Journal of Machine Learning Research 27 (2026) 1-76 Submitted 2/23; Revised 3/26; Published 4/26

Inference with non-differentiable surrogate loss in a general high-dimensional classification framework

Muxuan Liang MLIANG2@MDANDERSON.ORG
Department of Biostatistics
The University of Texas MD Anderson Cancer Center
Houston, Texas 77030, USA

Yang Ning YN265@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, New York 14853, USA

Maureen A Smith MAUREENSMITH@WISC.EDU
Departments of Population Health and Family Medicine
University of Wisconsin-Madison
Madison, Wisconsin 53706, USA

Ying-Qi Zhao YQZHAO@FREDHUTCH.ORG
Public Health Sciences Division
Fred Hutchinson Cancer Center
Seattle, Washington 98109, USA

Editor: Eric Laber

Abstract

Penalized empirical risk minimization with a surrogate loss function is often used to learn a high-dimensional linear decision rule in classification problems. Although much of the literature focus on the generalization error, there is a lack of inference procedures for identifying the driving factors of the estimated decision rule, especially when the surrogate loss is non-differentiable. We propose a kernel-smoothed decorrelated score to construct hypothesis tests and interval estimators for a linear decision rule estimated using a piecewise linear surrogate loss, which has a discontinuous gradient and non-regular Hessian. Specifically, we adopt kernel approximations to smooth the discontinuous gradient near discontinuity points and approximate the non-regular Hessian of the surrogate loss. In applications where additional nuisance parameters are involved, we propose a novel cross-fitted version to accommodate flexible nuisance estimates and kernel approximations. We establish the limiting distribution of the kernel-smoothed decorrelated score and its cross-fitted version in a high-dimensional setup. Simulation and real data analysis are conducted to demonstrate the validity and the superiority of the proposed method.

Keywords: Classification, Double machine learning, High-dimensional inference, Non-differentiable loss, Personalized medicine

1. Introduction

Classification identifies which of a set of labels an observation belongs to. Well-known classification methods include logistic regression, support vector machines, and many others.

When the dimensionality of the covariate space is high, which is common due to the increasing adoption of large datasets in biomedical applications, classification is more challenging (Fan and Fan, 2008; Dobriban and Wager, 2018; Bing and Wegkamp, 2023). Additionally, it has been shown that many problems can be formulated within a general classification framework (Bartlett et al., 2006; Bartlett and Wegkamp, 2008; Zhao et al., 2012; Zhou et al., 2017; Zhao et al., 2019). In these problems, the goal is to derive a data-driven decision rule that minimizes a loss function (or maximizes a utility function) defined according to the problem setup. For example, in prediction problems, the goal is to derive a data-driven decision rule that minimizes the prediction error for binary outcomes or labels; in precision medicine, the goal is to derive a data-driven decision rule that maximizes the averaged utility over entire population if the derived decision rule were implemented to recommend treatment options (Zhao et al., 2012; Zhou et al., 2017; Zhao et al., 2019).

Empirical risk minimization (ERM) is often used to estimate such a data-driven decision rule by minimizing a convex surrogate of the loss function. Statistical inference of the constructed decision rules not only provides uncertainty quantification of the data-driven decision rule, but also enables a data-driven paradigm for new scientific discovery, e.g., to identify the risk factors for the outcome of interest in prediction problems; and to identify the driving factors of the estimated data-driven decision rule to inform new treatment guidelines in precision medicine (Jeng et al., 2018; Wang et al., 2019; Ning and Liu, 2017; Liang et al., 2022). However, while there is a large literature on classification and the generalization error of an ERM with a convex surrogate, statistical inference within a high-dimensional classification framework is less well-studied. For regression problems, Van de Geer et al. (2014) proposed a debiased Lasso estimator for generalized linear models and established asymptotic normality under modest regularity conditions. Ning and Liu (2017) proposed a decorrelated score to test a low-dimensional projection of high-dimensional coefficient vectors, which can be applied to M-estimation under a strictly convex and differentiable loss function. Dezeure et al. (2017) proposed a bootstrap procedure to conduct simultaneous inference for parameters in groups with diverging group sizes. More recently, partial penalized tests proposed in Shi et al. (2019) can be applied to test hypotheses involving a growing number of coefficients. Ma et al. (2021) considered global hypothesis testing and multiple testing procedures for high-dimensional logistic regression models. Wu et al. (2023) proposed an inference procedure for single-index models with differentiable link functions. However, none of these methods can be applied to non-differentiable loss functions, such as the hinge loss, which is commonly adopted in classification problems.

Deriving an inference procedure for an ERM with a convex surrogate that is non-differentiable is more complicated. One such example is the popular classification method - support vector machine (SVM) (Cortes and Vapnik, 1995). It employs the hinge loss as a surrogate loss, which is continuous but non-differentiable. The majority of the existing SVM literature focused on its consistency, and the convergence rate of the risk under the derived classifiers to the Bayes risk (Lin, 2004; Zhang, 2004; Steinwart, 2005; Zhang, 2004; Bartlett et al., 2006; Steinwart et al., 2007; Vert and Vert, 2006; Blanchard et al., 2008). Peng et al. (2016) provided an error bound for a penalized SVM in ultra-high dimension; Zhang et al. (2016a) and Zhang et al. (2016b) focused on variable selection for SVMs in moderately high dimension. The literature on the asymptotic distribution of these estimators is limited. Koo et al. (2008) investigated the Bahadur representation of a linear SVM,

which implies the asymptotic normality of the estimator in a low-dimensional setup. Due to the non-differentiability of the hinge loss, they proposed a non-parametric estimator for the asymptotic variance. Wang et al. (2019) proposed a distributed inference procedure for a linear SVM. To handle the lack of differentiability, they used a smoothed loss function to approximate the hinge loss and showed the asymptotic normality of the estimator provided that $p/n \to 0$, where n is the total sample size, and p is the number of the covariates. However, an associated inference procedure in the high-dimensional setup is still lacking.

Furthermore, the classification framework has been adapted to learn an individualized treatment rule (ITR), which recommends treatment according to patient characteristics. There has been much literature proposing to learn ITRs from a weighted classification framework, where the weights are related to the observed clinical outcomes (Zhao et al., 2012; Zhou et al., 2017; Chen et al., 2016; Zhao et al., 2014, 2019; Pan and Zhao, 2021; Xue et al., 2022). In recent work, Zhao et al. (2019) and Liang et al. (2022) introduced both the outcome regression models and propensity score as nuisance parameters in the weights. To avoid model misspecification, the nuisance parameters are estimated via nonparametric or flexible machine learning algorithms. These algorithms may lead to nuisance parameter estimators with slow convergence rates. These flexible estimators of the nuisance parameters with possible slow convergence rates create a large barrier in statistical inference. Liang et al. (2022) proposed an inference procedure that can handle strictly convex differentiable loss functions. However, inference procedure for non-differentiable loss functions remain largely unexplored.

We propose a novel inference procedure for linear decision rules under a general classification framework in a high-dimensional setup, which can deal with non-differentiable convex surrogate loss functions. We introduce a kernel-smoothed decorrelated score for the inference procedure, which utilizes a local kernel function to smooth the discontinuous gradient near discontinuity points, where the loss function is not differentiable, and a global kernel function to approximate the non-regular Hessian. By using these kernel functions, the proposed procedure can be applied to any piece-wise linear convex loss functions. Furthermore, unlike the existing literature (Wang et al., 2019; Koo et al., 2008), the proposed procedure is valid even when $p/n \to +\infty$ and can be extended to test a hypothesis involving a growing number of projections. For the general classification problems, additional nuisance parameters may be involved in the loss function. Motivated by Chernozhukov et al. (2018), we further propose a new cross-fitting algorithm to efficiently accommodate these nuisance parameters. We show the uniform validity of the kernel-smoothed decorrelated score based procedure even when the nuisance parameters are estimated using nonparametric or flexible machine learning methods. Simulations and real data examples show the superiority of the proposed method.

Sections 2 and 3 introduce the kernel smoothed decorrelated score for the classification problem and the general classification framework. Section 4 provides theoretical justifications for the proposed procedure. Sections 5 and 6 present simulations and real data analyses. Section 7 concludes the paper and provides a discussion on future directions.

2. Statistical inference for classification problems

2.1 A classification problem

We observe the covariates $\boldsymbol{X} \in \mathbb{R}^p$ and a label $A \in \{-1, 1\}$. A decision rule, $d : \mathbb{R}^p \to \{-1, 1\}$, is a mapping from the covariate space $\mathbb{R}^p$ to the label space $\{-1, 1\}$. We mainly focus on the high-dimensional setting with $p/n \to +\infty$. For any rule, d, define the classification error for d as

$$L(d) = \mathbb{E}\left[1\left\{A \neq d(\boldsymbol{X})\right\}\right],$$

which is also the expected zero-one loss to compare $d(\boldsymbol{X})$ and A. The goal is to identify the decision rule which minimizes the classification error. This optimal rule, also known as the Bayes rule, is given by $d_{\text{opt}}(\mathbf{x}) = \text{sgn}\left\{P(A = 1|\mathbf{X} = \mathbf{x}) - 1/2\right\}$.

To estimate/infer this d_{opt}, we observe both $\boldsymbol{X}$ and A. With these observed data, we optimize the empirical analogue of $L(d)$, denoted by $\widehat{L}(d) = \widehat{\mathbb{E}}_n\left[1\left\{A \neq d(\boldsymbol{X})\right\}\right]$, where $\widehat{\mathbb{E}}_n[\cdot]$ is the empirical expectation. However, $\widehat{L}(d)$ is hard to optimize due to the discontinuity of the zero-one loss, and the space of decision rules is large. To tackle these challenges, a common strategy is to replace the zero-one loss by a piecewise convex loss function, often called a surrogate loss function, which is a real-valued map on $\mathbb{R}$. Examples of piece-wise convex loss functions include the hinge loss and the modified hinge loss used for classification with a rejection option (Bartlett and Wegkamp, 2008). Furthermore, we focus on linear decision rules. We thus optimize

$$\widehat{L}_\phi(\boldsymbol{\beta}) = \widehat{\mathbb{E}}_n[\phi(A\boldsymbol{X}^\top\boldsymbol{\beta})],$$

where $\phi(\cdot)$ is a surrogate loss function. Define the minimizer of $L_\phi(\boldsymbol{\beta})$ as $\boldsymbol{\beta}^*_\phi$, where $L_\phi(\boldsymbol{\beta}) = \mathbb{E}[\phi(A\boldsymbol{X}^\top\boldsymbol{\beta})]$. With high-dimensional data, we adopt penalized empirical risk minimization to estimate $\boldsymbol{\beta}^*_\phi$. Specifically, we consider

$$\widehat{L}^{\lambda_n}_\phi(\boldsymbol{\beta}) = \widehat{\mathbb{E}}_n[\phi(A\boldsymbol{X}^\top\boldsymbol{\beta})] + \lambda_n\|\boldsymbol{\beta}\|_1,$$

where λ_n is a tuning parameter. Denote the minimizer of $\widehat{L}^{\lambda_n}_\phi(\boldsymbol{\beta})$ as $\widehat{\boldsymbol{\beta}}_\phi$. Through the estimator $\widehat{\boldsymbol{\beta}}_\phi$, we aim to construct a hypothesis testing procedure and confidence interval for a low-dimensional projection of $\boldsymbol{\beta}^*_\phi$, i.e. $\boldsymbol{\eta}^\top\boldsymbol{\beta}^*_\phi$, where $\boldsymbol{\eta}$ is a known sparse vector.

Remark 1 *In the classification problem, the classification error is determined by the linear direction represented by $\boldsymbol{\beta}^*_\phi$. Thus, when comparing with other methods, we normalize the estimates such that its first coordinate equals* 1*, i.e., $|\beta^*_{\phi,1}| = 1$.*

Remark 2 *Although hinge loss is Fisher consistent, the $\boldsymbol{\beta}^*_\phi$ does not necessarily equal the optimal linear rule minimizing $L(d)$. In classification problems, the hinge loss leads to the maximum-margin linear separator, which can be different from the 0–1-optimal linear hyperplane. This requires additional assumptions, e.g., sufficient conditions in Liang et al. (2022); or a mixture of Gaussian model assumption (see Section 3 in Koo et al. (2008)). We assume that the users have chosen to use hinge loss to derive a data-driven decision rule, and our focus is on how to conduct inference based on these estimates.*

2.2 Kernel-smoothed decorrelated score

For the purpose of illustration, we consider a hypothesis testing problem $\mathcal{H}_0 : \beta^*_{\phi,l} = 0$ versus $\mathcal{H}_a : \beta^*_{\phi,l} \neq 0$, where $\beta^*_{\phi,l}$ is the l-th coordinate of the $\boldsymbol{\beta}^*_\phi$. The proposed approach can be easily applied to test $\mathcal{H}_0 : \boldsymbol{\eta}^\top \boldsymbol{\beta}^*_\phi = 0$ versus $\mathcal{H}_a : \boldsymbol{\eta}^\top \boldsymbol{\beta}^*_\phi \neq 0$.

We first review a decorrelated score, proposed in Ning and Liu (2017), which can be used to test this hypothesis when the surrogate loss ϕ is differentiable. The key of the decorrelated score is to decouple the estimation error of the high-dimensional component $\boldsymbol{\beta}^*_{\phi,-l}$ from the estimation of $\beta^*_{\phi,l}$, where $\boldsymbol{\beta}^*_{\phi,-l}$ is the sub-vector of $\boldsymbol{\beta}^*_\phi$ without the l-th coordinate. Define the minimizer of

$$L_{\phi'',l}(\boldsymbol{w}) = \mathbb{E}\left[\phi''\left(A\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\right)(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w})^2\right]$$

as $\boldsymbol{w}^*_{\phi,l}$. The decorrelated score is defined as

$$S_{\phi',l}(\boldsymbol{\beta}; \boldsymbol{w}^*_{\phi,l}) = \widehat{\mathbb{E}}_n\left[A\phi'\left(A\boldsymbol{X}^\top\boldsymbol{\beta}\right)(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\right].$$

Let $\widehat{\boldsymbol{\beta}}_{\phi,null(l)}$ be a vector equals $\widehat{\boldsymbol{\beta}}_\phi$ except the l-th coordinate is fixed at 0. Under the null hypothesis, it can be shown that $\sqrt{n}S_{\phi',l}(\widehat{\boldsymbol{\beta}}_{\phi,\mathrm{null}(l)}; \boldsymbol{w}^*_{\phi,l}) \xrightarrow{d} \mathcal{N}(0, \sigma_l^2)$, where σ_l^2 is some constant. However, this procedure cannot be applied to the non-differentiable loss functions due to the non-existence of regular $\phi'(\cdot)$ and $\phi''(\cdot)$.

For a piece-wise linear convex loss function ϕ, although ϕ is not differentiable at the discontinuity points, the gradient is well defined on any open intervals without discontinuity points. Mathematically, suppose that $-\infty < t_1 < \cdots < t_J < +\infty$ are the jump discontinuity points of ϕ', then ϕ' can be defined as $\Delta_0 + \sum_{j=1}^J \Delta_j 1\{t - t_j \geqslant 0\}$ on any open intervals. In addition, a Hessian ϕ'' can be defined using the Dirac function $\delta(t)$ as $\sum_{j=1}^J \Delta_j \delta(t - t_j)$, which is non-regular because it achieves $+\infty$ at 0 and vanishes at all other points.

We use kernel functions to smooth the gradient near discontinuity points and approximate the Hessian of the surrogate loss function. Specifically, we obtain a smoothed gradient ϕ' by a **local** kernel function, a smooth function whose derivative has a support contained in a compact interval. Consider $H(\cdot)$ satisfying $H(t) = 1$ if $t \geqslant 1$ and $H(t) = 0$ if $t \leqslant -1$; thus, its derivative has a support on $[-1, 1]$. As the bandwidth $h_{\mathrm{lo}} \to 0$, the functions $H(t/h_{\mathrm{lo}})$ and $H'(t/h_{\mathrm{lo}})/h_{\mathrm{lo}}$ approach the indicator function $1(t \geqslant 0)$ and the Dirac function $\delta(t)$, respectively. One example of such kernel is

$$H(t) = \begin{cases} 0 & \text{if } t \leqslant -1, \\ \frac{1}{2} + \frac{15}{16}(t - \frac{2}{3}t^3 + \frac{1}{5}t^5) & \text{if } |t| < 1, \\ 1 & \text{if } t \geqslant 1. \end{cases}$$

Define

$$\widetilde{\phi}'(t) = \Delta_0 + \sum_{j=1}^J \Delta_j H\left(\frac{t - t_j}{h_{\mathrm{lo}}}\right),$$

where $\Delta_j = \phi'(t_j+) - \phi'(t_j-)$. For any open interval where ϕ' exists, $\widetilde{\phi}'(t)$ is different from ϕ' only on $\bigcup_{j=1}^J [t_j - h_{\mathrm{lo}}, t_j + h_{\mathrm{lo}}]$. Thus, the smoothed gradient of $L_\phi(\boldsymbol{\beta})$ can be naturally

defined by $\mathbb{E}\left[A\widetilde{\phi}'\left(A\boldsymbol{X}^\top\boldsymbol{\beta}\right)\boldsymbol{X}\right]$. The smoothed score function of $\beta^*_{\phi,l}$ is

$$\mathbb{E}\left[A\widetilde{\phi}'\left(A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,\text{null(l)}}\right)X_l\right].$$

However, in high dimension, it is biased due to the estimation error of the high-dimensional component $\widehat{\boldsymbol{\beta}}_{\phi,-l}$. Following the idea of the decorrelated score, we decouple the smoothed score function of $\beta^*_{\phi,l}$ with the estimation error of $\widehat{\boldsymbol{\beta}}_{\phi,-l}$. Define $\boldsymbol{w}^*_{\phi,l}$ as the minimizer of

$$L_{\phi'',l}(\boldsymbol{w}) = \mathbb{E}\left[\sum_{j=1}^{J}\Delta_j\delta(t_j - A\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w})^2\right].$$

We construct the kernel-smoothed decorrelated score

$$S_{\widetilde{\phi}',l}(\boldsymbol{\beta};\boldsymbol{w}^*_{\phi,l}) = \widehat{\mathbb{E}}_n\left[A\widetilde{\phi}'\left(A\boldsymbol{X}^\top\boldsymbol{\beta}\right)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})\right].$$

To construct a score test for $\beta^*_{\phi,l} = 0$ using $S_{\widetilde{\phi}',l}(\boldsymbol{\beta};\boldsymbol{w}^*_{\phi,l})$, we need to estimate $\boldsymbol{w}^*_{\phi,l}$, which is nontrivial due to the non-regularity of the $\delta(t)$ function. We propose to use a **global** kernel function with a support on the entire real line to approximate $\delta(t)$. Suppose that $G(t)$ is a global kernel function satisfying that 1) $G(t) > 0$, $\forall t$; 2) $\int G(t)\mathrm{d}t = 1$; 3) $\int tG(t)\mathrm{d}t = 0$. Define $G_{h_{\text{gb}}}(t) = h_{\text{gb}}^{-1}G(t/h_{\text{gb}})$, where h_{gb} is the bandwidth of the global kernel. The estimated $\boldsymbol{w}^*_{\phi,l}$, denoted by $\widehat{\boldsymbol{w}}_{\phi,l}$, can be obtained by minimizing

$$\widehat{\mathbb{E}}_n\left[\sum_{j=1}^{J}\Delta_j G_{h_{\text{gb}}}(t_j - A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w})^2\right] + \mu_n\|\boldsymbol{w}\|_1,$$

where μ_n is a tuning parameter. The above objective function is strictly convex and smooth with respect to $\boldsymbol{w}$ given that $G(t) > 0$, and thus can be easily minimized.

Remark 3 *A straightforward idea is to use the gradient of the local kernel to approximate $\delta(t)$. However, only data points near the discontinuity points will contribute to the estimation of $\boldsymbol{w}^*_{\phi,l}$ in this case, because further derivative of $\tilde{\phi}'(t)$ is zero when t is away from the discontinuity points. In supplementary material, we show that using the local kernel function to approximate the non-regular Hessian will lead to a slower convergence rate of the estimator and require more restrictive conditions in the inference procedure.*

Remark 4 *Kernel functions are used to smoothen the gradient and hessian around non-differentiable points, which lead to gradient and hessian that can be easily computed. For non-differentiable loss, it may be possible to use sub-gradient of ϕ to replace $\phi'(\cdot)$ and sub-hessian of ϕ to replace $\phi''(\cdot)$; however, this approximation may be computationally challenging and problematic. For example, hinge loss has vanishing (sub-)hessian almost everywhere, and thus a more detailed approximation, e.g., using a global/local kernel function, is needed.*

Replacing the $\boldsymbol{w}^*_{\phi,l}$ by $\widehat{\boldsymbol{w}}_{\phi,l}$ in $S_{\widetilde{\phi}',l}(\boldsymbol{\beta};\boldsymbol{w}^*_{\phi,l})$, we can calculate the value of the estimated decorrelated score function $S_{\widetilde{\phi}',l}(\widehat{\boldsymbol{\beta}}_{\phi,null(l)};\widehat{\boldsymbol{w}}_{\phi,l})$ to test the null hypothesis. However, additional challenges arise from the correlation between $\widehat{\boldsymbol{\beta}}_\phi$ and the loss function used to

estimate $\boldsymbol{w}^*_{\phi,l}$, as well as the correlation between $\widehat{\boldsymbol{\beta}}_\phi$ and the kernel-smoothed decorrelated score function $S_{\widetilde{\phi}',l}(\cdot)$. We design a novel sample-splitting strategy, where the estimator $\widehat{\boldsymbol{\beta}}_\phi$ is independent from the data used to estimate $\boldsymbol{w}^*_{\phi,l}$, as well as the data used to construct $S_{\widetilde{\phi}',l}(\cdot)$. In addition, instead of averaging over multiple estimators as in the cross-fitting procedure (Chernozhukov et al., 2018), we average over the loss functions to estimate $\widehat{\boldsymbol{w}}_{\phi,l}$, which is more computationally robust. The entire inference procedure is summarized in Algorithm 1. In our simulation and real data analysis, we set $K = 2$. For the bandwidth selection, we choose $h_{\mathrm{lo}} = 1/\sqrt{n \log n}$ and $h_{\mathrm{gb}} = (\log p/n)^{-1/5}$ based on the theoretical results in Section 4.1 and 4.2.

Algorithm 1: Inference of $\boldsymbol{\beta}^*_\phi$.

Input: A random seed; n samples; a positive integer K.
Output: A p-value for $\mathcal{H}_0 : \beta^*_{\phi,l} = 0$.
Randomly split data into K parts $I_1, \cdots, I_K$ with equal size, and set $k = 1$;
Estimate $\boldsymbol{\beta}^*_\phi$ using data in I_k^c by

$$\widehat{\mathbb{E}}_n^{(-k)}[\phi(A\boldsymbol{X}^\top\boldsymbol{\beta})] + \lambda_n^{(-k)}\|\boldsymbol{\beta}\|_1,$$

where $\widehat{\mathbb{E}}_n^{(-k)}[\cdot]$ is the empirical average on I_k^c, and denote the estimator as $\widehat{\boldsymbol{\beta}}_\phi^{(-k)}$.
The parameter $\lambda_n^{(-k)}$ is tuned by cross-validation; we obtain $\widehat{\boldsymbol{\beta}}_\phi^{(-k)}$ for each k;
Obtain an estimator $\widehat{\boldsymbol{w}}_{\phi,l}$ for $\boldsymbol{w}^*_{\phi,l}$by minimizing

$$\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j G_{h_{\mathrm{gb}}}(t_j - A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)})\right\}(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w})^2\right] + \mu_n\|\boldsymbol{w}\|_1,$$

where μ_n is tuned by cross-validation, and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ is the empirical average on I_k;
Let $\left(\widehat{\boldsymbol{\beta}}^{(k)}_{\phi,null(l)}\right)^\top$ equal to $\widehat{\boldsymbol{\beta}}^{(k)}_\phi$ except its l-th coordinate replaced by 0. Construct the kernel-smoothed decorrelated score test statistic as

$$S_{\widetilde{\phi}',null(l)} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[A\widetilde{\phi}'\left(A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right],$$

and the estimator of the variance
$\widehat{\sigma}_l^2 = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\widetilde{\phi}'\left(A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right\}^2\right]$;
Calculate the p-value by $2\left(1 - \Phi(n^{-1/2}|S_{\widetilde{\phi}',null(l)}|/\widehat{\sigma}_l)\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

2.3 Construction of confidence interval

Due to the penalization, the estimator $\widehat{\beta}_{\phi,l}$ is biased and cannot be directly used to construct an interval estimator. To remove this bias, the key idea is to consider a one-step estimator

based on an unbiased estimating equation. Motivated by the fact that the kernel-smoothed decorrelated score is an asymptotically unbiased estimating equation for $\beta^*_{\phi,l}$, when the bandwidth shrinks to 0, we consider

$$\widetilde{\beta}_{\phi,l} = \bar{\beta}_{\phi,l} - S_{\widetilde{\phi}',l}(\widehat{\boldsymbol{\beta}}_{\phi}^{(-k)}; \widehat{\boldsymbol{w}}_{\phi,l})/\widehat{I}_l,$$

where

$$\begin{aligned}
\bar{\beta}_{\phi,l} &= \frac{1}{K}\sum_{k=1}^{K} \widehat{\beta}_{\phi,l}^{(-k)}, \\
S_{\widetilde{\phi}',l}(\boldsymbol{\beta}; \widehat{\boldsymbol{w}}_{\phi,l}) &= \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathbb{E}}_n^{(k)} \left[A\widetilde{\phi}' \left(A\boldsymbol{X}^\top \boldsymbol{\beta} \right) (X_l - \boldsymbol{X}_{-l}^\top \widehat{\boldsymbol{w}}_{\phi,l}) \right], \\
\widehat{I}_l &= \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathbb{E}}_n^{(k)} \left[\sum_j \Delta_j G_{h_{\mathrm{gb}}} \left(t_j - A\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_{\phi}^{(-k)} \right) (X_l - \boldsymbol{X}_{-l}^\top \widehat{\boldsymbol{w}}_{\phi,l})^2 \right],
\end{aligned}$$

can be calculated in Step 1 of Algorithm 1. Then 95%-confidence interval can be constructed by

$$\left(\widetilde{\beta}_{\phi,l} - 1.96 n^{-1/2} \widehat{\sigma}_l / \widehat{I}_l, \widetilde{\beta}_{\phi,l} + 1.96 n^{-1/2} \widehat{\sigma}_l / \widehat{I}_l \right).$$

3. A general classification framework

Weighted classification assigns class-specific weights, which reflect the relative importance of each class. More broadly, weights can depend on the covariate $\boldsymbol{X}$ that each individual has specific weights. Mathematically, the task is to learn a decision rule, $d(\boldsymbol{X})$, which minimizes a weighted zero-one loss

$$L(d; W_1, W_{-1}) = \mathbb{E}[W_1(\boldsymbol{X}) 1\{d(\boldsymbol{X}) \neq 1\} + W_{-1}(\boldsymbol{X}) 1\{d(\boldsymbol{X}) \neq -1\}], \tag{1}$$

where $W_a(\boldsymbol{X})$'s are pre-specified weights depending on the problem of interest.

Under the proposed general classification framework, the minimizer of the loss (1) can be derived as

$$d_{\mathrm{opt}}(\boldsymbol{X}) = \mathrm{sgn}\{W_1(\boldsymbol{X}) - W_{-1}(\boldsymbol{X})\},$$

where $\mathrm{sgn}(t) = 1$ if $t \geqslant 0$; $\mathrm{sgn}(t) = -1$, otherwise.

We are interested in developing an inference procedure for a low-dimensional projection of $\boldsymbol{\beta}^*_\phi$, where $\boldsymbol{\beta}^*_\phi$ is the minimizer of

$$L_\phi(\boldsymbol{\beta}; W_1, W_{-1}) = \mathbb{E}\left[W_1 \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) + W_{-1} \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \right],$$

where for simplicity we use the notation W_1 and W_{-1}. The weights are typically not directly observed, and need to be estimated. Specifically, to estimate $\boldsymbol{\beta}^*_\phi$, we minimize

$$\widehat{L}_\phi^{\lambda_n}(\boldsymbol{\beta}; \widehat{W}_1, \widehat{W}_{-1}) = \widehat{\mathbb{E}}_n \left[\widehat{W}_1 \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) + \widehat{W}_{-1} \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \right] + \lambda_n \|\boldsymbol{\beta}\|_1,$$

where $\widehat{W}_a$'s are the estimated weights using the observed data.

The general weighted classification framework has been broadly used in many applications. We provide two examples below.

3.1 Classification with missing labels

In a classification problem, it is likely that only partial samples are fully observed with $(\boldsymbol{X}, A)$, and we only observe the covariate information $\boldsymbol{X}$ for the remaining samples. For example, to predict patient-reported outcomes, covariate information is collected at baseline, whereas the outcomes are collected by a survey after intervention. We can only observe outcomes for those patients who fill out the survey, and other patients' outcomes are missing. This problem is also related to the semi-supervised learning literature, where the missing is often assumed to be completely at random (Wang and Shen, 2007; Hoffmann et al., 2020; Song et al., 2024; Deng et al., 2024; Cai et al., 2025). In this case, our proposed method can also be applied to infer the derived linear rule in semi-supervised learning.

