Information and Inference: A Journal of the IMA (2021) **00**, 1–46
https://doi.org/DOI HERE

High-Dimensional Single-Index Models: Link Estimation and Marginal Inference

KAZUMA SAWAYA
The University of Tokyo

YOSHIMASA UEMATSU
Hitotsubashi University

AND

MASAAKI IMAIZUMI*
The University of Tokyo / RIKEN Center of Advanced Intelligence Project
*Corresponding author: imaizumi@g.ecc.u-tokyo.ac.jp

[Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year]

This study proposes a novel method for estimation and hypothesis testing in high-dimensional single-index models. We address a common scenario where the sample size and the dimension of regression coefficients are large and comparable. Unlike previous approaches, which often overlook the estimation of the unknown link function, we introduce a new method for link function estimation. Leveraging the information from the estimated link function, we propose more efficient estimators that are better aligned with the underlying model. Furthermore, we rigorously establish the asymptotic normality of each coordinate of the estimator. This provides a valid construction of confidence intervals and p-values for any finite collection of coordinates. Numerical experiments validate our theoretical results.

Keywords: link function, observable adjustments, proportionally high dimensions, single-index model, statistical inference

1. Introduction

We consider n i.i.d. observations $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ with a p-dimensional Gaussian feature vector $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, and each scalar response y_i belongs to a set $\mathscr{Y}$ (e.g., $\mathbb{R}, \mathbb{R}_+, \{0,1\}, \mathbb{N}\cup\{0\}$), following the single-index model:

$$\mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}] = g(\boldsymbol{\beta}^\top \boldsymbol{x}), \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown deterministic coefficient vector, and $g(\cdot)$ is an unknown deterministic function, referred to as the link function, with $\boldsymbol{\beta}^\top \boldsymbol{x}$ being the index. To identify the scale of $\boldsymbol{\beta}$, we assume $\mathrm{Var}(\boldsymbol{\beta}^\top \boldsymbol{X}_i) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$. The model includes common scenarios such as:

- *Linear regression*: $y_i \mid \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\beta}^\top \boldsymbol{X}_i, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon > 0$ by setting $g(t) = t$.
- *Poisson regression*: $y_i \mid \boldsymbol{X}_i \sim \mathrm{Pois}(\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i))$ by setting $g(t) = \exp(t)$.
- *Binary choice models*: $y_i \mid \boldsymbol{X}_i \sim \mathrm{Bern}(g(\boldsymbol{\beta}^\top \boldsymbol{X}_i))$ with $g : \mathbb{R} \to [0,1]$. This includes logistic regression for $g(t) = 1/(1+\exp(-t))$ and the probit model by setting $g(\cdot)$ as the cumulative distribution function of the standard Gaussian distribution.

We are interested in a high-dimensional setting, where both the sample size n and the coefficient dimension $p := p(n)$ are large and comparable. Specifically, this study examines the proportionally

high-dimensional regime defined by:

$$n, p(n) \to \infty \quad \text{and} \quad p(n)/n =: \kappa \to \bar{\kappa}, \tag{1.2}$$

where $\bar{\kappa}$ is a positive constant.

The single-index model (3.1) possesses several practically important properties. First, it mitigates concerns about model misspecification, as it eliminates the need to specify $g(\cdot)$. Second, this model bypasses the curse of dimensionality associated with function estimation since the input index $\boldsymbol{\beta}^\top \boldsymbol{X}_i$ is a scalar. This advantage is particularly notable in comparison with nonparametric regression models, such as $y_i = \breve{g}(\boldsymbol{X}_i) + \varepsilon_i$, where $\breve{g} : \mathbb{R}^p \to \mathbb{R}$ remains unspecified. Third, the model facilitates the analysis of the contribution of each covariate, X_{ij} for $j = 1, \ldots, p$, to the response y_i by testing $\beta_j = 0$ against $\beta_j \neq 0$. Owing to these advantages, single-index models have been actively researched for decades [2, 15, 24, 29, 33, 38, 39, 42, 43, 44, 46, 48, 60, 61], attracting interest across a broad spectrum of fields, particularly in econometrics [41, 49].

In the proportionally high-dimensional regime as defined in (1.2), the single-index model and its variants have been extensively studied. For logistic regression, which is a particular instance of the single-index model, Salehi et al. [64], Sur et al. [68] have investigated the estimation and classification errors of the regression coefficient estimators $\boldsymbol{\beta}$. Furthermore, Sur and Candès [67], Yadlowsky et al. [74], Zhao et al. [76] have developed methods for asymptotically valid statistical inference. In the case of generalized linear models with a known link function $g(\cdot)$, Barbier et al. [5], Rangan [62] have characterized the asymptotic behavior of the coefficient estimator, while [65] have derived the coordinate-wise marginal asymptotic normality of an adjusted estimator of β_j. For the single-index model with an unknown link function $g(\cdot)$, the seminal work by Bellec [10] establishes the (non-marginal) asymptotic normality of estimators, even when there is link misspecification. However, the construction of an estimator for the link function $g(\cdot)$ and the marginal asymptotic normality of the coefficient estimator are issues that have not yet been fully resolved.

Inspired by these seminal works, the following questions naturally arise:

1. *Can we consistently estimate the unknown link function $g(\cdot)$?*
2. *Can we rigorously establish marginal statistical inference for each coordinate of $\boldsymbol{\beta}$?*
3. *Can we improve the estimation efficiency by utilizing the estimated link function?*

This paper aims to provide affirmative answers to these questions. Specifically, we propose a novel estimation methodology comprising three steps. First, we construct an estimator for the index $\boldsymbol{\beta}^\top \boldsymbol{X}_i$. Second, we develop an estimator for the link function $g(\cdot)$. Third, we design a new estimator for $\boldsymbol{\beta}$ with the estimated link function. To conduct statistical inference, we investigate the estimation problem of inferential parameters necessary for establishing coordinate-wise asymptotic normality in high-dimensional settings.

Our contributions are summarized as follows:

- *Link estimation*: We propose a consistent estimator for the link function $g(\cdot)$, which is of practical significance as well as estimating coefficients. This aids in interpreting the model via the link function and mitigates negative impacts on coefficient estimation due to link misspecification.
- *Marginal inference*: We establish the asymptotic normality for any finite subset of the coordinates of our estimator, facilitating coordinate-wise inference of $\boldsymbol{\beta}$. This approach allows us not only to test each variable's contribution to the response but also to conduct variable selection based on importance statistics for each feature.

- *Efficiency improvement*: By utilizing the consistently estimated link function, we anticipate that our estimator of $\boldsymbol{\beta}$ will be more efficient than previous estimators that rely on potentially misspecified link functions. We predominantly validate this efficiency through numerical simulations.

From a technical perspective, we leverage the proof strategy in Zhao et al. [76] to demonstrate the marginal asymptotic normality of our estimator for $\boldsymbol{\beta}$. Specifically, we extend the arguments to a broader class of unregularized M-estimators, whereas Zhao et al. [76] originally considered the maximum likelihood estimator (MLE) for logistic regression.

1.1. *Marginal Inference in High Dimensions*

We review key technical aspects of statistical inference for each coordinate β_j in the proportionally high-dimensional regime (1.2). We maintain $\boldsymbol{\beta}^\top \boldsymbol{X}_i$ of constant order by considering the setting $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = \Theta(1)$. We define $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ as the precision matrix for the distribution of $\boldsymbol{X}_i$ and set $\tau_j^2 := \Theta_{jj}^{-1} > 0$. An unbiased estimator of τ_j can be constructed using nodewise regression (cf. Section 5.1 in Zhao et al. [76]). For simplicity, we assume τ_j is known, following prior studies.

In the high-dimensional regime (1.2), statistical inference must address two components: the asymptotic distribution and the *inferential parameters* of an estimator. We review the asymptotic distribution of the MLE $\widehat{\boldsymbol{\beta}}^{\mathrm{m}}$ for logistic regression. According to Zhao et al. [76], for all $j \in \{1, \ldots, p\}$ such that $\sqrt{p}\tau_j \beta_j = O(1)$ as $n \to \infty$, the estimator achieves the following asymptotic normality:

$$\frac{\sqrt{p}(\widehat{\beta}_j^{\mathrm{m}} - \mu_* \beta_j)}{\sigma_* / \tau_j} \xrightarrow{d} \mathcal{N}(0,1). \tag{1.3}$$

Here, we define $\mu_* \in \mathbb{R}$ and $\sigma_* > 0$ as the asymptotic bias and variance, respectively, ensuring the convergence (1.3). It is crucial to note that both the estimator $\widehat{\beta}_j^{\mathrm{m}}$ and the target β_j scale as $O_p(1/\sqrt{p})$ here.

To perform statistical inference based on (1.3), it is necessary to estimate the inferential parameters μ_* and σ_*. Several studies including El Karoui et al. [31], Loureiro et al. [51], Sur and Candès [67], Thrampoulidis et al. [71] theoretically characterize these parameters as solutions to a system of nonlinear equations that depend on the data-generating process and the loss function. Additionally, various approaches have been developed to practically solve the equations by determining their hyperparameter $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$ under different conditions. Specifically, Sur and Candès [67] introduces *ProbeFrontier* for estimating $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$ based on the asymptotic existence/non-existence boundary of the maximum likelihood estimator (MLE) in logistic regression. *SLOE*, proposed by Yadlowsky et al. [74], enhances this estimation using a leave-one-out technique. Moreover, Sawaya et al. [65] takes a different approach to estimate $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$ for generalized linear models.

For single-index models, Bellec [10] introduces *observable adjustments* that estimate the inferential parameters directly under the identification condition $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$ irrespective of link misspecification, bypassing the system of equations. In our study, we develop an estimator for the single-index model satisfying the asymptotic normality (1.3), with corresponding estimators for the inferential parameters using observable adjustments.

1.2. *Related Works*

Research into the asymptotic behavior of statistical models in high-dimensional settings, where both n and p diverge proportionally, has gained momentum in recent years. Notable areas of exploration

include (regularized) linear regression models [6, 8, 27, 37, 40, 45, 50, 56, 59, 69, 71], robust estimation [11, 26, 31], generalized linear models [5, 62, 64, 65, 67, 68, 70, 76], low-rank matrix estimation [25, 52, 58], and various other models [51, 54, 57, 75]. These investigations focus primarily on the convergence limits of estimation and prediction errors. Theoretical analyses have shown that classical statistical estimation often fails to accurately estimate standard errors and may lack key desirable properties such as asymptotic unbiasedness and asymptotic normality.

In such analyses, the following theoretical tools have been employed: (i) the replica method [19, 55], (ii) approximate message passing algorithms [7, 16, 27, 34], (iii) the leave-one-out technique [30, 31], (iv) the convex Gaussian min-max theorem [71], (v) second-order Poincaré inequalities [20, 47], and (vi) second-order Stein's formulae [12, 13]. Although these tools were initially proposed for analyzing linear models with Gaussian design, they have been extensively adapted to a diverse range of models. In this study, we apply observable adjustments based on second-order Stein's formulae [10] to directly estimate the asymptotic bias and variance of coefficient estimators. Furthermore, we provide a comprehensive proof of marginal asymptotic normality, extending the work of Zhao et al. [76] to a wider array of estimators.

1.3. *Notation*

Define $[z] = \{1, \ldots, z\}$ for $z \in \mathbb{N}$. For a vector $\boldsymbol{b} = (b_1, \ldots, b_p) \in \mathbb{R}^p$, we write $\|\boldsymbol{b}\| := (\sum_{j=1}^p b_j^2)^{1/2}$ and $\|\boldsymbol{b}\|_{\boldsymbol{\Sigma}}^2 := \boldsymbol{b}^\top \boldsymbol{\Sigma} \boldsymbol{b}$. For a collection of indices $\mathscr{S} \subset [p]$, we define a sub-vector $\boldsymbol{b}_{\mathscr{S}} := (b_j)_{j \in \mathscr{S}}$ as a slice of $\boldsymbol{\beta}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, we define its minimum and maximum eigenvalues by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$, respectively. For a function $F : \mathbb{R} \to \mathbb{R}$, we say F' the derivative of F and $F^{(m)}$ the mth-order derivative. For a function $f : \mathbb{R} \to \mathbb{R}$ and a vector $\boldsymbol{b} \in \mathbb{R}^p$, $f(\boldsymbol{b}) = (f(b_1), f(b_2), \ldots, f(b_p))^\top \in \mathbb{R}^p$ denotes a vector by elementwise operations.

1.4. *Organization*

We organize the remainder of the paper as follows: Section 2 presents our estimation procedure. Section 3 describes the asymptotic properties of the proposed estimator and develops a statistical inference method. Section 4 provides several experiments to validate our estimation theory. Section 5 outlines the proofs of our theoretical results. Section 6 discusses alternative designs for estimators. Finally, Section 7 concludes with a discussion of our findings. The Appendix contains additional discussions and the complete proofs.

2. Statistical Estimation Procedure

In this section, we introduce a novel statistical estimation method for single-index models as defined in (3.1). To give an overview, our estimator $\widehat{\boldsymbol{\beta}}$ is constructed through the following steps:

(i) Construct an index estimator W_i for $\boldsymbol{\beta}^\top \boldsymbol{X}_i$ using the ridge regression estimator $\tilde{\boldsymbol{\beta}}$, referred to as a pilot estimator. This estimator is reasonable regardless of the misspecification of the link function.

(ii) Develop a function estimator $\widehat{g}(\cdot)$ for the link function $g(\cdot)$, based on the distributional characteristics of the index estimator W_i.

(iii) Construct our estimator $\widehat{\boldsymbol{\beta}}$ for the coefficients $\boldsymbol{\beta}$, using the estimated link $\widehat{g}(\cdot)$ function.

Furthermore, statistical inference additionally involves a fourth step:

(iv) Estimate the inferential parameters μ_* and σ_*, conditional on the estimated link function $\widehat{g}(\cdot)$.

In our estimation procedure, we divide the dataset $(\boldsymbol{X}_i, y_i)_{i=1}^n$ into two disjoint subsets $(\boldsymbol{X}_i, y_i)_{i\in I_1}$ and $(\boldsymbol{X}_i, y_i)_{i\in I_2}$, where $I_1, I_2 \subset [n]$ are index sets such that $I_1 \cap I_2 = \emptyset$ and $I_1 \cup I_2 = [n]$. Additionally, for $k = 1, 2$, let $\boldsymbol{X}^{(k)} \in \mathbb{R}^{n_k \times p}$ and $\boldsymbol{y}^{(k)} \in \mathbb{R}^{n_k}$ denote the design matrix and response vector of subset I_k, respectively. We utilize the first subset to estimate the link function (Steps (i) and (ii)), and the second subset to estimate the regression coefficients (Step (iii)) and inference parameters (Step (iv)). From a theoretical perspective, this division helps to manage the complicated dependency structure caused by data reuse. Nonetheless, for practical applications, we recommend employing all observations in each step to maximize the utilization of the data's inherent signal strength. Here, n_1 and n_2 are the sample sizes for each partition, satisfying $n = n_1 + n_2$ such that n_1 and n_2 are the same asymptotic order, i.e., we define $\kappa_1 = p/n_1$ and $\kappa_2 = p/n_2$ and there exist constants $c_1, c_2 > 0$ with $c_1 \le c_2$ independent of n such that $\kappa_1, \kappa_2 \in [c_1, c_2]$ holds.

2.1. *Index Estimation*

In this step, we use the first subset $(\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)})$. We define the pilot estimator as the ridge estimator, $\tilde{\boldsymbol{\beta}} = ((\boldsymbol{X}^{(1)})^\top \boldsymbol{X}^{(1)} + n_1 \lambda \boldsymbol{I}_p)^{-1} (\boldsymbol{X}^{(1)})^\top \boldsymbol{y}^{(1)}$ where $\lambda > 0$ is the regularization parameter. Further, we consider inferential parameters (μ_1, σ_1) of $\tilde{\boldsymbol{\beta}}$, which satisfy $\sqrt{p}\tau_j(\tilde{\beta}_j - \mu_1\beta_j)/\sigma_1 \xrightarrow{d} \mathcal{N}(0,1)$ for $j \in [p]$ such that $\sqrt{p}\tau_j\beta_j = O(1)$. Using these parameters, we develop an estimator W_i for the index $\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)}$ as follows:

$$W_i := \tilde{\mu}^{-1} \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} - \tilde{\mu}^{-1} \tilde{\gamma} \left(y_i^{(1)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} \right) \tag{2.1}$$

for each $i \in [n_1]$. Here, $\tilde{\mu}$ and $\tilde{\sigma}^2$ are estimators of μ_1 and σ_1, defined as

$$\tilde{\mu} = \left| \|\tilde{\boldsymbol{\beta}}\|^2 - \tilde{\sigma}^2 \right|^{1/2} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{n_1^{-1} \|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)} \tilde{\boldsymbol{\beta}}\|^2}{(\tilde{\nu} + \lambda)^2 / \kappa_1},$$

where $\tilde{\gamma} := \kappa_1 / (\tilde{\nu} + \lambda)$ and $\tilde{\nu} = n_1^{-1} \mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{X}^{(1)}((\boldsymbol{X}^{(1)})^\top \boldsymbol{X}^{(1)} + n_1 \lambda \boldsymbol{I}_p)^{-1} (\boldsymbol{X}^{(1)})^\top)$. These estimators are obtained by the observable adjustment technique described in Bellec [10].

This index estimator W_i is approximately unbiased for the index $\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)}$, yielding the following asymptotic result.

$$W_i \approx \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} + \mathcal{N}(0, \tilde{\sigma}^2 / \tilde{\mu}^2). \tag{2.2}$$

We will provide its rigorous statement in Proposition 5 in Section 5.1.

There are other options for the pilot estimator besides the ridge estimator $\tilde{\boldsymbol{\beta}}$. If $\kappa_1 \le 1$ holds, the least squares estimator can be an alternative. If y_i is a binary or non-negative integer, the MLE of logistic or Poisson regression can be a natural candidate, respectively, although the ridge estimator $\tilde{\boldsymbol{\beta}}$ is valid regardless of the form that y_i takes. In each case, the estimated inferential parameters $(\tilde{\gamma}, \tilde{\mu}, \tilde{\sigma}^2)$ should be updated accordingly. Details are presented in Section 6.

2.2. *Link Estimation*

We develop an estimator of the link function $g(\cdot)$ using W_i in (2.1). If we could observe the true index $\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)}$ with the unknown coefficient $\boldsymbol{\beta}$, it would be possible to estimate $g(x) = \mathbb{E}[y_1 \mid \boldsymbol{\beta}^\top \boldsymbol{X}_1 = x]$ by applying standard nonparametric methods to the pairs of responses and true indices $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)}\boldsymbol{\beta})$.

However, as the true index is unobservable, we must estimate $g(\cdot)$ using given pairs of responses and contaminated indices $(y_i^{(1)}, W_i)_{i=1}^{n_1}$, where $W_i \approx \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} + \mathcal{N}(0, \tilde{\varsigma}^2)$ with $\tilde{\varsigma}^2 = \tilde{\sigma}^2/\tilde{\mu}^2$ being a ratio of the inferential parameters. The type of error $\mathcal{N}(0, \tilde{\varsigma}^2)$ involving the regressor leads to an attenuation bias in the estimation of $g(\cdot)$, known as the *errors-in-variables* problem. To address this issue, we utilize a deconvolution technique [66] to remove the bias stemming from error-in-variables asymptotically. Further details of the deconvolution are provided in Supplementary Material.

We define an estimator of $g(\cdot)$. In preparation, we specify a kernel function $K : \mathbb{R} \to \mathbb{R}$, and define a deconvolution kernel $K_n : \mathbb{R} \to \mathbb{R}$ as follows:

$$K_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\mathrm{i}tx) \frac{\phi_K(t)}{\phi_{\tilde{\varsigma}}(t/h_n)} dt,$$

where $h_n > 0$ is a bandwidth, $\mathrm{i} = \sqrt{-1}$ is an imaginary unit, and $\phi_K : \mathbb{R} \to \mathbb{R}$ and $\phi_{\tilde{\varsigma}} : \mathbb{R} \to \mathbb{R}$ are the Fourier transform of $K(\cdot)$ and the density function of $\mathcal{N}(0, \tilde{\varsigma}^2)$, respectively. We then define our estimator of $g(\cdot)$ as

$$\widehat{g}(x) := \mathscr{R}[\breve{g}](x) \quad \text{with} \quad \breve{g}(x) = \frac{\sum_{i=1}^{n_1} y_i^{(1)} K_n\left((x - W_i)/h_n\right)}{\sum_{i=1}^{n_1} K_n\left((x - W_i)/h_n\right)}, \tag{2.3}$$

where $\mathscr{R}[\cdot]$ is a monotonization operator, specified later, which maps any measurable function to a monotonic function, and $\breve{g}(\cdot)$ is a Nadaraya-Watson estimator obtained by the deconvolution kernel. We will prove the consistency of this estimator in Theorem 1 in Section 3.

The monotonization operation $\mathscr{R}[\cdot]$ on $\breve{g}(\cdot)$ is justifiable because the true link function $g(\cdot)$ is assumed to be monotonic. One simple choice for $\mathscr{R}[\cdot]$, applicable to any measurable function $f : \mathbb{R} \to \mathbb{R}$, is

$$\mathscr{R}_{\mathrm{naive}}[f](x) = \sup_{x' \le x} f(x'), \quad x \in \mathbb{R}.$$

This definition holds for all $x \in \mathbb{R}$. Another effective alternative is the rearrangement operator [21]. This operator monotonizes a measurable function $f : \mathbb{R} \to \mathbb{R}$ within a compact interval $[\underline{x}, \bar{x}] \subset \mathbb{R}$:

$$\mathscr{R}_{\mathrm{r}}[f](x) = \inf\left\{ t \in \mathbb{R} : \int_{[0,1]} 1\left\{ f\left(\frac{u - \underline{x}}{\bar{x} - \underline{x}} \right) \le t \right\} du \ge \frac{x - \underline{x}}{\bar{x} - \underline{x}} \right\}, \; x \in [\underline{x}, \bar{x}]. \tag{2.4}$$

This operator, which sorts the values of $f(\cdot)$ in increasing order, is robust against local fluctuations such as function bumps. Thus, it effectively addresses bumps in $\breve{g}(\cdot)$ arising from kernel-based methods.

