Concentration Based Inference in High Dimensional Generalized Regression Models (I: Statistical Guarantees)

Zhu, Ying

Purdue University

17 August 2018

Concentration Based Inference in High Dimensional Generalized Regression Models (I: Statistical Guarantees)

Ying Zhu*

August 17, 2018

Abstract

We develop simple and non-asymptotically justified methods for hypothesis testing about the coefficients ($\theta^* \in \mathbb{R}^p$) in the high dimensional generalized regression models where p can exceed the sample size n. Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, we consider $H_0 : h(\theta^*) = \mathbf{0}_m$ against the alternative hypothesis $H_1 : h(\theta^*) \neq \mathbf{0}_m$, where m can be any integer in $[1, p]$ and h is allowed to be nonlinear in θ^*. Our test statistics is based on the sample "quasi score" vector evaluated at an estimate $\hat{\theta}_\alpha$ that satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$, where α is the prespecified Type I error. By exploiting the concentration phenomenon in Lipschitz functions, the key component reflecting the "dimension complexity" in our non-asymptotic thresholds uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated around and automatically captures the dependencies between the coordinates. We provide probabilistic guarantees in terms of the Type I and Type II errors for the "quasi score" test. In addition, confidence regions are constructed for the population quasi-score vector evaluated at θ^*. The first set of our results are specific to the standard Gaussian linear regression models; the second set of our results allow for reasonably flexible forms of non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients (which include the binary response models and certain nonlinear regressions), while only requiring the correct specification of $\mathbb{E}(Y_i|X_i)$s. The novelty of our methods is that their validity does not rely on good behavior of $\left\|\hat{\theta}_\alpha - \theta^*\right\|_2$ (or even $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ in the linear regression case) nonasymptotically or asymptotically.

*Corresponding Author. Email: yingzhu@purdue.edu. Assistant Professor of Statistics and Computer Science. Purdue University. West Lafayette, Indiana. A start-up fund from Purdue University partially supported this research. This draft was prepared during my appointment at Michigan State University (Department of Economics, Social Science Data Analytics Initiative) that also provided financial support. I thank Guang Cheng at Purdue University for discussions that improved the presentation of this work as well as pointing out several related papers. All errors are my own.

1 Introduction

In this paper, we consider non-asymptotic inference in the regression models

$$Y_i = \Upsilon\left(X_i; \theta^*\right) + W_i, \qquad i = 1, ..., n, \tag{1}$$

as well as the binary response models

$$\mathbb{P}\left(Y_i = 1 | X_i\right) = \Pi\left(X_i; \theta^*\right), \qquad i = 1, ..., n, \tag{2}$$

where $Y = \{Y_i\}_{i=1}^n$ consists of independent observations; $X = \{X_i\}_{i=1}^n \in \mathbb{R}^{n \times p}$ is the design matrix with the ith row specified by X_i; and θ^* is a $p-$dimensional vector of unknown coefficients and p is allowed to exceed the sample size n. We assume that the functional forms of $\Upsilon\left(X_i; \theta^*\right)$ and $\Pi\left(X_i; \theta^*\right)$ are known; for example, Π may be a "probit" or a "logit" in (2) and $\Pi\left(X_i; \theta^*\right) = \Pi\left(X_i \theta^*\right)$. Throughout the paper, we make our argument by conditioning on X. In (1), $W = \{W_i\}_{i=1}^n$ is a zero-mean noise vector with independently (but possibly non-identically) distributed entries.

Our goal is to find "practical" non-asymptotic methods for hypothesis testing about the coefficients in (1) and (2) under the regime of $p \geq n$ or even $p \gg n$. The form of our null hypothesis is rather general: Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, we consider

$$H_0 : h(\theta^*) = \mathbf{0}_m \; v.s. \; H_1 : h(\theta^*) \neq \mathbf{0}_m,$$

where m can be any integer in $[1, p]$ and h is allowed to be nonlinear in θ^*. We use $\mathbf{0}_m$ above to denote an $m-$dimensional vector of zeros. By making some changes in the notations, we can also test $H_0 : h(\theta^*) \leq \mathbf{0}_m$ or $H_0 : h(\theta^*) \geq \mathbf{0}_m$ using the procedures and analysis developed later in the paper.

This work is initially motivated by an important problem from intervention studies – testing for heterogeneity in treatment effects. Suppose V_i is a binary variable which equals 1 if individual i receives treatment and 0 otherwise; Z_i is a $p-$dimensional vector of covariates such that $\mathbb{E}(Z_i) = \mathbf{0}_p$ (this zero-mean condition can be relaxed but is assumed here to lighten the notations). We use Y_i^A to denote the (potential) outcome upon receiving treatment, Y_i^B to denote the (potential) outcome without treatment, and Y_i to denote the observed outcome; note that $Y_i = (1 - V_i) Y_i^B + V_i Y_i^A$. Let us assume $Y_i^A = \mathbb{E}\left(Y_i^A | Z_i\right) + W_i$ and $Y_i^B = \mathbb{E}\left(Y_i^B | Z_i\right) + W_i$ where

$$\begin{aligned} \mathbb{E}\left(Y_i^A | Z_i\right) &= \mu^* + \sum_{j=1}^p \alpha_j^* Z_{ij}, \\ \mathbb{E}\left(Y_i^A | Z_i\right) - \mathbb{E}\left(Y_i^B | Z_i\right) &= \sum_{j=1}^p \beta_j^* Z_{ij}. \end{aligned}$$

Under the "ignorability" assumption proposed by [15], i.e., $\mathbb{E}\left(Y_i^A|V_i,\, Z_i\right) = \mathbb{E}\left(Y_i^A|Z_i\right)$ and $\mathbb{E}\left(Y_i^B|V_i,\, Z_i\right) = \mathbb{E}\left(Y_i^B|Z_i\right)$, we can write

$$Y_i = \pi_0^* + \pi_1^* V_i + \sum_{j=1}^{p} \beta_j^* V_i Z_{ij} + \sum_{j=1}^{p} \alpha_j^* Z_{ij} + W_i \tag{3}$$

such that

$$TE(Z_i) := \mathbb{E}\left(Y_i^A - Y_i^B|Z_i\right) = \pi_1^* + \sum_{j=1}^{p} \beta_j^* Z_{ij}, \tag{4}$$

$$ATE := \mathbb{E}\left(Y_i^A - Y_i^B\right) = \pi_1^*. \tag{5}$$

Taking the expectation of $TE(Z_i)$ in (4) over Z_i gives (5), referred to as the Average Treatment Effect (ATE). For a realization $\{z_{ij}\}_{j=1}^p$ of $\{Z_{ij}\}_{j=1}^p$, the heterogeneity in the treatment effect corresponds to $\sum_{j=1}^p \beta_j^* z_{ij}$. In practice, we may be interested in two types of hypotheses about the form of heterogeneity. Given some $G \subseteq \{1, 2, ..., p\}$ of our interest, the first one considers

$$H_0 : \sum_{j \in G} \beta_j^* z_{ij} = b, \tag{6}$$

and the second one considers

$$H_0 : \beta_j^* = \beta_j^0 \ \forall j \in G \subseteq \{1, 2, ..., p\}. \tag{7}$$

These hypotheses can be handled by the methods developed in this paper since they are special cases of $H_0 : h(\theta^*) = \mathbf{0}_m$. Before this paper, some "practical" tests have been proposed in the literature of high dimensional inference. For example, [22] deal with $H_0 : \sum_{j=1}^p \theta_j^* x_{ij} = a$ concerning $Y_i = X_i\theta^* + W_i$ where $\theta^* \in \mathbb{R}^p$ need not be sparse; [6] deal with $H_{0,G} : \theta_j^* = \theta_j^0 \ \forall j \in G \subseteq \{1, 2, ..., p\}$ but require sparse θ^*. In view of their conditions on sparsity and the cardinality of $|G|$, if $\left\{\theta_j^*\right\}_{j \in G}$ is "dense" and the cardinality of G is "large", the method in [6] may fail.

A common feature of the current testing methods that are deemed "practical" (like the two papers mentioned above) is that they all hinge on asymptotic validity to some extent. This occurrence is perhaps not coincidental as asymptotic analysis often allows one to focus on the "leading" term(s) by assuming the "remainder" terms approach to zero faster, which can be quite convenient for determining the threshold in a test without imposing strong distributional assumptions. However, many real-world applications (some clinical trials, for example) have a limited sample size which renders any asymptotic argument questionable. On the other hand, exact and approximate inference methods that rely on properties of specific distributions can be too hard to generalize and therefore have limited applications.

As suggested by the title, this paper studies nonasymptotic inference by exploiting the sharp concentration phenomenon in Lipschitz functions, which should be distinguished from another line of literature based on normal approximations using the Stein's Method; see, e.g., [5] and [10]. In particular, [10] studies similar models (as this paper) and develops results for hypothesis testing in the regime of $n \gg p$; by contrast, our focus is on the regime of $p \geq n$ and possibly $p \gg n$. In [10], some of the results are still only asymptotically valid and the other results (even though nonasymptotically justified) come with probabilistic guarantees that contain rather loose constants and dimension-dependent components.

For the mean of a high-dimensional random vector, [1] study bootstrap confidence regions with the concentration approach. Beyond the inference for the mean of a high-dimensional random vector, is it possible to adapt a concentration approach for testing about the coefficients in a high-dimensional regression or classification problem? At first glance, there seems no lack of non-asymptotic bounds on the l_p-error (often $p \in [1, 2]$ or $p = \infty$) of some (regularized) estimator concerning (1) with $\Upsilon(X_i; \theta^*) = X_i\theta^*$ or (2) with $\Pi(X_i; \theta^*) = \Pi(X_i\theta^*)$. However, these bounds (even in the sharpest forms) tend to involve quite a few unknown nuisance parameters that are hard to estimate in practice. In order to adapt the existing bounds for the purpose of inference, prior knowledge on the sparsity of θ^* would be needed at a minimum; see, e.g., [9].

For this reason, we choose our test statistics to base on the sample "quasi score" vector evaluated at $\hat{\theta}_\alpha$ that satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$, where α is the prespecified Type I error. In terms of a linear regression model, the resulting procedure becomes a score test. More generally, our test statistics take the form

$$\Psi_q\left(\hat{\theta}_\alpha\right) := \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda\left(X_i; \hat{\theta}_\alpha\right) \right] \right\|_q \tag{8}$$

where $\Lambda = \Upsilon$ or $\Lambda = \Pi$, and $\hat{\theta}_\alpha$ is obtained by solving the following program:

$$\begin{aligned} \left(\hat{\theta}_\alpha, \hat{\mu}_\alpha\right) &\in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}^p} \|\mu_\alpha\|_{\tilde{q}} \\ \text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda\left(X_i; \theta_\alpha\right) \right] - \mu_\alpha \right\|_q &\leq r_{\alpha,q}, \\ h(\theta_\alpha) &= \mathbf{0}_m, \end{aligned} \tag{9}$$

with $q, \tilde{q} \in [1, \infty]$ chosen by the users. For $1 \leq q \leq \infty$, we write $\|v\|_q$ to mean the l_q-norm of a $k-$dimensional vector v, where $\|v\|_q := \left(\sum_{i=1}^{k} |v_i|^q\right)^{1/q}$ when $1 \leq q < \infty$ and $\|v\|_q := \max_{i=1,\ldots,k} |v_i|$ when $q = \infty$. The choice for $r_{\alpha,q}$ in the first constraint is to be specified in the subsequent sections.

