biased estimator using ψ as follows,

$$\widehat{\beta}^{\mathrm{db}}(\psi) := \widehat{\beta} - \widehat{\Theta}(\psi)\psi_n(\widehat{\beta}) \tag{30}$$

which is a one-step estimator starting at $\widehat{\beta}$. Here a matrix $\widehat{\Theta}(\psi)$ is an approximation of $\Theta(\psi)$.

We make the following assumption about the sparsity level of β_0 and $\Theta(\psi)$ similarly as in van de Geer et al. (2014). We define the column sparsity level of $\Theta(\psi)$ (except the diagonal entries) as $s_* := \max_{1\le j\le p} \|\Theta(\psi)_{j,-j}\|_0$, and recall the definition $s_0 := \|\beta_0\|_0$.

A5. *$s_*, s_0 = o(\sqrt{n}/\log p)$, and $\|\mathbf{X}\Theta(\psi)_j\|_\infty = O_p(1), \forall j$*

Also we state conditions regarding the estimation equation ψ. In particular, we assume that ψ and ψ' are bounded and ψ' is also Lipschitz continuous with respect to its first argument. Precisely,

A6. *(Lipschitz continuity of ψ' and boundedness of ψ and ψ') Both ψ'_R and ψ'_I are Lipschitz with respect to its first argument, i.e.,*

$$|\psi'_R(a,z) - \psi'_R(a',z)| \le L_\psi|a-a'|, \forall z \quad \textit{and} \quad |\psi'_I(a) - \psi'_I(a')| \le L_\psi|a-a'|.$$

In particular, ψ' is Lipschitz with Lipschitz constant $2L_\psi$. Furthermore, there exists $C_\psi < \infty$ such that $\max\{\|\psi\|_\infty, \|\psi'_I\|_\infty, \|\psi'_R\|_\infty\} \le C_\psi$.

Now we state Theorem 13 which gives the asymptotic distributions of de-biased estimators.

Theorem 13 *Assume the random design condition* **A1'**, **A5-A6**, *and $\|\beta_0\|_2 = O(1)$. Suppose $\widehat{\Theta}(\psi)$ is chosen to satisfy $\|\widehat{\Theta}(\psi)_j - \Theta(\psi)_j\|_1 = o_p(\sqrt{\frac{1}{\log p}})$ and $\|e_j - \psi'_{I,n}(\widehat{\beta})\widehat{\Theta}(\psi)_j\|_\infty = O_p(\sqrt{\frac{\log p}{n}}), \forall j$. For an initial estimate $\widehat{\beta}$ satisfying ℓ_1 and ℓ_2 bounds $\|\widehat{\beta}-\beta_0\|_1 = O_p(s_0\sqrt{\frac{\log p}{n}})$ and $\|\widehat{\beta}-\beta_0\|_2^2 = O_p(s_0\frac{\log p}{n})$, and $\|\widehat{\beta}-\beta_0\|_1/\|\widehat{\beta}-\beta_0\|_2 = O(\sqrt{s_0})$ a.s., we have for any $j \in \{1,\dots,p\}$,*

$$\sqrt{n}(\widehat{\beta}^{\mathrm{db}}(\psi)_j - \beta_{0j})/\sigma(\psi)_j = Z_j + o_p(1)$$

for Z_j which converges weakly to a $\mathcal{N}(0,1)$ distribution and for

$$\sigma(\psi)_j := \sqrt{\Theta(\psi)_j^\top \mathbb{E}[\psi(\mathbf{x}^\top\beta_0, \mathbf{z})^2\mathbf{x}\mathbf{x}^\top]\Theta(\psi)_j}.$$

Moreover, if the bound in **A5** *and the conditions in the theorem statement regarding* $\widehat{\Theta}(\psi)_j$ *hold uniformly in* j*, then the result also holds uniformly in* j*.*

We note that obtaining $\widehat{\Theta}(\psi)$ satisfying the conditions of Theorem 13 is possible by taking a similar approach as in van de Geer et al. (2014) using node-wise regressions. In Appendix B.6, we provide more details about such construction. We also note that an initial estimate $\widehat{\beta}$ which satisfies the following cone condition, $\|(\widehat{\beta} - \beta_0)_{S^c}\|_1 \leq L\|(\widehat{\beta} - \beta_0)_S\|_1$ for some $L > 0$ and $S \subseteq \{1, \ldots, p\}$ such that $|S| = s_0$, also satisfies the ℓ_1/ℓ_2 ratio condition of the error vector in Theorem 13, since $\|\widehat{\beta} - \beta_0\|_1 \leq (L+1)\sqrt{s_0}\|\widehat{\beta} - \beta_0\|_2$. The proof of Theorem 13 is deferred to Appendix B.5. The main argument follows similar lines as in the proof of Theorem 3.1 in van de Geer et al. (2014), with additional arguments to handle the potential non-monotonicity of ψ, which can arise from a non-convex loss function (due to a non-canonical GLM). Finally, we state the following Corollary 14 for the asymptotic distributions of the de-biased estimators for the label noise model.

Corollary 14 *Suppose we have a sample* $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$ *from a* **(GLM)** *with parameters* $(\log(1+\exp(\cdot)), g_{LN})$ *and* $\mathbf{z}_i \in \{0, 1\}$*. We assume the conditions of Proposition 9. We also assume that* $\widehat{\Theta}(\psi^\ell)$ *and* $\widehat{\Theta}(\psi^s)$ *satisfy the conditions about* $\widehat{\Theta}(\psi)$ *in Theorem 13, and* **A5** *holds. We consider two de-biased estimators:*

$$\widehat{\beta}_\ell^{\text{db}} := \widetilde{\beta}_\ell^H - \widehat{\Theta}(\psi^\ell)\psi_n^\ell(\widetilde{\beta}_\ell^H) \quad \textit{and} \quad \widehat{\beta}_s^{\text{db}} := \widehat{\beta}_s^H - \widehat{\Theta}(\psi^s)\psi_n^s(\widehat{\beta}_s^H).$$

We then have, for any $j \in \{1, \ldots, p\}$*,*

$$\sqrt{n}(\widehat{\beta}_{\ell,j}^{\text{db}} - \beta_{0j})/\sigma(\psi^\ell)_j = Z_j + o_p(1)$$
$$\sqrt{n}(\widehat{\beta}_{s,j}^{\text{db}} - \beta_{0j})/\sigma(\psi^s)_j = \widetilde{Z}_j + o_p(1)$$

for $Z_j, \widetilde{Z}_j$ *which converge weakly to a* $\mathcal{N}(0, 1)$ *distribution and for,*

$$\sigma(\psi^\ell)_j = \sqrt{\mathcal{I}^\ell(\beta_0)_{jj}^{-1}} \quad \textit{and} \quad \sigma(\psi^s)_j = \sqrt{\mathcal{I}^s(\beta_0)_{jj}^{-1}}$$

where $\mathcal{I}^\ell(\beta)$ *and* $\mathcal{I}^s(\beta)$ *are defined in Proposition 5.*

The conditions about ψ^ℓ and ψ^s can be checked similarly as in the proof of Corollary 7. The rate conditions about the initial estimators can be checked by Corollary 12. Also, it is well known that both $\widetilde{\beta}_\ell^H - \beta_0$ and $\widehat{\beta}_s^H - \beta_0$ belong to a cone $\{\Delta; \|\Delta_S\|_1 \leq 3\|\Delta_{S^c}\|_1\}$ where $S \subseteq \{1, \ldots, p\}$ is the support of β_0. We note that these results are analogous to Proposition 5 in the low-dimensional setting once penalization and de-biasing are introduced.

6. Empirical Study

In this section, we present results about the empirical behavior of the non-convex likelihood-based estimator and the convex surrogate estimator. Our focus in this section is two-fold. First, we study the relative efficiency of the two estimators when different design matrices and noise rates are considered. In particular, we empirically demonstrate that the gap between $\mathcal{C}(\mathbf{X})$ and $\mathcal{C}(W_z^{-1}W_y\mathbf{X})$ captures well the impact of design $\mathbf{X}$ and noise

rates (ρ_0, ρ_1) on the relative efficiency of the two estimators. Second, we study empirical performance of the two estimators in the low- and high-dimensional regimes, with and without regularization. As regards the estimation errors, the likelihood-based estimator is expected to perform better than the convex estimator in low dimensions. However, it is unclear whether this will continue to be true in high dimensions. Indeed, as we discuss hereafter, our simulation study shows that the convex estimator outperforms the likelihood-based estimator in terms of mean squared errors in sparse regimes, where signal strength is relatively low.

6.1 Methods

Based on the regime of each simulation, we obtain non-sparse estimates $\widehat{\beta}_\ell$ and $\widehat{\beta}_s$ from (16) in the low-dimensional regime or sparse estimates $\widetilde{\beta}_\ell^H$ and $\widehat{\beta}_s^H$ from (21) in the high-dimensional regime. We recall that we define $\widetilde{\beta}_\ell^H$ as a stationary point of the optimization problem in (21) due to the non-convexity of $\mathcal{L}_n^\ell(\beta)$. For non-convex problems, we initialize coefficients at the null model where $\beta = [0, \ldots, 0]^\top$ if a problem is in the low-dimensional regime, and we use a local initialization using a convex estimate otherwise. To compare with the uncorrupted regime, the coefficient estimates $\widehat{\beta}_{\text{ref}}$ and $\widehat{\beta}_{\text{ref}}^H$ are computed using logistic or ℓ_1-penalized logistic regression on the un-corrupted data $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$.

In terms of optimization, we use the proximal gradient method combined with a backtracking line search to solve optimization problems of (16) and (21). This approach guarantees that iterates converge to a stationary point of the objective function if the objective function is non-convex and converge to an optimum in the convex case (e.g., Chapter 10 in Beck, 2017). For $\widehat{\beta}_{\text{ref}}$ and $\widehat{\beta}_{\text{ref}}^H$ we used the 'glm()' function from R base package and 'glmnet()' from R package **glmnet** respectively.

6.2 Impact of Design

To study the relative efficiency of the two estimators in various designs, we fix dimensions $(n = 1000, p = 10)$ and consider a mixture of multivariate normal distributions with varying distances between the two mixture components. We will demonstrate that increase in distance between the means of the two mixture components leads to an increase in the gap between $\mathcal{C}(\mathbf{X})$ and the perturbed column space $\mathcal{C}(W_z^{-1} W_y \mathbf{X})$, and a larger gap between two subspaces is associated with greater efficiency differences in $\widehat{\beta}_\ell$ and $\widehat{\beta}_s$.

Now we describe our simulation set-up for this subsection. First, we generate a design matrix $\mathbf{X} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top]^\top$ by sampling each $\mathbf{x}_i$ from an equal mixture of multivariate Gaussian distribution centered at $\mu_1 = (d, \ldots, d)$ and $\mu_2 = (-d, \ldots, -d)$ with various d and covariance matrix Σ such that $\Sigma_{ij} = 0.2^{|i-j|}$. We let $\beta_0 := [1/\sqrt{p}, \ldots, 1/\sqrt{p}]^\top$ so that $\|\beta_0\|_2^2 = 1$. The true unobserved response $\mathbf{y}_i$ is drawn by $\mathbf{y}_i \sim \text{Ber}(p_{\beta_0}(\mathbf{x}_i))$ where $p_{\beta_0}(\mathbf{x}_i) = (1 + \exp(-\mathbf{x}_i^\top \beta_0))^{-1}$, and a noisy label $\mathbf{z}_i$ is obtained by flipping $\mathbf{y}_i$ based on noise rates $\rho_0 = 10\%$ and $\rho_1 = 5\%$. The range of $d^2 = (0, \ldots, 2.5)$ is considered so that $\text{dist}^2 := \|\mu_1 - \mu_2\|_2^2 = 4pd^2$ varies from 0 to 100. When $\text{dist} = 0$, $\mathbf{x}_i$ is from single Gaussian distribution, i.e., $\mathbf{x}_i \sim N(0, \Sigma), \forall i$. For each d, we repeat the experiment $B = 10000$ times. At each d and iteration $b = 1, ..., B$, we calculate

- relative ℓ_2 difference: $\text{rd}(\mathcal{I}_n^\ell(\beta_0)_b, \mathcal{I}_n^s(\beta_0)_b) := \|I_p - \mathcal{I}_n^\ell(\beta_0)_b{}^{-1/2} \mathcal{I}_n^s(\beta_0)_b \mathcal{I}_n^\ell(\beta_0)_b{}^{-1/2}\|_2$,

- gap: $\widehat{\delta}(\mathcal{C}(\mathbf{X}_b), \mathcal{C}(W_z^{-1}W_y\mathbf{X}_b)) = \|\mathcal{P}_{\mathcal{C}(\mathbf{X}_b)} - \mathcal{P}_{\mathcal{C}(W_z^{-1}W_y\mathbf{X}_b)}\|_2$ (Kato, 2013),
- mean squared errors: $\text{mse}_b^\ell := \|\widehat{\beta}_{\ell,b} - \beta_0\|_2^2$ and $\text{mse}_b^s := \|\widehat{\beta}_{s,b} - \beta_0\|_2^2$,
- asymptotic mean squared errors:[2] $\text{amse}_b^\ell := \frac{\text{tr}(\mathcal{I}_n^\ell(\beta_0)_b^{-1})}{n}$ and $\text{amse}_b^s := \frac{\text{tr}(\mathcal{I}_n^s(\beta_0)_b^{-1})}{n}$,

where subscripts of b mean corresponding quantities are from the bth experiment. We summarize results by taking an average of B values.

To compare the efficiency of the two estimators, we calculate $\widehat{r}_{\text{mse}}$, the ratio of estimated mean squared errors, and $\widehat{r}_{\text{amse}}$, the ratio of asymptotic mean squared errors. More concretely, we let $\widehat{r}_{\text{mse}} := \frac{\overline{\text{mse}^\ell}}{\overline{\text{mse}^s}}$ and $\widehat{r}_{\text{amse}} := \frac{\overline{\text{amse}^\ell}}{\overline{\text{amse}^s}}$. When n is sufficiently large, $\widehat{r}_{\text{mse}}$ is expected to be close to $\widehat{r}_{\text{amse}}$, and both to be close to $r_{\text{amse}} := \lim_n \frac{\mathbb{E}[\|\widehat{\beta}_\ell - \beta_0\|_2^2]}{\mathbb{E}[\|\widehat{\beta}_s - \beta_0\|_2^2]}$. Note if the two estimators have the same efficiency, ratios will be close to 1. If the ratios are strictly less than 1, we can conclude that $\widehat{\beta}_\ell$ is more efficient than $\widehat{\beta}_s$.

Figure 2 plots the ratios of the mean squared errors and asymptotic mean squared errors, as well as $1 - \text{rd}(\mathcal{I}_n^\ell(\beta_0), \mathcal{I}_n^s(\beta_0))$ with varying gap^2 values, i.e. $\widehat{\delta}^2(\mathcal{C}(\mathbf{X}), \mathcal{C}(W_z^{-1}W_y\mathbf{X}))$. We recall that $1 - \text{rd}(\mathcal{I}_n^\ell(\beta_0), \mathcal{I}_n^s(\beta_0)) = 1$ iff two estimators have the same asymptotic efficiency, i.e. $\mathcal{I}_n^\ell(\beta_0) = \mathcal{I}_n^s(\beta_0)$, and $1 - \text{rd}(\mathcal{I}_n^\ell(\beta_0), \mathcal{I}_n^s(\beta_0)) < 1$ if $\mathcal{I}_n^\ell(\beta_0) \succ \mathcal{I}_n^s(\beta_0)$. We see from Figure 2 that $1 - \text{rd}(\mathcal{I}_n^\ell(\beta_0), \mathcal{I}_n^s(\beta_0))$ linearly decreases with the gap^2, which aligns with the result of Corollary 3. Also, the efficiency of the surrogate estimator worsens compared to the likelihood-based estimator as the gap increases, but not in the linear fashion as in the case

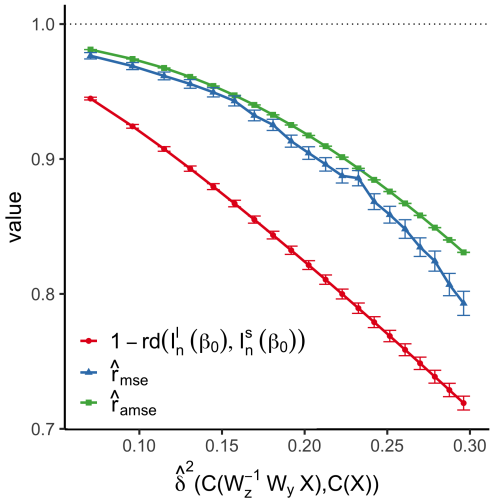


Figure 2: Ratios of mse and asymptotic mse and 1- relative ℓ_2 difference with varying gap^2. Error bars refer to 1se.

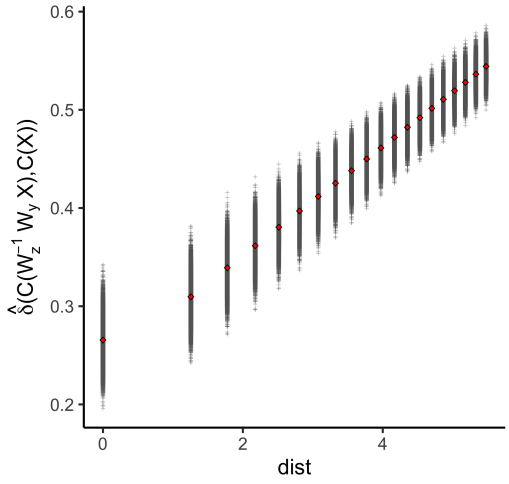


Figure 3: Plot of the distance between the means of two mixture distributions vs. the gap between the two column spaces.

2. $\mathbb{E}\|\widehat{\beta} - \beta_0\|_2^2 = \text{tr}(\mathbb{E}(\widehat{\beta} - \beta_0)(\widehat{\beta} - \beta_0)^\top) \approx \text{tr}(\mathcal{I}_n(\beta_0)^{-1}/n)$

of $1 - \mathrm{rd}(\mathcal{I}_n^\ell(\beta_0), \mathcal{I}_n^s(\beta_0))$. Unlike the relative ℓ_2 difference where we associated the quantity with variance ratio of the two estimators with respect to a particular direction u, variance ratios in all directions are considered in r_{amse} since $\widehat{r}_{\mathrm{amse}} = \mathrm{tr}(\mathcal{I}_n^\ell(\beta_0)^{-1})/\mathrm{tr}(\mathcal{I}_n^s(\beta_0)^{-1})$. Figure 3 plots the gap $\widehat{\delta}(\mathcal{C}(\mathbf{X}), \mathcal{C}(W_z^{-1}W_y\mathbf{X}))$ as functions of $\mathrm{dist} = \|\mu_1 - \mu_2\|_2$. We see that the gap between two subspaces increases as the distance between two mixture components increases.

6.3 Impact of Noise Rates

In this section, we study the relative efficiency of the two estimators with varying noise rates. From (19), for a given distribution $\mathbf{x}_i^\top \beta_0$, higher noise rates lead to a larger perturbation from $\mathcal{C}(\mathbf{X})$ to $\mathcal{C}(W_z^{-1}W_y\mathbf{X})$, and therefore a larger efficiency gap between the non-convex and the convex estimator is expected.

The distribution of the $\mathbf{x}_i^\top \beta_0$ plays a role in determining the gap between the two subspaces. For example, two samples from the models with the same noise rates can have very different gap values if the distributions of $\mathbf{X}$ are different. To illustrate this point concretely, suppose that two samples are from the models with the same rates of $\rho_0 = 0\%$ and $\rho_1 = 20\%$ but from different $\mathbf{X}$, where most $\mathbf{x}_i^\top \beta_0$ are negative in the first model but most $\mathbf{x}_i^\top \beta_0$ are positive in the second model. A larger amount of the variance misspecification using the convex approach will happen in the region when $\mathbf{x}_i^\top \beta_0$ is positive, since only the positive labels are flipped into the negative labels (Figure 4). This causes $W_z^{-1}W_y$ to deviate further from an identity matrix, and therefore, the gap between $\mathcal{C}(\mathbf{X})$ to $\mathcal{C}(W_z^{-1}W_y\mathbf{X})$ tends to be much larger for the second model despite the noise rates being the same in both models.