2.3. *Coefficient Estimation*

We next propose our estimator of $\boldsymbol{\beta}$ using $\widehat{g}(\cdot)$ obtained in (2.3). In this step, we consider the link estimator $\widehat{g}(\cdot)$ from $\boldsymbol{X}^{(1)}$ as given, and estimate $\boldsymbol{\beta}$ using $\boldsymbol{X}^{(2)}$. To facilitate this, we introduce the *surrogate loss function* for $\boldsymbol{b} \in \mathbb{R}^p$, with $\boldsymbol{x} \in \mathbb{R}^p$, $y \in \mathbb{R}$, and any measurable function $\bar{g} : \mathbb{R} \to \mathbb{R}$:

$$\ell(\boldsymbol{b}; \boldsymbol{x}, y, \bar{g}) := \bar{G}(\boldsymbol{x}^\top \boldsymbol{b}) - y \boldsymbol{x}^\top \boldsymbol{b},$$

where $\bar{G} : \mathbb{R} \to \mathbb{R}$ is a function such that $\bar{G}'(t) = \bar{g}(t)$. This function can be viewed as a natural extension of the matching loss [3] used in generalized linear models. If $\bar{g}(\cdot)$ is strictly increasing, then the loss is

strictly convex in $\boldsymbol{b}$. Moreover, the surrogate loss is justified by the characteristics of the true parameter as follows [1]:

$$\boldsymbol{\beta} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\, \mathbb{E}\left[\ell(\boldsymbol{b}; \boldsymbol{X}_1, y_1, g) | \boldsymbol{X}_1\right],$$

provided that $G(\cdot)$ is integrable. The surrogate loss aligns with the negative log-likelihood when $g(\cdot)$ is known and serves as a canonical link function, thereby making the surrogate loss minimizer a generalization of the MLEs in generalized linear models.

Using the second dataset $(\boldsymbol{X}^{(2)}, \boldsymbol{y}^{(2)})$ with any given function $\bar{g}(\cdot)$, we define our estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}(\bar{g}) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n_2} \ell(\boldsymbol{b}; \boldsymbol{X}_i^{(2)}, y_i^{(2)}, \bar{g}) + J(\boldsymbol{b}), \tag{2.5}$$

where $J : \mathbb{R}^p \to \mathbb{R}$ is a convex regularization function. Finally, we substitute the link estimator $\widehat{g}(\cdot)$ into (2.5) to obtain our estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$. The use of a nonzero regularization term, $J(\cdot)$, is beneficial in cases where the minimizer (2.5) is not unique or does not exist; see, for example, [18] for the logistic regression case.

2.4. *Inferential Parameter Estimation*

We finally study estimators for the inferential parameters of our estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$, which are essential for statistical inference as discussed in Section 1.1. As established in (1.3), it is necessary to estimate the asymptotic bias $\mu_*(\widehat{g})$ and variance $\sigma_*^2(\widehat{g})$ that satisfy the following relationship:

$$\frac{\sqrt{p}(\widehat{\beta}_j(\widehat{g}) - \mu_*(\widehat{g})\beta_j)}{\sigma_*(\widehat{g})} \xrightarrow{d} \mathcal{N}(0,1), \quad j \in [p],$$

conditional on $(\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)})$ and consequently on $\widehat{g}(\cdot)$.

We develop estimators for these inferential parameters using observable adjustments as suggested by Bellec [10], in accordance with the estimator (2.5). For any measurable function $\bar{g} : \mathbb{R} \to \mathbb{R}$, we define $\boldsymbol{D} = \mathrm{diag}(\bar{g}'(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g})))$ and $\widehat{v}_\lambda = n_2^{-1}\mathrm{tr}(\boldsymbol{D} - \boldsymbol{D}\boldsymbol{X}^{(2)}((\boldsymbol{X}^{(2)})^\top \boldsymbol{D}\boldsymbol{X}^{(2)} + n_2\lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^{(2)})^\top \boldsymbol{D})$ for $\lambda \geq 0$. When incorporating $J(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|^2/2$ into (2.5) with $\lambda > 0$, we propose the following estimators:

$$\widehat{\mu}(\bar{g}) = \left| \|\widehat{\boldsymbol{\beta}}(\bar{g})\|^2 - \widehat{\sigma}^2(\bar{g}) \right|^{1/2} \text{ and } \widehat{\sigma}^2(\bar{g}) = \frac{\|\boldsymbol{y}^{(2)} - \bar{g}(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g}))\|^2}{n_2(\widehat{v}_\lambda + \lambda)^2/\kappa_2}.$$

In the case where $J(\cdot) \equiv \boldsymbol{0}$ holds, we define

$$\widehat{\mu}_0(\bar{g}) = \left| \frac{\|\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g})\|^2}{n_2} - (1-\kappa_2)\widehat{\sigma}_0^2(\bar{g}) \right|^{1/2} \text{ and } \widehat{\sigma}_0^2(\bar{g}) = \frac{\|\boldsymbol{y}^{(2)} - \bar{g}(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g}))\|^2}{n_2\widehat{v}_0^2/\kappa_2}.$$

A theoretical justification for the asymptotic normality with these estimators and their application in inference is provided in Section 3.2.

3. Main Theoretical Results of Proposed Estimators

This section presents theoretical results for our estimation framework. Specifically, we prove the consistency of the estimator $\widehat{g}(\cdot)$ for the link function $g(\cdot)$, and the asymptotic normality of the estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ for the coefficient vector $\boldsymbol{\beta}$. Outlines of the proofs for each assertion will be provided in Section 5.

Assumption 1 (Gaussian covariates and identification) *Each row of the matrix $\boldsymbol{X}$ independently follows $\mathcal{N}_p(\mathbf{0},\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ obeying $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}=1$ and $0<c_\Sigma^{-1}\le\lambda_{\min}(\boldsymbol{\Sigma})\le\lambda_{\max}(\boldsymbol{\Sigma})\le c_\Sigma<\infty$ for some constant c_Σ.*

It is common to assume Gaussianity of covariates in the proportionally high-dimensional regime, as mentioned in Section 1.2. The condition $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}=1$ is necessary to identify the scale of $\boldsymbol{\beta}$, which ensures the uniqueness of the estimator in the single-index model with an unknown link function $g(\cdot)$. For example, without this condition, it would be impossible to distinguish between $g(\boldsymbol{X}_1^\top\boldsymbol{b})$ and $f(2\boldsymbol{X}_1^\top\boldsymbol{b})$, where $f(t)=g(t/2)$, for any $\boldsymbol{b}\in\mathbb{R}^p$. Furthermore, the assumption of upper and lower bounds on the eigenvalues of $\boldsymbol{\Sigma}$ implies that $\|\boldsymbol{\beta}\|=\Theta(1)$.

Assumption 2 (Coherency of the single-index model) *A distribution of $(y_i,\boldsymbol{X}_i)$, which follows the model (1.1), satisfies that there exist an (unknown) deterministic function $F:\mathbb{R}^2\to\mathbb{R}$ such that we have*

$$y_i=F(\boldsymbol{\beta}^\top\boldsymbol{X}_i,U_i), \tag{3.1}$$

where U_i is some random variable independent of $\boldsymbol{X}_i$.

This assumption ensures the coherency of the single-index model (1.1). The description has been commonly used [10, 48] and includes a wide class of models, e.g., the linear regression with $F(v,u)=v+u$ and $U_i\sim N(0,\sigma_U^2)$ with $\sigma_U^2>0$, and the logistic regression model with $F(v,u)=\mathbf{1}\{u\le 1/(1+e^{-v})\}$ and $U_i\sim\mathrm{Unif}[0,1]$. Specifically, this assumption excludes a case where $\boldsymbol{X}_i$ affects y_i in a way that is independent of the index $\boldsymbol{\beta}^\top\boldsymbol{X}_i$, such as $y_i=g(\boldsymbol{\beta}^\top\boldsymbol{X}_i)+\boldsymbol{\gamma}^\top\boldsymbol{X}_i\varepsilon_i$, where ε_i is a mean zero random variable independent of $\boldsymbol{X}_i$ and $\boldsymbol{\gamma}\in\mathbb{R}^p$ is an additional coefficient. A property of F is indirectly constrained by assumptions about the link g that follow.

Assumption 3 (Monotone and smooth link function) *There exists $m\in\mathbb{N}$ and constants $a<b$ such that $g^{(\ell)}(x)$ exists for every $\ell=0,1\ldots,m$ and $x\in[a,b]$. Also, there exists a constant $B>0$ such that $\max_{\ell=0,1,\ldots,m}\max_{x\in[a,b]}|g^{(\ell)}(x)|\le B$ holds. Furthermore, there exists $c_g\in(0,\infty)$ such that $c_g^{-1}\le\min_{x\in[a,b]}g'(x)$ holds.*

Assumption 3 restricts the class of link functions to those that are monotonic. This class has been extensively reviewed in the literature, with Balabdaoui et al. [4] providing a comprehensive discussion. It encompasses a wide range of applications, including utility functions, growth curves, and dose-response models [35, 53, 72]. Furthermore, under a monotonically increasing link function, the sign of $\boldsymbol{\beta}$ is identified, so that we can identify $\boldsymbol{\beta}$ only by the scale condition $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}=1$.

The lower boundedness of $g'(\cdot)$ on the closed interval implies that the loss function (2.5) for the coefficient estimation is strictly convex on the interval. This assumption holds for the negative log-likelihood of the logistic regression that is not strictly convex on the real line.

Assumption 4 (Moment conditions of $\boldsymbol{y}$) $\mathbb{E}[y_1^2] < \infty$ *holds. Further,* $m_2(x) := \mathbb{E}[y_1^2 \mid \boldsymbol{X}_1^\top \boldsymbol{\beta} = x]$ *is continuous in* $x \in [a,b]$ *for* $a,b \in \mathbb{R}$ *defined in Assumption 3.*

The continuity of $m_2(\cdot)$ is maintained in many commonly used models, particularly when $g(\cdot)$ is continuous. For instance, the Poisson regression model defines $m_2(x) = \exp(x)(1+\exp(x))$, and binary choice models specify $m_2(x) = g(x)$.

3.1. *Consistency of Link Estimation*

We demonstrate the uniform consistency of the link estimator $\widehat{g}(\cdot)$ in (2.3) over closed intervals. We consider the mth-order kernel $K(\cdot)$ that satisfies

$$\int_{-\infty}^{\infty} K(t)dt = 1, \quad \int_{-\infty}^{\infty} t^m K(t)dt \neq 0, \text{ and } \int_{-\infty}^{\infty} t^\ell K(t)dt = 0,$$

for $\ell \in [m-1]$. We then obtain the following:

Theorem 1 *Suppose that Assumptions 1–4 hold and the Fourier transform* $\phi_K(t)$ *of the kernel* $K(\cdot)$ *has a bounded support in* $[-M_0, M_0]$ *with some* $M_0 > 0$*, and the bandwidth* $h_n = (c_h \log n_1)^{-1/2}$ *satisfies* $2M_0^2(\sigma_1/\mu_1)^2 c_h < 1$*. Also, for* $Z_i \sim \mathcal{N}(0,1)$*,* $i \in [n]$ *defined in Proposition 5, suppose that* $(Z_1,\ldots,Z_n)^\top \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$ *holds. Then, we have the following as* $n_1 \to \infty$*:*

$$\sup_{a \le x \le b} |\widehat{g}(x) - g(x)| = O_{\mathrm{p}}\left(\frac{1}{(\log n_1)^{m/2}}\right). \tag{3.2}$$

This result shows the consistency of the link estimator. About the convergence rate, according to Fan and Truong [32], the logarithmic rate $O_{\mathrm{p}}(1/(\log n_1)^{m/2})$ reaches a lower bound, indicating that this rate cannot be improved.

Regarding the condition on $(Z_1,...,Z_n)^\top \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$, it ensures that the index estimator W_i can be regarded as asymptotically i.i.d. and is necessary for constructing a consistent link function estimator. We numerically discuss the condition in Appendix C.

3.2. *Marginal Asymptotic Normality of Coefficient Estimators*

This section demonstrates the marginal asymptotic normality of our estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ for $\boldsymbol{\beta}$, facilitated by the estimators of the inferential parameters, $\widehat{\mu}(\widehat{g})$ and $\widehat{\sigma}(\widehat{g})$. These results are directly applicable to hypothesis testing and the construction of confidence intervals for any finite subset of the β_j's.

3.2.1. Unit Covariance and $p > n$ Case

As previously noted, the inferential parameters vary depending on the estimator considered. In this section, we focus on the ridge regularized estimator with unit covariance $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. We will also present additional results for generalized covariance matrices and the ridgeless scenario later.

Theorem 2 *We consider the coefficient estimator* $\widehat{\boldsymbol{\beta}}(\widehat{g})$ *with* $J(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|^2/2$*, and the inferential estimators* $(\widehat{\mu}(\widehat{g}), \widehat{\sigma}(\widehat{g}))$*, associated with the link estimator* $\widehat{g}(\cdot)$*. Suppose that* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and Assumptions 1-3 hold. Then, a conditional distribution of* $(\widehat{\boldsymbol{\beta}}(\widehat{g}), \widehat{\mu}(\widehat{g}), \widehat{\sigma}(\widehat{g}))$ *with a fixed event on* $\widehat{g}(\cdot)$ *satisfies the*

following: for any coordinate $j \in [p]$ satisfying $\sqrt{p}\beta_j = O(1)$, we have

$$T_j := \frac{\sqrt{p}(\widehat{\beta}_j(\widehat{g}) - \widehat{\mu}(\widehat{g})\beta_j)}{\widehat{\sigma}(\widehat{g})} \xrightarrow{d} \mathcal{N}(0,1) \tag{3.3}$$

as $n, p \to \infty$ with the regime (1.2). Moreover, for any finite set of coordinates $\mathcal{S} \subset [p]$ satisfying $\sqrt{p}\|\boldsymbol{\beta}_{\mathcal{S}}\| = O(1)$, we have, as $n, p \to \infty$,

$$\frac{\sqrt{p}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}(\widehat{g}) - \widehat{\mu}(\widehat{g})\boldsymbol{\beta}_{\mathcal{S}})}{\widehat{\sigma}(\widehat{g})} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{|\mathcal{S}|}).$$

This result also implies that $\widehat{\beta}_j(\widehat{g})/\widehat{\mu}(\widehat{g})$ is an asymptotically unbiased estimator of β_j. Note that the convergence of the conditional distribution is ensured by the non-degeneracy property of the conditional event, as defined by $(\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)})$; see Goggin [36] for details. We can improve the condition $\sqrt{p}\beta_j = O(1)$ to $\beta_j = o(1)$ under the Assumption E in Bellec [10] which derives the explicit rate of convergence for the estimation error of $\widehat{\mu}(\cdot)$.

We highlight two key contributions of Theorem 2. First, it is the first result in the literature of a single-index model to demonstrate the marginal asymptotic normality of the coefficient estimator that leverages the link estimator. Second, it remains valid even when the ratio $\kappa = p/n$ exceeds one, a notable distinction from a similar marginal asymptotic result (Theorem 5.2 in [10]), which holds only when κ is less than one. Although Section 4 of [10] addresses the case $\kappa > 1$ and considers many penalty functions, its results pertain not to marginal asymptotic normality, but rather to the average behavior of the estimator.

Application to Statistical Inference: From Theorem 2, we construct a confidence interval $\mathrm{CI}^j_{1-\alpha}$ for each β_j with a confidence level $(1-\alpha)$ as follows:

$$\mathrm{CI}^j_{1-\alpha} := \frac{1}{\widehat{\mu}(\widehat{g})}\left[\widehat{\beta}_j(\widehat{g}) - z_{(1-\alpha/2)}\frac{\widehat{\sigma}(\widehat{g})}{\sqrt{p}},\ \widehat{\beta}_j(\widehat{g}) + z_{(1-\alpha/2)}\frac{\widehat{\sigma}(\widehat{g})}{\sqrt{p}}\right],$$

where $j \in [p]$ and $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. This construction ensures the coverage probability adheres to the specified confidence level asymptotically.

Corollary 3 *Under the settings of Theorem 2, for any $\alpha \in (0,1)$, we have the following as $n, p \to \infty$ with the regime (1.2):*

$$\sup_{1 \le j \le p}\left|\mathbb{P}\left(\beta_j \in \mathrm{CI}^j_{1-\alpha}\right) - (1-\alpha)\right| \to 0.$$

Hence, for testing the hypothesis $H^j_0 : \beta_j = 0$ against $H^j_1 : \beta_j \neq 0$ at level $\alpha \in (0,1)$, we can use the corrected t-statistics in (3.3). The test that rejects the null hypothesis H^j_0 if

$$\frac{\widehat{\sigma}(\widehat{g})z_{(1-\alpha/2)}}{\sqrt{p}\tau_j} \le |\widehat{\beta}_j(\widehat{g})|$$

controls the asymptotic size of the test at the level α. Additionally, it is feasible to develop a variable selection procedure that identifies variables related to the response. Specifically, the p-value associated

with H_0^j and the statistic $\sqrt{p}\widehat{\beta}_j(\widehat{g})/\widehat{\sigma}(\widehat{g})$ can serve as importance statistics for the jth covariate. This approach facilitates variable selection procedures that control the false discovery rate, as detailed in sources such as Benjamini and Hochberg [14], Candès et al. [17], Dai et al. [23], Xing et al. [73].

3.2.2. General Covariance and $p < n$ Case
We extend Theorem 2 to scenarios with a general covariance matrix $\mathbf{\Sigma}$ in unregularized settings. To this end, we utilize the estimators $(\widehat{\mu}_0(\widehat{g}), \widehat{\sigma}_0(\widehat{g}))$, which are defined for inferential parameters in Section 2.4. Recall that the precision matrix $\mathbf{\Theta}$ is defined as $\mathbf{\Sigma}^{-1}$.

Theorem 4 *We consider the coefficient estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ with $J(\boldsymbol{b}) \equiv 0$, and the inferential estimators $(\widehat{\mu}_0(\widehat{g}), \widehat{\sigma}_0(\widehat{g}))$, associated with the link estimator $\widehat{g}(\cdot)$. Suppose that Assumptions 1-3 hold. Then, a conditional distribution of $(\widehat{\boldsymbol{\beta}}(\widehat{g}), \widehat{\mu}_0(\widehat{g}), \widehat{\sigma}_0(\widehat{g}))$ with a fixed event on $\widehat{g}(\cdot)$ satisfies the following: for any coordinate $j \in [p]$ satisfying $\sqrt{p}\tau_j\beta_j = O(1)$, we have*

$$\frac{\sqrt{p}(\widehat{\beta}_j(\widehat{g}) - \widehat{\mu}_0(\widehat{g})\beta_j)}{\widehat{\sigma}_0(\widehat{g})/\tau_j} \xrightarrow{d} \mathcal{N}(0,1) \tag{3.4}$$

as $n, p \to \infty$ with the regime (1.2) with $\bar{\kappa} \in (0,1)$. Moreover, for a finite set of coordinates $\mathcal{S} \subset \{1, \ldots, p\}$, we have

$$\frac{\sqrt{p}\mathbf{\Theta}_{\mathcal{S}}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}(\widehat{g}) - \widehat{\mu}_0(\widehat{g})\boldsymbol{\beta}_{\mathcal{S}})}{\widehat{\sigma}_0(\widehat{g})} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{|\mathcal{S}|}), \tag{3.5}$$

where the submatrix $\mathbf{\Theta}_{\mathcal{S}}$ consists of Θ_{ij} for $i, j \in \mathcal{S}$.

We can also improve the condition $\sqrt{p}\tau_j\beta_j = O(1)$ to $\tau_j\beta_j = o(1)$, under the Assumption E in Bellec [10].

4. Experiments

This section provides numerical validations of our estimation framework as outlined in Section 2. The efficiency of our proposed estimator is subsequently compared with that of other estimators.

We examine two high-dimensional scenarios: $n < p$ and $n > p$. For the scenario where $n > p$, we assume the true coefficient vector follows a uniform distribution on the sphere: $\boldsymbol{\beta} \sim \mathrm{Unif}(\mathbb{S}^{p-1})$. In the case of $n < p$, we set $\beta_1 = \cdots = \beta_{100} = \sqrt{p/100}$ and $\beta_{101} = \cdots = \beta_p = 0$. We generate response variables y_i for Gaussian predictors $\boldsymbol{X}_i$ with an identity covariance matrix $\mathbf{\Sigma} = \boldsymbol{I}_p$, under the following four scenarios:

(i) Cloglog: $y_i \mid \boldsymbol{X}_i \sim \mathrm{Bern}(g_{(\mathrm{i})}(\boldsymbol{\beta}^\top \boldsymbol{X}_i))$ with $g_{(\mathrm{i})}(t) = 1 - \exp(-\exp(t))$;

(ii) xSqrt: $y_i \mid \boldsymbol{X}_i \sim \mathrm{Pois}(g_{(\mathrm{ii})}(\boldsymbol{\beta}^\top \boldsymbol{X}_i))$ with $g_{(\mathrm{ii})}(t) = t + \sqrt{t^2+1}$;

(iii) Cubic: cubic regression $y_i = g_{(\mathrm{iii})}(\boldsymbol{\beta}^\top \boldsymbol{X}_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, 1/2)$ and $g_{(\mathrm{iii})}(t) = t^3/3$;

(iv) Piecewise: piecewise regression $y_i = g_{(\mathrm{iv})}(\boldsymbol{\beta}^\top \boldsymbol{X}_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, 1/5)$ and $g_{(\mathrm{iv})}(t) = (0.2t - 2.3)1_{(-\infty,-1]} + 2.5t1_{(-1,1)} + (0.2t + 2.3)1_{[1,\infty)}$.

4.1. *Index Estimator*

We validate the normal approximation of the index estimator W_i as shown in (2.2). For cases where $n > p$, we set $(n, p) = (500, 50)$ for the Cloglog model and $(n, p) = (500, 200)$ for the other models. For cases where $n < p$, we set $(n, p) = (250, 500)$ and apply the ridge regularized estimator to all models. We assign the maximum likelihood estimator (MLE) of logistic regression to the pilot estimator for (i) Cloglog, the MLE of Poisson regression for (ii) xSqrt, and the least squares estimator for both (iii) Cubic and (iv) Piecewise models. We calculate $\tilde{\mu}(\boldsymbol{W} - \boldsymbol{X}\boldsymbol{\beta})/\tilde{\sigma}$ using 1,000 replications for each setup.

Figure 1 displays the results. In all settings, the difference between the index estimator and the index follows a Gaussian distribution, as expected.

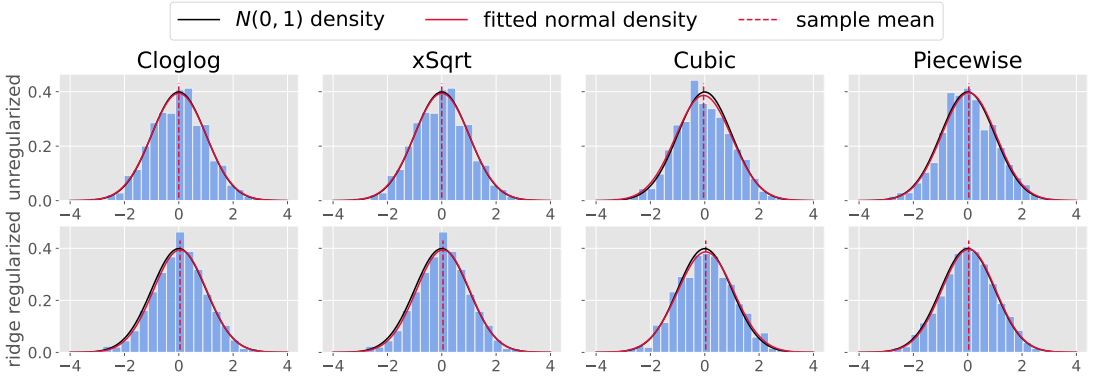

FIG. 1. Histograms of the first coordinate of the statistics $\tilde{\mu}(\boldsymbol{W} - \boldsymbol{X}\boldsymbol{\beta})/\tilde{\sigma}$ over 1,000 replications. According to Proposition 5, these histograms are expected to resemble the $\mathcal{N}(0, 1)$ density. The columns correspond to each model, ranging from (i) Cloglog to (iv) Piecewise, while the rows represent unregularized and ridge-regularized estimations for cases where $n > p$ and $n < p$, respectively.

