Journal of Machine Learning Research 24 (2023) 1-58 Submitted 9/21; Revised 11/22; Published 8/23

Surrogate Assisted Semi-supervised Inference for High Dimensional Risk Prediction

Jue Hou HOU00123@UMN.EDU
Division of Biostatistics
University of Minnesota School of Public Health
Minneapolis, MN 55455, USA

Zijian Guo ZIJGUO@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854-8019, USA

Tianxi Cai TCAI@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Editor: Mladen Kolar

Abstract

Risk modeling with electronic health records (EHR) data is challenging due to no direct observations of the disease outcome and the high-dimensional predictors. In this paper, we develop a surrogate assisted semi-supervised learning approach, leveraging small labeled data with annotated outcomes and extensive unlabeled data of outcome surrogates and high-dimensional predictors. We propose to impute the unobserved outcomes by constructing a sparse imputation model with outcome surrogates and high-dimensional predictors. We further conduct a one-step bias correction to enable interval estimation for the risk prediction. Our inference procedure is valid even if both the imputation and risk prediction models are misspecified. Our novel way of ultilizing unlabelled data enables the high-dimensional statistical inference for the challenging setting with a dense risk prediction model. We present an extensive simulation study to demonstrate the superiority of our approach compared to existing supervised methods. We apply the method to genetic risk prediction of type-2 diabetes mellitus using an EHR biobank cohort.

Keywords: generalized linear models, high dimensional inference, model mis-specification, risk prediction, semi-supervised learning.

1. Introduction

Precise risk prediction is vitally important for successful clinical care. High risk patients can be assigned to more intensive monitoring or intervention to improve outcome. Traditionally, risk prediction models are developed based on cohort studies or registry data. Population-based disease registries, while remain a critical source for epidemiological studies, collect information on a relatively small set of pre-specified variables and hence may limit researchers' ability to develop comprehensive risk prediction models (Warren and Yabroff, 2015). Most clinical care is delivered in healthcare systems (Thompson et al., 2015), and

©2023 Jue Hou, Zijian Guo and Tianxi Cai.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/21-1075.html.

electronic health records (EHR) embedded in healthcare systems accrue rich clinical data in broad patient populations. EHR systems centralize the data collected during routine patient care including structured elements such as codes for International Classification of Diseases, medication prescriptions, and medical procedures, as well as free-text narrative documents such as physician notes and pathology reports that can be processed through natural language processing for analysis. EHR data is also often linked with biobanks which provide additional rich molecular information to assist in developing comprehensive risk prediction models for a broad patient population.

Risk modeling with EHR data, however, is challenging due to several reasons. First, precise information on clinical outcome of interest, Y, is often embedded in free-text notes and requires manual efforts to extract accurately. Readily available outcome surrogates $\mathbf{S}$, such as the diagnostic codes or mentions of the outcome, may be predictive of the true outcome Y, can deviate from the true label Y . Here we consider the general situation that a vector of surrogates, $\mathbf{S}$, that are noisy error prone proxies of Y and may include non-informative surrogates. For example, using EHR data from Mass General Brigham, we found that the positive predictive value was only 0.48 and 0.19 for having at least 1 diagnosis code of Type II Diabetes Mellitus (T2DM) and for having at least 1 mention of T2DM in medical notes, respectively. Directly using these EHR proxies as true disease status to derive risk models may lead to substantial biases. On the other hand, extracting precise disease status requires manual chart review which is not feasible at a large scale. It is thus of great interest to develop risk prediction models under a semi-supervised learning (SSL) framework using both a large unlabeled dataset of size N containing information on predictors $\mathbf{X}$ along with surrogates $\mathbf{S}$ and a small labeled dataset of size n with additional observations on Y curated via chart review. Throughout the paper, we impose no stringent model assumptions on the triplet $(Y, \mathbf{X}, \mathbf{S})$ while using generalized linear working models to define and estimate the risk prediction model (see Section 2).

Additional challenges arise from the high dimensionality of the predictor vector $\mathbf{X}$, and the potential model mis-specifications. Although much progress has been made in high dimensional regression in recent years, there is a paucity of literature on high dimensional inference under the SSL setting. Precise estimation of the high dimensional risk model is even more challenging if the risk model is not sparse. Allowing the risk model to be dense is particularly important when $\mathbf{X}$ includes genomic markers since a large number of genetic markers appear to contribute to the risk of complex traits (Frazer et al., 2009). For example, Vujkovic et al. (2020) recently identified 558 genetic variants as significantly associated with T2DM risk. An additional challenge arises when the fitted risk model is mis-specified, which occurs frequently in practice especially in the high dimensional setting. Model mis-specifications can also lead to the fitted model of $Y \mid \mathbf{X}$ to be dense. There are limited methods currently available to make inference about high dimensional risk prediction models in the SSL setting especially under a possibly mis-specified dense model. In this paper, we fill in the gap by proposing an efficient surrogate assisted SSL (SAS) prediction procedure that leverages the fully observed surrogates $\mathbf{S}$ to make inference about a high dimensional risk model under such settings.

Our proposed estimation and inference procedures are as follows. For estimation, we first use the labelled data to fit a regularized imputation model with surrogates and high-dimensional covariates; then we impute the missing outcomes for the unlabeled data and fit the risk model using the imputed outcome and high-dimensional predictors. For inference, we devise a novel bias correction method, which corrects the bias due to the regularization for both imputation and estimation. Compared to existing literature, the key advantages of our proposed SAS procedure are

1. Applicable to dense risk model $Y \mid \mathbf{X}$: we allow the working risk model $Y \mid \mathbf{X}$ to be dense as long as the working imputation model $Y \mid \mathbf{S}, \mathbf{X}$ is sparse;

2. Robustness to model mis-specification: the working models on both risk prediction $Y \mid \mathbf{X}$ and imputation $Y \mid \mathbf{S}, \mathbf{X}$ can be mis-specified;

3. Requires no assumptions on the measurement error in $\mathbf{S}$ as proxies of Y and allows $\mathbf{S}$ itself to be of high dimension;

4. Our analysis on Lasso with estimated inputs in loss (see (6) and (20)) facilitates the consistency analysis for a dense model independent of the convergence rate of the consistently estimated inputs. The technique is an independent contribution to the high-dimensional statistics literatures.

The sparsity assumption on the imputation model is less stringent since we anticipate that most information on Y can be well captured by the low dimensional $\mathbf{S}$ while the fitted model of $Y \mid \mathbf{X}$ might be dense under possible model mis-specifications. Our theory uncovers that suitable use of unlabeled data may greatly relax the sparsity of $Y \mid \mathbf{X}$. As most literatures in SSL emphasized in the efficiency gain, our work opens a new direction of estimiability expansion through SSL.

1.1 Related Literatures

Under the supervised setting where both Y and $\mathbf{X}$ are fully observed, much progress has been made in recent years in the area of high dimensional inference. High dimensional regression methods have been developed for commonly used generalized linear models under sparsity assumptions on the regression parameters (van de Geer and Bühlmann, 2009; Negahban et al., 2010; Huang and Zhang, 2012). Recently, Zhu and Bradic (2018b) studied the inference of linear combination of coefficients under dense linear model and sparse precision matrix. Inference procedures have also been developed for both sparse (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014) and dense combinations of the regression parameters (Cai et al., 2019; Zhu and Bradic, 2018a). High-dimensional inference under the logistic regression model has also been studied recently (van de Geer et al., 2014; Ma et al., 2020; Guo et al., 2020).

Under the SSL setting with $n \ll N$, however, there is a paucity of literature on high dimensional inference. Although the SSL can be viewed as a missing data problem, it differs from the standard missing data setting in a critical way. Under the SSL setting, the missing probability tends to 1, which would violate a key assumption required in the missing data literature (Bang and Robins, 2005; Smucler et al., 2019; Chakrabortty et al., 2019, e.g.). Existing work on SSL with high-dimensional covariates largely focuses on the post-estimation inference on the global parameters under sparse linear models with examples including SSL estimation of population mean (Zhang et al., 2019; Zhang and Bradic, 2021), the explained variance (Cai and Guo, 2020), and the average treatment effect (Cheng et al., 2018; Kallus and Mao, 2020). Our SAS procedure is among the first attempts to conduct the semi-supervised inference of the high-dimensional coefficient and the individual prediction in the high-dimensional dense and possibly mis-specified risk prediction model. In a concurrent work, Deng et al. (2020) studied the efficient SSL estimation of high-dimensional linear models. Our work differs from them in at least three ways: 1) we consider the more flexible generalized linear models; 2) our setting involves the surrogates $\mathbf{S}$, characterizing the imprecise data in EHR; 3) we study dense coefficients whose number of nonzero elements exceeds the number of labels. In high-dimensional regression with missing data, another line of work studied the estimation of linear models with missing or noisy covariates $\mathbf{X}$ (Loh and Wainwright, 2011; Belloni et al., 2017; Chandrasekher et al., 2020).

The surrogates $\mathbf{S}$ can be viewed conceptually as "mis-measured" proxies of the true outcome Y. Semi-supervised methods have been developed under the assumption that $\mathbf{S}$ depends on $\mathbf{X}$ only through Y, which essentially assumes an independent measurement error in $\mathbf{S}$. For example, Gronsbell et al. (2019) studied the generalized linear risk prediction model using mis-measured $\mathbf{S}$. With a single S, Zhang et al. (2022) considered high-dimensional generalized linear model for the prediction model allowing the independence assumption to be slightly violated. Our SAS approach differs from the measurement error approach in two fundamental aspects: 1) typical measurement error approaches require S to be the single proxy outcome of the same type as Y while our SAS approach allow a vector $\mathbf{S}$ of arbitrary types as long as some of them are predictive for Y; 2) measurement error approaches impose stringent independence and model assumptions on the triplet $(S, \mathbf{X}, Y)$ while our SAS approach has neither. Violation of the two requirements may obstruct the deployment of measurement error methods or compromise its performance.