We empirically study the impact of the noise rates using a similar simulation set-up as in the previous subsection 6.2, except that we fix d—the varying parameter in the previous subsection—to be $d = 2/\sqrt{10}$ and instead vary the noise rates from 5% to 20%. In particular, the distribution of $\mathbf{x}_i^\top \beta_0$ is the same for all experiments in this section, which is an equal mixture of $\mathcal{N}(\beta_0^\top \mu_1, \beta_0^\top \Sigma \beta_0)$ and $\mathcal{N}(\beta_0^\top \mu_2, \beta_0^\top \Sigma \beta_0)$. We consider two settings for the noise rates to cover both symmetric and non-symmetric noise rates cases, where in the first setting we fix $\rho_1 = 5\%$ and vary ρ_0 from 5 to 20%, and in the second setting we let $\rho = \rho_0 = \rho_1$ and vary ρ from 5 to 20%. That is, for each noise rate setting and $b = 1, ..., B = 10000$, we generate $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)_{i=1}^n$ and obtain the gap value $\widehat{\delta}(\mathcal{C}(\mathbf{X}_b), \mathcal{C}(W_z^{-1}W_y\mathbf{X}_b)) = \|\mathcal{P}_{\mathcal{C}(\mathbf{X}_b)} -$

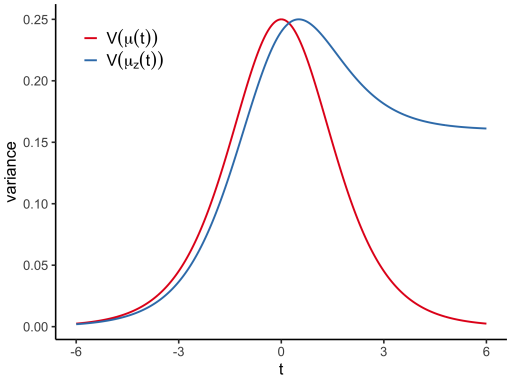
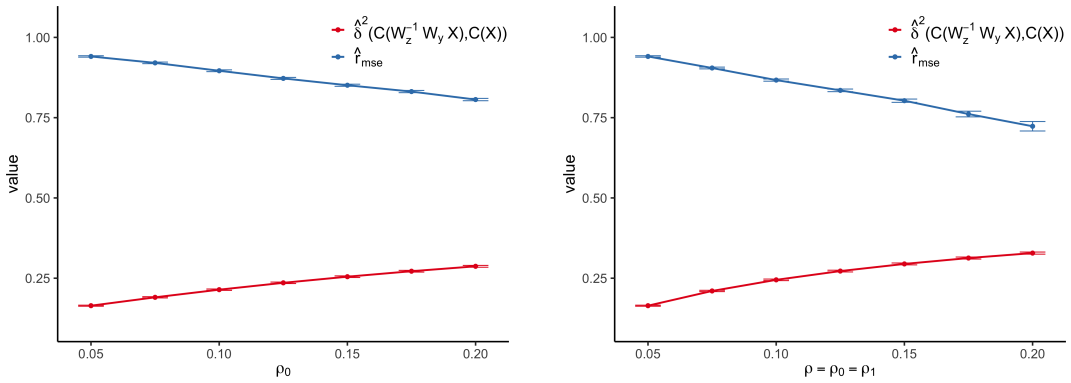


Figure 4: The plot of the variance of $\mathbf{y}$, $\mathrm{Var}(\mathbf{y}|\mathbf{x}) = \mathcal{V}(\mu(\mathbf{x}^\top\beta))$ and $\mathbf{z}$, $\mathrm{Var}(\mathbf{z}|\mathbf{x}) = \mathcal{V}_z(\mu(\mathbf{x}^\top\beta)) = \mathcal{V}((1 - \rho_1 - \rho_0)\mu(\mathbf{x}^\top\beta) + \rho_0)$, for varying $t = \mathbf{x}^\top\beta$ with $\rho_1 = 0.2$ and $\rho_0 = 0$. The difference between the two variances is larger when $t \gg 0$.



(a) Varying ρ_0 and fixed $\rho_1 = 5\%$

(b) Varying $\rho = \rho_0 = \rho_1$

Figure 5: Plots of ratios of mse and gap^2 with (a) varying ρ_0 and fixed $\rho_1 = 5\%$ and (b) varying $\rho = \rho_0 = \rho_1$. The error bars denote one standard error.

$\mathcal{P}_{\mathcal{C}(W_z^{-1}W_y\mathbf{X}_b)}\|_2$ and the mean squared errors $\text{mse}_b^\ell := \|\widehat{\beta}_{\ell,b} - \beta_0\|_2^2$ and $\text{mse}_b^s := \|\widehat{\beta}_{s,b} - \beta_0\|_2^2$, in the same way as we did in the previous subsection. The ratios of the estimated mean squared errors, $\widehat{r}_{\text{mse}} := \frac{\overline{\text{mse}^\ell}}{\overline{\text{mse}^s}}$ are then summarized over B values at each noise rate setting.

Figure 5 plots the ratios of the squared gaps $\widehat{\delta}^2(\mathcal{C}(\mathbf{X}), \mathcal{C}(W_z^{-1}W_y\mathbf{X}))$ and the mean squared errors as functions of noise rates. We note that $\widehat{r}_{\text{mse}} < 1$ implies the non-convex estimator had a smaller mse than the convex estimator. It can be seen from the plots the gap between two subspaces increases as the noise rates increase in both settings, and the convex approach performs worse than the non-convex approach.

6.4 Comparison of Estimation Errors in Low- and High-Dimensional Settings

To study estimation performances of the non-convex and convex approaches, we consider the following two regimes: (i) fixed $p = 10$ and growing n; (ii) growing (n, p) with $p = n$. Also, we consider two noise settings, where we use $\rho_1 = 5\%$ and $\rho_0 = 10\%$ for the first noise setting (low-noise) and double the noise rates for the second noise level setting (high-noise).

A sample $\mathbf{x}_i \in \mathbb{R}^p$ is generated from multivariate gaussian distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{p\times p}$ is given as $\Sigma_{i,j} = C_\Sigma(0.2)^{|i-j|}$, where C_Σ is chosen so that $\text{Var}(\mathbf{x}_i^\top \beta_0) = 5$. The sample size n varies from 1000 to 5000 where values in between are interpolated in a log scale. In both regimes, we first let 10 features be active ($s = 10$) and true parameter be $\beta_0 := [\underbrace{1, \ldots, 1}_{\text{s/2}}, \underbrace{-1 \ldots, -1}_{\text{s/2}}, 0, \ldots, 0]$. The true observed responses $\mathbf{y}_i$ and the noisy labels $\mathbf{z}_i$ are generated in the same way as in Section 6.2. Each experiment in the low-dimensional regimes is repeated $B = 300$ times and $B = 50$ times in the high-dimensional regimes. The mean and standard errors of B trials are reported in Figure 6. Tuning parameter λ needs to be chosen for the high-dimensional estimators. We choose λ in each simulation based on the testing loss from 5-fold cross validation.

Figure 6 shows the comparison results of the non-sparse and sparse estimators in both low and high-dimensional regimes. Not surprisingly, the likelihood-based estimator performs

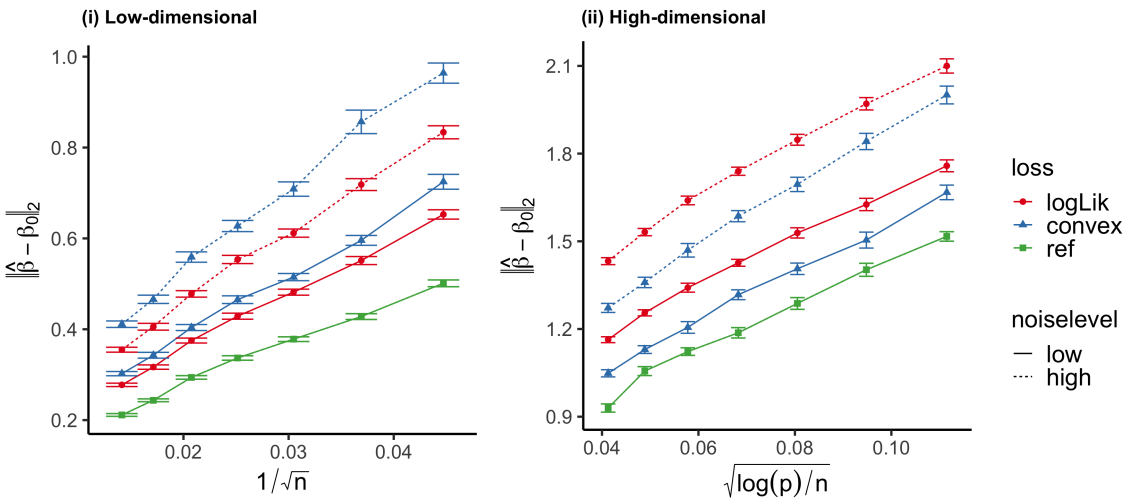
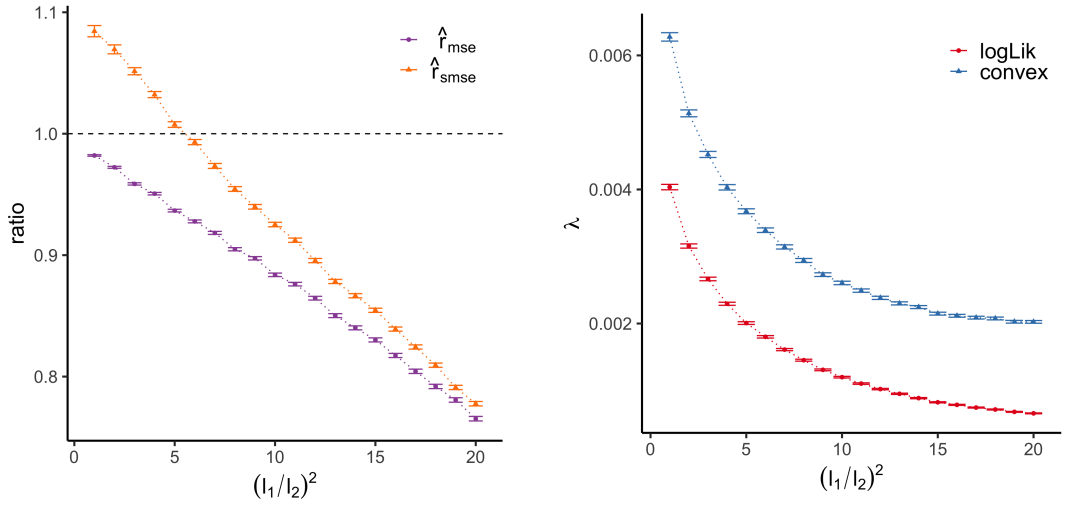


Figure 6: Comparison of the log-likelihood and surrogate loss based estimators in the low and high-dimensional regimes. Reference loss (ref) refers to the logistic loss when clean data is available.

uniformly better than the convex estimator in the low-dimensional regime without any regularization in the both noise settings. The loss of efficiency by using a convex surrogate loss appears to be relatively small when the noise level is low. The performance of the surrogate estimator worsens when noise rates increase since the squared gap $\widehat{\delta}^2(\mathcal{C}(\mathbf{X}), \mathcal{C}(W_z^{-1}W_y\mathbf{X}))$ increases, which agrees with the results in Section 6.3. On the contrary, the convex surrogate estimator appears to perform uniformly better than the likelihood-based loss in the high-dimensional setting.

It is well known that when no regularization is introduced, the likelihood function is the best function to optimize since the procedure results in the smallest asymptotic variance matrix (in regular problems). Its optimality (more precisely, the optimality of the score function) was also argued in classical estimating equation theory, where the score function is shown to be the best estimating equation function in the sense of minimizing the asymptotic variance (Godambe, 1960). The surrogate loss $\mathcal{L}_n^s(\beta)$ has a stronger curvature than $\mathcal{L}_n^\ell(\beta)$; in fact the curvature of $\mathcal{L}_n^s(\beta)$ is the same as $\mathcal{L}_n^c(\beta)$, the logistic loss from clean data. However, $\nabla\mathcal{L}_n^s(\beta)$ has also a larger variance than $\nabla\mathcal{L}_n^\ell(\beta)$ due to noise in the responses, resulting in the larger asymptotic variance matrix. We conjecture that in a penalized problem, especially when signal is relatively small compared to noise, regularization plays a role in reducing the variability in $\nabla\mathcal{L}_n^s(\beta)$ which leads to the better performance of the convex estimator in some cases.

To test our conjecture, we carry out an additional set of simulation. The set-up of the simulation mainly follows the set-up in Section 6.2 except we choose dimensions $(n = 10000, p = 20)$, fix $d = 3/\sqrt{p}$, and instead let the ℓ_1/ℓ_2 ratio of the true signal β_0 vary. More concretely, we fix $\|\beta_0\|_2 = 1$, and consider different sparsity levels



(a) $\widehat{r}_{\text{mse}}$ and $\widehat{r}_{\text{smse}}$ (b) λ for the likelihood and the convex losses

Figure 7: Plots of (a) $\widehat{r}_{\text{mse}}$ and $\widehat{r}_{\text{smse}}$ and (b) the amount of regularization chosen (λ) for the likelihood and the convex losses with varying ℓ_1/ℓ_2 ratios. The error bars denote one standard error. The dotted line represents when the ratio = 1, where the likelihood-based estimator and the convex estimator have the same efficiency.

s of $\beta_0 = [\underbrace{1/\sqrt{s}, \ldots, 1/\sqrt{s}}_{\text{s}}, \underbrace{0, \ldots, 0}_{\text{p-s}}]$ so that $\|\beta_0\|_1/\|\beta_0\|_2$ varies from 1 to $\sqrt{p}$. For each $s = 1, \ldots, p$, we obtain both non-sparse and sparse estimates. In the case of sparse estimates, tuning is performed by minimizing test loss on a test set of size n. At each s, the experiment is repeated $B = 10000$ times. Similarly as in Section 6.2, we calculate the two mean squared ratios each for the non-sparse and sparse estimates, $\widehat{r}_{\text{mse}} := \frac{\overline{\text{mse}^\ell}}{\overline{\text{mse}^s}}$ and $\widehat{r}_{\text{smse}} := \frac{\overline{\text{smse}^\ell}}{\overline{\text{smse}^s}}$, where we recall the definitions of $\text{mse}_b^\ell := \|\widehat{\beta}_{\ell,b} - \beta_0\|_2^2$ and $\text{mse}_b^s := \|\widehat{\beta}_{s,b} - \beta_0\|_2^2$ and define $\text{smse}_b^\ell := \|\widetilde{\beta}_{\ell,b}^H - \beta_0\|_2^2$ and $\text{smse}_b^s := \|\widehat{\beta}_{s,b}^H - \beta_0\|_2^2$.

As we can see from Figure 7, the likelihood-based estimator is always more efficient than the convex estimator in the case of non-sparse estimates. On the other hand, for the sparse estimators, the convex estimator outperforms the likelihood-based estimator when the ℓ_1/ℓ_2 ratio is small. As the ℓ_1/ℓ_2 ratio increases, we see from plot (b) that the amount of regularization decreases and the likelihood-based estimator starts to perform better than the convex estimator.

6.5 Comparison of Empirical Confidence Interval Coverage Rates in Low- and High-Dimensional Settings

In this section, we compare the empirical coverage rates of confidence intervals from the convex and non-convex methods. Similarly as in the previous sections, we consider the low and high-dimensional regimes, where we let $(n, p, s) = (2000, 20, 10)$ in the low-dimensional regime and $(n, p, s) = (500, 1000, 10)$ in the high-dimensional regime. The noise rates

(a) dimensions: (n=2000, p =20)

	coverage (all)	coverage (nzero)	coverage (zero)	CI length
logLik	0.951	0.951	0.951	0.362
	(0.005)	(0.008)	(0.007)	(0.001)
convex	0.962	0.961	0.962	0.387
	(0.005)	(0.007)	(0.006)	(0.002)
logLik (debiased)	0.944	0.942	0.946	0.340
	(0.006)	(0.009)	(0.008)	(0.001)
convex (debiased)	0.946	0.938	0.953	0.360
	(0.005)	(0.009)	(0.006)	(0.002)

(b) dimensions: (n=500, p =1000)

	coverage (all)	coverage (nzero)	coverage (zero)	CI length
logLik (debiased)	0.965	0.900	0.965	0.368
	(0.001)	(0.009)	(0.001)	(0.001)
convex (debiased)	0.964	0.925	0.964	0.388
	(0.001)	(0.008)	(0.001)	(0.001)

Table 2: Average coverage rates for 100 confidence interval realizations of a noisy labels model in low and high dimensional settings for all, active (nzero), and non-active (zero) features. Numbers in parentheses represent one standard error.

(ρ_0, ρ_1), true parameter β_0, and features $(\mathbf{x}_i)_{i=1}^n$ are set up in the same way as in the previous section 6.2.

We construct $1-\alpha$ confidence intervals in low- and high-dimensional settings, where we obtain both non-sparse and sparse estimates in the low-dimensional setting and obtain sparse estimates in the high-dimensional setting. The nominal level $1-\alpha$ is set to be 0.95. Confidence intervals for the non-sparse estimators are constructed based on their asymptotic normality results in Proposition 5. For sparse estimators, we first de-bias the estimates and construct confidence intervals based on the asymptotic normality results for de-biased estimators in Proposition 14. For the estimation of inverse Hessian matrices which are needed for de-biasing, the inverse matrix of the Hessian matrix and an approximate inverse matrix based on node-wise regressions are used in the low- and high- dimensional settings respectively. Mean empirical coverage rates for all features (all), active features (nzero), and non-active features (zero), as well as the mean lengths of the confidence intervals (CI length) are reported in Table 2 based on 100 realizations of confidence intervals.

The overall coverage rates appear to be good for the both convex and non-convex methods in all regimes where about 95% of the constructed intervals contained the true parameters. In all settings, both likelihood based methods (without and with penalization) result in confidence intervals with shorter lengths than those from the convex methods, which agrees with the results in Proposition 5 and 14. Confidence intervals from de-biased

estimators tend to be less conservative than those from the non-sparse estimators, which seem to cause lower empirical coverage rates than the nominal level for non-zero coefficients. Similar observations have also been made in Dezeure et al. (2015).

We conclude this section by remarking on the relative performance of the two approaches and some practical implications. In terms of estimation errors, the non-convex estimator performed better than the convex estimator in unpenalized, low-dimensional settings with large n. Also, in penalized schemes, the non-convex likelihood approach empirically performed better when the true model was not highly sparse, as shown in Figure 7. Confidence intervals from both methods showed good empirical coverage rates and the average confidence interval length was shorter for the likelihood approach than for the convex approach. Therefore, the non-convex approach is preferred in unpenalized and large n regimes, or in penalized and relatively not highly sparse regimes (ℓ_1/ℓ_2 ratio over $\sqrt{.3p}$ from Figure 7), potentially with multiple initializations. The convex approach provides a viable alternative to the non-convex likelihood-based approach in all settings, but it can be particularly advantageous in settings where the true model is highly sparse, or when running optimization algorithms multiple times with various initializations is computationally challenging. The convex objective is computationally attractive to work with since every stationary point of the objective is a global minimum.

7. Application to Beta-glucosidase Protein Data

In this section, we describe an application of our non-convex and convex methods to beta-glucosidase (BGL) protein sequence data. Beta-glucosidase is a key enzyme present in cellulase which converts cellobiose to glucose during cellulose hydrolysis. The BGL enzyme protein plays a significant role in bioethanol production (Singhania et al., 2013). Due to its industrial importance, it is of great interest to understand the effects of mutations of the protein and design a protein with improved functionality.

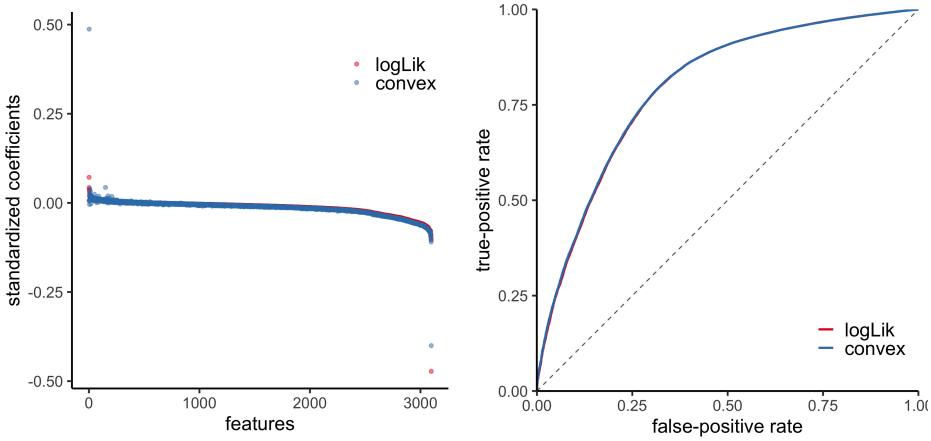
The data set we analyze is a positive and unlabeled beta-glucosidase protein sequence data set generated in the Romero Lab (Romero et al., 2015). Large-scale data were generated by deep mutational scanning (DMS) method, which applies the high-throughput screening method to sort out functional protein variants. Screening is based only on the enzyme functionality of a sequence. Unlabeled sequences from the initial library whose associated functionality is unknown are obtained together with screened sequences to be positive. The data consists of $n_\ell = 2533388$ functional (positive label) and $n_u = 1500277$ unlabeled sequences.[2] A sequence consists of 500 positions which takes one of 21 discrete values which correspond to 20 amino acid letter codes plus an additional letter for the alignment gap. From an alternative experiment, the prevalence of functional sequences in the unlabeled data set is known to be 0.35. This data set is previously analyzed in Song and Raskutti (2018) where the likelihood based approach was taken with the ℓ_1 penalization to obtain a sparse estimate. We obtained $p = 3097$ features using one-hot encoding of the sequences. Because each sequence contains only a few mutations, we obtained a sparse design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ by taking the amino acid levels in the WT sequence as the baseline levels. We note that the number of features ($p = 3097$) in the model is approximately one third of the number of maximum possible features ($10000 = 20 \times 500$). Since the sequences in

2. The raw data is available in https://github.com/RomeroLab/seq-fcn-data.git

the data set are local sequences around the wild-type sequence, some mutations were never observed.