We can also work with an alternative formulation:

$$\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right) \in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}} \mu_\alpha$$

$$\text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \theta_\alpha \right) \right] \right\|_q \leq r_{\alpha,q} + \mu_\alpha, \tag{10}$$

$$h(\theta_\alpha) = \mathbf{0}_m,$$

$$\mu_\alpha \geq 0.$$

Throughout this paper, we will slightly abuse the notations as in the above, where $\hat{\mu}_\alpha$ (also μ_α) in (9) is a vector and in (10) is a scalar. In addition, we suppress the dependence of $(\hat{\theta}_\alpha, \hat{\mu}_\alpha)$ in (9) on $(q,\, \tilde{q})$ and the dependence of $(\hat{\theta}_\alpha, \hat{\mu}_\alpha)$ in (10) on q for notational simplicity.

A solution $\hat{\theta}_\alpha$ to either (9) or (10) may not necessarily be unique: that is, there might be different $\hat{\theta}_\alpha$s that satisfy (9) (or (10)) while delivering the same (minimal) objective value $\|\hat{\mu}_\alpha\|_{\tilde{q}}$ (respectively, $\hat{\mu}_\alpha$). We refer to the vector μ_α in (9) (and the scalar μ_α in (10)) as the "slack" vector (respectively, the "slack" variable) that fills the "gap" between $\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \theta^* \right) \right] \right\|_q$ and $\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \theta_\alpha \right) \right] \right\|_q$ where $h(\theta_\alpha) = \mathbf{0}_m$. When the null hypothesis is true, i.e., $h(\theta^*) = \mathbf{0}_m$, the optimal value $\|\hat{\mu}_\alpha\|_{\tilde{q}}$ (respectively, $\hat{\mu}_\alpha$) must be zero with probability at least $1 - \alpha$. This fact does not imply that $\hat{\theta}_\alpha$ would necessarily be "close" to θ^* under H_0, but rather,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \hat{\theta}_\alpha \right) \right] \right\|_q \leq r_{\alpha,q}, \qquad \text{(under } H_0\text{)}$$

with the same probability guarantee $1 - \alpha$ for the event

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \theta^* \right) \right] \right\|_q \leq r_{\alpha,q}.$$

In Sections 2-4, we establish statistical guarantees (stated in terms of $(\alpha, \tilde{q}, q)$) for (9), and statistical guarantees (stated in terms of (α, q)) for (10). Moreover, our statistical theory is valid whether h (or Λ) is convex in θ^* or not; but h and Λ being convex would clearly make the optimization problems less difficult.

To compare (9) with (10) from the computational perspective, we let $\mathcal{F}_1^\alpha$ denote the set of $(\theta_\alpha, \mu_\alpha)$ that are feasible for (9) and $\mathcal{F}_{1,\theta}^\alpha$ denote the set of θ_α from $\mathcal{F}_1^\alpha$; similarly, $\mathcal{F}_2^\alpha$ and $\mathcal{F}_{2,\theta}^\alpha$ are defined with regard to (10). Note that an element $(\tilde{\theta}_\alpha, \tilde{\mu}_\alpha)$ in $\mathcal{F}_1^\alpha$ implies

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \tilde{\theta}_\alpha \right) \right] \right\|_q \leq r_{\alpha,q} + \left\| \tilde{\mu}_\alpha \right\|_q ;$$

that is, $(\tilde{\theta}_\alpha, \|\tilde{\mu}_\alpha\|_q) \in \mathcal{F}_2^\alpha$. Consequently, $\mathcal{F}_{1,\theta}^\alpha \subseteq \mathcal{F}_{2,\theta}^\alpha$. On the other hand, the objective function in (9) is minimized over a $p-$dimensional vector as opposed to a

scalar in (10). However, (9) does not require the slack vector to be positive while (10) require the slack variable to be positive. These facts suggest that the choice between (9) and (10) incurs some trade-offs in terms of computational cost. In a separate paper, we conduct simulation studies to compare the computational performance of (9) and (10) as well as examine various choices of $(\tilde{q},\, q)$ in (9) and q in (10).

Compared to basing the test statistics on a consistent estimator for θ^*, such as the existing Lasso estimators, Dantzig selectors, or the new variant (11) with $\tilde{q} = 1$ and $q = \infty$ (to be discussed later), the "quasi-score" statistics (8) using $\hat{\theta}_\alpha$ from (9) or (10) allow us to avoid the inherent challenges in an inverse problem. Like [22], our analysis does not require any sparsity condition on θ^*. Unlike these authors, we focus on the nonasymptotic inference and our motivation for bypassing the sparsity assumption in θ^* is mainly to make the concentration approach practical in the sense that our thresholds or confidence regions do not involve unknown parameters related to sparsity.

For the special case of a linear regression model, if we choose $q = \infty$, then (8) is reduced to

$$\Psi_\infty\left(\hat{\theta}_\alpha\right) := \left\| \frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - X_i\hat{\theta}_\alpha\right] \right\|_\infty .$$

This statistics shares some resemblance to the score-based correction term in the debiased Lasso literature (see, e.g., [6, 11, 17, 21]) as well as the decorrelated score in [14]. Unlike the debiased and decorrelated procedures which require an initial (consistent) estimator for (the sparse) θ^* in the correction term, our $\hat{\theta}_\alpha$ here need not be consistent and is directly used in the test statistics (requiring no further debiasing or decorrelating step). In addition, our methods are nonasymptotically valid and do not require θ^* to be sparse, whereas the aforementioned papers hinge on the asymptotic normality of the debiased or decorrelated procedure and require θ^* to be sufficiently sparse.

We derive implementable (non-asymptotic) thresholds $r_{\alpha,q}$ such that

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}\right\} \leq \alpha, \qquad \text{(Type I Error)}$$

where $\mathbb{P}_0$ means "under H_0". Our decision rule is that if $\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}$, we reject the null hypothesis H_0. To establish the claim for a given Type II error β, i.e.,

$$\mathbb{P}_1\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}\right\} \leq \beta, \qquad \text{(Type II Error)}$$

where $\mathbb{P}_1$ means "under H_1", we introduce a "Level$-\beta$ Separation Requirement" imposed upon the l_q-distances between the population quasi-score vectors evaluated at θ^* and θ_αs satisfying $h(\theta_\alpha) = \mathbf{0}_m$. In terms of a linear regression model, this requirement is simply about separation between the population score vectors. In addition to the guarantees on the Type I and Type II errors, we also construct confidence regions for the population quasi-score vector evaluated at θ^*.

Our non-asymptotic thresholds $r_{\alpha,q}$ consist of data-driven components which reflect the "dimension complexity", as well as components which are free of p. This form is a direct result of the concentration phenomenon in Lipschitz functions. The key data-driven component in our $r_{\alpha,q}$ uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated around and automatically captures the dependencies across coordinates. These facts put our framework in sharp contrast with the Bonferroni approach used in the estimation literature (e.g., [9]). In this perspective, our results share some similarity as those in [1] except that [1] concern inference for the mean of a random vector while we consider inference about the coefficients ($\theta^* \in \mathbb{R}^p$) in the high dimensional regression and binary response models.

Beyond the context of hypothesis testing, as a secondary contribution, the data-driven approach proposed in this paper for setting the thresholds $r_{a,q}$ also suggests a new class of regularized estimators, which solve

$$\min_{\theta_\alpha \in \mathbb{R}^p} \|\theta_\alpha\|_{\tilde{q}} \ \text{ subject to } \left\| \frac{1}{n} X^T (Y - X\theta_\alpha) \right\|_q \leq r_{\alpha,q}. \tag{11}$$

When $\tilde{q} = 1$ and $q = \infty$, (11) can be viewed as a variant of the Dantzig selector, for which we establish a complementary l_2-error bound and conditions for l_2-consistency. While (9) and (10) share some similarity as (11), they involve a second constraint $h(\theta_\alpha) = \mathbf{0}_m$ and a slack vector (or variable) μ_α in the first constraint, as well as a different objective function (minimizing the $l_{\tilde{q}}-$norm of the slack vector or minimizing the slack variable, instead of minimizing $\|\theta_\alpha\|_{\tilde{q}}$). Any resulting solution to (9) (or (10)) is a constrained estimator, in contrast to an unconstrained estimator obtained from (11).

The rest of the paper is organized as follows. We first consider the linear regression model with homoscedastic Gaussian noise in Section 2 and then move on to the binary response model (2) in Section 3. Section 4 considers the more general regression model (1) and extends our framework in Section 2 to allow for non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients. We postpone this extension until Section 4 because the underlying methods are similar to those in Section 3. Motivated by the data-driven feature of our concentration approach, Section 5 proposes a new class of regularized estimators along with a complementary l_2-error bound. Section 6 concludes the paper and discusses future directions. All technical details are deferred to Section 7.

2 Gaussian Linear Regressions

Let us begin with the linear regression model

$$Y_i = X_i\theta^* + W_i, \qquad i = 1, ..., n, \tag{12}$$

that is, when $\Upsilon(X_i;\theta^*) = X_i\theta^*$ in (1). We specialize (9) (or (10)) to the linear model by letting $\Lambda(X_i;\theta_\alpha) = X_i\theta_\alpha$. As a result, our test statistics (8) simply becomes

$$\Psi_q(\hat{\theta}_\alpha) := \left\| \frac{1}{n} X^T \left(Y - X\hat{\theta}_\alpha \right) \right\|_q .$$

A leading application of (12) concerns hypothesis testing about the coefficients in the standard Gaussian linear regression model. We first consider the scenario where $W \sim \mathcal{N}(0,\, \sigma^2 \boldsymbol{I}_n)$ and σ^2 is known, and then discuss the scenario where σ^2 is not known *a priori*. Throughout this section, we use $\mathbb{E}_W[\cdot]$ to denote the expectation over W only, conditioning on X.

By considering the concentration of $\left\| \frac{1}{n} X^T W \right\|_q$ around $\mathbb{E}_W\left[\left\| \frac{1}{n} X^T W \right\|_q \right]$, our first result establishes an "ideal" confidence region for the l_q-distance between the population score vectors evaluated at θ^* and a "theoretical" optimal solution, $\hat{\theta}^*_\alpha$; that is,

$$\begin{aligned} & \left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) \right\|_q \qquad (13) \\ = & \left\| \mathbb{E}_W\left[\frac{1}{n} X^T (Y - X\hat{\theta}^*_\alpha) \right] - \mathbb{E}_W\left[\frac{1}{n} X^T (Y - X\theta^*) \right] \right\|_q \\ = & \left\| \mathbb{E}_W\left[\frac{1}{n} X^T (Y - X\hat{\theta}^*_\alpha) \right] - \mathbb{E}_W\left[\frac{1}{n} X^T W \right] \right\|_q . \end{aligned}$$

This "theoretical" optimal solution above, $\hat{\theta}^*_\alpha$, is obtained by setting $r_{\alpha,q}$ in (9) (and (10)) to $\mathbb{E}_W\left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ plus a deviation. In practice, $\mathbb{E}_W\left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ may be replaced with its Monte Carlo approximation and a "small" remainder term. This approach results in a "practical" optimal solution, $\hat{\theta}_\alpha$, which can then be used to construct test statistics and a "practical" confidence region.