1.2 Organization of the Paper

The remainder of the paper is organized as follows. We introduce our population parameters and model assumptions in Section 2. In Section 3, we propose the SAS estimation method along with its associated inference procedures. In Section 4, we state the theoretical guarantees of the SAS procedures, whose proofs are provided in the Supplementary Materials. We also remark on the sparsity relaxation and the efficiency gain of the SSL. In Section 5, we present simulation results highlighting finite sample performance of the SAS estimators and comparisons to existing methods. In Section 6, we apply the proposed method to derive individual risk prediction for T2DM using EHR data from Mass General Brigham.

2. Settings and Notations

For the i-th observation, $Y_i \in \mathbb{R}$ denotes the outcome variable, $S_i \in \mathbb{R}^q$ denotes the surrogates for Y_i and $\mathbf{X}_i \in \mathbb{R}^{p+1}$ denotes the high-dimensional covariates with the first element being the intercept. Under the SSL setting, we observe n independent and identically distributed (i.i.d.) labeled observations, $\mathscr{L} = \{(Y_i, \mathbf{X}_i^\intercal, \mathbf{S}_i^\intercal)^\intercal, i = 1, ..., n\}$ and $(N - n)$ i.i.d unlabeled observations, $\mathscr{U} = \{\mathbf{W}_i = (\mathbf{X}_i^\intercal, \mathbf{S}_i^\intercal)^\intercal, i = n+1, ..., N\}$. We assume that the labeled subjects are randomly sampled by design and the proportion of labelled sample is $n/N = \rho \in (0, 1)$ with $\rho \to 0$ as $n \to \infty$. We focus on the high-dimensional setting where dimensions p and q grow with n and allow $p+q$ to be larger than n. Motivated by our application, our main focus is on the setting N much larger than p, but our approach can be extended to $N \le p$ under specific conditions.

To predict Y_i with $\mathbf{X}_i$, we consider a possibly mis-specified working regression model with a known monotone and smooth link function g,

$$Y_i \sim g(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i). \tag{1}$$

We identify the target parameter as the most predictive working model measured by the pseudo log-likelihood $\ell(y,x)$

$$\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} -\mathbb{E}\{\ell(Y_i, \boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\},\ \ell(y,x) = yx - G(x),\ G'(x) = g(x). \tag{2}$$

Here we do not assume any model for the true conditional expectation $\mathbb{E}(Y_i \mid \mathbf{X}_i)$. Our goal is to accurately estimate the high-dimensional parameter $\boldsymbol{\beta}_0$, alternatively characterized by the first order condition for (2),

$$\mathbb{E}[\mathbf{X}_i\{Y_i - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i)\}] = 0. \tag{3}$$

Our procedure generally allows for a wide range of link functions and detailed requirements on $g(\cdot)$ and its anti-derivative G are given in Section 4. In our motivating example, Y is a binary indicator of T2DM status and $g(x) = 1/(1+e^{-x})$ with $G(x) = \log(1+e^x)$. We shall further construct confidence intervals for $g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{x}_{\text{new}})$ with any $\mathbf{x}_{\text{new}} \in \mathbb{R}^{p+1}$. The predicted outcome $g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{x}_{\text{new}})$ can be interpreted as the maximum pseudo log-likelihood prediction under working model $g(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{\text{new}})$. We make no assumption on the sparsity of $\boldsymbol{\beta}_0$ relative to number of labels n, and hence it is not feasible to perform valid supervised learning for $\boldsymbol{\beta}_0$ when $s_{\boldsymbol{\beta}} = \|\boldsymbol{\beta}_0\|_0 > n$.

We shall derive an efficient SSL estimate for $\boldsymbol{\beta}_0$ by leveraging $\mathscr{U}$. To this end, we fit a working imputation model

$$Y_i \sim g(\boldsymbol{\gamma}^\mathsf{T}\mathbf{W}_i), \tag{4}$$

whose limiting parameter is likewise defined as the most predictive working model

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} -\mathbb{E}\{\ell(Y_i, \boldsymbol{\gamma}^\mathsf{T}\mathbf{W}_i)\} \Rightarrow \mathbb{E}[\mathbf{W}_i\{Y_i - g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\}] = 0. \tag{5}$$

The definition of $\boldsymbol{\gamma}$ guarantees

$$\mathbb{E}[\mathbf{X}_i\{Y_i - g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\}] = 0. \tag{6}$$

and hence if we impute Y_i as $\bar{Y}_i = g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)$, we have $\mathbb{E}[\mathbf{X}_i\{\bar{Y}_i - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i)\}] = 0$ regardless the adequacy of the imputation model (4) for the conditional mean $\mathbb{E}(Y_i \mid \mathbf{W}_i)$. It is thus feasible to carry out an SSL procedure by first deriving an estimate for $\bar{Y}_i$ using the labelled data $\mathscr{L}$ and then regressing the estimated $\bar{Y}_i$ against $\mathbf{X}_i$ using the whole data $\mathscr{L} \cup \mathscr{U}$. Although we do not require $\boldsymbol{\beta}_0$ to be sparse or any of the fitted models to hold, we do assume that $\boldsymbol{\gamma}_0$ defined in (5) to be sparse. When the surrogates $\mathbf{S}$ are strongly predictive for the outcome, the sparsity assumption on $\boldsymbol{\gamma}_0$ is reasonable since the majority of the information in Y can be captured in $\mathbf{S}$.

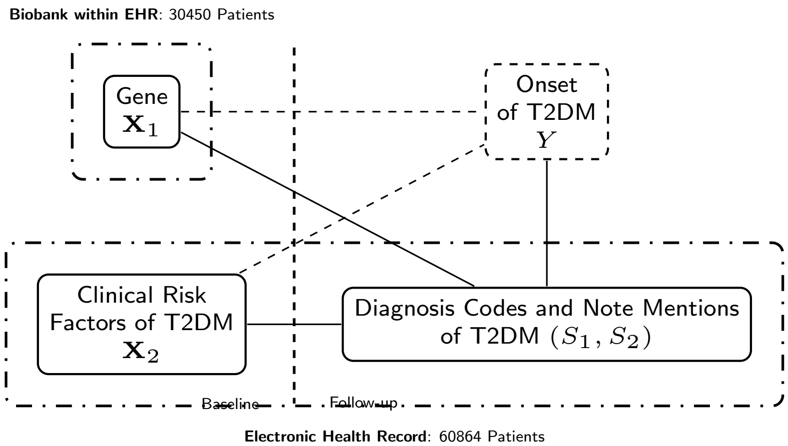

Figure 1: A dense prediction model (graph with dashed lines) can be compress to a sparse imputation model (through graph with solid lines) when the effect of most baseline covariates are reflected in a few variables in the EHR monitoring the development of the event of interest.

Notations. We focus on the setting where $\min\{n, p+q, N\} \to \infty$. For convenience, we shall use $n \to \infty$ in the asymptotic analysis. For two sequences of random variables A_n and B_n, we use $A_n = O_p(B_n)$ and $A_n = o_p(B_n)$ to denote $\lim_{c\to\infty} \lim_{n\to\infty} \mathbb{P}(|A| \geq c|B|) = 0$ and $\lim_{c\to 0} \lim_{n\to\infty} \mathbb{P}(|A| \geq c|B|) = 0$, respectively. For two positive sequences a_n and b_n, $a_n = O(b_n)$ or $b_n \gtrsim a_n$ means that $\exists C > 0$ such that $a_n \leq C b_n$ for all n; $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$, and $a_n \ll b_n$ or $a_n = o(b_n)$ if $\limsup_{n\to\infty} a_n/b_n = 0$. We use $Z_n \overset{\mathscr{L}}{\to} N(0,1)$ to denote the sequence of random variables Z_n converges in distribution to a standard normal random variable.

3. Methodology

3.1 SAS Estimation of $\boldsymbol{\beta}_0$

The SAS estimation procedure for $\boldsymbol{\beta}_0$ consists of two key steps: (i) fitting the imputation model to $\mathscr{L}$ to obtain estimate $\widehat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}_0$ defined in (5); and (ii) estimating $\boldsymbol{\beta}_0$ in (3) by fitting imputed outcome $\widehat{Y}_i = g(\widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{W}_i)$ against $\mathbf{X}_i$ to $\mathscr{U}$. In both steps, we devise the Lasso type estimator to deal with the high-dimensionality of $\mathbf{X}$. In principle, other types of variable selection methods, e.g. SCAD (Fan and Li, 2001) or square-root Lasso (Belloni et al., 2011), may also be used. We use the Lasso as the example for its simplicity. A further discussion on the choice of regularized estimators is given in Remark 6.