We apply both convex and non-convex methods to the data set to estimate each mutation effect of the BGL sequence. Estimated coefficients are obtained by fitting the model using all sequence examples. In addition, to compare predictive performance of the two methods, we split the data set into training and test sets using 90% and 10% of the sequence examples. The model is then refitted using .1%, 1%, 10%, and 100% of the examples in the training set to compare the performance of the two methods at various sample sizes. For a performance metric, we use the area under the ROC curve (AUC) for the comparison of classification performances. However, for positive and unlabeled data, the ROC curve and AUC value calculated using the observed labels as the responses are biased for the ROC curve and AUC value for the unobserved true responses (Jain et al., 2017). Following the approach in Jain et al. (2017), we report the corrected ROC curve and AUC values.

The results are provided in Figure 8 and 9. We observe that the convex approach performed similarly as well as the likelihood approach, where the two approaches produce similar coefficient estimates and comparable classification performance results. Both classifiers demonstrate good classification performance. Therefore, using the convex estimator does not appear to result in a substantial loss of efficiency in this example. The corrected AUC value (from the full training set) is 0.7977 from the likelihood approach and 0.7989 from the surrogate approach, which is a significant improvement from AUC= 0.5 in the case of random classification.



(a) Estimated coefficients (b) Corrected ROC curves

Figure 8: Plot of (a) the estimated coefficients and (b) the corrected ROC curves (trained using all examples in the training data set) from the non-convex and convex approaches. In (a), features are sorted based on the coefficients from the likelihood approach.

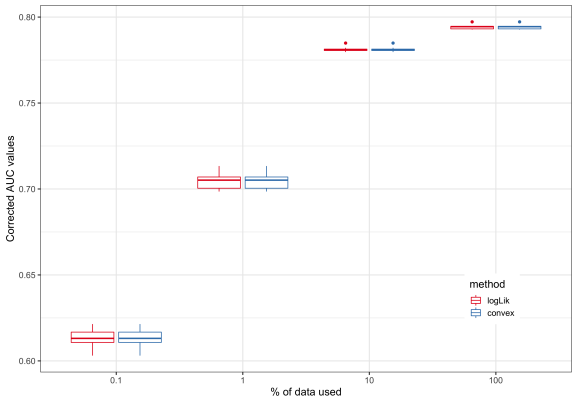


Figure 9: The corrected AUC values from models using various training sample sizes

8. Conclusion

We studied the binary regression problem in the presence of noise in labels in both the classic and high-dimension regimes. We demonstrated that the noisy label model belongs to a sub-class of generalized linear model family. We then discussed two approaches based on a convex surrogate loss for the sub-class of GLMs and a non-convex likelihood for the general class of GLMs. In the low-dimensional setting, the asymptotic distributions of the non-convex likelihood-based estimator and the convex surrogate estimator are derived. We also quantified the efficiency gap between the two approaches and argued that although the convex estimator is provably sub-optimal in terms of efficiency, the gap can be small in some applications. In the high-dimensional setting, we showed that both estimators, based on regularized non-convex and convex loss functions, achieve a minimax optimal $s \log p/n$ rate for the mean squared errors and derived the asymptotic distribution of the de-biased estimators which can be used for the hypothesis testing in a high-dimensional setting. We empirically demonstrated that both methods perform well in the simulation study and the real data analysis. In particular, although the estimator from the convex approach is sub-optimal in the low-dimensional regime, the efficiency gap between the two estimators is often small. Our empirical results suggest that in sparse regimes the convex surrogate estimator performs better than the likelihood-based estimator. It remains an open question to provide a theoretical justification for this claim.

Acknowledgments

H.S. was partially supported by the National Institute of Health via grant R01 GM131381-01. G.R. was partially supported by the National Science Foundation via grant DMS-1811767 and by the National Institute of Health via grant R01 GM131381-01. R.F.B. was partially supported by the National Science Foundation via grant DMS-1654076 and by an Alfred P. Sloan fellowship. We thank the three anonymous reviewers whose comments helped improve and clarify this manuscript.

Appendix A. Proofs for Results in Section 4

A.1 Proof of Proposition 2

First, since $\widehat{\beta}_\ell$ and $\widehat{\beta}_s$ are the minimizers of $\mathcal{L}_n^\ell(\beta)$ (over a compact region) and $\mathcal{L}_n^s(\beta)$, and both converge to $\mathbb{E}_{\beta_0}[\mathcal{L}_n^\ell(\beta)]$ and $\mathbb{E}_{\beta_0}[\mathcal{L}_n^s(\beta)]$ which have a unique maximizer at β_0, both $\widehat{\beta}_\ell$ and $\widehat{\beta}_s$ are consistent for β_0 under the Assumption **A1** (e.g., Theorem 2.1 and 2.7 in Newey and McFadden, 1994). Also, we note that $\widehat{\beta}_\ell$ and $\widehat{\beta}_s$ are zeros of

$$\nabla \mathcal{L}_n^\ell(\beta) = \frac{1}{n}\sum_{i=1}^n (\mu(h_{LN}(\mathbf{x}_i^\top \beta)) - \mathbf{z}_i) h'_{LN}(\mathbf{x}_i^\top \beta)\mathbf{x}_i = \frac{1}{n}\sum_{i=1}^n \psi^\ell(\beta, (\mathbf{x}_i, \mathbf{z}_i))\mathbf{x}_i \tag{31}$$

$$\nabla \mathcal{L}_n^s(\beta) = \frac{1}{n}\sum_{i=1}^n (\mu(\mathbf{x}_i^\top \beta) - T(\mathbf{z}_i))\mathbf{x}_i = \frac{1}{n}\sum_{i=1}^n \psi^s(\beta, (\mathbf{x}_i, \mathbf{z}_i))\mathbf{x}_i, \tag{32}$$

where we define

$$\psi^\ell(\beta, (\mathbf{x}, \mathbf{z})) := (\mu(h_{LN}(\mathbf{x}^\top \beta)) - \mathbf{z}) h'_{LN}(\mathbf{x}^\top \beta)$$
$$\psi^s(\beta, (\mathbf{x}, \mathbf{z})) := \mu(\mathbf{x}^\top \beta) - T(\mathbf{z}).$$

For notational simplicity, in what follows we write $\psi^{(i)}(\cdot) := \psi(\cdot, (\mathbf{x}_i, \mathbf{z}_i))\mathbf{x}_i$ for any $\psi \in \{\psi^\ell, \psi^s\}$. Also we define $\psi_n := n^{-1}\sum_{i=1}^n \psi^{(i)}$.

Then by the second order Taylor expansion, we can establish the asymptotic normality of the two estimators (e.g., Theorem 5.14 in Shao, 2003, Chapter 2.3.1 in Fahrmeir and Tutz, 2001). We first define the inverse of the asymptotic variance using an estimating equation ψ as

$$\mathcal{I}_n(\beta; \psi) \tag{33}$$
$$:= \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_\beta[\frac{d}{d\beta}\psi^{(i)}(\beta)|\mathbf{x}_i]\right)\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_\beta[\psi^{(i)}(\beta)\psi^{(i)}(\beta)^\top|\mathbf{x}_i]\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_\beta[\frac{d}{d\beta}\psi^{(i)}(\beta)|\mathbf{x}_i]\right).$$

Let $g_{CL}(\cdot) := \mu^{-1}(\cdot)$, which is a canonical link function. Recalling the fact $h_{LN}(t) = g_{CL} \circ g_{LN}^{-1}(t)$ and $g_{LN}^{-1}(t) = \mu_z(t) = (1-\rho_1-\rho_0)\mu(t) + \rho_0$, we have,

$$\psi^{\ell,(i)}(\beta) = (1-\rho_1-\rho_0)\left(\mu(h_{LN}(\mathbf{x}_i^\top \beta)) - \mathbf{z}_i\right) g'_{CL}(\mu_z(\mathbf{x}_i^\top \beta))\mu'(\mathbf{x}_i^\top \beta)\mathbf{x}_i \tag{34}$$

$$\psi^{s,(i)}(\beta) = \left(\mu(\mathbf{x}_i^\top \beta) - T(\mathbf{z}_i)\right)\mathbf{x}_i. \tag{35}$$

Also, $\mu'(t) = \mathcal{V}(\mu(t))$, since $\mu'(t) = A''(g_{CL} \circ \mu(t)) = \mathcal{V}(\mu(t))$ by the definition of $\mathcal{V}$. Also $g'_{CL}(t) = 1/\mathcal{V}(t)$ since $g'_{CL}(\mu(u)) = 1/\mu'(u) = 1/\mathcal{V}(\mu(u))$ by the chain rule. Plugging these expressions in (34), we obtain

$$\psi^{\ell,(i)}(\beta) = (1-\rho_1-\rho_0)\left(\mu(h_{LN}(\mathbf{x}_i^\top \beta)) - \mathbf{z}_i\right)\frac{\mathcal{V}(\mu(\mathbf{x}_i^\top \beta))}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top \beta))}\mathbf{x}_i.$$

Then since $\mathbb{E}_\beta[\mathbf{z}_i|\mathbf{x}_i] = \mu(h_{LN}(\mathbf{x}_i^\top\beta))$, we get

$$\begin{aligned}\mathbb{E}_\beta[\frac{d}{d\beta}\psi^{\ell,(i)}(\beta)|\mathbf{x}_i] &= A''(h_{LN}(\mathbf{x}_i^\top\beta))\left((1-\rho_1-\rho_0)\frac{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))}\right)^2\mathbf{x}_i\mathbf{x}_i^\top \\ &= \mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))\left((1-\rho_1-\rho_0)\frac{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))}\right)^2\mathbf{x}_i\mathbf{x}_i^\top \\ &= (1-\rho_1-\rho_0)^2\frac{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))^2}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))}\mathbf{x}_i\mathbf{x}_i^\top.\end{aligned}$$

Also, since $\psi^{\ell,(i)}(\beta)$ is a negative score function, the variance of the score function is the same as the expected negative derivative of the score function, i.e.,

$$\mathbb{E}_\beta[\psi^{\ell,(i)}(\beta)\psi^{\ell,(i)}(\beta)^\top|\mathbf{x}_i] = \mathbb{E}_\beta[\frac{d}{d\beta}\psi^{\ell,(i)}(\beta)|\mathbf{x}_i].$$

Then,

$$\mathcal{I}_n(\beta;\psi^\ell) = (1-\rho_1-\rho_0)^2\frac{1}{n}\sum_{i=1}^n\frac{\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))^2}{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))}\mathbf{x}_i\mathbf{x}_i^\top.$$

For the surrogate function, direct calculations give

$$\begin{aligned}\mathbb{E}_\beta[\frac{d}{d\beta}\psi^{s,(i)}(\beta)|\mathbf{x}_i] &= A''(\mathbf{x}_i^\top\beta)\mathbf{x}_i\mathbf{x}_i^\top = \mathcal{V}(\mu(\mathbf{x}_i^\top\beta))\mathbf{x}_i\mathbf{x}_i^\top \qquad (36)\\ \mathbb{E}_\beta[\psi^{s,(i)}(\beta)\psi^{s,(i)}(\beta)^\top|\mathbf{x}_i] &= \mathbb{E}_\beta[\Big(\mu(\mathbf{x}_i^\top\beta)-T(\mathbf{z}_i)\Big)^2\mathbf{x}_i\mathbf{x}_i^\top|\mathbf{x}_i] = \mathrm{Var}_\beta[T(\mathbf{z}_i)|\mathbf{x}_i]\mathbf{x}_i\mathbf{x}_i^\top\end{aligned}$$

since $\mathbb{E}_\beta[T(\mathbf{z}_i)|\mathbf{x}_i] = \mu(\mathbf{x}_i^\top\beta)$. Recalling the definition of $T(t) = (t-\rho_0)/(1-\rho_1-\rho_0)$ and $\mathcal{V}$,

$$\mathrm{Var}_\beta[T(\mathbf{z}_i)|\mathbf{x}_i] = (1-\rho_1-\rho_0)^{-2}\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta)). \qquad (37)$$

Thus we have

$$\begin{aligned}&\mathcal{I}_n^s(\beta;\psi^s)\\ &= \left(\frac{1}{n}\sum_{i=1}^n\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))\mathbf{x}_i\mathbf{x}_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^n\frac{\mathcal{V}(\mu_z(\mathbf{x}_i^\top\beta))}{(1-\rho_1-\rho_0)^2}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{V}(\mu(\mathbf{x}_i^\top\beta))\mathbf{x}_i\mathbf{x}_i^\top\right)\end{aligned}$$

by applying (36) and (37) to (33).

A.2 Proof of Corollary 3

First we show $\mathcal{I}_n^\ell(\beta_0) \succeq \mathcal{I}_n^s(\beta_0)$. By the definition of W_y and W_z, we have,

$$\begin{aligned}\mathcal{I}_n^\ell(\beta_0) &= (1-\rho_1-\rho_0)^2\mathbf{X}^\top W_y W_z^{-1} W_y\mathbf{X}/n\\ \mathcal{I}_n^s(\beta_0) &= (1-\rho_1-\rho_0)^2\mathbf{X}^\top W_y\mathbf{X}(\mathbf{X}^\top W_z\mathbf{X})^{-1}\mathbf{X}^\top W_y\mathbf{X}/n.\end{aligned}$$

Since the projection matrix $\mathcal{P}_{W_z^{1/2}\mathbf{X}}$ can be written as $\mathcal{P}_{W_z^{1/2}\mathbf{X}} = W_z^{1/2}\mathbf{X}(\mathbf{X}^\top W_z\mathbf{X})^{-1}\mathbf{X}^\top W_z^{1/2}$, we have $\mathcal{I}_n^s(\beta_0) = (1-\rho_1-\rho_0)^2\mathbf{X}^\top W_y W_z^{-1/2}\mathcal{P}_{W_z^{1/2}\mathbf{X}}W_z^{-1/2}W_y\mathbf{X}/n$. Then for any $v \in \mathbb{R}^n$,

$$\begin{aligned} v^\top(\mathcal{I}_n^\ell(\beta_0) - \mathcal{I}_n^s(\beta_0))v &= (1-\rho_1-\rho_0)^2 v^\top\mathbf{X}^\top W_y W_z^{-1/2}(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})W_z^{-1/2}W_y\mathbf{X}v/n \\ &= (1-\rho_1-\rho_0)^2\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})W_z^{-1/2}W_y\mathbf{X}v\|_2^2/n \geq 0 \end{aligned} \tag{38}$$

since $I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}}$ is idempotent. Thus, $\mathcal{I}_n^\ell(\beta_0) \succeq \mathcal{I}_n^s(\beta_0)$.

Now we address the inequality (18). First, we have

$$\|\mathcal{I}_n^\ell(\beta_0)^{-1/2}(\mathcal{I}_n^\ell(\beta_0) - \mathcal{I}_n^s(\beta_0))\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2 \leq \|\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2^2\|\mathcal{I}_n^\ell(\beta_0) - \mathcal{I}_n^s(\beta_0)\|_2,$$

and $\|\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2^2 = \|\mathcal{I}_n^\ell(\beta_0)^{-1}\|_2 = (1-\rho_1-\rho_0)^{-2}\sigma_{\min}(\mathbf{X}^\top W_y W_z^{-1}W_y\mathbf{X}/n)^{-1}$. Also, from (38),

$$\|\mathcal{I}_n^\ell(\beta_0) - \mathcal{I}_n^s(\beta_0)\|_2 = (1-\rho_1-\rho_0)^2\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})W_z^{-1/2}W_y\mathbf{X}\|_2^2/n.$$

Let $A := W_z^{-1/2}W_y\mathbf{X}$. Then,

$$\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})A\|_2^2 = \sup_{u\in\mathbb{R}^n}\frac{\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})Au\|_2^2}{\|u\|_2^2} = \sup_{u\in\mathcal{C}(A)}\frac{\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})u\|_2^2}{\|A^\dagger u\|_2^2}$$

where $A^\dagger$ is a Moore-Penrose inverse of A. Since $\|A^\dagger u\|_2^2 \geq \|u\|_2^2/\sigma_{\max}^2(A)$ for $u \in \mathcal{C}(A)$,

$$\begin{aligned} \|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})A\|_2^2 &\leq \sigma_{\max}^2(A)\sup_{u\in\mathcal{C}(A)}\frac{\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})u\|_2^2}{\|u\|_2^2} \\ &= \sigma_{\max}^2(A)\sup_{u\in\mathcal{C}(A),\|u\|_2=1}\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})u\|_2^2. \end{aligned}$$

Since $I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}}$ is a projection operator onto the orthogonal space of $\mathcal{C}(W_z^{1/2}\mathbf{X})$, $\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})u\|_2^2 = \inf_{v\in\mathcal{C}(W_z^{1/2}\mathbf{X})}\|u-v\|_2^2$. Therefore,

$$\begin{aligned} \|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})A\|_2^2 &\leq \sigma_{\max}^2(A)\sup_{u\in\mathcal{C}(A),\|u\|_2=1}\inf_{v\in\mathcal{C}(W_z^{1/2}\mathbf{X})}\|u-v\|_2^2 \\ &= \sigma_{\max}^2(W_z^{-1/2}W_y\mathbf{X})\delta^2(\mathcal{C}(W_z^{-1/2}W_y\mathbf{X}), \mathcal{C}(W_z^{1/2}\mathbf{X})), \end{aligned} \tag{39}$$

by the definition of the gap (17).

To proceed, we prove a lemma about the gap of two subspaces after linear transformation in relation to the original subspaces. Let $A(\mathcal{M}) := \{Av; v \in \mathcal{M}\}$, and note $\mathcal{C}(A\mathbf{X}) = A(\mathcal{C}(\mathbf{X}))$.

Lemma 15 *Let $\mathcal{M}, \mathcal{N} \subseteq \mathbb{R}^n$ be linear subspaces. Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix and $A(\mathcal{M}) := \{Av; v \in \mathcal{M}\}$. Then*

$$\delta(A(\mathcal{M}), A(\mathcal{N})) \leq \kappa(A)\delta(\mathcal{M}, \mathcal{N}),$$

where $\kappa(A) := \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$ is a condition number of A.

Proof By definition,

$$\begin{aligned}\delta(A(\mathcal{M}), A(\mathcal{N})) &= \sup_{u\in A(\mathcal{M}),\|u\|_2=1} \inf_{v\in A(\mathcal{N})} \|u-v\|_2 \\ &= \sup_{u\in \mathcal{M},\|Au\|_2=1} \inf_{v\in\mathcal{N}} \|Au-Av\|_2.\end{aligned}$$

For any $v \in \mathbb{R}^n$, we have $\sigma_{\min}(A)\|v\|_2 \le \|Av\|_2 \le \sigma_{\max}(A)\|v\|_2$ with $\sigma_{\min}(A) > 0$. Thus,

$$\begin{aligned}\sup_{u\in\mathcal{M},\|Au\|_2=1} \inf_{v\in\mathcal{N}} \|Au-Av\|_2 &\le \sup_{u\in\mathcal{M},\|u\|_2\le 1/\sigma_{\min}(A)} \inf_{v\in\mathcal{N}} \sigma_{\max}(A)\|u-v\|_2 \\ &\le \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \sup_{u\in\mathcal{M},\|u\|_2\le 1} \inf_{v\in\mathcal{N}} \|u-v\|_2\end{aligned}$$

where we use the fact for any $a \in \mathbb{R}$, $\inf_{v\in\mathcal{N}} \|u-v\|_2 = \inf_{v\in\mathcal{N}} \|u-av\|_2$ since if $v \in \mathcal{N}$ then $av \in \mathcal{N}$ by $\mathcal{N}$ being a linear subspace. We conclude by noting that $\sup_{u\in\mathcal{M},\|u\|_2\le 1} \inf_{v\in\mathcal{N}} \|u-v\|_2 = \sup_{u\in\mathcal{M},\|u\|_2=1} \inf_{v\in\mathcal{N}} \|u-v\|_2$. ■

By applying Lemma 15 to (39), we have

$$\delta(\mathcal{C}(W_z^{-1/2}W_y\mathbf{X}), \mathcal{C}(W_z^{1/2}\mathbf{X})) \le \kappa(W_z^{1/2})\delta(\mathcal{C}(W_z^{-1}W_y\mathbf{X}), \mathcal{C}(\mathbf{X})).$$

Therefore,

$$\begin{aligned}&\|\mathcal{I}_n^\ell(\beta_0)^{-1/2}(\mathcal{I}_n^\ell(\beta_0) - \mathcal{I}_n^s(\beta_0))\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2 \\ &\le \sigma_{\min}(\mathbf{X}^\top W_y W_z^{-1} W_y \mathbf{X}/n)^{-1}\|(I_n - \mathcal{P}_{W_z^{1/2}\mathbf{X}})W_z^{-1/2}W_y\mathbf{X}\|_2^2/n \\ &\le \sigma_{\min}(\mathbf{X}^\top W_y W_z^{-1} W_y \mathbf{X}/n)^{-1}\sigma_{\max}^2(W_z^{-1/2}W_y\mathbf{X}/\sqrt{n})\kappa(W_z)\delta^2(\mathcal{C}(W_z^{-1}W_y\mathbf{X}), \mathcal{C}(\mathbf{X})).\end{aligned}$$

Since $\sigma_{\min}(\mathbf{X}^\top W_y W_z^{-1} W_y \mathbf{X}/n)^{-1}\sigma_{\max}^2(W_z^{-1/2}W_y\mathbf{X}/\sqrt{n}) \le \kappa(\mathbf{X}^\top\mathbf{X}/n)\kappa(W_y^2)\kappa(W_z)$,

$$\|\mathcal{I}_n^\ell(\beta_0)^{-1/2}(\mathcal{I}_n^\ell(\beta_0) - \mathcal{I}_n^s(\beta_0))\mathcal{I}_n^\ell(\beta_0)^{-1/2}\|_2 \le \kappa(\mathbf{X}^\top\mathbf{X}/n)\kappa(W_y^2)\kappa(W_z^2)\delta^2(\mathcal{C}(W_z^{-1}W_y\mathbf{X}), \mathcal{C}(\mathbf{X})).$$

Note by Assumption **A1**, $\sup_i |\mathbf{x}_i^\top\beta|$ can be bounded by the term independent of n since $\sup_i |\mathbf{x}_i^\top\beta| \le \sup_i \|\mathbf{x}_i\|_2\|\beta\|_2 \le rC_X\sqrt{p}$. It follows that $\kappa(W_y^2), \kappa(W_z^2) = O(1)$. Also $\lambda_{\max}(\mathbf{X}^\top\mathbf{X}/n)$ is bounded by the term independent of n since p is fixed in the regime of interest and $\lambda_{\min}(\mathbf{X}^\top\mathbf{X}/n) \ge C_\lambda$ by Assumption **A1**.