Let R be the missing indicator. We assume missing at random (MAR), i.e., $A \perp R \mid \boldsymbol{X}$. By leveraging both the labeled and unlabeled samples, an estimator of the classification error is

$$\widehat{\mathbb{E}}_n \left[\widehat{W}_1(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) 1\{d(\boldsymbol{X}) \neq 1\} + \widehat{W}_{-1}(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) 1\{d(\boldsymbol{X}) \neq -1\} \right]. \tag{2}$$

Here,

$$\widehat{W}_a(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) = \frac{1\{R=1, A=a\}}{\widehat{\pi}_1(\boldsymbol{X})} - \frac{1\{R=1\} - \widehat{\pi}_1(\boldsymbol{X})}{\widehat{\pi}_1(\boldsymbol{X})} \widehat{p}_a(\boldsymbol{X}),$$

where $\widehat{\pi}_1(\boldsymbol{X})$ is an estimate of the nuisance parameter $\pi_1 = P(R=1 \mid \boldsymbol{X})$ and $\widehat{p}_a(\boldsymbol{X})$ is an estimate of the nuisance parameter $p_a(\boldsymbol{X}) = P(A=a|\boldsymbol{X})$. Let $\bar{\pi}_1(\boldsymbol{x})$ and $\bar{p}_a(\boldsymbol{x})$ be the point-wise limits of the $\widehat{\pi}_1(\boldsymbol{x})$ and $\widehat{p}_a(\boldsymbol{x})$. Define

$$\overline{W}_a(\boldsymbol{X}, A; \bar{\pi}, \bar{p}) = \frac{1\{R=1, A=a\}}{\bar{\pi}_1(\boldsymbol{X})} - \frac{1\{R=1\} - \bar{\pi}_1(\boldsymbol{X})}{\bar{\pi}_1(\boldsymbol{X})} \bar{p}_a(\boldsymbol{X}).$$

Under the class of linear decision rules, we can minimize

$$\begin{aligned} \widehat{L}_\phi^{\lambda_n}(\boldsymbol{\beta}; \widehat{W}_1, \widehat{W}_{-1}) = \widehat{\mathbb{E}}_n \Big[& \widehat{W}_1(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) \\ & + \widehat{W}_{-1}(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \Big] + \lambda_n \|\boldsymbol{\beta}\|_1. \end{aligned}$$

We assume that there exists a constant c such that $0 < c < \min\{\pi_1, \widehat{\pi}_1\}$. The target of the inference procedure is a low-dimensional projection of $\boldsymbol{\beta}_\phi^*$, which minimizes

$$L_\phi(\boldsymbol{\beta}; \overline{W}_1, \overline{W}_{-1}) = \mathbb{E} \left[\overline{W}_1(\boldsymbol{X}, A; \bar{\pi}, \bar{p}) \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) + \overline{W}_{-1}(\boldsymbol{X}, A; \bar{\pi}, \bar{p}) \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \right]. \tag{3}$$

Further, due to the construction of the $\overline{W}_a(\boldsymbol{X}, A; \bar{\pi}, \bar{p})$'s, if either $\bar{\pi}_1 = \pi_1$ or $\bar{p}_a = p_a$, then $\mathbb{E}[\overline{W}_a \mid \boldsymbol{X}] = W_a(\boldsymbol{X}) \equiv P(A=a|\boldsymbol{X})$ and the objective function (3) equals to

$$L_\phi(\boldsymbol{\beta}; W_1, W_{-1}) = \mathbb{E} \left[W_1 \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) + W_{-1} \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \right].$$

3.2 Estimation of individualized treatment rules

An individualized treatment rule $d(\boldsymbol{X})$ maps the covariate space $\mathbb{R}^p$ to the treatment space $\{-1, 1\}$, which is the label space here. To define the objective function, we adopt the

potential outcome framework (Rubin, 1974, 2005). Denote the potential outcome under treatment $a \in \{-1, 1\}$ as $Y(a)$, and the potential outcome under an individualized treatment rule d as $Y(d)$. Assume larger outcomes are more preferable. The goal is to learn the optimal individualized treatment rule that maximizes $\mathbb{E}[Y(d)]$.

We observe the covariate information $\boldsymbol{X}$, the assigned treatment A, and the outcome Y. We assume the following conditions: 1) the Stable Unit Treatment Value Assumption (SUTVA) (Imbens and Rubin, 2015); 2) the strong ignorability $Y(-1), Y(1) \perp A \mid \boldsymbol{X}$; 3) Consistency $Y = Y(a)$ if $A = a$. SUTVA condition assumes that the potential outcomes for a patient do not vary with the other patients' treatments. It also implies that there are no different versions of the treatment. The strong ignorability condition means that there is no unmeasured confounding between the potential outcomes and the treatment. The consistency ensures that the observed outcome is the potential outcome under the assigned treatment. Under these conditions, an augmented inverse probability weighted estimator of $\mathbb{E}[Y(d)]$ is

$$\widehat{\mathbb{E}}_n \left[\widehat{W}_1(Y, \boldsymbol{X}, A; \widehat{p}, \widehat{Q}) 1\{d(\boldsymbol{X}) = 1\} + \widehat{W}_{-1}(Y, \boldsymbol{X}, A; \widehat{p}, \widehat{Q}) 1\{d(\boldsymbol{X}) = -1\} \right]. \tag{4}$$

Here,

$$\widehat{W}_a(Y, \boldsymbol{X}, A; \widehat{p}, \widehat{Q}) = \frac{Y 1\{A = a\}}{\widehat{p}_a(\boldsymbol{X})} + \frac{1\{A = a\} - \widehat{p}_a(\boldsymbol{X})}{\widehat{p}_a(\boldsymbol{X})} \widehat{Q}_a(\boldsymbol{X}),$$

where $\widehat{p}_a(\boldsymbol{X})$ and $\widehat{Q}_a(\boldsymbol{X})$ are estimators for the nuisance parameters $p_a(\boldsymbol{X}) = P(A = a \mid \boldsymbol{X})$ and $Q_a(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X}, A = a)$, respectively. We assume that there exists a constant $c > 0$ such that $c < p_a(\boldsymbol{X}), \widehat{p}_a(\boldsymbol{X}) < 1 - c$. Let $\bar{p}_a$ and $\bar{Q}_a$ be the point-wise limit of $\widehat{p}_a$ and $\widehat{Q}_a$. Define

$$\overline{W}_a(Y, \boldsymbol{X}, A; \bar{p}, \bar{Q}) = \frac{Y 1\{A = a\}}{\bar{p}_a(\boldsymbol{X})} + \frac{1\{A = a\} - \bar{p}_a(\boldsymbol{X})}{\bar{p}_a(\boldsymbol{X})} \bar{Q}_a(\boldsymbol{X}).$$

Consider the class of linear decision rules, we can minimize

$$\widehat{L}_\phi^{\lambda_n}(\boldsymbol{\beta}; \widehat{W}_1, \widehat{W}_{-1}) = \widehat{\mathbb{E}}_n \left[\sum_{a \in \{-1,1\}} \widehat{W}_a(Y, \boldsymbol{X}, A; \widehat{p}, \widehat{Q}) \phi(\boldsymbol{a} X^\top \boldsymbol{\beta}) \right] + \lambda_n \|\boldsymbol{\beta}\|_1. \tag{5}$$

The target of the inference procedure is a low-dimensional projection of $\boldsymbol{\beta}_\phi^*$ defined as the minimizer of

$$L_\phi(\boldsymbol{\beta}; \overline{W}_1, \overline{W}_{-1}) = \mathbb{E}\left[\overline{W}_1(\boldsymbol{X}, A; \bar{p}, \bar{Q}) \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) + \overline{W}_{-1}(\boldsymbol{X}, A; \bar{p}, \bar{Q}) \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \right].$$

Furthermore, if either $\bar{p}_a(\boldsymbol{X}) = p_a(\boldsymbol{X})$ or $\bar{Q}_a(\boldsymbol{X}) = Q_a(\boldsymbol{X})$, then $\mathbb{E}[\overline{W}_a \mid \boldsymbol{X}] = W_a(\boldsymbol{X}) \equiv Q_a(\boldsymbol{X})$, and the above objective function is equivalent to

$$L_\phi(\boldsymbol{\beta}; W_1, W_{-1}) = \mathbb{E}\left[W_1 \phi(\boldsymbol{X}^\top \boldsymbol{\beta}) + W_{-1} \phi(-\boldsymbol{X}^\top \boldsymbol{\beta}) \right].$$

In the general classification framework where nuisance parameters are involved, non-parametric or machine learning algorithms are commonly employed to fit them to avoid

model misspecification. However, the convergence rates of the $\widehat{W}_a$'s may be slower than $O_p(n^{-1/2})$. We adopt a cross-fitting procedure (Chernozhukov et al., 2017) to tackle this issue. We split the entire dataset into two halves. The first half is used to fit the nuisance parameters, and to estimate the weights, $\widehat{W}_a$. We then implement Algorithm 1 on the second half of the data to obtain the estimated coefficients and p-values. Similarly, we can then fit the nuisance parameters on the second half and use the first half to estimate the coefficients and conduct inference. Finally, to compensate for the efficiency loss due to the cross-splitting, we can average the estimates and the kernel-smoothed decorrelated scores. The details of this algorithm can be found in Algorithm 2.

Algorithm 2: Inference of $\boldsymbol{\beta}^*_{\phi,1}$ with nuisance parameters.

Input: A random seed; n samples; a positive integer K.

Output: A p-value for $\mathcal{H}_0 : \beta^*_{\phi,l} = 0$.

Randomly split data into halves $\widetilde{I}$ and $\widetilde{J}$ with equal sizes;

Estimate nuisance parameters using data in $\widetilde{I}$ by kernel regression after variable screening, and construct the estimated weights $\widehat{W}_a^{(\widetilde{I})}$'s on the samples in $\widetilde{J}$ using the estimated nuisance parameters;

On $\widetilde{J}$, we implement Algorithm 1 with weights $\widehat{W}_a^{(\widetilde{I})}$'s and denote the kernel-smoothed decorrelated score and its variance estimate as $S^{(\widetilde{J})}_{\widetilde{\phi}',null(l)}$ and $\widehat{\sigma}^2_{(\widetilde{J}),l}$;

Similarly, we can obtain $S^{(\widetilde{I})}_{\widetilde{\phi}',null(l)}$ and $\widehat{\sigma}^2_{(\widetilde{I}),l}$. Aggregate them by

$$S_{\widetilde{\phi}',null(l)} = \left(S^{(\widetilde{I})}_{\widetilde{\phi}',null(l)} + S^{(\widetilde{J})}_{\widetilde{\phi}',null(l)}\right)/2, \quad \widehat{\sigma}^2_l = \left(\widehat{\sigma}^2_{(\widetilde{I}),l} + \widehat{\sigma}^2_{(\widetilde{J}),l}\right)/2.$$

Calculate the p-value by $2\left(1 - \Phi(n^{-1/2}|S_{\widetilde{\phi}',null(l)}|/\widehat{\sigma}_l)\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

4. Theoretical properties

In this section, we investigate the asymptotic properties of the proposed procedures under an ERM with non-differentiable loss, potentially involving nuisance parameters. We focus on the uniform validity of the proposed procedures to test a low-dimensional hypothesis. For testing a hypothesis with a growing dimension, we propose a bootstrap procedure and prove its validity in the supplementary material. In the supplementary material, we also provide the convergence rates of $\widehat{\boldsymbol{\beta}}_\phi$ and $\widehat{\boldsymbol{w}}_{\phi,l}$.

4.1 Asymptotic properties without nuisance parameters

First, we consider the situation without nuisance parameters. We assume the following conditions hold on each split dataset in the sample-splitting (and cross-fitting) procedure. For notation simplicity, we omit the subscript indicating the split dataset being used.

(a) The design matrix is bounded, i.e., $\|\boldsymbol{X}\|_\infty \leqslant M$ with probability 1; there is a constant C such that $|\boldsymbol{x}^\top \boldsymbol{\beta}_\phi^*| \leqslant C$, and a constant $c > 0$ such that $|\beta_{\phi,j_0}^*| \geqslant c$ for some index j_0. Let $f_{x_{j_0}|\boldsymbol{x}_{-j_0}}(x_{j_0}, a)$ be the conditional density function of X_{j_0} given $\boldsymbol{X}_{-j_0}$ and A. We assume that $f'_{x_{j_0}|\boldsymbol{x}_{-j_0},a}(x_{j_0})$ and $f''_{x_{j_0}|\boldsymbol{x}_{-j_0},a}(x_{j_0})$ are bounded for both $a = 1$ and -1.

(b) There exists a positive constant γ such that for all $t_0 > t > 0$,

$$\sup_{j,a\in\{-1,1\}} \mathbb{P}(|t_j - a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*| \leqslant t) \leqslant \tau t^\gamma,$$

where τ and t_0 are some constants.

(c) We assume that the eigenvalues of

$$\mathbb{E}\left[\left(\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{A\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right)^2 \boldsymbol{X}\boldsymbol{X}^\top\right]$$

is bounded away from $+\infty$ and 0 by some constants.

Condition (a) assumes bounded designs, a common condition in high-dimensional literature (Ning and Liu, 2017; Van de Geer et al., 2014; Dezeure et al., 2017). In addition, Condition (a) also assumes $\boldsymbol{\beta}_\phi^* \neq 0$ and some regularity conditions on the conditional density function of the covariates; these conditions are firstly introduced in Koo et al. (2008) and then adopted in Peng et al. (2016); Wang et al. (2019) to ensure that the hessian of the $L_\phi(\boldsymbol{\beta})$ is well defined and continuous in $\boldsymbol{\beta}$. Condition (b) assumes that samples do not concentrate on the jump discontinuity points. This is satisfied when at least one covariate with a non-zero coefficient is continuous and has a bounded density function. Condition (b) ensures the L-2 convergence of the $1\{t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi \geqslant 0\}$ to $1\{t_j - a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \geqslant 0\}$. Condition (c) guarantees the uniform convergence of the variance estimator $\widehat{\sigma}_l$.

Now, we provide the uniform validity of the kernel-smoothed decorrelated score under the null hypothesis.

Theorem 5 *Denote $s' = \max_l \|\boldsymbol{w}_{\phi,l}^*\|_0$. Assume that $\|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}_\phi^*\|_2 \leqslant \Delta_{\beta,2}$ with probability approaching to 1, $s'\sqrt{\log p/(nh_{\text{gb}})} = o(1)$ and $\max_l |\boldsymbol{x}_{-l}^\top \boldsymbol{w}_{\phi,l}^*|$ is bounded by $R = o(n^{1/6})$. Taking $\mu_n \asymp \delta_n + Rh_{\text{gb}}^2 + R\Delta_{\beta,2}$, where $\delta_n = R(\log p/(nh_{\text{gb}}))^{1/2}$. Further assume that $\sqrt{n}R(h_{\text{lo}} + \sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)}) = o(1)$, $\sqrt{n}(s'\mu_n)(\sqrt{\log p/n} + h_{\text{lo}} + \Delta_{\beta,2}) = o(1)$, and $(Rs'\mu_n + R^2\sqrt{\log p/n} + R^2\Delta_{\beta,2} + R^2 h_{\text{lo}})\sqrt{\log p} = o(1)$. If Conditions (a) - (c) are satisfied, under the null hypothesis, we have*

$$\max_{l\in\mathcal{H}_0} \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1} S_{\widetilde{\phi}',null(l)}\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) - (1-\alpha)\right| = o_p(1),$$

where $\mathcal{H}_0$ is the index set of zero coefficients in $\boldsymbol{\beta}_\phi^*$ under the null hypothesis.

Define

$$\sigma_l^2 = \mathbb{E}\left[\left(\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{A\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right)^2 (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)^2\right].$$

Theorem 5 implies that the asymptotic variance σ_l^2 can be estimated by $\widehat{\sigma}_l^2$ in Algorithm 1.

The following corollary provides the uniform validity of the confidence interval constructed through the one-step debiased estimator $\widetilde{\beta}_{\phi,l}$.

Corollary 6 *Assume the same conditions in Theorem 5, we have*

$$\max_l \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1}\widehat{I}_l\left(\widetilde{\beta}_{\phi,l}-\beta^*_{\phi,l}\right)\right|\leqslant \Phi^{-1}(1-\alpha/2)\right)-(1-\alpha)\right| = o_p(1).$$

Define

$$\widetilde{\sigma}_l^2 \;=\; \sigma_l^2/I_l^2, \quad I_l^2=\mathbb{E}\left[\sum_j \Delta_j\delta(t_j-A\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi)(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})^2\right].$$

Corollary 6 implies that the asymptotic variance $\widetilde{\sigma}_l^2$ can be estimated by $\widehat{\sigma}_l^2/\widehat{I}_l^2$.

Compared with the theoretical conditions for the marginal validity of the decorrelated score under a differentiable strictly convex loss function without nuisance parameters, our conditions for the uniform validity under a non-differentiable loss function assumes a more sparse model. To see this, Ning and Liu (2017) show that the condition required for the marginal validity of the decorrelated score is $\max\{s',s^*\}\log p/\sqrt{n}\to 0$; this is equivalent to $\sqrt{n}\|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2^2\to 0$ and $\sqrt{n}\Delta_{\beta,2}^2\to 0$. For the hinge loss, Theorem 5 requires that $\sqrt{n}s'\mu_n\Delta_{\beta,2}\to 0$ in addition to $\sqrt{n}\Delta_{\beta,2}^2\to 0$; this is equivalent to $\sqrt{n}\|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_1\Delta_{\beta,2}\to 0$ and $\sqrt{n}\Delta_{\beta,2}^2\to 0$ (see Appendix for the convergence rate of $\widehat{\boldsymbol{w}}$). If $\|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2^2\lesssim \|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_1\Delta_{\beta,2}$, the conditions in Theorem 5 and 7 are more restrictive than Ning and Liu (2017). In Appendix (the Proof of Theorem 7), we show that under a dedicated sample-splitting algorithm, we can reduce this requirement to a weaker condition than Ning and Liu (2017). With this sample-splitting algorithm, we only require that $\sqrt{n}\|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2\Delta_{\beta,2}\to 0$ in addition to $\sqrt{n}\Delta_{\beta,2}^2\to 0$. The requirements that $\sqrt{n}\|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2\Delta_{\beta,2}\to 0$ and $\sqrt{n}\Delta_{\beta,2}^2\to 0$ are weaker than $\sqrt{n}\|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2^2\to 0$ and $\sqrt{n}\Delta_{\beta,2}^2\to 0$, when $\Delta_{\beta,2}\lesssim \|\widehat{\boldsymbol{w}}_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2$. However, the dedicated sample-splitting algorithm leads to a significant increase in computation time. Hence, we mainly focus on the proposed procedure in the paper.

4.2 Asymptotic properties with nuisance parameters

In this section, we investigate the theoretical property of the proposed inference procedure when nuisance parameters exist.

(d) There is a constant C such that $\max\{W_1,W_{-1}\}\leqslant C$. $\partial_{x_{j_0}}W_a(\boldsymbol{x})$ and $\partial^2_{x_{j_0}}W_a(\boldsymbol{x})$ are bounded.

(e) There exist positive constants η and ζ such that

$$\begin{aligned}\sup_{\boldsymbol{x},a,y}\left|\widehat{W}_a-\overline{W}_a\right| &= O_p(n^{-\zeta}),\\ \sup_{\boldsymbol{x}}\left|\mathbb{E}\left[\widehat{W}_a\mid \boldsymbol{X}=\boldsymbol{x}\right]-W_a\right| &= O_p(n^{-\eta}),\end{aligned}$$

where $\overline{W}_a$ is the point-wise limit of $\widehat{W}_a$ as $n\to\infty$.

Condition (d) assumes that the weights are bounded and smooth. Condition (e) assumes that the convergence rates of the nuisance parameters are upper bounded by $O_p(n^{-\zeta})$; the expectation of the weights with estimated nuisance parameters approximates $W_a(\boldsymbol{X})$ faster than $O_p(n^{-\eta})$.

We now provide the uniform validity of the kernel-smoothed decorrelated score under the null hypothesis when there exist nuisance parameters.

Theorem 7 *Denote* $s' = \max_l \|\boldsymbol{w}^*_{\phi,l}\|_0$ *and* $\delta_n = R(\log p/(nh_{\rm gb}))^{1/2} + Rn^{-\eta}/h_{\rm gb}$. *Assume that* $\|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\|_2 \leqslant \Delta_{\beta,2}$ *with probability approaching to* 1, $s'\sqrt{\log p/(nh_{\rm gb})} = o(1)$ *and* $\max_l |\boldsymbol{x}_{-1}^\top \boldsymbol{w}^*_{\phi,l}|$ *is bounded by* $R = o(n^{1/6})$. *Taking* $\mu_n \asymp \delta_n + Rh^2_{\rm gb} + R\Delta_{\beta,2}$. *Further assume that* $\sqrt{n}R(h_{\rm lo}+n^{-\eta}+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)}) = o(1)$, $\sqrt{n}(s'\mu_n)(n^{-\eta}+\sqrt{\log p/n}+h_{\rm lo}+\Delta_{\beta,2}) = o(1)$, *and* $(Rs'\mu_n + R^2\sqrt{\log p/n} + R^2 n^{-\zeta} + R^2\Delta_{\beta,2} + R^2 h_{\rm lo})\sqrt{\log p} = o(1)$. *If Conditions (a) - (e), under the null hypothesis, we have*

$$\max_{l\in\mathcal{H}_0}\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1}S_{\widetilde{\phi}',null(l)}\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) - (1-\alpha)\right| = o_p(1),$$

where $\mathcal{H}_0$ *is the index set of zero coefficients in* $\boldsymbol{\beta}^*_\phi$ *under the null hypothesis and*

$$\sigma_l^2 \;=\; \mathbb{E}\left[\left\{\sum_a aW_a\left(\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j \geqslant 0\right\}\right)\right\}^2 (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})^2\right].$$

Compared with Theorem 5, the δ_n also involves the $Rn^{-\eta}/h_{\rm gb}$ due to the additional nuisance parameters.

Corollary 8 *Assume the same conditions in Theorem 7, we have*

$$\max_l \sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1}\widehat{I}_l\left(\widetilde{\beta}_{\phi,l} - \beta^*_{\phi,l}\right)\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) - (1-\alpha)\right| = o_p(1).$$

Define

$$\widetilde{\sigma}_l^2 \;=\; \sigma_l^2/I_l^2, \quad I_l^2 = \mathbb{E}\left[\sum_a\sum_j \Delta_j W_a \delta(t_j - a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})^2\right].$$

Theorem 7 and Corollary 8 imply that the asymptotic variance component σ_l^2 and σ_l^2/I_l^2 can be estimated by $\widehat{\sigma}_l^2$ and $\widehat{\sigma}_l^2/\widehat{I}_l^2$, respectively. They require that $R^2n^{-\zeta}\sqrt{\log p} = o(1)$, where $n^{-\zeta}$ corresponds to the slowest convergence rate of the nuisance parameters. This ensures that the $\widehat{I}_l$ converges fast enough to construct a uniformly valid testing procedure/debiased estimator. Furthermore, compared with Theorem 5, Theorem 7 also requires that $\sqrt{n}Rn^{-\eta} \to 0$ and $\sqrt{n}(s'\mu_n)n^{-\eta} \to 0$. These requirements ensure that the uncertainty of estimating the nuisance parameters is asymptotically ignorable, and does not affect the asymptotic variance. Under a doubly robust formulation, these requirements are not restrictive and can be satisfied even when the nuisance parameter estimates have a slow convergence rate (see below for concrete examples).

In Theorem 7 and Corollary 8, we require that $\sqrt{n}Rn^{-\eta} \to 0$. In classification with missing labels example, assume that there exists $\alpha, \beta > 0$ such that $\sup_{\boldsymbol{x}} |\widehat{\pi}_1(\boldsymbol{x}) - \bar{\pi}(\boldsymbol{x})| = O_p(n^{-\alpha})$ and $\sup_{\boldsymbol{x}} |\widehat{p}_a(\boldsymbol{x}) - \bar{p}_a(\boldsymbol{x})| = O_p(n^{-\beta})$. When we assume that both $\bar{\pi}_1 = \pi_1$ and $\bar{p}_a = p_a$, the condition that $\sqrt{n}Rn^{-\eta} \to 0$ holds with $\eta = \alpha + \beta$ if $\sqrt{n}Rn^{-\alpha-\beta} \to 0$, since

$$\sup_{\boldsymbol{x}} \left| \mathbb{E}[\widehat{W}_a(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) \mid \boldsymbol{X} = \boldsymbol{x}] - W_a(\boldsymbol{x}) \right| = O_p(n^{-\alpha-\beta}).$$

When the missing mechanism is known, i.e., $\widehat{\pi}_1 = \pi_1$, the condition that $\sqrt{n}Rn^{-\eta} \to 0$ automatically holds since

$$\sup_{\boldsymbol{x}} \left| \mathbb{E}[\widehat{W}_a(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p}) \mid \boldsymbol{X} = \boldsymbol{x}] - W_a(\boldsymbol{x}) \right| = 0,$$

for any $\widehat{p}_a$.

In the inference of individualized treatment rule example, assume that there exists $\alpha, \beta > 0$ such that $\sup_{\boldsymbol{x}} |\widehat{p}_a(\boldsymbol{x}) - \bar{p}_a(\boldsymbol{x})| = O_p(n^{-\alpha})$ and $\sup_{\boldsymbol{x}} |\widehat{Q}_a(\boldsymbol{x}) - \bar{Q}_a(\boldsymbol{x})| = O_p(n^{-\beta})$. When we assume that both $\bar{p}_a = p_a$ and $\bar{Q}_a = Q_a$, the condition $\sqrt{n}Rn^{-\eta} \to 0$ holds with $\eta = \alpha + \beta$ if $\sqrt{n}Rn^{-\alpha-\beta} \to 0$ since

$$\sup_{\boldsymbol{x}} \left| \mathbb{E}[\widehat{W}_a(\boldsymbol{X}, A; \widehat{p}, \widehat{Q}) \mid \boldsymbol{X} = \boldsymbol{x}] - W_a(\boldsymbol{x}) \right| = O_p(n^{-\alpha-\beta}).$$

When the treatment assignment mechanism is known, i.e., $\widehat{p}_a = p_a$ is known (for example, in randomized clinical trials), the condition that $\sqrt{n}Rn^{-\eta} \to 0$ automatically holds since

$$\sup_{\boldsymbol{x}} \left| \mathbb{E}[\widehat{W}_a(\boldsymbol{X}, A; \widehat{p}, \widehat{Q}) \mid \boldsymbol{X} = \boldsymbol{x}] - W_a(\boldsymbol{x}) \right| = 0,$$

for any $\widehat{Q}_a$.

With a bounded R, requirement on $\sqrt{n}Rn^{-\alpha-\beta} \to 0$ is equivalent to $\alpha + \beta > 1/2$, which can be satisfied by many estimation methods. For example, in the inference of the individualized treatment rule example, if we adopt generalized linear models with lasso penalties to estimate the propensity score and the outcome regression models, it is satisfied when $\widetilde{s} \log p / \sqrt{n} = o(1)$, where $\widetilde{s}$ is an upper bound of the number of the non-zero coefficients in the propensity score and the outcome regression models. If the propensity score is estimated by a regression spline estimator and is known to be p_π-dimensional (low-dimensional) by design, we have $\alpha = 1/3$ if π belongs to the Hölder class with a smoothness parameter greater than $5p_\pi$ (Newey, 1997). In this case, we only need $\beta > 1/6$.

5. Simulation

In this section, we conduct simulation studies to examine the performance of the proposed inference procedure. We consider two simulation scenarios: classification without nuisance parameters (see Section 2.1) and estimating ITR with nuisance parameters (see Section 3). Specifically, the population is a mixture of two subgroups with a probability 0.4 from Group I and probability 0.6 from Group II. For Group I, the covariate vector is generated from $N(\xi\boldsymbol{\mu}_0, I_{p\times p} - 0.1\boldsymbol{e}_1\boldsymbol{e}_1^\top)$; for Group II, the covariate vector is generated from $N(\xi\boldsymbol{\mu}_1, I_{p\times p})$, where $\mu_0 = (-1, 1, -0.5, 0.5, 0, \cdots, 0)^\top$ and $\mu_1 = (1, -1, -1, -1, 0, \cdots, 0)^\top$.

I. The label for Group I is $A = 1$, and the label for Group II is $A = -1$. The goal is to classify Group I from Group II based on the data;

II. The treatment assignment mechanism follows $P(A = 1 \mid \boldsymbol{X}) = \exp(0.25 \times (X_1^2 + X_2^2 + X_1X_2)/(1 + 0.25 \times (X_1^2 + X_2^2 + X_1X_2))$. The observed outcome $Y = Y(a)$ if the treatment a is assigned to the patient, $Y(a) = (\boldsymbol{X}^\top \boldsymbol{\gamma})^2 + C(\boldsymbol{X}; G)I(a = 1) + \epsilon$, where $\gamma = (-0.4, -0.4, 0.4, -0.4, 0, \cdots, 0)^\top$ and ϵ follows a standard normal distribution. Here, $C(\boldsymbol{X}; G = 1) = |X_1| + 0.5$ for Group I patients and $C(\boldsymbol{X}; G = 2) = -(|X_1| + 0.5)$ for Group II patients. The goal is to estimate the optimal individualized treatment rules that maximize the outcome. The CATE is given by $\mathbb{E}[C(\boldsymbol{X}; G) \mid \boldsymbol{X}]$; and thus, by calculation, the sign of CATE is the same as $\mathbb{P}(G = 1 \mid \boldsymbol{X}) - 0.5$.

Our Scenario II assumes that there are two subgroups with different treatment effects, and the group membership G is unobserved. Similar setups have been adopted in recent literature for heterogeneous subgroup-level treatment effects (Lin et al., 2021; Chandra et al., 2023). For each scenario, the parameter ξ controls the magnitude of the overlapping of two subgroups. Two subgroups are easier to separate by a linear decision rule for a larger ξ. We gradually increase ξ from 0.1 to 1 (by 0.1) and the dimension of the covariate p from 500 to 1600, which lead to 30 settings in total for each scenario.