4.2. *Link Function Estimator*

Next, we evaluate the numerical performance of the link estimator $\widehat{g}(\cdot)$, constructed from $(W_1, \ldots, W_n)$, using a fixed bandwidth for each n. Figure 2 (left panel) shows that the estimation error of $\widehat{g}(\cdot)$ for (iv) Piecewise uniformly approaches zero as the sample size increases. The right four panels in Figure 2 display the squared losses of $\widehat{g}(\cdot)$ evaluated over the interval $[-3, 3]$, which all decrease as n increases, while their convergence rate is slow as shown in (3.2).

4.3. *Our Coefficient Estimator*

We examine the asymptotic normality of each coordinate of the estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ for the true coefficients. As in Section 4.2, we construct the estimator using a fixed bandwidth and apply the rearrangement operator $\mathscr{R}_{\mathrm{r}}[\cdot]$ as defined in (2.4) over the interval $[-3, 3]$ to obtain $\widehat{g}(\cdot)$. We then compute $\widehat{\boldsymbol{\beta}}(\widehat{g})$ according to (2.5) using $J(\cdot) \equiv \boldsymbol{0}$ when $n > p$ and $J(\boldsymbol{b}) = \|\boldsymbol{b}\|^2$ when $n \leq p$. Figure 3 shows the marginal normal approximation of the estimators under these conditions. All histograms closely resemble the standard normal density, corroborating the asymptotic normality stated in Theorems 2 and 4.

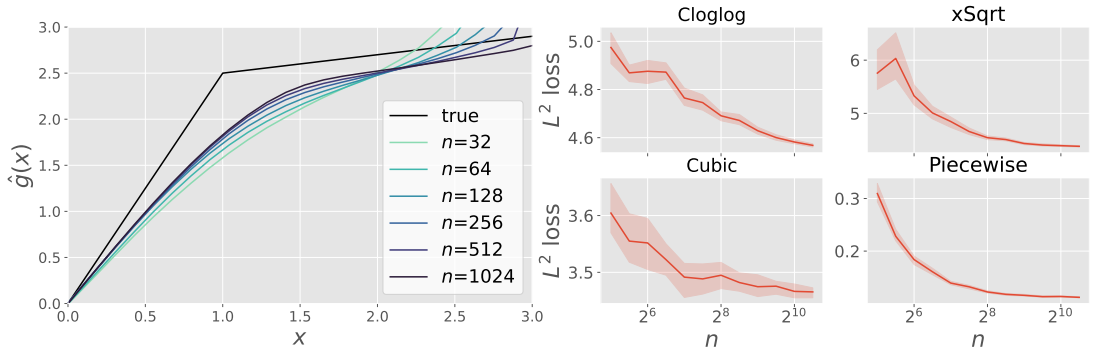

FIG. 2. Estimated link functions $\widehat{g}(\cdot)$ for (iv) Piecewise were obtained with a fixed ratio $p/n = 0.6$ and $n = 32, 64, \ldots, 1024$, averaged over 1,000 replications (left). The squared loss for $\widehat{g}(\cdot)$, evaluated over the interval $[-3, 3]$, for (i) Cloglog to (iv) Piecewise, as defined in the previous section, with a fixed ratio $p/n = 0.4$ averaged over 1,000 replications (right).

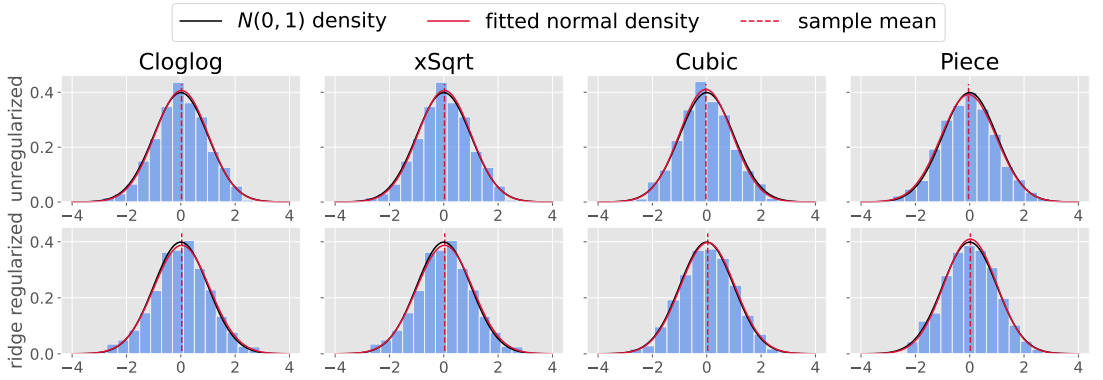

FIG. 3. Histograms of the first coordinate of the statistics $\sqrt{p}(\widehat{\beta}_1(\widehat{g}) - \widehat{\mu}(\widehat{g})\beta_1)/\widehat{\sigma}(\widehat{g})$ made from 1,000 replications. The columns correspond to each model from (i) Cloglog to (iv) Piecewise, and the rows correspond to unregularized and ridge-regularized estimation under $n > p$ and $n < p$, respectively. They are expected to approach $\mathcal{N}(0, 1)$ density by Theorems 2 and 4.

4.4. *Efficiency Comparison*

Finally, we compare the estimation efficiency of the proposed estimator with several pilot estimators. We use the effective asymptotic variance σ_*^2/μ_*^2 as an efficiency measure, which is the inverse of the effective signal-to-noise ratio as described in [34]. We estimate this variance using the statistic $\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\beta}}/(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}) - 1$ for an estimator $\widehat{\boldsymbol{\beta}}$. This statistic is a reasonable approximation of the asymptotic variance of the debiased version of $\widehat{\boldsymbol{\beta}}$ and converges almost surely to the effective asymptotic variance under certain conditions (see Section 5 for details).

From a practical perspective, we analyze the scatter plot of (W_i, y_i) and manually specify a functional form for $g(\cdot)$ to conduct parametric regression. We estimate parameters $a, b, c \in \mathbb{R}$ in different forms: $\check{g}(t) = 1/(1 + \exp(-at + b))$ for case (i), $\check{g}(t) = a\exp(t) + b$ for case (ii), $\check{g}(t) = at^3 + bt^2 + ct$ for case (iii), and $\check{g}(t) = a/(1 + \exp(-bt + c)) - a/2$ for case (iv). We then use these estimates to construct the link function. Additionally, we introduce new data-generating processes: Logit, where $y_i \mid \boldsymbol{X}_i \sim \text{Bern}(1/(1 + \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i)))$; Poisson, where $y_i \mid \boldsymbol{X}_i \sim \text{Pois}(\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i))$; Cubic+, where

$y_i = g_{(\mathrm{iii})}(\boldsymbol{\beta}^\top \boldsymbol{X}_i) + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(5, 1/2)$; and Piecewise+, where $y_i = g_{(\mathrm{iv})}(\boldsymbol{\beta}^\top \boldsymbol{X}_i) + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(5, 1/5)$.

Table 1 displays the efficiency measures for our proposed estimator and others across 100 replications. We find that our proposed estimator is generally more efficient in most settings, except when the estimators are specifically tailored to the models. This highlights the broad applicability of our estimator.

	LeastSquares	LogitMLE	PoisMLE	Proposed
Logit	$3.77 \pm .967$	**.525±.157**	–	$.527 \pm .157$
Cloglog	$3.13 \pm .599$	$.294 \pm .080$	–	**.271±.068**
Poisson	$3.77 \pm .967$	–	**.630±.124**	**.630±.124**
xSqrt	$2.32 \pm .692$	–	**1.12±.290**	**1.12±.290**
Cubic	**1.15±.258**	–	–	$1.74 \pm .440$
Cubic+	33.9 ± 50.7	–	–	**1.74±.439**
Piecewise	$.541 \pm .031$	–	–	**.391±.157**
Piecewise+	6.32 ± 3.26	–	–	**.330±.184**

TABLE 1 *Efficiency measure for each pair of model and estimator. We report the average ± standard deviation.*

4.5. *Real Data Applications*

We utilize two datasets from the UCI Machine Learning Repository [28] to illustrate the performance of the proposed estimator. The DARWIN dataset [22] comprises handwriting data from 174 participants, including both Alzheimer's disease patients and healthy individuals. The second dataset [63] features 753 attributes derived from the sustained phonation of the vowel sounds of patients, both with and without Alzheimer's disease. We employ a leave-one-out strategy for splitting each dataset. For each $n-1$ subset, we compute the regularized MLE of logistic regression alongside the proposed estimate derived from it. We then estimate the effective asymptotic variance, σ_*^2/μ_*^2, for each estimator. The results, presented in Tables 2–3, indicate that the proposed estimator consistently provides a more accurate estimation of the true coefficient vector compared to conventional logistic regression.

	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$
Logit	1.87±0.06	0.46±0.01	0.30±0.00
proposed	0.61±0.01	0.25±0.00	0.18±0.00

TABLE 2 *Estimated effective asymptotic variance of the MLE of logistic regression and the proposed estimator for DARWIN data. We provide the average ± standard deviation by using leave-one-out split datasets.*

	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$
Logit	2.22±0.03	0.30±0.00	0.16±0.00
proposed	0.15±0.00	0.06±0.00	0.05±0.00

TABLE 3 *Estimated effective asymptotic variance of the MLE of logistic regression and the proposed estimator for speech data. We provide the average ± standard deviation by using leave-one-out split datasets.*

5. Proof Outline

We outline the proofs for each theorem in Section 3.

5.1. *Consistency of Link Estimation (Theorem 1)*

We provide an overview of the proof for Theorem 1, which comprises two primary steps: (i) the asymptotic characteristics of the index estimator W_i discussed in Section 2.1, and (ii) demonstrating the consistency of the estimator $\widehat{g}(\cdot)$ in Section 2.2, related to W_i.

5.1.1. Error of Index Estimator

We consider the distributional approximation (2.2) for the index estimator W_i, established through observable adjustments by Bellec [10]. Theorems 4.3 and 4.4 in Bellec [10] support Proposition 5, i.e., under suitable assumptions we have the following statement:

Proposition 5 *Under Assumptions 1-1 and $\mathbb{E}[y_1^2] < \infty$, there exists $Z_i \sim \mathcal{N}(0,1)$ independent of $\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} \sim \mathcal{N}(0,1)$ such that, for each $i \in [n_1]$, as $n_1 \to \infty$, it holds that*

$$\left| \tilde{\mu} W_i - \tilde{\mu} \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} - \tilde{\sigma} Z_i \right| \xrightarrow{p} 0. \tag{5.1}$$

This asserts that each $\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)}$ for $i \in [n_1]$ is approximately equal to the sum of the biased true index $\tilde{\mu} \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)}$, a Gaussian error, and an additive bias term. We can see that the estimation error of the index estimator W_i is asymptotically represented by the Gaussian term as shown in Equation (2.2).

5.1.2. Error of Link Estimator

Next, we prove the consistency of the link estimator $\widehat{g}(\cdot)$ using the index estimator W_i. To this aim, we define a noise-contaminated index

$$\tilde{W}_i := \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} + \tilde{Z}_i,$$

where $\tilde{Z}_i \sim \mathcal{N}(0, \sigma_1^2/\mu_1^2)$ is an independent Gaussian variable. If W_i were exactly equivalent to $\tilde{W}_i$, we could apply the classical result of nonparametric error-in-variables regression [32] to demonstrate the uniform consistency of $\widehat{g}(\cdot)$. However, this equivalence is only asymptotic as shown in (5.1). Therefore, we establish that the error due to this asymptotic equivalence is negligibly small in the estimation of $\widehat{g}(\cdot)$ to complete the proof.

Specifically, we take the following steps. First, we decompose the error of $\widehat{g}(\cdot)$ into two terms. In preparation, we define $\tilde{g}(\cdot)$ as a deconvolution estimator using a deconvolution kernel $\tilde{K}_n(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}\exp(-\mathrm{i}tx)\phi_K(t)/\phi_\varsigma(t/h_n)dt$ using the true inferential parameters as $\varsigma = \sigma_*/\mu_*$ (its precise definition is given in Supplementary Material). This estimator corresponds to the estimator for the error-in-variable setup developed by [32]. Then, from the effect of the monotonization operator, we obtain the following decomposition:

$$\begin{aligned}\sup_{a\le x\le b}|\widehat{g}(x)-g(x)| &\le \sup_{a\le x\le b}|\breve{g}(x)-g(x)| \\ &\le \sup_{a\le x\le b}|\breve{g}(x)-\tilde{g}(x)| + \sup_{a\le x\le b}|\tilde{g}(x)-g(x)|. \end{aligned} \tag{5.2}$$

The second term $\sup_{a\le x\le b}|\tilde{g}(x)-g(x)|$ in (5.2) is the estimation error by the deconvolution estimator $\tilde{g}(\cdot)$, which is proven to be $o_{\mathrm{p}}(1)$ according to the result of [32].

On the other hand, the first term $\sup_{a\le x\le b}|\breve{g}(x)-\tilde{g}(x)|$ in (5.2) represents how our pre-monotonized estimator $\breve{g}(\cdot)$ in (2.3) approximates the estimator $\tilde{g}(\cdot)$. Rigorously, we obtain

$$\begin{aligned}|\breve{g}(x)-\tilde{g}(x)| \lesssim \underbrace{\frac{1}{n_1 h_n}\left|\sum_{i=1}^{n_1} K_n\left(\frac{\tilde{W}_i-x}{h_n}\right)-K_n\left(\frac{W_i-x}{h_n}\right)\right|}_{=:T_1} \\ + \underbrace{\frac{1}{n_1 h_n}\left|\sum_{i=1}^{n_1} K_n\left(\frac{W_i-x}{h_n}\right)-\tilde{K}_n\left(\frac{W_i-x}{h_n}\right)\right|}_{=:T_2},\end{aligned}$$

where $\lesssim$ is an inequality up to some universal constant. The first term T_1 describes the discrepancy between the estimator with the index estimator W_i and the contaminated index $\tilde{W}_i$. We develop an upper bound on T_1 by using the result of Proposition 5. The second term T_2 represents the discrepancy between the convolution kernels $K_n(\cdot)$ and $\tilde{K}_n(\cdot)$. Note that $K_n(\cdot)$ depends on the estimator $\tilde{\varsigma}^2 = \tilde{\sigma}^2/\tilde{\mu}^2$ of the inferential parameter, and $\tilde{K}_n(\cdot)$ depends on the true value of the inferential parameter $\varsigma = \sigma_*/\mu_*$. We derive its upper bound by evaluating the error of the estimators $\tilde{K}_n(\cdot)$.

By integrating these results into (5.2), we prove that the estimation error of $\widehat{g}(\cdot)$ is $O_p((\log n_1)^{-m/2})$

5.2. *Marginal Asymptotic Normality (Theorem 4)*

This section provides a proof sketch of Theorem 4. We specifically present a general theorem that characterizes the asymptotic normality of each coordinate of the unregularized estimator in high-dimensional settings. This discussion extends the proof provided by Zhao et al. [76] for logistic regression.

Consider the single-index model given by (3.1) and an arbitrary loss function $\bar{\ell}:\mathbb{R}\times\mathscr{Y}\to\mathbb{R}$. We define an M-estimator $\bar{\boldsymbol{\beta}}$, based on the loss function $\bar{\ell}(\cdot)$, as follows:

$$\bar{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p}\sum_{i=1}^{n}\bar{\ell}(\boldsymbol{b}^\top\boldsymbol{X}_i;y_i). \tag{5.3}$$

With this general setup, we establish the following statement:

Theorem 6 *Suppose that Assumption 1 holds. Also, suppose that the M-estimator* $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$ *in* (5.3) *is uniquely determined and there exists a constant* $C > 0$ *such that* $\mathbb{P}(\|\bar{\boldsymbol{\beta}}\| < C) \geq 1 - o(1)$ *holds. With the true parameter* $\boldsymbol{\beta} \in \mathbb{R}^p$*, define*

$$\mu_{\bar{\boldsymbol{\beta}}} = \frac{\bar{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}}{\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}}, \quad \text{and} \quad \sigma_{\bar{\boldsymbol{\beta}}}^2 = \|\boldsymbol{P}_{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}}^{\perp} \boldsymbol{\Sigma}^{1/2} \bar{\boldsymbol{\beta}}\|^2 = \|\bar{\boldsymbol{\beta}} - \mu_{\bar{\boldsymbol{\beta}}} \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2, \tag{5.4}$$

where $\boldsymbol{P}_{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}}^{\perp} = \boldsymbol{I}_p - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{1/2} / \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$*. Then, for any coordinates* $j \in [p]$ *with* $\sqrt{p}\tau_j \beta_j = O(1)$*, we obtain*

$$T_j := \frac{\sqrt{p}(\bar{\beta}_j - \mu_{\bar{\boldsymbol{\beta}}} \beta_j)}{\sigma_{\bar{\boldsymbol{\beta}}} / \tau_j} \xrightarrow{d} \mathcal{N}(0,1).$$

as $n, p \to \infty$ *with* $p/n \to \bar{\kappa} < 1$.

This theorem establishes the marginal asymptotic normality for a broad class of estimators defined by the minimization of convex loss functions. Additionally, it demonstrates that the limiting distributional behavior of $\bar{\boldsymbol{\beta}}$ is characterized by $\mu_{\bar{\boldsymbol{\beta}}}$ and $\sigma_{\bar{\boldsymbol{\beta}}}^2$ in the high-dimensional setting (1.2). Intuitively, $\mu_{\bar{\boldsymbol{\beta}}}$ is a scaled inner product of $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, and $\sigma_{\bar{\boldsymbol{\beta}}}^2$ denotes the magnitude of the orthogonal component of $\bar{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$.

The rigorous proof in Supplementary Material is conducted in the following steps:

(i) Since we have $\boldsymbol{\beta}^\top \boldsymbol{X}_i = (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i)^\top (\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta})$, we achieve the replacements $\boldsymbol{X}_i$ to $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{\beta}$ to $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$, and $\bar{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\eta}} = \boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{\beta}}$. From the Cholesky factorization of $\boldsymbol{\Sigma}$, we have

$$T_j = \frac{\sqrt{p}(\bar{\beta}_j - \mu_{\bar{\boldsymbol{\beta}}} \beta_j)}{\sigma_{\bar{\boldsymbol{\beta}}} / \tau_j} = \frac{\sqrt{p}(\widehat{\eta}_j - \mu_{\bar{\boldsymbol{\beta}}} \eta_j)}{\sigma_{\bar{\boldsymbol{\beta}}}}.$$

(ii) Considering the rotation $\boldsymbol{U}$ around $\boldsymbol{\eta}$ (i.e., $\boldsymbol{U}\boldsymbol{\eta} = \boldsymbol{\eta}$ and $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_p$), several calculations give, for $\boldsymbol{T} := (T_1, \ldots, T_p)^\top / \sqrt{p}$,

$$\boldsymbol{T} = \frac{\boldsymbol{P}_{\boldsymbol{\eta}}^{\perp} \widehat{\boldsymbol{\eta}}}{\|\boldsymbol{P}_{\boldsymbol{\eta}}^{\perp} \widehat{\boldsymbol{\eta}}\|} \overset{\mathrm{d}}{=} \frac{\boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\eta}}^{\perp} \widehat{\boldsymbol{\eta}}}{\|\boldsymbol{P}_{\boldsymbol{\eta}}^{\perp} \widehat{\boldsymbol{\eta}}\|}.$$

This means that $\boldsymbol{T}$ is uniformly distributed on the unit sphere in $\boldsymbol{\eta}^{\perp}$ (See Figure 4).

(iii) Drawing on the analogy to the asymptotic equivalence between the p-dimensional standard normal distribution and $\mathrm{Unif}(\sqrt{p}\mathbb{S}^{p-1})$, we obtain the asymptotic normality of T_j.

We apply this general theorem to obtain Theorem 4. A similar argument implies Theorem 2 for the regularized estimator.

6. Other Design of Pilot Estimator

We can consider alternative estimators as the pilot estimator discussed in Section 2.1. Depending on the context, choosing an appropriate pilot estimator can enhance the asymptotic efficiency of the overall estimation process. Below, we list the various estimator options and their associated values necessary for estimating their inferential parameters.

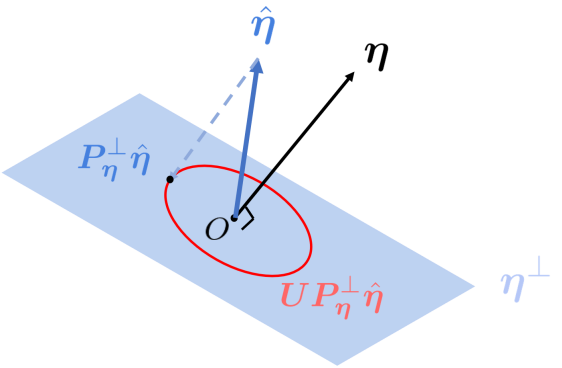

FIG. 4. Illustration of the proof technique of Theorem 6. $\mu_{\tilde{\beta}}$ is the inner product of $(\boldsymbol{\eta}, \widehat{\boldsymbol{\eta}})$. The radius of the set depicted by the red circle corresponds to $\sigma_{\tilde{\beta}}$.

6.1. *Least Squares Estimators*

In the case of $\kappa_1 < 1$, we can use the least squares estimator

$$\tilde{\boldsymbol{\beta}}_{\mathrm{LS}} = ((\boldsymbol{X}^{(1)})^\top \boldsymbol{X}^{(1)})^{-1} (\boldsymbol{X}^{(1)})^\top \boldsymbol{y}.$$

In this case, there exist corresponding inferential parameters of $\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}$.

We obtain the following marginal asymptotic normality of the least-squares estimator. We recall the definition of inferential parameters in (5.4) and consider the corresponding parameter $\mu_{\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}}$ and $\sigma_{\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}}$ by substituting $\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}$. Then, we obtain the following result by a straightforward application of Theorem 6.