A.3 Proof of Proposition 6

To show that there exists a unique stationary point in the interior of $\mathbb{B}_2(r)$ w.h.p, we show that there exists an ℓ_2 ball of radius ϵ_0 centered at β_0 in which $\mathcal{L}_n^\ell(\beta)$ is strongly convex w.h.p and has at least one local minimum, and no stationary point exists in $\mathbb{B}_2(r)\backslash\mathbb{B}_2(\epsilon_0;\beta_0)$.

We use the following three lemmas to establish the result, whose proofs are provided at the end of this sub-section. The first lemma is about the gradient and Hessian of the population risk. The second and third lemma establish the uniform convergence of the empirical loss, gradient and Hessian to their population counterparts, respectively. We let $\mathcal{L}^\ell(\beta) := \mathbb{E}[\mathcal{L}_n^\ell(\beta)]$.

Lemma 16 *There exist an $\epsilon_0 > 0$ and a constant $\gamma_\ell > 0$ such that*

$$\inf_{\beta \in \mathbb{B}_2(r) \setminus \mathbb{B}_2(\epsilon_0; \beta_0)} \|\nabla \mathcal{L}^\ell(\beta)\|_2 \wedge \inf_{\beta \in \mathbb{B}_2(\epsilon_0; \beta_0)} \lambda_{\min}(\nabla^2 \mathcal{L}^\ell(\beta)) \geq \gamma_\ell.$$

Lemma 17 *For any given $\delta > 0$, we have,*

$$\mathbb{P}\left(\sup_{\beta \in \mathbb{B}_2(r)} \left| \mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta) \right| \leq \tau \sqrt{\frac{Cp \log p}{n}} \right) \geq 1 - \delta,$$

where $C = \log(1/\delta)$ and τ is a constant depending on the model parameters $(K_X, K_Z, r, C_X, C_\ell)$.

Lemma 18 (Theorem 1 in Mei et al., 2018) *For $n \geq Cp \log p$ where $C = c_0 \cdot (\log(r\tau/\delta) \vee 1)$ for an absolute constant c_0 and $\tau = K_X \max\{C_\ell, L_\ell^{1/3}\}$,*

$$\mathbb{P}\left(\sup_{\beta \in \mathbb{B}_2(r)} \|\nabla \mathcal{L}_n^\ell(\beta) - \nabla \mathcal{L}^\ell(\beta)\|_2 \leq \tau \sqrt{\frac{Cp \log n}{n}} \right) \geq 1 - \delta$$
$$\mathbb{P}\left(\sup_{\beta \in \mathbb{B}_2(r)} \|\nabla^2 \mathcal{L}_n^\ell(\beta) - \nabla^2 \mathcal{L}^\ell(\beta)\|_2 \leq \tau^2 \sqrt{\frac{Cp \log n}{n}} \right) \geq 1 - \delta.$$

First we establish the result of Proposition 6 given Lemma 16, 17, and 18. By Lemma 17 and 18, the following inequalities

$$\sup_{\beta \in \mathbb{B}_2(r)} \|\nabla \mathcal{L}_n^\ell(\beta) - \nabla \mathcal{L}^\ell(\beta)\|_2 \vee \sup_{\beta \in \mathbb{B}_2(r)} \|\nabla^2 \mathcal{L}_n^\ell(\beta) - \nabla^2 \mathcal{L}^\ell(\beta)\|_2 \leq (\gamma_\ell/2) \wedge (\epsilon_g/4r)$$
$$\sup_{\beta \in \mathbb{B}_2(r)} \left| \mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta) \right| \leq \epsilon_L/4 \tag{40}$$

hold with at least probability $1 - 3\delta$ given a sufficiently large sample size, for γ_ℓ defined in Lemma 16, ϵ_L and ϵ_g defined in (41) and (42), respectively. We show that on the event (40) there exists a unique global minimum inside $\mathbb{B}_2(r)$.

First, on the event (40), we see that $\mathcal{L}_n^\ell(\beta)$ is strongly convex over $\mathbb{B}_2(\epsilon_0; \beta_0)$, since

$$\begin{aligned}
&\inf_{\beta \in \mathbb{B}_2(\epsilon_0; \beta_0)} \lambda_{\min}(\nabla^2 \mathcal{L}_n^\ell(\beta)) \\
&\geq \inf_{\beta \in \mathbb{B}_2(\epsilon_0; \beta_0)} \lambda_{\min}(\nabla^2 \mathcal{L}^\ell(\beta)) - \sup_{\beta \in \mathbb{B}_2(\epsilon_0; \beta_0)} \|\nabla^2 \mathcal{L}_n^\ell(\beta) - \nabla^2 \mathcal{L}^\ell(\beta)\|_2 \\
&\geq \gamma_\ell/2.
\end{aligned}$$

Then we argue that there exists a local minimum inside the ball $\mathbb{B}_2(\epsilon_0; \beta_0)$. It is sufficient to show that there exists $\beta \in \mathbb{B}_2(\epsilon_0; \beta_0) \setminus \partial \mathbb{B}_2(\epsilon_0; \beta_0)$ such that $\mathcal{L}_n^\ell(\beta) < \inf_{\beta \in \partial \mathbb{B}_2(\epsilon_0; \beta_0)} \mathcal{L}_n^\ell(\beta)$. Take $\beta = \beta_0$. Note there exists $\epsilon_L > 0$ such that

$$\inf_{\beta \in \partial \mathbb{B}_2(\epsilon_0; \beta_0)} \mathcal{L}^\ell(\beta) - \mathcal{L}^\ell(\beta_0) = \epsilon_L, \tag{41}$$

since $\partial \mathbb{B}_2(\epsilon_0; \beta_0)$ is compact and A is a strictly convex function. Then on the event (40),

$$\inf_{\beta \in \partial \mathbb{B}_2(\epsilon_0; \beta_0)} \mathcal{L}_n^\ell(\beta) \geq \inf_{\beta \in \partial \mathbb{B}_2(\epsilon_0; \beta_0)} \mathcal{L}^\ell(\beta) - \epsilon_L/4 \quad \text{and} \quad \mathcal{L}_n^\ell(\beta_0) \leq \mathcal{L}^\ell(\beta_0) + \epsilon_L/4.$$

Therefore,

$$\inf_{\beta\in\partial\mathbb{B}_2(\epsilon_0;\beta_0)} \mathcal{L}_n^\ell(\beta) - \mathcal{L}_n^\ell(\beta_0) \geq \inf_{\beta\in\partial\mathbb{B}_2(\epsilon_0;\beta_0)} \mathcal{L}^\ell(\beta) - \mathcal{L}^\ell(\beta_0) - \epsilon_L/2 \geq \epsilon_L/2,$$

where we use (41) for the last inequality. Also, the empirical gradient in $\mathbb{B}_2(r)$ does not vanish outside of $\mathbb{B}_2(\epsilon_0;\beta_0)$, since

$$\begin{aligned}
&\inf_{\beta\in\mathbb{B}_2(r)\setminus\mathbb{B}_2(\epsilon_0;\beta_0)} \|\nabla\mathcal{L}_n^\ell(\beta)\|_2 \\
&\quad\geq \inf_{\beta\in\mathbb{B}_2(r)\setminus\mathbb{B}_2(\epsilon_0;\beta_0)} \|\nabla\mathcal{L}^\ell(\beta)\|_2 - \sup_{\beta\in\mathbb{B}_2(r)} \|\nabla\mathcal{L}_n^\ell(\beta) - \nabla\mathcal{L}^\ell(\beta)\|_2 \\
&\quad\geq \gamma_\ell/2.
\end{aligned}$$

Finally, there exists no stationary point on the boundary of $\mathbb{B}_2(r)$. Note there is no stationary point of $\mathcal{L}^\ell(\beta)$ on $\partial\mathbb{B}_2(r)$ since $\langle\nabla\mathcal{L}^\ell(\beta), \beta_0 - \beta\rangle < 0$ for any $\beta \in \partial\mathbb{B}_2(r)$. Since $\partial\mathbb{B}_2(r)$ is compact, we have $\epsilon_g > 0$ such that

$$\sup_{\beta\in\partial\mathbb{B}_2(r)} \langle\nabla\mathcal{L}^\ell(\beta), \beta_0 - \beta\rangle < -\epsilon_g. \tag{42}$$

Then for any $\beta \in \partial\mathbb{B}_2(r)$,

$$\begin{aligned}
\langle\nabla\mathcal{L}_n^\ell(\beta), \beta_0 - \beta\rangle &= \langle\nabla\mathcal{L}^\ell(\beta), \beta_0 - \beta\rangle + \langle\nabla\mathcal{L}_n^\ell(\beta) - \nabla\mathcal{L}^\ell(\beta), \beta_0 - \beta\rangle \\
&\leq -\epsilon_g + \|\nabla\mathcal{L}_n^\ell(\beta) - \nabla\mathcal{L}^\ell(\beta)\|_2\|\beta - \beta_0\|_2 \leq -\epsilon_g/2.
\end{aligned}$$

Hence, on the event (40), there exists a unique stationary point in $\mathbb{B}_2(\epsilon_0;\beta_0) \subsetneq \mathbb{B}_2(r)$ which is a global minimum.

Now we turn to the proofs of three lemmas.

Proof [Proof of Lemma 16] First, we lower bound the minimum eigenvalue of the Hessian.

$$\begin{aligned}
\inf_{u;\|u\|_2=1} u^\top\nabla^2\mathcal{L}^\ell(\beta)u &= \inf_{u;\|u\|_2=1} \left(u^\top\nabla^2\mathcal{L}^\ell(\beta_0)u + u^\top\left(\nabla^2\mathcal{L}^\ell(\beta) - \nabla^2\mathcal{L}^\ell(\beta_0)\right)u\right) \\
&\geq \inf_{u;\|u\|_2=1} u^\top\nabla^2\mathcal{L}^\ell(\beta_0)u - \sup_{u;\|u\|_2=1} \left|u^\top\left(\nabla^2\mathcal{L}^\ell(\beta) - \nabla^2\mathcal{L}^\ell(\beta_0)\right)u\right|
\end{aligned}$$

At $\beta = \beta_0$,

$$\inf_{u;\|u\|_2=1} u^\top\nabla^2\mathcal{L}^\ell(\beta_0)u = \mathbb{E}[\ell''(\mathbf{x}^\top\beta_0, \mathbf{z})(\mathbf{x}^\top u)^2] = \mathbb{E}[\rho_I(\mathbf{x}^\top\beta_0)(\mathbf{x}^\top u)^2]$$

since $\mathbb{E}[\rho_R(\mathbf{x}^\top\beta_0, \mathbf{z})|\mathbf{x}] = 0$. Recalling the fact that $\rho_I(t) = A''(h(t))h'(t)^2 \geq 0$ for all t, we have a lower bound

$$\mathbb{E}[\rho_I(\mathbf{x}^\top\beta_0)(\mathbf{x}^\top u)^2] \geq \mathbb{E}[\rho_I(\mathbf{x}^\top\beta_0)(\mathbf{x}^\top u)^2\mathbb{1}\{|\mathbf{x}^\top\beta_0| \leq \tau_c\}] \geq \inf_{|t|\leq\tau_c} \rho_I(t)\mathbb{E}[(\mathbf{x}^\top u)^2\mathbb{1}\{|\mathbf{x}^\top\beta_0| \leq \tau_c\}]$$

for any $\tau_c > 0$. We let $\tau_c := \left(r^2K_X^2 \log\frac{16^2K_X^4}{C_\lambda^2}\right)^{1/2}$. Then by Cauchy-Schwarz, Assumption **A1'** and Lemma 19,

$$\begin{aligned}
\mathbb{E}[(\mathbf{x}^\top u)^2\mathbb{1}\{|\mathbf{x}^\top\beta_0| \leq \tau_c\}] &= \mathbb{E}[(\mathbf{x}^\top u)^2] - \mathbb{E}[(\mathbf{x}^\top u)^2\mathbb{1}\{|\mathbf{x}^\top\beta_0| \geq \tau_c\}] \\
&\geq C_\lambda - \mathbb{E}[(\mathbf{x}^\top u)^4]^{1/2}\mathbb{P}(|\mathbf{x}^\top\beta_0| \geq \tau_c)^{1/2} \geq C_\lambda/2.
\end{aligned} \tag{43}$$

Now we bound the difference term. Using Lipschitz assumption in **A2**, we have

$$\begin{aligned}\left|u^\top \left(\nabla^2 \mathcal{L}^\ell(\beta) - \nabla^2 \mathcal{L}^\ell(\beta_0)\right) u\right| &\leq \mathbb{E}[\left|\ell''(\mathbf{x}^\top\beta, \mathbf{z}) - \ell''(\mathbf{x}^\top\beta_0, \mathbf{z})\right| (\mathbf{x}^\top u)^2] \\ &\leq L_\ell \mathbb{E}[|\mathbf{x}^\top(\beta - \beta_0)|(\mathbf{x}^\top u)^2].\end{aligned}$$

Then by Cauchy-Schwarz and sub-Gaussian moment property,

$$L_\ell \mathbb{E}[|\mathbf{x}^\top(\beta - \beta_0)|(\mathbf{x}^\top u)^2] \leq L_\ell \|\Delta_0\|_2 \mathbb{E}[(\mathbf{x}^\top \Delta_0/\|\Delta_0\|_2)^2]^{1/2} \mathbb{E}[(\mathbf{x}^\top u)^4]^{1/2} \leq 4\sqrt{2} K_X^3 L_\ell \|\Delta_0\|_2 \tag{44}$$

where $\Delta_0 = \beta - \beta_0$. Hence, combining (43), (44), we conclude for β such that $\|\beta - \beta_0\|_2 \leq \epsilon_0$, for $\epsilon_0 := (\inf_{|t|\leq\tau_c} \rho_I(t) C_\lambda)/(16\sqrt{2} K_X^3 L_\ell)$,

$$\inf_{u;\|u\|_2=1} u^\top \nabla^2 \mathcal{L}^\ell(\beta) u \geq \inf_{|t|\leq\tau_c} \rho_I(t) C_\lambda/4.$$

Now we address the lower bound of the gradient. Let $\beta \in \mathbb{B}_2(r) \setminus \mathbb{B}_2(\epsilon_0; \beta_0)$ be fixed.

$$\begin{aligned}\langle \beta - \beta_0, \nabla \mathcal{L}^\ell(\beta)\rangle &= \mathbb{E}[\{A'(h(\mathbf{x}^\top\beta)) - A'(h(\mathbf{x}^\top\beta_0))\} h'(\mathbf{x}^\top\beta) \mathbf{x}^\top(\beta - \beta_0)] \\ &= \mathbb{E}[A''(h(\mathbf{x}^\top\beta_i)) h'(\mathbf{x}^\top\beta_i) h'(\mathbf{x}^\top\beta)(\mathbf{x}^\top(\beta - \beta_0))^2]\end{aligned}$$

for $\beta_i = \beta_0 + v(\beta - \beta_0)$ where $v \in [0, 1]$ by the mean value theorem.

Define an event $\mathcal{E} := \{|\mathbf{x}^\top\beta_0| \leq \tau_c, |\mathbf{x}^\top\Delta_0| \leq 2\tau_c\}$ where $\Delta_0 := \beta - \beta_0$.

$$\langle \beta - \beta_0, \nabla \mathcal{L}^\ell(\beta)\rangle \geq C_r \mathbb{E}[(\mathbf{x}^\top(\beta - \beta_0))^2 \mathbb{1}_\mathcal{E}] \geq C_r C_\lambda \|\Delta_0\|_2^2/2 \tag{45}$$

for $C_r := \left(\inf_{t;|t|\leq 3\tau_c} A''(h(t)) h'(t)\right) \left(\inf_{t;|t|\leq 3\tau_c} h'(t)\right) > 0$, since

$$\begin{aligned}\mathbb{E}[(\mathbf{x}^\top(\beta - \beta_0))^2 \mathbb{1}_\mathcal{E}] &= \mathbb{E}[(\mathbf{x}^\top(\beta - \beta_0))^2] - \mathbb{E}[(\mathbf{x}^\top(\beta - \beta_0))^2 \mathbb{1}_{\mathcal{E}^c}] \\ &\geq \|\Delta_0\|_2^2 \left(C_\lambda - \mathbb{E}[(\mathbf{x}^\top\Delta_0/\|\Delta_0\|_2)^4]^{1/2} \mathbb{P}(\mathcal{E}^c)^{1/2}\right) \\ &\geq \|\Delta_0\|_2^2 C_\lambda/2\end{aligned}$$

by Lemma 19. We apply Cauchy-Schwarz inequality to (45) to obtain

$$\|\nabla \mathcal{L}^\ell(\beta)\|_2 \geq C_\lambda C_r \|\beta - \beta_0\|_2/2 \geq (C_\lambda C_r \epsilon_0)/2$$

since $\|\beta - \beta_0\|_2 \geq \epsilon_0$. Finally, take $\gamma_\ell := C_\lambda \left(\inf_{|t|\leq\tau_c} \rho_I(t)/4 \wedge (C_r\epsilon_0)/2\right)$ to conclude. ■

Lemma 19 *Let $\mathbf{x} \in \mathbb{R}^p$ be a random vector which satisfies the sub-gaussian tail condition with the parameter K_X, and also let $u_2, u_2, u_3 \in \mathbb{R}^p$ be non-random vectors such that $\|u_1\|_2 = 1$, $\|u_2\|_2 \leq c_1 r$, and $\|u_3\|_2 \leq c_2 r$ for some $c_1, c_2 > 0$. For $\tau_c := \left(r^2 K_X^2 \log \frac{16^2 K_X^4}{C_\lambda^2}\right)^{1/2}$, we have*

$$\mathbb{E}[(\mathbf{x}^\top u_1)^4]^{1/2} \{\mathbb{P}(|\mathbf{x}^\top u_2| \geq c_1\tau_c) + \mathbb{P}(|\mathbf{x}^\top u_3| \geq c_2\tau_c)\}^{1/2} \leq \frac{C_\lambda}{2}.$$

Proof Moment and tail properties of sub-Gaussian distribution give $\mathbb{E}[(\mathbf{x}^\top u_1)^4]^{1/2} \leq 4K_X^2$ and $\mathbb{P}(|\mathbf{x}^\top u_2| \geq c_1\tau_c) + \mathbb{P}(|\mathbf{x}^\top u_3| \geq c_2\tau_c) \leq 4\exp(-\tau_c^2/r^2K_X^2)$. Then the choice of τ_c gives the desirable bound. ■

Proof [Proof of Lemma 17] First, we use an extension of McDiarmid's inequality to obtain

$$\mathbb{P}\left(\sup_{\beta\in\mathbb{B}_2(r)}\left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right| \geq \mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)}\left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right|\right] + t\right) \leq \exp(-Ct^2n),$$

for some constant $C > 0$. We will use the following extension of McDiarmids inequality, due to Kontorovich (2014).