In Step (i), we estimate $\boldsymbol{\gamma}_0$ by the L_1 regularized pseudo log-likelihood estimator $\widehat{\boldsymbol{\gamma}}$, defined as

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p+q+1}}{\operatorname{argmin}} \ \ell_{\mathsf{imp}}(\boldsymbol{\gamma}) + \lambda_{\gamma} \|\boldsymbol{\gamma}_{-1}\|_1 \quad \text{with} \quad \lambda_{\gamma} \asymp \sqrt{\log(p+q)/n}, \tag{7}$$

where $\mathbf{a}_{-1}$ denotes the sub-vector of all the coefficients except for the intercept and

$$\ell_{\mathsf{imp}}(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\gamma}^{\mathsf{T}}\mathbf{W}_i) \quad \text{with} \quad \ell(y,x) \text{ defined in } (2). \tag{8}$$

The imputation loss (8) corresponds to the negative log-likelihood when Y is binary and the imputation model holds with g being anti-logit. With $\widehat{\boldsymbol{\gamma}}$, we impute the unobserved outcomes for subjects in $\mathscr{U}$ as $\widehat{Y}_i = g(\widehat{\boldsymbol{\gamma}}^{\mathsf{T}}\mathbf{W}_i)$, for $n+1 \le i \le N$.

In Step (ii), we estimate $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})$, defined as,

$$\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \ell^{\dagger}(\boldsymbol{\beta};\widehat{\boldsymbol{\gamma}}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_{-1}\|_1 \quad \text{with} \quad \lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log(p)/N}, \tag{9}$$

where $\ell^{\dagger}(\boldsymbol{\beta};\widehat{\boldsymbol{\gamma}})$ is the imputed pseudo log-likelihood:

$$\ell^{\dagger}(\boldsymbol{\beta};\widehat{\boldsymbol{\gamma}}) = \frac{1}{N}\sum_{i>n} \ell(\widehat{Y}_i, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i) + \frac{1}{N}\sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i) \quad \text{with} \quad \ell(y,x) \text{ defined in } (2). \tag{10}$$

We denote the complete data pseudo log-likelihood of the full data as

$$\ell_{\mathsf{PL}}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \ell(Y_i, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i). \tag{11}$$

and define the gradients of the various losses (8)-(11) as

$$\dot{\boldsymbol{\ell}}_{\mathsf{imp}}(\boldsymbol{\gamma}) = \nabla\ell_{\mathsf{imp}}(\boldsymbol{\gamma}),\ \dot{\boldsymbol{\ell}}_{\mathsf{PL}}(\boldsymbol{\beta}) = \nabla\ell_{\mathsf{PL}}(\boldsymbol{\beta}),\ \dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta};\boldsymbol{\gamma}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell^{\dagger}(\boldsymbol{\beta};\boldsymbol{\gamma}). \tag{12}$$

3.2 SAS Inference for Individual Prediction

Since $g(\cdot)$ is specified, the inference on $g(\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta})$ immediately follows from the inference on $\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta}$. We shall consider the inference on standardized linear prediction $\mathbf{x}_{\text{std}}^{\mathsf{T}}\boldsymbol{\beta}$ with the standardized covariates

$$\mathbf{x}_{\text{std}} = \mathbf{x}_{\text{new}}/\|\mathbf{x}_{\text{new}}\|_2$$

and then scale the confidence interval back. This way, the scaling with $\|\mathbf{x}_{\text{new}}\|_2$ is made explicit in the expression of the confidence interval.

The estimation error of $\widehat{\boldsymbol{\beta}}$ can be decomposed into two components corresponding to the respective errors associated with (7) and (9). Specifically, we write

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \{\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0\} + \{\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})\}, \tag{13}$$

where $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})$ is defined as the minimizer of the expected imputed loss conditionally on the labeled data, that is,

$$\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \mathbb{E}[\ell^{\dagger}(\boldsymbol{\beta};\widehat{\boldsymbol{\gamma}}) \mid \mathscr{L}]. \tag{14}$$

The term $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0$ denotes the error from the imputation model in (7) while the term $\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})$ denotes the error from the prediction model in (9) given the imputation model parameter $\widehat{\boldsymbol{\gamma}}$. As ℓ_1 penalization is involved in both steps, we shall correct the regularization bias from the two sources. Following from the typical one-step debiasing LASSO (Zhang and Zhang, 2014), the bias $\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})$ is estimated by $\widehat{\boldsymbol{\Theta}}\dot{\boldsymbol{\ell}}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})$ where $\widehat{\boldsymbol{\Theta}}$ is an estimator of $[\mathbb{E}\{g'(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\intercal}\}]^{-1}$, the inverse Hessian of $\ell^{\dagger}(\cdot;\widehat{\boldsymbol{\gamma}})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

The bias correction for $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0$ requires some innovation since we need to conduct the bias correction for a nonlinear functional $\bar{\boldsymbol{\beta}}(\cdot)$ of LASSO estimator $\widehat{\boldsymbol{\gamma}}$, which has not been studied in the literature. We identify $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})$ and $\boldsymbol{\beta}_0$ by the first order moment conditions,

$$\begin{aligned}\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) &: \mathbb{E}_{i>n}[\mathbf{X}_i\{g(\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})^{\intercal}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)\} \mid \mathscr{L}] \approx 0, \\ \boldsymbol{\beta}_0 &: \mathbb{E}[\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i) - Y_i\}] = \mathbb{E}[\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i) - g(\boldsymbol{\gamma}_0^{\intercal}\mathbf{W}_i)\}] = 0. \end{aligned} \tag{15}$$

Here $\mathbb{E}_{i>n}[\cdot \mid \mathscr{L}]$ denotes the conditional expectation of a single copy of the unlabeled data given the labelled data. By equating the two estimating equations in (15), we apply the first order approximation and approximate the difference $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0$ by

$$\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0 \approx - [\mathbb{E}\{g'(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\intercal}\}]^{-1}\mathbb{E}_{i>n}\left[\mathbf{X}_i\{g(\boldsymbol{\gamma}_0^{\intercal}\mathbf{W}_i) - g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)\} \mid \mathscr{L}\right] \tag{16}$$

Together with the bias correction for $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0$, this motivates the debiasing procedure

$$\widehat{\boldsymbol{\beta}} - \frac{1-\rho}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}\mathbf{X}_i\left\{g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i) - Y_i\right\} - \widehat{\boldsymbol{\Theta}}\dot{\boldsymbol{\ell}}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}).$$

The $1-\rho$ factor, which tends to one when n much smaller than N, comes from the proportion of unlabeled data whose missing outcome are imputed.

For theoretical considerations, we devise a cross-fitting scheme in our debiasing process. We split the labelled and unlabeled data into K folds of approximately equal size, respectively. The number of folds does not grow with dimension (e.g. $K = 10$). We denote the indices sets for each fold of the labelled data $\mathscr{L}$ as $\mathcal{I}_1, \ldots, \mathcal{I}_K$, and those of the unlabeled data $\mathscr{U}$ as $\mathcal{J}_1, \ldots, \mathcal{J}_K$. We denote the respective sizes of each fold in the labelled data and full data as $n_k = |\mathcal{I}_k|$ and $N_k = n_k + |\mathcal{J}_k|$, where $|\mathcal{I}|$ denotes the carnality of $\mathcal{I}$. Define $\mathcal{I}_k^c = \{1, \ldots, n\} \setminus \mathcal{I}_k$ and $\mathcal{J}_k^c = \{n+1, \ldots, N\} \setminus \mathcal{J}_k$. For each labelled fold k, we fit the imputation model with out-of-fold labelled samples:

$$\widehat{\boldsymbol{\gamma}}^{(k)} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p+q+1}}{\operatorname{argmin}} \frac{1}{n-n_k}\sum_{i\in\mathcal{I}_k^c}\ell(Y_i, \boldsymbol{\gamma}^{\intercal}\mathbf{W}_i) + \lambda_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma}_{-1}\|_1. \tag{17}$$

Using $\widehat{\boldsymbol{\gamma}}^{(k)}$, we fit the prediction model with the out-of-fold data $\mathcal{I}_k^c \cup \mathcal{J}_k^c$:

$$\widehat{\boldsymbol{\beta}}^{(k)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{1}{N-N_k}\left[\sum_{i\in\mathcal{J}_k^c}\ell(g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i), \boldsymbol{\beta}^{\intercal}\mathbf{X}_i) + \sum_{i\in\mathcal{I}_k^c}\ell(Y_i, \boldsymbol{\beta}^{\intercal}\mathbf{X}_i)\right] + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_{-1}\|_1. \tag{18}$$

To estimate the projection

$$\mathbf{u}_0 = \mathbb{E}\{g'(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\intercal}\}]^{-1}\mathbf{x}_{\text{std}}, \tag{19}$$

we propose an L_1-penalized estimator

$$\widehat{\mathbf{u}}^{(k)} = \underset{\mathbf{u}\in\mathbb{R}^{p}}{\operatorname{argmin}} \frac{1}{N-N_k}\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left[\frac{1}{2}g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)(\mathbf{X}_i^{\intercal}\mathbf{u})^2 - \mathbf{u}^{\intercal}\mathbf{x}_{\text{std}}\right] + \lambda_{\mathbf{u}}\|\mathbf{u}\|_1, \tag{20}$$

where $\widehat{\boldsymbol{\beta}}^{(k,k')}$ is trained with samples out of folds k and k',

$$\widehat{\boldsymbol{\beta}}^{(k,k')} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \frac{\sum_{i\in(\mathcal{J}_k\cup\mathcal{J}_{k'})^c} \ell\left(g\left(\widehat{\boldsymbol{\gamma}}^{(k,k')\intercal}\mathbf{W}_i\right), \boldsymbol{\beta}^\intercal\mathbf{X}_i\right) + \sum_{i\in(\mathcal{I}_k\cup\mathcal{I}_{k'})^c} \ell(Y_i, \boldsymbol{\beta}^\intercal\mathbf{X}_i)}{N - N_k - N_{k'}} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_{-1}\|_1, \tag{21}$$

$$\text{with} \quad \widehat{\boldsymbol{\gamma}}^{(k,k')} = \operatorname*{argmin}_{\boldsymbol{\gamma}\in\mathbb{R}^{p+q+1}} \frac{\sum_{i\in\mathcal{I}_k^c\cap\mathcal{I}_{k'}^c} \ell(Y_i, \boldsymbol{\gamma}^\intercal\mathbf{W}_i)}{n - n_k - n_{k'}} + \lambda_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma}_{-1}\|_1.$$