Corollary 7 *Suppose Assumption 1 holds. Then, for any coordinates $j = 1, \ldots, p$ obeying $\sqrt{p}\tau_j \beta_j = O(1)$, we have the following as $n, p \to \infty$ with $\lim_{n\to\infty} \kappa_1 \in (0,1)$:*

$$\frac{\sqrt{p}(\tilde{\beta}_{\mathrm{LS},j} - \mu_{\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}} \beta_j)}{\sigma_{\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}} / \tau_j} \xrightarrow{d} \mathcal{N}(0,1).$$

We also define the following values $(\tilde{\gamma}_{\mathrm{LS}}, \tilde{\mu}_{\mathrm{LS}}, \tilde{\sigma}^2_{\mathrm{LS}})$ to estimate the inferential parameters $\mu_{\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}}$ and $\sigma_{\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}}$. Namely, we define $\tilde{\gamma}_{\mathrm{LS}} = \kappa_1/(1-\kappa_1)$, and

$$\tilde{\mu}_{\mathrm{LS}} = \left| \frac{\|\boldsymbol{X}^{(1)} \tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2}{n_1} - (1-\kappa_1)\tilde{\sigma}^2_{\mathrm{LS}} \right|^{1/2}, \tilde{\sigma}^2_{\mathrm{LS}} = \frac{\tilde{\gamma}_{\mathrm{LS}}}{n_1(1-\kappa_1)} \|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)} \tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2.$$

If we employ the least squares estimator $\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}$ as the pilot estimator in Section 2.1, we replace $(\tilde{\mu}, \tilde{\sigma}^2)$ for the index estimator W_i in (2.1) by $(\tilde{\mu}_{\mathrm{LS}}, \tilde{\sigma}^2_{\mathrm{LS}})$.

6.2. *Maximum Likelihood Estimators*

When y_i takes discrete values, a more appropriate pilot estimator can be proposed. For binary outcomes such as in classification problems where $y_i \in \{0,1\}$, we can employ MLE for logistic regression:

$$\tilde{\boldsymbol{\beta}}_{\mathrm{mle}} \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i=1}^{n_1} \log(1+\exp(\boldsymbol{b}^\top \boldsymbol{X}_i^{(1)})) - y_i \boldsymbol{b}^\top \boldsymbol{X}_i^{(1)}.$$

In the case with $y_i \in \mathbb{N}\cup\{0\}$, we can consider the MLE for the Poisson regression

$$\tilde{\boldsymbol{\beta}}_{\mathrm{mle}} \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i=1}^{n_1} \exp(\boldsymbol{b}^\top \boldsymbol{X}_i^{(1)}) - y_i \boldsymbol{b}^\top \boldsymbol{X}_i^{(1)}.$$

Its asymptotic normality is obtained by applying Theorem 6.

Corollary 8 *Under Assumption 1, suppose that $\tilde{\boldsymbol{\beta}}_{\mathrm{mle}}$ is uniquely determined and there exists a constant $C>0$ such that $\mathbb{P}(\|\tilde{\boldsymbol{\beta}}_{\mathrm{mle}}\| < C) \geq 1-o(1)$ holds as $n,p\to\infty$. Then, for any coordinates $j=1,\ldots,p$ obeying $\sqrt{p}\tau_j\beta_j = O(1)$, we have the following as $n,p\to\infty$ with $\lim_{n\to\infty}\kappa_1 \in (0,1)$:*

$$\frac{\sqrt{p}(\tilde{\beta}_{\mathrm{mle},j} - \mu_{\tilde{\boldsymbol{\beta}}_{\mathrm{mle}}}\beta_j)}{\sigma_{\tilde{\boldsymbol{\beta}}_{\mathrm{mle}}}/\tau_j} \xrightarrow{d} \mathcal{N}(0,1),$$

In these cases, we can define values $(\tilde{\gamma}_{\mathrm{mle}}, \tilde{\mu}_{\mathrm{mle}}, \tilde{\sigma}_{\mathrm{mle}})$ for estimating their inferential parameters. Define $g_0(x) = 1/(1+\exp(-x))$ for logistic regression and $g_0(x)=\exp(x)$ for Poisson regression. Then, we define the values as $\tilde{\gamma}_{\mathrm{mle}} = \kappa_1 \tilde{v}_{\mathrm{mle}}^{-1}$ and

$$\tilde{\mu}_{\mathrm{mle}} = \left| \frac{\|\boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathrm{mle}}\|^2}{n_1} - (1-\kappa_1)\tilde{\sigma}_{\mathrm{mle}}^2 \right|^{1/2}, \quad \tilde{\sigma}_{\mathrm{mle}}^2 = \frac{\|\boldsymbol{y}^{(1)} - g_0(\boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathrm{mle}})\|^2}{n_1 \tilde{v}_{\mathrm{mle}}/\kappa_1},$$

where we define $\tilde{v}_{\mathrm{mle}} = n_1^{-1}\mathrm{tr}(\tilde{\boldsymbol{D}} - \tilde{\boldsymbol{D}}\boldsymbol{X}^{(1)}((\boldsymbol{X}^{(1)})^\top \tilde{\boldsymbol{D}}\boldsymbol{X}^{(1)})^{-1}(\boldsymbol{X}^{(1)})^\top \tilde{\boldsymbol{D}})$ and $\tilde{\boldsymbol{D}} = \mathrm{diag}(g_0'(\boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathrm{mle}}))$. Based on this definition, we can develop a corresponding index estimator by replacing $(\tilde{\mu},\tilde{\sigma})$ in (2.2) by $\tilde{\mu}_{\mathrm{mle}}$ and $\tilde{\sigma}_{\mathrm{mle}}$.

7. Conclusion and Discussion

This study establishes a novel statistical inference procedure for high-dimensional single-index models. Specifically, we develop a consistent estimation method for the link function. Furthermore, using the estimated link function, we formulate an efficient estimator and confirm its marginal asymptotic normality. This verification allows for the accurate construction of confidence intervals and p-values for any finite collection of coordinates.

We identify several avenues for future research: (a) extending these results to cases where the covariate distribution is non-Gaussian, (b) generalizing our findings to multi-index models, and (c) confirming the marginal asymptotic normality of our proposed estimators under any form of regularization and covariance. These prospects offer intriguing possibilities for further exploration. Particularly, (c) can be achieved by imposing the exchangeability condition on the true coefficient $\boldsymbol{\beta}$ and employing a technique similar to the proof of Corollary 3.8 in Li and Sur [50].

A. Effect of Link Estimation on Inferential Parameters

The following theorem reveals that the estimation error of the link function is asymptotically negligible with respect to the observable adjustments.

Specifically, we consider a slightly modified version of the inferential estimator. In preparation, we define a censoring operator $\iota : \mathbb{R} \to \mathbb{R}$ on a interval $[a,b] \subset \mathbb{R}$ as $\iota(z) = \max(a, \min(b,z))$. Then, for any $\bar{g} : \mathbb{R} \to \mathbb{R}$, we define a truncation version of $\boldsymbol{D}$ as $\boldsymbol{D}_c = \mathrm{diag}(\bar{g}'(\iota(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g}))))$, and $\widehat{v}_{0c} = n_2^{-1}\mathrm{tr}(\boldsymbol{D}_c - \boldsymbol{D}_c\boldsymbol{X}^{(2)}((\boldsymbol{X}^{(2)})^\top \boldsymbol{D}_c \boldsymbol{X}^{(2)})^{-1}(\boldsymbol{X}^{(2)})^\top \boldsymbol{D}_c)$. Further, in the case of $J(\cdot) \equiv \boldsymbol{0}$, we define the modified estimator as

$$\widehat{\mu}_{0c}(\bar{g}) = \left| \|\iota(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g}))\|^2/n_2 - (1-\kappa_2)\widehat{\sigma}_{0c}^2(\bar{g}) \right|^{1/2}, \text{ and}$$

$$\widehat{\sigma}_{0c}^2(\bar{g}) = \frac{\|\boldsymbol{y}^{(2)} - \bar{g}(\iota(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\bar{g})))\|^2}{n_2 \widehat{v}_{0c}^2/\kappa_2}.$$

Using the modified definition, we obtain the following result.

Theorem 9 *Suppose that $J(\cdot) \equiv \boldsymbol{0}$ holds and the estimator (2.5) exists. Further, suppose that Assumptions 1-4 hold. Then, we have the following as $n_1 \to \infty$:*

$$|\widehat{\mu}_{0c}(\widehat{g}) - \widehat{\mu}_{0c}(g)| \xrightarrow{p} 0, \quad \text{and} \quad \left|\widehat{\sigma}_{0c}^2(\widehat{g}) - \widehat{\sigma}_{0c}^2(g)\right| \xrightarrow{p} 0.$$

This result indicates that, since the link estimator $\widehat{g}(\cdot)$ is consistent, we can estimate the inferential parameters under the true link $g(\cdot)$.

The difficulty in this proof arises from the dependence between the elements of the estimator, which cannot be handled by the triangle inequality or Hölder's inequality, To overcome the difficulty, we utilize the Azuma-Hoeffding inequality for martingale difference sequences.

B. Theoretical Efficiency Comparison

We compare the efficiency of our estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ with that of the ridge estimator $\tilde{\boldsymbol{\beta}}$ as the pilot. As shown in Bellec [10], the ridge estimator is a valid estimator for the single-index model in the high-dimensional scheme (1.2) even without estimating the link function $g(\cdot)$.

To the aim, we define the *effective asymptotic variance* based on inferential parameters, which is a ratio of the asymptotic bias and the asymptotic variance. That is, our estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ has its effective asymptotic variance $\widehat{\sigma}^2(\widehat{g})/\widehat{\mu}^2(\widehat{g})$, and the ridge estimator $\tilde{\boldsymbol{\beta}}$ has $\tilde{\sigma}^2/\tilde{\mu}^2$. The effective asymptotic variance corresponds to the asymptotic variance of each coordinate of the estimators with bias correction.

We give the following result for necessary and sufficient conditions for the proposed estimator to be more efficient than the least squares estimator and the ridge estimator.

Proposition 10 *We consider the coefficient estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ with $J(\boldsymbol{b}) \equiv \lambda\|\boldsymbol{b}\|^2$ and the setup $n_1 = n_2$. We use the regularization parameter $\lambda_1 > 0$ for the pilot estimator $\tilde{\boldsymbol{\beta}}$. Suppose that Assumptions 1-3 are fulfilled. Then, $\widehat{\sigma}^2(\widehat{g})/\widehat{\mu}^2(\widehat{g}) < \tilde{\sigma}^2/\tilde{\mu}^2$ holds if and only if we have*

$$\frac{\|\widehat{\boldsymbol{\beta}}(\widehat{g})\|}{\|\tilde{\boldsymbol{\beta}}\|} \cdot \frac{|\widehat{v}_\lambda + \lambda|}{|\tilde{v} + \lambda_1|} \cdot \frac{\|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}\|}{\|\boldsymbol{y}^{(2)} - \widehat{g}(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|} > 1.$$

TABLE C.4 *p-values of statistical tests on independence*

	logit	Pois	Cubic
$\kappa = 0.1$	0.759	0.441	0.478
$\kappa = 0.5$	0.912	0.123	0.383
$\kappa = 1$	0.602	0.060	0.825
$\kappa = 2$	0.931	0.531	0.704

This necessary and sufficient condition suggests that our estimator may have an advantage by exploiting the nonlinearity of the link function $g(\cdot)$. The first reason is that, when $\boldsymbol{y}$ has nonlinearity in $\boldsymbol{X}\boldsymbol{\beta}$, the residual $\|\boldsymbol{y} - \widehat{g}(\boldsymbol{X}\widehat{\boldsymbol{\beta}})\|^2$ of the proposed method is expected to be asymptotically smaller than $\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2$. The second reason is that $\widehat{v}_\lambda$ approximates the gradient mean $n^{-1}\sum_{i=1}^n \widehat{g}'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g}))$, so this element increases when $g(\cdot)$ has a large gradient. Using these facts, the proposed method incorporates the nonlinearity of $g(\cdot)$ and helps improve efficiency.

Proposition 11 *If $J(\cdot) \equiv \mathbf{0}$, $n_1 = n_2$, and are fulfilled, then $\widehat{\sigma}_0^2(\widehat{g})/\widehat{\mu}_0^2(\widehat{g}) < \tilde{\sigma}_{\mathrm{LS}}^2/\tilde{\mu}_{\mathrm{LS}}^2$ if and only if*

$$\frac{\|\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\widehat{g})\|}{\|\boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|} \cdot \frac{|\widehat{v}_0|}{1-\kappa_1} \cdot \frac{\|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|}{\|\boldsymbol{y}^{(2)} - \widehat{g}(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|} > 1.$$

This result does not imply that the estimator $\widehat{\boldsymbol{\beta}}(\widehat{g})$ always improves the pilot estimator $\tilde{\boldsymbol{\beta}}$. Indeed, in the setting of linear regression models, Bean et al. [9] proves that the maximum likelihood estimator is not efficient in the regime (1.2). Instead, they provide the form of the optimal loss function that depends on the dimensionality κ.

C. Remarks on the condition in Theorem 1

In this section, we discuss the validity of the condition on $(Z_1, ..., Z_n)$ in Theorem 1 by numerical experiments. To this end, we conduct independence tests on $(W_1, \ldots, W_n)$ across various models and estimators. Specifically, we perform a permutation test as described below. For each model defined in Section 4, we fix $n = 1000$ and vary p for $\kappa = p/n = 0.1, 0.5, 1, 2$.

1. We generate samples following each model, and compute the corresponding $W_i,\ i \in [n]$.
2. We repeat it 100 times for independently generated samples and make a $100 \times n$ matrix of W_i's, say $\breve{\boldsymbol{W}}$.
3. We randomly permute the columns of the matrix $\breve{\boldsymbol{W}}$ 1000 times. For each permutation, we compute the dependence measure $\frac{1}{n}\sum_{i=1}^n \max_{i' \neq i} |\widehat{\boldsymbol{C}}_{i,i'}|$ where $\widehat{\boldsymbol{C}} = \frac{1}{100}\breve{\boldsymbol{W}}^\top \breve{\boldsymbol{W}}$, called Mean Maximum (absolute) Correlation (MMC).
4. We compute p-values as the proportion of times the MMC of the original $\breve{\boldsymbol{W}}$ falls below the MMC from 1000 permutations.

From the result displayed in Table C.4, the null hypotheses of independence are not rejected at conventional significance levels, thus numerically, there is no evidence suggesting that the condition does not hold.

D. Nonparametric Regression with Deconvolution

In this section, we review the concept of nonparametric regression with deconvolution to address the errors-in-variable problem. To begin with, we redefine the notation only for this section. For a pair of random variables (X,Y,Z), suppose that the model is

$$\mathbb{E}[Z \mid X = x] = m(x),$$

and that we can only observe n i.i.d. realizations of $Y = X + \varepsilon$ and Z. Here, ε is a random variable called measurement error or error in variables. For the identification, we assume that the distribution of ε is known. Let the joint distribution of (X,Z) be $f(x,z)$. By the definition of the conditional expectations, $m(x) = r(x)/f(x)$ with

$$r(x) = \int_{-\infty}^{\infty} z f(x,z) dz, \quad f(x) = \int_{-\infty}^{\infty} f(x,z) dz,$$

for the continuous random variables. The goal of the problem is to estimate the function $m(\cdot)$.

If we could observe X, a popular estimator of $m(x)$ is Nadaraya-Watson estimator $\tilde{r}(x)/\tilde{f}(x)$ with

$$\tilde{r}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Z_i K\left(\frac{x - X_i}{h_n}\right), \quad \tilde{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

where $K(\cdot)$ is a kernel function and h_n is the bandwidth. Since X is unobservable, we alternatively construct the *deconvolution* estimator [66]. Let the characteristic function of X, Y and ε be $\phi_X(\cdot)$, $\phi_Y(\cdot)$ and $\phi_\varepsilon(\cdot)$, respectively. Since the density of Y is the convolution of that of X and ε, and the convolution in the frequency domain is just a multiplication, we have $\phi_X(t) = \phi_Y(t)/\phi_\varepsilon(t)$. Thus, the inverse Fourier transform of $\phi_Y(t)/\phi_\varepsilon(t)$ gives the density of X. Since we know the distribution of ε and we can approximate $\phi_Y(t)$ by the characteristic function of the kernel density estimator of Y, we can construct an estimator of $f(x)$ as

$$\widehat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\mathrm{i}tx) \phi_K(th_n) \frac{\widehat{\phi}_Y(t)}{\phi_\varepsilon(t)} dt, \tag{D.1}$$

where we use the fact that the Fourier transform of $\tilde{f}_Y(y) = (nh_n)^{-1} \sum_{i=1}^{n} K((y - Y_i)/h_n)$ is $\phi_K(th_n)\widehat{\phi}_Y(t)$, which approximates $\phi_Y(\cdot)$. Here, $\widehat{\phi}_Y(t)$ is the empirical characteristic function:

$$\widehat{\phi}_Y(t) = \frac{1}{n} \sum_{i=1}^{n} \exp(\mathrm{i}tY_i).$$

We can rewrite (D.1) in a kernel form

$$\widehat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K_n\left(\frac{x - Y_i}{h_n}\right),$$

with

$$K_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\mathrm{i}tx) \frac{\phi_K(t)}{\phi_\varepsilon(t/h_n)} dt.$$

Using this, Fan and Truong [32] proposes a kernel regression estimator $\widehat{m}(x) = \widehat{r}(x)/\widehat{f}(x)$ involving errors in variables with

$$\widehat{r}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} Z_i K_n\left(\frac{x-Y_i}{h_n}\right).$$

To establish the theoretical guarantee, we impose the following assumptions:

(N1) (Super-smoothness of the distribution of ε) There exists constants $d_0, d_1, \beta, \gamma > 0$ and $\beta_0, \beta_1 \in \mathbb{R}$ satisfying, as $t \to \infty$,

$$d_0 |t|^{\beta_0} \exp(-|t|^{\beta}/\gamma) \le |\phi_\varepsilon(t)| \le d_1 |t|^{\beta_1} \exp(-|t|^{\beta}/\gamma).$$

(N2) The characteristic function of the error distribution $\phi_\varepsilon(\cdot)$ does not vanish.
(N3) Let $a < b$. The marginal density $f_X(\cdot)$ of the unobserved X is bounded away from zero on the interval $[a,b]$, and has a bounded k-th derivative.
(N4) The true regression function $m(\cdot)$ has a continuous k-th derivative on $[a,b]$.
(N5) The conditional second moment $\mathbb{E}[Z^2 \mid X = x]$ is continuous on $[a,b]$, and $\mathbb{E}[Z^2] < \infty$.
(N6) The kernel $K(\cdot)$ is a k-th order kernel. Namely,for $j = 1, \ldots, k-1$, it holds that

$$\int_{-\infty}^{\infty} K(t)dt = 1, \quad \int_{-\infty}^{\infty} t^k K(t)dt \neq 0, \quad \int_{-\infty}^{\infty} t^j K(t)dt = 0.$$

(N1) includes Gaussian distributions for $\beta = 2$ and Cauchy distributions for $\beta = 1$. For a positive constant B, define a set of function

$$\mathscr{F} = \left\{ f(x,z) : \left| f_X^{(k)}(\cdot) \right| \le B, \min_{a \le x \le b} f_X(x) \ge B^{-1}, \sup_{a \le x \le b} \left| m^{(j)}(x) \right| \le B, j = 0, 1, \ldots, k \right\}.$$

In this setting, we have the uniform consistency of $\widehat{m}(\cdot)$ and its rate of convergence.

Lemma 12 (Theorem 2 in Fan and Truong [32]) *Assume (N1)-(N6) and that $\phi_K(t)$ has a bounded support on $|t| < M_0$. Then, for bandwidth $h_n = c(\log n)^{-1/\beta}$ with $c > M_0(2/\gamma)^{1/\beta}$,*

$$\lim_{d \to \infty} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{a \le x \le b} |\widehat{m}(x) - m(x)| \ge d(\log n)^{-k/\beta} \right) = 0,$$

holds for any $f \in \mathscr{F}$.

Furthermore, we can show the uniform convergence of the derivative of $\widehat{m}(\cdot)$. We use the following result in the proof of Theorem 1.

Lemma 13 *Under the condition of Lemma 12, we have, for any $f \in \mathscr{F}$,*

$$\sup_{a \le x \le b} |\widehat{m}'(x) - m'(x)| \xrightarrow{p} 0.$$

To prove this, we use the following two lemmas.