Lemma 20 (Theorem 1 in Kontorovich, 2014) *Let $(\mathcal{X}_i, \rho_i, \mu_i)$ be a sequence of metric spaces, $i = 1, \ldots, n$. Let $\mathcal{X}^n = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $\mu^n = \mu_1 \times \cdots \times \mu_n$, and $\rho^n(x, x') = \sum_{i=1}^n \rho_i(x, x')$ be the product probability space, the product measure, and ℓ_1 product metric. Let X_i be $\mathcal{X}_i$-valued random variables where $X_i \sim \mu_i$. Suppose $\varphi : \mathcal{X}^n \to \mathbb{R}$ is 1-Lipschitz with respect to ρ^n metric, i.e., $|\varphi(x) - \varphi(x')| \leq \sum_{i=1}^n \rho_i(x, x')$ for $x, x' \in \mathcal{X}^n$, and there exists a sub-gaussian parameter $\Delta_{SG}(\mathcal{X}_i) < \infty$ such that*

$$\mathbb{E}[\exp(\lambda\sigma_i\rho_i(X_i, X_i'))] \leq \exp(\lambda^2\Delta_{SG}(\mathcal{X}_i)^2/2), \text{ for all } \lambda \in \mathbb{R},$$

where $X_i, X_i' \sim \mu_i$ are independent and σ_i is a Rademacher variable independent of (X_i, X_i'). Then $\mathbb{E}[\varphi] < \infty$, and

$$\mathbb{P}\left(|\varphi(X_1, \ldots, X_n) - \mathbb{E}[\varphi(X_1, \ldots, X_n)]| > t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \Delta_{SG}^2(\mathcal{X}_i)}\right).$$

Let $\mathbf{u}_i := [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip}, \mathbf{z}_i] \in \mathcal{X}_i$, for $\mathcal{X}_i := \mathbb{R}^p \times \mathcal{Z}$, $\forall i$. We define $\phi : \mathcal{X}^n \to \mathbb{R}$ as

$$\phi(\mathbf{u}_1, \ldots, \mathbf{u}_n) := \sup_{\beta\in\mathbb{B}_2(r)}\left|\mathcal{L}_n^\ell(\beta) - \mathbb{E}[\mathcal{L}_n^\ell(\beta)]\right| = \sup_{\beta\in\mathbb{B}_2(r)}\left|\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i^\top\beta, \mathbf{z}_i) - \mathbb{E}[\ell(\mathbf{x}_i^\top\beta, \mathbf{z}_i)]\right|.$$

We show that ϕ is L_0/n Lipschitz, for L_0 to be defined later. Let μ_i be the joint distribution of $(\mathbf{x}_i, \mathbf{z}_i)$, i.e., $\mu_i = \mathbb{P}_\mathbf{x} \times \mathbb{P}_{\mathbf{z}|\mathbf{x}}$, $\forall i$. For $(\mathbf{u}_1, \ldots, \mathbf{u}_n), (\mathbf{u}_1', \ldots, \mathbf{u}_n') \sim \mu^n$,

$$\begin{aligned}
&|\phi(\mathbf{u}_1, \ldots, \mathbf{u}_n) - \phi(\mathbf{u}_1', \ldots, \mathbf{u}_n')| \\
&\leq \sup_{\beta\in\mathbb{B}_2(r)}\left|\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i^\top\beta, \mathbf{z}_i) - \ell(\mathbf{x}_i'^\top\beta, \mathbf{z}_i')]\right| \\
&= \sup_{\beta\in\mathbb{B}_2(r)}\left|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\tilde{\mathbf{x}}_i^\top\beta, \tilde{\mathbf{z}}_i)^\top\begin{bmatrix}(\mathbf{x}_i - \mathbf{x}_i')^\top\beta \\ \mathbf{z}_i - \mathbf{z}_i'\end{bmatrix}\right| \\
&\leq \sup_{\beta\in\mathbb{B}_2(r)}\left|\frac{1}{n}\sum_{i=1}^n (\|\beta\|_\infty \vee 1)\max\{|\nabla_1\ell(\tilde{\mathbf{x}}_i^\top\beta, \tilde{\mathbf{z}}_i)|, |\nabla_2\ell(\tilde{\mathbf{x}}_i^\top\beta, \tilde{\mathbf{z}}_i)|\}\|\mathbf{u}_i - \mathbf{u}_i'\|_1\right|
\end{aligned}$$

where the first equality uses mean value theorem, the second inequality uses Hölder's inequality, $\tilde{\mathbf{x}}_i^\top\beta \in [\mathbf{x}_i^\top\beta, \mathbf{x}_i'^\top\beta]$ and $\tilde{\mathbf{z}}_i \in [\mathbf{z}_i, \mathbf{z}_i']$, and ∇_i refers to a derivative with respect

to ith argument. We note that $|\nabla_1 \ell(\tilde{\mathbf{x}}_i^\top \beta, \tilde{\mathbf{z}}_i)| = |\ell'(\tilde{\mathbf{x}}_i^\top \beta, \tilde{\mathbf{z}}_i)| \leq C_\ell$ a.s. by Assumption **A2**. On the other hand, for all $\beta \in \mathbb{B}_2(r)$, $|\nabla_2 \ell(\tilde{\mathbf{x}}_i^\top \beta, \tilde{\mathbf{z}}_i)| = |A(h(\tilde{\mathbf{x}}_i^\top \beta)) - h(\tilde{\mathbf{x}}_i^\top \beta)| \leq C_{\ell,2}$, for $C_{\ell,2} := \sup_{t;|t|\leq 2rC_X\sqrt{p}} |A(h(t)) - h(t)|$ since $|\tilde{\mathbf{x}}_i^\top \beta| \leq |\mathbf{x}_i^\top \beta| + |\mathbf{x}_i'^\top \beta| \leq 2rC_X\sqrt{p}$ by Assumption **A1'**. Then,

$$\begin{aligned}|\phi(\mathbf{u}_1, \ldots, \mathbf{u}_n) - \phi(\mathbf{u}_1', \ldots, \mathbf{u}_n')| &\leq \frac{1}{n}(r \vee 1) \cdot (C_\ell \vee C_{\ell,2}) \sum_{i=1}^n \|\mathbf{u}_i - \mathbf{u}_i'\|_1 \\ &= \frac{L_0}{n} \rho^n(\mathbf{u}_i, \mathbf{u}_i')\end{aligned}$$

where $L_0 := (r \vee 1) \cdot (C_\ell \vee C_{\ell,2})$, and ρ^n is an ℓ_1 product metric for $\rho_i(x, x') = \|x - x'\|_1$. In particular, $\phi(\cdot)$ is L_0/n Lipschitz. Then by Lemma 20,

$$\mathbb{P}\left(\sup_{\beta\in\mathbb{B}_2(r)} \left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right| \geq \mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right|\right] + t\right) \leq \exp(-\frac{t^2 n^2}{2L_0^2 \sum_{i=1}^n \Delta_{SG}^2(\mathcal{X}_i)}), \tag{46}$$

provided that $\Delta_{SG}^2(\mathcal{X}_i) < \infty$. Now we calculate $\Delta_{SG}^2(\mathcal{X}_i)$. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[\exp(\lambda\sigma_i \|\mathbf{u}_i - \mathbf{u}_i'\|_1)] &= \mathbb{E}[\exp(\lambda\sigma_i \{\sum_{j=1}^p |\mathbf{x}_{ij} - \mathbf{x}_{ij}'| + |\mathbf{z}_i - \mathbf{z}_i'|\})] \\ &= \mathbb{E}[\exp(\lambda\sigma_i \sum_{j=1}^p |\mathbf{x}_{ij} - \mathbf{x}_{ij}'|)\mathbb{E}[\exp(\lambda\sigma_i |\mathbf{z}_i - \mathbf{z}_i'|)|\mathbf{x}_i]] \\ &\leq \exp(\lambda^2(pK_X^2 + K_Z^2)),\end{aligned}$$

thus $\Delta_{SG}^2(\mathcal{X}_i) = 2(pK_X^2 + K_Z^2)$. Then for $t \geq \sqrt{2}L_0(K_X + K_Z)\sqrt{\frac{p\log(1/\delta)}{n}}$, the probability of the LHS is bounded below by $1 - \delta$.

We bound the expectation term by the standard arguments using symmetrization and contraction inequality. We have,

$$\begin{aligned}\mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right|\right] &= \mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \left|\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i^\top \beta, \mathbf{z}_i) - \mathbb{E}[\ell(\mathbf{x}_i^\top \beta, \mathbf{z}_i)]\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \ell(\mathbf{x}_i^\top \beta, \mathbf{z}_i)\right|\right],\end{aligned}$$

where we let $(\sigma_i)_{i=1}^n$ be i.i.d. Rademacher variables independent from $(\mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$. Since $|\ell(t, \mathbf{z}_i) - \ell(s, \mathbf{z}_i)| \leq C_\ell |t - s|$ a.s. by Assumption **A2**, contraction inequality gives

$$\begin{aligned}\mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right|\right] &\leq 4C_\ell \mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{x}_i^\top \beta\right|\right] + 2\mathbb{E}\left[\frac{1}{n}\left|\sum_{i=1}^n \sigma_i \ell(0, \mathbf{z}_i)\right|\right] \\ &\leq 4C_\ell \mathbb{E}\left[\sup_{\beta\in\mathbb{B}_2(r)} \|\beta\|_1 \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{x}_i\right\|_\infty\right] + 2\mathbb{E}\left[\frac{1}{n}\left|\sum_{i=1}^n \sigma_i \ell(0, \mathbf{z}_i)\right|\right] \\ &\leq 4rC_\ell\sqrt{p}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{x}_i\right\|_\infty\right] + 2\mathbb{E}\left[\frac{1}{n}\left|\sum_{i=1}^n \sigma_i \ell(0, \mathbf{z}_i)\right|\right].\end{aligned}$$

Since $|\sigma_i| \leq 1$ a.s. and $\mathbb{E}[\sigma_i \mathbf{x}_{ij}] = 0$, $\sigma_i \mathbf{x}_{ij}$ is mean-zero cK_X sub-gaussian with some absolute constant $c > 0$, $\forall i, j$. As $(\sigma_i \mathbf{x}_{ij})_{i=1}^n$ are independent for any j, $\sum_{i=1}^n \sigma_i \mathbf{x}_{ij}/n$ is also sub-gaussian with parameter $cK_X/\sqrt{n}$. Similarly, $\sigma_i \ell(0, \mathbf{z}_i)$ is mean-zero $c' K_Z$ sub-gaussian where c' only depends on (A, g), and $\sum_{i=1}^n \sigma_i \ell(0, \mathbf{z}_i)/n$ is sub-gaussian with parameter $c' K_Z/\sqrt{n}$. Therefore, by the bound on the maximum of sub-gaussian variables,

$$\mathbb{E}\left[\sup_{\beta \in \mathbb{B}_2(r)} \left|\mathcal{L}_n^\ell(\beta) - \mathcal{L}^\ell(\beta)\right|\right] \leq c'' r(K_X \vee K_Z) C_\ell \sqrt{\frac{p \log p}{n}}. \tag{47}$$

where c'' is a constant depending only on the choice of model (A, g). Combining (46) and (47), we obtain the desired inequality. ■

Proof [Proof of Lemma 18| We verify Assumptions 1-3 in Mei et al. (2018). The first assumption is to verify whether the gradient of the loss has a sub-Gaussian tail. The second assumption is to show that the Hessian evaluated on a unit vector is sub-Exponential. The third assumption is about the Lipschitz continuity of the Hessian. We mainly check whether quantities in interest satisfy a sub-gaussian/exponential moment bounds.

A1 For any $u \in \mathbb{R}^p$ such that $\|u\|_2 = 1$, $\langle \ell'(\mathbf{x}^\top \beta, \mathbf{z})\mathbf{x}, u\rangle$ is sub-gaussian since $\mathbb{E}(|\ell'(\mathbf{x}^\top \beta, \mathbf{z})\mathbf{x}^\top u|^k)^{1/k} \leq \|\ell'\|_\infty \mathbb{E}[|\mathbf{x}^\top u|^k]^{1/k} \leq C_\ell K_X \sqrt{k}$ for any $k \geq 1$.

A2 Similarly for any $u \in \mathbb{R}^p$ such that $\|u\|_2 = 1$ $\langle u, \ell''(\mathbf{x}^\top \beta, \mathbf{z})\mathbf{x}\mathbf{x}^\top u\rangle$ is sub-exponential since
$\mathbb{E}[|\ell''(\mathbf{x}^\top \beta, \mathbf{z})(\mathbf{x}^\top u)^2|^k]^{1/k} \leq 2C_\ell \mathbb{E}[(\mathbf{x}^\top u)^{2k}]^{1/k} \leq 4C_\ell K_X^2 k$ for any $k \geq 1$.

A3 $\|\nabla^2 \mathcal{L}^\ell(\beta_0)\|_2 = \sup_{u;\|u\|_2=1} \mathbb{E}[\rho_I(\mathbf{x}^\top \beta_0)(\mathbf{x}^\top u)^2] \leq 2C_\ell K_X^2$. Also from the Lipschitz continuity assumption of ℓ'',

$$\begin{aligned}\mathbb{E}\left[\sup_{\beta_1 \neq \beta_2} \frac{\|\nabla^2 \mathcal{L}^\ell(\beta_1) - \nabla^2 \mathcal{L}^\ell(\beta_2)\|_2}{\|\beta_1 - \beta_2\|_2}\right] &= \mathbb{E}\left[\sup_{\substack{\beta_1 \neq \beta_2, \\ u;\|u\|_2=1}} \frac{|\ell''(\mathbf{x}^\top \beta_1, \mathbf{z}) - \ell''(\mathbf{x}^\top \beta_2, \mathbf{z})|(\mathbf{x}^\top u)^2}{\|\beta_1 - \beta_2\|_2}\right] \\ &\leq L_\ell \mathbb{E}\left[\sup_{\substack{\beta_1 \neq \beta_2, \\ u;\|u\|_2=1}} \frac{|\mathbf{x}^\top \beta_1 - \mathbf{x}^\top \beta_2|(\mathbf{x}^\top u)^2}{\|\beta_1 - \beta_2\|_2}\right].\end{aligned}$$

By Cauchy-Schwarz, $|\mathbf{x}^\top(\beta_1 - \beta_2)| \leq \|\mathbf{x}\|_2 \|\beta_1 - \beta_2\|_2$ and $(\mathbf{x}^\top u)^2 \leq \|\mathbf{x}\|_2^2$ since $\|u\|_2 = 1$. Thus

$$\mathbb{E}\left[\sup_{\beta_1 \neq \beta_2} \frac{\|\nabla^2 \mathcal{L}^\ell(\beta_1) - \nabla^2 \mathcal{L}^\ell(\beta_2)\|_2}{\|\beta_1 - \beta_2\|_2}\right] \leq L_\ell \mathbb{E}\left[\|\mathbf{x}\|_2^3\right] \leq 3^{3/2} L_\ell K_X^3 p^{3/2},$$

since $\mathbb{E}[\|\mathbf{x}\|_2^3] = \mathbb{E}[(\sum_{i=1}^p x_i^2)^{3/2}] \leq p^{1/2}\mathbb{E}[(\sum_{i=1}^p |x_i|^3)] \leq 3^{3/2} K_X^3 p^{3/2}$. ■

A.4 Proof of Corollary 7

For a **(GLM)** with parameters $(\log(1+\exp(\cdot)), g_{LN})$ with $\mathbf{z}_i \in \{0,1\}$, we first note that the sub-gaussian tail condition for $\mathbf{z}_i$ is satisfied with $K_Z = 1$ since $|\mathbf{z}_i - \mathbb{E}[\mathbf{z}_i|\mathbf{x}_i]| \leq 1$, almost surely. Now we show that **A2** is satisfied. Then the result follows from the Proposition 6.

From (11) and (12), we have

$$\ell'(t,z) = \left(A'(h_{LN}(t)) - z\right) h'_{LN}(t)$$
$$\ell''(t,z) = \rho_I(t) + \rho_R(t,z),$$

for $A(t) = \log(1+\exp(t))$ and $\rho_I(t)$ and $\rho_R(t,z)$ such that

$$\rho_I(t) = A''(h_{LN}(t))h'_{LN}(t)^2, \quad \text{and} \quad \rho_R(t,z) = (A'(h_{LN}(t)) - z)h''_{LN}(t).$$

From Lemma 21 which is presented at the end of this subsection, $\|h'_{LN}\|_\infty \leq 1$ and $\|h''_{LN}\|_\infty \leq 2$. Also $A''(t) = e^t/(1+e^t)^2$ is bounded by $1/4$ and $|A'(h_{LN}(t)) - z| \leq 1$ for any t and $z \in \{0,1\}$, since $0 \leq A'(h_{LN}(t)) \leq 1, \forall t$. Thus

$$|\ell'(t,z)| \leq 1, \quad |\rho_I(t)| \leq \frac{1}{4}, \quad \text{and} \quad |\rho_R(t,z)| \leq \|h''_{LN}\|_\infty \leq 2, \quad \forall z \in \{0,1\}, \forall t,$$

and $\max\{\|\ell'\|_\infty, \|\rho_I\|_\infty, \|\rho_R\|_\infty\}$ is bounded by 2.

To verify that ℓ'' is L_ℓ-Lipschitz where L_ℓ does not depend on t, it is sufficient to show that the gradients of ρ_I and ρ_R are bounded independent of t. By calculation, we have

$$\rho'_I(t) = A'''(h_{LN}(t))h'_{LN}(t)^3 + 2A''(h_{LN}(t))h'_{LN}(t)^2 h''_{LN}(t)$$
$$\rho'_R(t,z) = A''(h_{LN}(t))h'_{LN}(t)h''_{LN}(t) + \{A'(h_{LN}(t)) - z\}h'''_{LN}(t).$$

We bound each term separately. As other terms can be bounded similarly other than the term involving A''', it is sufficient to show that $A'''(t)$ is bounded by an absolute constant. We have,

$$|A'''(t)| = \left|\frac{e^t}{(1+e^t)^2} - \frac{2e^{2t}}{(1+e^t)^3}\right| \leq \left|\frac{e^t}{(1+e^t)^2}\right| + \left|\left(\frac{2e^t}{(1+e^t)^2}\right)\left(\frac{e^t}{1+e^t}\right)\right| \leq \frac{1}{4} + \frac{1}{2} \leq 1.$$

Finally, we present the Lemma about the boundedness of h'_{LN}, h''_{LN} and h'''_{LN}.

Lemma 21 *There exists $C \leq 7$ such that $\max\{\|h'_{LN}\|_\infty, \|h''_{LN}\|_\infty, \|h'''_{LN}\|_\infty\} \leq C$ for $h_{LN} = (A')^{-1} \circ g_{LN}^{-1}$.*

Proof From the definition of g_{LN} and h_{LN} in (9), we have

$$h_{LN}(t) := \log\left(\frac{(1-\rho_1-\rho_0)\mu(t) + \rho_0}{1 - (1-\rho_1-\rho_0)\mu(t) - \rho_0}\right).$$

Let $a = 1-\rho_1-\rho_0$ and $b = \rho_0$. We have $a\mu(t) + b \leq a + b < 1$ and $a\mu(t) \leq a\mu(t) + b$ for any t. Then $\forall t$,

$$\frac{a\mu(t)}{a\mu(t)+b} \leq 1 \quad \text{and} \quad \frac{a(1-\mu(t))}{1-a\mu(t)-b} < 1 \tag{48}$$

By definition of $h_{LN}(t) = \log(a\mu(t)+b) - \log(1-a\mu(t)-b)$,

$$\begin{aligned} h'_{LN}(t) &= \frac{d}{d\mu(t)} \log\left(\frac{a\mu(t)+b}{1-(a\mu(t)+b)}\right) \frac{d\mu(t)}{dt} \\ &= \frac{a\mu(t)(1-\mu(t))}{(a\mu(t)+b)(1-a\mu(t)-b)} \leq 1 \end{aligned}$$

by the fact that

$$\frac{d\mu(t)}{dt} = A''(t) = \mu(t)(1-\mu(t))$$

and the inequalities (48). In particular, $h'_{LN} \geq 0$ and $\|h'_{LN}\|_\infty \leq 1$.