The estimators in (21) take similar forms as those in (17) and (18) except that their training samples exclude two folds of data $\mathcal{I}_k \cup \mathcal{J}_k$ and $\mathcal{I}_{k'} \cup \mathcal{J}_{k'}$. In the summand of (20), the data $(Y_i, \mathbf{X}_i, \mathbf{S}_i)$ in fold k' $\mathcal{I}_{k'} \cup \mathcal{J}_{k'}$ is independent of $\widehat{\boldsymbol{\beta}}^{(k,k')}$ trained without folds k and k'. The estimation of $\mathbf{u}$ requires an estimator of $\boldsymbol{\beta}$ and both estimators are subsequently used for the debiasing step. Using the same set of data multiple times for $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{u}}$, debiasing and variance estimation may induce over-fitting bias, so we implemented the cross-fitting scheme to reduce the over-fitting bias. As a remark, cross-fitting might not be necessary for theory with additional assumptions and/or empirical process techniques.

We obtain the cross-fitted debiased estimator for $\mathbf{x}_{\text{std}}^\intercal\boldsymbol{\beta}$ as $\widehat{\mathbf{x}_{\text{std}}^\intercal\boldsymbol{\beta}}$, defined as

$$\begin{aligned} &\frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_{\text{std}}^\intercal \widehat{\boldsymbol{\beta}}^{(k)} - \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k} \widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)\} \\ &- \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i\left\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i) + \rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - Y_i\right\}. \end{aligned} \tag{22}$$

The second term is used to correct the bias $\bar{\boldsymbol{\beta}}(\widehat{\gamma}) - \boldsymbol{\beta}_0$ and the third term is used to correct the bias $\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}(\widehat{\gamma})$. The corresponding variance estimator is

$$\begin{aligned} \widehat{V}_{\mathsf{SAS}} =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} (\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\left\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i) + \rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - Y_i\right\}^2 \\ &+ \frac{\rho^2}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k} (\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\left\{g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^\intercal\mathbf{W}_i)\right\}^2 \end{aligned} \tag{23}$$

Through the link g and the scaling factor $\|\mathbf{x}_{\text{new}}\|_2$, we estimate $g(\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_0)$ by $g\left(\|\mathbf{x}_{\text{new}}\|_2\widehat{\mathbf{x}_{\text{std}}^\intercal\boldsymbol{\beta}}\right)$ and construct the $(1-\alpha)\times 100\%$ confidence interval for $g(\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_0)$ as

$$\left[g\left\{\|\mathbf{x}_{\text{new}}\|_2\left(\widehat{\mathbf{x}_{\text{std}}^\intercal\boldsymbol{\beta}} - \mathcal{Z}_{\alpha/2}\sqrt{\widehat{V}_{\mathsf{SAS}}/n}\right\}\right), g\left\{\|\mathbf{x}_{\text{new}}\|_2\left(\widehat{\mathbf{x}_{\text{std}}^\intercal\boldsymbol{\beta}} + \mathcal{Z}_{\alpha/2}\sqrt{\widehat{V}_{\mathsf{SAS}}/n}\right)\right\}\right], \tag{24}$$

where $Z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

4. Theory

We introduce assumptions required for both estimation and inference in Section 4.1. We state our theories for estimation and inference, respectively in Sections 4.2 and 4.3.

4.1 Assumptions

We assume the complete data consist of i.i.d. copies of $(Y_i, \mathbf{X}_i, \mathbf{S}_i)$, for $i = 1, \ldots, N$. For our focused SSL settings, only the first n outcome labels $Y_i, \ldots, Y_n$ are observed. Under the i.i.d assumption, our SSL setting is equivalent to the missing completely at random (MCAR) assumption. The sparsities of $\boldsymbol{\gamma}_0$, $\boldsymbol{\beta}_0$ and $\mathbf{u}_0$ are denoted as

$$s_{\boldsymbol{\gamma}} = \|\boldsymbol{\gamma}_0\|_0,\ s_{\boldsymbol{\beta}} = \|\boldsymbol{\beta}_0\|_0,\ s_{\mathsf{u}} = \|\mathbf{u}_0\|_0.$$

We focus on the setting with $n, p+q, N \to \infty$ with n being allowed to be smaller than $p+q$. We allow that $s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\beta}}$ and s_{u} grow with $n, p+q, N$ and satisfy $s_{\boldsymbol{\gamma}} \ll n$ and $s_{\boldsymbol{\beta}} + s_{\mathsf{u}} \ll N$. While our method and theory adaptively applies to both SSL ($N \gg n$) and missing data ($N \asymp n$) settings without prior knowledge on the limit of n/N, we emphasize on the SSL ($N \gg n$) setting that matches our motivating EHR studies and is also less studied in the literature. To achieve the sharper dimension conditions, we consider the sub-Gaussian design as in Portnoy (1984, 1985); Negahban et al. (2010). We denote the sub-Gaussian norm for random variables and random vectors both as $\|\cdot\|_{\psi_2}$. The detailed definition is given in Appendix D.

Assumption 1 *For constants ν_1, ν_2 and M independent of n, p and N,*

a) *the residuals $Y_i - g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)$ and $Y_i - g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)$ are sub-Gaussian random variables with sub-Gaussian norm bounded by $\|Y_i - g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)\|_{\psi_2} \le \nu_1$ and $\|Y_i - g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\|_{\psi_2} \le \nu_2$;*

b) *The link function g satisfies the monotonicity and smoothness conditions: $\inf_{x\in\mathbb{R}} g'(x) \ge 0$, $\sup_{x\in\mathbb{R}} g'(x) < M$ and $\sup_{x\in\mathbb{R}} g''(x) < M$.*

Under our motivating example with a binary Y_i and $g(x) = e^x/(1+e^x)$, 1a and 1b are satisfied. The condition is also satisfied for the probit link function and the identity link function. Condition 1a is universal for high-dimensional regression. Admittedly, Lipschitz requirement in 1b rules out some generalized linear model links with unbounded derivatives like the exponential link, but we may substitute the condition by assuming a bounded $\|\mathbf{X}_i\|_\infty$.

Assumption 2 *For constants σ^2_{max} and σ^2_{min} independent of n, p, N,*

a) *$\mathbf{W}_i$ is a sub-Gaussian vector with sub-Gaussian norm $\|\mathbf{W}_i\|_{\psi_2} \le \sigma_{\mathsf{max}}/\sqrt{2}$;*

b) *The weak overlapping condition at the population parameter $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$,*

   (i) $\inf_{\|\mathbf{v}\|_2=1} \mathbf{v}^{\mathsf{T}}\mathbb{E}([g'(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i) \wedge 1]\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}})\mathbf{v} \ge \sigma^2_{\mathsf{min}}$,

   (ii) $\inf_{\|\mathbf{v}\|_2=1} \mathbf{v}^{\mathsf{T}}\mathbb{E}[\{g'(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i) \wedge 1\}\mathbf{W}_i\mathbf{W}_i^{\mathsf{T}}]\mathbf{v} \ge \sigma^2_{\mathsf{min}}$;

c) *The non-degeneracy of average residual variance:*

$$\inf_{\|\mathbf{v}\|_2=1} \mathbb{E}[\{Y_i - (1-\rho)\cdot g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i) - \rho\cdot g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\}^2(\mathbf{X}_i^{\mathsf{T}}\mathbf{v})^2] \ge \sigma^2_{\mathsf{min}}.$$

Assumption 2a is typical for high-dimensional regression (Negahban et al., 2010), which also implies the bounded maximal eigenvalue of the second moment

$$\sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^{\intercal}\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^{\intercal}]\mathbf{v} \leq \sigma_{\max}^2.$$

Notably, we do not require two common conditions under high-dimensional generalized linear models (Huang and Zhang, 2012; van de Geer et al., 2014): 1) the upper bound on $\sup_{i=1,\ldots,N}\|\mathbf{X}_i\|_\infty$; 2) the lower bound on $\inf_{i=1,\ldots,N} g'(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)$, often known as the overlapping condition for logistic regression model. Compared to the overlapping condition under logistic regression that $g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)$ and $g(\boldsymbol{\gamma}_0^{\intercal}\mathbf{W}_i)$ are bounded away from zero, our Assumptions 2b and 2c are weaker because they are implied by the typical minimal eigenvalue condition

$$\inf_{\|\mathbf{v}\|_2=1} \mathbf{v}^{\intercal}\mathbb{E}(\mathbf{W}_i\mathbf{W}_i^{\intercal})\mathbf{v} \geq \sigma_{\min}^2$$

plus the overlapping condition.

4.2 Consistency of the SAS Estimation

We now state the L_2 and L_1 convergence rates of our proposed SAS estimator.