We compare our proposed method with an ad-hoc method and the method in Liang et al. (2022) using a logistic loss. For the ad-hoc approach, we first use the hinge loss with the lasso penalty to identify covariates with non-zero coefficient estimates. Then, we refit the hinge loss without any penalty using the identified covariates and first 8 covariates, and employ the inference procedure for low-dimensional settings (Koo et al., 2008) to construct test statistics and confidence intervals for identified and first 8 covariates. In Scenario II, we further combine the ad-hoc method with the cross-fitting algorithm proposed in Chernozhukov et al. (2017) as the competitor. For Scenario II, we use the kernel regression after the variable screening to estimate nuisance parameters for both the proposed and ad-hoc methods. For each simulation setting, we repeatedly simulate the training samples 500 times, the sample sizes of which vary from 500 to 1600. To evaluate the inference procedure, we report type I error (testing $\mathcal{H}_0 : \beta^*_{\phi,l} = 0$, where $l = 5, 6, 7, 8$), power (testing $\mathcal{H}_0 : \beta^*_{\phi,l} = 0$, where $l = 1, 2, 3, 4$), the averaged coverage of the 95% confidence intervals for the eight coordinates in $\boldsymbol{\beta}^*_\phi$, and the averaged length of the 95% confidence intervals. The true value of $\boldsymbol{\beta}^*_\phi$ is determined by the average over 500 replicates with $n = 2500$. The coordinates of $\boldsymbol{\beta}^*_\phi$ are set to zero if the absolute value of its average estimates is less than 0.01. When comparing interval length with the method using a logistic loss, we normalize the interval length based on the first coordinate of $\boldsymbol{\beta}^*_\phi$. In addition, we also report the classification accuracy (for Scenario I), value function (for Scenario II), and estimation errors of debiased one-step estimator (against $\boldsymbol{\beta}^*_\phi$ under hinge loss) to compare the methods using a hinge loss vs. a logistic loss.

Figures 1 - 3 show the simulation results for Scenario I. The results show that the proposed method has controlled Type-I errors and higher powers than the ad-hoc method. The ad-hoc method has inflated type-I errors, which indicate that the decorrelation and sample-splitting procedures to construct valid test statistics are necessary. From Figure 2, the proposed method achieves nominal coverages and has shorter confidence intervals than the ad-hoc method across all settings. When comparing with the method using a logistic

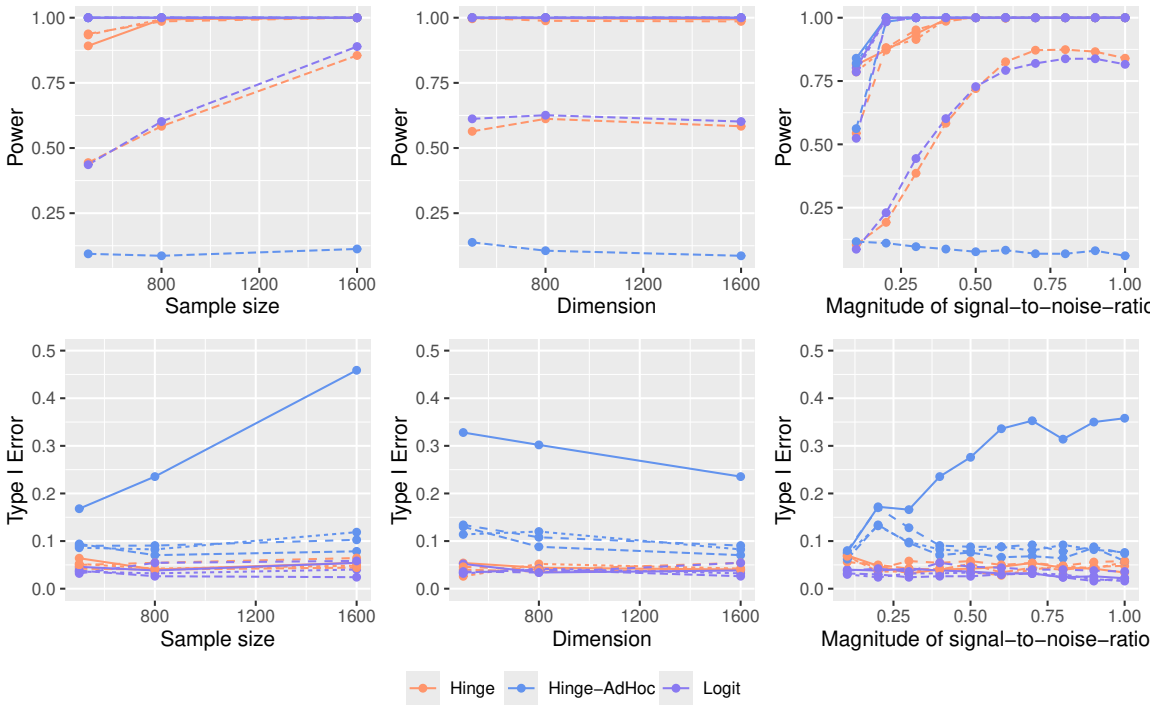

Figure 1: Testing results for Scenario I with the change of sample size when $\xi = 0.4$ and $p = 800$, the change of ξ ($n = 800$; $p = 1600$) and the change of p ($n = 800$; $\xi = 0.4$). Line styles represent different coefficients.

loss, our method yields slightly lower power but comparable or higher accuracy, as shown in Figure 3. We also observe that the averaged coverage shows a non-monotonic pattern for our proposed method. When ξ changes, the value of $\boldsymbol{\beta}_\phi^*$ may also change. The varying $\boldsymbol{\beta}_\phi^*$' may contribute to this non-monotonic pattern. The simulation results for Scenario II are presented in Figures 4 - 6. Again, the proposed method yields better results in terms of Type I error control, coverage, and the length of confidence intervals. The proposed method also achieves a higher value compared with the method using a logistic loss. Results of more simulation settings can be found in Appendix D.

6. Real data examples

In this section, we apply our proposed inference approach in real world problems. Specifically, we consider two scientific questions: 1) whether we can identify the risk factors associated with uncontrolled HbA1c in a year, given the patients baseline characteristics; 2) whether we can identify driving factors to inform better treatment strategies.

The data we used comes from the electronic medical records linked with Medicare claims data on Type-II diabetic patients with complex commodities. These data are collected through the Heath Innovation Program at the University of Wisconsin. It contains $n = 9101$ patients with many covariate information. In order to answer the research questions high-

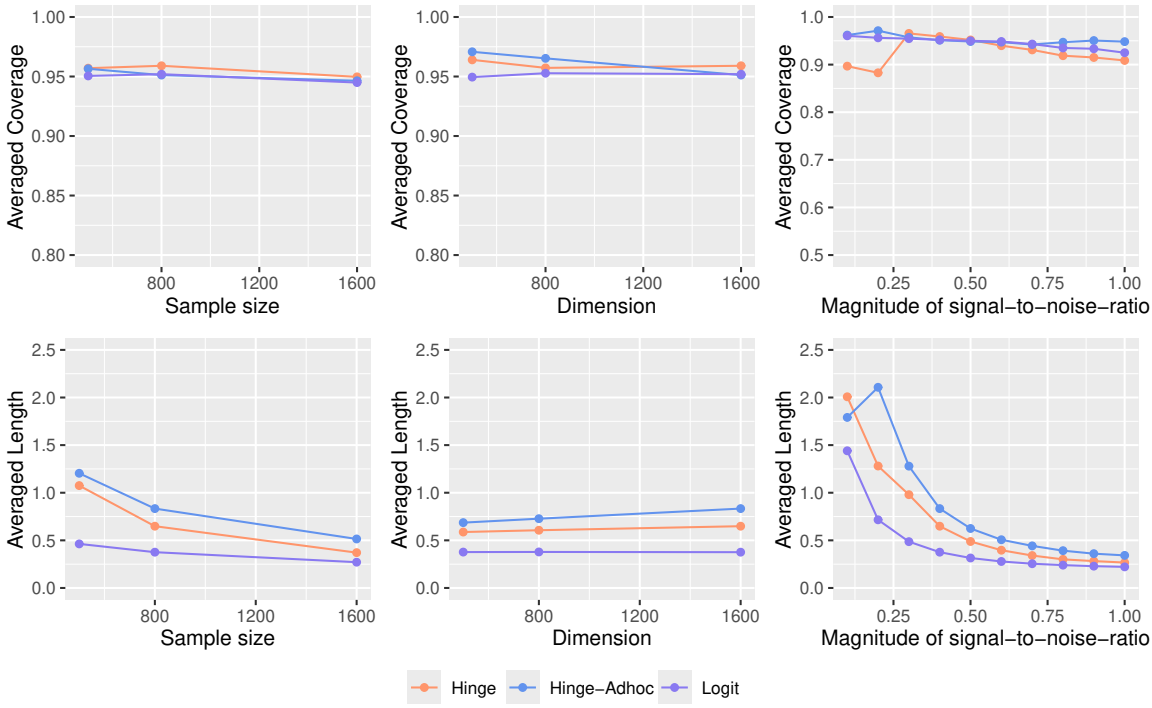

Figure 2: Coverage results for Scenario I with the change of sample size when $\xi = 0.4$ and $p = 800$, the change of ξ ($n = 800$; $p = 1600$) and the change of p ($n = 800$; $\xi = 0.4$).

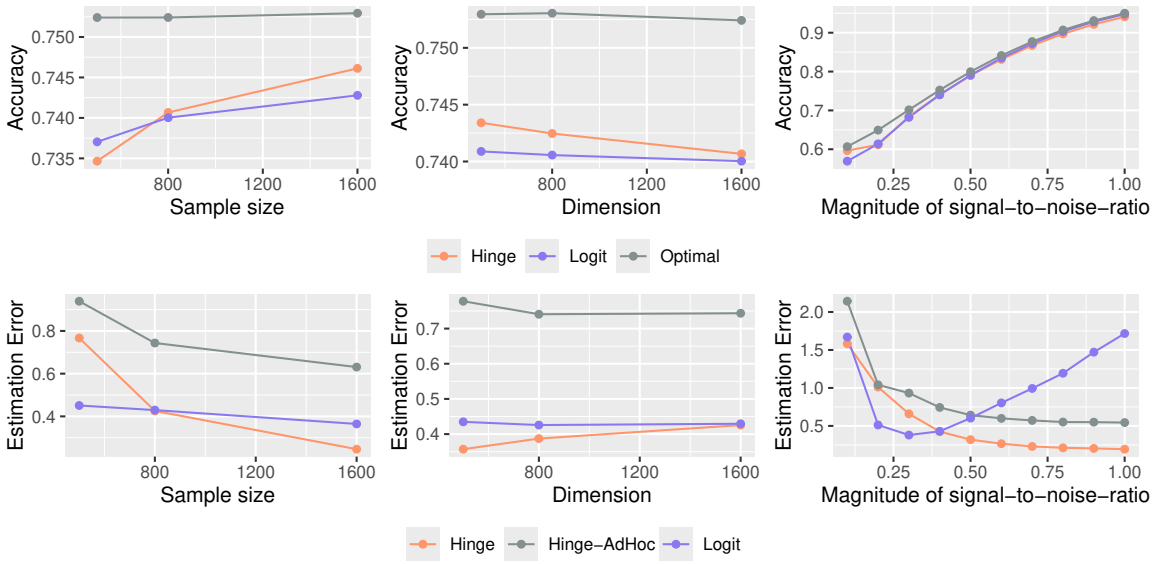

Figure 3: Classification accuracy and estimation error for Scenario I with the change of sample size when $\xi = 0.4$ and $p = 800$, the change of ξ ($n = 800$; $p = 1600$) and the change of p ($n = 800$; $\xi = 0.4$).

lighted above, we include 40 covariates including sociodemographic variables, disease history, baseline HbA1c levels, etc., and their first-order interactions in our analysis. As a variable screening procedure, we rank these covariates and their interactions by the variances and select $p = 120$ covariates with highest variances.The outcome of interest in both questions is whether the patient successfully controlled his or her HbA1c below 8% after one year.

6.1 Identify risk factors associated with uncontrolled HbA1c for Type-II diabetic patients

Our goal is to identify patients and risk factors associated with uncontrolled HbA1c after one-year of follow-up if following current clinical guideline. This can be considered as a classification problem. The label to predict, denoted as A, indicates whether patients have uncontrolled HbA1c after one-year of follow-up. Specifically, we set $A = 1$ if patients successfully control the HbA1c level after one-year follow-up; and set $A = -1$, otherwise.

We first use the linear SVM with a lasso penalty to estimate the decision rule and then conduct inference on the estimated decision rule. Under a cross-validation procedure, the estimated decision rule achieves a prediction accuracy of 0.895 with a standard deviation of 0.003. After controlling the false discovery rate (FDR) at 0.05 by Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), we find that patients with valvular disease are more likely to suffer from uncontrolled HbA1c's (see Table 1); minorities with hypertension are more likely to have HbA1c under control after treatment, which is probably due to the full consideration for patients with hypertension as one common comorbidity in cur-

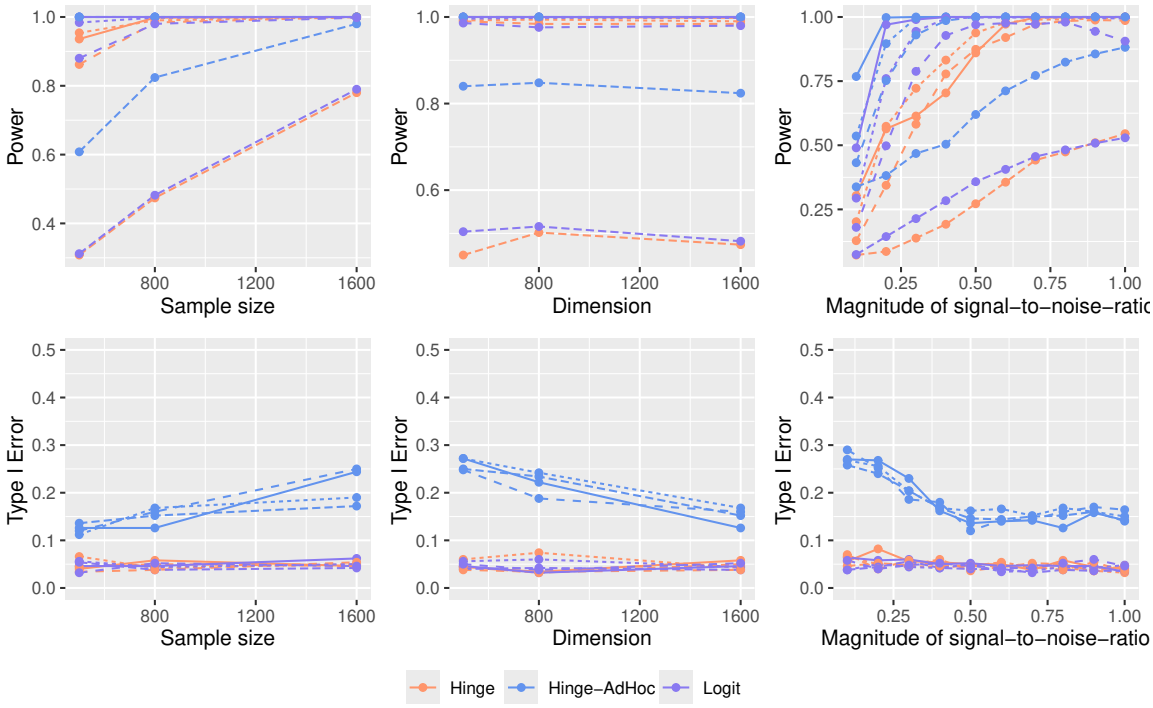

Figure 4: Testing results for Scenario II with the change of sample size when $\xi = 0.8$ and $p = 800$, the change of ξ ($n = 800$; $p = 1600$) and the change of p ($n = 800$; $\xi = 0.8$). Line styles represent different coefficients.

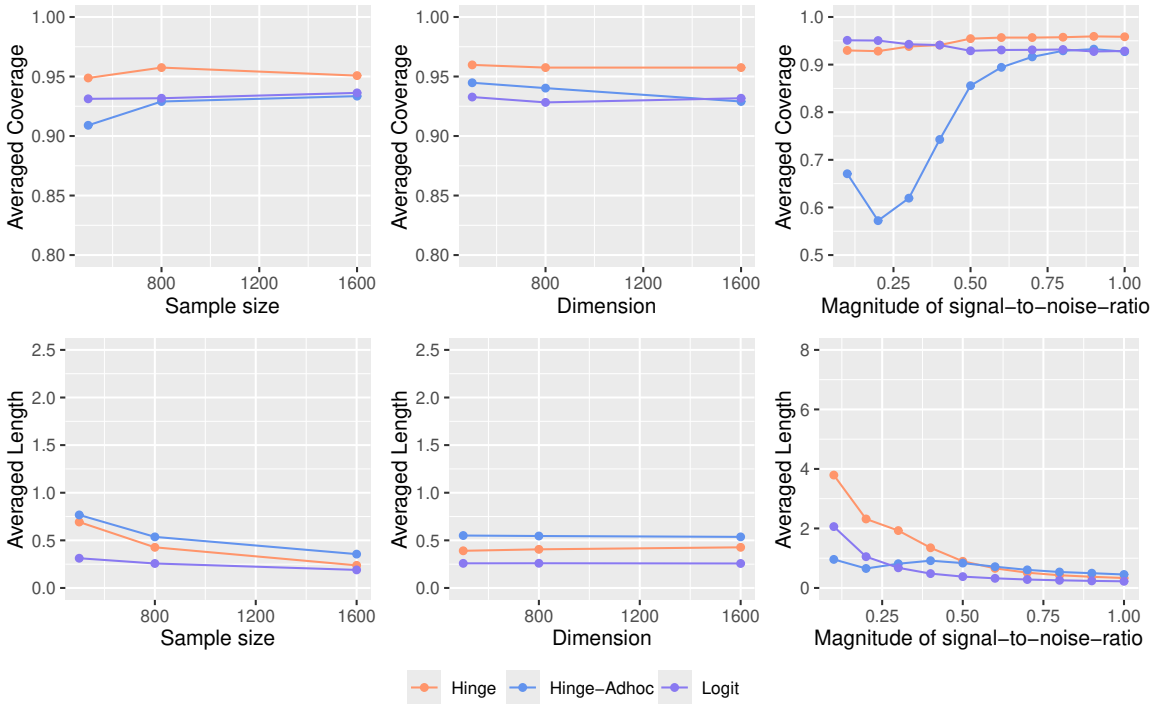

Figure 5: Coverage results for Scenario II with the change of sample size when $\xi = 0.8$ and $p = 800$, the change of ξ ($n = 800$; $p = 1600$) and the change of p ($n = 800$; $\xi = 0.8$).

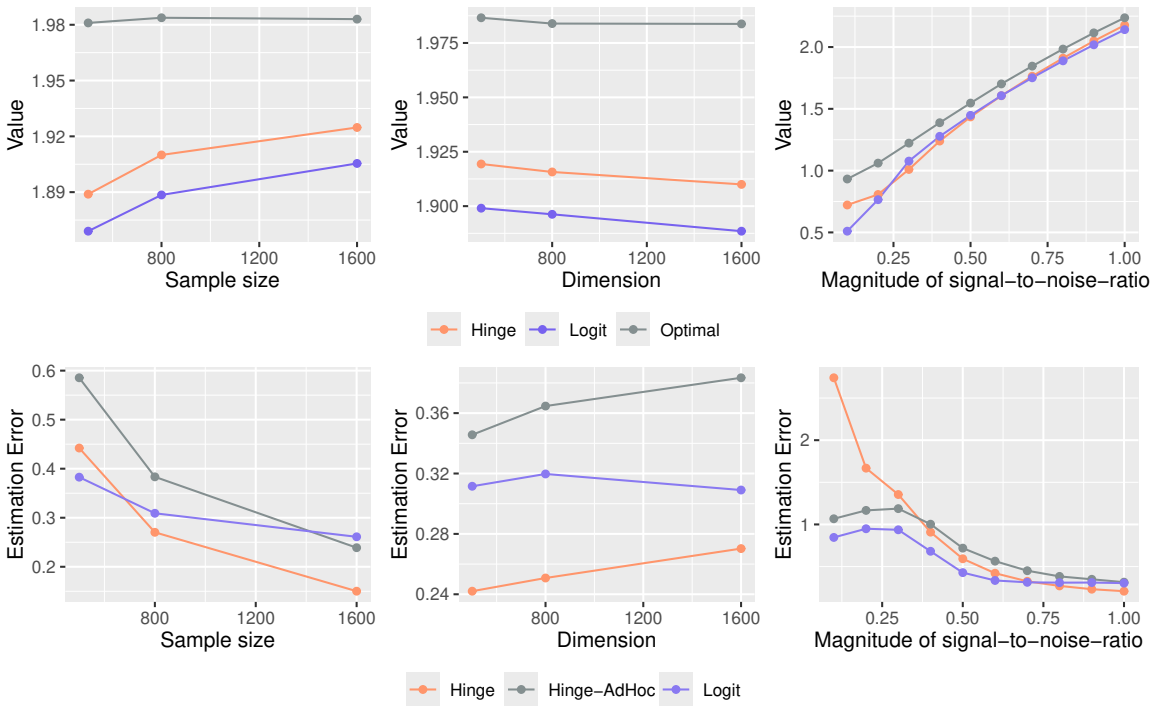

Figure 6: Value function and estimation error for Scenario II with the change of sample size when $\xi = 0.8$ and $p = 800$, the change of ξ ($n = 800$; $p = 1600$) and the change of p ($n = 800$; $\xi = 0.8$).

Table 1: Coefficients and p-value for the identified significant covariates of the estimated decision rule after FDR control.

Covariate	Coef	P-value	95% - CI
Valvular disease (Yes)	-2.079	1.528×10^{-99}	[-2.118, -2.041]
Hypertension (Yes) : Race (other)	2.126	3.687×10^{-88}	[2.091, 2.161]

rent clinical guideline. After controlling for the FDR, the ad-hoc method reveals that 112 covariates out of 120 covariates are significant. However, when evaluating the length of the constructed confidence intervals, the average interval length for the proposed method is much shorter at 0.122 compared to that of the ad-hoc method, which is 0.244. The high number of identified risk factors and long length of confidence intervals may be attributed to biased test statistics and interval estimations, as discussed in Section 5.

6.2 Identify driving factors of the estimated ITR to inform better clinical guideline for Type-II diabetic patients

We aim to estimate the optimal ITR and identify the driving factors to inform future clinical guideline for Type-II diabetic Medicare patients. We consider two treatment options: hypoglycemic agents versus usual care. We set $A = 1$ if the patient received hypoglycemic agent at baseline and $A = -1$ otherwise. The outcome Y is whether the patient successfully controlled his or her HbA1c below 8% after one year. Under these specifications, the weighted classification problem proposed in Section 3 aims to estimate an ITR such that under the derived ITR, the chance of successfully controlling HbA1c below 8% after one year is maximized for each individual.

To demonstrate that the estimated ITR informs a better treatment strategy than the current clinical practice, we use the cross-validation procedure to calculate the chance of successfully controlling HbA1c if the estimated ITR were implemented. The results show that the current clinical practice has a success rate of 0.860 with a standard deviation of 0.008; the estimated ITR has a success rate of 0.871 with a standard deviation of 0.013. In terms of the identified driving factors of the estimated ITR, using the proposed inference method, after controlling FDR at 0.05, we find that patients having hypertension, especially with higher A1c levels at baseline, are more likely to benefit from hypoglycemic agents in controlling HbA1c, which confirmed the importance of hypertension in Type-II diabetes care. In addition, we also find that female patients with chronic complications, are more likely to benefit from hypoglycemic agents in controlling HbA1c, see Table 2. We also implement the ad-hoc method as a comparison. After controlling for the FDR, the ad-hoc method identifies 107 significant driving factors out of 120 predictors. The averaged length of confidence intervals for the ad-hoc method is 0.423, longer than that of the proposed method, which is 0.122.

Table 2: Coefficients and p-value for the identified significant covariates of the estimated optimal ITR after FDR control. Special chronic conditions refer to chronic conditions including amputation, chronic blood loss, drug abuse, lymphoma, metastatic cancer, and peptic ulcer disease.

Covariate	Coef	P-value	95% - CI
Hypertension (Yes)	-0.0451	8.32×10^{-4}	[-0.0918, 0.0016]
Chronic Complications (Yes) : Female	0.1244	4.51×10^{-6}	[0.0947, 0.154]
Hypertension (Yes) : Baseline A1c	0.0627	6.95×10^{-7}	[0.0205, 0.105]

7. Discussion

We propose a high-dimensional inference procedure for a non-differentiable convex loss function in a general classification framework, which can be utilized to discover the driving factors in decision making. In particular, combined with the cross-fitting algorithm, our procedure can accommodate weights involving additional nuisance parameters, which may be estimated via nonparametric or other machine learning algorithms.

There are multiple directions that could be further studied in the future. First, although we allow a non-differentiable surrogate loss, we still require the convexity of the surrogate loss. It would be interesting to see how the proposed method can be extended to deal with a non-convex surrogate loss or even zero-one loss without any relaxation. Second, we may consider distributed inference or online inference. When the sample size is large, the computation may be infeasible on a single machine due to limited resources. In this case, distributed inference can leverage the computation power of a cluster of machines and reduce the runtime. When the sample size is limited, it would be important to understand how to utilize newly collected samples to improve the learning process of the decision rule. Third, in this work, we only consider bounded designs; we may consider extending these results to sub-gaussian designs or designs with heavy tails.

Appendix A.

In this appendix, we provide the proof of the theorems in Section 4. To unify the theoretical analysis for different scenarios, we consider the following formulation to our problem. The target of the inference $\boldsymbol{\beta}_\phi^*$ is defined as

$$L_\phi(\boldsymbol{\beta}; W_1, W_{-1}) = \mathbb{E}\left[W_1\phi(\boldsymbol{X}^\top\boldsymbol{\beta}) + W_{-1}\phi(\boldsymbol{X}^\top\boldsymbol{\beta})\right],$$

where W_a's are the weights depending the problem solved. For each problem, the choice of the weights are not unique; as a simple example, we have that $L_\phi(\boldsymbol{\beta}; W_1, W_{-1}) = L_\phi(\boldsymbol{\beta}; \mathbb{E}[W_1 \mid \boldsymbol{X}], \mathbb{E}[W_{-1} \mid \boldsymbol{X}])$. To avoid ambiguity, we specify the choice of W_a's corresponds to 1) common classification problem; 2) classification with missing labels; 3) estimation of ITR. In common classification and classification with missing labels, the weight $W_a = p_a(\boldsymbol{X})$; in estimation of ITR, the weight $W_a = \mathbb{E}[Y(a) \mid \boldsymbol{X}]$. Details are discussed in Section 3.1 and 3.2 in the main text. Under these choices, the weight W_a's are functions of $\boldsymbol{X}$.

The proof is organized as follows. To start with, we discuss the approximation error of the local and global kernel functions. Then we provide the estimation error for $\boldsymbol{w}_{\phi,l}^*$ defined as the minimizer of

$$L_{\phi''}(\boldsymbol{w}) = \mathbb{E}\left[\sum_{j=1}^{J} \Delta_j \delta(A\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j)(X_1 - \boldsymbol{X}_{-l}^\top\boldsymbol{w})^2\right].$$

Finally, using these results, we derive the asymptotic property of the kernel-smoothed decorrelated score with the nuisance parameter under the null hypothesis. As a corollary, we also provide the asymptotic normality of the one-step debiased estimator.

Approximation error of local and global kernel functions

In this section, we discuss the approximation error of using the local and global kernel functions.

(C1) We assume that the design is bounded, i.e., $\|\boldsymbol{X}\|_\infty \leqslant M$ with probability 1; there is a constant C such that $\max\left\{W_1, W_{-1}, |\boldsymbol{x}^\top\boldsymbol{\beta}_\phi^*|\right\} \leqslant C$. Let $f_{x_1|\boldsymbol{X}_{-l}}(x_1)$ be the conditional density function of X_1 given $\boldsymbol{X}_{-l}$. We assume that $f'_{x_1|\boldsymbol{X}_{-l}}(x_1)$ and $f''_{x_1|\boldsymbol{X}_{-l}}(x_1)$ are bounded; $\partial_{x_1} W_a(\boldsymbol{x})$ and $\partial^2_{x_1} W_a(\boldsymbol{x})$ are bounded. We also assume that there exists a constant $c > 0$ such that $|\beta_{\phi,1}^*| \geqslant c$.

Under Condition (C1), we have the following lemmas.

Lemma 9 *We have*

$$\sup_j \left\|\mathbb{E}\left[\left\{W_1\widetilde{\phi}'\left(1 - \boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right) - W_{-1}\widetilde{\phi}'\left(1 + \boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)\right\}\boldsymbol{X}_j\right]\right\|_\infty = O(h_l),$$

where W_1 and W_{-1} are bounded functions of $\boldsymbol{X}$.