Lemma 14 *We have, for any $t \in \mathbb{R}$,*

$$\mathbb{E}\left[\left|\widehat{\phi}_Y(t) - \phi_Y(t)\right|^2\right] \leq n^{-1},$$

and

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i \exp(\mathrm{i}tY_i) - \mathbb{E}[Z\exp(\mathrm{i}tY)]\right|^2\right] \leq n^{-1}\mathbb{E}[Z^2].$$

Proof of Lemma 14 We decompose the term on the left-hand side in the first statement by Euler's formula as

$$\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\phi}_Y(t) - \phi_Y(t)\right|^2\right] \\
&= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} e^{\mathrm{i}tY_i} - \mathbb{E}e^{\mathrm{i}tY}\right|^2\right] \\
&= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\{\cos(tY_i) - \mathbb{E}\cos(tY)\} - \frac{\mathrm{i}}{n}\sum_{i=1}^{n}\{\sin(tY_i) - \mathbb{E}\sin(tY)\}\right|^2\right] \\
&= \mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\cos(tY_i) - \mathbb{E}\cos(tY)\right\}^2 - \left\{\frac{1}{n}\sum_{i=1}^{n}\sin(tY_i) - \mathbb{E}\sin(tY)\right\}^2\right] \\
&\leq \mathrm{Var}\left(n^{-1}\sum_{i=1}^{n}\cos(tY_i)\right) + \mathrm{Var}\left(n^{-1}\sum_{i=1}^{n}\sin(tY_i)\right) \\
&\leq n^{-1}\mathbb{E}\left[\cos(tY_1)^2 + \sin(tY_1)^2\right] = n^{-1}.
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i \exp(\mathrm{i}tY_i) - \mathbb{E}[Z\exp(\mathrm{i}tY)]\right|^2\right] \\
&= \frac{1}{n}\mathrm{Var}(Z_1\cos(tY_1)) + \frac{1}{n}\mathrm{Var}(Z_1\sin(tY_1)) \\
&\leq n^{-1}\mathbb{E}[Z^2].
\end{aligned}$$

This completes the proof. □

Lemma 15 *Under the setting of Lemma 12, for bandwidth $h_n = c(\log n)^{-1/\beta}$ with $c > M_0(2/\gamma)^{1/\beta}$, we have*

$$n^{-1}\sup_x |K_n(x)|^2 = o(1).$$

Proof of Lemma 15 At first, (N1) implies that there exists a constant M such that

$$|\phi_\varepsilon(t)| > \frac{d_0}{2}|t|^{\beta_0}\exp(-|t|^\beta/\gamma),$$

for $|t| > M$. By the fact that $|\exp(-itx)| = 1$ and that the support of $\phi_K(\cdot)$ is bounded by M_0, we have

$$\begin{aligned}
\sup_x |K_n(x)| &\le \int_{-\infty}^{\infty} \frac{|\phi_K(t)|}{|\phi_\varepsilon(t/h_n)|}dt \\
&\le 2\int_0^{Mh_n} \frac{|\phi_K(t)|}{|\phi_\varepsilon(t/h_n)|}dt + \frac{4}{d_0}\int_{Mh_n}^{M_0} |\phi_K(t)| \left|\frac{t}{h_n}\right|^{-\beta_0} \exp\left(\frac{|t/h_n|^\beta}{\gamma}\right)dt \\
&\le 2h_n\int_0^M \frac{1}{|\phi_\varepsilon(u)|}du + \frac{4}{d_0}(M_0 - Mh_n)h_n^{\beta_0}M_0^{-\beta_0}\exp\left(\frac{|M_0/h_n|^\beta}{\gamma}\right) \\
&= O(h_n) + O(h_n^{\beta_0}\exp(|M_0/h_n|^\beta/\gamma)).
\end{aligned}$$

Here, we use the fact that $|\phi_K(t)| \le \int |e^{-itx}||K(x)|dx < \infty$. Since we choose $h_n = c(\log n)^{-1/\beta}$ with $c > M_0(2/\gamma)^{1/\beta}$, we obtain the conclusion. □

Proof of Lemma 13 Let $a \le x \le b$. To begin with, by the triangle inequality, we have

$$\begin{aligned}
&\sup_{a\le x\le b}\left|\widehat{m}'(x) - m'(x)\right| \\
&= \sup_{a\le x\le b}\left|\frac{\widehat{r}'(x)\widehat{f}(x) - \widehat{r}(x)\widehat{f}'(x)}{\widehat{f}(x)^2} - \frac{r'(x)f(x) - r(x)f'(x)}{f(x)^2}\right| \\
&\le \sup_{a\le x\le b}\left|\frac{\widehat{r}'(x)\widehat{f}(x) - \widehat{r}(x)\widehat{f}'(x) - r'(x)f(x) + r(x)f'(x)}{f(x)^2}\right| \\
&\quad + \sup_{a\le x\le b}\left|\frac{\widehat{r}'(x)\widehat{f}(x) - \widehat{r}(x)\widehat{f}'(x)}{f(x)^2}\left(\frac{f(x)^2}{\widehat{f}(x)^2} - 1\right)\right| \\
&\le B^2 \sup_{a\le x\le b}\left|\widehat{r}'(x)\widehat{f}(x) - \widehat{r}(x)\widehat{f}'(x) - r'(x)f(x) + r(x)f'(x)\right| \\
&\quad + B^2 \sup_{a\le x\le b}\left|\widehat{r}'(x)\widehat{f}(x) - \widehat{r}(x)\widehat{f}'(x)\right|\left|\frac{f(x)^2 - \widehat{f}(x)^2}{\widehat{f}(x)^2}\right|, \qquad \text{(D.2)}
\end{aligned}$$

where the last inequality uses the assumption $\min_{a\le x\le b}|f(x)| \ge B^{-1}$. We consider showing the convergence in probability by showing the L^1 convergence. Using the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\sup_{a\le x\le b}\left|\widehat{r}'(x)\widehat{f}(x) - \widehat{r}(x)\widehat{f}'(x) - r'(x)f(x) + r(x)f'(x)\right|$$

$$
\begin{aligned}
&\leq \mathbb{E}\sup_{a\leq x\leq b}\left|\widehat{f}(x)\left(\widehat{r}'(x)-r'(x)\right)\right|+\mathbb{E}\sup_{a\leq x\leq b}\left|r'(x)\left(\widehat{f}(x)-f(x)\right)\right| \\
&\quad+\mathbb{E}\sup_{a\leq x\leq b}\left|f'(x)\left(r(x)-\widehat{r}(x)\right)\right|+\mathbb{E}\sup_{a\leq x\leq b}\left|\widehat{r}(x)\left(f'(x)-\widehat{f}'(x)\right)\right| \\
&\leq \sqrt{\mathbb{E}\left[\sup_{a\leq x\leq b}\left|\widehat{f}(x)\right|^2\right]}\sqrt{\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{r}'(x)-r'(x)|^2\right]} \\
&\quad+\sup_{a\leq x\leq b}|r'(x)|\sqrt{\mathbb{E}\left[\sup_{a\leq x\leq b}\left|\widehat{f}(x)-f(x)\right|^2\right]} \\
&\quad+\sup_{a\leq x\leq b}|f'(x)|\sqrt{\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{r}(x)-r(x)|^2\right]} \\
&\quad+\sqrt{\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{r}(x)|^2\right]}\sqrt{\mathbb{E}\left[\sup_{a\leq x\leq b}\left|\widehat{f}'(x)-f'(x)\right|^2\right]}.
\end{aligned}
$$

Thus, to bound the right-hand side of (D.2), we need to show that $\mathbb{E}[\sup_x|\widehat{f}(x)|^2]$ and $\mathbb{E}[\sup_x|\widehat{r}(x)|^2]$ are bounded by constants and that $\mathbb{E}[\sup_x|\widehat{f}(x)-f(x)|^2]$, $\mathbb{E}[\sup_x|\widehat{r}(x)-r(x)|^2]$, $\mathbb{E}[\sup_x|\widehat{f}'(x)-f'(x)|^2]$, and $\mathbb{E}[\sup_x|\widehat{r}'(x)-r'(x)|^2]$ converge to zero.

• **Bound for** $\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{f}(x)-f(x)|^2\right]$. By triangle inequality and the fact that $(x+y)^2\leq 2x^2+2y^2$ for $x,y\in\mathbb{R}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{f}(x)-f(x)|^2\right] \\
&\leq 2\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{f}(x)-\mathbb{E}\widehat{f}(x)|^2\right]+2\sup_{a\leq x\leq b}|\mathbb{E}\widehat{f}(x)-f(x)|^2. \qquad \text{(D.3)}
\end{aligned}
$$

For the first term of the left-hand side of (D.3), the Cauchy-Schwarz inequality gives

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{a\leq x\leq b}|\widehat{f}(x)-\mathbb{E}\widehat{f}(x)|^2\right] \\
&\leq \frac{1}{(2\pi)^2}\mathbb{E}\left[\left\{\int_{-\infty}^{\infty}\frac{|\phi_K(th_n)|}{|\phi_\varepsilon(t)|}\left|\widehat{\phi}_Y(t)-\phi_Y(t)\right|dt\right\}^2\right] \\
&\leq \frac{1}{(2\pi)^2}\left\{\int_{-\infty}^{\infty}\frac{|\phi_K(th_n)|}{|\phi_\varepsilon(t)|}dt\right\}\left\{\int_{-\infty}^{\infty}\mathbb{E}\left[\left|\widehat{\phi}_Y(t)-\phi_Y(t)\right|^2\right]\frac{|\phi_K(th_n)|}{|\phi_\varepsilon(t)|}dt\right\}.
\end{aligned}
$$

Lemma 14 and the proof of Lemma 15 imply that this converges to zero as $n\to\infty$. Next, we consider the second term in (D.3). We obtain

$$
\mathbb{E}\left[\widehat{f}(x)\right]=\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-\mathrm{i}tx)\phi_K(th_n)\frac{\mathbb{E}_Y[\exp(\mathrm{i}tY)]}{\phi_\varepsilon(t)}dt
$$

$$
\begin{aligned}
&= \frac{1}{2\pi}\mathbb{E}_X\left[\int_{-\infty}^{\infty}\exp(\mathrm{i}tx)\phi_K(th_n)\exp(-\mathrm{i}tX)dt\right] \\
&= \mathbb{E}_X\left[\frac{1}{2\pi h_n}\int_{-\infty}^{\infty}\exp\left(\mathrm{i}t\frac{x-X}{h}\right)\phi_K(t)dt\right] \\
&= \frac{1}{h_n}\mathbb{E}_X\left[K\left(\frac{x-X}{h_n}\right)\right].
\end{aligned}
$$

Thus, a classical result for the kernel density estimation gives $\sup_x |\mathbb{E}[\widehat{f}(x)] - f(x)| \to 0$ as $n \to 0$.
• **Bound for** $\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}(x) - r(x)|^2\right]$. By triangle inequality and the fact that $(x+y)^2 \le 2x^2 + 2y^2$,

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}(x) - r(x)|^2\right] \\
&\le 2\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}(x) - \mathbb{E}\widehat{r}(x)|^2\right] + 2\sup_{a\le x\le b}|\mathbb{E}\widehat{r}(x) - r(x)|^2. \qquad \text{(D.4)}
\end{aligned}
$$

For the first term of the left-hand side of (D.4), Cauchy-Schwarz inequality gives

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}(x) - \mathbb{E}\widehat{r}(x)|^2\right] \\
&\le \frac{1}{(2\pi)^2}\mathbb{E}\left[\left\{\int_{-\infty}^{\infty}\frac{|\phi_K(th_n)|}{|\phi_\varepsilon(t)|}\left|\frac{1}{n}\sum_{i=1}^{n}Z_i\exp(\mathrm{i}tY_j) - \mathbb{E}[Z\exp(\mathrm{i}tY)]\right|dt\right\}^2\right] \\
&\le \frac{1}{(2\pi)^2}\left\{\int_{-\infty}^{\infty}\frac{|\phi_K(th_n)|}{|\phi_\varepsilon(t)|}dt\right\}\left\{\frac{1}{n}\int_{-\infty}^{\infty}\frac{|\phi_K(th_n)|}{|\phi_\varepsilon(t)|}dt\right\},
\end{aligned}
$$

where we use the proof of Lemma 15 for the last inequality. Lemma 14 implies that this term converges to zero as $n \to \infty$. Next, we consider the second term in (D.4). We have

$$
\mathbb{E}[\widehat{r}(x)] = \frac{1}{h_n}\mathbb{E}_{X,Z}\left[ZK\left(\frac{x-X}{h_n}\right)\right].
$$

Thus we have $\sup_{a\le x\le b}|\mathbb{E}[\widehat{r}(x)] - r(x)| \to 0$.
• **Bound for** $\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}'(x) - f'(x)|^2\right]$. By triangle inequality and the fact that $(x+y)^2 \le 2x^2 + 2y^2$ for $x, y \in \mathbb{R}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}'(x) - f'(x)|^2\right] \\
&\le 2\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}'(x) - \mathbb{E}\widehat{f}'(x)|^2\right] + 2\sup_{a\le x\le b}|\mathbb{E}\widehat{f}'(x) - f'(x)|^2. \qquad \text{(D.5)}
\end{aligned}
$$

For the first term of the left-hand side of (D.5), since $\partial \exp(-\mathrm{i}tx)/(\partial x) = -\mathrm{i}t\exp(-\mathrm{i}tx)$ and $|\mathrm{i}| = |\exp(-\mathrm{i}tx)| = 1$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}'(x)-\mathbb{E}\widehat{f}'(x)|^2\right] \\
&= \mathbb{E}\left[\sup_{a\le x\le b}\left|\frac{1}{2\pi}\int_{-\infty}^{\infty}-\mathrm{i}t\exp(-\mathrm{i}tx)\frac{\phi_K(th_n)}{\phi_\varepsilon(t)}\left\{\widehat{\phi}_Y(t)-\phi_Y(t)\right\}dt\right|^2\right] \\
&\le \frac{1}{(2\pi)^2}\mathbb{E}\left[\left\{\int_{-\infty}^{\infty}\frac{|t\phi_K(th_n)|}{|\phi_\varepsilon(t)|}\left|\widehat{\phi}_Y(t)-\phi_Y(t)\right|dt\right\}^2\right].
\end{aligned}$$

Thus, this converges to zero in the same way as (D.3). For the second term in (D.5), by the integration by parts,

$$\begin{aligned}
\mathbb{E}\left[\widehat{f}'(x)\right] &= \frac{1}{h_n^2}\int_{-\infty}^{\infty}K'\left(\frac{x-y}{h_n}\right)f(y)dy \\
&= \frac{1}{h_n}\int_{-\infty}^{\infty}K\left(\frac{x-y}{h_n}\right)f'(y)dy - \frac{1}{h_n}\left[K\left(\frac{x-y}{h_n}\right)f'(y)\right]_{-\infty}^{\infty}.
\end{aligned}$$

Here, the second term is zero and the first term converges to $f'(x)$ uniformly.

• **Bound for** $\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}'(x)-r'(x)|^2\right]$. By triangle inequality and the fact that $(x+y)^2 \le 2x^2+2y^2$ for $x,y\in\mathbb{R}$, we have

$$\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}'(x)-r'(x)|^2\right] \le 2\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}'(x)-\mathbb{E}\widehat{r}'(x)|^2\right] + 2\sup_{a\le x\le b}|\mathbb{E}\widehat{r}'(x)-r'(x)|^2. \tag{D.6}$$

For the first term of the left-hand side of (D.6), since $|\mathrm{i}| = |\exp(-\mathrm{i}tx)| = 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_x|\widehat{f}'(x)-\mathbb{E}\widehat{f}'(x)|^2\right] \\
&= \mathbb{E}\left[\sup_x\left|\frac{1}{2\pi}\int_{-\infty}^{\infty}-\mathrm{i}t\exp(-\mathrm{i}tx)\frac{\phi_K(th_n)}{\phi_\varepsilon(t)}\left\{\frac{1}{n}\sum_{i=1}^{n}Z_i\exp(-\mathrm{i}tY_i)-\mathbb{E}[Z\exp(-\mathrm{i}tZ)]\right\}dt\right|^2\right] \\
&\le \frac{1}{(2\pi)^2}\mathbb{E}\left[\left\{\int_{-\infty}^{\infty}\frac{|t\phi_K(th_n)|}{|\phi_\varepsilon(t)|}\left|\frac{1}{n}\sum_{i=1}^{n}Z_i\exp(-\mathrm{i}tY_i)-\mathbb{E}[Z\exp(-\mathrm{i}tZ)]\right|dt\right\}^2\right].
\end{aligned}$$

Thus, this converges to zero in the same way as (D.4). For the second term in (D.6), by the integration by parts,

$$\begin{aligned}
\mathbb{E}\left[\widehat{r}'(x)\right] &= \frac{1}{h_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}zK'\left(\frac{x-y}{h_n}\right)f(y,z)dydz \\
&= \frac{1}{h_n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}zK\left(\frac{x-y}{h_n}\right)\frac{\partial}{\partial y}f(y,z)dydz
\end{aligned}$$

$$-\frac{1}{h_n}\int_{-\infty}^{\infty}\left[zK\left(\frac{x-y}{h_n}\right)\frac{\partial}{\partial y}f(y,z)\right]_{-\infty}^{\infty}dz.$$

Here, the second term is zero and the first term converges to $r'(x)=(\partial/\partial x)\int zf(x,z)dz$ uniformly.

• **Bound for** $\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}(x)|^2\right]$ **and** $\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}(x)|^2\right]$. By triangle inequality and the fact that $(x+y)^2\le 2x^2+2y^2$,

$$\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}(x)|^2\right]\le 2\sup_{a\le x\le b}|f(x)|^2+2\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{f}(x)-f(x)|^2\right].$$

We have already shown $\mathbb{E}[\sup_{a\le x\le b}|\widehat{f}(x)-f(x)|^2]=o(1)$, $\mathbb{E}[\sup_{a\le x\le b}|\widehat{f}(x)|^2]$ is asymptotically bounded by a constant. Similarly, we can show that the other expectation $\mathbb{E}\left[\sup_{a\le x\le b}|\widehat{r}(x)|^2\right]$ is asymptotically bounded by a constant. Combining these results together, we conclude that the first term of (D.2) is $o(1)$.

Next, we consider the second term of (D.2). Since $\widehat{f}(x)$ is asymptotically bounded uniformly on $[a,b]$ by the results above, we have only to show that $\sup_{a\le x\le b}|\widehat{f}(x)^2-f(x)^2|=o(1)$. This holds since

$$\sup_{a\le x\le b}|\widehat{f}(x)^2-f(x)^2|\le\sup_{a\le x\le b}|\widehat{f}(x)+f(x)|\sup_{a\le x\le b}|\widehat{f}(x)-f(x)|=o_{\mathrm{p}}(1).$$

This concludes that $\sup_{a\le x\le b}|\widehat{m}'(x)-m'(x)|\xrightarrow{p}0$ as $n\to\infty$. □

E. Proofs of the Results

For a convex function $f:\mathbb{R}\to\mathbb{R}$ and a constant $\gamma>0$, define the proximal operator $\mathrm{prox}_{\gamma f}:\mathbb{R}\to\mathbb{R}$ as

$$\mathrm{prox}_{\gamma f}(x)=\operatorname*{arg\,min}_{z\in\mathbb{R}}\left\{\gamma f(z)+\frac{1}{2}(x-z)^2\right\}.$$

E.1. *Proof of Master Theorem*

First, we define the notation used in the proof. We consider an invertible matrix $\boldsymbol{L}\in\mathbb{R}^{p\times p}$ that satisfies $\boldsymbol{\Sigma}=\boldsymbol{L}\boldsymbol{L}^\top$. Define, for each $i\in\{1,\dots,n\}$,

$$\tilde{\boldsymbol{X}}_i=\boldsymbol{L}^{-1}\boldsymbol{X}_i,\quad \boldsymbol{\eta}=\boldsymbol{L}^\top\boldsymbol{\beta},\quad \widehat{\boldsymbol{\eta}}=\boldsymbol{L}^\top\bar{\boldsymbol{\beta}}. \tag{E.1}$$

Proof of Theorem 6 We consider the following three steps.

Step 1: Reduction to standard Gaussian features. Note that the single-index model $y_i=g(\boldsymbol{X}_i^\top\boldsymbol{\beta})+\varepsilon_i$ is equivalent to $y_i=g(\tilde{\boldsymbol{X}}_i^\top\boldsymbol{\eta})+\varepsilon_i$. Since $\boldsymbol{X}_i^\top\boldsymbol{b}=\tilde{\boldsymbol{X}}_i^\top(\boldsymbol{L}^\top\boldsymbol{b})$ holds, we have $\bar{\ell}(\boldsymbol{X}_i^\top\boldsymbol{b};y_i)=\bar{\ell}(\tilde{\boldsymbol{X}}_i^\top\boldsymbol{L}^\top\boldsymbol{b};y_i)$. Hence, $\widehat{\boldsymbol{\eta}}\in\operatorname{arg\,min}_{\tilde{\boldsymbol{b}}\in\mathbb{R}^p}\sum_{i=1}^n\bar{\ell}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{b}};y_i)$ is the estimator corresponding to the true parameter $\boldsymbol{\eta}\in\mathbb{R}^p$ and features $\tilde{\boldsymbol{X}}_i\sim\mathcal{N}_p(\boldsymbol{0},\boldsymbol{I}_p)$.

We can choose $\boldsymbol{\Sigma}=\boldsymbol{L}\boldsymbol{L}^\top$ to be a Cholesky factorization so that $\eta_p=\tau_p\beta_p$ and $\widehat{\eta}_p=\tau_p\bar{\beta}_p$ with $\tau_p=(\boldsymbol{\Sigma}^{-1})_{pp}^{-1/2}$ by (E.1). This follows from the fact that $L_{pp}=\tau_p$ since $\tau_p^2=\mathrm{Var}(X_{ip}\mid\boldsymbol{X}_{i\setminus p})=\mathrm{Var}(X_{ip}\mid\tilde{\boldsymbol{X}}_{i\setminus p})$,

where $\boldsymbol{X}_{i\backslash p} \in \mathbb{R}^{p-1}$ denotes the vector $\boldsymbol{X}_i$ without pth coordinate. Since we can generalize this to any coordinate by permutation, we obtain

$$\tau_j \frac{\bar{\beta}_j - \mu_{\bar{\boldsymbol{\beta}}}\beta_j}{\sigma_{\bar{\boldsymbol{\beta}}}} = \frac{\widehat{\eta}_j - \mu_{\bar{\boldsymbol{\beta}}}\eta_j}{\sigma_{\bar{\boldsymbol{\beta}}}},$$

for each $j \in \{1,\dots,p\}$ and any pair $(\mu_{\bar{\boldsymbol{\beta}}}, \sigma_{\bar{\boldsymbol{\beta}}})$.

Step 2: Reduction to uniform distribution on sphere. Define an orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{\eta}} = \boldsymbol{\eta}\boldsymbol{\eta}^\top / \|\boldsymbol{\eta}\|^2$ onto $\boldsymbol{\eta}$, and an orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{\eta}}^\perp = \boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{\eta}}$ onto the orthogonal complement of $\boldsymbol{\eta}$. Let $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ be any orthogonal matrix obeying $\boldsymbol{U}\boldsymbol{\eta} = \boldsymbol{\eta}$, namely, any rotation operator about $\boldsymbol{\eta}$. Then, since $\widehat{\boldsymbol{\eta}} = \boldsymbol{P}_{\boldsymbol{\eta}}\widehat{\boldsymbol{\eta}} + \boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}}$, we have

$$\boldsymbol{U}\widehat{\boldsymbol{\eta}} = \boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\eta}}\widehat{\boldsymbol{\eta}} + \boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}} = \boldsymbol{P}_{\boldsymbol{\eta}}\widehat{\boldsymbol{\eta}} + \boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}}.$$

Using this, we obtain

$$\frac{\boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}}}{\|\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}}\|} \overset{\mathrm{d}}{=} \frac{\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}}}{\|\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\widehat{\boldsymbol{\eta}}\|} = \frac{\widehat{\boldsymbol{\eta}} - \mu_{\bar{\boldsymbol{\beta}}}\boldsymbol{\eta}}{\sigma_{\bar{\boldsymbol{\beta}}}}, \tag{E.2}$$

where the first identity follows from the fact that $\boldsymbol{U}\widehat{\boldsymbol{\eta}} \overset{\mathrm{d}}{=} \widehat{\boldsymbol{\eta}}$ since $\boldsymbol{U}\widehat{\boldsymbol{\eta}}$ is the estimator with a true coefficient $\boldsymbol{U}\boldsymbol{\eta} = \boldsymbol{\eta}$ and features drawn iid from $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, by $\bar{\ell}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{b}}; y_i) = \bar{\ell}((\boldsymbol{U}^\top\tilde{\boldsymbol{X}}_i)^\top\boldsymbol{U}\tilde{\boldsymbol{b}}; y_i)$ and $\boldsymbol{U}^\top\tilde{\boldsymbol{X}}_i \overset{\mathrm{d}}{=} \tilde{\boldsymbol{X}}_i$. (E.2) reveals that $(\widehat{\boldsymbol{\eta}} - \mu_{\bar{\boldsymbol{\beta}}}\boldsymbol{\eta})/\sigma_{\bar{\boldsymbol{\beta}}}$ is rotationally invariant about $\boldsymbol{\eta}$, lies in $\boldsymbol{\eta}^\perp$, and has a unit norm. This means $(\widehat{\boldsymbol{\eta}} - \mu_{\bar{\boldsymbol{\beta}}}\boldsymbol{\eta})/\sigma_{\bar{\boldsymbol{\beta}}}$ is uniformly distributed on the unit sphere lying in $\boldsymbol{\eta}^\perp$.