Now we bound h''_{LN}. From elementary calculation, it can be shown that

$$\begin{aligned} h''_{LN}(t) &= h'_{LN}(t)(1-2\mu(t)) - h'_{LN}(t)^2(1-2(a\mu(t)+b)), \\ h'''_{LN}(t) &= h''_{LN}(t)\left\{1-2\mu(t)-2h'_{LN}(t)(1-2(a\mu(t)+b))\right\} \\ &\quad - 2\mu(t)(1-\mu(t))h'_{LN}(t)(1-ah'_{LN}(t)). \end{aligned}$$

In particular,

$$|h''_{LN}(t)| \leq h'_{LN}(t)|1-2\mu(t)| + h'_{LN}(t)^2|1-2(a\mu(t)+b)| \leq 2\|h'_{LN}\|_\infty \leq 2$$

since $\max_{0\leq\mu\leq 1}|1-2\mu| = 1$ and $0 \leq \mu(t), a\mu(t)+b \leq 1$, for all t. Also,

$$\begin{aligned} |h'''_{LN}(t)| &\leq |h''_{LN}(t)|\left\{|1-2\mu(t)| + 2h'_{LN}(t)|1-2(a\mu(t)+b)|\right\} \\ &\quad + 2\mu(t)(1-\mu(t))h'_{LN}(t)(1-ah'_{LN}(t)) \\ &\leq 3\|h''_{LN}\|_\infty + \frac{1}{2}. \end{aligned}$$

■

Appendix B. Proofs for Results in Section 5

B.1 Proof of Proposition 9

First, we note that the inequality (23) holds trivially for $\beta = \beta_0$. For any $\beta \in \mathbb{B}_2(r) \setminus \{\beta_0\}$ and $\Delta_0 := \beta - \beta_0$,

$$\begin{aligned} &\langle \nabla\mathcal{L}_n^\ell(\beta) - \nabla\mathcal{L}_n^\ell(\beta_0), \Delta_0\rangle \\ &= \langle \nabla\mathcal{L}^\ell(\beta) - \nabla\mathcal{L}^\ell(\beta_0), \Delta_0\rangle + \langle \nabla\mathcal{L}_n^\ell(\beta) - \nabla\mathcal{L}^\ell(\beta), \Delta_0\rangle + \langle \nabla\mathcal{L}^\ell(\beta_0) - \nabla\mathcal{L}_n^\ell(\beta_0), \Delta_0\rangle \\ &\geq \langle \nabla\mathcal{L}^\ell(\beta), \Delta_0\rangle - \left(\left|\frac{\langle \nabla\mathcal{L}_n^\ell(\beta) - \nabla\mathcal{L}^\ell(\beta), \Delta_0\rangle}{\|\Delta_0\|_1}\right| + \|\nabla\mathcal{L}_n^\ell(\beta_0)\|_\infty\right)\|\Delta_0\|_1 \end{aligned}$$

using $\nabla\mathcal{L}^\ell(\beta_0) = 0$ and Hölder's inequality. From (45), we have

$$\langle \nabla\mathcal{L}^\ell(\beta), \Delta_0\rangle \geq \alpha_\ell\|\Delta_0\|_2^2,$$

where $\alpha_\ell := C_r C_\lambda/2$, for C_r defined in (45). Let

$$\mathcal{E} := \left\{ \left| \frac{\langle \nabla \mathcal{L}_n^\ell(\beta) - \nabla \mathcal{L}^\ell(\beta), \Delta_0 \rangle}{\|\Delta_0\|_1} \right| + \|\nabla \mathcal{L}_n^\ell(\beta_0)\|_\infty \leq \tau_\ell \sqrt{\frac{\log p}{n}}, \forall \beta \in \mathbb{B}_2(r) \setminus \{\beta_0\} \right\}, \quad (49)$$

for $\tau_\ell := c \cdot (C_\ell K_X + C_1(T_\star + L_\star \tau))$, where $c > 0$ is an absolute constant, and $T_\star, L_\star, \tau$, and C_1 are constants which are defined in Lemma 23. On $\mathcal{E}$, we note that

$$\langle \nabla \mathcal{L}_n^\ell(\beta) - \nabla \mathcal{L}_n^\ell(\beta_0), \beta - \beta_0 \rangle \geq \alpha_\ell \|\beta - \beta_0\|_2^2 - \tau_\ell \sqrt{\frac{\log p}{n}} \|\beta - \beta_0\|_1.$$

Therefore, it is sufficient to show that $\mathbb{P}(\mathcal{E}) \geq 1 - \epsilon$.

First, we show $\|\nabla \mathcal{L}_n^\ell(\beta_0)\|_\infty \leq c_0 C_\ell K_X \sqrt{\frac{\log p}{n}}$ w.p $1 - \epsilon/2$, where c_0 is an absolute constant. Note,

$$\|\nabla \mathcal{L}_n^\ell(\beta_0)\|_\infty = \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \ell'(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) \mathbf{x}_{ij} \right|.$$

We use the following Lemma 22 to bound $\|\nabla \mathcal{L}_n^\ell(\beta_0)\|_\infty$.

Lemma 22 *Suppose $(\xi_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ are random variables such that ξ_{ij} is a mean-zero sub-gaussian with parameter C_ξ and $(\xi_{ij})_{i=1}^n$ are independent for any $j \in \{1, \ldots, p\}$. Then,*

$$\mathbb{P}\left(\max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \xi_{ij} \right| \geq 3 C_\xi \sqrt{\frac{\log p}{n}} \right) \leq \frac{1}{p^7}.$$

Proof $\frac{1}{n} \sum_{i=1}^n \xi_{ij}$ is sub-gaussian with parameter $C_\xi/\sqrt{n}$. By taking a union bound, for any $t \geq 0$ we have

$$\mathbb{P}\left(\max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \xi_{ij} \right| \geq t \sqrt{\frac{\log p}{n}} \right) \leq \exp(-t^2 \log p / C_\xi^2 + \log 2p)$$

Take $t^2 = 9 C_\xi^2$. Then $\|\nabla \mathcal{L}_n(\beta_0)\|_\infty \leq t \sqrt{\frac{\log p}{n}}$ with probability at least $1 - 1/p^7$. ■

Taking $\xi_{ij} = \ell'(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) \mathbf{x}_{ij}$ in Lemma 22, we have, $\mathbb{E}[|\xi_{ij}|^k]^{1/k} \leq C_\ell \mathbb{E}[|\mathbf{x}_{ij}|^k]^{1/k} \leq \sqrt{k} C_\ell K_X$ for any $k \geq 1$ by Assumption **A1'** and **A2**. Also, $\mathbb{E}[\xi_{ij}] = 0$ since $\mathbb{E}[\mathbf{z}_i | \mathbf{x}_i] = A'(h(\mathbf{x}_i^\top \beta_0))$. Therefore, ξ_{ij} is a mean-zero sub-gaussian variable with parameter $c_0 C_\ell K_X$, where c_0 is an absolute constant. Then from Lemma 22, $\|\nabla \mathcal{L}_n(\beta_0)\|_\infty \leq c_0 C_\ell K_X \sqrt{\frac{\log p}{n}}$ with probability at least $1 - 1/p^7$. Thus, for $n \geq C \cdot (2/\epsilon)^{1/7}$ for a sufficiently large constant C, we have $\|\nabla \mathcal{L}_n(\beta_0)\|_\infty \leq c_0 C_\ell K_X \sqrt{\frac{\log p}{n}}$ w.p. at least $1 - \epsilon/2$, in the regime of interest $p \gg n$.

The bound for the second term can be obtained by taking advantage of the uniform convergence result of the directional derivative of the loss function in Mei et al. (2018), and we summarize the result for the case of $\mathcal{L}_n(\beta)$ in Lemma 23. Taking $\delta = \epsilon/2$ in Lemma 23, we obtain $\mathbb{P}(\mathcal{E}) \geq 1 - \epsilon$, as desired.

Lemma 23 (Theorem 3 in Mei et al., 2018) *There exists a constant $C_1 > 0$, which depends on model parameters (r, K_X, C_ℓ, L_ℓ) and δ such that*

$$\mathbb{P}\left(\sup_{\beta\in\mathbb{B}_2(r)\setminus\{\beta_0\}}\left|\frac{\langle\nabla\mathcal{L}_n^\ell(\beta)-\nabla\mathcal{L}^\ell(\beta),\beta-\beta_0\rangle}{\|\beta-\beta_0\|_1}\right|\le C_1(T_\star+L_\star\tau)\sqrt{\frac{\log(np)}{n}}\right)\ge 1-\delta, \quad (50)$$

where $\tau = K_X\max\{C_\ell, L_\ell^{1/3}\}$, $T_\star = C_\ell C_X$, and $L_\star = C_\rho + C_\ell(C_bC_d+1)$.

Proof [Proof of Lemma 23] We verify Assumptions 2-5 in Mei et al. (2018). From the proof of Lemma 18, we have already checked that Assumption 2 and 3 in Mei et al. (2018) are satisfied under **A1'** and **A2**. We thus check Assumptions 4 and 5 in Mei et al. (2018), which verify the existence of $T_\star$ and $L_\star$, where $T_\star$ is a constant such that $\|\ell'(\mathbf{x}^\top\beta,\mathbf{z})\mathbf{x}\|_\infty \le T_\star$ a.s., and $L_\star$ is a Lipschitz constant for the function $g(\cdot,\cdot)\to\mathbb{R}$, which is defined as follows:

$$g(\mathbf{x}^\top(\beta-\beta_0),(\mathbf{x},\mathbf{z})) = \langle\ell'(\mathbf{x}^\top\beta,\mathbf{z})\mathbf{x},\beta-\beta_0\rangle.$$

For the existence of $T_\star$, $\|\ell'(\mathbf{x}^\top\beta,\mathbf{z})\mathbf{x}\|_\infty \le C_\ell C_X$, by Assumption **A1'** and **A2**. Thus we can let $T_\star = C_\ell C_X$. For Assumption 5 in Mei et al. (2018),

$$\begin{aligned}\langle\ell'(\mathbf{x}^\top\beta,\mathbf{z})\mathbf{x},\beta-\beta_0\rangle &= (A'(h(\mathbf{x}^\top\beta))-\mathbf{z})h'(\mathbf{x}^\top\beta)\mathbf{x}^\top(\beta-\beta_0)\\ &= g(\mathbf{x}^\top(\beta-\beta_0);(\mathbf{x},\mathbf{z}))\end{aligned}$$

for $g(t;(\mathbf{x},\mathbf{z})) := \ell'(t+\mathbf{x}^\top\beta_0,\mathbf{z})t$. We show that $g(t;(\mathbf{x},\mathbf{z}))$ is Lipschitz with respect to t under Assumption **A3**. Taking a derivative with respect to t,

$$g'(t;(\mathbf{x},\mathbf{z})) = \ell''(t+\mathbf{x}^\top\beta_0,\mathbf{z})t + \ell'(t+\mathbf{x}^\top\beta_0,\mathbf{z}).$$

Then,

$$\begin{aligned}|g'(t;(\mathbf{x},\mathbf{z}))| &\le |\ell''(t+\mathbf{x}^\top\beta_0,\mathbf{z})(t+\mathbf{x}^\top\beta_0)| + |\ell''(t+\mathbf{x}^\top\beta_0,\mathbf{z})\mathbf{x}^\top\beta_0| + |\ell'(t+\mathbf{x}^\top\beta_0,\mathbf{z})|\\ &\le C_\rho + C_\ell(C_bC_d+1)\end{aligned}$$

by Assumptions **A1'** and **A2**, noting $|\mathbf{x}^\top\beta_0| \le C_bC_d$ a.s. by Assumption **A4**. Therefore $L_\star$ can be taken as $L_\star = C_\rho + C_\ell(C_bC_d+1)$. ■

B.2 Proof of Proposition 10

$$\begin{aligned}\langle\nabla\mathcal{L}_n^s(\beta)-\nabla\mathcal{L}_n^s(\beta_0),\beta-\beta_0\rangle &= \frac{1}{n}\sum_{i=1}^n(\mu(\mathbf{x}_i^\top\beta)-\mu(\mathbf{x}_i^\top\beta_0))\mathbf{x}_i^\top(\beta-\beta_0)\\ &= \frac{1}{n}\sum_{i=1}^n\mu'(\mathbf{x}_i^\top\beta_0+v\mathbf{x}_i^\top(\beta-\beta_0))(\mathbf{x}_i^\top(\beta-\beta_0))^2\end{aligned}$$

Then from the proof of Proposition 2 in Negahban et al. (2012), there exist positive constants κ_1 and κ_2 such that

$$\langle\nabla\mathcal{L}_n^s(\beta)-\nabla\mathcal{L}_n^s(\beta_0),\beta-\beta_0\rangle \ge \kappa_1\|\Delta\|_2\left(\|\Delta\|_2-\kappa_2\sqrt{\frac{\log p}{n}}\|\Delta\|_1\right), \quad \forall\beta\in\mathbb{B}_2(1;\beta_0)$$

with probability at least $1-c_1\exp(-c_2n)$, for some $c_1, c_2 > 0$. The result (24) follows from the basic arithmetic inequality $2ab \le (a+b)^2$.

B.3 Proof of Theorem 11

First, we address the ℓ_1 and ℓ_2 error bounds for the non-convex estimator. We characterize the ℓ_1 and ℓ_2 error bounds of a stationary point following similar lines as in the proof of Theorem 1 in Loh (2017), which established the result with a different tolerance function and penalty. Since β_0 is feasible, by the first order optimality condition, we have the following inequality

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}_\ell^H) + \lambda v(\widetilde{\beta}_\ell^H), \beta_0 - \widetilde{\beta}_\ell^H \rangle \geq 0,$$

for $v(\widetilde{\beta}_\ell^H) \in \partial \|\widetilde{\beta}_\ell^H\|_1$. We let $\widetilde{\Delta} := \widetilde{\beta}_\ell^H - \beta_0$. By applying RSC condition (23),

$$\alpha_\ell \|\widetilde{\Delta}\|_2^2 - \tau_\ell \sqrt{\frac{\log p}{n}} \|\widetilde{\Delta}\|_1 + \langle \nabla \mathcal{L}_n(\beta_0) + \lambda v(\widetilde{\beta}_\ell^H), \widetilde{\beta}_\ell^H - \beta_0 \rangle \leq 0,$$

By convexity of $\|\cdot\|_1$,

$$\lambda \|\beta_0\|_1 - \lambda \|\widetilde{\beta}_\ell^H\|_1 \geq -\lambda v(\widetilde{\beta}_\ell^H)^\top \widetilde{\Delta}. \tag{51}$$

Therefore,

$$\alpha_\ell \|\widetilde{\Delta}\|_2^2 - \tau_\ell \sqrt{\frac{\log p}{n}} \|\widetilde{\Delta}\|_1 + \langle \nabla \mathcal{L}_n(\beta_0), \widetilde{\beta}_\ell^H - \beta_0 \rangle + \lambda(\|\widetilde{\beta}_\ell^H\|_1 - \|\beta_0\|_1) \leq 0,$$

That is,

$$\begin{aligned}
\alpha_\ell \|\widetilde{\Delta}\|_2^2 &\leq \tau_\ell \sqrt{\frac{\log p}{n}} \|\widetilde{\Delta}\|_1 + |\langle \nabla \mathcal{L}_n(\beta_0), \widetilde{\beta}_\ell^H - \beta_0 \rangle| + \lambda(\|\beta_0\|_1 - \|\widetilde{\beta}_\ell^H\|_1) \\
&\leq \tau_\ell \sqrt{\frac{\log p}{n}} \|\widetilde{\Delta}\|_1 + \|\nabla \mathcal{L}_n(\beta_0)\|_\infty \|\widetilde{\Delta}\|_1 + \lambda(\|\widetilde{\Delta}_S\|_1 - \|(\widetilde{\beta}_\ell^H)_{S^c}\|_1)
\end{aligned}$$

Since $\tau_\ell \sqrt{\frac{\log p}{n}} + \|\nabla \mathcal{L}_n(\beta_0)\|_\infty \leq \frac{\lambda}{2}$,

$$\begin{aligned}
\alpha_\ell \|\widetilde{\Delta}\|_2^2 &\leq \frac{\lambda}{2}(\|\widetilde{\Delta}_S\|_1 + \|\widetilde{\Delta}_{S^c}\|_1) + \lambda(\|\widetilde{\Delta}_S\|_1 - \|(\widetilde{\beta}_\ell^H)_{S^c}\|_1) \\
&= \frac{3\lambda}{2} \|\widetilde{\Delta}_S\|_1 - \frac{\lambda}{2} \|\widetilde{\Delta}_{S^c}\|_1.
\end{aligned}$$

In particular,

$$\alpha_\ell \|\widetilde{\Delta}\|_2^2 \leq \frac{3\lambda}{2} \|\widetilde{\Delta}_S\|_1 \leq \frac{3\sqrt{s_0}\lambda}{2} \|\widetilde{\Delta}\|_2, \tag{52}$$

$$\|\widetilde{\Delta}_{S^c}\|_1 \leq 3\|\widetilde{\Delta}_S\|_1 \tag{53}$$

ℓ_2 bound follows from (52) and

$$\|\widetilde{\Delta}\|_1 = \|\widetilde{\Delta}_S\|_1 + \|\widetilde{\Delta}_{S^c}\|_1 \leq 4\|\widetilde{\Delta}_S\|_1 \leq 4\sqrt{s_0} \|\widetilde{\Delta}\|_2.$$

Now, we address the ℓ_1 and ℓ_2 error bounds for the convex estimator. To do so, we need to establish a different RSC condition, introduced by Negahban et al. (2012) as follows:

Definition 24 (restricted strong convexity in Negahban et al., 2012) *For a given set $\mathbb{S}$, the loss function $\mathcal{L}_n$ satisfies restricted strong convexity (RSC) with parameter $\alpha > 0$ if*

$$\mathcal{L}_n(\beta) - \mathcal{L}_n(\beta_0) - \langle \nabla \mathcal{L}_n(\beta_0), \beta - \beta_0 \rangle \geq \alpha \|\beta - \beta_0\|_2^2 \quad \text{for all } \beta - \beta_0 \in \mathbb{S}. \tag{54}$$

In the following Lemma 25, we show that the RSC condition 8 with $\tau_{n,p}(t) = \tau(\log p/n)t^2$ and $\Omega = \mathbb{B}_2(\delta; \beta_0)$ implies the RSC condition in Negahban et al. (2012).

Lemma 25 *The RSC condition 8 with $\tau_{n,p}(t) = \tau(\log p/n)t^2$ and $\Omega = \mathbb{B}_2(\delta; \beta_0)$ implies (54) with parameter $\alpha/4$ and*

$$\mathbb{S} = \{\Delta \in \mathbb{R}^p; \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\} \cap \{\Delta \in \mathbb{R}^p; \|\Delta\|_2 \leq \delta\},$$

where $S \subseteq \{1, \ldots, p\}$ is the support of β_0 and $s_0 := |S|$, given the sample size $n \geq (32\tau s_0/\alpha) \log p$.

Provided that Lemma 25 is true and given the condition of λ_s in Theorem 11, the ℓ_2 error bound

$$\|\widehat{\beta}_s^H - \beta_0\|_2 \leq \frac{8\sqrt{s_0}\lambda_s}{\alpha_s} \tag{55}$$

can be obtained by applying Theorem 1 in Negahban et al. (2012). Also it is well known that an error vector $\widehat{\beta} - \beta_0$, where $\widehat{\beta}$ is a solution of Lasso optimization problem, belongs to the cone $\{\Delta \in \mathbb{R}^p; \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$. Thus $\|\widehat{\beta}_s^H - \beta_0\|_1 \leq 4\|(\widehat{\beta}_s^H - \beta_0)_S\|_1 \leq 4\sqrt{s_0}\|\widehat{\beta}_s^H - \beta_0\|_2$. Applying this inequality to (55) gives an ℓ_1 bound. Now we present the proof of Lemma 25.

Proof [Proof of Lemma 25] For any β such that $\beta - \beta_0 \in \mathbb{S}$, we have,

$$\begin{aligned}
\mathcal{L}_n(\beta) &= \mathcal{L}_n(\beta_0) + \int \nabla \mathcal{L}_n(\beta_0 + t(\beta - \beta_0))^\top (\beta - \beta_0) dt \\
&= \mathcal{L}_n(\beta_0) + \nabla \mathcal{L}_n(\beta_0)^\top (\beta - \beta_0) + \int_0^1 \frac{1}{t} (\nabla \mathcal{L}_n(\beta_0 + t(\beta - \beta_0)) - \nabla \mathcal{L}_n(\beta_0))^\top t(\beta - \beta_0) dt.
\end{aligned} \tag{56}$$

By the RSC condition 8 with $\tau_{n,p}(t) = \tau(\log p/n)t^2$ and $\Omega = \mathbb{B}_2(\delta; \beta_0)$, for any $\beta \in \mathbb{B}_2(\delta; \beta_0)$ it holds that

$$(\nabla \mathcal{L}_n(\beta_0 + t(\beta - \beta_0)) - \nabla \mathcal{L}_n(\beta_0))^\top t(\beta - \beta_0) \geq t^2 \left(\alpha \|\beta - \beta_0\|_2^2 - \tau \left(\frac{\log p}{n} \right) \|\beta - \beta_0\|_1^2 \right). \tag{57}$$

Applying (57) to (56),

$$\begin{aligned}
\mathcal{L}_n(\beta) - \mathcal{L}_n(\beta_0) - \nabla \mathcal{L}_n(\beta_0)^\top (\beta - \beta_0) &\geq \int_0^1 t \left(\alpha \|\beta - \beta_0\|_2^2 - \tau \left(\frac{\log p}{n} \right) \|\beta - \beta_0\|_1^2 \right) dt \\
&= \frac{\alpha}{2} \|\beta - \beta_0\|_2^2 - \frac{\tau}{2} \left(\frac{\log p}{n} \right) \|\beta - \beta_0\|_1^2.
\end{aligned}$$

Since $\beta - \beta_0 \in \mathbb{S}$, $\|\beta - \beta_0\|_1 \leq 4\sqrt{s_0}\|\beta - \beta_0\|_2$. Therefore,

$$\mathcal{L}_n(\beta) - \mathcal{L}_n(\beta_0) - \nabla\mathcal{L}_n(\beta_0)^\top(\beta - \beta_0) \geq \left(\frac{\alpha}{2} - 8\tau s_0 \frac{\log p}{n}\right)\|\beta - \beta_0\|_2^2 \geq \frac{\alpha}{4}\|\beta - \beta_0\|_2^2$$

where the last inequality is from a given sample condition $n \geq (32\tau s_0/\alpha)\log p$. ■

B.4 Proof of Corollary 12

The Corollary 12 essentially follows from Proposition 9, Proposition 10, and Theorem 11. The main conditions to verify are $\|\nabla\mathcal{L}_n^\ell(\beta_0)\|_\infty, \|\nabla\mathcal{L}_n^s(\beta_0)\|_\infty = O(\sqrt{\frac{\log p}{n}})$ with high probability and Assumption **A3**, since we have already shown that Assumption **A2** is satisfied for the noisy labels problem in the proof of Corollary 7.