Proof of Lemma 9. To show that

$$\sup_j \left\| \mathbb{E}\left[\left\{ W_1 \widetilde{\phi}'\left(1 - \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) - W_{-1} \widetilde{\phi}'\left(1 + \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) \right\} \boldsymbol{X}_j \right] \right\|_\infty = O(h_l),$$

we will show that for any $\|\boldsymbol{v}\|_2 = 1$, we have that

$$\sup_{\boldsymbol{v}} \left| \mathbb{E}\left[\left\{ W_1 \widetilde{\phi}'\left(1 - \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) - W_{-1} \widetilde{\phi}'\left(1 + \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) \right\} \boldsymbol{X}^\top \boldsymbol{v} \right] \right| = O(h_l),$$

uniformly holds. Let us directly compute.

$$\begin{aligned} & \mathbb{E}\left[\left\{ W_1 \widetilde{\phi}'\left(1 - \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) - W_{-1} \widetilde{\phi}'\left(1 + \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) \right\} \boldsymbol{X}^\top \boldsymbol{v} \right] \\ = & \sum_j \Delta_j \mathbb{E}\left[\left\{ W_1 H\left(\frac{1 - \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j}{h_l} \right) - W_{-1} H\left(\frac{1 + \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j}{h_l} \right) \right\} \boldsymbol{X}^\top \boldsymbol{v} \right]. \end{aligned}$$

We will compute $\mathbb{E}\left[W_1 H\left(\frac{1 - \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j}{h_l} \right) \boldsymbol{X}^\top \boldsymbol{v} \right]$ and $\mathbb{E}\left[W_{-1} H\left(\frac{1 + \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j}{h_l} \right) \boldsymbol{X}^\top \boldsymbol{v} \right]$ separately.

$$\begin{aligned} & \mathbb{E}\left[W_1 H\left(\frac{1 - \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j}{h_l} \right) \boldsymbol{X}^\top \boldsymbol{v} \right] \\ = & \int W_1(x_1, \boldsymbol{X}_{-l}) H\left(\frac{1 - x_1 \beta_{\phi,1}^* - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}_{\phi,-1}^* - t_j}{h_l} \right) \boldsymbol{x}^\top \boldsymbol{v} \\ & \times f_{x_1|\boldsymbol{X}_{-l}}(x_1) f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l}) \mathrm{d}x_1 \mathrm{d}\boldsymbol{X}_{-l} \\ = & -\frac{h_l}{\beta_{\phi,1}^*} \int W_1\left(\frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}_{\phi,-1}^* - t_j - h_l y}{\beta_{\phi,1}^*}, \boldsymbol{X}_{-l} \right) H(y) \\ & \left\{ v_1 \frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}_{\phi,-1}^* - t_j - h_l y}{\beta_{\phi,1}^*} + \boldsymbol{X}_{-l}^\top \boldsymbol{v}_{-1} \right\} \\ & f_{x_1|\boldsymbol{X}_{-l}}\left(\frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}_{\phi,-1}^* - t_j - h_l y}{\beta_{\phi,1}^*} \right) f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l}) \mathrm{d}y \mathrm{d}\boldsymbol{X}_{-l}. \end{aligned}$$

Let

$$M_a(t; \boldsymbol{X}_{-l}) = \int_{-\infty}^t W_a(t, \boldsymbol{X}_{-l}) t f_{x_1|\boldsymbol{X}_{-l}}(t) \mathrm{d}t$$

and

$$N_a(t; \boldsymbol{X}_{-l}) = \int_{-\infty}^t W_a(t, \boldsymbol{X}_{-l}) f_{x_1|\boldsymbol{X}_{-l}}(t) \mathrm{d}t.$$

We have

$$
\begin{aligned}
& -\int W_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j-h_l y}{\beta_{\phi,1}^*},\boldsymbol{X}_{-l}\right)H(y) \\
& \left\{v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j-h_l y}{\beta_{\phi,1}^*}+\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}\right\} \\
& f_{x_1|\boldsymbol{X}_{-l}}\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j-h_l y}{\beta_{\phi,1}^*}\right)\mathrm{d}y \\
= \; & \frac{\beta_{\phi,1}^*}{h_l}v_1\int H(y)\mathrm{d}M_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j-h_l y}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)+ \\
& \frac{\beta_{\phi,1}^*}{h_l}\int\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}H(y)\mathrm{d}N_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}^*-t_j-h_l y}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right) \\
= \; & -\frac{\beta_1^*}{h_l}v_1\int_{-1}^1 M_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j-h_l y}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)H'(y)\mathrm{d}y- \\
& \frac{\beta_1^*}{h_l}\int\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}N_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j-h_l y}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)H'(y)\mathrm{d}y \\
= \; & -\frac{\beta_1^*}{h_l}v_1 M_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right) \\
& -\frac{\beta_{\phi,1}^*}{h_l}\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}N_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)+R_1,
\end{aligned}
$$

where R_1 can be bounded by a constant,

$$
\begin{aligned}
|R_1| \; \leqslant \; & |\beta_{\phi,1}^* v_1|\int_{-1}^1\sup_{t\in[-M,M]}|\partial_t M_1(t;\boldsymbol{X}_{-l})||H'(y)|\mathrm{d}y+ \\
& |\beta_{\phi,1}^*||\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}|\int_{-1}^1\sup_{t\in[-M,M]}|\partial_t N_1(t;\boldsymbol{X}_{-l})||H'(y)|\mathrm{d}y \\
\leqslant \; & |\beta_{\phi,1}^*||\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}|\left(\sup_{t\in[-M,M]}|\partial_t M_1(t;\boldsymbol{X}_{-l})|+\sup_{t\in[-M,M]}|\partial_t N_1(t;\boldsymbol{X}_{-l})|\right) \\
& \int_{-1}^1|H'(y)|\mathrm{d}y.
\end{aligned}
$$

Combining these equations and inequalities, we have

$$\begin{aligned}
&\mathbb{E}\left[W_1 H\left(\frac{1-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*-t_j}{h_l}\right)\boldsymbol{X}^\top\boldsymbol{v}\right]\\
=&\int\left\{v_1 M_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)+\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}N_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)\right\}\\
&f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}-\frac{h_l}{\beta_{\phi,1}^*}\int R_1 f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l},
\end{aligned}$$

and $|\int R_1 f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}| \leqslant \int |R_1| f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} = O(\{E[(\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1})^2]\}^{1/2})$ is bounded. Similarly, we can show

$$\begin{aligned}
&\sup_v\Bigg|\mathbb{E}\left[W_{-1}H\left(\frac{1+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*-t_j}{h_l}\right)\boldsymbol{X}^\top\boldsymbol{v}\right]\\
&-\int\left\{v_1 M_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l}\right)+\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}N_1\left(\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,-1}^*};\boldsymbol{X}_{-l}\right)\right\}\\
&f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}\Big|\\
=&\ O(h_l).
\end{aligned}$$

Combining these results. we have

$$\begin{aligned}
\sup_v\Bigg|\sum_j\Delta_j\mathbb{E}\Big[&\{W_{-1}H(\tfrac{1+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*-t_j}{h_l})-W_{-1}H(\tfrac{1+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*-t_j}{h_l})\}\boldsymbol{X}^\top\boldsymbol{v}\Big]\\
&-\sum_j\Delta_j\int\Big\{v_1M_1(\tfrac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l})+\boldsymbol{X}_{-l}^\top\boldsymbol{v}_{-1}\\
&\cdot N_1(\tfrac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{\phi,-1}^*-t_j}{\beta_{\phi,1}^*};\boldsymbol{X}_{-l})\Big\}f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\,\mathrm{d}\boldsymbol{X}_{-l}\Bigg|=O(h_l)
\end{aligned}$$

By definition of $\boldsymbol{\beta}_\phi^*$, we have

$$\sup_v\left|\sum_j\Delta_j\mathbb{E}\left[W_{-1}\left\{H\left(\frac{1+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*-t_j}{h_l}\right)-H\left(\frac{1+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*-t_j}{h_l}\right)\right\}\boldsymbol{X}^\top\boldsymbol{v}\right]\right|=O(h_l).$$

□

Lemma 10 *We have that*

$$\begin{aligned}
&\mathbb{E}\left[\left\{W_1\tilde{\phi}''(1-\boldsymbol{X}^\top\boldsymbol{\beta})+W_{-1}\tilde{\phi}''(1+\boldsymbol{X}^\top\boldsymbol{\beta})\right\}X_{j_1}X_{j_2}\right]\\
=&\sum_{j=1}^J\mathbb{E}\left[\left\{W_1\sum_{j=1}^J\Delta_j\delta(t_j-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)+W_{-1}\sum_{j=1}^J\Delta_j\delta(t_j+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right\}X_{j_1}X_{j_2}\right]\\
&+O(h_l+\|\boldsymbol{\beta}-\boldsymbol{\beta}_\phi^*\|_2),
\end{aligned}$$

uniformly holds over all (j_1, j_2)*for any* $\boldsymbol{\beta}$ *with* $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_\phi\|_2 \to 0$. *Especially, we have that*

$$\begin{aligned} & \sup_{\boldsymbol{v}\in\{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1\}} \left| \mathbb{E}\left[\left\{ W_1 \tilde{\phi}''(1 - \boldsymbol{X}^\top\boldsymbol{\beta}) + W_{-1}\tilde{\phi}''(1 + \boldsymbol{X}^\top\boldsymbol{\beta}) \right\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \right] \right. \\ & \left. - \mathbb{E}\left[\left\{ W_1 \sum_{j=1}^J \Delta_j \delta(t_j - \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) + W_{-1} \sum_{j=1}^J \Delta_j \delta(t_j + \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) \right\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \right] \right| \\ = \;& O(h_l + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_\phi\|_2). \end{aligned}$$

Proof of Lemma 10. To show the result in Lemma 10, we will equivalently show that

$$\begin{aligned} & \sup_{\boldsymbol{v}\in\{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1\}} \left| \mathbb{E}\left[\left\{ W_1 \tilde{\phi}''(1 - \boldsymbol{X}^\top\boldsymbol{\beta}) + W_{-1}\tilde{\phi}''(1 + \boldsymbol{X}^\top\boldsymbol{\beta}) \right\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \right] \right. \\ & \left. - \mathbb{E}\left[\left\{ W_1 \sum_{j=1}^J \Delta_j \delta(t_j - \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) + W_{-1} \sum_{j=1}^J \Delta_j \delta(t_j + \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) \right\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \right] \right| \\ = \;& O(h_l + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_\phi\|_2). \end{aligned}$$

Notice that

$$\begin{aligned} & \mathbb{E}\Big[\Big\{ W_1 \tilde{\phi}''(1 - \boldsymbol{X}^\top\boldsymbol{\beta}) + W_{-1}\tilde{\phi}''(1 + \boldsymbol{X}^\top\boldsymbol{\beta}) \Big\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \Big] \\ = \;& \sum_j \frac{\Delta_j}{h_l} \mathbb{E}\Big[\Big\{ W_1 H'\Big(\frac{1 - \boldsymbol{X}^\top\boldsymbol{\beta} - t_j}{h_l} \Big) + W_{-1} H'\Big(\frac{1 + \boldsymbol{X}^\top\boldsymbol{\beta} - t_j}{h_l} \Big) \Big\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \Big]. \end{aligned}$$

Let us compute

$$I_1 := \mathbb{E}\left[\left\{ W_1 H'\left(\frac{1 - \boldsymbol{X}^\top\boldsymbol{\beta} - t_j}{h_l} \right) + W_{-1} H'\left(\frac{1 + \boldsymbol{X}^\top\boldsymbol{\beta} - t_j}{h_l} \right) \right\} \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \right].$$

Because

$$\begin{aligned} I_1 \;&=\; \mathbb{E}\Big[W_1 H'\Big(\frac{1 - \boldsymbol{X}^\top\boldsymbol{\beta} - t_j}{h_l} \Big) \Big] \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \\ &\quad + \mathbb{E}\Big[W_{-1} H'\Big(\frac{1 + \boldsymbol{X}^\top\boldsymbol{\beta} - t_j}{h_l} \Big) \Big] \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} \\ &=:\; I_{11} + I_{12}. \end{aligned}$$

we calculate I_{11} and I_{12} separately. For I_{11}, we have

$$\begin{aligned}
I_{11} &= \int W_1 H'\left(\frac{1-\boldsymbol{x}^\top\boldsymbol{\beta}-t_j}{h_l}\right)\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v} f_{x_1|\boldsymbol{X}_{-l}}(x_1) f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}x_1\mathrm{d}\boldsymbol{X}_{-l} \\
&= \int\int W_1 H'\left(\frac{1-x_1\beta_1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}-t_j}{h_l}\right)(v_1x_1+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l})^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}(x_1)\mathrm{d}x_1 f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&= -\frac{h_l}{\beta_1}\int\int W_1 H'(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}-t_j-h_l y}{\beta_1}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}-t_j-hy}{\beta_1}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&= -\frac{h_l}{\beta^*_{\phi,1}}\int\int W_1 H'(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_l y}{\beta^*_{\phi,-1}}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_l y}{\beta^*_{\phi,1}}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}+\mathcal{R}.
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{R} &= \frac{h_l}{\breve{\beta}_1^2}(\beta_1-\beta^*_{\phi,1})\int\int W_1 H'(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\breve{\boldsymbol{\beta}}_{-1}-t_j-h_l y}{\breve{\beta}_1}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\breve{\boldsymbol{\beta}}_{-1}-t_j-h_l y}{\breve{\beta}_1}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&\quad -\frac{h_l}{\breve{\beta}_1}\int\int W_1 H'(y)\,2\Delta^\top(\breve{\boldsymbol{\beta}})(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\phi)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\breve{\boldsymbol{\beta}}_{-1}-t_j-h_l y}{\breve{\beta}_1}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right) \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\breve{\boldsymbol{\beta}}_{-1}-t_j-h_l y}{\breve{\beta}_1}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&\quad -\frac{h_l}{\breve{\beta}_1}\int\int W_1 H'(y)\,\Delta^\top(\breve{\boldsymbol{\beta}})(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\phi)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\breve{\boldsymbol{\beta}}_{-1}-t_j-h_l y}{\breve{\beta}_1}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f'_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\breve{\boldsymbol{\beta}}_{-1}-t_j-h_l y}{\breve{\beta}_1}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l},
\end{aligned}$$

$\breve{\boldsymbol{\beta}}$ is a vector on the segment $t_0\boldsymbol{\beta}+(1-t_0)\boldsymbol{\beta}^*_\phi$ with $t_0\in[0,1]$, and $\Delta(\boldsymbol{\beta})=\left(-\frac{\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}}{\beta_1^2},\frac{\boldsymbol{X}_{-l}}{\beta_1}\right)$. By Condition (C1) and $|\beta_1-\beta^*_{\phi,1}|\leqslant\|\boldsymbol{\beta}-\boldsymbol{\beta}^*_\phi\|_2$, further notice that $H'(y)\geqslant 0$ and $\int H'(y)\mathrm{d}y=1$, we have that

$$\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}\Delta_j/h_l|\mathcal{R}|\lesssim E[|\boldsymbol{X}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\phi)|]=O(\|\boldsymbol{\beta}-\boldsymbol{\beta}^*_\phi\|_2).$$

Next, we compute

$$
\begin{aligned}
& -\frac{h_l}{\beta^*_{\phi,1}} \int\int W_1 H'(y) \left(-v_1 \frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}^*_{\phi,-1} - t_j - h_l y}{\beta^*_{\phi,1}} + \boldsymbol{v}_{-1} \boldsymbol{X}_{-l} \right)^2 \\
& f_{x_1|\boldsymbol{X}_{-l}} \left(-\frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}^*_{\phi,-1} - t_j - h_l y}{\beta^*_{\phi,1}} \right) \mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l}) \mathrm{d}\boldsymbol{X}_{-l} \\
= \; & -\frac{h_l}{\beta^*_{\phi,1}} \int\int W_1 H'(y) \left(-v_1 \frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}^*_{\phi,-1} - t_j}{\beta^*_{\phi,1}} + \boldsymbol{v}_{-1} \boldsymbol{X}_{-l} \right)^2 \\
& f_{x_1|\boldsymbol{X}_{-l}} \left(-\frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}^*_{\phi,-1} - t_j}{\beta^*_{\phi,1}} \right) \mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l}) \mathrm{d}\boldsymbol{X}_{-l} - \frac{h_l}{\beta^*_{\phi,1}} \widetilde{\mathcal{R}},
\end{aligned}
$$

where

$$
\begin{aligned}
\widetilde{\mathcal{R}} \; = \; & \int\int W_1 H^{'}(y) \left\{ 2 \frac{v_1}{\beta^*_{\phi,1}} \left(-v_1 \frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}^*_{\phi,-1} - t_j - h_l \breve{y}}{\beta^*_{\phi,1}} + \boldsymbol{v}_{-1} \boldsymbol{X}_{-l} \right) + \right. \\
& \left. \frac{1}{\beta^*_{\phi,1}} f^{'}_{x_1|\boldsymbol{X}_{-l}} \left(-\frac{1 - \boldsymbol{X}_{-l}^\top \boldsymbol{\beta}^*_{\phi,-1} - t_j - h_l \breve{y}}{\beta^*_{\phi,1}} \right) \right\} h_l y \mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l}) \mathrm{d}\boldsymbol{X}_{-l},
\end{aligned}
$$

and $\breve{y}$ is between $[-1, 1]$. By Condition (C1) and properties of $H^{'}(y)$, we have that $\widetilde{\mathcal{R}} = O(h_l)$ uniformly for all $\boldsymbol{v}$. Likewise, we can calculate I_{12}. By summarizing I_{11} and I_{12}, we can conclude the proof. □

Lemma 11 *We have that*

$$
\begin{aligned}
& \mathbb{E}\left[\left\{ W_1 \sum_{j=1}^{J} \Delta_j G_{h_g}(t_j - \boldsymbol{X}^\top \boldsymbol{\beta}) + W_{-1} \sum_{j=1}^{J} \Delta_j G_{h_g}(t_j + \boldsymbol{X}^\top \boldsymbol{\beta}) \right\} X_{j_1} X_{j_2} \right] \\
= \; & \sum_{j=1}^{J} \mathbb{E}\left[\left\{ W_1 \sum_{j=1}^{J} \Delta_j \delta(t_j - \boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi) + W_{-1} \sum_{j=1}^{J} \Delta_j \delta(t_j + \boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi) \right\} X_{j_1} X_{j_2} \right] \\
& + O(h_g^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_\phi\|_2),
\end{aligned}
$$

uniformly holds over all (j_1, j_2) *for any* $\boldsymbol{\beta}$ *with* $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \to 0$. *Especially, we have that*

$$
\begin{aligned}
& \sup_{\boldsymbol{v} \in \{\boldsymbol{v}: \|\boldsymbol{v}\|_2 = 1\}} \left| \mathbb{E}\left[\left\{ W_1 G_{h_g}(1 - \boldsymbol{X}^\top \boldsymbol{\beta}) + W_{-1} G_{h_g}(1 + \boldsymbol{X}^\top \boldsymbol{\beta}) \right\} \boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{v} \right] \right. \\
& \left. - \mathbb{E}\left[\left\{ W_1 \sum_{j=1}^{J} \Delta_j \delta(t_j - \boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi) + W_{-1} \sum_{j=1}^{J} \Delta_j \delta(t_j + \boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi) \right\} \boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{v} \right] \right| \\
= \; & O(h_g^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_\phi\|_2).
\end{aligned}
$$

Proof of Lemma 11. The proof of Lemma 11 is similar to that of Lemma 10. More specifically, we can replace h_l with h_g, replace the local kernel function H' with the global

kernel function G, and have that

$$\begin{aligned}
I_{11} &= \int W_1 G\left(\frac{1-\boldsymbol{x}^\top\boldsymbol{\beta}-t_j}{h_g}\right)\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v} f_{x_1|\boldsymbol{X}_{-l}}(x_1) f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}x_1\mathrm{d}\boldsymbol{X}_{-l} \\
&= \int\int W_1 G\left(\frac{1-x_1\beta_1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}-t_j}{h_g}\right)(v_1x_1+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l})^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}(x_1)\mathrm{d}x_1 f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&= -\frac{h_g}{\beta_1}\int\int W_1 G(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}-t_j-h_gy}{\beta_1}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}_{-1}-t_j-hy}{\beta_1}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&= -\frac{h_g}{\beta^*_{\phi,1}}\int\int W_1 G(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_gy}{\beta^*_{\phi,1}}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_gy}{\beta^*_{\phi,1}}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}+\mathcal{R}.
\end{aligned}$$

where $\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}\Delta_j/h_g|\mathcal{R}| = O(\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2)$, similar to Lemma 10. For the first term, we have

$$\begin{aligned}
&-\frac{h_g}{\beta^*_{\phi,1}}\int\int W_1 G(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_ly}{\beta^*_{\phi,1}}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_ly}{\beta^*_{\phi,1}}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
=\ & -\frac{h_g}{\beta^*_{\phi,1}}\int\int W_1 G(y)\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j}{\beta^*_{\phi,1}}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)^2 \\
&\quad f_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j}{\beta^*_{\phi,1}}\right)\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}-\frac{h_g}{\beta^*_{\phi,1}}\widetilde{\mathcal{R}},
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\mathcal{R}} &= \int\int W_1 G(y)\left\{2\frac{v_1}{\beta^*_{\phi,1}}\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j}{\beta^*_{\phi,1}}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)+\right. \\
&\quad \left.\frac{1}{\beta^*_{\phi,1}}f'_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j}{\beta^*_{\phi,1}}\right)\right\}h_ly\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l} \\
&\quad +\int\int W_1 G(y)\partial_{y=\breve{y}}\left\{2\frac{v_1}{\beta^*_{\phi,1}}\left(-v_1\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_ly}{\beta^*_{\phi,1}}+\boldsymbol{v}_{-1}\boldsymbol{X}_{-l}\right)+\right. \\
&\quad \left.\frac{1}{\beta^*_{\phi,1}}f'_{x_1|\boldsymbol{X}_{-l}}\left(-\frac{1-\boldsymbol{X}_{-l}^\top\boldsymbol{\beta}^*_{\phi,-1}-t_j-h_ly}{\beta^*_{\phi,1}}\right)\right\}h_l^2y^2\mathrm{d}y f_{\boldsymbol{X}_{-l}}(\boldsymbol{X}_{-l})\mathrm{d}\boldsymbol{X}_{-l}
\end{aligned}$$

By $\int G(y)y\mathrm{d}y = 0$, we have that $\tilde{R} = O(h_l^2)$. Comparing with the derivation of Lemma 10, we can conclude this lemma. □

Estimation error of $\boldsymbol{w}^*_{\phi,l}$

In this section, we investigate the convergence rate of $\widehat{\boldsymbol{w}}_{\phi,l}$ assuming a consistent estimation of $\boldsymbol{\beta}^*_\phi$, i.e., $\|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\| = o_p(1)$. We first provide our theoretical result on the convergence rate of $\widehat{\boldsymbol{w}}_{\phi,l}$ without nuisance parameters.

Lemma 12 *Denote $s' = \max_l \|\boldsymbol{w}^*_{\phi,l}\|_0$. Assume that $\|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\|_2 \leqslant \Delta_{\beta,2}$ with probability approaching to 1, $s'\sqrt{\log p/(nh_{\mathrm{gb}})} = o(1)$ and $\max_l |\boldsymbol{x}_{-l}^\top \boldsymbol{w}^*_{\phi,l}|$ is bounded by R. Taking $\mu_n \asymp \delta_n + Rh^2_{\mathrm{gb}} + R\Delta_{\beta,2}$, where $\delta_n = R(\log p/(nh_{\mathrm{gb}}))^{1/2}$. Then we have*

$$\begin{aligned} \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_2 &= O_p(\sqrt{s'}\mu_n), \\ \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,}\|_1 &= O_p(\sqrt{s'}\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_2), \end{aligned}$$

uniformly hold over all l's.

Suppose that $R = O(1)$, Lemma 12 shows that under the choice of the bandwidth $h_{\mathrm{gb}} = (\log p/n)^{1/5}$, due to the global kernel approximation, the uniform l_1-convergence rate of $\widehat{\boldsymbol{w}}_{\phi,l}$ is $O_p(s'\,(\log p/n)^{2/5})$ if $\Delta_{\beta,2} = O((\log p/n)^{2/5})$, which is different from the standard convergence rate in high-dimensional inference for generalized linear models. Due to the slow convergence rate of $\widehat{\boldsymbol{w}}_{\phi,l}$, we may need more restrictive conditions than those assumed in generalized linear models in order to obtain a valid inference procedure. In particular, an ERM under a logistic surrogate loss can be interpreted as a logistic regression under a possibly misspecified model; this implies that our conditions for piece-wise linear surrogate loss functions to obtain a valid inference procedure may be more restrictive than those for differentiable surrogate loss functions.

Lemma 13 provides our theoretical result on the convergence rate of $\widehat{\boldsymbol{w}}_{\phi,l}$ with additional nuisance parameters.

Lemma 13 *Denote $s' = \max_l \|\boldsymbol{w}^*_{\phi,l}\|_0$ and $\delta_n = R(\log p/(nh_{\mathrm{gb}}))^{1/2} + Rn^{-\eta}/h_{\mathrm{gb}}$. Assume that $\|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\|_2 \leqslant \Delta_{\beta,2}$ with probability approaching to 1, $s'\sqrt{\log p/(nh_{\mathrm{gb}})} = o(1)$ and $\max_l |\boldsymbol{x}_{-1}^\top \boldsymbol{w}^*_{\phi,l}|$ is bounded by R. Taking $\mu_n \asymp \delta_n + Rh^2_{\mathrm{gb}} + R\Delta_{\beta,2}$, we have*

$$\begin{aligned} \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_2 &= O_p(\sqrt{s'}\mu_n) \\ \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,}\|_1 &= O_p(\sqrt{s'}\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_2), \end{aligned}$$

uniformly hold over all l's.

Compared with Lemma 12, the δ_n also involves the $Rn^{-\eta}/h_{\mathrm{gb}}$ due to the additional nuisance parameters. It implies that 1) the best bandwidth choice h_{gb} may depend on $n^{-\eta}$; 2) the convergence rate of $\widehat{\boldsymbol{w}}_{\phi,l}$ depends on $n^{-\eta}$.

To prove these results, we start from the definition of the target $\boldsymbol{w}^*_{\phi,l}$ which is defined as

$$L_{\phi''}(\boldsymbol{w}_l) = \mathbb{E}\left[\sum_{j=1}^{J}\Delta_j \sum_{a\in\{-1,1\}} W_a \delta(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j)(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_l)^2\right].$$

In estimation of $\boldsymbol{w}^*_{\phi,l}$, we replace $\boldsymbol{\beta}^*_\phi$ with $\widehat{\boldsymbol{\beta}}_\phi$ and W_a with $\widehat{W}_a$. As discussed in the main text, the $\widehat{W}_a$'s depend on different problems. In common classification problem, the $\widehat{W}_a = 1\{A = a\}$; in classification problems with missing labels, the weight $\widehat{W}_a = W_a(\boldsymbol{X}, A; \widehat{\pi}, \widehat{p})$; the $\widehat{\pi}$ and $\widehat{p}$ are nuisance parameters. In estimating ITRs, the weight $\widehat{W}_a = W_a(Y, \boldsymbol{X}, A; \widehat{\pi}, \widehat{Q})$; the $\widehat{\pi}$ and $\widehat{Q}$ are nuisance parameters. Replacing the δ function with a global kernel, we obtain the estimator $\widehat{\boldsymbol{w}}_{\phi,l}$ by minimizing

$$\ell_{\phi''}(\boldsymbol{w}_l) = \widehat{\mathbb{E}}_n \left[\sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \widehat{W}_a G_{h_g}(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi)(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_l)^2 \right] + \mu_n \|\boldsymbol{w}_l\|_1.$$

In both Algorithm 1 and 2, we split the data such that estimating $\widehat{\boldsymbol{\beta}}_\phi$ and $\widehat{\boldsymbol{w}}_{\phi,l}$ are implemented on independent datasets. To unify two algorithms, we split the dataset into k folds and consider the following optimization.

$$\begin{aligned} \ell_{\phi''}(\boldsymbol{w}_l; \widehat{W}_a) \quad = \quad & \sum_{k=1}^{K} \widehat{\mathbb{E}}_n^{(k)} \Big[\sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \widehat{W}_a G_{h_g}(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)})(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_l)^2 \Big] \\ & + \mu_n \|\boldsymbol{w}_l\|_1. \end{aligned}$$

where the nuisance parameters in $\widehat{W}_a$ are assumed to be estimated on an independent dataset; $\widehat{\boldsymbol{\beta}}_\phi^{(-k)}$ is estimated excluding the k-th fold; and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ is the empirical average on the k-th fold. We can check that both Step 3 in Algorithm 1 and Step 5 in Algorithm 5 solve an optimization following this general formulation.

To derive the asymptotic property of the minimizer $\widehat{w}_\phi$, we introduce the notation $\overline{W}_a$ as the weight with the limits of the estimators of nuisance parameters. Suppose that Θ represents the nuisance parameters, the true nuisance parameter is denoted as Θ_0 and the estimated nuisance parameter is denoted as $\widehat{\Theta}$. For example, in classification problem with missing labels, the true nuisance parameter $\Theta_0 = (\pi, p_a)$; the estimated nuisance parameter $\widehat{\Theta} = (\widehat{\pi}, \widehat{p}_a)$. In this context, the weight $\widehat{W}_a$ can be written as $\widehat{W}_a = W_a(\boldsymbol{X}, A; \widehat{\Theta})$. Denote the limits of the estimated nuisance parameter $\widehat{\Theta}$ as $\bar{\Theta}$. We define the weight $\overline{W}_a = W_a(\boldsymbol{X}, A; \bar{\Theta})$.

Under these notations, we assume that

(C2) There exists a positive constant η such that

$$\sup_{\boldsymbol{x}} \left\| \mathbb{E}\left[\widehat{W}_a - \overline{W}_a \mid \boldsymbol{X} = \boldsymbol{x} \right] \right\| = O_p(n^{-\eta}), \quad \mathbb{E}\left[\overline{W}_a \mid \boldsymbol{X} \right] = W_a.$$

Especially, when $\widehat{W}_a = \overline{W}_a = 1\{A = a\}$ and $W_a(\boldsymbol{X}) = P(A = a \mid \boldsymbol{X})$.