Step 3: Deriving asymptotic normality. The result of the previous step gives us

$$\frac{\widehat{\boldsymbol{\eta}} - \mu_{\bar{\boldsymbol{\beta}}}\boldsymbol{\eta}}{\sigma_{\bar{\boldsymbol{\beta}}}} \overset{\mathrm{d}}{=} \frac{\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\boldsymbol{Z}}{\|\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\boldsymbol{Z}\|}, \tag{E.3}$$

where $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Triangle inequalities yield that

$$\frac{\|\boldsymbol{Z}\|}{\sqrt{p}} - \frac{|\boldsymbol{\eta}^\top\boldsymbol{Z}|}{\sqrt{p}\,\|\boldsymbol{\eta}\|} \le \frac{\|\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\boldsymbol{Z}\|}{\sqrt{p}} \le \frac{\|\boldsymbol{Z}\|}{\sqrt{p}} + \frac{|\boldsymbol{\eta}^\top\boldsymbol{Z}|}{\sqrt{p}\,\|\boldsymbol{\eta}\|}.$$

Since $|\boldsymbol{\eta}^\top\boldsymbol{Z}|/(\sqrt{p}\,\|\boldsymbol{\eta}\|) \xrightarrow{\text{a.s.}} 0$ and $\|\boldsymbol{Z}\|/\sqrt{p} \xrightarrow{\text{a.s.}} 1$, we obtain $\|\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\boldsymbol{Z}\|/\sqrt{p} \xrightarrow{\text{a.s.}} 1$. Therefore, this fact and (E.3) imply that

$$\sqrt{p}\frac{\widehat{\eta}_j - \mu_{\bar{\boldsymbol{\beta}}}\eta_j}{\sigma_{\bar{\boldsymbol{\beta}}}} \overset{\mathrm{d}}{=} \breve{\sigma}_j Q + o_{\mathrm{p}}(1), \quad \breve{\sigma}_j^2 = 1 - \frac{\eta_j^2}{\|\boldsymbol{\eta}\|^2},$$

where $Q \sim \mathcal{N}(0,1)$. Here we use the fact that the covariance matrix of $\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\boldsymbol{Z}$ is $\boldsymbol{P}_{\boldsymbol{\eta}}^\perp\boldsymbol{P}_{\boldsymbol{\eta}}^\perp = \boldsymbol{I}_p - \boldsymbol{\eta}\boldsymbol{\eta}^\top/\|\boldsymbol{\eta}\|^2$. The assumptions $\eta_j = o(1)$ and $\|\boldsymbol{\eta}\| = 1$ complete the proof. □

E.2. *Proof of Theorem 1*

Let ς^2 be the ratio σ_1^2/μ_1^2, where μ_1^2 and σ_1^2 are the true inferential parameters of the pilot estimator $\tilde{\boldsymbol{\beta}}$.

Proof of Proposition 5 First, define $\gamma_1 = \mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{X}^{(1)\top}\boldsymbol{X}^{(1)} + n\lambda \boldsymbol{I}_p)^{-1})$. We also define

$$\mu_1 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}, \quad \text{and} \quad \sigma_1^2 = \tilde{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} - \mu_1^2.$$

Let $\tilde{r}^2$ be the mean squared error $n_1^{-1}\|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\tilde{\boldsymbol{\beta}}\|^2$. Since $\tilde{\boldsymbol{\beta}}$ is a ridge estimator, Theorem 4.3 in Bellec [10] implies,

$$\max_{1\le i\le n_1} \mathbb{E}\left[\tilde{r}^{-2}\left|\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} - \mathrm{prox}_{\gamma_1 f}\left(\mu_1 \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} + \sigma_1 Z_i\right)\right|^2\right] \le \frac{C}{n_1}, \tag{E.4}$$

with $f(t) = t^2/2$ and $Z_i \sim \mathcal{N}(0,1)$ independent of $\boldsymbol{\beta}^\top \boldsymbol{X}^{(1)}$. Since ridge regression satisfies $\|\tilde{\boldsymbol{\beta}}\| \le C_\lambda$ with a constant $C_\lambda > 0$ depending on the regularization parameter $\lambda > 0$ by the KKT condition, we have $\tilde{r}^2 = O_{\mathrm{p}}(1)$. Hence,

$$\left|\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} - \mathrm{prox}_{\gamma_1 f}\left(\mu_1 \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} + \sigma_1 Z_i\right)\right| \xrightarrow{p} 0,$$

as $n_1 \to \infty$ for each $i \in [n_1]$. By using the fact that $\mathrm{prox}_{\gamma f}(a) = a - f'(\mathrm{prox}_{\gamma f}(a))$ for $a \in \mathbb{R}$, $\gamma > 0$, and $f : \mathbb{R} \to \mathbb{R}$ by the definition of the proximal operator, we obtain

$$\left|\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} + \gamma_1\left(y_i^{(1)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)}\right) - \mu_1 \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} - \sigma_1 Z_i\right| \xrightarrow{p} 0,$$

as $n_1 \to \infty$. Next, we consider to replace $(\mu_1, \sigma_1, \gamma_1)$ with observable adjustments $(\tilde{\mu}, \tilde{\sigma}, \tilde{\gamma})$. Theorem 4.4 in Bellec [10] gives their consistency:

$$\begin{aligned}
&\mathbb{E}\left[\tilde{v}|\tilde{\gamma} - \gamma_1|\right] \le C_1 n^{-1/2}, \\
&\mathbb{E}\left[\tilde{v}^2\tilde{t}^2\tilde{r}^{-2}\left(\left|\tilde{\mu}^2 - \mu_1^2\right| + \left|\tilde{\sigma}^2 - \sigma_1^2\right|\right)\right] \le C_2 n^{-1/2},
\end{aligned}$$

where we define $\tilde{t}^2 = \|(\lambda\boldsymbol{\Sigma}^{-1/2} + \tilde{v}\boldsymbol{\Sigma}^{1/2})\tilde{\boldsymbol{\beta}}\|^2 - \kappa_1\tilde{r}^2$ and C_1, C_2 are positive constants. Proposition 3.1 in Bellec [10] implies that $\tilde{v} \ge 1/(1+\bar{c}) - 4\bar{c}/n_1$ for a constant $\bar{c} > 0$. Also, Theorem 4.4 in Bellec [10] implies that $\tilde{t}^2 \xrightarrow{p} (\boldsymbol{\beta}^\top(\tilde{v}\boldsymbol{\Sigma} + \lambda)\tilde{\boldsymbol{\beta}})^2$. By using these results, we have $\tilde{\gamma} \xrightarrow{p} \gamma_1$, $\tilde{\mu} \xrightarrow{p} \mu_1$, and $\tilde{\sigma}^2 \xrightarrow{p} \sigma_1^2$ as $n_1 \to \infty$ since the sign of μ_1 is specified by an assumption. Then, triangle inequality implies

$$\begin{aligned}
&\left|\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} + \tilde{\gamma}\left(y_i^{(1)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)}\right) - \tilde{\mu}\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} - \tilde{\sigma} Z_i\right| \\
&\le \left|\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)} + \gamma_1\left(y_i^{(1)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)}\right) - \mu_1 \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} - \sigma_1 Z_i\right| \\
&+ \left|(\gamma_1 - \tilde{\gamma})\left(y_i^{(1)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)}\right)\right| + \left|(\mu_1 - \tilde{\mu})\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i^{(1)}\right| + |(\sigma_1 - \tilde{\sigma})Z_i|,
\end{aligned}$$

which converges in probability to zero. □

To prove Theorem 1, we first introduce an approximation $\tilde{g}(\cdot)$ for the link estimator $\widehat{g}(\cdot)$ defined in (2.3).

Lemma 16 *For $i = 1,\dots,n_1$, define $\tilde{W}_i = \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} + \varsigma Z_i$ with $\varsigma = \sigma_1/|\mu_1|$ and Z_i independent of $\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)}$ satisfying $(Z_1,\dots,Z_n) \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$. We also define*

$$\tilde{g}(x) := \frac{\sum_{i=1}^{n_1} y_i^{(1)} \int_{-\infty}^{\infty} \exp\left(t^2\varsigma^2/(2h_n^2) - \mathrm{i}t(x-\tilde{W}_i)/h_n\right) \phi_K(t)dt}{\sum_{i=1}^{n_1} \int_{-\infty}^{\infty} \exp\left(t^2\varsigma^2/(2h_n^2) - \mathrm{i}t(x-\tilde{W}_i)/h_n\right) \phi_K(t)dt}. \tag{E.5}$$

Then, under the setting of Theorem 1, we have, as $n_1 \to \infty$,

$$\sup_{a\le x\le b} |\check{g}(x) - \tilde{g}(x)| = O_{\mathrm{p}}\left(\frac{1}{(\log n_1)^{m/2}}\right).$$

Proof of Lemma 16 We rewrite the kernel function for deconvolution in (2.3) as

$$K_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-\mathrm{i}tx) \frac{\phi_K(t)}{\exp(-t^2\widehat{\varsigma}^2/(2h_n^2))} dt,$$

and also introduce an approximated version of the kernel function as

$$\tilde{K}_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-\mathrm{i}tx) \frac{\phi_K(t)}{\exp(-t^2\varsigma^2/(2h_n^2))} dt.$$

The difference here is that the parameter $\widehat{\varsigma}$ is replaced by ς. We also define $\phi_\varsigma(t) = \exp(-t^2\varsigma^2/2)$.

At first, we have

$$\begin{aligned}
&|\check{g}(x) - \tilde{g}(x)| \\
&= \left| \frac{\sum_{i=1}^{n_1} y_i^{(1)} K_n\left(\frac{W_i - x}{h_n}\right)}{\sum_{i=1}^{n_1} K_n\left(\frac{W_i - x}{h_n}\right)} - \frac{\sum_{i=1}^{n_1} y_i^{(1)} \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right)}{\sum_{i=1}^{n_1} \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right)} \right| \\
&\le C_{1,n}C_{2,n} \left| \frac{1}{n_1 h_n} \sum_{i=1}^{n_1} y_i^{(1)} \left\{ K_n\left(\frac{W_i - x}{h_n}\right) - \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right) \right\} \right| \\
&\quad + C_{1,n}C_{3,n} \frac{1}{n_1 h_n} \left| \sum_{i=1}^{n_1} \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right) - K_n\left(\frac{W_i - x}{h_n}\right) \right| \\
&\le C_{1,n}C_{2,n} \left| \frac{1}{n_1 h_n} \sum_{i=1}^{n_1} y_i^{(1)} \left\{ \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right) - \tilde{K}_n\left(\frac{W_i - x}{h_n}\right) \right\} \right| \qquad \text{(E.6)} \\
&\quad + C_{1,n}C_{2,n} \left| \frac{1}{n_1 h_n} \sum_{i=1}^{n_1} y_i^{(1)} \left\{ \tilde{K}_n\left(\frac{W_i - x}{h_n}\right) - K_n\left(\frac{W_i - x}{h_n}\right) \right\} \right| \qquad \text{(E.7)} \\
&\quad + C_{1,n}C_{3,n} \frac{1}{n_1 h_n} \left| \sum_{i=1}^{n_1} \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right) - \tilde{K}_n\left(\frac{W_i - x}{h_n}\right) \right| \qquad \text{(E.8)}
\end{aligned}$$

$$+C_{1,n}C_{3,n}\frac{1}{n_1h_n}\left|\sum_{i=1}^{n_1}\tilde{K}_n\left(\frac{W_i-x}{h_n}\right)-K_n\left(\frac{W_i-x}{h_n}\right)\right|, \tag{E.9}$$

where we define $C_{1,n}=\left|n^2h_n^2(\sum_{i=1}^{n_1}\tilde{K}_n\left(\frac{\tilde{W}_i-x}{h_n}\right))^{-1}(\sum_{i=1}^{n_1}K_n\left(\frac{W_i-x}{h_n}\right))^{-1}\right|$, $C_{2,n}=\left|\frac{1}{n_1h_n}\sum_{i=1}^{n_1}\tilde{K}_n\left(\frac{\tilde{W}_i-x}{h_n}\right)\right|$, and $C_{3,n}=\left|\frac{1}{n_1h_n}\sum_{i=1}^{n_1}y_i^{(1)}\tilde{K}_n\left(\frac{\tilde{W}_i-x}{h_n}\right)\right|$. Here, $C_{2,n}$ and $C_{3,n}$ converge to upper-bounded non-negative functions of x from the consistency of the deconvoluted kernel density estimator by the i.i.d. assumption of $Z_1,\ldots,Z_n$. We proceed to bound $C_{1,n}$ and each term on the right-hand side. First, we bound (E.9). $(|t|e^{-1}/\sqrt{2})$-Lipschitz continuity of $\phi_\varsigma(t)$ with respect to ς yields

$$\begin{aligned}
&\frac{1}{n_1h_n}\left|\sum_{i=1}^{n_1}\tilde{K}_n\left(\frac{W_i-x}{h_n}\right)-K_n\left(\frac{W_i-x}{h_n}\right)\right|\\
&=\left|\frac{1}{2\pi n_1h_n}\sum_{i=1}^{n_1}\int_{-M_0}^{M_0}\exp\left(-\mathrm{i}t\frac{W_i-x}{h_n}\right)\frac{\phi_K(t)}{\phi_\varsigma(t/h_n)\phi_{\tilde{\varsigma}}(t/h_n)}\left\{\phi_\varsigma(t/h_n)-\phi_{\tilde{\varsigma}}(t/h_n)\right\}dt\right|\\
&\le\frac{1}{\sqrt{2}e\pi h_n^2}|\varsigma-\tilde{\varsigma}|\int_0^{M_0}|t\phi_K(t)|\exp\left(\frac{t^2(\varsigma^2+\tilde{\varsigma}^2)}{2h_n^2}\right)dt.
\end{aligned}$$

Theorem 4.4 in Bellec [10] implies that $|\varsigma-\tilde{\varsigma}|=O_{\mathrm{p}}(n_1^{-1/2})$ since

$$|\varsigma-\tilde{\varsigma}|\,(\varsigma+\tilde{\varsigma})=\left|\varsigma^2-\tilde{\varsigma}^2\right|\le\frac{1}{\tilde{\mu}^2\mu_1^2}\left\{\mu_1^2\left|\tilde{\sigma}^2-\sigma_1^2\right|+\sigma_1^2\left|\mu_1^2-\tilde{\mu}^2\right|\right\}. \tag{E.10}$$

Hence, as we choose $h_n=(c_h\log n_1)^{-1/2}$ such that $M_0^2(\varsigma^2+\tilde{\varsigma}^2)c_h/2+c\le 1/2$ for some $c>0$, we obtain

$$\frac{1}{n_1h_n}\left|\sum_{i=1}^{n_1}\tilde{K}_n\left(\frac{W_i-x}{h_n}\right)-K_n\left(\frac{W_i-x}{h_n}\right)\right|=O_{\mathrm{p}}\left((\log n_1)n_1^{-c}\right).$$

Next, we bound (E.8). For any $x,x'\in\mathbb{R}$, we have

$$\begin{aligned}
|e^{-\mathrm{i}tx}-e^{-\mathrm{i}tx'}|&=\left(\left\{\cos(-tx)-\cos(-tx')\right\}^2+\left\{\sin(-tx)-\sin(-tx')\right\}^2\right)^{1/2}\\
&\le\sqrt{2}t|x-x'|,
\end{aligned}$$

where the last inequality follows from 1-Lipschitz continuity of $\cos(\cdot)$ and $\sin(\cdot)$. Since $\phi_K(\cdot)$ is supported on $[-M_0,M_0]$, we have,

$$\begin{aligned}
&\frac{1}{n_1h_n}\left|\sum_{i=1}^{n_1}\tilde{K}_n\left(\frac{\tilde{W}_i-x}{h_n}\right)-\tilde{K}_n\left(\frac{W_i-x}{h_n}\right)\right|\\
&=\left|\frac{1}{2\pi n_1h_n}\sum_{i=1}^{n_1}\int_{-M_0}^{M_0}\frac{\phi_K(t)}{\phi_\varsigma(t/h_n)}\left\{\exp\left(-\mathrm{i}t\frac{(\tilde{W}_i-x)}{h_n}\right)-\exp\left(-\mathrm{i}t\frac{(W_i-x)}{h_n}\right)\right\}dt\right|
\end{aligned}$$

$$\leq \frac{\sqrt{2}}{\pi n_1 h_n^2} \sum_{i=1}^{n_1} \left|\tilde{W}_i - W_i\right| \int_0^{M_0} |t\phi_K(t)| \exp\left(\frac{t^2\varsigma^2}{2h_n^2}\right) dt.$$

Here, we can use the fact that, by the triangle inequality,

$$|\tilde{W}_i - W_i| \leq |\varsigma - \tilde{\varsigma}|Z_i + |W_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} - \tilde{\varsigma} Z_i| = O_{\mathrm{p}}(n_1^{-1/2}),$$

where the equality follows from (E.10) and (E.4). Thus, since we choose $h_n = (c_h \log n)^{-1/2}$ such that $M_0^2\varsigma^2 c_h/2 + c \leq 1/2$ for some $c > 0$, we have

$$\begin{aligned}
&\frac{1}{n_1 h_n} \left| \sum_{i=1}^{n_1} \tilde{K}_n\left(\frac{\tilde{W}_i - x}{h_n}\right) - \tilde{K}_n\left(\frac{W_i - x}{h_n}\right) \right| \\
&= O_{\mathrm{p}}\left(\frac{1}{n_1^{1/2} h_n^2} \exp\left(\frac{M_0^2\varsigma^2}{2} c_h \log n_1\right)\right) \\
&= O_{\mathrm{p}}((\log n_1) n_1^{-c}).
\end{aligned}$$

This concludes the convergence of $C_{1,n}$ and the term (E.8). Repeating the arguments above for (E.6)-(E.7) implies that $\sup_{a\leq x\leq b} |\widehat{g}(x) - g(x)| = O_p((\log n_1)^{-m/2})$. □

Proof of Theorem 1 Since $\boldsymbol{\beta}^\top \boldsymbol{X}_i^{(1)} \sim \mathcal{N}(0,1)$ by Assumption 1, Lemma 12 implies that, for $\tilde{g}(\cdot)$ defined in (E.5),

$$\sup_{a\leq x\leq b} |\tilde{g}(x) - g(x)| = O_{\mathrm{p}}\left(\frac{1}{(\log n_1)^{m/2}}\right). \tag{E.11}$$

Thus, we obtain

$$\begin{aligned}
\sup_{a\leq x\leq b} |\widehat{g}(x) - g(x)| &\leq \sup_{a\leq x\leq b} |\breve{g}(x) - \tilde{g}(x)| + \sup_{a\leq x\leq b} |\tilde{g}(x) - g(x)| \\
&= O_{\mathrm{p}}\left(\frac{1}{(\log n_1)^{m/2}}\right).
\end{aligned}$$

The last equality follows Lemma 16 and (E.11). Also, the first inequality follows the triangle inequality and a property of each choice of the monotonization operator $\mathscr{R}[\cdot]$. If we select the naive $\mathscr{R}_{\mathrm{naive}}[\cdot]$, we obtain the following for $x \in [a,b]$:

$$\begin{aligned}
|\widehat{g}(x) - g(x)| &= \left| \sup_{x'\in[a,x]} \breve{g}(x') - g(x) \right| \\
&= \left| \sup_{x'\in[a,x]} \breve{g}(x') - \sup_{x'\in[a,x]} g(x') \right| \\
&\leq \sup_{x'\in[a,x]} |\breve{g}(x') - g(x')|,
\end{aligned}$$

by the monotonicity of $g(\cdot)$. If we select the rearrangement operator $\mathscr{R}^a[\cdot]$, Proposition 1 in Chernozhukov et al. [21] yields the same result for $x \in [a,b]$. Thus, whichever monotonization is chosen, we obtain the statement. □

E.3. *Proof of Theorem 2*

Lemma 17 *Let Assumption 1-3 hold. Define*

$$\mu_n = \boldsymbol{\beta}^\top \widehat{\boldsymbol{\beta}}(\widehat{g}), \quad \sigma_n^2 = \|\boldsymbol{P}_{\boldsymbol{\beta}}^{\perp} \widehat{\boldsymbol{\beta}}(\widehat{g})\|^2, \tag{E.12}$$

where $\boldsymbol{P}_{\boldsymbol{\beta}}^{\perp} = \boldsymbol{I}_p - \boldsymbol{\beta}\boldsymbol{\beta}^\top$. *Then, we have*

$$|\widehat{\mu}(\widehat{g}) - \mu_n| \xrightarrow{p} 0, \quad \text{and} \quad \left|\widehat{\sigma}^2(\widehat{g}) - \sigma_n^2\right| \xrightarrow{p} 0.$$

Proof of Lemma 17 Theorem 4.4 in Bellec [10] implies that as $n_2 \to \infty$, we have

$$\widehat{v}_\lambda^2 \widehat{t}^2 \dot{r}^{-4} \left|\widehat{\mu}^2(\widehat{g}) - \mu_n^2\right| \xrightarrow{p} 0, \quad \widehat{v}_\lambda^2 \widehat{t}^2 \dot{r}^{-4} \left|\widehat{\sigma}^2(\widehat{g}) - \sigma_n^2\right| \xrightarrow{p} 0,$$

with $\widehat{t}^2 = (\widehat{v}_\lambda + \lambda)^2 \|\widehat{\boldsymbol{\beta}}(\widehat{g})\|^2 - \kappa_2 \dot{r}^2$ and $\dot{r}^2 = n_2^{-1} \|\boldsymbol{y}^{(2)} - \widehat{g}(\boldsymbol{X}^{(2)} \widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2$. Recall that $\widehat{v}_\lambda$ and $\boldsymbol{D}$ are defined in Section 2.4. Thus, it is sufficient to show that $\widehat{v}_\lambda^2, \widehat{t}^2$, and $\dot{r}^{-4}$ are asymptotically lower bounded away from zero. First, the fact that $\mathrm{tr}(\boldsymbol{D}) \geq n_2 c_g^{-1} > 0$ holds by Assumption 3 and Proposition 3.1 in Bellec [10] imply that there exists a constant $\widehat{c} > 0$ such that $\widehat{v}_\lambda \geq c_g^{-1}/(1+\widehat{c}) - 4\widehat{c}/n_2$ holds. Next, since ridge penalized regression estimators satisfy $\|\widehat{\boldsymbol{\beta}}(\widehat{g})\| \leq C'_\lambda$ with a constant $C'_\lambda > 0$ depending on the regularization parameter $\lambda > 0$, we have $\dot{r}^2 = O_{\mathrm{p}}(1)$. Also, Theorem 4.4 in Bellec [10] implies that $\widehat{t}^2 \xrightarrow{p} ((\widehat{v}_\lambda + \lambda)\boldsymbol{\beta}^\top \widehat{\boldsymbol{\beta}}(\widehat{g}))^2$. Thus, we have $|\widehat{\mu}(\widehat{g}) - \mu_n| \xrightarrow{p} 0$ and $\left|\widehat{\sigma}^2(\widehat{g}) - \sigma_n^2\right| \xrightarrow{p} 0$ as $n_2 \to \infty$ since the sign of μ_n is specified by an assumption. □

Proof of Theorem 2 We use the notations defined in (E.12). First, we can apply Theorem 6 and obtain

$$\frac{\sqrt{p}(\widehat{\boldsymbol{\beta}}_j(\widehat{g}) - \mu_n \boldsymbol{\beta}_j)}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1).$$

This is because we can skip Step 1 in the proof of Theorem 6 by $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and repeat Steps 2–3 since $J(\tilde{\boldsymbol{U}}\boldsymbol{b}) = J(\boldsymbol{b})$ for any orthogonal matrices $\tilde{\boldsymbol{U}} \in \mathbb{R}^{p \times p}$. Hence, we have

$$\sqrt{p}\frac{\widehat{\beta}_j(\widehat{g}) - \widehat{\mu}(\widehat{g})\beta_j}{\widehat{\sigma}(\widehat{g})} = \sqrt{p}\frac{\widehat{\beta}_j(\widehat{g}) - \mu_n \beta_j}{\sigma_n}\frac{\sigma_n}{\widehat{\sigma}(\widehat{g})} + \sqrt{p}\frac{(\mu_n - \widehat{\mu}(\widehat{g}))\beta_j}{\widehat{\sigma}(\widehat{g})} \xrightarrow{d} \mathcal{N}(0,1),$$

where the convergence follows from the facts that $\widehat{\mu}(\widehat{g}) \xrightarrow{p} \mu_n$ and $\widehat{\sigma}^2(\widehat{g}) \xrightarrow{p} \sigma_n^2$ by Lemma 17. This concludes the proof of (3.3).