To state the first result, we introduce the following notation (which will appear in many places throughout this paper):

$$\begin{aligned} \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q &= \sqrt[q]{\sum_{j=1}^p \left(\sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij}^2} \right)^q}, \qquad q \in [1,\, \infty) \\ \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q &= \max_{j\in\{1,\ldots,p\}} \sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij}^2}, \qquad q = \infty. \end{aligned}$$

Proposition 2.1. *Assume (12) where $W \sim \mathcal{N}(0,\, \sigma^2 \boldsymbol{I}_n)$ and is independent of X. Then for any $q \in [1,\, \infty]$, we have*

$$\mathbb{P}\left\{ \left\| \frac{1}{n} X^T W \right\|_q \geq \mathbb{E}_W\left[\left\| \frac{1}{n} X^T W \right\|_q \right] + t \right\} \leq \exp\left(\frac{-nt^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q^2} \right). \qquad (14)$$

Moreover, for $\alpha \in (0,\, 1)$, let

$$r_{\alpha,q} = r^*_{\alpha,q} := \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + \sigma \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q \sqrt{\frac{2}{n} \log \frac{1}{\alpha}} \tag{15}$$

in (9) (or (10)). Then, an optimal solution $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ to (9) must satisfy

$$\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) \right\|_{\tilde{q}} \;\; \geq \;\; \|\hat{\mu}^*_\alpha\|_{\tilde{q}}\,, \tag{16}$$

$$\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) - \hat{\mu}^*_\alpha \right\|_q \;\; \leq \;\; 2r^*_{\alpha,q}, \tag{17}$$

with probability at least $1 - \alpha$. Similarly, an optimal solution $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ to (10) must satisfy

$$\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) \right\|_q \;\; \geq \;\; \hat{\mu}^*_\alpha, \tag{18}$$

$$\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) \right\|_q \;\; \leq \;\; 2r^*_{\alpha,q} + \hat{\mu}^*_\alpha, \tag{19}$$

with probability at least $1 - \alpha$.

Hypothesis Testing

For the moment, suppose we set $r_{\alpha,q} = r^*_{\alpha,q}$ in (9) (or (10)) according to (15) as in Theorem 2.1. Under H_0, $(\theta^*,\, \mathbf{0})$ $((\theta^*,\, 0))$ is an optimal solution to (9) (respectively, (10)). Consequently, for a chosen $\alpha \in (0,\, 1)$, an optimal solution to (9) (and (10)) must satisfy

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}^*_\alpha) \geq r^*_{\alpha,q} \right\} \leq \alpha \tag{20}$$

where $\mathbb{P}_0$ means "under H_0".

The claim in (20) suggests a test (with level α) based on the statistics $\Psi_q(\hat{\theta}^*_\alpha)$ and an "ideal" critical value, $r^*_{\alpha,q}$, given in (15). When $W \sim \mathcal{N}(0,\, \sigma^2 \boldsymbol{I}_n)$ and σ^2 is known, the first term $\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ in $r^*_{\alpha,q}$ can be approximated by Monte-Carlo as follows. Let $Z \in \mathbb{R}^{R \times n}$ be a matrix consisting of independent entries randomly drawn from $\mathcal{N}(0,\, 1)$ and the rth-row of Z is denoted by Z_r. By (69) and (70), note that $\sigma R^{-1} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q$ is sub-Gaussian with parameter at most

$(nR)^{-1/2}\sigma \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q$. Consequently, (67) yields the following concentration

$$\mathbb{P}\left\{ \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] \geq \frac{\sigma}{R} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q + t \right\} \leq \exp\left(\frac{-nRt^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q^2} \right). \tag{21}$$

Combining (14) and (21) yields

$$\begin{aligned} &\mathbb{P}\left\{ \left\| \frac{1}{n} X^T W \right\|_q \geq \frac{\sigma}{R} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q + t_1 + t_2 \right\} \\ \leq &\exp\left(\frac{-nt_1^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q^2} \right) + \exp\left(\frac{-nRt_2^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q^2} \right). \end{aligned} \tag{22}$$

Construction of Critical Values ($r_{\alpha,q}$) and Type I Error

For some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0, 1)$, we let in (22),

$$\begin{aligned} t_1 &= \sigma \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q \sqrt{\frac{2}{n} \log \frac{1}{\alpha_1}} := \tau_{\alpha_1,q}, \\ t_2 &= \sigma \left\| \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \right\|_q \sqrt{\frac{2}{nR} \log \frac{1}{\alpha_2}} := \sqrt{\frac{1}{R}} \tau_{\alpha_2,q}. \end{aligned} \tag{23}$$

Based on (22) along with the choices of t_1 and t_2 above, we set the RHS of the first constraint in (9) (or (10)) with

$$r_{\alpha,q} = \frac{\sigma}{R} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}} \tau_{\alpha_2,q}. \tag{24}$$

Under H_0, $(\theta^*, \mathbf{0})$ $((\theta^*, 0))$ is an optimal solution to (9) (respectively, (10)) with $r_{\alpha,q}$ specified in (24). Consequently, a (practical) optimal solution to (9) (and (10)) must satisfy

$$\mathbb{P}_0\left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha \qquad \textit{(Type I Error).} \tag{25}$$

Practical Confidence Regions

Let $\left(\hat{\theta}_\alpha, \hat{\mu}_\alpha\right)$ be an optimal solution to (9) with $r_{\alpha,q}$ specified in (24). Our previous analysis implies that

$$
\begin{aligned}
& \left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}_\alpha) - \hat{\mu}_\alpha \right\|_q \\
\leq & \left\| \frac{1}{n} X^T (Y - X\hat{\theta}_\alpha) - \hat{\mu}_\alpha \right\|_q + \left\| \frac{1}{n} X^T W \right\|_q \\
\leq & \frac{2\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q} \qquad (26)
\end{aligned}
$$

and

$$
\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}_\alpha) \right\|_{\tilde{q}} \geq \|\hat{\mu}_\alpha\|_{\tilde{q}} \qquad (27)
$$

with probability at least $1 - \alpha$; similarly, in terms of (10), we have

$$
\begin{aligned}
& \left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}_\alpha) \right\|_q - \hat{\mu}_\alpha \\
\leq & \left\| \frac{1}{n} X^T (Y - X\hat{\theta}_\alpha) \right\|_q - \hat{\mu}_\alpha + \left\| \frac{1}{n} X^T W \right\|_q \\
\leq & \frac{2\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q} \qquad (28)
\end{aligned}
$$

and

$$
\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}_\alpha) \right\|_q \geq \hat{\mu}_\alpha \qquad (29)
$$

with probability at least $1 - \alpha$. The argument for (27) and (29) is identical to what is used to show (16) and (18). As we have pointed out in the introduction, there might be different $\hat{\theta}_\alpha$s that satisfy (9) (or (10)) while producing the same (minimal) objective value $\|\hat{\mu}_\alpha\|_{\tilde{q}}$ (respectively, $\hat{\mu}_\alpha$). Consequently, there is more than one confidence region in the form of (26)-(27) or (28)-(29).

If $\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ can be known exactly and we were able to set $r_{\alpha_1,q} = r^*_{\alpha_1,q}$ in (9) (or (10)) as in Theorem 2.1, then any resulting optimal solution $\left(\hat{\theta}^*_\alpha, \hat{\mu}^*_\alpha\right)$ to (9) (respectively (10)) should satisfy

$$
\begin{aligned}
\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) - \hat{\mu}^*_\alpha \right\|_q & \leq 2\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + 2\tau_{\alpha_1,q}, \qquad (30) \\
\left\| \frac{1}{n} X^T X(\theta^* - \hat{\theta}^*_\alpha) \right\|_q - \hat{\mu}^*_\alpha & \leq 2\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + 2\tau_{\alpha_1,q}, \qquad (31)
\end{aligned}
$$

both with probability at least $1 - \alpha_1$. Comparing (26) with (30) and (28) with (31), note that the confidence interval based on (the practical) $\hat{\theta}_\alpha$ is widened by

$$
2 \left(\frac{\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q - \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] \right) + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q},
$$

which can be made arbitrarily small with a large number of random draws in the Monte-Carlo approximation. Because of such an approximation, the probabilistic guarantees for (26) and (28) are bounded from below by $1-\alpha$ instead of $1-\alpha_1$.

Given the statistics $\Psi_q(\hat{\theta}_\alpha)$ based on (a practical) $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (24), we have constructed a test with level α as shown in (25). For some chosen $\beta \in (0,\, 1)$, when can this test correctly detect an alternative with probability at least $1-\beta$? To answer this question, we introduce the "Separation Requirement" in the following section.

Separation Requirement and Type II Error

Letting $\Theta_0 := \{\theta \in \mathbb{R}^p \,:\, h(\theta) = \mathbf{0}_m\}$, we choose $\beta_1,\, \beta_2 > 0$ such that $\beta_1 + \beta_2 = \beta \in (0,\, 1)$, and assume

$$\inf_{\theta \in \Theta_0} \left\| \frac{1}{n} X^T X(\theta^* - \theta) \right\|_q \geq \delta_{\beta,q} \tag{32}$$

with

$$\delta_{\beta,q} = 2\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}} \tau_{\alpha_2,q} + \sqrt{\frac{1}{R}} \tau_{\beta_1,q} + \tau_{\beta_2,q} \tag{33}$$

for the prespecified $\alpha_1,\, \alpha_2 > 0$ (as used in (24)) such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$. We will refer to (32) as the "Separation Requirement" (SR) at the level β. In view of (13), note that the SR is imposed upon the l_q-distance between the population score vectors evaluated at θ^* and $\theta (\in \Theta_0)$.

Remarks. If we are interested in testing $H_0 : \sum_{j \in G} \theta_j^* x_{ij} = a$ or $H_{0,G} : \theta_j^* = \theta_j^0 \; \forall j \in G \subseteq \{1, 2, ..., p\}$, it may be helpful in practice to multiply the covariates X by M, a diagonal $p \times p$ matrix with diagonal entries $d_j = \left(\max_{i \in \{1,\ldots,n\}} |X_{ij}| \right)^{-1}$ $(j = 1, ..., p)$. Consequently, we will work with the rescaled hypothesis $H_{0,G} : d_j^{-1}\theta_j^* = d_j^{-1}\theta_j^0 \; \forall j \in G \subseteq \{1, 2, ..., p\}$; for the other hypothesis $H_0 : \sum_{j \in G} \theta_j^* x_{ij} = a$, since x_{ij}s (and θ_j^*s) will be multiplied by d_j (respectively, divided by d_j), there is no change on the form of the original hypothesis.