Theorem 1 (Consistency of SAS estimation) *Under Assumptions 1, 2 and with*

$$s_{\boldsymbol{\gamma}} = o(n/\log(p+q)),\ s_{\boldsymbol{\beta}} = o(N/\log(p)),\ \lambda_{\boldsymbol{\beta}} \gtrsim \sqrt{\log(p)/N}, \tag{25}$$

we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\beta}}}\lambda_{\boldsymbol{\beta}} + (1-\rho)\sqrt{s_{\boldsymbol{\gamma}}\log(p+q)/n}\right),$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O_p\left(s_{\boldsymbol{\beta}}\lambda_{\boldsymbol{\beta}} + (1-\rho)^2 s_{\boldsymbol{\gamma}}\log(p+q)/(n\lambda_{\boldsymbol{\beta}})\right).$$

Remark 2 *The dimension requirement for our SAS estimator achieving L_2 consistency significantly weakens the existing dimension requirement in the supervised setting (Negahban et al., 2010; Huang and Zhang, 2012; Bühlmann and Van De Geer, 2011; Bickel et al., 2009) With $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log(p)/N}$, Theorem 1 implies the L_2 consistency of $\widehat{\boldsymbol{\beta}}$ under the dimension condition,*

$$(1-\rho)^2 s_{\boldsymbol{\gamma}}\log(p+q)/n + s_{\boldsymbol{\beta}}\log(p)/N = o(1). \tag{26}$$

When $N \gg n$, our requirement on the sparsity of β, $s_{\boldsymbol{\beta}} = o(N/\log(p))$ is significantly weaker than $s_{\boldsymbol{\beta}} = o(n/\log(p))$, which is known as the fundamental sparsity limit to identify the high-dimensional regression vector in the supervised setting. Theorem 1 indicates that with assistance from observed $\mathbf{S} \in \mathscr{U}$, the SAS procedure allows $s_{\boldsymbol{\beta}} > n$ provided that N is sufficiently large and the imputation model is sparse. This distinguishes our result from most estimation results in high-dimensional supervised settings. Among SSL literatures, the utility of unlabeled data in relaxation of sparsity condition has never been conceived before.

Remark 3 *In the context of Theorem 1, a sparse imputation, often induced by a small number of highly predictive surrogates, is essential for an optimal estimation rate. When $s_{\boldsymbol{\beta}} > s_{\boldsymbol{\gamma}}$, the L_2 rate in Theorem 1 has two components, $\sqrt{s_{\boldsymbol{\beta}}\log(p)/N}$ regarding the minimax rate to learn $\boldsymbol{\beta}$ from all data and $\sqrt{s_{\boldsymbol{\gamma}}\log(p+q)/n}$ regarding the minimax rate to learn $\boldsymbol{\gamma}$ in the labeled data (Raskutti et al., 2011). Thus, the rate cannot be further improved if the sparser imputation model is used to identify the denser $\boldsymbol{\beta}$ without additional conditions.*

Remark 4 *If the L_1 consistency is of interest, the penalty levels are chosen as*

$$\lambda_{\boldsymbol{\beta}} \asymp \max\left\{\sqrt{\log(p)/N}, \sqrt{s_{\boldsymbol{\gamma}}/s_{\boldsymbol{\beta}}}\lambda_{\boldsymbol{\gamma}}\right\}, \tag{27}$$

which produces the L_1 estimation rate from Theorem 1

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O_p\left(s_{\boldsymbol{\beta}}\sqrt{\log(p)/N} + \sqrt{s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} \log(p)/n}\right).$$

Compared to the condition for L_1 consistency under supervised learning, $s_{\boldsymbol{\beta}} = o\left(\sqrt{n/\log(p)}\right)$, the condition from SAS estimation $s_{\boldsymbol{\beta}} = o\left((n/s_{\boldsymbol{\gamma}} + N)/\log(p)\right)$ allows a denser $\boldsymbol{\beta}_0$ in the setting with a very sparse $\boldsymbol{\gamma}_0$ and a large unlabeled data. On the other hand, the L_2 estimation rate in Theorem 1 remains the same if

$$\sqrt{\log(p)/N} \lesssim \lambda_{\boldsymbol{\beta}} \lesssim \max\left\{\sqrt{\log(p)/N}, \sqrt{s_{\boldsymbol{\gamma}}/s_{\boldsymbol{\beta}}}\lambda_{\boldsymbol{\gamma}}\right\}.$$

Our theory on the SAS inference procedure uses the L_2 instead of the L_1 consistency.

Theorem 1 implies the following prediction consistency result.

Corollary 5 (Consistency of individual prediction) *Suppose $\mathbf{x}_{\text{new}}$ is sub-Gaussian random vector satisfying $\sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^\mathsf{T}\mathbb{E}[\mathbf{x}_{\text{new}}\mathbf{x}_{\text{new}}^\mathsf{T}]\mathbf{v} \le \sigma_{\mathsf{max}}^2$. Under the conditions of Theorem 1, we have*

$$g\left(\widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_{\text{new}}\right) - g\left(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{x}_{\text{new}}\right) = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\right) = o_p(1).$$

The concentration result of Corollary 5 is established with respect to the joint distribution of the data and the new observation $\mathbf{x}_{\text{new}}$. This is in a sharp contrast to the individual prediction conditioning on any new observation $\mathbf{x}_{\text{new}}$. If the goal is to conduct inference for any given $\mathbf{x}_{\text{new}}$, the theoretical justification is provided in the following Theorem 7 and Corollary 8.

Remark 6 *Other types of penalties shown to provide consistent estimation in L_2 for the working imputation model can substitute the Lasso penalty in (7), since the L_2 rate $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2$ is the only property invoked for $\widehat{\boldsymbol{\gamma}}$ in the proof of Theorem 1. For example, we may choose the square-root Lasso (Belloni et al., 2011) with pivotal recovery under linear models with identity link $g(x) = x$. Changing the Lasso penalty in (9), however, might require a different proof to produce the stated estimation rate adaptive to arbitrary $s_{\boldsymbol{\beta}}/N$ and $s_{\boldsymbol{\gamma}}/n$, covering both $s_{\boldsymbol{\beta}}/N \ll s_{\boldsymbol{\gamma}}/n$ and $s_{\boldsymbol{\beta}}/N \gtrsim s_{\boldsymbol{\gamma}}/n$ settings (Case 1 and 2, respectively, in the Proof of Theorem 1). If the $s_{\boldsymbol{\beta}}/N \ll s_{\boldsymbol{\gamma}}/n$ setting guaranteed by a very large N alone is of interest, other penalties for $\widehat{\boldsymbol{\beta}}$ can work equally well (by adapting Case 1 in the Proof of Theorem 1).*

4.3 $\sqrt{n}$-inference with Debiased SAS Estimator

We state the validity of our SSL inference in Theorem 7. We use to $A \xrightarrow{\mathscr{L}} B$ to denote that random variable A converges in distribution to a distribution B.

Theorem 7 (SAS Inference) *Let $\mathbf{x}_{\rm new}$ be the random vector representing the covariate of a new individual. Under Assumptions 1, 2 and the dimension condition*

$$(1-\rho)^4 \frac{s_{\gamma}^2 \log(p+q)^2}{n} + \frac{\rho(s_{\beta}^2 + s_{\beta}s_{\mathsf{u}})\log(p)^2}{N} + (1-\rho)^2 \frac{s_{\gamma}s_{\mathsf{u}}\log(p+q)\log(p)}{N} = o(1), \quad (28)$$

we draw inference on $\mathbf{x}_{\rm new}^{\mathsf{T}}\boldsymbol{\beta}_0$ conditionally on $\mathbf{x}_{\rm new}$ according to

$$\sqrt{n}\widehat{V}_{\mathsf{SAS}}^{-1/2}\left(\widehat{\mathbf{x}_{\rm std}^{\mathsf{T}}\boldsymbol{\beta}} - \frac{\mathbf{x}_{\rm new}^{\mathsf{T}}\boldsymbol{\beta}_0}{\|\mathbf{x}_{\rm new}\|_2}\right) \mid \mathbf{x}_{\rm new} \xrightarrow{\mathscr{L}} N(0,1),$$

where $\widehat{V}_{\mathsf{SAS}}^2$ defined in (23) *is the estimator of the asymptotic variance*

$$\begin{aligned} V_{\mathsf{SAS}} =& \mathbb{E}[(\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i)^2\{Y - (1-\rho)\cdot g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i) - \rho\cdot g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\}^2] \\ &+ \rho(1-\rho)\mathbb{E}[(\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i)^2\{g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i) - g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\}^2], \\ \textit{with} \quad & \mathbf{u}_0 = \boldsymbol{\Theta}_0 \frac{\mathbf{x}_{\rm new}}{\|\mathbf{x}_{\rm new}\|_2} = [\mathbb{E}\{g'(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}\}]^{-1}\frac{\mathbf{x}_{\rm new}}{\|\mathbf{x}_{\rm new}\|_2}. \end{aligned} \quad (29)$$

By the Young's inequality, the condition (28) is implied by

$$(1-\rho)^4 \frac{s_{\gamma}^2 \log(p+q)^2}{n} + \frac{\sqrt{\rho}(s_{\beta} + s_{\mathsf{u}})\log(p)}{\sqrt{N}} = o(1), \quad (30)$$

When p is much smaller than the full sample size N, our condition (30) allows the sparsity levels of $\boldsymbol{\beta}_0$ and $\mathbf{u}_0$ to be as large as p. Even if p is larger than N, our SAS inference procedure is valid if $s_{\beta} + s_{\mathsf{u}} \lesssim \sqrt{N}/\log(p)$. In the literature on confidence interval construction in high-dimensional supervised setting, the valid inference procedure for a single regression coefficient in the linear regression requires $s_{\beta} \lesssim \sqrt{n}/\log(p)$ (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014). Such a sparsity condition has been shown to be necessary to construct a confidence interval of a parametric rate (Cai and Guo, 2017). We have leveraged the unlabeled data to significantly relax the fundamental limit of statistical inference from $s_{\beta} \lesssim \sqrt{n}/\log(p)$ to $s_{\beta} \lesssim N/\{\log(p)\sqrt{n}\}$. The amount of labelled data validates the statistical inference for a dense model in high dimensions.