Lemma 14 *Let $s' = \max \|\boldsymbol{w}^*_{\phi,l}\|_0$ and $\delta_n = R(\log p/(nh_g))^{1/2} + Rn^{-\eta}/h_g$. Assume that $s'\sqrt{\log p/(nh_g)} = o(1)$ and $\max_j |\boldsymbol{w}^*_{\phi,l}|_1 \leqslant R$, take $\mu_n \asymp \delta_n + +Rh_g^2 + R\Delta_{\beta,2}$, we have*

$$\begin{aligned} \max_j \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_2 \quad &= \quad O_p(\sqrt{s'}\mu_n) \\ \max_j \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_1 \quad &= \quad O(\sqrt{s'}\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_2). \end{aligned}$$

Proof of Lemma 14. By definition, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})^2\right]+\mu_n\|\widehat{\boldsymbol{w}}_{\phi,l}\|_1\\
\leqslant\;&\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)^2\right]+\mu_n\|\boldsymbol{w}_{\phi,l}^*\|_1.
\end{aligned}$$

Let $\widehat{\boldsymbol{\Delta}}_l:=\widehat{\boldsymbol{w}}_{\phi,l}^*-\boldsymbol{w}_{\phi,l}^*$, $t_l:=\|\widehat{\boldsymbol{\Delta}}_l\|_2\wedge 1$, and $\boldsymbol{\delta}_l:=\widehat{\boldsymbol{\Delta}}_l/\|\widehat{\boldsymbol{\Delta}}_l\|_2$. By the convexity, we have that

$$\begin{aligned}
&\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top(\boldsymbol{w}_\phi^*+t_l\boldsymbol{\delta}))^2\right]\\
&+\mu_n\|\boldsymbol{w}_{\phi.j}^*+t_l\boldsymbol{\delta}_l\|_1\\
\leqslant\;&\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)^2\right]+\mu_n\|\boldsymbol{w}_{\phi,l}^*\|_1.
\end{aligned}$$

By rearranging terms, we have equivalently

$$\begin{aligned}
&\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(t_l\boldsymbol{\delta}_l^\top\boldsymbol{X}_{-l})^2\right]\\
\leqslant\;&2\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)t_l\boldsymbol{\delta}_l^\top\boldsymbol{X}_{-l}\right]\\
&+\mu_n\|\boldsymbol{w}_{\phi,l}^*\|_1-\mu_n\|\boldsymbol{w}_{\phi,l}^*+t_l\boldsymbol{\delta}_l\|_1.
\end{aligned}$$

Let $\mathcal{S}'$ denote the non-zero indexes of $\boldsymbol{w}_{\phi,l}^*$. By the definition of $\mathcal{S}'$, we have

$$\|\boldsymbol{w}_{\phi,l}^*\|_1-\|\boldsymbol{w}_{\phi,l}^*+t_l\boldsymbol{\delta}_l\|_1\leqslant\|\boldsymbol{w}_{\phi,\mathcal{S}',l}^*\|_1-\|\boldsymbol{w}_{\phi,\mathcal{S}',l}^*+t_l\boldsymbol{\delta}_{\mathcal{S}',l}\|_1-t\|\boldsymbol{\delta}_{\bar{\mathcal{S}}',l}\|_1\leqslant t_l\|\boldsymbol{\delta}_{\mathcal{S}',l}\|_1-t_l\|\boldsymbol{\delta}_{\bar{\mathcal{S}}',l}\|_1.$$

Combining the two inequalities, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\widehat{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(t_l\boldsymbol{\delta}^\top\boldsymbol{X}_{-l})^2\right]\\
\leqslant\;&2\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)t_l\boldsymbol{\delta}_l^\top\boldsymbol{X}_{-l}\right]\\
&+2\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\left(\widehat{W}_a-\overline{W}_a\right)G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}\right.\\
&\left.(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)t_l\boldsymbol{\delta}^\top\boldsymbol{X}_{-l}\right]+\mu_n\left(t_l\|\boldsymbol{\delta}_{\mathcal{S}'}\|_1-t_l\|\boldsymbol{\delta}_{\bar{\mathcal{S}}'}\|_1\right)\\
\leqslant\;&2I_{11}+2I_{12}+\mu_n t_l\left(\|\boldsymbol{\delta}_{\mathcal{S}',l}\|_1-\|\boldsymbol{\delta}_{\bar{\mathcal{S}}'}\|_1\right).
\end{aligned}$$

For I_{11}, we have

$$\begin{aligned} & I_{11} \\ \leqslant\ & t\left|\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\boldsymbol{X}_{-l}^\top\boldsymbol{\delta}_l\right]\right| \\ \leqslant\ & t\sum_{k=1}^{K}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\boldsymbol{X}_{-l}^\top\boldsymbol{\delta}_l\right]\right|. \end{aligned}$$

Notice that

$$\begin{aligned} & \left|\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\boldsymbol{X}_{-l}^\top\boldsymbol{\delta}_l\right]\right| \\ \leqslant\ & \left\|(\widehat{\mathbb{E}}_n^{(k)}-\mathbb{E})\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\boldsymbol{X}_{-l}\right]\right\|_\infty \\ & \cdot\|\boldsymbol{\delta}_l\|_1 \\ & +\left|\mathbb{E}\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\boldsymbol{X}_{-l}^\top\boldsymbol{\delta}_l\right]\right|. \end{aligned}$$

Due to the independence between $\widehat{\boldsymbol{\beta}}_\phi^{(-k)}$ and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, by Hoeffding's inequality and

$$\sup_t G_{h_g}(t)=O(h_g^{-1}),$$

we have

$$\begin{aligned} & \max_l\left\|(\widehat{\mathbb{E}}_n^{(k)}-\mathbb{E})\left[\left\{\sum_{j=1}^{J}\Delta_j\sum_{a\in\{-1,1\}}\overline{W}_a G_{h_g}\left(t_j-a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right\}\right. \\ & \qquad\left.\cdot(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\boldsymbol{X}_{-l}\right]\right\|_\infty \\ =\ & O_p\Big(R(\log p/(nh_g))^{-1/2}\Big). \end{aligned}$$

By Lemma 11 and $\max_j \|\boldsymbol{w}^*_{\phi,j}\|_1 \leqslant R$, we have that

$$\begin{aligned}
& \left| \mathbb{E}\left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \overline{W}_a G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\} (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}^\top_{-l}\boldsymbol{\delta}_l \right]\right| \\
= & \left| \mathbb{E}\left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} W_a G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\} (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}^\top_{-l}\boldsymbol{\delta}_l \right]\right| \\
\leqslant & \left| \mathbb{E}\left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} W_a \delta\left(t_j - a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\right)\right\} (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}^\top_{-l}\boldsymbol{\delta} \right]\right| + \\
& \left| \mathbb{E}\left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} W_a \left(G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right) - \delta\left(t_j - a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\right)\right)\right\} \right.\right. \\
& \left.\left. (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}^\top_{-l}\boldsymbol{\delta}_l \right]\right| \\
\leqslant & \sqrt{1 + \|\boldsymbol{w}^*_{\phi,l}\|_2^2}\, O(h_g^2 + \|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\|_2)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
& \max_l \left| \mathbb{E}\left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \overline{W}_a G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\} (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}^\top_{-l}\boldsymbol{\delta}_l \right]\right| \\
= & O(R\left\{h_g^2 + \|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\|_2\right\}).
\end{aligned}$$

Combining these inequalities, we have that

$$I_{11} \leqslant O_p((R\log p/(nh_g))^{-1/2}) t_l \|\boldsymbol{\delta}_l\|_1 + O(R\left\{h_g^2 + \|\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi\|_2\right\}) t_l,$$

uniformly holds over all l's.

For I_{12}, we have

$$\begin{aligned}
& I_{12} \\
\leqslant & \sum_{k=1}^{K} \left\| \widehat{\mathbb{E}}^{(k)}_n \left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \left(\widehat{W}_a - \overline{W}_a\right) G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\} \right.\right. \\
& \left.\left. (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}_{-l} \right]\right\|_\infty \|t_l\boldsymbol{\delta}_l\|_1 \\
\leqslant & \sum_{k=1}^{K} \left\| (\widehat{\mathbb{E}}^{(k)}_n - \mathbb{E}) \left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \left(\widehat{W}_a - \overline{W}_a\right) G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\} \right.\right. \\
& \left.\left. (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}_{-l} \right]\right\|_\infty \|t_l\boldsymbol{\delta}_l\|_1 \\
& + \sum_{k=1}^{K} \left\| \mathbb{E}\left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a\in\{-1,1\}} \left(\widehat{W}_a - \overline{W}_a\right) G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\} \right.\right. \\
& \left.\left. (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\boldsymbol{X}_{-l} \right]\right\|_\infty \|t_l\boldsymbol{\delta}_l\|_1 \\
\leqslant & O_p\left(\sup|\widehat{W}_a - \overline{W}_a| R\sqrt{\log p/(nh_g)} + Rn^{-\eta}/h_g\right) \|t_l\boldsymbol{\delta}_l\|_1,
\end{aligned}$$

uniformly holds over l's. Because $\widehat{\Theta} \to \bar{\Theta}$ in probability, we have that $\sup |\widehat{W}_a - \overline{W}_a| = o_p(1)$.

Combining these inequalities, we can show that there exists a constant C_ϵ such that with probability at least $1-\epsilon$, for some positive constant C_M, we have

$$\begin{aligned} & \sum_{k=1}^{K} \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{j=1}^{J} \Delta_j \sum_{a \in \{-1,1\}} \widehat{W}_a G_{h_g} \left(t_j - a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \right\} (t_l \boldsymbol{\delta}_l^\top \boldsymbol{X}_{-l})^2 \right] \\ \leqslant \; & 2C_\epsilon \delta_n t_l \|\boldsymbol{\delta}_l\|_1 + C_M (h_g^2 + \Delta_{\boldsymbol{\beta},2}) t_l + \mu_n t_l \left(\|\boldsymbol{\delta}_{\mathcal{S}',l}\|_1 - \|\boldsymbol{\delta}_{\bar{\mathcal{S}}',l}\|_1 \right), \end{aligned}$$

uniformly holds over l's, where $\delta_n = R(\log p/(nh_g))^{-1/2} + Rn^{-\eta}/h_g$.

Next, we will establish the lower bound. Especially, we will show the following claim.

Lemma 15 *Denote* $\widehat{F}^{(k)}(\boldsymbol{\beta}) = \widehat{\mathbb{E}}_n^{(k)}[\widehat{U}(\boldsymbol{\beta}) \boldsymbol{X} \boldsymbol{X}^\top]$, *where*

$$\widehat{U}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \Delta_j \sum_{a \in \{-1,1\}} \widehat{W}_a G_{h_g}(t_j - a \boldsymbol{X}^\top \boldsymbol{\beta}).$$

Assuming Conditions (a) - (e), with high probability, we have that

$$\boldsymbol{v}^\top \widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)}) \boldsymbol{v} \geqslant \frac{\kappa^2}{2} \|\boldsymbol{v}\|_2 (\|\boldsymbol{v}\|_2 - C_\kappa \sqrt{\log p/(nh_g)} \|\boldsymbol{v}\|_1),$$

for all $\|\boldsymbol{v}\|_2 \leqslant 1$ *and* k*'s, where* C_κ *and* κ *are positive constants.*

Assume that the claim is true, we can combine this lower bound and the derived upper bound,

$$\begin{aligned} \frac{\kappa^2}{2} t_l \left(t_l - C_\kappa \sqrt{\frac{\log p}{nh_g}} \|\boldsymbol{\delta}_l\|_1 \right) \; \leqslant \; & 2C_\epsilon \delta_n t_l \|\boldsymbol{\delta}_l\|_1 + C_M R (h_g^2 + \Delta_{\boldsymbol{\beta},2}) t_l \\ & + \mu_n t_l \left(\|\boldsymbol{\delta}_{\mathcal{S}',l}\|_1 - \|\boldsymbol{\delta}_{\bar{\mathcal{S}}',l}\|_1 \right). \end{aligned}$$

Thus, for a sufficient large C_κ, we have that

$$\frac{\kappa^2}{2} t_l \leqslant \left(\frac{\kappa^2}{2} C_\kappa + 2C_\epsilon \right) \delta_n \|\boldsymbol{\delta}_l\|_1 + C_M R (h_g^2 + \Delta_{\boldsymbol{\beta},2}) + \mu_n \left(\|\boldsymbol{\delta}_{\mathcal{S}',l}\|_1 - \|\boldsymbol{\delta}_{\bar{\mathcal{S}}',l}\|_1 \right).$$

Given the complexity of the best choice of h_g and μ_n, we take

$$\mu_n = 2 \left[\left(\frac{\kappa^2}{2} C_\kappa + 2C_\epsilon \right) \delta_n + C_M R (h_g^2 + \Delta_{\beta,2}) \right],$$

and notice that $\|\boldsymbol{\delta}\|_1 \geqslant 1$, we have that

$$\frac{\kappa^2}{2} t_l \leqslant 3\mu_n \|\boldsymbol{\delta}_{\mathcal{S}',l}\|_1 - \mu_n \|\boldsymbol{\delta}_{\bar{\mathcal{S}}',l}\|_1.$$

Thus, we have that

$$\begin{aligned}\|\widehat{\boldsymbol{w}}^*_{\phi,l}\|_1 &= t_l\|\boldsymbol{\delta}_l\|_1 \lesssim O(s'\mu_n),\\ \|\widehat{\boldsymbol{w}}^*_{\phi,l}\|_2 &= t_l \lesssim O(\sqrt{s'}\mu_n),\\ \|\widehat{\boldsymbol{w}}^*_{\phi,l}\|_1/\|\widehat{\boldsymbol{w}}^*_{\phi,l}\|_2 &= \|\boldsymbol{\delta}_l\|_1 \lesssim O(\sqrt{s'}),\end{aligned}$$

uniformly holds for all l's.

□

Proof of Lemma 15. Let

$$I_{\gamma\gamma} = \mathbb{E}\left[\left\{W_1\sum_{j=1}^J \Delta_j\delta(t_j - \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) + W_{-1}\sum_{j=1}^J \Delta_j\delta(t_j + \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi)\right\}\boldsymbol{X}\boldsymbol{X}^\top\right].$$

We have

$$\boldsymbol{v}^\top\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{v} = \boldsymbol{v}^\top I_{\gamma\gamma}\boldsymbol{v} + \boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi) - I_{\gamma\gamma}\right)\boldsymbol{v}.$$

We have

$$\begin{aligned}&\boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi) - I_{\gamma\gamma}\right)\boldsymbol{v}\\ =\;& \boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi) - \mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{X}\boldsymbol{X}^\top]\right)\boldsymbol{v}\\ &+\boldsymbol{v}^\top\left(\mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{X}\boldsymbol{X}^\top] - I_{\gamma\gamma}\right)\boldsymbol{v}\\ \leqslant\;& \sup_{\|\boldsymbol{v}\|_2\leqslant 1}\left|\boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi) - \mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{X}\boldsymbol{X}^\top]\right)\boldsymbol{v}\right|\\ &+\sup_{\|\widetilde{\boldsymbol{v}}\|_2=1}\left|\widetilde{\boldsymbol{v}}^\top\left(\mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{X}\boldsymbol{X}^\top] - I_{\gamma\gamma}\right)\widetilde{\boldsymbol{v}}\right|\|\boldsymbol{v}\|_2^2.\end{aligned}$$

For $\sup_{\|\widetilde{\boldsymbol{v}}\|_2=1}\left|\widetilde{\boldsymbol{v}}^\top\left(\mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{X}\boldsymbol{X}^\top] - I_{\gamma\gamma}\right)\widetilde{\boldsymbol{v}}\right|$, we have

$$\begin{aligned}&\widetilde{\boldsymbol{v}}^\top\left(\mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\boldsymbol{X}\boldsymbol{X}^\top] - I_{\gamma\gamma}\right)\widetilde{\boldsymbol{v}}\\ \leqslant\;& \mathbb{E}\Bigg[\Big\{(\widehat{W}_1 - \overline{W}_1)\sum_{j=1}^J \Delta_j G_{h_g}(t_j - \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\\ &\qquad +(\widehat{W}_{-1} - \overline{W}_{-1})\sum_{j=1}^J \Delta_j G_{h_g}(t_j + \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\Big\}(\boldsymbol{X}^\top\widetilde{\boldsymbol{v}})^2\Bigg]\\ &+\mathbb{E}\Bigg[\Big\{\overline{W}_1\sum_{j=1}^J \Delta_j G_{h_g}(t_j - \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\\ &\qquad +\overline{W}_{-1}\sum_{j=1}^J \Delta_j G_{h_g}(t_j + \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi)\Big\}(\boldsymbol{X}^\top\widetilde{\boldsymbol{v}})^2\Bigg]\\ &-\widetilde{\boldsymbol{v}}^\top I_{\gamma\gamma}\widetilde{\boldsymbol{v}}.\end{aligned}$$

By Lemma 11, we have

$$\sup_{\|\widetilde{\boldsymbol{v}}\|_2=1}\left|\mathbb{E}\left[\left\{\overline{W}_1\sum_{j=1}^{J}\Delta_j G_{h_g}(t_j-\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)})+\overline{W}_{-1}\sum_{j=1}^{J}\Delta_j G_{h_g}(t_j+\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)})\right\}\right.$$
$$\left.(\boldsymbol{X}^\top\widetilde{\boldsymbol{v}})^2\right]-\widetilde{\boldsymbol{v}}^\top I_{\gamma\gamma}\widetilde{\boldsymbol{v}}\Big|\, O_p(h_g^2+\|\widehat{\boldsymbol{\beta}}_\phi^{(-k)}-\boldsymbol{\beta}_\phi^*\|_2).$$

The first term is $O_p(n^{-\eta}/h_g)$.

For $\sup_{\|\boldsymbol{v}\|_2\leqslant 1}\left|\boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)})-\mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)})\boldsymbol{X}\boldsymbol{X}^\top]\right)\boldsymbol{v}\right|$, by the proof of Proposition 2 in Negahban et al. (2012), with probability goes to 1, we have that

$$\sup_{\|\boldsymbol{v}\|_2\leqslant 1}\left|\boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)})-\mathbb{E}[\widehat{U}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)})\boldsymbol{X}\boldsymbol{X}^\top]\right)\boldsymbol{v}\right|\leqslant\frac{\kappa^2}{2}\|\boldsymbol{v}\|_2^2+C_\kappa\|\boldsymbol{v}\|_1\|\boldsymbol{v}\|_2\sqrt{\log p/(nh_g)},$$

where C_κ is a positive constant.

Combining these inequalities, we have

$$\boldsymbol{v}^\top\left(\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)})-I_{\gamma\gamma}\right)\boldsymbol{v}\geqslant-C_\kappa\sqrt{\log p/(nh_g)}\|\boldsymbol{v}\|_1\|\boldsymbol{v}\|_2-O_p(h_g^2+\Delta_{\beta,2})\|\boldsymbol{v}\|_2^2,$$

where C_κ and κ are positive constants. Given $h_g^2+\Delta_{\beta,2}\to 0$, we have that

$$\boldsymbol{v}^\top\widehat{F}^{(k)}(\widehat{\boldsymbol{\beta}}_\phi^{(-k)})\boldsymbol{v}\geqslant\frac{\kappa^2}{2}\|\boldsymbol{v}\|_2(\|\boldsymbol{v}\|_2-C_\kappa\sqrt{\log p/(nh_g)}\|\boldsymbol{v}\|_1),$$

for all $\|\boldsymbol{v}\|_2\leqslant 1$ and k's. □

Asymptotic property of the kernel-smoothed decorrelated score under the null hypothesis

To derive the asymptotic of the proposed test statistics, we assume the following condition.

(C3) There exists a positive constant γ such that for all $t_0>t>0$,

$$\sup_{j,a\in\{-1,1\}}\mathbb{P}(|t_j-a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*|\leqslant t)\leqslant\tau t^\gamma,$$

where τ and t_0 are some constants.

Theorem 16 *Assume that* $\sqrt{n}R(h_l+n^{-\eta}+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)})=o(1)$, $\sqrt{n}(s'\mu_n)(n^{-\eta}+\sqrt{\log p/n}+h_l+\Delta_{\beta,2})=o(1)$, *and* $(Rs'\mu_n+R^2n^{-\zeta}+R^2h_l)\sqrt{\log p}=o(1)$, *where* $\|\widehat{\boldsymbol{\beta}}_\phi-\boldsymbol{\beta}_\phi^*\|_2\leqslant\Delta_{\beta,2}$ *with probability approaching to* 1. *Assume Conditions (C1) - (C3) with* $R=o(n^{1/6})$, *under the null hypothesis, we have*

$$\max_{l\in\mathcal{H}_0}\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1}S_{\widetilde{\phi}',null(l)}\right|\leqslant\Phi^{-1}(1-\alpha/2)\right)-(1-\alpha)\right|=o_p(1).$$

Especially, for a common classification problem without nuisance parameters, we just require that $\sqrt{n}R(h_l+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)})=o(1)$, $\sqrt{n}(s'\mu_n)(\sqrt{\log p/n}+h_l+\Delta_{\beta,2})=o(1)$, *and* $(Rs'\mu_n+h_l)\log p=o(1)$.

Proof of Theorem 16. In this proof, we consider a unified analysis for both Algorithm 1 and 2. In Algorithm 1, the kernel-smoothed decorrelated score is defined as

$$S_{\widetilde{\phi}',null(l)} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[A\widetilde{\phi}'\left(A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,null(l)}^{(-k)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right].$$

With bespoke weights, the kernel-smoothed decorrelated score is the average of the two scores on split datasets. For investigation of the theoretical property, we just need to focus on one score defined on the split dataset. Specifically, under the notation of Algorithm 2, the kernel-smoothed decorrelated score on a split dataset is defined as

$$S_{\widetilde{\phi}',null(l)}^{(\widetilde{J})} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(\widetilde{J}_k)}\left[\sum_{a\in\{-1,1\}}a\widehat{W}_a^{(\widetilde{I})}\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,null(l)}^{(-\widetilde{J}_k)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l}^{(\widetilde{J})})\right].$$

To summarize the two scenarios, we consider the following construction of the kernel-smoothed decorrelated score. Assuming $\widehat{W}_a$'s are estimated on an independent dataset, we split the entire dataset into K folds; the $\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)}$ is estimated using the data excluding the k-th fold; the $\widehat{\boldsymbol{w}}_{\phi,l}^*$ is obtained by minimizing $\ell_{\phi''}\left(\boldsymbol{w}_l;\widehat{W}_a\right)$; the kernel-smoothed decorrelated score is defined as

$$S_{\widetilde{\phi}',null(l)} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{-1,1\}}a\widehat{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,null(l)}^{(-k)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right].$$

By showing the asymptotic property of $S_{\widetilde{\phi}',null(l)}$ derived in this procedure under the null hypothesis, we can derive the asymptotic properties of the scores in both Algorithm 1 and 2.

In addition to Algorithm 1 and 2, following the Discussion under Corollary 5, we also consider the following inference procedure. Take the modification of Algorithm 2 as an example, we split the data in $\widetilde{J}$ into K folds denoted as $\widetilde{J}_1,\cdots,\widetilde{J}_K$; we obtain the estimator $\widehat{\boldsymbol{\beta}}_\phi^{(-\widetilde{J}_k,-\widetilde{J}_{k'})}$ by minimizing

$$\widehat{\mathbb{E}}_n^{(-\widetilde{J}_k,-\widetilde{J}_{k'})}\left[\sum_{a\in\{1,-1\}}\widehat{W}_a^{(\widetilde{I})}\phi(a\boldsymbol{X}^\top\boldsymbol{\beta})\right] + \lambda_n\|\boldsymbol{\beta}\|_1,$$

where $\widehat{\mathbb{E}}_n^{(-\widetilde{J}_k,-\widetilde{J}_{k'})}[\cdot]$ represents the empirical average over the data excluding $\widetilde{J}_k\cup\widetilde{J}_{k'}$, and obtain $\widehat{\boldsymbol{\beta}}_\phi^{(-\widetilde{J}_k)}$ by minimizing

$$\widehat{\mathbb{E}}_n^{(-\widetilde{J}_k)}\left[\sum_{a\in\{1,-1\}}\widehat{W}_a^{(\widetilde{I})}\phi(a\boldsymbol{X}^\top\boldsymbol{\beta})\right] + \lambda_n\|\boldsymbol{\beta}\|_1.$$

Further, we estimate $\widehat{\boldsymbol{w}}_{\phi,l}^{(-\widetilde{J}_k)}$ by minimizing

$$\frac{1}{K-1}\sum_{k'\neq k}\widehat{\mathbb{E}}_n^{(\widetilde{J}_{k'})}\left[\left\{\sum_{a\in\{1,-1\}}\sum_{j=1}^{J}\Delta_j\widehat{W}_a^{(\widetilde{I})}G_{h_g}(t_j - a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-\widetilde{J}_k,-\widetilde{J}_{k'})})\right\}(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_l)^2\right]$$
$$+\mu_n\|\boldsymbol{w}_l\|_1.$$

Then we construct the kernel-smoothed decorrelated score by

$$S^{(\tilde{J})}_{\tilde{\phi}',null(l)} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(\tilde{J}_k)}\left[\sum_{a\in\{-1,1\}} a\widehat{W}_a^{(\tilde{I})}\tilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-\tilde{J}_k)}_{\phi,null(l)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}^{(-\tilde{J}_k)}_{\phi,l})\right].$$

Compared with the Algorithm 2, the key difference is that the data evaluating the score is independent with $\widehat{\boldsymbol{w}}^{(-\tilde{J}_k)}_{\phi,l}$ in the modified procedure.

To start with, let

$$S^{(k)}_{\tilde{\phi}',null(l)} = \widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{-1,1\}} a\widehat{W}_a\tilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\right)(X_1 - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right].$$

We consider the asymptotic property of this quantity given the assumption that $\widehat{W}_a$ is estimated indepdently; $\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$;$\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ is correlated with $\widehat{\boldsymbol{w}}_{\phi,l}$ (Algorithm 1 and 2) or $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ is independent with $\widehat{\boldsymbol{w}}_{\phi,l}$ (modified algorithm). We decompose the constructed score into three terms,

$$\begin{aligned}
S^{(k)}_{\tilde{\phi}',null(l)} &= \widehat{\mathbb{E}}_n^{(k)}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\tilde{\phi}'\Big(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)}\Big) \\
&\qquad - \sum_{a\in\{-1,1\}} a\widehat{W}_a\tilde{\phi}'\big(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\big)\Big\}(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\Bigg] \\
&\quad + \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\tilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right)\right\}\boldsymbol{X}_{-l}^\top(\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l})\right] \\
&\quad + \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\tilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right)\right\}(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})\right] \\
&= I_1 + I_2 + I_3.
\end{aligned}$$

Under the null hypothesis, we will show that $\max_l \sqrt{n}I_1$ and $\max_l \sqrt{n}I_2$ are $o_p(1)$, and $\sqrt{n}I_3$ converges to a Gaussian distribution uniformly in l's. First, we will show the asymptotic

distribution of I_3. Consider

$$\begin{aligned} I_3 &= \widehat{\mathbb{E}}_n^{(k)}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\Big(\widetilde{\phi}'(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*) - \Delta_0 \\ &\qquad - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)\Big\}(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\Bigg] \\ &\quad + \widehat{\mathbb{E}}_n^{(k)}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a(\widehat{W}_a - \overline{W}_a)\Big(\Delta_0 + \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)\Big\} \\ &\qquad \cdot(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\Bigg] \\ &\quad + \widehat{\mathbb{E}}_n^{(k)}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\overline{W}_a\Big(\Delta_0 + \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)\Big\} \\ &\qquad \cdot(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\Bigg] \\ &= I_{31} + I_{32} + I_{33}. \end{aligned}$$

For I_{33}, it coverages to a Gaussian distribution. For I_{31}, by Bernstein inequality, we have

$$\begin{aligned} &\mathbb{P}\Bigg\{\Bigg|(\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E})\Bigg[\Bigg\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\Bigg[\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right) - \Delta_0 - \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\Bigg]\Bigg\} \\ &(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\Bigg]\Bigg| \geqslant t\Bigg\} \leqslant \exp\left\{-\frac{nt^2}{2(V_1 + M_1 t/3)}\right\}, \end{aligned}$$

where

$$\begin{aligned} V_1 = \mathbb{E}\Bigg[&\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\Big(\widetilde{\phi}'(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*) - \Delta_0 \\ &- \sum_{j=1}^J \Delta_j\, 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)\Big\}\,(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\Bigg]^2, \end{aligned}$$

and M_1 is the upper-bound of

$$\left|\left\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\left[\widetilde{\phi}'\,(at) - \Delta_0 - \sum_{j=1}^J \Delta_j 1\,\{at - t_j \geqslant 0\}\right]\right\}(X_1 - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right|$$

on $t \in [-h_l, h_l]$. Direct calculation of V_1 yields that for a sufficient large constant C_1,

$$V_1 \leqslant C_1 R^2 \mathbb{E}\left[\sum_{a\in\{-1,1\}} a\left[\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right) - \Delta_0 - \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right]^2.$$

Notice that $\widetilde{\phi}'(t)$ has the same value as $\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\{t \geqslant t_j\}$ if $t - t_j \notin [-h_l, h_l]$. By Condition (b), we have $V_1 \leqslant 4C_1 R^2 h_l^\gamma$. Thus, the inequality yields that

$$\begin{aligned}
& \widehat{\mathbb{E}}_n^{(k)}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\Big(\widetilde{\phi}'(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*) - \Delta_0 \\
& \qquad - \sum_{j=1}^{J}\Delta_j 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)\Big\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\Bigg] \\
= \;& \mathbb{E}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\Big(\widetilde{\phi}'(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*) - \Delta_0 \\
& \qquad - \sum_{j=1}^{J}\Delta_j 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)\Big\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\Bigg] \\
& + O_p\left(R\sqrt{\frac{h_l^\gamma \log p}{n}}\right).
\end{aligned}$$

uniformly holds over all l's.