Our next result concerns the Type II error of the test based on $\Psi_q(\hat{\theta}_\alpha)$ and $r_{\alpha,q}$ defined in (24). For completeness, we also include the claim for the Type I error.

Theorem 2.1. *Assume (12) where $W \sim \mathcal{N}(0,\, \sigma^2 \boldsymbol{I}_n)$ and is independent of X. For some chosen $\alpha_1,\, \alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$, consider the statistics $\Psi_q(\hat{\theta}_\alpha)$ based on (a practical) $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (24). For any $q \in [1,\, \infty]$, we have*

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha, \qquad \textit{(Type I Error)} \tag{34}$$

where $\mathbb{P}_0$ means "under H_0". For the same $r_{\alpha,q}$ used in (34) and some chosen $\beta_1,\, \beta_2 > 0$ such that $\beta_1 + \beta_2 = \beta \in (0,\, 1)$, if $h(\theta^*) \neq \mathbf{0}_m$ and (32) is satisfied, we

have

$$\mathbb{P}_1\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}\right\} \leq \beta, \qquad \textit{(Type II Error)} \tag{35}$$

where $\mathbb{P}_1$ *means "under* H_1*".*

Some interesting implications can be drawn from Theorem 2.1. First, the results above do not rely on good behavior of $\left\|\hat{\theta}_\alpha - \theta^*\right\|_2$ or even $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ nonasymptotically or asymptotically. Second, we observe from (33) and (23) that the quantities taking the form of $\sqrt{\log\frac{1}{\pi}}$ in $\delta_{\beta,q}$ are dimension free while the leading term $2\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ in (33) captures the "dimension complexity". This result is a direct consequence of the concentration phenomenon in Lipschitz functions of Gaussians. At the expense of incurring a small deviation term, we have demonstrated that $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$ can be well approximated by the data-driven threshold $\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_\infty$, which automatically takes into consideration the dependencies between the coordinates. In particular, for the case $q = \infty$, we show in Section 7.4 that[1]

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \asymp \sqrt{\frac{\log p}{n}} \tag{36}$$

when $W \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. While the nonasymptotic validity of our testing procedure does not require any growth restrictions on the dimensionality, we see from (36) that the separation $\delta_{\beta,q}$ tends to zero only when $\frac{\log p}{n} = o(1)$.

As an alternative, the Bonferroni approach can also be used to construct a testing procedure. In particular, we can solve (9) (or (10)) with $q = \infty$ and

$$r_{\alpha,\infty} = \sqrt{\max_{j\in\{1,\ldots,p\}}\frac{2\sigma^2}{n}\sum_{i=1}^{n}X_{ij}^2}\sqrt{\frac{1}{n}\log\frac{2p}{\alpha}}. \tag{37}$$

Consequently, the separation distance in (32) that allows us to correctly detect an alternative with probability at least $1-\beta$ takes the form

$$\begin{aligned}\delta_{\beta,\infty} &= r_{\alpha,\infty} + r_{\beta,\infty} \\ &= \sqrt{\max_{j\in\{1,\ldots,p\}}\frac{2\sigma^2}{n}\sum_{i=1}^{n}X_{ij}^2}\left(\sqrt{\frac{1}{n}\log\frac{2p}{\alpha}} + \sqrt{\frac{1}{n}\log\frac{2p}{\beta}}\right).\end{aligned} \tag{38}$$

In contrast to our previous concentration approach, the Bonferroni alternative derives the upper bound (37) from a simple union bound on $\left\|\frac{1}{n}X^TW\right\|_\infty$; as a consequence, the resulting threshold $r_{\alpha,\infty}$ depends on p and fails to capture the dependencies between the coordinates.

[1]We write $f(n) \succsim g(n)$ if $f(n) \geq C_0 g(n)$ for some constant $C_0 \in (0,\infty)$, $f(n) \precsim g(n)$ if $f(n) \leq C_1 g(n)$ for some constant $C_1 \in (0,\infty)$, and $f(n) \asymp g(n)$ if $f(n) \succsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously.

Unknown Noise Variance

When there is no prior information on σ, $\sqrt{\mathrm{Var}\,(Y_i)}$ may be used as an upper bound. We can easily estimate $\sqrt{\mathrm{Var}\,(Y_i)}$ by $\hat{\sigma}_Y = \sqrt{n^{-1}\sum\left(Y_i - \bar{Y}\right)^2}$. Proposition 4.1 in [1] implies that

$$\sqrt{\mathrm{Var}\,(Y_i)} \leq \left(C_n - \frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{\gamma}{2}\right)\right)^{-1}\hat{\sigma}_Y := \bar{B}_\gamma$$

with probability at least $1-\gamma$, where $C_n = \sqrt{\frac{2}{n}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)} = 1 - O(n^{-1})$.

In problems where X is a fixed design and the only source of randomness in Y comes from W, replacing σ with $\hat{\sigma}_Y$ does not make $r_{\alpha,q}$ a more conservative threshold for constructing confidence regions if $\mathrm{Var}(W_i)$ is a constant over i. In problems with a random design, using $\hat{\sigma}_Y$ could result in confidence regions that are more conservative.

We find it rather challenging to estimate σ precisely and obtain a sharp threshold simultaneously within the non-asymptotic framework. The main issue is that our procedure does not guarantee a small $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ with high probability, which seems to be needed for consistent estimation of σ. On the other hand, if we were able to ensure a small error with respect to the prediction norm, our nonasymptotic control is likely to become less sharper and also involves unknown nuisance parameters that are hard to estimate.

So far our analysis has focused on Gaussian linear regressions with homoscedastic noise. Is it possible to extend our non-asymptotic framework to allow for non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients? We answer this question in Section 4.

3 Binary Classifications

In this section, we specialize (9) (or (10)) to (2) by letting $\mathit{\Lambda}\left(X_i;\theta_\alpha\right) = \mathit{\Pi}\left(X_i;\theta_\alpha\right)$. Our test statistics (8) now becomes

$$\Psi_q\left(\hat{\theta}_\alpha\right) := \left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right]\right\|_q.$$

Throughout this section (and also Section 5), we use $\mathbb{E}_{Y|X}[\cdot]$ to denote the expectation over the distribution of Y conditioning on X; for an i.i.d. sequence of Radamacher random variables, $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ (independent of Y and X), we use $\mathbb{E}_\varepsilon[\cdot]$ to denote the expectation over ε only, conditioning on Y and X, and $\mathbb{E}_{\varepsilon,Y|X}[\cdot]$ to denote the expectation over the distribution of $(\varepsilon,\, Y)$ conditioning on X.

Like in the regression problem, we first establish the concentration of

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \mathit{\Pi}\left(X_i;\theta^*\right)\right]\right\|_q$$

around its expectation

$$S_{\theta^*} := \mathbb{E}_{Y|X}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q\right]. \tag{39}$$

Previously we have simply replaced $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ in (14) with its Monte Carlo approximation $\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q$ (or $\frac{\bar{B}_\gamma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q$ when prior information on σ is not available) and a "small" deviation. This strategy cannot be applied to the expectation S_{θ^*} directly. Instead, we first seek a reasonable upper bound which involves only $\{Y,\, X\}$ and random variables from a known distribution. These results are stated in the following proposition.

Proposition 3.1. *For any $q \in [1,\, \infty]$, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q \geq S_{\theta^*} + t\right\} \leq \exp\left(\frac{-nt^2}{2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q^2}\right). \tag{40}$$

Let $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables independent of Y and X. Under (2), we have

$$\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_i X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q\right\} \leq S_{\theta^*} \leq 2\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i Y_i\right\|_q\right\}. \tag{41}$$

Remarks. Note that bound (40) holds for any fixed θ (not just the true coefficient vector, θ^*). However, (41) relies crucially on the model assumption $\mathbb{E}_{Y|X}\left(Y_i\right) = \Pi\left(X_i;\theta^*\right)$, as implied by (2).

The upper bound in (41) can be viewed as the symmetrized version of S_{θ^*}. Considering a collection of i.i.d. Radamacher random draws (independent of Y and X),

$$\left\{\varepsilon_{ir} :\, i = 1, ..., n,\, r = 1, ..., R\right\}, \tag{42}$$

we can replace S_{θ^*} with $\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q$ (a Monte-Carlo approximation of the symmetrized version) and some "small" deviations. The complementary lower bound in (41) suggests that S_{θ^*} and its symmetrized version have the magnitude. As a consequence, our replacement strategy is not an overly conservative approach for constructing critical values.

Hypothesis Testing

To avoid repetition, we omit the discussion on the "ideal" confidence regions and directly jump to the construction of the test statistics $\Psi_q(\hat{\theta}_\alpha)$ based on (a practical)

$r_{\alpha,q}$ and $\hat{\theta}_\alpha$. The first step is to relate S_{θ^*} with $\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q$ as shown in the following proposition. Like $\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q$, we define

$$\begin{aligned}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_iX_i)^2}\right\|_q &= \sqrt[q]{\sum_{j=1}^{p}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2X_{ij}^2}\right)^q}, \qquad q\in[1,\,\infty)\\ \left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_iX_i)^2}\right\|_q &= \max_{j\in\{1,\ldots,p\}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2X_{ij}^2}, \qquad q=\infty.\end{aligned}$$

Proposition 3.2. *Given (2) and (42) which is independent of Y and X, for any $q\in[1,\,\infty]$, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i-\mathit{\Pi}\left(X_i;\theta^*\right)\right]\right\|_q \geq \frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q + t_1+2t_2+2t_3 \tag{43}$$

with probability no greater than $\alpha\in(0,\,1)$, where

$$\begin{aligned}t_1 &= \tau_{\alpha_1,q}=\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_1}},\\ t_2 &= \tau_{\alpha_2,q}=\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_2}},\\ t_3 &= \left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_iX_i)^2}\right\|_q\sqrt{\frac{8}{nR}\log\frac{1}{\alpha_3}}:=\tau^{\dagger}_{\alpha_3,q},\end{aligned}$$

for some chosen α_1, α_2, $\alpha_3>0$ such that $\sum_{k=1}^{3}\alpha_k=\alpha$.