Next, we consider an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{p \times p}$ with the first row $\boldsymbol{U}_1 = \boldsymbol{v}^\top$. Since $\boldsymbol{U}\widehat{\boldsymbol{\beta}}(\widehat{g})$ is the estimator given by (2.5) with covariates $\boldsymbol{U}\boldsymbol{X}_i^{(2)}$ and the true coefficient vector $\boldsymbol{U}\boldsymbol{\beta}$, applying (3.4) to this with $j = 1$ yields that, for any sequence of non-random vectors $\boldsymbol{v}_n$ such that $\|\boldsymbol{v}_n\| = 1$ and $\sqrt{p}\tau(\boldsymbol{v}_n)\boldsymbol{v}_n^\top \boldsymbol{\beta} = O(1)$,

$$\frac{\sqrt{p}\boldsymbol{v}_n^\top(\widehat{\boldsymbol{\beta}}(\widehat{g}) - \widehat{\mu}(\widehat{g})\boldsymbol{\beta})}{\widehat{\sigma}(\widehat{g})/\tau(\boldsymbol{v}_n)} \xrightarrow{d} \mathcal{N}(0,1), \tag{E.13}$$

where $\tau^2(\boldsymbol{v}_n) = (\boldsymbol{v}_n^\top \boldsymbol{\Theta} \boldsymbol{v}_n)^{-1}$. Finally, (3.5) follows from (E.13) and the Cramér-Wold device. □

Remark 1 *Under the Assumption E in [10], we obtain the explicit rate of convergence* $|\mu_n - \widehat{\mu}(\widehat{g})| = O_p(n^{-1/2})$. *Then, we can improve the condition* $\sqrt{p}\beta_j = O(1)$ *to* $\beta_j = o(1)$ *since*

$$\sqrt{p}\frac{(\mu_n - \widehat{\mu}(\widehat{g}))\beta_j}{\widehat{\sigma}(\widehat{g})} = O_p(\beta_j),$$

under the Assumption E in [10].

E.4. *Proof of Theorem 4*

First, we define the notations used in the proof. We consider an invertible matrix $\boldsymbol{L} \in \mathbb{R}^{p\times p}$ satisfying $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$. Define, for each $i \in \{1,\ldots,n\}$,

$$\tilde{\boldsymbol{X}}_i = \boldsymbol{L}^{-1}\boldsymbol{X}_i^{(2)}, \quad \boldsymbol{\theta} = \boldsymbol{L}^\top \boldsymbol{\beta}, \quad \widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}(\widehat{g}) = \boldsymbol{L}^\top \widehat{\boldsymbol{\beta}}(\widehat{g}). \tag{E.14}$$

Lemma 18 *Let Assumption 1-3(1) hold. Using the notations* (E.14)*, define*

$$\mu_0 = \boldsymbol{\theta}^\top \widehat{\boldsymbol{\theta}}, \quad \sigma_0^2 = \|\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp}\widehat{\boldsymbol{\theta}}\|^2,$$

where $\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp} = \boldsymbol{I}_p - \boldsymbol{\theta}\boldsymbol{\theta}^\top$. *Then, we have*

$$|\widehat{\mu}_0(\widehat{g}) - \mu_0| \xrightarrow{p} 0, \quad \text{and} \quad \left|\widehat{\sigma}_0^2(\widehat{g}) - \sigma_0^2\right| \xrightarrow{p} 0.$$

Proof of Lemma 18 Theorem 4.4 in Bellec [10] implies that as $n_2 \to \infty$, we have

$$\widehat{v}_0^2 \widehat{t}_0^2 \dot{r}_0^{-4} |\widehat{\mu}_0(\widehat{g}) - \mu_0| \xrightarrow{p} 0, \quad \widehat{v}_0^2 \widehat{t}_0^2 \dot{r}_0^{-4} \left|\widehat{\sigma}_0^2(\widehat{g}) - \sigma_0^2\right| \xrightarrow{p} 0,$$

with $\widehat{t}_0^2 = n_2^{-1}\|\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\widehat{g})\|^2 \widehat{v}_0^2 - \kappa_2(1-\kappa_2)\dot{r}_0^2$ and $\dot{r}_0^2 = n_2^{-1}\|\boldsymbol{y}^{(2)} - \widehat{g}(\boldsymbol{X}^{(2)}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2$. Recall that $\widehat{v}_0$ is obtain by the definition of $\widehat{v}_\lambda$ in Section 2.4 and setting $\lambda = 0$. Thus, it is sufficient to show that $\widehat{v}_0^2, \widehat{t}_0^2$, and $\dot{r}_0^{-4}$ are asymptotically lower bounded away from zero. First, $\mathrm{tr}(\boldsymbol{D}) \geq n_2 c_g^{-1} > 0$ by Assumption 3 and Proposition 3.1 in Bellec [10] imply that there exists a constant $\widehat{c}' > 0$ such that $\widehat{v}_0 \geq c_g^{-1}/(1+\widehat{c}') - 4\widehat{c}'/n_2$. Next, we assume that $\|\widehat{\boldsymbol{\beta}}(\widehat{g})\| \leq C$ with probability approaching one, we have $\dot{r}_0^2 = O_{\mathrm{p}}(1)$. Also, Theorem 4.4 in Bellec [10] implies that $\widehat{t}_0^2 \xrightarrow{p} \widehat{v}_0\mu_0$. Thus, we have $|\widehat{\mu}_0(\widehat{g}) - \mu_0| \xrightarrow{p} 0$ and $\left|\widehat{\sigma}_0^2(\widehat{g}) - \sigma_0^2\right| \xrightarrow{p} 0$ as $n_2 \to \infty$ since the sign of μ_0 is specified by an assumption. □

Proof of Theorem 4 At first, the first step of the proof of Theorem 6 implies that, for any coordinate $j = 1,\ldots,p$,

$$\tau_j \frac{\widehat{\beta}_j - \widehat{\mu}(\widehat{g})\beta_j}{\widehat{\sigma}(\widehat{g})} = \frac{\widehat{\theta}_j - \widehat{\mu}(\widehat{g})\theta_j}{\widehat{\sigma}(\widehat{g})},$$

where $\tau_j^{-2} = (\boldsymbol{\Sigma}^{-1})_{jj}$. Here, $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ are defined in (E.14). Thus, we consider $\widehat{\boldsymbol{\theta}}$ instead of $\widehat{\boldsymbol{\beta}}(\widehat{g})$. We have

$$\sqrt{p}\frac{\widehat{\theta}_j - \widehat{\mu}(\widehat{g})\theta_j}{\widehat{\sigma}(\widehat{g})} = \sqrt{p}\frac{\widehat{\theta}_j - \mu_0\theta_j}{\sigma_0}\frac{\sigma_n}{\widehat{\sigma}(\widehat{g})} + \sqrt{p}\frac{(\mu_0 - \widehat{\mu}(\widehat{g}))\theta_j}{\widehat{\sigma}(\widehat{g})}.$$

Thus, the facts that $\widehat{\mu}_0(\widehat{g}) \xrightarrow{p} \mu_0$ and $\widehat{\sigma}_0^2(\widehat{g}) \xrightarrow{p} \sigma_0^2$ by Lemma 18 conclude the proof of (3.4). The rest of the proof follows from repeating the arguments in the proof of Theorem 2. □

E.5. *Proof of Theorem 9*

Lemma 19 *Let $c_g^{-1} \le g'(\cdot)$ hold. Consider censoring of $\widehat{\boldsymbol{\beta}}(\widehat{g})^\top \boldsymbol{X}_i^{(2)}$ and $\widehat{\boldsymbol{\beta}}(g)^\top \boldsymbol{X}_i^{(2)}$ for all $i \in [n_2]$ in $[a,b]$. Under the setting of Lemma 12 with $k = 3$, we have*

$$\max_{i=1,\ldots,n_2} \left| \widehat{\boldsymbol{\beta}}(g)^\top \boldsymbol{X}_i^{(2)} - \widehat{\boldsymbol{\beta}}(\widehat{g})^\top \boldsymbol{X}_i^{(2)} \right| \xrightarrow{p} 0.$$

Proof of Lemma 19 We can assume $\boldsymbol{X}_i^{(2)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ for each $i = 1, \ldots, n_2$ without loss of generality by the first step of the proof of Theorem 6. In this proof, we omit (2) on $\boldsymbol{X}^{(2)}$ for simplicity of the notation. To begin with, we write the KKT condition of the estimation:

$$f(\widehat{\boldsymbol{\beta}}(g), g) = \boldsymbol{0},$$

where we define $f(\boldsymbol{b}, g) = n_2^{-1/2}\boldsymbol{X}^\top(g(\boldsymbol{X}\boldsymbol{b}) - \boldsymbol{y})$. We write $f_j(\boldsymbol{b}, g) = n_2^{-1/2}\boldsymbol{X}_{\cdot j}^\top(g(\boldsymbol{X}\boldsymbol{b}) - \boldsymbol{y})$ for $j = 1, \ldots, p$. Since $\partial/(\partial b_j) f_j(\boldsymbol{b}, g) = n_2^{-1/2}\boldsymbol{X}_{\cdot j}^\top \boldsymbol{D}(\boldsymbol{b})\boldsymbol{X}_{\cdot j}$ with $\boldsymbol{D}(\boldsymbol{b}) = \mathrm{diag}(g'(\boldsymbol{X}\boldsymbol{b}))$, by the mean value theorem, there exists a constant $c \in [0,1]$ such that $\bar{\boldsymbol{b}} = c\widehat{\boldsymbol{\beta}}(g) + (1-c)\widehat{\boldsymbol{\beta}}(\widehat{g})$ satisfies

$$\frac{f_j(\widehat{\boldsymbol{\beta}}(g), g) - f_j(\widehat{\boldsymbol{\beta}}(\widehat{g}), g)}{\widehat{\beta}_j(g) - \widehat{\beta}_j(\widehat{g})} = \left(\frac{1}{\sqrt{n_2}} \boldsymbol{X}_{\cdot j}^\top \boldsymbol{D}(\bar{\boldsymbol{b}}) \boldsymbol{X}_{\cdot j} \right) > 0.$$

Define $R_k := \widehat{g}(\boldsymbol{X}_k^\top \widehat{\boldsymbol{\beta}}(\widehat{g})) - g(\boldsymbol{X}_k^\top \widehat{\boldsymbol{\beta}}(\widehat{g}))$. We have

$$\begin{aligned}
&\sqrt{n_2}\left(\widehat{\beta}_j(g) - \widehat{\beta}_j(\widehat{g})\right) \\
&= \left(n_2^{-1}\boldsymbol{X}_{\cdot j}^\top \boldsymbol{D}(\bar{\boldsymbol{b}})\boldsymbol{X}_{\cdot j}\right)^{-1} \left(f_j(\widehat{\beta}_j(g), g) - f_j(\widehat{\beta}_j(\widehat{g}), g)\right) \\
&= \left(n_2^{-1}\boldsymbol{X}_{\cdot j}^\top \boldsymbol{D}(\bar{\boldsymbol{b}})\boldsymbol{X}_{\cdot j}\right)^{-1} \left\{ f_j(\widehat{\beta}_j(g), g) + \left(f_j(\widehat{\beta}_j(\widehat{g}), \widehat{g}) - f_j(\widehat{\beta}_j(\widehat{g}), g)\right) - f_j(\widehat{\beta}_j(\widehat{g}), \widehat{g}) \right\} \\
&= \left(n_2^{-1}\boldsymbol{X}_{\cdot j}^\top \boldsymbol{D}(\bar{\boldsymbol{b}})\boldsymbol{X}_{\cdot j}\right)^{-1} \left(\frac{1}{\sqrt{n_2}} \sum_{k=1}^{n_2} X_{kj} R_k \right),
\end{aligned}$$

where the second equality follows from the first-order conditions. In sequel, for simplicity, we consider the leave-one-out estimator $\widehat{\boldsymbol{\beta}}_{-i}$ and $\widehat{g}_{-i}$ constructed by the observations without the i-th sample. Define

$$\tilde{R}_k := (\log n_2)\left(\widehat{g}_{-1}(\boldsymbol{X}_k^\top \widehat{\boldsymbol{\beta}}_{-1}(\widehat{g}_{-1})) - g(\boldsymbol{X}_k^\top \widehat{\boldsymbol{\beta}}_{-1}(\widehat{g}_{-1}))\right),$$

and

$$T_j := \left(n_2^{-1} \boldsymbol{X}_{-1,j}^\top \boldsymbol{D}_{-1}(\bar{\boldsymbol{b}}) \boldsymbol{X}_{-1,j} \right)^{-1},$$

where we define the leave-one-out design terms as $\boldsymbol{X}_{-1,j} := (X_{2j}, \ldots, X_{n_2 j})^\top \in \mathbb{R}^{n_2-1}$, and $\boldsymbol{D}_{-1}(\bar{\boldsymbol{b}}) := \mathrm{diag}(g'_{-1}(\boldsymbol{X}_2^\top \bar{\boldsymbol{b}}), \ldots, g'_{-1}(\boldsymbol{X}_{n_2}^\top \bar{\boldsymbol{b}})) \in \mathbb{R}^{(n_2-1)\times(n_2-1)}$. We obtain

$$\begin{aligned}
\left| \boldsymbol{X}_1^\top \widehat{\boldsymbol{\beta}}_{-1}(g) - \boldsymbol{X}_1^\top \widehat{\boldsymbol{\beta}}_{-1}(\widehat{g}_{-1}) \right| &= \left| \sum_{j=1}^{p} X_{1j} \left(\widehat{\beta}_{-1,j}(g) - \widehat{\beta}_{-1,j}(\widehat{g}_{-1}) \right) \right| \\
&\leq \left| \frac{1}{n_2 \log n_2} \sum_{j=1}^{p} X_{1j} T_j \sum_{k=2}^{n_2} X_{kj} \tilde{R}_k \right|.
\end{aligned}$$

Here, define a filtration $\mathscr{F}_k = \sigma(\{\widehat{g}_{-1}, \widehat{\boldsymbol{\beta}}_{-1}(\widehat{g}_{-1}), T_1, \ldots, T_p, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{k+1}\})$ with an initialization $\mathscr{F}_0 = \sigma(\{\widehat{g}_{-1}, \widehat{\boldsymbol{\beta}}_{-1}(\widehat{g}_{-1}), T_1, \ldots, T_p\})$. Define a random variable $\tilde{S}_k = n_2^{-1} \sum_{j=1}^{p} T_j X_{1j} X_{kj}$. Then, $\tilde{R}_k \tilde{S}_k$ is a martingale difference sequence since $\mathbb{E}[\tilde{R}_k \tilde{S}_k \mid \mathscr{F}_{k-1}] = 0$ and $\mathbb{E}|\tilde{R}_k \tilde{S}_k| \leq \mathbb{E}[\tilde{R}_k^2]^{1/2} \mathbb{E}[\tilde{S}_k^2]^{1/2} < \infty$. This follows from the fact that

$$\begin{aligned}
\mathbb{E}\left[\tilde{S}_k^2\right] &= \mathbb{E}\left[\left(n_2^{-1} \sum_{j=1}^{p} T_j X_{1j} X_{kj} \right)^2 \right] \\
&= n_2^{-2} \sum_{j=1}^{p} \mathbb{E}\left[T_j^2 X_{1j}^2 X_{kj}^2 \right] + 2 \sum_{j<j'} \mathbb{E}\left[T_j X_{1j} X_{kj} T_{j'} X_{1j'} X_{kj'} \right] \\
&= n_2^{-2} \sum_{j=1}^{p} \mathbb{E}[T_j^2 X_{kj}^2] \leq n_2^{-2} \sum_{j=1}^{p} \mathbb{E}[T_j^4]^{1/2} \mathbb{E}[X_{kj}^4]^{1/2}.
\end{aligned}$$

The last inequality follows from the Cauchy-Schwartz inequality. Since X_{kj} is the standard Gaussian, $\mathbb{E}[X_{kj}^4] = 3$ holds. Also, we have $0 < T_j \leq c_g (n_2^{-1} \sum_{l=2}^{n_2} X_{lj}^2)^{-1}$, where $((n_2-1)^{-1} \sum_{l=2}^{n_2} X_{lj}^2)^{-1}$ follows the inverse gamma distribution with parameters $((n_2-1)/2, 2/(n_2-1))$ and the bounded fourth moment $(2/(n_2-1))^4 \Gamma(2/(n_2-1)-4)/\Gamma(2/(n_2-1))$.

Let $(\log n_2)^{-m/2} \tilde{R} := \sup_{a \leq x \leq b} |\widehat{g}_{-1}(x) - g(x)|$. Note that, since $\tilde{R} = O_{\mathrm{p}}(1)$ by Lemma 12, for any $\varepsilon_1 > 0$, there exists $\bar{c} > 0$ such that we have $\mathbb{P}(R_k > \bar{c}) \leq \varepsilon_1$. Also, note that censoring does not affect this fact since $\widehat{g}(\cdot)$ is given independent of $\boldsymbol{X}$. Hence, we obtain, for $\bar{c}$ and any $t_n > 0$ satisfying $t_n = o(\sqrt{n_2})$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\sqrt{n_2}} \max_{2 \leq k \leq n_2} \left| \tilde{R}_k \sum_{j=1}^{p} T_j X_{1j} X_{kj} \right| > t_n \;\middle|\; |\tilde{R}| \leq \bar{c} \right) \\
&\leq \mathbb{P}\left(\frac{\bar{c}}{\sqrt{n_2}} \max_{2 \leq k \leq n_2} \left| \sum_{j=1}^{p} T_j X_{1j} X_{kj} \right| > t_n \right) \\
&\leq \mathbb{P}\left(\frac{\bar{c}}{\sqrt{n_2}} \max_{2 \leq k \leq n_2} \left| \sum_{j=1}^{p} X_{1j} X_{kj} T_j \right| > t_n \;\middle|\; \max_{1 \leq j \leq p} |T_j| \leq u \right) + \mathbb{P}\left(\max_{1 \leq j \leq p} |T_j| > u \right)
\end{aligned}$$

$$\leq 2n_2 \exp\left(-\frac{ct_n^2}{\bar{c}^2K^2}\right) + \mathbb{P}\left(\max_{1\leq j\leq p}|T_j| > u\right). \tag{E.15}$$

with some $c>0$ depending on u, where the last inequality follows from the union bound and Bernstein's inequality. Here, K is the sub-exponential norm of $uX_{11}X_{21}$. Since we have $T_j \leq c_g(n_2{}^{-1}\sum_{l=2}^{n_2} X_{lj}^2)$, it holds that

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j\leq p}|T_j| > u\right) &\leq \mathbb{P}\left(\max_{1\leq j\leq p}\frac{1}{n_2}\sum_{l=2}^{n_2} X_{lj}^2 - 1 > u_*\right)\\ &\leq p\exp\left(-c\left(\frac{u_*^2}{K^2}\wedge\frac{u_*}{K}\right)n_2\right),\end{aligned} \tag{E.16}$$

where $u_* = c_g/u - 1$. Using the bounds, Azuma-Hoeffding's inequality yields, for any $x, u_n > 0$ and $t_n > 0$ satisfying $t_n = o(\sqrt{n_2})$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{(\log n_2)^{m/2}}\max_{1\leq i\leq n_2}\left|\frac{1}{n_2}\sum_{k\neq i}^{n_2}\tilde{R}_k\sum_{j=1}^{n_2}X_{ij}X_{kj}\right| > x\right)\\
&\leq \mathbb{P}\left(\frac{1}{(\log n_2)^{m/2}}\max_{1\leq i\leq n_2}\left|\frac{1}{n_2}\sum_{k\neq i}^{n_2}\tilde{R}_k\sum_{j=1}^{n_2}X_{ij}X_{kj}\right| > x \;\middle|\; |\tilde{R}|\leq\bar{c}\right) + \varepsilon_1\\
&\leq n_2\mathbb{P}\left(\frac{1}{(\log n_2)^{m/2}}\left|\frac{1}{n_2}\sum_{k=2}^{n_2}\tilde{R}_k\sum_{j=1}^{n_2}X_{1j}X_{kj}\right| > x \;\middle|\; |\tilde{R}|\leq\bar{c}\right) + \varepsilon_1\\
&\leq n_2\mathbb{P}\left(\frac{1}{(\log n_2)^{m/2}}\left|\frac{1}{n_2}\sum_{k=2}^{n_2}\tilde{R}_k\sum_{j=1}^{n_2}X_{1j}X_{kj}\right| > x \;\middle|\; \frac{1}{n}\max_{2\leq k\leq n_2}\left|\tilde{R}_k\sum_{j=1}^{p}X_{1j}X_{kj}\right| \leq \frac{t_n}{\sqrt{n_2}}, |\tilde{R}|\leq\bar{c}\right)\\
&\quad + n_2\mathbb{P}\left(\frac{1}{n_2}\max_{2\leq k\leq n_2}\left|\tilde{R}_k\sum_{j=1}^{p}X_{1j}X_{kj}\right| > \frac{t_n}{\sqrt{n_2}} \;\middle|\; |\tilde{R}|\leq\bar{c}\right) + \varepsilon_1\\
&\leq 2n_2\exp\left(-\frac{x^2(\log n_2)^m}{2t_n^2}\right) + 2n_2^2\exp\left(-\frac{ct_n^2}{\bar{c}^2K^2}\right)\\
&\quad + n_2^2\exp\left(-c\left(\frac{u_*^2}{K^2}\wedge\frac{u_*}{K}\right)n_2\right) + \varepsilon_1,\end{aligned}$$

where the last inequality follows from (E.15) and (E.16). Thus, one can choose, for instance, $m=3$, $t_n = (\log n_2)^{3/5}$, and $\bar{c} = \log\log n_2$ so that we have $\frac{1}{(\log n_2)^{m/2}}\max_{1\leq i\leq n_2}|\frac{1}{n_2}\sum_{k\neq i}^{n_2}\tilde{R}_k\sum_{j=1}^{n_2}X_{ij}X_{kj}| = o_{\mathrm{p}}(1)$ and $\varepsilon_1\to 0$. □

Proof of Theorem 9 In this proof, we omit the superscript (2) on $\boldsymbol{X}^{(2)}$ and $\boldsymbol{y}^{(2)}$ for simplicity of the notation. We firstly rewrite the inferential parameters defined in Section A as

$$\widehat{\mu}_{0c}^2(g) = \frac{\|\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g))\|^2}{n_2} - \kappa_2(1-\kappa_2)\widehat{\sigma}^2(g), \quad \widehat{\sigma}_{0c}^2(g) = \frac{n_2^{-1}\|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g))\|^2}{\left(n_2^{-1}\mathrm{tr}(\boldsymbol{V}(g))\right)^2},$$

where $\boldsymbol{V}_c(g) = \boldsymbol{D}_c(g) - \boldsymbol{D}_c(g)\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{D}_c(g)$. Since we have

$$
\begin{aligned}
&\left|\widehat{\sigma}^2_{0c}(\widehat{g}) - \widehat{\sigma}^2_{0c}(g)\right| \\
&\leq \frac{1}{\left(n_2^{-1}\mathrm{tr}(\boldsymbol{V}_c(\widehat{g}))\right)^2 \left(n_2^{-1}\mathrm{tr}(\boldsymbol{V}_c(g))\right)^2} \\
&\quad \times \left\{ \frac{\|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g))\|^2}{n_2} \left| \left(\frac{\mathrm{tr}(\boldsymbol{V}_c(\widehat{g}))}{n_2}\right)^2 - \left(\frac{\mathrm{tr}(\boldsymbol{V}_c(g))}{n_2}\right)^2 \right| \right. \\
&\quad \left. + \left(\frac{\mathrm{tr}(\boldsymbol{V}_c(g))}{n_2}\right)^2 \left| \frac{\|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g))\|^2}{n_2} - \frac{\|\boldsymbol{y} - g(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2}{n_2} \right| \right\},
\end{aligned}
$$

It is sufficient to show the following properties:

$$
n_2^{-1} \left| \|\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2 - \|\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g))\|^2 \right| = o_{\mathrm{p}}(1), \tag{E.17}
$$

$$
n_2^{-1} \left| \|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g})))\|^2 - \|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g)))\|^2 \right| = o_{\mathrm{p}}(1), \tag{E.18}
$$

$$
n_2^{-1} \left| \mathrm{tr}(\boldsymbol{V}_c(\widehat{g})) - \mathrm{tr}(\boldsymbol{V}_c(g)) \right| = o_{\mathrm{p}}(1). \tag{E.19}
$$

For (E.17), immediately we have

$$
\begin{aligned}
&n_2^{-1} \left| \|\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2 - \|\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g))\|^2 \right| \\
&= \left| n_2^{-1} \sum_{i=1}^{n_2} \left(\iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g})) - \iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(g)) \right) \left(\iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g})) + \iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(g)) \right) \right|.
\end{aligned}
$$

Since $\max_{i=1,\ldots,n_2} |\iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(g)) - \iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g}))| \xrightarrow{p} 0$ as $n_2 \to \infty$ by Lemma 19, this term converges in probability to zero.