We first address bounds for $\|\nabla\mathcal{L}_n^\ell(\beta_0)\|_\infty$ and $\|\nabla\mathcal{L}_n^s(\beta_0)\|_\infty$. We note that $\|\nabla\mathcal{L}_n(\beta_0)\|_\infty$, for $\mathcal{L}_n \in (\mathcal{L}_n^\ell, \mathcal{L}_n^s)$, has the form

$$\|\nabla\mathcal{L}_n(\beta_0)\|_\infty = \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^n \xi_{ij}\right|,$$

where $\xi_{ij} = \{(A'(h_{LN}(\mathbf{x}_i^\top\beta_0)) - \mathbf{z}_i)h'_{LN}(\mathbf{x}_i^\top\beta_0)\}\mathbf{x}_{ij}$ if $\mathcal{L}_n = \mathcal{L}_n^\ell$, and $\xi_{ij} = \{A'(\mathbf{x}_i^\top\beta_0) - T(\mathbf{z}_i)\}\mathbf{x}_{ij}$ if $\mathcal{L}_n = \mathcal{L}_n^s$. Also $\mathbb{E}[\xi_{ij}] = 0, \forall i, j$ and $(\xi_{ij})_{i=1}^n$ are independent for any $j \in \{1, \ldots, p\}$.

From Lemma 21, we have $|(A'(h_{LN}(\mathbf{x}_i^\top\beta_0)) - \mathbf{z}_i)h'_{LN}(\mathbf{x}_i^\top\beta_0)| \leq 1$ and $|A'(\mathbf{x}_i^\top\beta_0) - T(\mathbf{z}_i)| \leq 1$ a.s. Thus $\mathbb{E}[|\xi_{ij}|^k]^{1/k} \leq \mathbb{E}[|\mathbf{x}_{ij}|^k]^{1/k} \leq \sqrt{k}K_X$ for any $k \geq 1$ by Assumption **A1'**. In particular, ξ_{ij} is mean-zero sub-gaussian with parameter cK_X where $c > 0$ is an absolute constant. Therefore, by Lemma 22, $\|\nabla\mathcal{L}_n(\beta_0)\|_\infty \leq c'K_X\sqrt{\frac{\log p}{n}}$ with probability at least $1 - 1/p^7$ for a different constant $c' > 0$.

Now we show that Assumption **A3** holds. We recall $\ell''(t, z) = \rho_I(t) + \rho_R(t, z)$ for $\rho_I(t) = A''(h_{LN}(t))h'_{LN}(t)^2$, $\rho_R(t, z) = (A'(h_{LN}(t)) - z)h''_{LN}(t)$ where $A(t) = \log(1+\exp(t))$ and $h_{LN}(\cdot)$ defined in Section 3.2. In the following, we show that both $\sup_t |\rho_I(t)t|$ and $\sup_t |\rho_R(t, z)t|$ are bounded by an absolute constant. First, we let $a = 1 - \rho_1 - \rho_0$, $b = \rho_0$. Since

$$h'_{LN}(t) = \frac{a\mu(t)(1-\mu(t))}{(a\mu(t)+b)(1-a\mu(t)-b)},$$

we have,

$$\begin{aligned}
t\rho_I(t) &= tA''(h_{LN}(t))h'_{LN}(t)^2 \\
&= t(a\mu(t)+b)(1-a\mu(t)-b)\left(\frac{a\mu(t)(1-\mu(t))}{(a\mu(t)+b)(1-a\mu(t)-b)}\right)^2 \\
&= at\mu(t)(1-\mu(t))h'_{LN}(t).
\end{aligned}$$

By Lemma 21, $\|h'_{LN}\|_\infty \leq 1$. Also, with an elementary calculation, it can be shown that

$$|t\mu(t)(1-\mu(t))| = \frac{|t|e^t}{(1+e^t)^2} \leq 2, \forall t.$$

Therefore, $\sup_t |\rho_I(t)t| \leq 2$. Now we address $\sup_t |\rho_R(t,z)t|$. Since $|A'(h_{LN}(t)) - z| \leq 1$ for all t and $z \in \{0,1\}$, we have,

$$|t\rho_R(t,z)| = |(A'(h_{LN}(t)) - z)h''(t)t| \leq |h''(t)t|.$$

Therefore, it is sufficient to bound $h''(t)t$. Note if $\rho_1 = \rho_0 = 0$, $h''(t) = 0$. Therefore $th''(t)$ is trivially bounded. Otherwise, we discuss three cases separately: 1. $\rho_1, \rho_0 > 0$, 2. $\rho_1 > 0, \rho_0 = 0$, and 3. $\rho_1 = 0, \rho_0 > 0$. First, we note from the proof of Lemma 21, we have,

$$th''_{LN}(t) = th'_{LN}(t)\{1 - 2\mu(t) - h'_{LN}(t)(1 - 2(a\mu(t) + b))\}.$$

Case 1: $\rho_1, \rho_0 > 0$
In this case, $\sup_t |th'_{LN}(t)| < \infty$, since

$$\sup_t |th'_{LN}(t)| \leq \frac{a \sup_t |t\mu(t)(1-\mu(t))|}{\inf_t (a\mu(t)+b)(1-a\mu(t)-b)} \leq \frac{2a}{\inf_{(\rho_0 \wedge \rho_1) \leq p \leq 1-(\rho_0 \wedge \rho_1)} x(1-x)} < \infty. \tag{58}$$

Also, $|1 - 2\mu(t) - h'_{LN}(t)(1 - 2(a\mu(t) + b)| \leq 1 + \|h'_{LN}\|_\infty \leq 2$.

For Case 2 and 3, we cannot use the bound (58) since the denominator becomes zero. With elementary calculations, we can obtain

$$th''_{LN}(t) = -\frac{a(1-b-a)t\mu(t)^2(1-\mu(t))}{(a\mu(t)+b)(1-a\mu(t)-b)^2} + \frac{abt\mu(t)(1-\mu(t))^2}{(a\mu(t)+b)^2(1-a\mu(t)-b)} \tag{59}$$

Case 2: $\rho_1 > 0, \rho_0 = 0$
Equivalently, $b = 0, a = 1 - \rho_1$, therefore the second term in (59) does not exist, and

$$th''_{LN}(t) = -\frac{(1-a)t\mu(t)(1-\mu(t))}{(1-a\mu(t))^2}$$

Therefore $|th''_{LN}(t)| \leq 2/(1-a) = 2/\rho_1$.
Case 3: $\rho_1 = 0, \rho_0 > 0$
In Case 3, $a = 1 - \rho_0, b = \rho_0, a + b = 1$. The first term in (59) does not exist, and

$$th''_{LN}(t) = \frac{bt\mu(t)(1-\mu(t))}{(a\mu(t)+b)^2},$$

noting $1 - a\mu(t) - b = a(1 - \mu(t))$. Therefore, $|th''_{LN}(t)| \leq 2/b = 2/\rho_0$.

B.5 Proof of Theorem 13

First, for a given ψ, we let $\widehat{\beta}^{\mathrm{db}} = \widehat{\beta}^{\mathrm{db}}(\psi)$, $\widehat{\Theta} = \widehat{\Theta}(\psi)$, and $\Theta = \Theta(\psi)$ for ease of notation. For any fixed $j \in \{1, \ldots, p\}$, we have

$$\widehat{\beta}^{\mathrm{db}}_j - \beta_{0j} = \widehat{\beta}_j - \beta_{0j} - \widehat{\Theta}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \psi(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\mathbf{x}_i \right). \tag{60}$$

Let $\widehat{\Delta} := \widehat{\beta} - \beta_0$. By the Taylor expansion,

$$\begin{aligned}
&\widehat{\Theta}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \psi(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i) \mathbf{x}_i \right) \\
&\quad = n^{-1} \sum_{i=1}^n \left(\psi(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) + \psi'(v_i, \mathbf{z}_i)(\mathbf{x}_i^\top \widehat{\beta} - \mathbf{x}_i^\top \beta_0) \right) \widehat{\Theta}_j^\top \mathbf{x}_i \\
&\quad = n^{-1} \sum_{i=1}^n \left(\psi(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) + \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i) \mathbf{x}_i^\top \widehat{\Delta} + \{\psi'(v_i, \mathbf{z}_i) - \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\} \mathbf{x}_i^\top \widehat{\Delta} \right) \widehat{\Theta}_j^\top \mathbf{x}_i
\end{aligned}$$

for v_i such that $|v_i - \mathbf{x}_i^\top \widehat{\beta}| \leq |\mathbf{x}_i^\top (\widehat{\beta} - \beta_0)|$.

First, we address the last term and show that it is $o_p(n^{-1/2})$.

$$\begin{aligned}
&n^{-1} \sum_{i=1}^n \{\psi'(v_i, \mathbf{z}_i) - \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\} \mathbf{x}_i^\top \widehat{\Delta} \widehat{\Theta}_j^\top \mathbf{x}_i \\
&\leq n^{-1} \sum_{i=1}^n |\psi'(v_i, \mathbf{z}_i) - \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)| |\mathbf{x}_i^\top \widehat{\Delta}| |\widehat{\Theta}_j^\top \mathbf{x}_i| \qquad (61)
\end{aligned}$$

From **A6**, $\psi'(t, z)$ is Lipschitz in t with the Lipschitz constant $2L_\psi$,$\forall z$. Thus we have,

$$|\psi'(v_i, \mathbf{z}_i) - \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)| \leq 2L_\psi |v_i - \mathbf{x}_i^\top \widehat{\beta}| \leq 2L_\psi |\mathbf{x}_i^\top \beta_0 - \mathbf{x}_i^\top \widehat{\beta}|, \qquad (62)$$

and by combining (61), (62), we obtain

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n (\psi'(v_i, \mathbf{z}_i) - \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)) \mathbf{x}_i^\top \widehat{\Delta} \widehat{\Theta}_j^\top \mathbf{x}_i &\leq \frac{2L_\psi}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \widehat{\Delta})^2 |\widehat{\Theta}_j^\top \mathbf{x}_i| \\
&\leq \frac{2L_\psi}{n} \|\mathbf{X}\widehat{\Delta}\|_2^2 \max_{1 \leq i \leq n} |\widehat{\Theta}_j^\top \mathbf{x}_i|.
\end{aligned}$$

To bound $\|\mathbf{X}\widehat{\Delta}\|_2^2$, we use the following result, which can be obtained by combining Lemma 12 and 15 in Loh and Wainwright (2012).

Lemma 26 *Suppose $\mathbf{x}_i$ satisfies the sub-gaussian tail condition with the parameter K_X, for all $i = 1, \ldots, n$. For any $u > 0$, the following inequality holds with probability at least $1 - 2\exp(-c' n u(1 \wedge u)/2)$,*

$$\sup_{\|v\|_1 \leq \sqrt{s(u)} \|v\|_2} \left| v^\top \left(\sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{n} - \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top) \right) v \right| \leq 27 u K_X^2 \|v\|_2^2 \qquad (63)$$

where $s(u) := (c' n / 4 \log p)(u \wedge u^2)$ and c' is a universal constant in Bernstein's inequality (see Corollary 2.8.3 in Vershynin, 2018), given a sufficient sample size $n \geq (4 \log p / c') \max\{(u \wedge u^2), (u \wedge u^2)^{-1}\}$.

Then from an application of Lemma 26, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \Delta)^2 \leq \alpha' \|\Delta\|_2^2 + \tau' \frac{\log p}{n}\|\Delta\|_1^2, \quad \forall \Delta \in \mathbb{R}^p \tag{64}$$

with probability at least $1 - 2\exp(-c'n)$ where α', τ', and c' are constants which only depend on model parameter K_X and not dimensions (n, p). Thus we have $\|\mathbf{X}\widehat{\Delta}\|_2^2/n = O_p(s_0(\log p/n)) + O_p(s_0^2(\log p/n)^2) = o_p(n^{-1/2})$ by the rate assumption of s_0 in **A5**. Also,

$$\begin{aligned}\max_{1\leq i\leq n} |\mathbf{x}_i^\top \widehat{\Theta}_j| &\leq \max_{1\leq i\leq n} |\mathbf{x}_i^\top (\widehat{\Theta}_j - \Theta_j)| + \max_{1\leq i\leq n} |\mathbf{x}_i^\top \Theta_j| \\ &\leq \max_{1\leq i\leq n} \|\mathbf{x}_i\|_\infty \|\widehat{\Theta}_j - \Theta_j\|_1 + \|\mathbf{X}\Theta_j\|_\infty = O_p(1).\end{aligned}$$

It holds because $\max_{i,j}|\mathbf{x}_{ij}| \leq C_X$ by **A1'**, $\|\widehat{\Theta}_j - \Theta_j\|_1 = o_p(\sqrt{1/\log p})$ from the assumption about $\widehat{\Theta}$, and $\|\mathbf{X}\Theta_j\|_\infty = O_p(1)$ from **A5**. Therefore,

$$\frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \|\mathbf{X}\widehat{\Theta}_j\|_\infty = o_p(n^{-1/2}), \tag{65}$$

and we have,

$$\widehat{\Theta}_j^\top \psi_n(\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\psi(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) + \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\mathbf{x}_i^\top \widehat{\Delta}\right)\widehat{\Theta}_j^\top \mathbf{x}_i + o_p(n^{-1/2}). \tag{66}$$

Combining (60) with (66),

$$\begin{aligned}\widehat{\beta}_j^{\mathrm{db}} - \beta_{0j} &= \widehat{\beta}_j - \beta_{0j} - n^{-1}\sum_{i=1}^{n}\left(\psi(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) + \psi'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\mathbf{x}_i^\top \widehat{\Delta}\right)\widehat{\Theta}_j^\top \mathbf{x}_i + o_p(n^{-1/2}) \\ &= e_j^\top \widehat{\Delta} - \widehat{\Theta}_j^\top \psi_n(\beta_0) - n^{-1}\sum_{i=1}^{n}\left(\psi_I'(\mathbf{x}_i^\top \widehat{\beta}) + \psi_R'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\right)\widehat{\Theta}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \widehat{\Delta} + o_p(n^{-1/2}),\end{aligned}$$

where we use the relationship $\psi'(t, z) = \psi_I'(t) + \psi_R'(t, z)$ in (28). Recalling the definition $\psi_{I,n}'(\beta) := n^{-1}\sum_{i=1}^{n}\psi_I'(\mathbf{x}_i^\top \beta)\mathbf{x}_i\mathbf{x}_i^\top$,

$$n^{-1}\sum_{i=1}^{n}\psi_I'(\mathbf{x}_i^\top \widehat{\beta})\widehat{\Theta}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \widehat{\Delta} = \widehat{\Theta}_j^\top \left(n^{-1}\sum_{i=1}^{n}\psi_I'(\mathbf{x}_i^\top \widehat{\beta})\mathbf{x}_i\mathbf{x}_i^\top\right)\widehat{\Delta} = \widehat{\Theta}_j^\top \psi_{I,n}'(\widehat{\beta})\widehat{\Delta}$$

thus we have

$$\widehat{\beta}_j^{\mathrm{db}} - \beta_{0j} = -\widehat{\Theta}_j^\top \psi_n(\beta_0) \underbrace{- n^{-1}\sum_{i=1}^{n}\psi_R'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i)\widehat{\Theta}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \widehat{\Delta}}_{\text{Term-I}} + \underbrace{\widehat{\Delta}^\top(e_j - \psi_{I,n}'(\widehat{\beta})\widehat{\Theta}_j)}_{\text{Term-II}} + o_p(n^{-1/2}).$$

We will show that the first term $\sqrt{n}\widehat{\Theta}_j^\top \psi_n(\beta_0)$ will converge to the normal distribution. Both remainder terms (Term-I and Term-II) need to be $o_p(n^{-1/2})$. For the second remainder term (Term-II), we have $|\widehat{\Delta}^\top(e_j - \psi_{I,n}'(\widehat{\beta})\widehat{\Theta}_j)| \leq \|\widehat{\Delta}\|_1 \|e_j - \psi_{I,n}'(\widehat{\beta})\widehat{\Theta}_j\|_\infty =$

$O_p(s_0\sqrt{\log p/n})O_p(\sqrt{\log p/n}) = o_p(n^{-1/2})$ by the rate condition **A5** and the assumptions in the theorem. Now we address the first remainder term (Term-I):

$$n^{-1}\sum_{i=1}^{n}\psi_R'(\mathbf{x}_i^\top\widehat{\beta},\mathbf{z}_i)\widehat{\Theta}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\widehat{\Delta}$$
$$= n^{-1}\sum_{i=1}^{n}\left\{\psi_R'(\mathbf{x}_i^\top\beta_0,\mathbf{z}_i) + \left(\psi_R'(\mathbf{x}_i^\top\widehat{\beta},\mathbf{z}_i) - \psi_R'(\mathbf{x}_i^\top\beta_0,\mathbf{z}_i)\right)\right\}\widehat{\Theta}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\widehat{\Delta}. \tag{67}$$

We need the following Lemma which establishes a kind of sparse eigenvalue condition.

Lemma 27 *Let $E \in \mathbb{R}^{n\times n}$ be a random matrix which has a representation $E = \frac{1}{n}\sum_{i=1}^n e_i\mathbf{x}_i\mathbf{x}_i^\top$, for random $(e_i)_{i=1}^n$ such that $\mathbb{E}[e_i|\mathbf{x}_i] = 0$ and $|e_i| \le c_e$ a.s., and $\mathbf{x}_i$ satisfies the sub-gaussian tail condition with the parameter K_X for all i. Then for any $s, s' \ge 1$, if $n \ge C(s+s')\log p$ for an absolute constant C, there exist constants $c_1, c_2 > 0$ such that*

$$P\left(\sup_{\substack{u\in\mathbb{B}_1(\sqrt{s})\cap\mathbb{B}_2(1),\\ v\in\mathbb{B}_1(\sqrt{s'})\cap\mathbb{B}_2(1)}}|u^\top Ev| \ge c_1\sqrt{(s+s')\frac{\log p}{n}}\right) \le \frac{c_2}{p^{s+s'}}.$$

The proof of the Lemma is presented at the end of this section. Now we apply Lemma 27 to show that $n^{-1}\sum_{i=1}^n\psi_R'(\mathbf{x}_i^\top\widehat{\beta},\mathbf{z}_i)\widehat{\Theta}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\widehat{\Delta}$ is $o_p(n^{-1/2})$. We have,

$$n^{-1}\sum_{i=1}^{n}\psi_R'(\mathbf{x}_i^\top\beta_0,\mathbf{z}_i)\widehat{\Theta}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\widehat{\Delta} = \widehat{\Theta}_j^\top\left(n^{-1}\sum_{i=1}^{n}\psi_R'(\mathbf{x}_i^\top\beta_0,\mathbf{z}_i)\mathbf{x}_i\mathbf{x}_i^\top\right)\widehat{\Delta} = \widehat{\Theta}_j^\top E^R\widehat{\Delta}$$

where we define $E^R := n^{-1}\sum_{i=1}^n\psi_R'(\mathbf{x}_i^\top\beta_0,\mathbf{z}_i)\mathbf{x}_i\mathbf{x}_i^\top$. From the condition of $\widehat{\beta}$ in Theorem 13, we have $\widehat{\Delta}/\|\widehat{\Delta}\|_2 \in \mathbb{B}_1(\sqrt{cs_0})\cap\mathbb{B}_2(1)$ for a constant $c > 0$. Also, $\|\widehat{\Theta}_j\|_1 \le \|\widehat{\Theta}_j - \Theta_j\|_1 + \|\Theta_j\|_1$ and $\|\widehat{\Theta}_j\|_2 \ge \|\Theta_j\|_2 - \|\widehat{\Theta}_j - \Theta_j\|_2 \ge \|\Theta_j\|_2 - \|\widehat{\Theta}_j - \Theta_j\|_1$. Define an event $\mathcal{E}_n := \{\|\widehat{\Theta}_j - \Theta_j\|_1 \le 0.5\|\Theta_j\|_2\}$. Then

$$\frac{\|\widehat{\Theta}_j\|_1}{\|\widehat{\Theta}_j\|_2} \le \frac{\|\Theta_j\|_1 + \|\widehat{\Theta}_j - \Theta_j\|_1}{\|\Theta_j\|_2 - \|\widehat{\Theta}_j - \Theta_j\|_1} \le 3\frac{\|\Theta_j\|_1}{\|\Theta_j\|_2}$$

on $\mathcal{E}_n$. We note that Θ_j is at most $s_* + 1$ sparse vector, recalling the definition $s_* := \max_{1\le j\le p}\|\Theta_{j,-j}\|_0$. Also, $\|\Theta\|_2 \asymp 1$, since $\|\Theta\|_2 = \lambda_{\min}^{-1}(\mathbb{E}[\psi_I'(\mathbf{x}^\top\beta_0)\mathbf{x}\mathbf{x}^\top])$ and the minimum eigenvalue of $\mathbb{E}[\psi_I'(\mathbf{x}^\top\beta_0)\mathbf{x}\mathbf{x}^\top]$ can be shown to be bounded above and also bounded below by a positive constant. More concretely, for any unit vector u,

$$u^T\mathbb{E}[\psi_I'(\mathbf{x}^T\beta_0)\mathbf{x}\mathbf{x}^\top]u \ge \mathbb{E}[\psi_I'(\mathbf{x}^\top\beta_0)(\mathbf{x}^\top u)^2\mathbb{1}\{|\mathbf{x}^\top\beta_0| \le \tau_c\}] \ge \inf_{|t|\le\tau_c}\psi_I'(t)C_\lambda/2,$$

and

$$u^T\mathbb{E}[\psi_I'(\mathbf{x}^T\beta_0)\mathbf{x}\mathbf{x}^\top]u \le C_\psi\mathbb{E}[(\mathbf{x}^\top u)^2] \le 2C_\psi K_X^2$$

for $\tau_c := (2c_b^2K_X^2\log(16K_X^2/C_\lambda))^{1/2}$ using Lemma 19, Assumptions **A1'** and **A6**, where c_b is a constant such that $\|\beta_0\|_2 \le c_b$, which exists by the condition $\|\beta_0\|_2 = O(1)$.