Construction of Critical Values ($r_{\alpha,q}$) and Type I Error

Based on (43) along with the choices of t_1, t_2 and t_3 above, we set in (9) (or (10)),

$$r_{\alpha,q}=\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q+\tau_{\alpha_1,q}+2\tau_{\alpha_2,q}+2\tau^{\dagger}_{\alpha_3,q}. \tag{44}$$

Under H_0, $(\theta^*,\,\mathbf{0}_p)$ $((\theta^*,\,0))$ is an optimal solution to (9) (respectively, (10)) with $r_{\alpha,q}$ specified in (44). Consequently, a (practical) optimal solution to (9) (and (10)) must satisfy

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha)\geq r_{\alpha,q}\right\}\leq\alpha \qquad \textit{(Type I Error)}. \tag{45}$$

Separation Requirement and Type II Error

Letting $\Theta_0 := \{\theta \in \mathbb{R}^p : h(\theta) = \mathbf{0}_m\}$, we choose $\beta_1, \beta_2, \beta_3 > 0$ such that $\sum_{k=1}^3 \beta_k = \beta \in (0, 1)$, and assume

$$\inf_{\theta \in \Theta_0} \left\| \frac{1}{n} \sum_{i=1}^n X_i \left[\Pi\left(X_i; \theta^*\right) - \Pi\left(X_i; \theta\right) \right] \right\|_q \geq \delta_{\beta,q} \tag{46}$$

with

$$\begin{aligned} \delta_{\beta,q} = \mathbb{E}_{Y|X} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\|_q \right] + \mathbb{E}_{\varepsilon, Y|X} \left[\left\| \frac{2}{n} \sum_{i=1}^n \varepsilon_i X_i Y_i \right\|_q \right] \\ + \tau_{\alpha_1,q} + 2\tau_{\alpha_2,q} + \sqrt{\frac{16}{R}} \tau_{\alpha_3,q} + 2\tau_{\beta_1,q} + \sqrt{\frac{16}{R}} \tau_{\beta_2,q} + \tau_{\beta_3,q}, \end{aligned} \tag{47}$$

for the prespecified $\alpha_1, \alpha_2, \alpha_3 > 0$ (as used in (44)) such that $\sum_{k=1}^3 \alpha_k = \alpha \in (0, 1)$. Note that the SR is imposed upon the l_q-distance between the "quasi score" vectors evaluated at θ^* and $\theta (\in \Theta_0)$, since

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n X_i \left[\Pi\left(X_i; \theta^*\right) - \Pi\left(X_i; \theta\right) \right] \right\|_q \\ = & \left\| \mathbb{E}_{Y|X} \left\{ \frac{1}{n} \sum_{i=1}^n X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\} - \mathbb{E}_{Y|X} \left\{ \frac{1}{n} \sum_{i=1}^n X_i \left[Y_i - \Pi\left(X_i; \theta\right) \right] \right\} \right\|_q . \end{aligned}$$

Our next result concerns the Type II error of the test based on $\Psi_q(\hat{\theta}_\alpha)$ and $r_{\alpha,q}$ defined in (44). For completeness, we also exhibit the Type I error and the practical confidence regions in this result.

Theorem 3.1. *Suppose the conditions in Proposition 3.2 hold. For some chosen* $\alpha_1, \alpha_2, \alpha_3 > 0$ *such that* $\sum_{k=1}^3 \alpha_k = \alpha \in (0, 1)$, *consider the statistics* $\Psi_q(\hat{\theta}_\alpha)$ *based on (a practical)* $\hat{\theta}_\alpha$ *and the critical value* $r_{\alpha,q}$ *defined in (44). For any* $q \in [1, \infty]$, *we have*

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha, \qquad \textit{(Type I Error)} \tag{48}$$

where $\mathbb{P}_0$ *means "under* H_0*". For the same* $r_{\alpha,q}$ *used in (48) and some chosen* $\beta_1, \beta_2, \beta_3 > 0$ *such that* $\sum_{k=1}^3 \beta_k = \beta \in (0, 1)$, *if* $h(\theta^*) \neq \mathbf{0}_m$ *and (46) is satisfied, we have*

$$\mathbb{P}_1 \left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} \right\} \leq \beta, \qquad \textit{(Type II Error)} \tag{49}$$

where $\mathbb{P}_1$ *means "under* H_1*".*

Furthermore, an optimal solution $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ *to (9) must satisfy*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right] \right\|_{\tilde{q}} \;\geq\; \left\|\hat{\mu}_\alpha\right\|_{\tilde{q}}, \tag{50}$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right] - \hat{\mu}_\alpha \right\|_{q} \;\leq\; 2r_{\alpha,q}, \tag{51}$$

with probability at least $1-\alpha$*. Similarly, an optimal solution* $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ *to (10) must satisfy*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right] \right\|_{q} \;\geq\; \hat{\mu}_\alpha, \tag{52}$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right] \right\|_{q} \;\leq\; 2r_{\alpha,q} + \hat{\mu}_\alpha, \tag{53}$$

with probability at least $1-\alpha$*.*

4 Regression with Heteroscedasticity, Non-Gaussian Responses, and Nonlinearity

In this section, we extend our framework in Section 2 to allow for non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients. Often we would have more information on the distribution of Y than the distribution of W. In some applications, we might only know Y consists of entries supported on $[a,\, b]$. For example, [19] estimate the effect of spending on math pass rates ($Y_i \in [0,\, 1]$) under the assumption $\mathbb{E}\left(Y_i|X_i\right) = \Phi\left(X_i\theta^*\right)$, where $\Phi(\cdot)$ denotes the standard normal c.d.f. and X_i include the spending variable as well as other covariates. We could use (1) with $\mathit{\Upsilon}\left(X_i;\theta^*\right) = \Phi\left(X_i\theta^*\right)$ to model this problem. In other applications, we might know that Y has a strongly log-concave distribution with parameter $\varphi > 0$ (so the entries of Y are possibly unbounded).

In either case, without knowing the exact distribution of Y, we can still obtain the following analogues of (40).

Lemma 4.1*. Assume (1) where* Y *has a strongly log-concave distribution*[2] *with*

[2] A strongly log-concave distribution is a distribution with density $\mathsf{p}(z) = \exp\left(-\psi(z)\right)$ such that for some $\varphi > 0$ and all $\lambda \in [0,\, 1]$, $z,\, z^{'} \in \mathbb{R}^n$,

$$\lambda\psi(z) + (1-\lambda)\psi(z^{'}) - \psi(\lambda z + (1-\lambda)z^{'}) \geq \frac{\varphi}{2}\lambda(1-\lambda)\left\|z - z^{'}\right\|_2^2.$$

parameter φ. Then for any $q \in [1, \infty]$, we have

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q \geq \mathbb{E}_{Y|X}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q\right] + t\right\}$$
$$\leq \exp\left(\frac{-n\varphi t^2}{2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}\right\|_q^2}\right). \tag{54}$$

Remarks. For a fixed design X, if $Y \sim \mathcal{N}(X\theta^*, \Sigma)$ and $\Sigma \succ 0$, φ can be set to the smallest eigenvalue of Σ^{-1}. Beyond a normal distribution, [16] discuss quite a few examples of strongly log-concave distributions.

Lemma 4.2. *Assume (1) where Y consists of independent random variables, all of which are supported on $[a, b]$. Then for any $q \in [1, \infty]$, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q \geq \mathbb{E}_{Y|X}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q\right] + t\right\}$$
$$\leq \exp\left(\frac{-nt^2}{2\left(b-a\right)^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}\right\|_q^2}\right). \tag{55}$$

The trick we have exploited to bound $\mathbb{E}_{Y|X}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \mathit{\Pi}\left(X_i;\theta^*\right)\right]\right\|_q\right]$ in Section 3 can be used here to bound

$$Q_{\theta^*} = \mathbb{E}_{Y|X}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q\right]. \tag{56}$$

Consequently, we have the following result, for which we will assume the availability of an upper bound ζ on $|b-a|^2$ (respectively, on φ^{-1}).

Proposition 4.1. *Let $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables, independent of Y and X. Under (1), we have*

$$\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_i X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q\right\} \leq Q_{\theta^*} \leq 2\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i Y_i\right\|_q\right\} \tag{57}$$

for any $q \in [1, \infty]$. Suppose Y consists of independent random variables, all of which are supported on $[a, b]$; or Y has a strongly log-concave distribution with parameter φ. Given Radamacher random draws, i.e., (42), independent of Y and X, for any

$q \in [1, \infty]$, *we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Upsilon \left(X_i; \theta^* \right) \right] \right\|_q \geq \frac{2}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q + t_1 + 2t_2 + 2t_3 \tag{58}$$

with probability no greater than $\alpha \in (0, 1)$, *where*

$$\begin{aligned} t_1 &= \left\| \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2} \right\|_q \sqrt{\frac{2\zeta}{n} \log \frac{1}{\alpha_1}}, \\ t_2 &= \left\| \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2} \right\|_q \sqrt{\frac{2\zeta}{n} \log \frac{1}{\alpha_2}}, \\ t_3 &= \left\| \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(Y_i X_i \right)^2} \right\|_q \sqrt{\frac{8}{nR} \log \frac{1}{\alpha_3}}, \end{aligned}$$

for some chosen α_1, α_2, $\alpha_3 > 0$ *such that* $\sum_{k=1}^{3} \alpha_k = \alpha$.

Note that the assumptions in Proposition 4.1 allow for the possibilities of heteroscedastic noise as well as nonlinearity in θ^*, while requiring no specific knowledge on the distribution for Y (other than it is bounded or has a strongly log-concave distribution). We can specialize (9) (or (10)) to (1) by letting $\Lambda \left(X_i; \theta_\alpha \right) = \Upsilon \left(X_i; \theta_\alpha \right)$. Our test statistics (8) then takes the form

$$\Psi_q \left(\hat{\theta}_\alpha \right) := \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Upsilon \left(X_i; \hat{\theta}_\alpha \right) \right] \right\|_q .$$

Based on (58), we can apply our previous argument in Section 3 to construct tests and confidence regions here.

In the linear regression model $Y = X\theta^* + W$, [9] resolve the issues of heteroscedasticity and non-Gaussian responses by tailoring the Bonferroni approach to self-normalized sums. Their confidence regions involve several unknown nuisance parameters that are hard to estimate in practice. Even in the case where the noise variances are known and homoscedastic, to apply the confidence sets in [9] for testing hypotheses of the form $H_{0,G} : \theta_j^* = \theta_j^0 \ \forall j \in G \subseteq \{1, 2, ..., p\}$, one would require sufficient sparsity in θ^* as well as prior knowledge on the underlying sparsity (e.g., an upper bound on the number of zero coefficients in θ^*).

For deriving $r_{a,q}$ in (9) or (10), the strategy where we replace S_{θ^*} in (39) and Q_{θ^*} in (56) by $\frac{2}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q$ plus some "small" deviations only requires the correct specification of the conditional mean of Y_i; that is, $\mathbb{E}_{Y|X} \left(Y_i \right) = \Pi \left(X_i; \theta^* \right)$, as implied by (2), and $\mathbb{E}_{Y|X} \left(Y_i \right) = \Upsilon \left(X_i; \theta^* \right)$, as implied by (1). This treatment

delivers generic confidence regions in the form of (50)-(51) or (52)-(53). While these general results could be applied to the linear regression model with homoscedastic Gaussian noise, the method we have developed in Section 2 yields confidence regions that are more accurate in terms of constants as (26)-(27) or (28)-(29) exploit the Gaussian distribution of the noise vector as well as the structure of linear models (where the quasi-score vector coincides with the score vector).

5 A New Class of Regularized Estimators

Beyond the context of hypothesis testing, the data-driven approach proposed in this paper for setting $r_{\alpha,q}$ suggests a new class of regularized estimators, which solve the following program:

$$\min_{\theta_\alpha \in \mathbb{R}^p} \|\theta_\alpha\|_{\tilde{q}} \ \text{ subject to} \left\| \frac{1}{n} X^T (Y - X\theta_\alpha) \right\|_q \leq r_{\alpha,q}, \tag{59}$$

where

$$r_{\alpha,q} = \frac{\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}} \tau_{\alpha_2,q}, \tag{60}$$

$$\tau_{\alpha_1,q} = \sigma \left\| \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2} \right\|_q \sqrt{\frac{2}{n} \log \frac{1}{\alpha_1}}, \tag{61}$$

for some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$.