The sparsity of $\mathbf{u}_0$ is determined by $\mathbf{x}_{\rm new}$ and the precision matrix $\boldsymbol{\Theta}_0$. In the supervised learning setting, for confidence interval construction for a single regression coefficient, van de Geer et al. (2014) requires $s_{\mathsf{u}} \lesssim n/\log(p)$ is required. According to (30), our SAS inference requires $s_{\mathsf{u}} \lesssim N/\{\log(p)\sqrt{n}\}$, which can be weaker than $s_{\mathsf{u}} \lesssim n/\log(p)$ if the amount of unlabeled data is larger than n^2. Theorem 7 implies that our proposed CI in (24) is valid in terms of coverage, which is summarized in the following corollary.

Corollary 8 *Under Assumptions 1 and 2, as well as* (28)*, the CI defined in* (24) *satisfies,*

$$\begin{aligned} \mathbb{P}\Bigg\{ g\left(\|\mathbf{x}_{\rm new}\|_2\left(\widehat{\mathbf{x}_{\rm std}^{\mathsf{T}}\boldsymbol{\beta}} - \mathcal{Z}_{\alpha/2}\sqrt{\widehat{V}_{\mathsf{SAS}}/n}\right)\right) &\le g\left(\mathbf{x}_{\rm new}^{\mathsf{T}}\boldsymbol{\beta}_0\right) \\ \le g\left(\left(\|\mathbf{x}_{\rm new}\|_2\widehat{\mathbf{x}_{\rm std}^{\mathsf{T}}\boldsymbol{\beta}} + \mathcal{Z}_{\alpha/2}\sqrt{\widehat{V}_{\mathsf{SAS}}/n}\right)\right)\Bigg\} &= 1-\alpha+o(1). \end{aligned}$$

$$2g'(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{\rm new})\|\mathbf{x}_{\rm new}\|_2\mathcal{Z}_{\alpha/2}\sqrt{V_{\mathsf{SAS}}/n} \lesssim \|\mathbf{x}_{\rm new}\|_2/\sqrt{n},$$

where V_{SAS} is the the asymptotic variance defined in (29)*.*

Confidence interval construction for $g\left(\mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_0\right)$ in high-dimensional supervised setting has been recently studied in Guo et al. (2020). Guo et al. (2020) assumes the prediction model to be correctly specified as a high-dimensional sparse logistic regression and the inference procedure is valid if $s_{\boldsymbol{\beta}} \lesssim \sqrt{n}/\log p$. In contrast, we leverage the unlabeled data to allow for mis-specified prediction model and a dense regression vector, as long as the dimension requirement in (28) is satisfied.

4.4 Efficiency comparison of SAS Inference

Efficiency in high-dimensional setting or SSL setting in which the proportion of labelled data decays to zero is yet to be formalized. Here we use the efficiency bound in the classical low-dimensional with a fixed ρ as the benchmark. Apart from the relaxation of various sparsity conditions, we illustrate next that our SAS inference achieves a decent efficiency with properly specified imputation model compared to the supervised learning and the benchmark.

Similar to the phenomenon discovered by Chakrabortty and Cai (2018), if the imputation model is correct, we can guarantee the efficiency gain by SAS inference in comparison to the asymptotic variance of the supervised learning,

$$V_{\mathsf{SL}} = \mathbb{E}[(\mathbf{u}_0^{\intercal}\mathbf{X}_i)^2\{Y_i - g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)\}^2]. \tag{31}$$

Proposition 9 *If $\mathbb{E}(Y_i \mid \mathbf{S}_i, \mathbf{X}_i) = g(\boldsymbol{\gamma}_0^{\intercal}\mathbf{W}_i)$, we have $V_{\mathsf{SL}} \geq V_{\mathsf{SAS}}$.*

Moreover, we can show that our SAS inference attains the benchmark efficiency derived from classical fixed ρ setting (Tsiatis, 2007). To simplify the derivation, we describe the missing-completely-at-random mechanism through the binary observation indicator R_i, $i = 1, \dots, N$, independent of Y_i, $\mathbf{X}_i$ and $\mathbf{S}_i$. We still denote the proportion of labelled data as $\rho = \mathbb{E}(R_i)$. The unsorted data take the form

$$\mathscr{D} = \left\{\mathbf{D}_i = (\mathbf{X}_i^{\intercal}, \mathbf{S}_i^{\intercal}, R_i, R_iY_i)^{\intercal}, i = 1, \dots, N\right\}.$$

We consider the following class of complete data semi-parametric models

$$\begin{aligned}\mathcal{M}_{\mathsf{comp}} = \Big\{ & f_{\mathbf{X},Y,\mathbf{S},R}(\mathbf{x}, y, \mathbf{s}, r) = f_{\mathbf{X}}(\mathbf{x})\rho^r(1-\rho)^{1-r} f_{Y|\mathbf{S},\mathbf{X}}(y|\mathbf{s}, \mathbf{x}) f_{\mathbf{S}|\mathbf{X}}(\mathbf{s}|\mathbf{x}) : \\ & f_{Y|\mathbf{S},\mathbf{X}}, f_{\mathbf{X}}, f_{\mathbf{S}|\mathbf{X}} \text{ are arbitrary density}\Big\},\end{aligned} \tag{32}$$

and establish the efficiency bounds for RAL estimators under $\mathcal{M}_{\mathsf{comp}}$ by deriving the associated efficient influence function in the following proposition. We denote the nuisance parameters for $f_{Y|\mathbf{S},\mathbf{X}}$, $f_{\mathbf{X}}$ and $f_{\mathbf{S}|\mathbf{X}}$ as $\boldsymbol{\eta}$. We use $\boldsymbol{\eta}_0$ to denote the true underlying nuisance parameter that generates the data. The parameter of interest $\boldsymbol{\beta}_0$ is not part of the model $\mathcal{M}_{\mathsf{comp}}$ but defined by the implicit function through the moment condition (3).

Proposition 10 *The efficient influence function for $\theta = \mathbf{x}_{\text{std}}^{\intercal}\boldsymbol{\beta}$ under $\mathcal{M}_{\mathsf{comp}}$ is*

$$\phi_{\mathsf{eff}}(\mathbf{D}_i; \theta_0, \boldsymbol{\eta}_0) = \frac{R_i}{\rho}\mathbf{u}_0^{\intercal}\mathbf{X}_i\{Y_i - \mathbb{E}(Y_i \mid \mathbf{S}_i, \mathbf{X}_i)\} - \mathbf{u}_0^{\intercal}\mathbf{X}_i\{\mathbb{E}(Y_i \mid \mathbf{S}_i, \mathbf{X}_i) - g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)\}.$$

Under the Assumptions of Theorem 7 and additionally $\mathbb{E}(Y_i \mid \mathbf{S}_i, \mathbf{X}_i) = g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)$, our SAS debiased estimator admits the same influence function

$$\widehat{\mathbf{x}_{\text{std}}^\mathsf{T}\boldsymbol{\beta}} - \frac{\mathbf{x}_{\text{new}}^\mathsf{T}\boldsymbol{\beta}_0}{\|\mathbf{x}_{\text{new}}\|_2} = \frac{1}{N}\sum_{i=1}^N \phi_{\text{eff}}(\mathbf{D}_i;\theta_0,\boldsymbol{\eta}_0) + o_p\left((\rho N)^{-1/2}\right)$$

according to Appendix B3 Step 2 (A.31).

5. Simulation

We have conducted extensive simulation studies to evaluate the finite sample performance of the SAS estimation and inference procedures under various scenarios. Throughout, we let $p = 500$, $q = 100$, $N = 20000$ and consider $n = 500$. The signals in $\boldsymbol{\beta}$ are varied to be approximately sparse or fully dense with a mixture of strong and weak signals. The surrogates $\mathbf{S}$ are either moderately and strongly predictive of Y as specified below. For each configuration, we summarize the results based on 500 simulated datasets. We compare our SAS procedure with the supervised LASSO (SLASSO) that (1) estimates the $\boldsymbol{\beta}_0$ by regressing Y to $\mathbf{X}$ over the labeled data with Lasso; (2) draw inference on $\mathbf{x}_{\text{new}}^\mathsf{T}\boldsymbol{\beta}_0$ with the one-step debiased Lasso van de Geer et al. (2014).