Thus $\|\Theta_j\|_1 \le \sqrt{s_*+1}\|\Theta_j\|_2$, $\widehat{\Theta}_j/\|\widehat{\Theta}_j\|_2 \in \mathbb{B}_1(\sqrt{9(s_*+1)}) \cap \mathbb{B}_2(1)$ on $\mathcal{E}_n$. Also, we have $\mathbb{P}(\mathcal{E}_n) \underset{n}{\to} 1$ by $\|\widehat{\Theta}_j - \Theta_j\|_1 = o_p(1/\sqrt{\log p})$ and $\|\Theta_j\|_2 \asymp 1$. Then on $\mathcal{E}_n$,

$$|\widehat{\Delta}^\top E^R \widehat{\Theta}_j| \le \|\widehat{\Delta}\|_2 \|\widehat{\Theta}_j\|_2 \sup_{\substack{u \in \mathbb{B}_1(\sqrt{s}) \cap \mathbb{B}_2(1), \\ v \in \mathbb{B}_1(\sqrt{s'}) \cap \mathbb{B}_2(1)}} |u^\top E v|,$$

for $s = cs_0$ and $s' = 9(s_*+1)$. Since $\|\widehat{\Delta}\|_2 = O_p(\sqrt{s_0 \log p/n})$ and $\|\widehat{\Theta}_j\|_2 = O_p(1)$,

$$\|\widehat{\Delta}\|_2 \|\widehat{\Theta}_j\|_2 \sup_{\substack{u \in \mathbb{B}_1(\sqrt{s}) \cap \mathbb{B}_2(1), \\ v \in \mathbb{B}_1(\sqrt{s'}) \cap \mathbb{B}_2(1)}} |u^\top E v| = O_p(\sqrt{\frac{s_0 \log p}{n}}) \cdot O_p(\sqrt{\frac{(s_0+s_*)\log p}{n}}) = o_p(n^{-1/2})$$

on $\mathcal{E}_n$, where the last inequality is from the rate conditions $s_0, s_* = o(\sqrt{n}/\log p)$ from **A5**. Since $\mathbb{P}(\mathcal{E}_n) \underset{n}{\to} 1$, we conclude $|\widehat{\Delta}^\top E^R \widehat{\Theta}_j| = o_p(n^{-1/2})$.

For the second term in (67),

$$n^{-1} \sum_{i=1}^n \left| \psi_R'(\mathbf{x}_i^\top \widehat{\beta}, \mathbf{z}_i) - \psi_R'(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) \right| |\widehat{\Theta}_j^\top \mathbf{x}_i| |\mathbf{x}_i^\top \widehat{\Delta}| \le L_\psi \|\mathbf{X}\widehat{\Theta}_j\|_\infty \frac{1}{n} \|\mathbf{X}\widehat{\Delta}\|_2^2$$

where we use **A6** that $\psi_R'(t,z)$ is L_ψ-Lipschitz in t for any z. Then from (65), we have that the second term is $o_p(n^{-1/2})$. Therefore, combining the results we obtain

$$n^{-1} \sum_{i=1}^n \psi_R'(x_i^\top \widehat{\beta}, \mathbf{z}_i) \widehat{\Theta}_j^\top x_i x_i^\top \widehat{\Delta} = o_p(n^{-1/2}).$$

So far, we have obtained,

$$\widehat{\beta}_j^{\mathrm{db}} - \beta_{0j} = -\widehat{\Theta}_j^\top \psi_n(\beta_0) + o_p(n^{-1/2}).$$

It remains to show that

$$\frac{\sqrt{n}\widehat{\Theta}_j^\top \psi_n(\beta_0)}{\sqrt{(\Theta^\top \mathbb{E}[\psi(\mathbf{x}^\top \beta_0, \mathbf{z})^2 \mathbf{x}\mathbf{x}^\top]\Theta)_{jj}}} \xrightarrow{d} \mathcal{N}(0,1).$$

By CLT,

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n \psi(\mathbf{x}_i^\top \beta_0, \mathbf{z}_i) \mathbf{x}_i^\top \Theta_j \to \mathcal{N}(0,1)$$

where

$$\sigma^2 = \mathrm{Var}(\psi(\mathbf{x}^\top \beta_0, \mathbf{z}) \mathbf{x}^\top \Theta_j) = \mathbb{E}[\psi(\mathbf{x}^\top \beta_0, \mathbf{z})^2 (\mathbf{x}^\top \Theta_j)^2]$$

since $\mathbb{E}[\psi(\mathbf{x}^\top \beta_0, \mathbf{z})] = 0$ by (27). Thus it is sufficient to show $\sqrt{n}\widehat{\Theta}_j^\top \psi_n(\beta_0) = \sqrt{n}\Theta_j^\top \psi_n(\beta_0) + o_p(1)$ to conclude. Indeed, we have,

$$|\sqrt{n}(\widehat{\Theta}_j - \Theta_j)^\top \psi_n(\beta_0)| \le \sqrt{n}\|\widehat{\Theta}_j - \Theta_j\|_1 \|\psi_n(\beta_0)\|_\infty = o_p(1).$$

This holds because, by the condition of $\widehat{\Theta}$, $\|\widehat{\Theta}_j - \Theta_j\|_1 = o_p(\sqrt{1/\log p})$ and $\|\psi_n(\beta_0)\|_\infty = O_p(\sqrt{\log p/n})$. Recalling the definition of ψ_n, we have,

$$\|\psi_n(\beta_0)\|_\infty = \max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^n \psi(\mathbf{x}_i^\top\beta_0, \mathbf{z}_i)\mathbf{x}_{ij}\right|.$$

From **A6**, we have $\|\psi\|_\infty \le C_\psi$. Also $\mathbb{E}[\psi(\mathbf{x}_i^\top\beta_0, \mathbf{z}_i)\mathbf{x}_{ij}] = 0$ by (27). Thus $\frac{1}{n}\sum_{i=1}^n \psi(\mathbf{x}_i^\top\beta_0, \mathbf{z}_i)\mathbf{x}_{ij}$ is mean-zero sub-gaussian with a parameter $cC_\psi K_X/\sqrt{n}$ for an absolute constant $c > 0$. Thus from Lemma 22, $\|\psi_n(\beta_0)\|_\infty = O_p(\sqrt{\log p/n})$.

Proof [Proof of Lemma 27] First we establish the following inequality. For any $\tilde{s} \ge 1$, there exists $c_0 > 0$ which do not depend on dimensions (n, p) such that

$$\mathbb{P}\left(\sup_{u\in\mathbb{B}_0(\tilde{s})\cap\mathbb{B}_2(1)} |u^\top E u| \ge c_0\sqrt{\tilde{s}\frac{\log p}{n}}\right) \le c_2/p^{\tilde{s}}, \tag{68}$$

holds where c_2 is an absolute constant.

Since for any unit vector $u \in \mathbb{R}^p$ and i, $\mathbb{E}[e_i(\mathbf{x}_i^\top u)^2] = 0$ and $\mathbb{E}[|e_i(\mathbf{x}_i^\top u)^2|^k]^{1/k} \le c_e\mathbb{E}[(\mathbf{x}_i^\top u)^{2k}]^{1/k} \le 2c_e K_X^2 k$, $\forall k \ge 1$, $e_i(\mathbf{x}_i^\top u)^2$ is mean-zero sub-exponential whose parameter is $cc_e K_X^2$ for an absolute constant c. From Bernstein's inequality, for every $t \ge 0$, we have

$$\mathbb{P}(|u^\top E u| \ge t c_e K_X^2) \le \exp(-c' n(t^2 \wedge t)), \tag{69}$$

where $c' > 0$ is an absolute constant. Note,

$$\begin{aligned}\mathbb{B}_0(\tilde{s})\cap\mathbb{B}_2(1) &= \bigcup_{k=0}^{\tilde{s}}\{v\in\mathbb{B}_2(1); \|v\|_0 = k\}\\ &= \bigcup_{k=0}^{\tilde{s}}\bigcup_{S;|S|=k}\{v\in\mathbb{B}_2(1); \mathrm{supp}(v) = S\}.\end{aligned}$$

Taking a union bound,

$$\mathbb{P}(\sup_{u\in\mathbb{B}_0(\tilde{s})\cap\mathbb{B}_2(1)} |u^\top E u| \ge t c_e K_X^2) \le \sum_{k=0}^{\tilde{s}}\sum_{S;|S|=k} P(\|E_{S,S}\|_2 \ge t c_e K_X^2),$$

where $E_{S,S}$ is a sub-matrix of E supported on S. Letting $\mathcal{N}_\epsilon$ is an ϵ-net of the sphere $\mathcal{S}^{|S|-1}$, we have

$$\|E_{S,S}\|_2 \le \frac{1}{1-2\epsilon}\sup_{v\in\mathcal{N}_\epsilon}|v^\top E_{S,S} v|$$

by the covering argument (e.g., Vershynin, 2018). Take $\epsilon = 1/4$. Then,

$$\begin{aligned}\mathbb{P}(\sup_{u\in\mathbb{B}_0(\tilde{s})\cap\mathbb{B}_2(1)} |u^\top E u| \ge t c_e K_X^2) &\le \sum_{k=0}^{\tilde{s}}\binom{p}{k}9^k P(|v^\top E_{S,S} v| \ge t c_e K_X^2/2)\\ &\le 2\exp(-c'' n(t\wedge t^2) + \tilde{s}\log(9p)),\end{aligned}$$

where we use the bounds $|\mathcal{N}_{1/4}| \leq 9^{|S|}$, $\binom{p}{k} \leq p^k$, and (69), and c'' is a universal constant. Taking $t^2 = 2\tilde{s}\log(9p)/c''n$, we have

$$\mathbb{P}(\sup_{u\in\mathbb{B}_0(\tilde{s})\cap\mathbb{B}_2(1)} |u^\top Eu| \geq c_e K_X^2 \sqrt{2\tilde{s}\frac{\log 9p}{n}}) \leq 2/(9p^{\tilde{s}}),$$

given a sample size condition $n \geq 2\tilde{s}\log 9p/c''$. The we obtain the inequality (68) with $c_0 = 4c_eK_X^2$ and $c_2 = 2/9$, where we use the inequality $p^5 \geq 9p$ for $p \geq 2$.

Now we show that on the event that

$$\sup_{u\in\mathbb{B}_0(s+s')\cap\mathbb{B}_2(1)} |u^\top Eu| \leq c_0\sqrt{(s+s')\frac{\log p}{n}}, \tag{70}$$

we have

$$\sup_{\substack{u\in\mathbb{B}_1(\sqrt{s})\cap\mathbb{B}_2(1),\\ v\in\mathbb{B}_1(\sqrt{s'})\cap\mathbb{B}_2(1)}} |u^\top Ev| \leq c_1\sqrt{(s+s')\frac{\log p}{n}} \tag{71}$$

where c_1 is a multiple of c_0.

From Lemma 11 in Loh and Wainwright (2012), we have

$$\mathbb{B}_1(\sqrt{s})\cap\mathbb{B}_2(1) \subseteq 3\overline{\text{conv}(\mathbb{B}_0(s)\cap\mathbb{B}_2(1))}$$

where $\text{conv}(D)$ denotes a convex hull of $D \subseteq \mathbb{R}^p$. Using this Lemma, for any $u \in \mathbb{B}_1(\sqrt{s})\cap\mathbb{B}_2(1)$ and $v \in \mathbb{B}_1(\sqrt{s'})\cap\mathbb{B}_2(1)$, we have the following representation

$$u = \sum_{i=1}^p \alpha_i u_i \quad \text{and} \quad v = \sum_{j=1}^p \beta_j v_j$$

for $\alpha_i \geq 0, \beta_j \geq 0, u_i, v_j$ such that $\sum_i \alpha_i = \sum_j \beta_j = 1$, $u_i \in \mathbb{B}_0(s)\cap\mathbb{B}_2(3)$ and $v_j \in \mathbb{B}_0(s')\cap\mathbb{B}_2(3)$, $\forall i, j$. Then,

$$u^\top Ev = (\sum_{i=1}^p \alpha_i u_i)^\top E(\sum_{j=1}^p \beta_j v_j) = \sum_{i,j} 9\alpha_i\beta_j \tilde{u}_i^\top E\tilde{v}_j$$

for $\tilde{u}_i, \tilde{v}_j \in \mathbb{B}_0(s+s')\cap\mathbb{B}_2(1), \forall i, j$. Since,

$$|\tilde{u}_i^\top E\tilde{v}_j| \leq \frac{1}{2}\left\{|(\tilde{u}_i+\tilde{v}_j)^\top E(\tilde{u}_i+\tilde{v}_j)| + |\tilde{u}_i^\top E\tilde{u}_i| + |\tilde{v}_j^\top E\tilde{v}_j|\right\},$$

by the basic inequality $2x^\top Ey = (x+y)^\top E(x+y) - x^\top Ex - y^\top Ey$ for any $x, y \in \mathbb{R}^p$, we have

$$\begin{aligned}\sup_{\substack{u\in\mathbb{B}_1(\sqrt{s})\cap\mathbb{B}_2(1),\\ v\in\mathbb{B}_1(\sqrt{s'})\cap\mathbb{B}_2(1)}} |u^\top Ev| &\leq \sum_{i,j}\frac{27}{2}(\alpha_i\beta_j) \sup_{u\in\mathbb{B}_0(s+s')\cap\mathbb{B}_2(1)} |u^\top Eu| \\ &\leq \sum_{i,j}(\alpha_i\beta_j)\left(\frac{27}{2}c_0\sqrt{(s+s')\frac{\log p}{n}}\right) = c_1\sqrt{(s+s')\frac{\log p}{n}}\end{aligned}$$

where for the second inequality we use (70) and $c_1 = 27c_0/2$. Thus (71) holds with probability at least $1 - c_2/p^{s+s'}$. ■

B.6 Construction of an Approximate Inverse of Fisher Information Matrix Using Node-Wise Regression

First we let $W(\beta) := \mathrm{diag}(\{\psi_I'(\mathbf{x}_i^\top \beta)\}_{i=1}^n)$. We note the square root of $W(\beta)$ exists since $\psi_I(t) \geq 0$ for all t. Following the node-wise lasso construction in van de Geer et al. (2014), we define

$$\widehat{\gamma}_j := \underset{\gamma \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \|W(\widehat{\beta})^{1/2}\mathbf{X}_j - W(\widehat{\beta})^{1/2}\mathbf{X}_{-j}\gamma\|_2 + \lambda_j \|\gamma\|_1$$
$$\widehat{\tau}_j^2 := \|W(\widehat{\beta})^{1/2}\mathbf{X}_j - W(\widehat{\beta})^{1/2}\mathbf{X}_{-j}\widehat{\gamma}\|_2^2/n + \lambda_j \|\widehat{\gamma}_j\|_1.$$

We construct $\widehat{\Theta}(\psi)$ by taking $\widehat{\Theta}(\psi)_j^\top := \widehat{\tau}_j^{-2}[-\widehat{\gamma}_{j,1}, \ldots, 1, -\widehat{\gamma}_{j,p}] \in \mathbb{R}^{1\times p}$.

Lemma 28 (Theorem 3.2 in van de Geer et al., 2014) *Assume* **A1', A3, A5-A6** *and* $\lambda_j \asymp \sqrt{\log p/n}$ *for all* j. *In addition we assume there exists* $C_X > 0$ *such that* $\|\mathbf{x}_i\|_\infty \leq C_X$ *a.s. for all* i. *Then for any* $j \in \{1, \ldots, p\}$, *we have*

$$\|\widehat{\Theta}_j(\psi) - \Theta_j(\psi)\|_1 = o_p(1/\sqrt{\log p}), \qquad \|\widehat{\Theta}_j(\psi) - \Theta_j(\psi)\|_2 = o_p(n^{-1/4}).$$

Proof The result follows by checking the conditions of Theorem 3.2 in van de Geer et al. (2014). ■

References

D. Angluin and P. Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, Apr. 1988.

A. Beck. *First-Order Methods in Optimization*. SIAM, Oct. 2017.

C. R. Bollinger and M. H. David. Modeling discrete choice with response error: Food stamp participation. *J. Am. Stat. Assoc.*, 92(439):827–835, 1997.

J. Bootkrajang and A. Kabán. Label-Noise robust logistic regression and its applications. In *Machine Learning and Knowledge Discovery in Databases*, pages 143–158. Springer Berlin Heidelberg, 2012.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall/CRC, 2 edition edition, June 2006.

A. T. Chaganty and P. Liang. Estimating Latent-Variable graphical models using moments and likelihoods. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1872–1880, Bejing, China, Jan. 2014.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-Dimensional inference: Confidence intervals, p-Values and R-Software hdi. *Stat. Sci.*, 30(4):533–558, 2015.

M. Du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pages 1386–1394, June 2015.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer-Verlag New York, 2 edition, 2001.

D. M. Fowler and S. Fields. Deep mutational scanning: a new style of protein science. *Nat. Methods*, 11(8):801–807, Aug. 2014.

B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Trans Neural Netw Learn Syst*, 25(5):845–869, May 2014.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.*, 31(4):1208–1211, 1960.

T. Hastie and W. Fithian. Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867, Aug. 2013.

J. A. Hausman, J. Abrevaya, and F. M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *J. Econom.*, 87(2):239–269, Dec. 1998.

P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA, 2010.

S. Jain, M. White, and P. Radivojac. Recovering true classifier performance in positive-unlabeled learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

T. Kato. *Perturbation theory for linear operators*. Springer Science & Business Media, June 2013.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Stat.*, 28(5):1356–1378, Oct. 2000.

A. Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36, Jan. 2014.

A. H. Li and J. Bradic. Boosting in the presence of outliers: Adaptive classification with nonconvex loss functions. *Journal of the American Statistical Association*, 113(522):660–674, 2018.

B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186, Nov. 2003.

P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Stat.*, 45(2):866–896, Apr. 2017.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Stat.*, 40(3):1637–1664, June 2012.

P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, 16(1):559–616, Jan. 2015.

R. H. Lyles, L. Tang, H. M. Superak, C. C. King, D. D. Celentano, Y. Lo, and J. D. Sobel. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology*, 22(4):589–597, July 2011.

L. S. Magder and J. P. Hughes. Logistic regression when the outcome is measured with uncertainty. *Am. J. Epidemiol.*, 146(2):195–203, July 1997.

P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition*. CRC Press, Aug. 1989.

S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *Ann. Stat.*, 46(6A):2747–2774, Dec. 2018.

R. Morton. Efficiency of estimating equations and the use of pivots. *Biometrika*, 68(1): 227–233, 1981. ISSN 0006-3444.

N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Cost-Sensitive learning with noisy labels. *J. Mach. Learn. Res.*, 18(155):1–33, 2018.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for High-Dimensional analysis of M-Estimators with decomposable regularizers. *Stat. Sci.*, 27(4):538–557, Nov. 2012.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, Jan. 2009.

J. M. Neuhaus. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855, 1999.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, Jan. 1994.

M. S. Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2):355–365, June 1992.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for High-Dimensional linear regression Over ℓ_q-Balls. *IEEE Trans. Inf. Theory*, 57(10):6976–6994, Oct. 2011.

P. A. Romero, T. M. Tran, and A. R. Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl. Acad. Sci. U. S. A.*, 112(23):7159–7164, June 2015.

D. Sculley and G. V. Cormack. Filtering email spam in the presence of noisy user feedback. In *CEAS*, 2008.

J. Shao. *Mathematical Statistics*. Springer Science & Business Media, July 2003.

R. R. Singhania, A. K. Patel, R. K. Sukumaran, C. Larroche, and A. Pandey. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. *Bioresour. Technol.*, 127:500–507, Jan. 2013.

P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 1085–1092. MIT Press, 1995.

H. Song and G. Raskutti. PUlasso: High-dimensional variable selection with presence-only data. *J. Am. Stat. Assoc.*, pages 1–41, Dec. 2018.

S. van de Geer, P. Bhlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014.

A. van den Hout and P. G. M. van der Heijden. Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review / Revue Internationale de Statistique*, 70(2):269–288, 2002.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

R. Vershynin. *High-Dimensional Probability by Roman Vershynin*. Cambridge University Press, Sept. 2018.

G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. Presence-only data and the em algorithm. *Biometrics*, 65(2):554–563, June 2009.

P. Yang, X. Li, H.-N. Chua, C.-K. Kwoh, and S.-K. Ng. Ensemble positive unlabeled learning for disease gene identification. *PLoS One*, 9(5):e97079, May 2014.