Unlike (9) or (10), (59) and later (65) have a different objective, $\min_{\theta_\alpha \in \mathbb{R}^p} \|\theta_\alpha\|_{\tilde{q}}$, and do not involve the slack vector (or variable) μ_α in $\left\| \frac{1}{n} X^T (Y - X\theta_\alpha) \right\|_q \leq r_{\alpha,q}$. Consequently, the resulting solution to (59) is an unconstrained estimator, in contrast to our constrained estimator $\hat{\theta}_\alpha$ in Sections 2-4 where $h(\hat{\theta}_\alpha) = \mathbf{0}_m$ needs to be satisfied. When $\tilde{q} = 1$ and $q = \infty$, we may view (59) as a variant of the Dantzig selector.

In what follows, let $\hat{\theta}_\alpha^{new}$ be a solution to the program (59) with $\tilde{q} = 1$ and $q = \infty$. We can establish an upper bound on $\left\| \hat{\theta}_\alpha^{new} - \theta^* \right\|_2$ using the l_2-sensitivity defined as follows:

$$\kappa_{J_*} := \inf_{\Delta \in \mathbb{C}_{J_*} : \|\Delta\|_2 = 1} \left\| \frac{1}{n} X^T X \Delta \right\|_\infty \tag{62}$$

where

$$\begin{aligned} J_* &:= \left\{ j \in \{1, ..., p\} : \theta_j^* \neq 0 \right\}, \\ \mathbb{C}_{J_*} &:= \left\{ \Delta \in \mathbb{R}^p : \|\Delta_{J_*^c}\|_1 \leq \|\Delta_{J_*}\|_1 \right\}, \end{aligned}$$

where Δ_{J_*} denotes the vector in $\mathbb{R}^p$ that has the same coordinates as Δ on J_* and zero coordinates on the complement J_*^c of J_*. The l_2-sensitivity is introduced by

[8][3] and similar to the cone invertibility factors defined in [20]. In particular, under a coherence condition introduced by [7], Proposition 4.2 in [8] shows that

$$\kappa_{J_*} \gtrsim \frac{1}{\sqrt{|J_*|}} \tag{63}$$

where $|J_*|$ denotes the cardinality of J_*.

The following result concerns the l_2-error bound for $\hat{\theta}_\alpha^{new}$.

Theorem 5.1. *Assume (12) where $W \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ and is independent of X. Choosing $\tilde{q} = 1$ and $q = \infty$ in (59) and setting $r_{\alpha,q}$ according to (60) with $q = \infty$, we have*

$$\mathbb{P}\left(\left\|\hat{\theta}_\alpha^{new} - \theta^*\right\|_2 \leq \frac{2r_{\alpha,\infty}}{\kappa_{J_*}}\right) \geq 1 - \alpha \tag{64}$$

where κ_{J_*} is defined in (62).

In view of (36) and (63), we see that the rate of our $\hat{\theta}_\alpha^{new}$, i.e., $\kappa_{J_*}^{-1}\sqrt{\frac{\log p}{n}}$, is not worse than the typical rate $\sqrt{\frac{|J_*|\log p}{n}}$ for estimation (see, e.g., [2]). For a fixed $\alpha > 0$, $\kappa_{J_*}^{-1}\sqrt{\frac{\log p}{n}} = o(1)$ is required for the l_2-consistency of $\hat{\theta}_\alpha^{new}$. If $|J_*|$ is large relative to n (lack of sparsity), then $\kappa_{J_*}^{-1}$ could diverge faster than (or no slower than) $\sqrt{\frac{n}{\log p}}$.

The innovation of (59) lies in the use of (60) which can accurately approximate the term $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$ in (14) via Monte-Carlo and automatically take into consideration the dependencies across coordinates. This fact makes (59) in contrast with the Bonferroni approach which would set $r_{\alpha,\infty}$ proportional to $\sqrt{\frac{1}{n}\log\frac{2p}{\alpha}}$. In the situation where the noise variance σ is not known *a priori*, we can always modify (59) by adopting the approach in Section 2. Alternatively, it is also possible to modify the optimization procedure in [9] with our data-driven approach for setting the constraint on $\left\|\frac{1}{n}X^T(Y - X\theta_\alpha)\right\|_\infty$.

Remarks. Note that the confidence interval in (64) cannot be computed easily as κ_{J_*} is unknown and hard to estimate. Even for testing a simple hypothesis such as $H_0 : \theta^* = \mathbf{0}$, deriving a practical critical value for the statistics $\left\|\hat{\theta}_\alpha^{new}\right\|_2$ is a challenging task. For this reason, we have chosen to work with the quasi-score tests (which require no conditions on sparsity) as demonstrated in Sections 2-4.

For the binary response models considered in Section 3 and the regression models considered in Section 4, (43) and (58) allow us to also develop a new class of l_1-regularized estimators based on the data-driven approach for setting $r_{\alpha,\infty}$. These

[3]In contrast to (59), the estimators in [8] and [9] rely on the Bonferroni approach tailored to the self-normalized sums.

estimators solve the following program

$$\min_{\theta_\alpha \in \mathbb{R}^p} \|\theta_\alpha\|_1 \text{ subject to } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Lambda \left(X_i; \theta_\alpha \right) \right] \right\|_\infty \leq r_{\alpha,\infty}, \tag{65}$$

where $\Lambda = \Pi$ (for the binary response models in Section 3) or $\Lambda = \Upsilon$ (for the regression models in Section 4), and given (42),

$$r_{\alpha,\infty} = \frac{2}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_\infty + \tau_{\alpha_1,\infty} + 2\tau_{\alpha_2,\infty} + 2\tau^{\dagger}_{\alpha_3,\infty},$$

for some chosen α_1, α_2, $\alpha_3 > 0$ such that $\sum_{k=1}^{3} \alpha_k = \alpha \in (0, 1)$.

With additional effort, we can establish an analogue of Theorem 5.1 and a solution to (65) will have similar implications; we omit the detail of such a result here to focus on the non-asymptotic inference.

6 Conclusions

We have developed non-asymptotically justified methods for hypothesis testing about the coefficients ($\theta^* \in \mathbb{R}^p$) in the high dimensional generalized regression models where p can exceed the sample size n. Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, we consider

$$H_0 : h(\theta^*) = \mathbf{0}_m \; v.s. \; H_1 : h(\theta^*) \neq \mathbf{0}_m,$$

where m can be any integer in $[1, p]$ and h is allowed to be nonlinear in θ^*.

Our test statistics is based on the sample "quasi score" vector evaluated at an estimate $\hat{\theta}_\alpha$ that satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$, where α is the prespecified Type I error. By exploiting the concentration phenomenon in Lipschitz functions, the key component reflecting the "dimension complexity" in our non-asymptotic thresholds uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated around and automatically captures the dependencies between the coordinates. We provide probabilistic guarantees in terms of the Type I and Type II errors for the "quasi score" test. In addition, confidence regions are constructed for the population quasi-score vector evaluated at θ^*.

The results in Section 2 are specific to the standard Gaussian linear regression models; the results in Sections 3 and 4 allow for reasonably flexible forms of non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients (including the binary response models and certain nonlinear regressions), while only requiring the correct specification of $\mathbb{E}\left(Y_i | X_i\right)$s. The novelty of our methods is that their validity does not rely on good behavior of $\left\| \hat{\theta}_\alpha - \theta^* \right\|_2$ (or even $n^{-1/2} \left\| X \left(\hat{\theta}_\alpha - \theta^* \right) \right\|_2$ in the linear regression case) nonasymptotically or asymptotically.

7 Proofs

7.1 Preliminary

Here we include several classical results which are used in the main proofs. We first introduce a definition of sub-Gaussian variables.

Definition 7.1. *A zero-mean random variable U_1 is sub-Gaussian if there is a $\nu > 0$ such that*

$$\mathbb{E}\left[\exp\left(\lambda U_1\right)\right] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right) \tag{66}$$

for all $\lambda \in \mathbb{R}$, and we refer to ν as the sub-Gaussian parameter.

Remarks.

1. Using the Chernoff bound, one can show that any zero-mean random variable U_1 obeying (66) satisfies

$$\mathbb{P}\left(U_1 \leq -t\right) \leq \exp\left(-\frac{t^2}{2\nu^2}\right), \tag{67}$$

$$\mathbb{P}\left(U_1 \geq t\right) \leq \exp\left(-\frac{t^2}{2\nu^2}\right), \tag{68}$$

   for all $t \geq 0$.

2. Let $\{U_i\}_{i=1}^R$ be independent zero-mean sub-Gaussian random variables, each with parameter at most ν. Then $R^{-1}\sum_{i=1}^R U_i$ is sub-Gaussian with parameter at $\nu/\sqrt{R}$. To see this, note that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\lambda}{R}\sum_{i=1}^R U_i\right)\right] &= \prod_{i=1}^R \mathbb{E}\left[\exp\left(\frac{\lambda U_i}{R}\right)\right] \\ &\leq \prod_{i=1}^R \exp\left(\frac{\lambda^2\nu^2}{2R^2}\right) \\ &= \exp\left(\frac{\lambda^2\nu^2}{2R}\right).\end{aligned} \tag{69}$$

The following result exhibits the type of sub-Gaussian variables that are of interest to our analysis.

Lemma 7.1. *Suppose $U = \{U_i\}_{i=1}^n$ has a strongly log-concave distribution with parameter $\varphi > 0$ and $f : \mathbb{R}^n \to \mathbb{R}$ is $L-$Lipschitz with respect to Euclidean norm.*

Then for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[\exp\left(\lambda\left\{f(U) - \mathbb{E}\left[f(U)\right]\right\}\right)\right] \leq \exp\left(\frac{\lambda^2 L^2}{2\varphi}\right). \tag{70}$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left\{f(U) - \mathbb{E}\left[f(U)\right] \leq -t\right\} &\leq \exp\left(-\frac{\varphi t^2}{2L^2}\right), \\
\mathbb{P}\left\{f(U) - \mathbb{E}\left[f(U)\right] \geq t\right\} &\leq \exp\left(-\frac{\varphi t^2}{2L^2}\right).
\end{aligned}$$

Remarks. The proof involves the so-called "inf-convolution" argument and an application of the Brunn-Minkowski inequality; see [3] and [13].

Lemma 7.2. *Assume $U = \{U_i\}_{i=1}^n$ consists of independent random variables, all of which are supported on $[a, b]$. If $f : \mathbb{R}^n \to \mathbb{R}$ is separately convex*[4] *and $L-$Lipschitz with respect to the Euclidean norm, then for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(\lambda\left\{f(U) - \mathbb{E}\left[f(U)\right]\right\}\right)\right] \leq \exp\left[\frac{\lambda^2 (b-a)^2 L^2}{2}\right]. \tag{71}$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left[f(X) - \mathbb{E}\left[f(X)\right] \leq -t\right] &\leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right), \\
\mathbb{P}\left[f(X) - \mathbb{E}\left[f(X)\right] \geq t\right] &\leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).
\end{aligned}$$

Remarks. One proof for Lemma 7.2 involves the entropy method; see [4]. Talagrand and Ledoux have contributed to the result above in different papers.