To mimic the zero-inflated discrete distribution of EHR features, we first generate $Z^x_{i,1}, \ldots, Z^x_{i,p}, Z^u_i, Z^s_{i,1}, \ldots, Z^s_{i,q}$ independently from $N(0, 25)$. Then we construct $\mathbf{X}_i$ from $\{Z^u_i, \mathbf{Z}^x_i = (Z^x_{i,1}, ..., Z^x_{i,1})^\mathsf{T}\}$ via the transformation $\varsigma(z) = \lfloor \log\{1 + \exp(z)\}\rfloor$:

$$X_{i,1} = \left\{\varsigma\left(\textstyle\sum_{j=2}^p 2X_{i,j}/\sqrt{p-1} + Z^x_{i,1}/\sqrt{2}\right) - \mu_{\mathbf{X}}\right\}/\sigma_{\mathbf{X}},$$
$$X_{i,j} = [\varsigma(Z^x_{i,j}\sqrt{1-p^{-1}} + Z^u_i/\sqrt{p}) - \mu_{\mathbf{X}}]/\sigma_{\mathbf{X}},\ j = 2, \ldots, p.$$

We standardize $X_{i,j}$ to roughly mean zero and unit variance with $\mu_{\mathbf{X}} = 1.80$ and $\sigma_{\mathbf{X}} = 2.74$. The shared term Z^u_i induces correlation among the covariates.

For $\mathbf{S}$ and Y, we consider two scenarios under which the imputation model is either correctly or incorrectly specified. We present the "Scenario I: neither the risk prediction model nor the imputation model is correctly specified" in the main text and the "Scenario II: The imputation model is correctly specified and exactly sparse" in Section A of the Supplementary materials.

Scenario I: neither the risk prediction model nor the imputation model is correctly specified. In this scenario, we first generate Y_i from the probit model

$$\mathbb{P}(Y_i = 1|\mathbf{Z}^x_i) = \Phi(\boldsymbol{\alpha}^\mathsf{T}\mathbf{Z}^x_i) \quad \text{with} \quad \Phi(x) = \int_{-\infty}^x (2\pi)^{-1/2}e^{-x^2/2}dx,$$

and then generate $\mathbf{S}$ from

$$S_{i,1} = \left\{\varsigma(Z^s_{i,1}/2 + \theta Y_i) - \mu_{\mathbf{S}}\right\}\sigma_{\mathbf{S}}^{-1} + \boldsymbol{\xi}^\mathsf{T}\mathbf{X}_i, \quad \text{and} \quad S_{i,j} = \{\varsigma(Z^s_{i,j}) - \mu_{\mathbf{X}}\}\sigma_{\mathbf{X}}^{-1},\ j = 2, \ldots, p.$$

We chose $\mu_{\mathbf{S}}$ and $\sigma_{\mathbf{S}}$ depending on $\boldsymbol{\alpha}$ such that $S_{i,1}$ is roughly mean 0 and variance 1. Under this setting, a logistic imputation model would be misspecified but nevertheless approximately sparse with appropriately chosen $\boldsymbol{\xi}$. The coefficients $\boldsymbol{\alpha}$ control the optimal prediction accuracy of $\mathbf{X}$ for Y while θ controls the optimal prediction accuracy of $\mathbf{S}$ for Y. We consider two $\boldsymbol{\alpha}$ of different sparsity patterns, which also determine the rest of parameters

$$\begin{aligned}
&\text{Sparse } (s_{\boldsymbol{\alpha}} = 3): && \boldsymbol{\alpha} = (0.45, 0.318, 0.318, \mathbf{0}^{\intercal}_{497\times 1})^{\intercal},\ \mu_{\mathbf{S}} = 1.82,\ \sigma_{\mathbf{S}} = 2.01,\\
&\text{Dense } (s_{\boldsymbol{\alpha}} = 500): && \boldsymbol{\alpha} = (0.316, \mathbf{0.059}^{\intercal}_{29\times 1}, \mathbf{0.007}^{\intercal}_{470\times 1})^{\intercal},\ \mu_{\mathbf{S}} = 2.71,\ \sigma_{\mathbf{S}} = 2.68,
\end{aligned}$$

where $\mathbf{a}_{k\times 1} = (a, ..., a)^{\intercal}_{k\times 1}$ for any a. The sparsity of $\boldsymbol{\alpha}$ affects the approximate sparsity of $\boldsymbol{\beta}$ subsequently (Table 1), which we measured by the squared ratio between ℓ_1 norm and ℓ_2 norm

$$\mathcal{S}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1^2/\|\boldsymbol{\beta}\|_2^2,\ \min_{j:\beta_j\neq 0} |\beta_j| \leq \mathcal{S}(\boldsymbol{\beta})/\|\boldsymbol{\beta}\|_0 \leq 1. \tag{33}$$

We consider two θ: (a) $\theta = 0.6$ for $\mathbf{S}$ to be moderately predictive of Y; and (b) $\theta = 1$ for strong surrogates. The parameter $\boldsymbol{\xi}$ depends on both the choices of $\boldsymbol{\alpha}$ and θ:

$$\begin{aligned}
&s_{\boldsymbol{\alpha}} = 3,\ \theta = 0.6: && \boldsymbol{\xi} = (0.407, 0.330, 0.330, \mathbf{0.005}^{\intercal}_{497\times 1})^{\intercal},\\
&s_{\boldsymbol{\alpha}} = 3,\ \theta = 1: && \boldsymbol{\xi} = (0.199, 0.163, 0.163, \mathbf{0.002}^{\intercal}_{497\times 1})^{\intercal},\\
&s_{\boldsymbol{\alpha}} = 500,\ \theta = 0.6: && \boldsymbol{\xi} = (0.350, \mathbf{0.064}^{\intercal}_{29\times 1}, \mathbf{0.011}^{\intercal}_{470\times 1})^{\intercal},\\
&s_{\boldsymbol{\alpha}} = 500,\ \theta = 1: && \boldsymbol{\xi} = (0.169, \mathbf{0.032}^{\intercal}_{29\times 1}, \mathbf{0.005}^{\intercal}_{470\times 1})^{\intercal}.
\end{aligned}$$

Due to the complexity of the data generating process and the noncollapsibility of the logistic regression models, we cannot analytically express the true $\boldsymbol{\beta}_0$ in both scenarios. Instead, we numerically evaluate $\boldsymbol{\beta}_0$ with a large simulated data using the oracle knowledge of the ex-changeability among covariates according to the model

$$\operatorname{logit}\{\mathbb{P}(Y_i = 1|S_{i,1})\} \sim \eta_0 + \eta_1 X_{i,1} + \eta_2 \sum_{j=2}^{s_\alpha} X_{i,j} + \eta_3 \sum_{j=s_\alpha+1}^{p} X_{i,j}.$$

We derive the true $\boldsymbol{\beta}_0$ as

$$\boldsymbol{\beta}_0 = (\eta_0, \eta_1, (\boldsymbol{\eta_2})^{\intercal}_{s_\alpha\times 1}, (\boldsymbol{\eta_3})^{\intercal}_{(p-s_\alpha)\times 1})^{\intercal}.$$

We report the simulation settings under Scenario I in Table 1, where we present the predictive power of the oracle estimation and the lasso estimation. We also report the average area-under-curve (AUC) of the receiver operating characteristic (ROC) curve for oracle $\boldsymbol{\beta}_0$, SLASSO and the proposed SAS estimation. Our SAS estimation achieves a better AUC compared to supervised LASSO across all scenarios, and is comparable to the AUC with the true coefficient $\boldsymbol{\beta}_0$. Besides, we observe that the AUC of supervised LASSO is sensitive to the approximate sparsity $\mathcal{S}(\boldsymbol{\beta}_0)$, while the AUC of SAS estimation does not seem to be affected by $\mathcal{S}(\boldsymbol{\beta}_0)$.

To evaluate the SAS inference for the individualized prediction, we consider six different choices of $\mathbf{x}_{\text{new}}$. We first select $\{\mathbf{x}^{\mathsf{L}}_{\text{new}}, \mathbf{x}^{\mathsf{M}}_{\text{new}}, \mathbf{x}^{\mathsf{H}}_{\text{new}}\}$ from a random sample of $\mathbf{x}_{\text{new}}$ generated from the distribution of $\mathbf{X}_i$ such that their predicted risks are around 0.2, 0.5, and 0.7, corresponding to low, moderate and high risk. We additionally consider three sets of $\mathbf{x}_{\text{new}}$

Table 1: AUC Table for simulations with 500 labels under Scenario I. The AUCs are evaluated on an independent testing set of size 100. We approximately measure the sparsity by $\mathcal{S}(\mathbf{v}) = \|\mathbf{v}\|_1^2/\|\mathbf{v}\|_2^2$.

Scenario			Prediction Accuracy (AUC)		
Surrogate	$\mathcal{S}(\boldsymbol{\beta}_0)$	$\mathcal{S}(\boldsymbol{\gamma}_0)$	Oracle	SLASSO	SAS
Strong	174	1.32	0.724	0.660	0.711
Moderate	174	1.26	0.724	0.660	0.713
Strong	28.3	1.33	0.719	0.694	0.713
Moderate	28.3	1.24	0.719	0.694	0.711

Table 2: Comparison of SAS Estimation to the supervised LASSO (SLASSO) with Bias, Empirical standard error (ESE) and root mean-squared error (rMSE) of the linear predictions $\mathbf{x}_{\rm new}^{\sf T}\boldsymbol{\beta}_0$ under Scenario I 500 labels, moderate or large $\mathcal{S}(\boldsymbol{\beta}_0)$ and strong or moderate surrogates.