Further, Lemma 9 yields

$$\begin{aligned}
& \max_l \Bigg|\mathbb{E}\left[\left\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\left[\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right) - \Delta_0 - \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}\right. \\
& \left.(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right]\Bigg| = O(Rh_l).
\end{aligned}$$

Thus, we have

$$I_{31} = O(Rh_l) + O_p\left(R\sqrt{\frac{h_l^\gamma \log p}{n}}\right).$$

For I_{32}, by Hoeffding's inequality and the definition of $\overline{W}_a$, we have

$$\begin{aligned}
& \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}} a\left[\widehat{W}_a - \overline{W}_a\right]\left[\Delta_0 + \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right] \\
= \;& \mathbb{E}\left[\left\{\sum_{a\in\{-1,1\}} a\left[\widehat{W}_a - \overline{W}_a\right]\left[\Delta_0 + \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right] \\
& + o_p(Rn^{-\zeta}\sqrt{\log p/n}),
\end{aligned}$$

uniformly holds over all l's.

Further, under Condition (e), we have

$$\begin{aligned}
&\max_l \mathbb{E}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a(\widehat{W}_a - \overline{W}_a)\Big(\Delta_0 + \sum_{j=1}^{J}\Delta_j 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j \geqslant 0\}\Big)\Big\} \\
&\qquad\qquad \times(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})\Bigg] \\
= \;& O_p(Rn^{-\eta}).
\end{aligned}$$

As a special case, for a common classification problem, we have that $\widehat{W}_a = \overline{W}_a$; that is, in this case, $I_{32} = 0$ (or $\eta = \infty$). Given $Rn^{-\eta+1/2} \to 0$ and $Rn^{-\zeta}\sqrt{\log p} = O(1)$, we have $\max_l \sqrt{n} I_{32} \to 0$. Combining with results for I_{31} and I_{33}, we have that

$$\max_l \sqrt{n}|I_3 - I_{33}| = o_p(1).$$

In addition , $\sqrt{n} I_{33} \to N(0, \sigma_l^2)$, where

$$\sigma_l^2 = \mathbb{E}\left[\left\{\sum_{a\in\{-1,1\}} a\overline{W}_a\left[\Delta_0 + \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j \geqslant 0\right\}\right]\right\}^2 (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})^2\right].$$

For I_2, we have

$$\begin{aligned}
I_2 \;=\; & \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\left[\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right) - \Delta_0 - \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j \geqslant 0\right\}\right]\right\}\right. \\
& \left.\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l})\right] \\
& + \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}} a\left[\widehat{W}_a - \overline{W}_a\right]\left[\Delta_0 + \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j \geqslant 0\right\}\right]\right\}\right. \\
& \left.\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l})\right] \\
& + \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}} a\overline{W}_a\left[\Delta_0 + \sum_{j=1}^{J}\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi - t_j \geqslant 0\right\}\right]\right\}\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}^*_{\phi,l})\right] \\
= \;& I_{21} + I_{22} + I_{23}.
\end{aligned}$$

For I_{21}, when $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, by Bernstein inequality, we have

$$\begin{aligned}
\max_l |I_{21}| \leqslant & \left\| \widehat{\mathbb{E}}_n^{(k)} \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a(\widehat{W}_a - \overline{W}_a) \Big\{ \widetilde{\phi}'(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*) - \Delta_0 \right. \\
& \left. - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\} \Big\} \Big\} \boldsymbol{X} \Bigg] \right\|_\infty \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \\
& + \left\| \widehat{\mathbb{E}}_n^{(k)} \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \Big\{ \widetilde{\phi}'(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*) - \Delta_0 \right. \\
& \left. - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\} \Big\} \Big\} \boldsymbol{X} \Bigg] \right\|_\infty \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \\
\leqslant & \Big(o_p(\sqrt{h_l^\gamma \log p/n}) + O_p(n^{-\eta}) \Big) \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \\
& + O_p(\sqrt{h_l^\gamma \log p/n} + h_l) \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \\
= & O_p\Big((n^{-\eta} + \sqrt{h_l^\gamma \log p/n} + h_l) \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \Big).
\end{aligned}$$

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, by Bernstein inequality and notice that

$$\max_l \frac{\left\| \widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^* \right\|_1}{\left\| \widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^* \right\|_2} \leqslant \sqrt{s'},$$

we have

$$\begin{aligned}
& \max_l |I_{21}| \\
\leqslant & \max_l \left| \widehat{\mathbb{E}}_n^{(k)} \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a(\widehat{W}_a - \overline{W}_a) \Big\{ \widetilde{\phi}'(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*) - \Delta_0 \right. \\
& \left. - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\} \Big\} \Big\} \boldsymbol{X}_{-l}^\top \frac{\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*}{\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2} \Bigg] \right| \max_j \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2 \\
& + \max_l \left| \widehat{\mathbb{E}}_n^{(k)} \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \Big\{ \widetilde{\phi}'(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*) - \Delta_0 \right. \\
& \left. - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\} \Big\} \Big\} \boldsymbol{X}_{-l}^\top \frac{\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*}{\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2} \Bigg] \right| \max_j \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2 \\
\leqslant & \Big(o_p(\sqrt{h_l^{\gamma/2} \log p/n}(1 \vee \sqrt{s' \log p/(nh_l^{\gamma/2})})) + O_p(n^{-\eta}) \Big) \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 \\
& + O_p\Big(\sqrt{h_l^{\gamma/2} \log p/n}(1 \vee \sqrt{s' \log p/(nh_l^{\gamma/2})}) + h_l \Big) \max_j \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2 \\
= & O_p\Big((n^{-\eta} + \sqrt{h_l^{\gamma/2} \log p/n}(1 \vee \sqrt{s' \log p/(nh_l^{\gamma/2})}) + h_l) \max_j \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2 \Big).
\end{aligned}$$

uniformly holds over l's. Assuming that $n^{1/2}(s'\mu_n)((n^{-\eta}+\sqrt{h_l^{\gamma}\log p/n}+h_l)) \to 0$ when $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ (Algorithm 1 and 2), and $n^{1/2}(\sqrt{s'}\mu_n)(n^{-\eta}+\sqrt{h_l^{\gamma/2}/n}(1\vee\sqrt{s'/(nh_l^{\gamma/2})})+h_l) \to 0$ when $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ (modified algorithm), we have that $\max_l \sqrt{n}|I_{21}| = o_p(1)$. For I_{22} and I_{23}, we also separate the discussion. When $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ (Algorithm 1 and 2), by Hoeffding's inequality, we have

$$
\begin{aligned}
&\max_l |I_{22}| \\
\leqslant\ & \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{-1,1\}} a\left[\widehat{W}_a - \overline{W}_a\right] \left[\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right] \right\} \boldsymbol{X} \right] \right\|_\infty \\
&\max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_1 \\
=\ & \left(o_p\left(\sqrt{\log p/n}\right) + O_p(n^{-\eta})\right) \max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_1 ;
\end{aligned}
$$

$$
\begin{aligned}
&\max_l |I_{23}| \\
\leqslant\ & \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \left[\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right] \right\} \boldsymbol{X} \right] \right\|_\infty \\
&\max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_1 \\
=\ & O_p\left(\sqrt{\log p/n}\right) \max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_1 .
\end{aligned}
$$

Thus, assuming that $n^{1/2}(s'\mu_n)(\sqrt{\log p/n}+n^{-\eta}) \to 0$, we have that $\max_l \sqrt{n}|I_{22}| = o_p(1)$ and $\max_l \sqrt{n}|I_{23}| = o_p(1)$.

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ (modified algorithm), by Bernstein inequality, we have

$$
\begin{aligned}
&\max_l |I_{22}| \\
\leqslant\ & \max_l \left| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\left\{ \sum_{a\in\{-1,1\}} a\left[\widehat{W}_a - \overline{W}_a\right] \left[\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right] \right\} \right. \right. \\
&\left. \left. \boldsymbol{X}_{-l}^\top \frac{\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*}{\left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_2} \right] \right| \max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_2 \\
&+ \max_l \left| \mathbb{E} \left[\left\{ \sum_{a\in\{-1,1\}} a\left[\widehat{W}_a - \overline{W}_a\right] \left[\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right] \right\} \right. \right. \\
&\left. \left. \boldsymbol{X}_{-l}^\top \frac{\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*}{\left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_2} \right] \right| \max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_2 \\
=\ & \left(o_p\left((1\vee\sqrt{s'\log p/n})\sqrt{\log p/n}\right) + O_p(n^{-\eta})\right) \max_l \left\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\right\|_2 ;
\end{aligned}
$$

$$
\begin{aligned}
& \max_l |I_{23}| \\
\leqslant\ & \max_l \left| \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \left[\Delta_0 + \sum_{j=1}^{J} \Delta_j 1\left\{ a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* - t_j \geqslant 0 \right\} \right] \right\} \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
=\ & O_p\left((1 \vee \sqrt{s' \log p/n}) \sqrt{\log p/n} \right) \max_l \left\| \widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^* \right\|_2 .
\end{aligned}
$$

Thus, assuming that $n^{1/2}(\sqrt{s'}\mu_n)((1 \vee \sqrt{s' \log p/n})\sqrt{\log p/n} + n^{-\eta}) \to 0$, we have that $\max_l \sqrt{n}|I_{22}| = o_p(1)$ and $\max_l \sqrt{n}|I_{23}| = o_p(1)$.

For I_1, we have

$$
\begin{aligned}
I_1 =\ & \widehat{\mathbb{E}}_n^{(k)} \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} \Big) \\
& \qquad - \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\big(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \big) \Big\} \boldsymbol{X}_{-l}^\top (\boldsymbol{w}_{\phi,l}^* - \widehat{\boldsymbol{w}}_{\phi,l}) \Bigg] \\
& + \widehat{\mathbb{E}}_n^{(k)} \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} \Big) \\
& \qquad - \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\big(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \big) \Big\} (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \Bigg] \\
=\ & I_{11} + I_{12}.
\end{aligned}
$$

Because that I_{11} is related to the estimation error of $\widehat{\boldsymbol{w}}_{\phi,l}$, we will discuss separately depending on the relationship between $\widehat{\boldsymbol{w}}_{\phi,l}$ and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$. When $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, we have

$$
\begin{aligned}
& \max_l |I_{11}| \\
\leqslant\ & \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}'\left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} \right) - \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}'\left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right\} \boldsymbol{X} \right] \right\|_\infty \\
& \max_l \left\| \boldsymbol{w}_{\phi,l}^* - \widehat{\boldsymbol{w}}_{\phi,l} \right\|_1 \\
& + \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{-1,1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \left(\widetilde{\phi}'\left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} \right) - \widetilde{\phi}'\left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right) \right\} \boldsymbol{X} \right] \right\|_\infty \\
& \max_l \left\| \boldsymbol{w}_{\phi,l}^* - \widehat{\boldsymbol{w}}_{\phi,l} \right\|_1 \\
=\ & I_{111} + I_{112}.
\end{aligned}
$$

Under Condition (C2), by Hoeffding's inequality, we have that

$$
I_{112} = o_p(\sqrt{\log p/n} \max_l \left\| \boldsymbol{w}_{\phi,l}^* - \widehat{\boldsymbol{w}}_{\phi,l} \right\|_1) + O_p(n^{-\eta} \max_l \left\| \boldsymbol{w}_{\phi,l}^* - \widehat{\boldsymbol{w}}_{\phi,l} \right\|_1).
$$

For I_{111}, by Taylor's expansion and Lemma 10, we have that

$$\begin{aligned}
& I_{111} \\
\leqslant & \left\| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top\widetilde{\boldsymbol{\beta}}\right)\right\} \boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,\text{null}} - \boldsymbol{\beta}^*_\phi)\boldsymbol{X}\right]\right\|_\infty \max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1 \\
& + O_p(\sqrt{\log p/n}\max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1) \\
\leqslant & \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top\widetilde{\boldsymbol{\beta}}\right)\right\} \boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,\text{null}} - \boldsymbol{\beta}^*_\phi)\boldsymbol{X}^\top\boldsymbol{\nu}\right]\right| \max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1 \\
& + O_p(\sqrt{\log p/n}\max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l}\right\|_1) \\
\leqslant & \left(\left|\mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top\widetilde{\boldsymbol{\beta}}\right)\right\} \left(\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,\text{null}} - \boldsymbol{\beta}^*_\phi)\right)^2\right]\right|^{1/2}\right. \\
& \left. \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top\widetilde{\boldsymbol{\beta}}\right)\right\} \left(\boldsymbol{X}^\top\boldsymbol{\nu}\right)^2\right]\right|^{1/2}\right) \max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1 \\
& + O_p(\sqrt{\log p/n}\left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1) \\
\leqslant & O_p(\Delta_{\beta,2}\max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1) + O_p(\sqrt{\log p/n}\max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}\right\|_1)
\end{aligned}$$

Thus, assuming that $n^{1/2}(s'\mu_n)(\Delta_{\beta,2} + \sqrt{\log p/n}) \to 0$, we have $\max_l \sqrt{n}|I_{11}| \to 0$ in probability.

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}^{(k)}_n[\cdot]$, we have

$$\begin{aligned}
& I_{11} \\
\leqslant & \left|\widehat{\mathbb{E}}^{(k)}_n\left[\Big\{ \sum_{a\in\{-1,1\}} a\overline{W}_a\widetilde{\phi}'\Big(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\Big)\right.\right. \\
& \qquad \left.\left. - \sum_{a\in\{-1,1\}} a\overline{W}_a\widetilde{\phi}'\big(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\big)\Big\}\boldsymbol{X}^\top_{-l}(\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l})\right]\right| \\
& + \left|\widehat{\mathbb{E}}^{(k)}_n\left[\Big\{ \sum_{a\in\{-1,1\}} a(\widehat{W}_a - \overline{W}_a)\Big(\widetilde{\phi}'\Big(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\Big)\right.\right. \\
& \qquad \left.\left. -\widetilde{\phi}'\big(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\big)\Big)\Big\}\boldsymbol{X}^\top_{-l}(\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l})\right]\right| \\
= & I_{111} + I_{112}.
\end{aligned}$$

Under Condition (C2), by Bernstein's inequality, we have that

$$\begin{aligned}
\max_l |I_{112}| &= o_p((1\vee\sqrt{s'\log p/n})\sqrt{\log p/n}\max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l}\right\|_2) \\
& \quad + O_p(n^{-\eta}\max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l}\right\|_2).
\end{aligned}$$

For I_{111}, by Bernstein's inequality, Taylor's expansion and Lemma 10, we have that

$$\begin{aligned}
& \max_l |I_{111}| \\
\leqslant\ & \max_l \left| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} \boldsymbol{X}^\top (\widehat{\boldsymbol{\beta}}_{\phi,\text{null}} - \boldsymbol{\beta}^*_\phi) \boldsymbol{X}_{-l}^\top (\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l}) \right] \right| \\
& + O_p((1 \vee \sqrt{s' \log p/n}) \sqrt{\log p/n}\, \|\boldsymbol{w}^* - \widehat{\boldsymbol{w}}\|_2) \\
\leqslant\ & \left(\left| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} \left(\boldsymbol{X}^\top (\widehat{\boldsymbol{\beta}}_{\phi,\text{null}} - \boldsymbol{\beta}^*_\phi)\right)^2 \right] \right|^{1/2} \right. \\
& \left. \max_l \left| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} \left(\boldsymbol{X}_{-l}^\top (\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l})\right)^2 \right] \right|^{1/2} \right) \\
& + O_p((1 \vee \sqrt{s' \log p/n}) \sqrt{\log p/n} \max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l}\right\|_2) \\
\leqslant\ & O_p(\Delta_{\beta,2} \max_l \left\|\boldsymbol{w}^*_{\phi} - \widehat{\boldsymbol{w}}^*_{\phi,l}\right\|_2) + O_p((1 \vee \sqrt{s' \log p/n}) \sqrt{\log p/n} \max_l \left\|\boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}^*_{\phi,l}\right\|_2)
\end{aligned}$$

Thus, assuming that $n^{1/2}(\sqrt{s'}\mu_n)(\Delta_{\beta,2} + (1 \vee \sqrt{s' \log p/n})\sqrt{\log p/n}) \to 0$, we have

$$\max_l \sqrt{n} |I_{11}| \to 0,$$

in probability.

Now, we will focus on I_{12}.

$$\begin{aligned}
& I_{12} \\
=\ & (\widehat{\mathbb{E}}_n - \mathbb{E}) \Bigg[\Big\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\Big) \\
& \qquad - \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\big(a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\big) \Big\} (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l}) \Bigg] \\
& + \mathbb{E}\Bigg[\Big\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\Big) \\
& \qquad - \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}'\big(a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\big) \Big\} (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l}) \Bigg] \\
=\ & I_{121} + I_{122}.
\end{aligned}$$

For I_{122}, by Condition (C2) and Taylor expansion, we have

$$\begin{aligned}
& I_{122} \\
= \; & \mathbb{E}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}'\Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\Big) \\
& \qquad - \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}'\big(a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\big)\Big\}\,(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})\Bigg] \\
& + O_p(Rn^{-\eta}) \\
= \; & \mathbb{E}\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\Big(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\Big)\Big\} \\
& \qquad \boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{\beta}}_{\phi,-1} - \boldsymbol{\beta}^*_{\phi,-1})\,(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})\Bigg] + O_p(Rn^{-\eta}).
\end{aligned}$$

uniformly holds over all l's.

By Lemma 10, we can show that

$$\begin{aligned}
& \max_l \mathbb{E}\left[\left\{\sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right)\right\} \boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{\beta}}_{\phi,-1} - \boldsymbol{\beta}^*_{\phi,-1})(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})\right] \\
= \; & O_p(R(h_l + \Delta_{\beta,2})\Delta_{\beta,2}).
\end{aligned}$$

Due to $n^{1/2}R((h_l + \Delta_{\beta,2})\Delta_{\beta,2} + n^{-\eta}) \to 0$, we have that $\max_l |I_{122}| = o_p(1/\sqrt{n})$. For I_{121}, we have the following.

$$\begin{aligned}
& I_{121} \\
= \; & (\widehat{\mathbb{E}}^{(k)}_n - \mathbb{E})\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a\Big(\widetilde{\phi}'\Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null}\Big) - \Delta_0 \\
& \qquad - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null} - t_j \geqslant 0\}\Big) \\
& \qquad - \sum_{a\in\{-1,1\}} a\widehat{W}_a\Big(\widetilde{\phi}'\big(a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi\big) - \Delta_0 \\
& \qquad - \sum_{j=1}^J \Delta_j 1\{a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi - t_j \geqslant 0\}\Big)\Big\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})\Bigg] \\
& + (\widehat{\mathbb{E}}^{(k)}_n - \mathbb{E})\Bigg[\Big\{\sum_{a\in\{-1,1\}} a\widehat{W}_a \sum_{j=1}^J \Delta_j\Big(1\{a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null} - t_j \geqslant 0\} \\
& \qquad - 1\{a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi - t_j \geqslant 0\}\Big)\Big\}\,(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})\Bigg] \\
= \; & I_{1211} + I_{1212}.
\end{aligned}$$

By Bernstein inequality, similar to I_{31}, we have that

$$\max_l |I_{1211}| \lesssim O_p\left(R\sqrt{\frac{h^{\gamma/2}\log p}{n}} \vee R\frac{\log p}{n}\right).$$

For I_{1212}, by Bernstein inequality, we have

$$\mathbb{P}\left\{\left|(\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E})\left[\left\{a\widehat{W}_a \sum_{j=1}^J \Delta_j\left(1\left\{a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - t_j \geqslant 0\right\} - 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right)\right\}\right.\right.$$
$$\left.\left.(X_1 - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right]\right| \geqslant t\right\} \leqslant \exp\left\{-\frac{nt^2}{2(V_2 + M_2 t/3)}\right\},$$

where

$$\begin{aligned} V_2 &= \mathbb{E}\Bigg[\widehat{W}_a \sum_{j=1}^J \Delta_j\Big(1\{a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - t_j \geqslant 0\} \\ &\qquad -1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big)^2 (X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)^2\Bigg]. \end{aligned}$$

and $M_2 = O(R)$ is some constant. Notice that by Condition (C3), we have

$$\begin{aligned} V_2 &\lesssim 2^J R^2 \sum_{j=1}^J \mathbb{E}\Big[1\{a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - t_j \geqslant 0\} - 1\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\}\Big]^2 \\ &\leqslant 2^J R^2 \sum_{j=1}^J \mathbb{P}\Big(|t_j - \boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*| \leqslant |\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - \boldsymbol{\beta}_\phi^*)|\Big) \\ &\leqslant 2^J R^2 \sum_{j=1}^J \Big(\mathbb{P}\big(s \leqslant |t_j - \boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*| \leqslant |\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - \boldsymbol{\beta}_\phi^*)|\big) \\ &\qquad +\mathbb{P}\big(|t_j - \boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*| \leqslant s\big)\Big) \\ &\leqslant 2^J R^2 \sum_{j=1}^J \Big(\mathbb{E}\big[|\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - \boldsymbol{\beta}_\phi^*)|^2\big]/s^2 + \tau s^\gamma\Big). \end{aligned}$$

Taking $s = O(\|\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - \boldsymbol{\beta}_\phi^*\|_2^{2/(\gamma+2)})$, we have

$$\max_l V_2 \leqslant CJR^2\|\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - \boldsymbol{\beta}_\phi^*\|_2^{2\gamma/(\gamma+2)}.$$

Thus, we have

$$\max_l |I_{1212}| = O_p\left(R\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}_{\phi,null}^{(-k)} - \boldsymbol{\beta}_\phi^*\|_2^{2\gamma/(\gamma+2)} \vee R\frac{\log p}{n}\right).$$

To summarize the results for I_1, I_2, and I_3, when $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, assuming that $\sqrt{n}R(h_l+n^{-\eta}+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)})\to 0$ and $\sqrt{n}(s'\mu_n)(n^{-\eta}+\sqrt{\log p/n}+h_l+\Delta_{\beta,2})\to 0$, we have that $\max_l\sqrt{n}|I_1|=o_p(1)$, $\max_l\sqrt{n}|I_2|=o_p(1)$ and

$$\sqrt{n}\max_l\left|S^{(k)}_{\widetilde{\phi}',null(l)}-\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}}a\overline{W}_a\left[\Delta_0+\sum_{j=1}^J\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi-t_j\geqslant 0\right\}\right]\right\}\right.$$
$$\left.(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})\right]\Big|=o_p(1).$$

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, assuming that

$$\sqrt{n}R(h_l+n^{-\eta}+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)})\to 0,$$

and

$$\sqrt{n}(\sqrt{s'}\,\mu_n)\Big(n^{-\eta}+(1\vee\sqrt{s'\log p/n})\sqrt{\log p/n}+h_l+\Delta_{\beta,2}$$
$$+\sqrt{h_l^{\gamma/2}\log p/n}\,(1\vee\sqrt{s'\log p/(nh_l^{\gamma/2})})\Big)\to 0,$$

we have that $\max_l\sqrt{n}|I_1|=o_p(1)$, $\max_l\sqrt{n}|I_2|=o_p(1)$ and

$$\sqrt{n}\max_l\left|S^{(k)}_{\widetilde{\phi}',null(l)}-\widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}}a\overline{W}_a\left[\Delta_0+\sum_{j=1}^J\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi-t_j\geqslant 0\right\}\right]\right\}\right.$$
$$\left.(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})\right]\Big|=o_p(1).$$

As a special case, for a common classification problem, when $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, we require that $\sqrt{n}R(h_l+n^{-\eta}+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)})\to 0$ and $\sqrt{n}(s'\mu_n)(\sqrt{\log p/n}+h_l+\Delta_{\beta,2})\to 0$. When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, we require that $\sqrt{n}R(h_l+n^{-\eta}+\sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)})\to 0$ and $\sqrt{n}(\sqrt{s'}\mu_n)((1\vee\sqrt{s'\log p/n})\sqrt{\log p/n}+h_l+\Delta_{\beta,2}+\sqrt{h_l^{\gamma/2}\log p/n}(1\vee\sqrt{s'\log p/(nh_l^{\gamma/2})}))\to 0$.

Under these conditions, we have shown the asymptotic property of $S^{(k)}_{\widetilde{\phi}',null}$. Because $S_{\widetilde{\phi}',null}$ is the average of $S^{(k)}_{\widetilde{\phi}',null}$'s which are asymptotically independent, the $S_{\widetilde{\phi}',null}$ is asymptotically normal under the null hypothesis. Thus, we have

$$\sqrt{n}\max_l\left|S_{\widetilde{\phi}',null(l)}-\widehat{\mathbb{E}}_n\left[\left\{\sum_{a\in\{-1,1\}}a\overline{W}_a\left[\Delta_0+\sum_{j=1}^J\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi-t_j\geqslant 0\right\}\right]\right\}\right.$$
$$\left.(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})\right]\Big|=o_p(1).$$

To conclude, define

$$(\widehat{\sigma}_l^{(k)})^2 = \widehat{\mathbb{E}}_n^{(k)}\left[\left\{\sum_{a\in\{-1,1\}}a\widehat{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)}\right)\right\}^2(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})^2\right],$$

$$\sigma_l^2 = \mathbb{E}\left[\left\{\sum_{a\in\{-1,1\}}a\overline{W}_a\left[\Delta_0+\sum_{j=1}^J\Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi-t_j\geqslant 0\right\}\right]\right\}^2(X_l-\boldsymbol{X}_{-l}^\top\boldsymbol{w}^*_{\phi,l})^2\right],$$

and we will show

$$\max_l |(\widehat{\sigma}_l^{(k)})^2 - \sigma_l^2| = O_p(Rs'\mu_n + R^2 n^{-\zeta} + R^2 h_l).$$

To show this, notice that

$$\begin{aligned}
& (\widehat{\sigma}_l^{(k)})^2 - \sigma_l^2 \\
= \; & \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)} \right) \right\}^2 \left\{ (X_l - \boldsymbol{X}_{-l}^\top \widehat{\boldsymbol{w}}_{\phi,l})^2 - (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})^2 \right\} \right] \\
& + (\widehat{\mathbb{E}}_n^{(k)} - E) \left[\left\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)} \right) \right\}^2 (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})^2 \right] \\
& + \mathbb{E} \left[\left[\left\{ \sum_{a\in\{-1,1\}} a\widehat{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)} \right) \right\}^2 \right. \right. \\
& \left. \left. - \left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \left[\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{ a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi - t_j \geqslant 0 \right\} \right] \right\}^2 \right] (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})^2 \right] \\
= \; & I_{41} + I_{42} + I_{43}.
\end{aligned}$$

We can see that

$$\begin{aligned}
\max_l |I_{42}| &\leqslant CR^2 \sqrt{\log p/n}, \\
\max_l |I_{43}| &\leqslant CR^2 \mathbb{E} \Bigg| \sum_{a\in\{-1,1\}} a\widehat{W}_a \Big\{ \widetilde{\phi}' \Big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)} \Big) \\
& \qquad - \Big[\Delta_0 + \sum_{j=1}^J \Delta_j 1\{ a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi - t_j \geqslant 0 \} \Big] \Big\} \Bigg|.
\end{aligned}$$

Further, we have

$$\begin{aligned}
& \mathbb{E} \left| \sum_{a\in\{-1,1\}} a\widehat{W}_a \left\{ \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)} \right) - \left[\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{ a\boldsymbol{X}^\top \boldsymbol{\beta}^*_\phi - t_j \geqslant 0 \right\} \right] \right\} \right| \\
= \; & O_p(n^{-\zeta} + \Delta_{\beta,2} + h_l).
\end{aligned}$$

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, we have

$$\max_l |I_{41}| \leqslant CR \max_l \left\| \boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l} \right\|_1.$$

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, we have

$$\max_l |I_{41}| \leqslant CR \max_l \left\| \boldsymbol{w}^*_{\phi,l} - \widehat{\boldsymbol{w}}_{\phi,l} \right\|_2.$$

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is correlated with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, assuming $\sqrt{\log p}(Rs'\mu_n + R^2 n^{-\zeta} + R^2 h_l) \to 0$, we have

$$\sqrt{\log p}\max_l |(\widehat{\sigma}_l^{(k)})^2 - \sigma_l^2| = o_p(1).$$

When $\widehat{\boldsymbol{w}}_{\phi,l}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$, assuming $\sqrt{\log p}(R\sqrt{s'}\mu_n + R^2 n^{-\zeta} + R^2 h_l) \to 0$, we have

$$\sqrt{\log p}\max_l |(\widehat{\sigma}_l^{(k)})^2 - \sigma_l^2| = o_p(1).$$

In addition,

$$\min_l \sigma_l^2 \geqslant \lambda_{\min},$$

where $\lambda_{\min}$ is the smallest eigen value of

$$\mathbb{E}\left[\left\{\sum_{a\in\{-1,1\}} a\overline{W}_a\left[\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}^2 \boldsymbol{X}\boldsymbol{X}^\top\right]$$

Combining this result with the uniform convergence of $S_{\widetilde{\phi}',null(l)}$, we have

$$\sqrt{n}\max_l \left|\widehat{\sigma}_l^{-1} S_{\widetilde{\phi}',null(l)} - \sigma_l^{-1}\widehat{\mathbb{E}}_n\left[\left\{\sum_{a\in\{-1,1\}} a\overline{W}_a\left[\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}\right.\right.$$
$$\left.\left.(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right]\right| = o_p(1).$$

By the Berry-Esseen bound for CLT, there exist a universal constant c_0 such that

$$\max_j \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left(\left|\sigma_l^{-1} n^{1/2}\widehat{\mathbb{E}}_n\left[\left\{\sum_{a\in\{-1,1\}} a\overline{W}_a\left[\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}\right.\right.\right.$$
$$\left.\left.\left.(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right]\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) - (1-\alpha)\Big|$$
$$\leqslant \frac{c_0}{\sqrt{n}}\max_l \mathbb{E}[|M_l|^3],$$

where

$$M_l = \left\{\sum_{a\in\{-1,1\}} a\overline{W}_a\left[\Delta_0 + \sum_{j=1}^J \Delta_j 1\left\{a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* - t_j \geqslant 0\right\}\right]\right\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*).$$

By $\max_l \mathbb{E}[|M_l|^3] \leqslant CR^3$, we have

$$\max_l \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1} S_{\widetilde{\phi}',null(l)}\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) - (1-\alpha)\right|$$
$$\leqslant \frac{c_0}{\sqrt{n}} R^3,$$

for a sufficient large c_0. When $\frac{R^3}{\sqrt{n}} \to 0$, we can conclude the theorem. □

Corollary 17 *Assume the same conditions in Theorem 16, we have*

$$\max_l \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1}\widehat{I}_l\left(\widetilde{\beta}_{\phi,l}-\beta^*_{\phi,l}\right)\right|\leqslant \Phi^{-1}(1-\alpha/2)\right)-(1-\alpha)\right| = o_p(1).$$

Proof of Corollary 17. To unify the proof for the construction of confidence interval with and without nuisance parameters. To start with, we review how the confidence interval are constructed in both cases. When there is no nuisance parameters, the one-step de-biased estimator is based on

$$\widetilde{\beta}_{\phi,l} = \bar{\beta}_{\phi,l} - S_{\tilde{\phi}',l}/\widehat{I}_l,$$

where

$$\begin{aligned}
\bar{\beta}_{\phi,l} &= \frac{1}{K}\sum_{k=1}^K \widehat{\beta}_{\phi,l},\\
S_{\tilde{\phi}',l} &= \frac{1}{K}\sum_{k=1}^K \widehat{\mathbb{E}}_n^{(k)}\left[A\widetilde{\phi}'\left(A\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right],\\
\widehat{I}_l &= \frac{1}{K}\sum_{k=1}^K \widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{-1,1\}}\sum_{j=1}^J \Delta_j G_{h_g}\left(t_j - a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)X_l(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right].
\end{aligned}$$

When there exists nuisance parameters, the one-step de-biased estimator is based on

$$\begin{aligned}
\widetilde{\beta}_{\phi,l} &= \left(\widetilde{\beta}_{\phi,l}^{(\tilde{I})}+\widetilde{\beta}_{\phi,l}^{(\tilde{J})}\right)/2\\
\widetilde{\beta}_{\phi,l}^{(\tilde{J})} &= \bar{\beta}_{\phi,l}^{(\tilde{J})} - S^{\tilde{J}}_{\tilde{\phi}',l}/\widehat{I}_l^{(\tilde{J})},
\end{aligned}$$

where

$$\begin{aligned}
\bar{\beta}_{\phi,l}^{(\tilde{J})} &= \frac{1}{K}\sum_{k=1}^K \widehat{\beta}_{\phi,l}^{(-\tilde{J}_k)},\\
S_{\tilde{\phi}',l}^{(\tilde{J})} &= \frac{1}{K}\sum_{k=1}^K \widehat{\mathbb{E}}_n^{(\tilde{J}_k)}\left[\sum_{a\in\{1,-1\}} a\widehat{W}_a^{(\tilde{I})}\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-\tilde{J}_k)}\right)(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l}^{(\tilde{J})})\right],\\
\widehat{I}_l^{(\tilde{J})} &= \frac{1}{K}\sum_{k=1}^K \widehat{\mathbb{E}}_n^{(\tilde{J}_k)}\left[\sum_{a\in\{-1,1\}} \widehat{W}_a^{(\tilde{I})}\sum_{j=1}^J \Delta_j G_{h_g}\left(t_j - a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-\tilde{J}_k)}\right)X_l(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l}^{(\tilde{J})})\right],
\end{aligned}$$

and $\widetilde{\beta}_{\phi,l}^{(\tilde{I})}$, $\bar{\beta}_{\phi,l}^{(\tilde{I})}$, $S^{\tilde{I}}_{\tilde{\phi}',l}$, and $\widehat{I}_l^{(\tilde{I})}$ are defined similarly.