7.2 Proof of Proposition 2.1

For any $q \in [1, \infty]$, $\left\|\frac{1}{n} X^T W\right\|_q$ is Lipschitz in W with respect to the Euclidean norm. To see this, note that a triangle inequality and a Cauchy-Schwarz inequality

[4]Let the function $f_j : \mathbb{R} \to \mathbb{R}$ be defined by varying only the jth co-ordinate of a function $f : \mathbb{R}^n \to \mathbb{R}$; f is *separately convex* if for each $j \in \{1, 2, ..., n\}$, f_j is a convex function of the jth coordinate.

yield

$$\begin{aligned}\left|\left\|\frac{1}{n}X^TW\right\|_q-\left\|\frac{1}{n}X^TW^{'}\right\|_q\right| &\leq \left\|\frac{1}{n}X^T\left(W-W^{'}\right)\right\|_q \\ &\leq \frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\left\|W-W^{'}\right\|_2. \qquad (72)\end{aligned}$$

As a result of Lemma 7.1, we have the concentration in (14).

If $h(\theta^*)=\mathbf{0}_m$, (14) then implies that $(\theta^*,\,\mathbf{0}_p)$ $((\theta^*,\,0))$ is an optimal solution to (9) (respectively, (10)). If $h(\theta^*)\neq\mathbf{0}_m$, as long as $\{\theta\in\mathbb{R}^p\,:\,h(\theta)=\mathbf{0}_m\}\neq\emptyset$, we can find some $\tilde{\theta}_\alpha$ such that $h(\tilde{\theta}_\alpha)=\mathbf{0}_m$. Letting

$$\tilde{\mu}_\alpha=\frac{1}{n}X^T(Y-X\tilde{\theta}_\alpha)-\frac{1}{n}X^T(Y-X\theta^*)=\frac{1}{n}X^T(X\theta^*-X\tilde{\theta}_\alpha),$$

(14) then implies that $\left(\tilde{\theta}_\alpha,\,\tilde{\mu}_\alpha\right)$ $(\left(\tilde{\theta}_\alpha,\,\|\tilde{\mu}_\alpha\|_q\right))$ is a feasible solution to (9) (respectively, (10)) with probability at least $1-\alpha$. In any case, an optimal solution $\left(\hat{\theta}_\alpha^*,\,\hat{\mu}_\alpha^*\right)$ to (9) must satisfy

$$\left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha^*)-\frac{1}{n}X^T(Y-X\theta^*)\right\|_{\tilde{q}}=\left\|\frac{1}{n}X^T(X\theta^*-X\hat{\theta}_\alpha^*)\right\|_{\tilde{q}}\geq\|\hat{\mu}_\alpha^*\|_{\tilde{q}}$$

with probability at least $1-\alpha$. Similarly, an optimal solution $\left(\hat{\theta}_\alpha^*,\,\hat{\mu}_\alpha^*\right)$ to (10) must satisfy

$$\left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha^*)-\frac{1}{n}X^T(Y-X\theta^*)\right\|_q=\left\|\frac{1}{n}X^T(X\theta^*-X\hat{\theta}_\alpha^*)\right\|_q\geq\hat{\mu}_\alpha^*$$

with probability at least $1-\alpha$. On the other hand, in terms of (9), applying the triangle inequality yields

$$\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}_\alpha^*)-\hat{\mu}_\alpha^*\right\|_q\leq\left\|\frac{1}{n}X^TW\right\|_q+\left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha^*)-\hat{\mu}_\alpha^*\right\|_q\leq 2r_{\alpha,q}^*$$

with probability at least $1-\alpha$. In terms of (10), we simply have

$$\mathbb{P}\left(\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}_\alpha^*)\right\|_q\leq 2r_{\alpha,q}^*+\hat{\mu}_\alpha^*\right)\geq 1-\alpha.$$

7.3 Proof of Theorem 2.1

We have already derived (34) in Section 2. To show (35), we define the event

$$\mathcal{E}=\left\{\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q\geq\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+\sqrt{\frac{1}{R}}\tau_{\beta_1,q}\right\}.$$

As we have argued for (21), we also have the upper deviation inequality

$$\mathbb{P}\left\{\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+t\right\} \leq \exp\left(\frac{-nRt^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q^2}\right)$$

and consequently, $\mathbb{P}(\mathcal{E}) \leq \beta_1$. Let $\mathcal{E}^c$ denote the complement of $\mathcal{E}$. Under H_1, we have

$$\begin{aligned}
&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\beta,q}\right\}\\
=&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\beta,q}|\mathcal{E}^c\right\}\mathbb{P}(\mathcal{E}^c)+\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\beta,q}|\mathcal{E}\right\}\mathbb{P}(\mathcal{E})\\
\leq&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\beta,q}|\mathcal{E}^c\right\}+\mathbb{P}(\mathcal{E})\\
\leq&\mathbb{P}\left\{\left\|\frac{1}{n}X^T(X\theta^*-X\hat{\theta}_\alpha)\right\|_q-\left\|\frac{1}{n}X^TW\right\|_q \leq r_{\beta,q}|\mathcal{E}^c\right\}+\beta_1\\
\leq&\mathbb{P}\left\{\delta_{\beta,q}-\left\|\frac{1}{n}X^TW\right\|_q \leq r_{\beta,q}|\mathcal{E}^c\right\}+\beta_1\\
\leq&\mathbb{P}\left\{\left\|\frac{1}{n}X^TW\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+\tau_{\beta_2,q}|\mathcal{E}^c\right\}+\beta_1\\
\leq&\beta
\end{aligned}$$

where the fifth line follows from (32) and the sixth line follows from (33), the fact that we are conditioning on $\mathcal{E}^c$, and (14).

7.4 Additional Derivations

To show (36), we define an i.i.d. sequence of Gaussian random variables

$$\widetilde{W}_k \sim \mathcal{N}\left(0, \min_{j,l\in\{1,...,p\}}\frac{1}{2n^2}\sum_{i=1}^{n}(X_{ij}-X_{il})^2\right)$$

for $k = 1, ..., p$. Since our design matrix X does not contain identical columns,

$$\mathrm{Var}\left(\widetilde{W}_k\right) = \min_{j,l\in\{1,...,p\}}\frac{1}{2n^2}\sum_{i=1}^{n}(X_{ij}-X_{il})^2 \neq 0.$$

For all $j \neq l$, we have

$$\mathbb{E}_W\left[\left(\frac{1}{n}X_j^TW-\frac{1}{n}X_l^TW\right)^2\right] \geq \mathbb{E}_{\widetilde{W}}\left[\left(\widetilde{W}_j-\widetilde{W}_l\right)^2\right].$$

By the Sudakov-Fernique Gaussian comparison result (see Corollary 3.14 in [12]), we obtain

$$\begin{aligned}\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] &\geq \mathbb{E}_W\left[\max_{j\in\{1,\ldots,p\}}\frac{1}{n}X_j^TW\right]\\ &\geq \frac{1}{2}\mathbb{E}_W\left[\max_{j\in\{1,\ldots,p\}}\widetilde{W}_j\right]\\ &\geq \frac{1}{2}\left(1-\frac{1}{e}\right)\sqrt{\frac{\log p}{4n^2}\min_{j,l\in\{1,\ldots,p\}}\sum_{i=1}^n\left(X_{ij}-X_{il}\right)^2}\end{aligned}$$

(for all $p \geq 20$), where the last line follows from a classical lower bound on the Gaussian maximum (see, e.g., [12]). The upper bound

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \leq \sqrt{\frac{2\log p}{n^2}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^n X_{ij}^2}+\sqrt{\frac{8}{n^2\log p}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^n X_{ij}^2}$$

(for all $p \geq 2$) is another classical result on the Gaussian maximum (see, e.g., [18]).

Remarks. To obtain the lower bound on $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$, we first compare the dependent sequence $\left\{\frac{1}{n}X_j^TW\right\}_{j=1}^p$ with another independent Gaussian sequence $\left\{\widetilde{W}\right\}_{j=1}^p$ and then apply a lower bound on $\mathbb{E}_{\widetilde{W}}\left[\max_{j\in\{1,\ldots,p\}}\widetilde{W}_j\right]$. In contrast, the upper bound on $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$ is obtained by applying $\sum_{j=1}^p\mathbb{P}\left(\left|\frac{1}{n}X_j^TW\right|\geq t\right)$, where independence is not needed. Moreover, the result $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]\precsim\sqrt{\frac{\log p}{n}}$ also holds when W is a sequence of sub-Gaussian variables while $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]\succsim\sqrt{\frac{\log p}{n}}$ requires W to be a sequence of Gaussian variables.

7.5 Proof of Proposition 3.1

Using the argument that leads to (72), we can show $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i-\Pi\left(X_i;\theta^*\right)\right]\right\|_q$ is Lipschitz in Y with respect to the Euclidean norm for any $q\in[1,\infty]$. That is,

$$\begin{aligned}&\left|\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i-\Pi\left(X_i;\theta^*\right)\right]\right\|_q-\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i^{'}-\Pi\left(X_i;\theta^*\right)\right]\right\|_q\right|\\ \leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q\left\|Y-Y^{'}\right\|_2.\end{aligned}\tag{73}$$

Note that $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i-\Pi\left(X_i;\theta^*\right)\right]\right\|_q$ is separately convex in terms of Y. As a result of Lemma 7.2, we have the concentration in (40).

To establish (41), we exploit the convexity of l_q-norms and the fact that $\mathbb{E}_{Y|X}\left(Y_i\right) = \mathit{\Pi}\left(X_i;\theta^*\right)$. Let $Y^{'} = \left\{Y_i^{'}\right\}_{i=1}^n$ be an i.i.d. sequence identical to but independent of Y conditioning on X, and $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be i.i.d. Radamacher random variables independent of Y, $Y^{'}$, and X. We obtain

$$
\begin{aligned}
&\mathbb{E}_{Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \mathit{\Pi}\left(X_i;\theta^*\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \mathbb{E}_{Y^{'}|X}\left(Y_i^{'}\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{Y|X}\left\{\left\|\mathbb{E}_{Y^{'}|X}\left[\frac{1}{n}\sum_{i=1}^n X_i\left(Y_i - Y_i^{'}\right)\right]\right\|_q\right\}\\
\leq&\mathbb{E}_{Y^{'},Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^n X_i\left(Y_i - Y_i^{'}\right)\right\|_q\right\}\\
=&\mathbb{E}_{\varepsilon,Y^{'},Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\left(Y_i - Y_i^{'}\right)\right\|_q\right\}\\
\leq&2\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right\}.
\end{aligned}
\tag{74}
$$

where the second line follows since $\mathbb{E}_{Y^{'}|X}\left(Y_i^{'}\right) = \mathit{\Pi}\left(X_i;\theta^*\right)$, the fourth line follows from Jensen's inequality, and the sixth line follows from the fact that $\varepsilon_i X_i\left(Y_i - Y_i^{'}\right)$ and $X_i\left(Y_i - Y_i^{'}\right)$ have the same distribution.