Next, for (E.18), since we have

$$
\begin{aligned}
&n_2^{-1} \left| \|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g})))\|^2 - \|\boldsymbol{y} - g(\iota(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(g)))\|^2 \right| \\
&= \left| n_2^{-1} \sum_{i=1}^{n_2} \left(\widehat{g}(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) - g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(g))) \right) \left(2y_i - \widehat{g}(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) - g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(g))) \right) \right|,
\end{aligned}
$$

we should bound $\widehat{g}(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) - g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(g)))$. Indeed, using the triangle inequality reveals

$$
\begin{aligned}
&\left| \widehat{g}(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) - g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(g))) \right| \\
&\leq \left| g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) - g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(g))) \right| + \left| \widehat{g}(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) - g(\iota(\boldsymbol{X}_i^\top \boldsymbol{\beta}(\widehat{g}))) \right|.
\end{aligned}
$$

The first term on the right-hand side is $o_{\mathrm{p}}(1)$ by the Lipschitz continuity of $g(\cdot)$ and Lemma 19. Also, the second term is upper bounded by $\sup_x |\widehat{g}(x) - g(x)|$, and is $o_{\mathrm{p}}(1)$ by Lemma 12.

To achieve (E.19), we first have

$$\begin{aligned}
&n_2^{-1}\left|\mathrm{tr}(\boldsymbol{V}_c(\widehat{g}))-\mathrm{tr}(\boldsymbol{V}_c(g))\right| \\
&\le n_2^{-1}\left|\mathrm{tr}\left(\boldsymbol{D}_c(\widehat{g})-\boldsymbol{D}_c(g)\right)\right| \\
&\quad + n_2^{-1}\left|\mathrm{tr}(\boldsymbol{D}_c(\widehat{g})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\right. \\
&\qquad \left. -\boldsymbol{D}_c(g)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)\right)|.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&n_2^{-1}\left|\mathrm{tr}\left(\boldsymbol{D}_c(\widehat{g})-\boldsymbol{D}_c(g)\right)\right| \\
&\le n_2^{-1}\sum_{i=1}^{n_2}\left|\widehat{g}'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g})))-g(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g)))\right| \\
&\le \sup_{a\le x\le b}|\widehat{g}'(x)-g'(x)| + Bn_2^{-1}\sum_{i=1}^{n_2}\left|\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))-\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g))\right| \\
&= o_{\mathrm{p}}(1),
\end{aligned}$$

by Lemma 13 and Lemma 19. For the second term, the triangle inequality yields

$$\begin{aligned}
&n_2^{-1}|\mathrm{tr}(\boldsymbol{D}_c(\widehat{g})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g}) \\
&\quad -\boldsymbol{D}_c(g)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g))| \\
&\le n_2^{-1}\left|\mathrm{tr}\left(\{\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)\right)\right| && \text{(E.20)} \\
&\quad + n_2^{-1}\left|\mathrm{tr}\left(\boldsymbol{D}_c(\widehat{g})\boldsymbol{X}\left\{(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}-(\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\boldsymbol{X})^{-1}\right\}\boldsymbol{X}^\top\boldsymbol{D}_c(g)\right)\right| && \text{(E.21)} \\
&\quad + n_2^{-1}\left|\mathrm{tr}\left(\boldsymbol{D}_c(\widehat{g})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\boldsymbol{X})^{-1}\boldsymbol{X}^\top\{\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\}\right)\right|. && \text{(E.22)}
\end{aligned}$$

Using the Cauchy-Schwartz inequality, (E.20) is bounded by

$$n_2^{-1}\left\|\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\right\|_F\left\|\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)\right\|_F.$$

Here, we have

$$\begin{aligned}
&n_2^{-1/2}\left\|\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\right\|_F \\
&\le\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\left\{g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g)))-\widehat{g}'(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))\right\}^2\right)^{1/2} \\
&\le\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\left[2\left\{g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g)))-g'(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))\right\}^2+2\left\{g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g})))-\widehat{g}'(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))\right\}^2\right]\right)^{1/2} \\
&\le\left(\frac{2}{n_2}\sum_{i=1}^{n_2}\left\{g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g)))-g'(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))\right\}^2\right)^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{2}{n_2}\sum_{i=1}^{n_2}\left\{g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g})))-\widehat{g}'(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))\right\}^2\right)^{1/2}\\
&\le 2B\max_{i=1,\dots,n_2}\left|\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g))-\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g}))\right|+2\sup_{a\le x\le b}|\widehat{g}'(x)-g'(x)|\\
&=o_{\mathrm{p}}(1).
\end{aligned}
$$

by Lemma 13 and Lemma 19. Also, we have

$$
\begin{aligned}
&n_2^{-1/2}\left\|\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)\right\|_F\\
&=n_2^{-1/2}\left\|\boldsymbol{D}_c(g)^{-1/2}\boldsymbol{D}_c(g)^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)^{1/2}\boldsymbol{D}_c(g)^{1/2}\right\|_F\\
&\le n_2^{-1/2}\left\|\boldsymbol{D}_c(g)^{-1/2}\right\|_{\mathrm{op}}\left\|\boldsymbol{D}_c(g)^{1/2}\right\|_{\mathrm{op}}\left\|\boldsymbol{D}_c(g)^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)^{1/2}\right\|_F\\
&=n_2^{-1/2}\left\|\boldsymbol{D}_c(g)^{-1/2}\right\|_{\mathrm{op}}\left\|\boldsymbol{D}_c(g)^{1/2}\right\|_{\mathrm{op}}\sqrt{\mathrm{tr}(\boldsymbol{I}_{n_2})}\\
&=\left\|\boldsymbol{D}_c(g)^{-1/2}\right\|_{\mathrm{op}}\left\|\boldsymbol{D}_c(g)^{1/2}\right\|_{\mathrm{op}}.
\end{aligned}
$$

Since $\left\|\boldsymbol{D}_c(g)^{1/2}\right\|_{\mathrm{op}}\le\sup_x g'(x)^{1/2}$ and $\left\|\boldsymbol{D}_c(g)^{-1/2}\right\|_{\mathrm{op}}\le(\inf_x g'(x))^{-1/2}$ are constants by an assumption of $g(\cdot)$, we conclude that (E.20) is $o_{\mathrm{p}}(1)$. (E.22) is also shown to be $o_{\mathrm{p}}(1)$ in a similar manner. Since $\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}=-\boldsymbol{A}^{-1}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{B}^{-1}$ for two invertible matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, (E.21) can be rewritten as

$$
n_2^{-1}\left|\mathrm{tr}\left(\boldsymbol{D}_c(\widehat{g})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\boldsymbol{X})^{-1}\boldsymbol{X}^\top\left\{\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\right\}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)\right)\right|,
$$

and a similar technique used above provides the upper bound,

$$
\begin{aligned}
&n_2^{-1}\left\|\boldsymbol{D}_c(\widehat{g})\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(\widehat{g})^{-1}\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)^{-1}\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\right\|_{\mathrm{op}}\\
&\quad\times\left\|\boldsymbol{D}_c(\widehat{g})^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(\widehat{g})^{1/2}\right\|_F\\
&\quad\times\left\|\boldsymbol{D}_c(g)^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}_c(g)\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}_c(g)^{1/2}\right\|_F\\
&=\left\|\boldsymbol{D}_c(\widehat{g})\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(\widehat{g})^{-1}\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)^{-1}\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\right\|_{\mathrm{op}}.
\end{aligned}
$$

Here, $\left\|\boldsymbol{D}_c(g)\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(g)^{-1}\right\|_{\mathrm{op}}^{1/2}$ is a constant by an assumption, and also the term $\left\|\boldsymbol{D}_c(\widehat{g})\right\|_{\mathrm{op}}^{1/2}\left\|\boldsymbol{D}_c(\widehat{g})^{-1}\right\|_{\mathrm{op}}^{1/2}$ is asymptotically bounded by the uniform consistency of $\widehat{g}'$ for g' by Lemma 13. Finally, we have

$$
\begin{aligned}
&\left\|\boldsymbol{D}_c(g)-\boldsymbol{D}_c(\widehat{g})\right\|_{\mathrm{op}}\\
&=\max_{i=1,\dots,n_2}\left|g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g)))-\widehat{g}'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g})))\right|\\
&\le\max_{i=1,\dots,n_2}\left|g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(g)))-g'(\iota(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\widehat{g})))\right|
\end{aligned}
$$

$$
\begin{aligned}
&+ \max_{i=1,\dots,n_2} \left| g'(\iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g}))) - \widehat{g}'(\iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g}))) \right| \\
&\le B \max_{i=1,\dots,n_2} \left| \iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(g)) - \iota(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\widehat{g})) \right| + \sup_{a \le x \le b} |g'(x) - \widehat{g}'(x)| \\
&= o_{\mathrm{p}}(1),
\end{aligned}
$$

by Lemma 13 and Lemma 19. Thus, (E.21) is $o_{\mathrm{p}}(1)$. Combining these results concludes the proof. □

E.6. *Proof of Proposition 10*

Proof of Proposition 10 Let $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}(\widehat{g})$ for simplicity of the notation. Recall that when $J(\cdot) \equiv \mathbf{0}$,

$$
\begin{aligned}
\tilde{\mu}_{\mathrm{LS}}^2 &= n_1^{-1} \|\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2 - (1-\kappa_1)\tilde{\sigma}_{\mathrm{LS}}^2, \quad \tilde{\sigma}_{\mathrm{LS}}^2 = \frac{\kappa_1}{n_1(1-\kappa_1)^2} \|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2, \\
\widehat{\mu}^2(\widehat{g}) &= n_2^{-1} \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g})\|^2 - (1-\kappa_2)\widehat{\sigma}^2(\widehat{g}), \quad \widehat{\sigma}^2(\widehat{g}) = \frac{\kappa_2}{n_2 \widehat{v}_\lambda^2} \|\boldsymbol{y} - \widehat{g}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2.
\end{aligned}
$$

Since we have

$$
\frac{\tilde{\mu}_{\mathrm{LS}}^2}{\tilde{\sigma}_{\mathrm{LS}}^2} = \frac{\|\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2}{\frac{\kappa_1}{(1-\kappa_1)^2}\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2} - (1-\kappa_1),
$$

and

$$
\frac{\widehat{\mu}^2(\widehat{g})}{\widehat{\sigma}^2(\widehat{g})} = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g})\|^2}{\frac{\kappa_2}{\widehat{v}_\lambda^2}\|\boldsymbol{y} - \widehat{g}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|^2} - (1-\kappa_2),
$$

$\widehat{\sigma}^2(\widehat{g})/\widehat{\mu}^2(\widehat{g}) < \tilde{\sigma}_{\mathrm{LS}}^2/\tilde{\mu}_{\mathrm{LS}}^2$ is equivalent to

$$
\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g})\|}{\|\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|} \cdot \frac{|\widehat{v}_\lambda|}{1-\kappa_1} \cdot \frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{LS}}\|}{\|\boldsymbol{y} - \widehat{g}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g}))\|} > 1.
$$

Next, when $J(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|^2$, recall that

$$
\begin{aligned}
\tilde{\mu}^2 &= \|\tilde{\boldsymbol{\beta}}\|^2 - \tilde{\sigma}^2, \quad \tilde{\sigma}^2 = \kappa_1 n_1^{-1} \|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2 (\tilde{v}^2 + \lambda_1)^{-2} \\
\widehat{\mu}^2(\widehat{g}) &= \|\widehat{\boldsymbol{\beta}}(\widehat{g})\|^2 - \widehat{\sigma}^2(\widehat{g}), \quad \widehat{\sigma}^2(\widehat{g}) = \kappa_2 n_1^{-1} \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{g})\|^2 (\widehat{v}_\lambda^2 + \lambda)^{-2}.
\end{aligned}
$$

Thus, in a similar way as when $J(\cdot) \equiv \mathbf{0}$, we conclude the proof. □

REFERENCES

1. Agarwal, A., S. Kakade, N. Karampatziakis, L. Song, and G. Valiant (2014). Least squares revisited: Scalable approaches for multi-class prediction. In *International Conference on Machine Learning*, pp. 541–549. PMLR.
2. Alquier, P. and G. Biau (2013). Sparse single-index model. *Journal of Machine Learning Research 14*(1).
3. Auer, P., M. Herbster, and M. K. Warmuth (1995). Exponentially many local minima for single neurons. *Advances in neural information processing systems 8*.

4. Balabdaoui, F., C. Durot, and H. Jankowski (2019). Least squares estimation in the monotone single index model. *Bernoulli 25*(4B), 3276–3310.
5. Barbier, J., F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences 116*(12), 5451–5460.
6. Bayati, M., M. A. Erdogdu, and A. Montanari (2013). Estimating lasso risk and noise level. *Advances in Neural Information Processing Systems 26*.
7. Bayati, M. and A. Montanari (2011a). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory 57*(2), 764–785.
8. Bayati, M. and A. Montanari (2011b). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory 58*(4), 1997–2017.
9. Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences 110*(36), 14563–14568.
10. Bellec, P. C. (2022). Observable adjustments in single-index models for regularized m-estimators. *arXiv preprint arXiv:2204.06990*.
11. Bellec, P. C. and T. Koriyama (2025). Error estimation and adaptive tuning for unregularized robust m-estimator. *Journal of Machine Learning Research 26*(16), 1–40.
12. Bellec, P. C. and Y. Shen (2022). Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Conference on Learning Theory*, pp. 1912–1947. PMLR.
13. Bellec, P. C. and C.-H. Zhang (2021). Second-order stein: Sure for sure and other applications in high-dimensional inference. *The Annals of Statistics 49*(4), 1864–1903.
14. Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.
15. Bietti, A., J. Bruna, C. Sanford, and M. J. Song (2022). Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems 35*, 9768–9783.
16. Bolthausen, E. (2014). An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics 325*(1), 333–366.
17. Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(3), 551–577.
18. Candès, E. J. and P. Sur (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics 48*(1), 27–42.
19. Charbonneau, P., E. Marinari, G. Parisi, F. Ricci-tersenghi, G. Sicuro, F. Zamponi, and M. Mezard (2023). *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking after 40 Years*. World Scientific.
20. Chatterjee, S. (2009). Fluctuations of eigenvalues and second order poincaré inequalities. *Probability Theory and Related Fields 143*(1-2), 1–40.
21. Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika 96*(3), 559–575.
22. Cilia, N. D., C. De Stefano, F. Fontanella, and A. S. Di Freca (2018). An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. *Procedia Computer Science 141*, 466–471.
23. Dai, C., B. Lin, X. Xing, and J. S. Liu (2023). False discovery rate control via data splitting. *Journal of the American Statistical Association 118*(544), 2503–2520.
24. Dalalyan, A. S., A. Juditsky, and V. Spokoiny (2008). A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research 9*, 1647–1678.
25. Deshpande, Y., E. Abbe, and A. Montanari (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA 6*(2), 125–170.
26. Donoho, D. and A. Montanari (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields 166*(3), 935–969.
27. Donoho, D. L., A. Maleki, and A. Montanari (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences 106*(45), 18914–18919.

28. Dua, D. and C. Graff (2017). UCI machine learning repository.
29. Eftekhari, H., M. Banerjee, and Y. Ritov (2021). Inference in high-dimensional single-index models under symmetric designs. *The Journal of Machine Learning Research 22*(1), 1247–1309.
30. El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields 170*, 95–175.
31. El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences 110*(36), 14557–14562.
32. Fan, J. and Y. K. Truong (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, 1900–1925.
33. Fan, J., Z. Yang, and M. Yu (2023). Understanding implicit regularization in over-parameterized single index model. *Journal of the American Statistical Association 118*(544), 2315–2328.
34. Feng, O. Y., R. Venkataramanan, C. Rush, and R. J. Samworth (2022). A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning 15*(4), 335–536.
35. Foster, J. C., J. M. Taylor, and B. Nan (2013). Variable selection in monotone single-index models via the adaptive lasso. *Statistics in medicine 32*(22), 3944–3954.
36. Goggin, E. M. (1994). Convergence in distribution of conditional expectations. *The Annals of Probability*, 1097–1114.
37. Guo, X. and G. Cheng (2022). Moderate-dimensional inferences on quadratic functionals in ordinary least squares. *Journal of the American Statistical Association 117*(540), 1931–1950.
38. Härdle, W., V. Spokoiny, and S. Sperlich (1997). Semiparametric single index versus fixed link function modelling. *The Annals of Statistics 25*(1), 212–243.
39. Härdle, W. and T. M. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association 84*(408), 986–995.
40. Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics 50*(2), 949–986.
41. Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*, Volume 12. Springer.
42. Hristache, M., A. Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, 595–623.
43. Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics 58*(1-2), 71–120.
44. Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, 387–421.
45. Krzakala, F., M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment 2012*(08), P08009.
46. Kuchibhotla, A. K., R. K. Patra, and B. Sen (2023). Semiparametric efficiency in convexity constrained single-index model. *Journal of the American Statistical Association 118*(541), 272–286.
47. Lei, L., P. J. Bickel, and N. El Karoui (2018). Asymptotics for high dimensional regression m-estimates: fixed design results. *Probability Theory and Related Fields 172*, 983–1079.
48. Li, K.-C. and N. Duan (1989). Regression analysis under link violation. *The Annals of Statistics 17*(3), 1009–1052.
49. Li, Q. and J. S. Racine (2023). *Nonparametric econometrics: theory and practice*. Princeton University Press.
50. Li, Y. and P. Sur (2023). Spectrum-aware adjustment: A new debiasing framework with applications to principal components regression. *arXiv preprint arXiv:2309.07810*.
51. Loureiro, B., C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems 34*, 18137–18151.
52. Macris, N., C. Rush, et al. (2020). All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. *Advances in Neural Information Processing Systems 33*, 14915–14926.

53. Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society*, 1315–1327.
54. Mei, S. and A. Montanari (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics 75*(4), 667–766.
55. Mézard, M., G. Parisi, and M. A. Virasoro (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, Volume 9. World Scientific Publishing Company.
56. Miolane, L. and A. Montanari (2021). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics 49*(4), 2313–2335.
57. Montanari, A., F. Ruan, Y. Sohn, and J. Yan (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.
58. Montanari, A. and R. Venkataramanan (2021). Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics 49*(1), 321–345.
59. Mousavi, A., A. Maleki, and R. G. Baraniuk (2018). Consistent parameter estimation for LASSO and approximate message passing. *The Annals of Statistics 46*(1), 119 – 148.
60. Nishiyama, Y. and P. M. Robinson (2005). The bootstrap and the edgeworth correction for semiparametric averaged derivatives. *Econometrica 73*(3), 903–948.
61. Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 1403–1430.
62. Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2168–2172. IEEE.
63. Sakar, C., G. Serbes, A. Gunduz, H. Nizam, and B. Sakar (2018). Parkinson's Disease Classification. UCI Machine Learning Repository.
64. Salehi, F., E. Abbasi, and B. Hassibi (2019). The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems 32*.
65. Sawaya, K., Y. Uematsu, and M. Imaizumi (2023). Feasible adjustments of statistical inference in high-dimensional generalized linear models. *arXiv preprint arXiv:2305.17731*.
66. Stefanski, L. A. and R. J. Carroll (1990). Deconvolving kernel density estimators. *Statistics 21*(2), 169–184.
67. Sur, P. and E. J. Candès (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences 116*(29), 14516–14525.
68. Sur, P., Y. Chen, and E. J. Candès (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields 175*(1), 487–558.
69. Takahashi, T. and Y. Kabashima (2018). A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment 2018*(7), 073405.
70. Tan, K. and P. C. Bellec (2023). Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. *arXiv preprint arXiv:2305.17825*.
71. Thrampoulidis, C., E. Abbasi, and B. Hassibi (2018). Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory 64*(8), 5592–5628.
72. Wan, Y., S. Datta, J. J. Lee, and M. Kong (2017). Monotonic single-index models to assess drug interactions. *Statistics in medicine 36*(4), 655–670.
73. Xing, X., Z. Zhao, and J. S. Liu (2023). Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association 118*(541), 222–241.
74. Yadlowsky, S., T. Yun, C. Y. McLean, and A. D'Amour (2021). Sloe: A faster method for statistical inference in high-dimensional logistic regression. *Advances in Neural Information Processing Systems 34*, 29517–29528.
75. Yang, G. and E. J. Hu (2021). Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR.
76. Zhao, Q., P. Sur, and E. J. Candès (2022). The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli 28*(3), 1835–1861.