In the modified algorithm 2 as illustrated in the proof of Theorem 1, the major difference is that

$$\begin{aligned}
S_{\tilde{\phi}',l}^{(\tilde{J})} &= \frac{1}{K}\sum_{k=1}^K \widehat{\mathbb{E}}_n^{(\tilde{J}_k)}\left[\sum_{a\in\{1,-1\}} a\widehat{W}_a^{(\tilde{I})}\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-\tilde{J}_k)}\right)(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l}^{(-\tilde{J}_k)})\right],\\
\widehat{I}_l^{(\tilde{J})} &= \frac{1}{K}\sum_{k=1}^K \widehat{\mathbb{E}}_n^{(\tilde{J}_k)}\left[\sum_{a\in\{-1,1\}} \widehat{W}_a^{(\tilde{I})}\sum_{j=1}^J \Delta_j G_{h_g}\left(t_j - a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-\tilde{J}_k)}\right)X_l(X_l-\boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l}^{(-\tilde{J}_k)})\right],
\end{aligned}$$

where $\widehat{\mathbb{E}}_n^{(\tilde{J}_k)}[\cdot]$ and $\widehat{\boldsymbol{w}}_{\phi,l}^{(-\tilde{J}_k)}$ are independent. On the contrast, in the construction of one-step de-biased estimator in Algorithm 2, $\widehat{\mathbb{E}}_n^{(\tilde{J}_k)}[\cdot]$ and $\widehat{\boldsymbol{w}}_{\phi,l}^{(\tilde{J})}$ are correlated.

To unify the proofs, we consider the same studied procedure in the proof of Theorem 1. The one-step de-biased estimator is constructed by

$$\widetilde{\beta}_{\phi,l} \quad = \quad \bar{\beta}_{\phi,l} - S_{\tilde{\phi}',l}/\widehat{I}_l,$$

where

$$\begin{aligned}
\bar{\beta}_{\phi,l} &= \frac{1}{K}\sum_{k=1}^{K}\widehat{\beta}_{\phi,l}^{(-k)}, \\
S_{\tilde{\phi}',l} &= \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\widehat{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right], \\
\widehat{I}_l &= \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{-1,1\}} a\widehat{W}_a\sum_{j=1}^{J}\Delta_j G_{h_g}\left(t_j - a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)X_l(X_l - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right].
\end{aligned}$$

We assume that $\widehat{W}_a$ is estimated on an independent dataset; $\widehat{\boldsymbol{\beta}}_\phi^{(-k)}$ is independent with $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$;$\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ is correlated with $\widehat{\boldsymbol{w}}_{\phi,l}$ (Algorithm 1 and 2) or $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ is independent with $\widehat{\boldsymbol{w}}_{\phi,l}$ (modified algorithm). By investigation on the asymptotic property of $\widetilde{\beta}_{\phi,1}$, we can show the asymptotic properties for both Algorithm 1 and 2 as well as the modified algorithm.

Define

$$S_{\tilde{\phi}'}^{(k)} \quad = \quad \widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\widehat{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)(X_1 - \boldsymbol{X}_{-l}^\top\widehat{\boldsymbol{w}}_{\phi,l})\right].$$

We will study the asymptotic property of

$$\widetilde{\beta}_{\phi,l}^{(k)} = \widehat{\beta}_{\phi,l}^{(-k)} - S_{\tilde{\phi}',l}^{(k)}/\widehat{I}_l,$$

because that $\widetilde{\beta}_{\phi,l} = \sum_{k=1}^{K}\widetilde{\beta}_{\phi,l}^{(k)}$ and $\widetilde{\beta}_{\phi,l}^{(k)}$ are asymptotically independent.

Specifically, we will firstly show that

$$\max_l n^{1/2}\left|I_l^*\left(\widetilde{\beta}_{\phi,l}^{(k)} - \beta_{\phi,l}^*\right) + \widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right]\right| = o_p(1),$$

where

$$I_l^* = \mathbb{E}\left[\sum_{a\in\{1,-1\}}\overline{W}_a\sum_{j=1}^{J}\Delta_j\delta(t_j - a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)X_l(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right].$$

Notice that $\widetilde{\beta}_{\phi,l}^{(k)} = \widehat{\beta}_{\phi,l}^{(k)} - S_{\widetilde{\phi}',l}^{(k)}/\widehat{I}_l$, we need to show that

$$\max_l n^{1/2} \left| I_l^* \left(\widehat{\beta}_{\phi,l}^{(k)} - \beta_{\phi,l}^* \right) + \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] \right.$$
$$\left. - S_{\widetilde{\phi}',l}^{(k)} I_l^* \widehat{I}_l^{-1} \right| = o_p(1).$$

To show this, we will show the following results:

$$\max_l n^{1/2} \left| I_l^* \big(\widehat{\beta}_{\phi,l}^{(k)} - \beta_{\phi,l}^* \big) + \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}' \big(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \big) (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] \right.$$
$$\left. - \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\widehat{W}_a \widetilde{\phi}' \big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \big) (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] \right| = o_p(1), \tag{6}$$

$$\max_l n^{1/2} \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\widehat{W}_a \widetilde{\phi}' \big(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \big) \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| = o_p(1), \tag{7}$$

$$\max_l n^{1/2} \left| S_{\widetilde{\phi}',l}^{(k)} \big(I_l^* - \widehat{I}_l \big) \right| = o_p(1). \tag{8}$$

For (7) and (8), because both terms involves $\widehat{w}_{\phi,l}$, we separate the discussion based on the relationship between $\widehat{w}_{\phi,l}$ and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$.

First, we consider that $\widehat{w}_{\phi,l}$ and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ are correlated. For (7), we have

$$\begin{aligned}
(7) \quad \leqslant \quad & \max_l n^{1/2} \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
& + \max_l n^{1/2} \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \left\{ \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) - \widetilde{\phi}' \left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right\} \right. \right. \\
& \left. \left. \times \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
& + \max_l n^{1/2} \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
\leqslant \quad & n^{1/2} \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \boldsymbol{X} \right] \right\|_\infty \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \\
& + n^{1/2} \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \left\{ \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) - \widetilde{\phi}' \left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right\} \boldsymbol{X} \right] \right\|_\infty \\
& \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1 \\
& + n^{1/2} \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \boldsymbol{X} \right] \right\|_\infty \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_1.
\end{aligned}$$

For the first term, from the proof of Theorem 16, we have

$$\begin{aligned}
& \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \widetilde{\phi}' \left(a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \boldsymbol{X} \right] \right\|_\infty \\
\leqslant & \left\| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\sum_{a\in\{1,-1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \widetilde{\phi}' \left(a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \boldsymbol{X} \right] \right\|_\infty \\
& + \left\| \mathbb{E} \left[\sum_{a\in\{1,-1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \widetilde{\phi}' \left(a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \boldsymbol{X} \right] \right\|_\infty \\
= & o_p(\sqrt{\log p/n}) + O(n^{-\eta}).
\end{aligned}$$

For the second term, by the Hoeffding's inequality and Lemma 10, we have

$$\begin{aligned}
& \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a \overline{W}_a \left\{ \widetilde{\phi}' \left(a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) - \widetilde{\phi}' \left(a \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right\} \boldsymbol{X} \right] \right\|_\infty \\
\leqslant & \left\| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\sum_{a\in\{1,-1\}} a \overline{W}_a \left\{ \widetilde{\phi}' \left(a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) - \widetilde{\phi}' \left(a \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right\} \boldsymbol{X} \right] \right\|_\infty \\
& + \left\| \mathbb{E} \left[\sum_{a\in\{1,-1\}} a \overline{W}_a \left\{ \widetilde{\phi}' \left(a \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) - \widetilde{\phi}' \left(a \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \right\} \boldsymbol{X} \right] \right\|_\infty \\
= & O_p(\sqrt{\log p/n}) + O(h_l + \Delta_{\beta,2}).
\end{aligned}$$

For the third term, by the Hoeffding's inequality and Lemma 9, we have

$$\begin{aligned}
& \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{1,-1\}} a \overline{W}_a \widetilde{\phi}' \left(a \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \boldsymbol{X} \right] \right\|_\infty \\
\leqslant & \left\| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\sum_{a\in\{1,-1\}} a \overline{W}_a \widetilde{\phi}' \left(a \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \boldsymbol{X} \right] \right\|_\infty \\
& + \left\| \mathbb{E} \left[\sum_{a\in\{1,-1\}} a \overline{W}_a \widetilde{\phi}' \left(a \boldsymbol{X}^\top \boldsymbol{\beta}_\phi^* \right) \boldsymbol{X} \right] \right\|_\infty \\
= & O_p(\sqrt{\log p/n}) + O(h_l).
\end{aligned}$$

Assume that $n^{1/2}(s'\mu_n)(\Delta_{\beta,2} + \sqrt{\log p/n} + h_l) \to 0$, we have (7) $\to 0$ in probability.

For (8), we have

$$
\begin{aligned}
&\max_l |I_l^* - \widehat{I}_l| \\
\leqslant\ & \max_l \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{-1,1\}} a\widehat{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) X_l \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
& + \max_l \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{-1,1\}} a \left(\widehat{W}_a - \overline{W}_a \right) \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \right. \right. \\
& \left. \left. \times X_l (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
& + \max_l \left| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \right. \right. \\
& \left. \left. \times X_l (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
& + \max_l \left| \mathbb{E} \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) X_l (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] - I_l^* \right|.
\end{aligned}
$$

For the first term, we have

$$
\begin{aligned}
& \max_l \left| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{-1,1\}} a\widehat{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) X_l \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
\leqslant\ & \left\| \widehat{\mathbb{E}}_n^{(k)} \left[\sum_{a\in\{-1,1\}} a\widehat{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) \boldsymbol{X}\boldsymbol{X}^\top \right] \right\|_{\max} \max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\|_1 \\
=\ & O_p(\max_l \|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\|_1).
\end{aligned}
$$

Similarly, for the second term, we have

$$
\begin{aligned}
& \max_l \left| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) X_l (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] \right| \\
=\ & O_p(R\sqrt{\frac{n^{-\zeta}\log p}{nh_g}}) + O(Rn^{-\eta}/h_g).
\end{aligned}
$$

By Hoeffding's inequality, for the third term, we have

$$
\begin{aligned}
& \max_l \left| (\widehat{\mathbb{E}}_n^{(k)} - \mathbb{E}) \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g} \left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)} \right) X_l (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}, l_\phi^*) \right] \right| \\
=\ & O_p\left(R\sqrt{\frac{\log p}{nh_g}} \right).
\end{aligned}
$$

By Lemma 11, we have

$$\begin{aligned}
&\max_l \left| \mathbb{E}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a \sum_{j=1}^{J} \Delta_j G_{h_g}\left(t_j - a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right) X_l(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*) \right] - I_l^* \right| \\
= \;& O_p(R(h_g^2 + \Delta_{\beta,2})).
\end{aligned}$$

Next, we will bound $S^{(k)}_{\tilde{\phi}',l}$. By exam the proof of Theorem 16 carefully, the convergence rate of $S^{(k)}_{\tilde{\phi}',l}$ is different from $S^{(k)}_{\tilde{\phi}',null(l)}$ due to the I_{122} replacing $\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,null(l)}$ by $\widehat{\boldsymbol{\beta}}^{(-k)}_{\phi,l}$. More specifically,

$$\begin{aligned}
& I_{122} \\
= \;& \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}'\left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right) - \sum_{a\in\{-1,1\}} a\overline{W}_a \widetilde{\phi}'\left(a\boldsymbol{X}^\top \boldsymbol{\beta}_\phi^*\right) \right\} (X_1 - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_\phi^*) \right] \\
& + O_p(Rn^{-\eta}) \\
= \;& \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} \overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} \boldsymbol{X}_{-l}^\top (\widehat{\boldsymbol{\beta}}_{\phi,-l} - \boldsymbol{\beta}^*_{\phi,-l})(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l}) \right] \\
& + \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} \overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} X_l(\widehat{\beta}_{\phi,l} - \beta^*_{\phi,l})(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l}) \right] + O_p(Rn^{-\eta}) \\
= \;& \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} \overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} X_l(\widehat{\beta}_{\phi,l} - \beta^*_{\phi,l})(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l}) \right] \\
& + O_p(R(h_l + \Delta_{\beta,2})\Delta_{\beta,2} + Rn^{-\eta}),
\end{aligned}$$

uniformly holds for all l's.

Because

$$\begin{aligned}
& \left| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} \overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} X_l(\widehat{\boldsymbol{\beta}}_{\phi,l} - \boldsymbol{\beta}^*_{\phi,l})(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l}) \right] \right| \\
\leqslant \;& \left| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} \overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} \left[\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}_\phi - \boldsymbol{\beta}^*_\phi)\right]^2 \right] \right|^{1/2} \\
& \left| \mathbb{E}\left[\left\{ \sum_{a\in\{-1,1\}} \overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top \widetilde{\boldsymbol{\beta}}\right) \right\} (X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}^*_{\phi,l})^2 \right] \right|^{1/2} \\
= \;& O(\Delta_{\beta,2}),
\end{aligned}$$

we have $\max_l |S^{(k)}_{\tilde{\phi}',l}| = O_p(\Delta_{\beta,2})$ assuming that $R(h_l + \Delta_{\beta,2})\Delta_{\beta,2} + Rn^{-\eta} = o_p(n^{-1/2})$. Thus, we have that (8) $= O_p(\sqrt{n}(\max_l \|\widehat{\boldsymbol{w}}^*_{\phi,l} - \boldsymbol{w}^*_{\phi,l}\|_1 + Rn^{-\eta}/h_g + R\sqrt{\frac{\log p}{nh_g}} + R\Delta_{\beta,2} +$

$Rh_g^2)\Delta_{\beta,2})$. Assuming the same condition in Theorem 16, we have that (8) $= o_p(1)$. As a summary, assuming that the same condition in Theorem 16, we have that both (7) and (8) are negligible.

When $\widehat{w}_\phi$ and $\widehat{\mathbb{E}}_n^{(k)}[\cdot]$ are independent. For (7), we have

$$\begin{aligned}(7) \quad \leqslant \quad & \max_l n^{1/2}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a(\widehat{W}_a - \overline{W}_a)\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right| \\ & + \max_l n^{1/2}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\left\{\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\right.\right.\right. \\ & \qquad \left.\left.\left. -\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)\right\}\cdot\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right| \\ & + \max_l n^{1/2}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right|.\end{aligned}$$

For the first term, by Bernstein's inequality, we have

$$\begin{aligned}& \max_l n^{1/2}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\left(\widehat{W}_a - \overline{W}_a\right)\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right| \\ = \quad & o_p\left(\sqrt{\log p/n}(1\vee\sqrt{s'\log p/n})\max_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2\right) + O(n^{-\eta}\max_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2).\end{aligned}$$

For the second term, similarly, we have

$$\begin{aligned}& \max_l n^{1/2}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\left\{\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right) - \widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)\right\}\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right| \\ = \quad & O_p\left(\sqrt{\log p/n}(1\vee\sqrt{s'\log p/n})\max_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2\right) + O(\Delta_{\beta,2}\max_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2).\end{aligned}$$

For the third term, we have

$$\begin{aligned}& \max_l n^{1/2}\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right| \\ = \quad & O_p\left(\sqrt{\log p/n}(1\vee\sqrt{s'\log p/n})\max_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2\right) + O(h_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2).\end{aligned}$$

For (8), we only need to derive an alternative bound for the term involving $\widehat{\boldsymbol{w}}_{\phi,l}$. Thus, by Bernstein's inequality and Lemma 11, we have

$$\begin{aligned}& \max_l\left|\widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{-1,1\}} a\widehat{W}_a\sum_{j=1}^J\Delta_j G_{h_g}\left(t_j - a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_\phi^{(-k)}\right)X_l\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*)\right]\right| \\ = \quad & O_p(\max_l\|\widehat{\boldsymbol{w}}_{\phi,l} - \boldsymbol{w}_{\phi,l}^*\|_2).\end{aligned}$$

By the discussion in the correlated case, we have $\max_l |S^{(k)}_{\tilde{\phi}',l}| = O_p(\Delta_{\beta,2})$ assuming that $R(h_l+\Delta_{\beta,2})\Delta_{\beta,2}+Rn^{-\eta} = o_p(n^{-1/2})$. Thus, we have that (8) $= O_p(\sqrt{n}(\max_l \|\widehat{\boldsymbol{w}}^*_{\phi,l}-\boldsymbol{w}^*_{\phi,l}\|_2+ Rn^{-\eta}/h_g + R\sqrt{\frac{\log p}{nh_g}} + R\Delta_{\beta,2} + Rh_g^2)\Delta_{\beta,2})$. Assuming the same condition in Theorem 16, we have that (8) $= o_p(1)$. As a summary, assuming that the same condition in Theorem 16, we have that both (7) and (8) are negligible.

For (6), we have

$$\begin{aligned}
(6) &= n^{1/2}\Bigg|(\widehat{\mathbb{E}}^{(k)}_n - \mathbb{E})\Bigg[\sum_{a\in\{1,-1\}} a\Big\{\overline{W}_a\widetilde{\phi}'\big(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\big) \\
&\qquad -\widehat{W}_a\widetilde{\phi}'\Big(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi\Big)\Big\}(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\Bigg]\Bigg| \\
&\quad +n^{1/2}\Bigg|I^*_l\left(\widehat{\beta}^{(k)}_{\phi,l} - \beta^*_{\phi,l}\right) \\
&\qquad +\mathbb{E}\Bigg[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\big(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\big)\ (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\Bigg] \\
&\qquad -\mathbb{E}\Bigg[\sum_{a\in\{1,-1\}} a\widehat{W}_a\widetilde{\phi}'\Big(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi\Big)\ (X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\Bigg]\Bigg| \\
&= I_{41} + I_{42}.
\end{aligned}$$

For I_{41}, we have that

$$\begin{aligned}
I_{41} &\leqslant n^{1/2}\left|(\widehat{\mathbb{E}}^{(k)}_n - \mathbb{E})\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\left\{\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right) - \widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)\right\}(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\right]\right| \\
&\quad +n^{1/2}\left|(\widehat{\mathbb{E}}^{(k)}_n - \mathbb{E})\left[\sum_{a\in\{1,-1\}} a\left\{\overline{W}_a - \widehat{W}_a\right\}\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\right]\right|
\end{aligned}$$

Following the same proof for I_{121}, we can show that the first term is $o_p(1)$ uniformly over all l's; the second term is also $o_p(1)$ due to that $\sup|\overline{W}_a - \widehat{W}_a| = O_p(n^{-\zeta})$ and $R\sqrt{\log p}n^{-\zeta} = o(1)$.

For I_{42}, we have

$$\begin{aligned}
I_{42} &\leqslant n^{1/2}\Bigg|I^*\left(\widehat{\beta}^{(k)}_{\phi,1} - \beta^*_{\phi,1}\right) + \mathbb{E}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right)(X_1 - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\right] \\
&\qquad -\mathbb{E}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)(X_1 - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\right]\Bigg| \\
&\qquad +n^{1/2}\left|\mathbb{E}\left[\sum_{a\in\{1,-1\}} a\left(\widehat{W}_a - \overline{W}_a\right)\widetilde{\phi}'\left(a\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}^{(-k)}_\phi\right)(X_1 - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\right]\right| \\
&= I_{421} + I_{422}.
\end{aligned}$$

The last term is $I_{422} = O_p(Rn^{-\eta+1/2})$. For I_{421}, by Taylor's expansion, we have

$$\begin{aligned} I_{421} \leqslant\ & n^{1/2}\left| I_l^* \left(\widehat{\beta}_{\phi,l}^{(k)} - \beta_{\phi,l}^*\right) + \mathbb{E}\left[\left\{\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top\widetilde{\boldsymbol{\beta}}\right)\right\} X_l(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right]\right. \\ & \left.(\widehat{\boldsymbol{\beta}}_{\phi,l}^{(k)} - \boldsymbol{\beta}_{\phi,l}^*)\right| \\ & + n^{1/2}\left|\mathbb{E}\left[\left\{\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}''\left(a\boldsymbol{X}^\top\widetilde{\boldsymbol{\beta}}\right)\right\}(X_l - \boldsymbol{X}_{-l}^\top \boldsymbol{w}_{\phi,l}^*)\right]\right. \\ & \left.\boldsymbol{X}_{-l}^\top(\widehat{\boldsymbol{\beta}}_{\phi,-l}^{(k)} - \boldsymbol{\beta}_{\phi,-l}^*)\right| \end{aligned}$$

By Lemma 1.2 and the definition of $\boldsymbol{w}_{\phi,l}^*$, we have that the both terms are $O_p(\sqrt{n}R(h_l + \Delta_{\beta,2}^2))$. Thus, under the condition that $\sqrt{n}R(h_l + \Delta_{\beta,2}^2) \to 0$ (assumed in Theorem 16), we have that (6) $\to 0$ in probability.

In conclusion, assuming that the same condition in Theorem 16, we have that (6), (7), and (8) are negligible. Thus, under these conditions, we have that

$$\max_l n^{1/2}\left| I_l^*\left(\widetilde{\beta}_{\phi,l}^{(k)} - \beta_{\phi,l}^*\right) + \widehat{\mathbb{E}}_n^{(k)}\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right]\right| = o_p(1).$$

This implies the asymptotic normality of $\widetilde{\beta}_{\phi,l}^{(k)}$. In addition, it also implies that $\widetilde{\beta}_{\phi,l}^{(k)}$'s are asymptotically independent for different k's. Thus, we can conclude the asymptotic normality of $\widetilde{\beta}_{\phi,l}$. Further, due to $\min_l I_l^* \geqslant \lambda_{\min}$, we have that

$$\max_l n^{1/2}\left|\widehat{I}_l\left(\widetilde{\beta}_{\phi,l} - \beta_{\phi,l}^*\right) + \widehat{\mathbb{E}}_n\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right]\right| = o_p(1).$$

Combining with the uniform bound

$$\log p \max_l |(\widehat{\sigma}_l)^2 - \sigma_l^2| = o_p(1),$$

we have

$$\begin{aligned} \max_l n^{1/2}\Bigg|&\widehat{\sigma}_l^{-1}\widehat{I}_l(\widetilde{\beta}_{\phi,l} - \beta_{\phi,l}^*) \\ &+\sigma_l^{-1}\widehat{\mathbb{E}}_n\left[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*\right)(X_l - \boldsymbol{X}_{-l}^\top\boldsymbol{w}_{\phi,l}^*)\right]\Bigg| = o_p(1). \end{aligned}$$

By the Berry-Esseen bound for CLT, we have

$$\begin{aligned} & \max_l \sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(\left|n^{1/2}\widehat{\sigma}_l^{-1}\widehat{I}_l\left(\widetilde{\beta}_{\phi,l} - \beta_{\phi,l}^*\right)\right| \leqslant \Phi^{-1}(1-\alpha/2)\right) - (1-\alpha)\right| \\ \leqslant\ & \frac{c_0}{\sqrt{n}}R^3, \end{aligned}$$

for a sufficient large c_0. □

Appendix B.

In this section, we extend our proposed method to testing a hypothesis with a growing dimension. Via a bootstrap procedure, a valid inference procedure is developed for a group hypothesis such as $\mathcal{H}_0 : \beta^*_{\phi,l} = 0, \forall l \in \mathcal{G}$, where $\mathcal{G}$ is an index set with its cardinality growing as $n \to +\infty$. Specifically, to test this hypothesis based on Algorithm 2, we consider the following multiplier bootstrap procedure. Define

$$\delta^*_l \equiv \frac{1}{K}\sum_k \widehat{\mathbb{E}}^{(k)}_n \left[r \sum_{a\in\{1,-1\}} a\widehat{W}_a \widetilde{\phi}' \left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{(k)}_\phi \right) (X_l - \boldsymbol{X}^\top_{-l}\widehat{\boldsymbol{w}}_{\phi,l}) \right],$$

where r follows an independent standard normal distribution. Let $c^*_{1-\alpha}$ be the upper α-quantile of the distribution of $\max_{l\in\mathcal{G}} |\delta^*_l|$ conditional on the training samples, which can be calculated by bootstrapping the weight r. We will reject the $\mathcal{H}_0$ if $\max_{l\in\mathcal{G}} |\widetilde{\beta}_{\phi,l}| > c^*_{1-\alpha}$. The following theorem provides the validity of the testing procedure.

Theorem 18 *Assume the same Conditions in Theorem 17, under the null hypothesis, we have*

$$\sup_{\alpha\in(0,1)} \left| \mathbb{P}\left(n^{1/2} \max_{l\in\mathcal{G}} \left|\widetilde{\beta}_{\phi,l}\right| \leqslant c^*_{1-\alpha} \right) - (1-\alpha) \right| = o(1),$$

if

$$\begin{aligned} n^{-1}(\log(n|\mathcal{G}|))^7 &= o(1) \\ \Delta_{n,p}\sqrt{\log|\mathcal{G}|} &= o(1) \\ \Delta^{1/2}\log|\mathcal{G}| &= o(1), \end{aligned}$$

where

$$\begin{aligned} \Delta_{n,p} &= n^{1/2}(s'\mu_n)(\Delta_{\beta,2} + \sqrt{\log p/n} + h_l + n^{-\eta}) \\ &\quad + \sqrt{n}R(h_l + n^{-\eta} + \sqrt{\log p}\Delta^{2\gamma/(\gamma+2)}_{\beta,2}) \\ &\quad + (Rs'\mu_n + R^2 n^{-\zeta} + R^2 h_l)\sqrt{\log p}, \\ \Delta &= Rs'\mu_n + R^2 n^{-\zeta} + R^2 h_l. \end{aligned}$$

From Theorem 18, when we have additional nuisance parameters, we can see that the requirement on $\log|\mathcal{G}|$ is related to the convergence rate of the nuisance parameters $n^{-\zeta}$.