On the other hand, similar argument from above also yields

$$
\begin{aligned}
&\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^n \varepsilon_i X_i\left[Y_i - \mathit{\Pi}\left(X_i;\theta^*\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^n \varepsilon_i X_i\left[Y_i - \mathbb{E}_{Y^{'}|X}\left(Y_i^{'}\right)\right]\right\|_q\right\}\\
\leq&\mathbb{E}_{\varepsilon,Y^{'},Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^n \varepsilon_i X_i\left(Y_i - Y_i^{'}\right)\right\|_q\right\}\\
=&\mathbb{E}_{Y^{'},Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^n X_i\left(Y_i - Y_i^{'}\right)\right\|_q\right\}.
\end{aligned}
$$

Applying the following inequality

$$\begin{aligned}&\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\\ \leq&\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-\Pi\left(X_i;\theta^*\right)\right)\right\|_q+\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i^{'}-\Pi\left(X_i;\theta^*\right)\right)\right\|_q,\end{aligned}$$

and taking expectations gives

$$\mathbb{E}_{Y^{'}Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\leq\mathbb{E}_{Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-\Pi\left(X_i;\theta^*\right)\right)\right\|_q\right\}.$$

Putting the pieces together, we obtain the result in (41).

7.6 Proof of Proposition 3.2

We first show that $\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}$ is Lipschitz in Y with respect to the Euclidean norm for any $q\in[1,\,\infty]$. That is,

$$\begin{aligned}&\left|\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}-\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i^{'}\right\|_q\right\}\right|\\ \leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\mathbb{E}_\varepsilon\left[\sum_{i=1}^{n}\varepsilon_i^2\left(Y_i-Y_i^{'}\right)^2\right]}\\ \leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\left\|Y-Y^{'}\right\|_2.\end{aligned}$$

Note that $\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}$ is separately convex in terms of Y. As a result of Lemma 7.2, we have the following concentration

$$\mathbb{P}\left\{\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}\geq\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}+\tau_{\alpha_2,q}\right\}\leq\alpha_2\qquad(75)$$

Let $\varepsilon=\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables independent of Y and X. Conditioning on Y and X, we can again show that $\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iY_iX_i\right\|_q$ is Lipschitz in ε with respect to the Euclidean norm for any

$q \in [1, \infty]$ and the Lipschitz constant is $\frac{1}{\sqrt{n}} \left\| \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i X_i)^2} \right\|_q$, where

$$\begin{aligned} \left\| \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i X_i)^2} \right\|_q &= \sqrt[q]{\sum_{j=1}^{p} \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 X_{ij}^2} \right)^q}, \qquad q \in [1, \infty) \\ \left\| \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i X_i)^2} \right\|_q &= \max_{j \in \{1,...,p\}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 X_{ij}^2}, \qquad q = \infty. \end{aligned}$$

Let $\{\varepsilon_{ir} : i = 1, ..., n,\ r = 1, ..., R\}$ be a collection of i.i.d. Radamacher random draws. Conditioning on Y and X, (69) and (71) imply $\frac{1}{R}\sum_{r=1}^{R} \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q$ is sub-Gaussian with parameter at most $\frac{2}{\sqrt{nR}} \left\| \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i X_i)^2} \right\|_q$. Consequently, (67) yields the following concentration

$$\mathbb{P}\left\{ \mathbb{E}_\varepsilon \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i Y_i \right\|_q \right\} \geq \frac{1}{R}\sum_{r=1}^{R} \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q + \tau^{\dagger}_{\alpha_3, q} \right\} \leq \alpha_3 \tag{76}$$

Combining (40), (74), (75) and (76) yields (43).

7.7 Proof of Theorem 3.1

We have already derived (48) in Section 3. For the confidence regions in Theorem 3.1, we simply follow the same argument used in the proof for Proposition 2.1.

To show (49), let us define the event

$$\mathcal{E} = \left\{ \frac{1}{R}\sum_{r=1}^{R} \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q \geq \mathbb{E}_{\varepsilon, Y|X} \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i Y_i \right\|_q \right\} + \tau_{\beta_1, q} + \tau^{\dagger}_{\beta_2, q} \right\}$$

for some chosen β_1, $\beta_2 > 0$ such that $\beta_1 + \beta_2 \in (0, 1)$. As we have argued for (75) and (76), we also have the upper deviation result $\mathbb{P}\{\mathcal{E}\} \leq \beta_1 + \beta_2$. We use $\mathcal{E}^c$ to denote the complement of $\mathcal{E}$. Note that

$$\mathbb{P}\left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\beta, q} \right\} \leq \mathbb{P}\left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\beta, q} | \mathcal{E}^c \right\} + \mathbb{P}(\mathcal{E}).$$

Let $\beta_3 = \beta - \beta_1 - \beta_2$. Since $\mathbb{P}(\mathcal{E}) \leq \beta_1 + \beta_2$, it suffices to show that

$$
\begin{aligned}
&\mathbb{P}_1\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\beta,q} | \mathcal{E}^c\right\} \\
\leq &\mathbb{P}_1\left\{\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[\Pi\left(X_i;\theta^*\right) - \Pi\left(X_i;\hat{\theta}_\alpha\right)\right]\right\|_q - \left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q \leq r_{\beta,q} | \mathcal{E}^c\right\} \\
\leq &\mathbb{P}_1\left\{\delta_{\beta,q} - \left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q \leq r_{\beta,q} | \mathcal{E}^c\right\} \\
\leq &\mathbb{P}_1\left\{\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q \geq S_{\theta^*} + \tau_{\beta_3,q} | \mathcal{E}^c\right\} \\
\leq &\beta_3,
\end{aligned}
$$

where the third line follows from (46) and the fourth line follows from (47), the fact that we are conditioning on $\mathcal{E}^c$, (40), and the inequality $\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i X_i\right)^2}\right\|_q \leq \left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q$.

7.8 Proof of Lemmas 4.1-4.2 and Proposition 4.1

As a result of Lemma 7.1 and (73), we have the concentration in Lemma 4.1. Because $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \Upsilon\left(X_i;\theta^*\right)\right]\right\|_q$ is separately convex in terms of Y, Lemma 7.2 implies the concentration in Lemma 4.2. Except for a few changes in the notations, the argument to show Proposition 4.1 is nearly identical to what has been used to show Propositions 3.1 and 3.2.

7.9 Proof of Theorem 5.1

For some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0, 1)$, let us define the event

$$\mathcal{E} = \left\{\left\|\frac{1}{n}X^T W\right\|_\infty \leq r_{\alpha,\infty}\right\}$$

where $r_{\alpha,\infty}$ is defined in (60). Bound (22) implies that $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha$. We use the notation $\hat{\Delta} = \hat{\theta}_\alpha^{new} - \theta^*$ in the following. On the event $\mathcal{E}$, we obtain

$$\left\|\frac{1}{n}X^T X\hat{\Delta}\right\|_\infty \leq \left\|\frac{1}{n}X^T W\right\|_\infty + \left\|\frac{1}{n}X^T(Y - X\hat{\theta}_\alpha^{new})\right\|_\infty \leq 2r_{\alpha,\infty}. \tag{77}$$

Given $\mathcal{E}$, θ^* is feasible for (59) and consequently,

$$\left\|\hat{\theta}_\alpha^{new}\right\|_1 \leq \left\|\theta^*\right\|_1,$$

which implies that

$$\left\|\hat{\Delta}_{J_*^c}\right\|_1 \leq \|\theta^*_{J_*}\|_1 - \left\|\hat{\theta}^{new}_{\alpha J_*}\right\|_1 \leq \left\|\hat{\theta}^{new}_{\alpha J_*} - \theta^*_{J_*}\right\|_1 = \left\|\hat{\Delta}_{J_*}\right\|_1; \tag{78}$$

that is, $\hat{\Delta} \in \mathbb{C}_{J_*}$. Using the definition of κ_{J_*} in (62), (77) and (78) imply that

$$\left\|\hat{\theta}^{new}_{\alpha} - \theta^*\right\|_2 \leq \frac{2r_{\alpha,\infty}}{\kappa_{J_*}}$$

with probability at least $1 - \alpha$.

References

[1] Arlot, S., G. Blanchard, and E. Roquain (2010). "Some Nonasymptotic Results on Resampling in High Dimension, I: Confidence Regions." *Annals of Statistics*, 38, 51-82.

[2] Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics*, 37, 1705-1732.

[3] Bobkov, S. G. and M. Ledoux (2000). "From Brunn-Minkowski to Brascamp-Lieb and to Logarithmic Sobolev Inequalities." *Geometric and Functional Analysis*. 10, 1028-1052.

[4] Boucheron, S, G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. Oxford.

[5] Chernozhukov, V., D. Chetverikov, and K. Kato (2013). "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors." *Annals of Statistics*, 41, 2786-2819.

[6] Dezeure, R., P. Bühlmann, and C.-H. Zhang (2017). "High-Dimensional Simultaneous Inference with the Bootstrap." *Test*, 26, 685-719.

[7] Donoho, D. L., M. Elad, and V. N. Temlyakov (2006). "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise". *IEEE Transactions on Information Theory*, 52, 6-18.

[8] Gautier, E. and A. B. Tsybakov (2011). "High-Dimensional Instrumental Variables Regression and Confidence Sets." Manuscript. CREST (ENSAE).

[9] Gautier, E. and A. B. Tsybakov (2014). "High-Dimensional Instrumental Variables Regression and Confidence Sets." Manuscript. CREST (ENSAE).

[10] Horowitz, J. L. (2017). "Non-Asymptotic Inference in Instrumental Variables Estimation." Manuscript. Northwestern University.

[11] Javanmard, A. and A. Montanari (2014). "Confidence Intervals and Hypothesis Testing for High- Dimensional Regression." *Journal of Machine Learning Research*, 15, 2869-2909.

[12] Ledoux, M., and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY.

[13] Maurey, B. (1991). "Some Deviation Inequalities." *Geometric and Functional Analysis*. 1, 188-197.

[14] Ning, Y and H. Liu (2017). "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models." *Annals of Statistics*, 45, 158-195.

[15] Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70, 41-55.

[16] Saumard, A. and J. A. Wellner (2014). "Log-Concavity and Strong Log-Concavity: A Review." *Statistics Surveys,* 8, 45-114.

[17] van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models." *Annals of Statistics*, 42, 1166-1202.

[18] Wainwright, M. J. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. In preparation. University of California, Berkeley.

[19] Wooldridge, J. M. and Y. Zhu (2017). "Inference in Approximately Sparse Correlated Random Effects Probit Models." Forthcoming in *Journal of Business and Economic Statistics*.

[20] Ye, F., and C.-H. Zhang (2010). "Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls". *Journal of Machine Learning Research*, 11, 3519-3540.

[21] Zhang C.-H. and S. S. Zhang (2014). "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217-242.

[22] Zhu, Y. and J. Bradic (2017). "Linear Hypothesis Testing in Dense High-Dimensional Linear Models." Forthcoming in *Journal of the American Statistical Association*.