Proof of Theorem 18. Under the null hypothesis, from Theorem 17, we know that $\forall\epsilon$, there is a C_ϵ such that

$$\begin{aligned} &\mathbb{P}\Bigg(\max_l n^{1/2} \Bigg| \left(\widetilde{\beta}_{\phi,l} - \beta^*_{\phi,l}\right) \\ &\qquad + I_l^{-1}\widehat{\mathbb{E}}_n \left[\sum_{a\in\{1,-1\}} a\overline{W}_a \widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right)(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l}) \right] \Bigg| \geqslant C_\epsilon \Delta_{n,p} \Bigg) \\ &\leqslant \epsilon/3. \end{aligned}$$

Define $\boldsymbol{\xi} = (\xi_l)_{l\in\mathcal{G}}$ as a $|\mathcal{G}|$-dimensional multi-variant Gaussian distribution with

$$\mathbb{E}[\xi_l] = 0$$

$$\mathbb{E}[\xi_{l_1}\xi_{l_2}] = \mathbb{E}\left[\left\{\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\left(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\right)\right\}^2 (X_{l_1} - \boldsymbol{X}^\top_{-l_1}\boldsymbol{w}^*_{\phi,l_1})(X_{l_2} - \boldsymbol{X}^\top_{-l_2}\boldsymbol{w}^*_{\phi,l_2})\right].$$

By Theorem 14 in Wasserman (2014), under the null hypothesis, we have

$$\begin{aligned}
&\mathbb{P}\left(n^{1/2}\max_{l\in\mathcal{G}}|\widetilde{\beta}_{\phi,l}| \leqslant c^*_{1-\alpha}\right)\\
\leqslant\ &\mathbb{P}\Bigg(n^{1/2}\max_{l\in\mathcal{G}} I_l^{-1}\Bigg|\widehat{\mathbb{E}}_n\Bigg[\sum_{a\in\{1,-1\}} a\overline{W}_a\widetilde{\phi}'\big(a\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi\big)\\
&\qquad\cdot(X_l - \boldsymbol{X}^\top_{-l}\boldsymbol{w}^*_{\phi,l})\Bigg]\Bigg| \leqslant c^*_{1-\alpha} + C_\epsilon\Delta_{n,p}\Bigg) + \epsilon/3\\
\leqslant\ &\mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha} + C_\epsilon\Delta_{n,p}\right) + M\left(\frac{(\log(n|\mathcal{G}|))^7}{n}\right)^{1/8} + \epsilon/4.
\end{aligned}$$

uniformly holds for all α.

By $n^{-1}(\log(n|\mathcal{G}|))^7 \to 0$, for large enough n, we have

$$\begin{aligned}
&\mathbb{P}\left(n^{1/2}\max_{l\in\mathcal{G}}|\widetilde{\beta}_{\phi,l}| \leqslant c^*_{1-\alpha}\right)\\
\leqslant\ &\mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha} + C_\epsilon\Delta_{n,p}\right) + \epsilon/2.
\end{aligned}$$

By Corollary 16 in Wasserman (2014), we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha} + C_\epsilon\Delta_{n,p}\right)\\
\leqslant\ &\mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha}\right) + C_\epsilon\Delta_{n,p}\sqrt{1\vee\log|\mathcal{G}|/C_\epsilon\Delta_{n,p}}.
\end{aligned}$$

Due to $\Delta_{n,p}\sqrt{\log|\mathcal{G}|} = o(1)$, for large enough n, we have

$$\mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha} + C_\epsilon\Delta_{n,p}\right) \leqslant \mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha}\right) + \epsilon/4.$$

Thus, we have

$$\mathbb{P}\left(n^{1/2}\max_{l\in\mathcal{G}}|\widetilde{\beta}_{\phi,l}| \leqslant c^*_{1-\alpha}\right) \leqslant \mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha}\right) + 3\epsilon/4.$$

To conclude, we just need to show that

$$\mathbb{P}\left(\max_{l\in\mathcal{G}}|\xi_l| \leqslant c^*_{1-\alpha}\right) \leqslant \mathbb{P}\left(\max_{l\in\mathcal{G}}|\delta^*_l| \leqslant c^*_{1-\alpha}\right) + \epsilon/4.$$

Let Ω represent the entire training data. Notice that

$$\begin{aligned}
&\mathbb{E}[\delta_l^* \mid \Omega] = 0 \\
&\mathbb{E}[\delta_{l_1}^* \delta_{l_2}^* \mid \Omega] = \frac{1}{K}\sum_k \widehat{\mathbb{E}}_n^{(k)} \left[\left\{ \sum_{a\in\{1,-1\}} a\widehat{W}_a \widetilde{\phi}'\left(a\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_\phi^{(k)}\right) \right\}^2 \right. \\
&\left. (X_{l_1} - \boldsymbol{X}_{-l_1}^\top \widehat{\boldsymbol{w}}_{\phi,l_1})(X_{l_2} - \boldsymbol{X}_{-l_2}^\top \widehat{\boldsymbol{w}}_{\phi,l_2}) \right].
\end{aligned}$$

By the proof of Theorem 17, we know that

$$\left\|\mathbb{E}[\delta_{l_1}^* \delta_{l_2}^* \mid \Omega] - \mathbb{E}[\xi_{l_1}\xi_{l_2}]\right\|_{\max} = O_p(Rs'\mu_n + R^2 n^{-\zeta} + R^2 h_l).$$

By Theorem 17 in Wasserman (2014), we have

$$\mathbb{P}\left(\max_{l\in\mathcal{G}} |\xi_l| \leqslant c_{1-\alpha}^*\right) \leqslant \mathbb{P}\left(\max_{l\in\mathcal{G}} |\delta_l^*| \leqslant c_{1-\alpha}^*\right) + C\Delta^{1/3}(1 \vee \log|\mathcal{G}|/\Delta)^{2/3},$$

where $\Delta = Rs'\mu_n + R^2 n^{-\zeta} + R^2 h_l$. Assuming $\Delta^{1/2}\log|\mathcal{G}| = o(1)$, we have

$$\mathbb{P}\left(\max_{l\in\mathcal{G}} |\xi_l| \leqslant c_{1-\alpha}^*\right) \leqslant \mathbb{P}\left(\max_{l\in\mathcal{G}} |\delta_l^*| \leqslant c_{1-\alpha}^*\right) + \epsilon/4.$$

Similarly, we can show that

$$\mathbb{P}\left(n^{1/2}\max_{l\in\mathcal{G}} |\widetilde{\beta}_{\phi,l}| \leqslant c_{1-\alpha}^*\right) \geqslant 1 - \alpha - \epsilon.$$

This concludes the proof. For the algorithm with dedicated sample-splitting strategy, we require that

$$\begin{aligned}
n^{-1}(\log(n|\mathcal{G}|))^7 &= o(1) \\
\Delta_{n,p}\sqrt{\log|\mathcal{G}|} &= o(1) \\
\Delta^{1/2}\log|\mathcal{G}| &= o(1),
\end{aligned}$$

with

$$\begin{aligned}
\Delta_{n,p} &= \sqrt{n}(\sqrt{s'}\mu_n)((1 \vee \sqrt{s'\log p/n})\sqrt{\log p/n} + h_l + \Delta_{\beta,2} \\
&\quad + \sqrt{h_l^{\gamma/2}\log p/n}(1 \vee \sqrt{s'\log p/(nh_l^{\gamma/2})})) \\
&\quad + \sqrt{n}R(h_l + n^{-\eta} + \sqrt{\log p}\Delta_{\beta,2}^{2\gamma/(\gamma+2)}) \\
&\quad + (R\sqrt{s'}\mu_n + R^2 n^{-\zeta} + R^2 h_l)\sqrt{\log p}, \\
\Delta &= R\sqrt{s'}\mu_n + R^2 n^{-\zeta} + R^2 h_l.
\end{aligned}$$

□

Appendix C.

The following corollary shows the convergence rate of $\widehat{\boldsymbol{\beta}}_\phi$ when ϕ is chosen as a hinge loss. Under this corollary, Theorem 16 holds under the derived convergence rate. We assume

(A1) The density functions of $\boldsymbol{X}_{j_0}$ conditional on $A = 1$ and $A = -1$ are continious and have common support and finishte second moments.

(A2) $s^* = O(n^{c_1})$ for some $c_1 \in [0, 1/2)$.

(A3) There exists a constant M_1 such that

$$\max_{\boldsymbol{d}\in\mathbb{R}^p:\|\boldsymbol{d}\|_0\leqslant 2s^*} \frac{\widehat{\mathbb{E}}_n[(\boldsymbol{X}^\top \boldsymbol{d})^2]}{n\|\boldsymbol{d}\|_2^2} \leqslant M_1$$

almost surely.

(A4) $n^{(1-c_2)/2} \min_{\beta^*_{\phi,l}\neq 0} |\beta^*_{\phi,l}| \geqslant M_2$ for some constants $M_2 > 0$ and $2c_1 < c_2 < 1$.

(A5) The conditional density functions of $\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi$ given $A = 1$ and $A = -1$ are uniformly bounded away from 0 and $+\infty$ in a neighborhood of 1 and -1, respectively.

Conditions (A1)-(A5) are the conditions in Peng et al. (2016). In Peng et al. (2016), these regularity conditions are required for classification problems; In Lemma 4.5 where additional nuisance parameters are present, we also need these regularity conditions.

Lemma 19 *Under Conditions (A1) - (A5) and (C1) - (C3) with $\alpha + \beta > 1/2$, if ϕ is a hinge loss, then $\Delta_{\beta,1} = O(s^*\sqrt{\log p/n})$ and $\Delta_{\beta,2} = O(\sqrt{s^* \log p/n})$, where s^* is the number of non-zero entries in $\boldsymbol{\beta}^*_\phi$.*

Proof of Lemma 19. When W_a's are known, the proof of Thoerem 4 in Peng et al. (2016) can be directly applied to show this result. Thus, in this proof, we focus on the case where W_a's are estimated on an independent dataset. Let $\boldsymbol{h} = \boldsymbol{\beta}^*_\phi - \widehat{\boldsymbol{\beta}}_\phi$. Since $\widehat{\boldsymbol{\beta}}_\phi$ minimizes $\ell_\phi(\boldsymbol{\beta}; \widehat{W}_1, \widehat{W}_{-1}) + \lambda_n\|\boldsymbol{\beta}\|_1$, we have

$$\begin{aligned} & \widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}^*_\phi + \boldsymbol{h})) + \widehat{W}_{-1}\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}^*_\phi + \boldsymbol{h}))\right\}\right] + \lambda_n\|\widehat{\boldsymbol{\beta}}_\phi\|_1 \qquad (9) \\ \leqslant\ & \widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1\phi(\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) + \widehat{W}_{-1}\phi(-\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi)\right\}\right] + \lambda_n\|\boldsymbol{\beta}^*_\phi\|_1. \qquad (10) \end{aligned}$$

Let $S = \{1 \leqslant j \leqslant p : \beta^*_{\phi,j} \neq 0\}$, we have

$$\|\boldsymbol{\beta}^*_\phi\|_1 - \|\widehat{\boldsymbol{\beta}}_\phi\|_1 \leqslant \|\boldsymbol{\beta}^*_{\phi,S}\|_1 - \|\widehat{\boldsymbol{\beta}}_\phi\|_1 \leqslant \|\boldsymbol{h}_S\|_1 - \|\boldsymbol{h}_{S^c}\|_1.$$

By convexity, we have

$$\begin{aligned} & \widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}^*_\phi + \boldsymbol{h})) + \widehat{W}_{-1}\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}^*_\phi + \boldsymbol{h}))\right\}\right] \\ & -\widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1\phi(\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi) + \widehat{W}_{-1}\phi(-\boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi)\right\}\right] \\ \geqslant\ & -\left\|\widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1 1\left\{1 - \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi \geqslant 0\right\} - \widehat{W}_{-1}1\left\{1 + \boldsymbol{X}^\top\boldsymbol{\beta}^*_\phi \geqslant 0\right\}\right\}\right]\right\|_\infty \|\boldsymbol{h}\|_1 \end{aligned}$$

By Hoeffding's inequality, with $\lambda_n = C(\alpha)\sqrt{2\log p/n}$, similar to Lemma 1 in Peng et al. (2016), we have

$$\mathbb{P}\left(\left\|\widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1 1\left\{1-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* \geqslant 0\right\} - \widehat{W}_{-1}1\left\{1+\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^* \geqslant 0\right\}\right\}\right]\right\|_\infty \geqslant \lambda_n/2\right) \leqslant p^{-\alpha},$$

for a pre-specified α, where $C(\alpha)$ is some constant. Hence, we have

$$\begin{aligned}\lambda_n(\|\boldsymbol{h}_S\|_1 - \|\boldsymbol{h}_{S^c}\|_1) &\geqslant -\lambda_n/2(\|\boldsymbol{h}_S\|_1 + \|\boldsymbol{h}_{S^c}\|_1),\\ 3\|\boldsymbol{h}_S\|_1 &\geqslant \|\boldsymbol{h}_{S^c}\|.\end{aligned}$$

Define

$$\begin{aligned}B(\boldsymbol{h}) &= \widehat{\mathbb{E}}_n\left[\left\{\widehat{W}_1\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h})) + \widehat{W}_{-1}\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h}))\right\}\right.\\ &\quad \left.-\left\{\widehat{W}_1\phi(\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*) + \widehat{W}_{-1}\phi(-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right\}\right]\\ &\quad -\mathbb{E}\left[\left\{W_1\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h})) + W_{-1}\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h}))\right\}\right.\\ &\quad \left.-\left\{W_1\phi(\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*) + W_{-1}\phi(-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right\}\right].\end{aligned}$$

We will show that assuming $p > n$, then for all n sufficiently large, we have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{h}\|_0\leqslant s, \|\boldsymbol{h}\|_2\neq 0} |B(\boldsymbol{h})|/\|\boldsymbol{h}\|_2 \geqslant C_2\sqrt{\log p/n}\right) \leqslant p^{-c_2},$$

where C_2 and c_2 are some positive constants. We decompose

$$\begin{aligned}B(\boldsymbol{h}) &= \widehat{\mathbb{E}}_n\left[\left(\widehat{W}_1 - W_1\right)\left(\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h})) - \phi(\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right)\right.\\ &\quad \left.+\left(\widehat{W}_{-1} - W_{-1}\right)\left(\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h})) - \phi(-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right)\right]\\ &\quad +\widetilde{B}(\boldsymbol{h}).\end{aligned}$$

By Lemma 3 in Peng et al. (2016), there exists positive constants C_1 and c_1 such that

$$\mathbb{P}\left(\sup_{\|\boldsymbol{h}\|_0\leqslant s, \|\boldsymbol{h}\|_2\neq 0} |\widetilde{B}(\boldsymbol{h})|/\|\boldsymbol{h}\|_2 \geqslant C_1/3\sqrt{\log p/n}\right) \leqslant p^{-c_1}/3. \tag{11}$$

Denote

$$\begin{aligned}G(\boldsymbol{h}) &:= \left(\widehat{\mathbb{E}}_n - \mathbb{E}\right)\left[\left(\widehat{W}_1 - W_1\right)\left(\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h})) - \phi(\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right)\right.\\ &\quad \left.+\left(\widehat{W}_{-1} - W_{-1}\right)\left(\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^*+\boldsymbol{h})) - \phi(-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right)\right]\end{aligned}$$

Because $\sup_{\boldsymbol{X},A,Y}|\widehat{W}_a - W_a| = O_p(\max\{n^{-\alpha}, n^{-\beta}\})$, by Lemma 3 in Peng et al. (2016), we have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{h}\|_0\leqslant s, \|\boldsymbol{h}\|_2\neq 0} |G(\boldsymbol{h})|/\|\boldsymbol{h}\|_2 \geqslant C_1/3\sqrt{\log p/n}\right) \leqslant p^{-c_1}/3. \tag{12}$$

Assuming that $\lambda_{\max}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top])$ is bounded, we have that

$$\mathbb{E}\left[\left(\widehat{W}_1 - W_1\right)\left(\phi(\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^* + \boldsymbol{h})) - \phi(\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right)\right. \tag{13}$$

$$\left. + \left(\widehat{W}_{-1} - W_{-1}\right)\left(\phi(-\boldsymbol{X}^\top(\boldsymbol{\beta}_\phi^* + \boldsymbol{h})) - \phi(-\boldsymbol{X}^\top\boldsymbol{\beta}_\phi^*)\right)\right] / \|\boldsymbol{h}\|_2 = O_p(n^{-\alpha-\beta}), \tag{14}$$

uniformly holds for all $\boldsymbol{h}$. Combining (11), (12), and (13), we have that

$$\mathbb{P}\left(\sup_{\|\boldsymbol{h}\|_0 \leqslant s, \|\boldsymbol{h}\|_2 \neq 0} |B(\boldsymbol{h})| / \|\boldsymbol{h}\|_2 \geqslant C_2\sqrt{\log p/n}\right) \leqslant p^{-c_2}, \tag{15}$$

with $C_2 = C_1$ and $c_2 = c_1$. Based on (9) and (15), following the proof of Theorem 4 in Peng et al. (2016), we have the desired result. □

Appendix D.

In this section, we generate the data following the simulation studies in Liang et al. (2022), and compare different methods using simulations. We refer readers to Scenario (II) in Liang et al. (2022) for details on data generation. In addition, to show whether our method is sensitive to the specification of bandwidth, we add a scale factor to the current specification and check the performance of the methods under different scale factors. Figures 7 - 9 summarize the results in this simulation scenario. We can see that while logistic loss leads to higher power in detecting non-zero coefficients, the value functions are lower than those using the hinge loss. Additionally, the power, Type I error, and power remain similar under different bandwidth specifications. Thus, our method is robust to bandwidth selection.

References

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8):1823–1840, 2008.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

Xin Bing and Marten Wegkamp. Optimal discriminant analysis in high-dimensional latent factor models. *Annals of Statistics*, 51(3):1232–1257, 2023.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36(2):489–531, 2008.

Tianxi Cai, Mengyan Li, and Molei Liu. Semi-supervised triply robust inductive transfer learning. *Journal of the American Statistical Association*, 120(550):1037–1047, 2025.

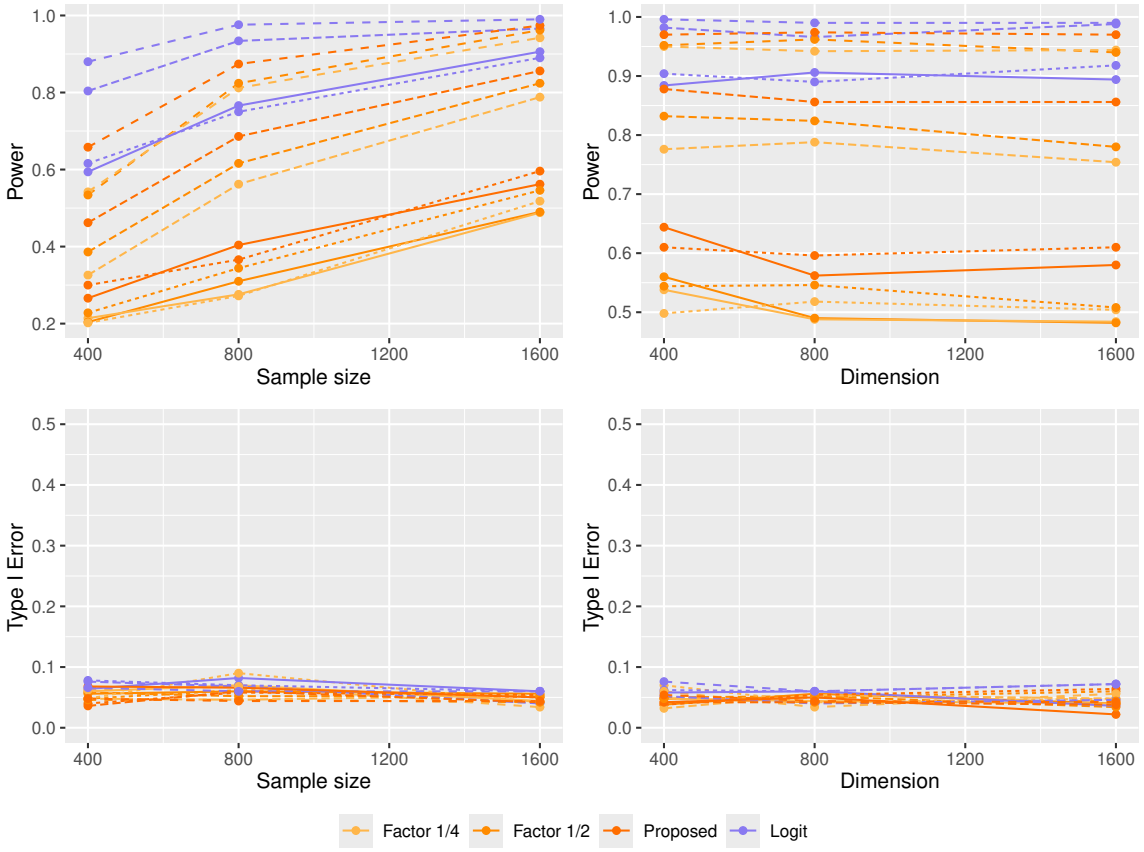

Figure 7: Testing results with the change of sample size when $\xi = 0.8$ and $p = 800$ and the change of p ($n = 800$; $\xi = 0.8$). Line styles represent different coefficients.

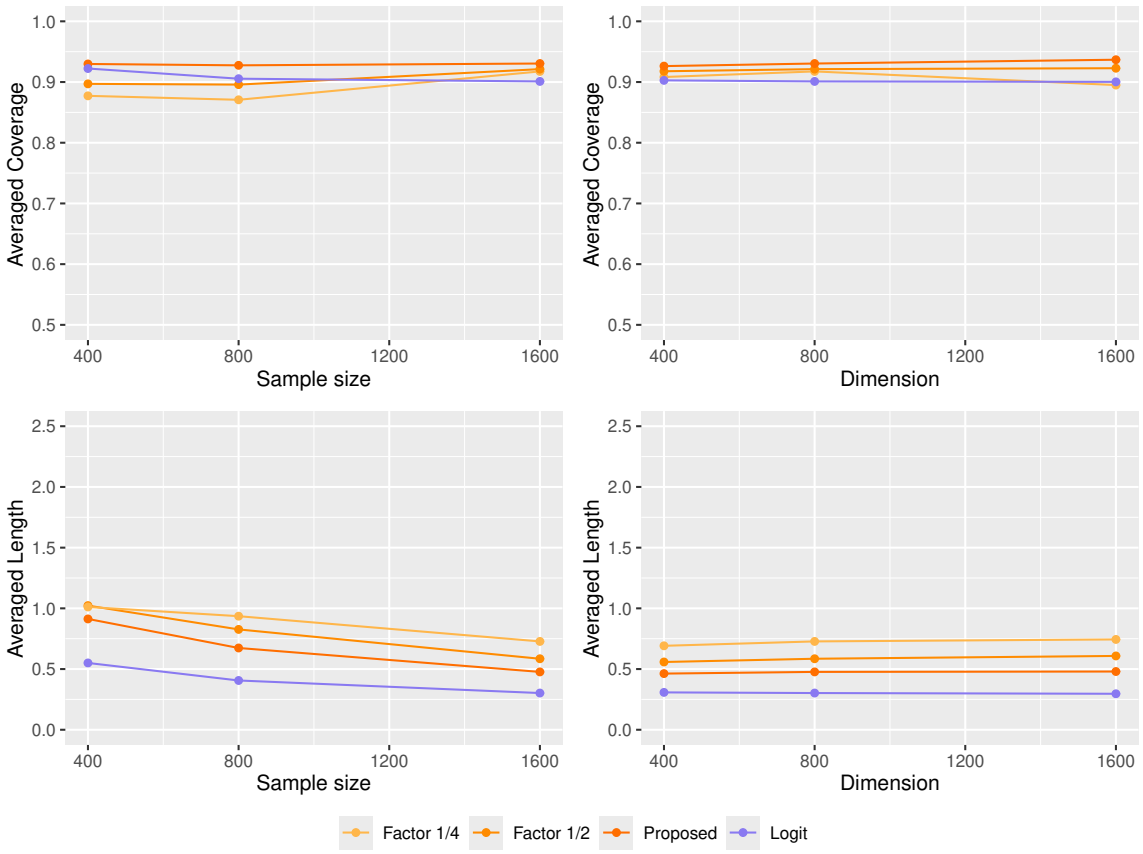

Figure 8: Coverage results with the change of sample size when $\xi = 0.8$ and $p = 800$ and the change of p ($n = 800$; $\xi = 0.8$).

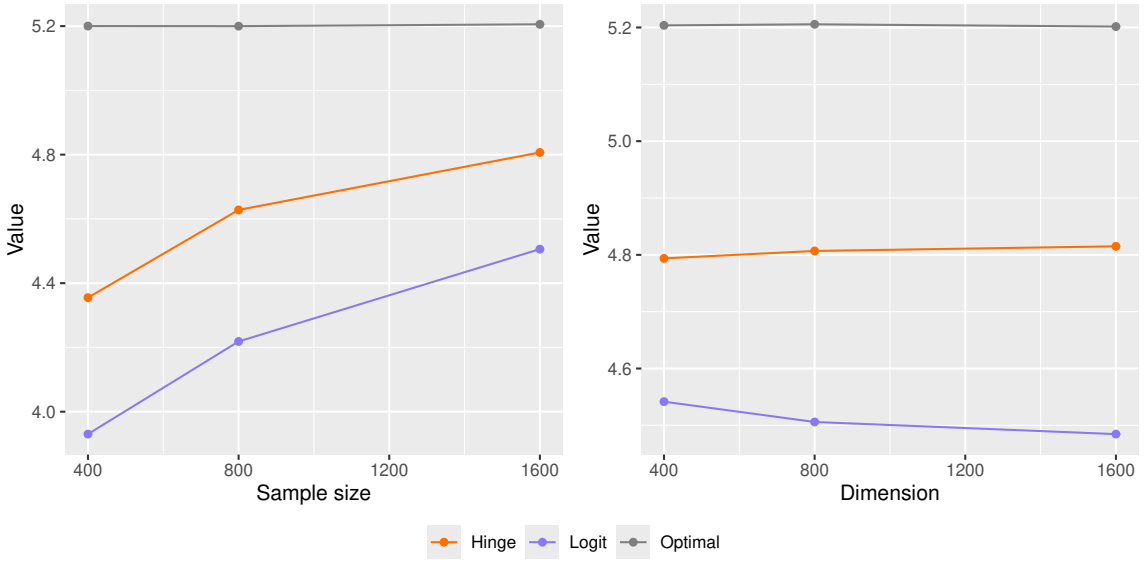

Figure 9: Value comparison with the change of sample size when $\xi = 0.8$ and $p = 800$ and the change of p ($n = 800$; $\xi = 0.8$).

Noirrit Kiran Chandra, Abhra Sarkar, John F de Groot, Ying Yuan, and Peter Müller. Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials. *Journal of the American Statistical Association*, 118(544):2301–2314, 2023.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516): 1509–1521, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal and safe estimation for high-dimensional semi-supervised learning. *Journal of the American Statistical Association*, 119(548):2748–2759, 2024.

Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, 2017.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

Franca Hoffmann, Bamdad Hosseini, Zhi Ren, and Andrew M Stuart. Consistency of semi-supervised learning algorithms on graphs: Probit and one-hot methods. *Journal of Machine Learning Research*, 21(186):1–55, 2020.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

X. Jessie Jeng, Wenbin Lu, and Huimin Peng. High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12(1):2074–2089, 2018.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9(44): 1343–1368, 2008.

Muxuan Liang, Young-Geun Choi, Yang Ning, Maureen A Smith, and Ying-Qi Zhao. Estimation and inference on high-dimensional individualized treatment rule in observational data using split-and-pooled de-correlated score. *Journal of Machine Learning Research*, 23(262):1–65, 2022.

Ruitao Lin, Peter F Thall, and Ying Yuan. Bags: A bayesian adaptive group sequential trial design with subgroup-specific survival comparisons. *Journal of the American Statistical Association*, 116(533):322–334, 2021.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004.

Rong Ma, T Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, 116(534):984–998, 2021.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics*, 45(1):158–195, 2017.

Yinghao Pan and Ying-Qi Zhao. Improved doubly robust estimation in learning optimal individualized treatment rules. *Journal of the American Statistical Association*, 116(533): 283–294, 2021.

Bo Peng, Lan Wang, and Yichao Wu. An error bound for l1-norm support vector machine coefficients in ultra-high dimension. *Journal of Machine Learning Research*, 17(1):8279–8304, 2016.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Chengchun Shi, Rui Song, Zhao Chen, Runze Li, et al. Linear hypothesis testing for high dimensional generalized linear models. *Annals of Statistics*, 47(5):2671–2703, 2019.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 119(546):1065–1075, 2024.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

Ingo Steinwart, Clint Scovel, et al. Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

Régis Vert and Jean-Philippe Vert. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7(29):817–854, 2006.

Junhui Wang and Xiaotong Shen. Large margin semi-supervised learning. *Journal of Machine Learning Research*, 8(65):1867–1891, 2007.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20(113):1–41, 2019.

Larry Wasserman. Stein's method and the bootstrap in low and high dimensions: A tutorial. Technical report, 2014.

Yunan Wu, Lan Wang, and Haoda Fu. Model-assisted uniformly honest inference for optimal treatment regimes in high dimension. *Journal of the American Statistical Association*, 118(541):305–314, 2023.

Fei Xue, Yanqing Zhang, Wenzhuo Zhou, Haoda Fu, and Annie Qu. Multicategory angle-based learning for estimating optimal dynamic treatment regimes with censored data. *Journal of the American Statistical Association*, 117(539):1438–1451, 2022.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. A consistent information criterion for support vector machines in diverging model spaces. *Journal of Machine Learning Research*, 17(1):466–491, 2016a.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):53–76, 2016b.

Ying-Qi Zhao, Donglin Zeng, Eric B Laber, Rui Song, Ming Yuan, and Michael Rene Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 2014.

Ying-Qi Zhao, Eric B Laber, Yang Ning, Sumona Saha, and Bruce E Sands. Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, 20(1):1821–1843, 2019.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R. Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.