	SLASSO			SAS: Moderate			SAS: Strong		
Type	Bias	ESE	rMSE	Bias	ESE	rMSE	Bias	ESE	rMSE
				Moderate $\mathcal{S}(\boldsymbol{\beta}_0)$					
$\mathbf{x}_{\rm new}^{\sf L}$	0.605	0.387	0.719	0.165	0.249	0.298	0.118	0.196	0.229
$\mathbf{x}_{\rm new}^{\sf M}$	-0.083	0.337	0.347	-0.008	0.246	0.246	-0.016	0.195	0.196
$\mathbf{x}_{\rm new}^{\sf H}$	-0.718	0.521	0.887	-0.234	0.294	0.376	-0.176	0.225	0.286
$\mathbf{x}_{\rm new}^{\sf S}$	-0.072	0.144	0.161	-0.080	0.094	0.123	-0.018	0.078	0.080
$\mathbf{x}_{\rm new}^{\sf I}$	-0.460	0.096	0.470	-0.110	0.093	0.143	-0.055	0.071	0.090
$\mathbf{x}_{\rm new}^{\sf D}$	-0.413	0.091	0.423	-0.110	0.089	0.141	-0.114	0.069	0.133
				Large $\mathcal{S}(\boldsymbol{\beta}_0)$					
$\mathbf{x}_{\rm new}^{\sf L}$	0.389	0.275	0.477	0.161	0.215	0.269	0.133	0.264	0.296
$\mathbf{x}_{\rm new}^{\sf M}$	-0.017	0.280	0.280	-0.014	0.213	0.213	-0.017	0.268	0.268
$\mathbf{x}_{\rm new}^{\sf H}$	-0.600	0.481	0.769	-0.251	0.271	0.370	-0.164	0.296	0.339
$\mathbf{x}_{\rm new}^{\sf S}$	-0.202	0.140	0.246	-0.074	0.097	0.122	-0.009	0.078	0.079
$\mathbf{x}_{\rm new}^{\sf I}$	-0.178	0.098	0.203	-0.075	0.086	0.115	-0.071	0.075	0.103
$\mathbf{x}_{\rm new}^{\sf D}$	-0.185	0.090	0.206	-0.109	0.084	0.138	-0.113	0.073	0.135

with different levels of sparsity:

$$\begin{aligned}
&\text{Sparse:} && \mathbf{x}_{\rm new}^{\sf S} = (1, 1, \mathbf{0}_{499\times 1}^{\sf T})^{\sf T};\\
&\text{Intermediate:} && \mathbf{x}_{\rm new}^{\sf I} = (1, \mathbf{0.183}_{30\times 1}^{\sf T}, \mathbf{0}_{470\times 1}^{\sf T})^{\sf T};\\
&\text{Dense:} && \mathbf{x}_{\rm new}^{\sf D} = (1, \mathbf{0.045}_{500\times 1}^{\sf T})^{\sf T}.
\end{aligned}$$

In Table 2, we compare our SAS estimator of $\mathbf{x}_{\rm new}^{\sf T}\boldsymbol{\beta}_0$ with the corresponding SLASSO across all settings under Scenario I. The root mean-squared-error (rMSE) of the SAS estimation decays proportionally with the sample size, while the rMSE of the supervised LASSO provides evidence of inconsistency for moderate and dense deterministic $\mathbf{x}_{\rm new}$. The bias of the supervised LASSO is also significantly larger than that of the SAS estimation. The performance of the SAS estimation is insensitive to sparsity of $\boldsymbol{\beta}_0$, while that of supervised LASSO severely deteriorate with dense $\boldsymbol{\beta}_0$. The improvement from the supervised LASSO to the SAS estimation is regulated by the surrogate strength.

In Table 3, we compare our SAS inference with supervised debiased LASSO across the settings under Scenario I. Our SAS inference procedure attains approximately honest

Table 3: Bias, Empirical standard error (ESE), average of the estimated standard error (ASE) along with empirical coverage of the 95% confidence intervals (CP) for the debiased supervised LASSO (SLASSO) and debiased SAS estimator of linear predictions $\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta}_0$ under Scenario I with 500 labels, moderate or large $\mathcal{S}(\boldsymbol{\beta}_0)$ and strong or moderate surrogates.

	Debiased SLASSO				Debiased SAS Moderate Surrogates				Debiased SAS Strong Surrogates			
Type	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
Risk prediction model approximatedly sparse												
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	-0.290	1.901	1.896	0.948	0.021	1.873	1.864	0.949	0.018	1.531	1.531	0.950
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.091	1.994	1.981	0.947	-0.007	1.961	1.954	0.950	-0.015	1.560	1.570	0.953
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	0.348	2.106	2.074	0.942	-0.050	2.036	2.039	0.950	-0.011	1.632	1.623	0.950
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	0.171	0.157	0.128	0.694	-0.019	0.149	0.150	0.950	-0.001	0.132	0.125	0.924
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.001	0.129	0.125	0.938	-0.013	0.123	0.116	0.932	0.010	0.101	0.094	0.920
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	0.141	0.137	0.138	0.812	-0.011	0.123	0.118	0.944	-0.001	0.096	0.095	0.940
Large $\mathcal{S}(\boldsymbol{\beta}_0)$												
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	-0.134	1.918	1.914	0.951	0.018	1.875	1.878	0.951	0.018	1.529	1.524	0.948
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.056	1.970	1.962	0.948	-0.020	1.911	1.927	0.952	0.005	1.603	1.597	0.950
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	0.109	2.051	2.029	0.945	-0.022	1.997	1.991	0.950	-0.040	1.671	1.668	0.951
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	0.029	0.155	0.127	0.892	-0.008	0.153	0.147	0.946	-0.013	0.133	0.131	0.938
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	0.002	0.131	0.125	0.930	0.001	0.122	0.114	0.936	0.002	0.101	0.098	0.936
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	0.113	0.135	0.139	0.874	-0.007	0.119	0.116	0.938	-0.003	0.099	0.097	0.960

coverage of 95 % confidence intervals for all types of $\mathbf{x}_{\text{new}}$ under all scenarios. Unsurprisingly, the debiased SLASSO has under coverage for the deterministic $\mathbf{x}_{\text{new}}$ as the consequence of violation to the sparsity assumption for $\boldsymbol{\beta}_0$ and precision matrix. Under our design, the first covariate X_1 has the strongest dependence upon the other covariates, whose associated row in the precision matrix is thus densest. Consequently, the inference for $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{\text{new}}^{\mathsf{S}} = \beta_0 + \beta_1$ The debiased SLASSO also has an acceptable coverage for random $\mathbf{x}_{\text{new}}^{\mathsf{L}}$, $\mathbf{x}_{\text{new}}^{\mathsf{M}}$, $\mathbf{x}_{\text{new}}^{\mathsf{H}}$ sampled from the covariate distribution despite the presence of substantial bias, which we attribute to the even larger variance that dominates the bias. In contrast, our SAS inference has small bias across all scenarios and improved variance from the strong surrogate.

According to Tables A1, A2 and A3 in the Appendix A, the results under Scenario II are consistent with our findings under Scenario I. We also compares SAS to an unsupervised learning approach using proxy outcome derived from surrogates in the Appendix A. Under the Scenario III very similar to Scenario I, SAS performs well as in Scenario I while the unsupervised learning approach fails completely. This is expected since the unsupervised approach requires that the deviation of the surrogates from the true outcome $S - Y$ is uncorrelated with the risk factors $\mathbf{X}$. Otherwise, spurious association between outcome Y and risk factors $\mathbf{X}$ can be induced, creating bias in estimation of risk prediction model.

6. Application of SAS to EHR Study

We applied the proposed SAS method to the risk prediction of Type II Diabetes Mellitus (T2DM) using EHR and genomic data of participants of the Mass General Brigham Biobank study. Number of genetic risk factors among single nucleotide polymorphism for T2DM has grown exponentially following the expansion of genome-wide association studies. As an incomplete summary, Voight et al. (2010),Morris et al. (2012) and Scott et al. (2017) each discovered around a dozen new risk SNPs for T2DM, and the recent studies by Mahajan et al. (2018) and Vujkovic et al. (2020) discovered 135 and 558 new risk SNPs, respectively. Some new risk SNPs in Mahajan et al. (2018) even had large coefficients in the poly genetic risk score. The ever growing number of risk SNPs suggest that the genetic risk prediction model for T2DM may be dense. Compared to the large biobank data that generated the genome-wide association studies, EHR captures the temporal information of T2DM onset and other phenotypes predictive for T2DM and thus may provide a more accurate forecasting for T2DM. As we mentioned in the introduction, direct extraction of disease onset from EHR by diagnosis code or mention in medical notes may contain substantial false positives. From an expert annotation of the medical histories for 271 patients, we found 38 patients with T2DM diagnosis code and 161 patients with mention of T2DM in medical notes who actually had never developed T2DM. The annotation process requires intensive labor of highly skilled medical experts, leading to the limited number of labels.

To define the study cohort, we extracted from the EHR of each patient their date of first EHR encounter (t_{ini}), follow up period (C), the counts and dates for the diagnosis codes and note mentions of clinical concepts related to T2DM as well as its risk factors. We only included patients who do not have any diagnosis code or note mention of T2DM up to baseline, where the baseline time is defined as 1990 if t_{ini} is prior to 1990 and as their first year if $t_{ini} \geq 1990$. Although neither the diagnosis code nor note mention of T2DM is sufficiently specific, they are highly sensitive and can be used to accurately remove patients who have already developed T2DM at baseline. This exclusion criterion resulted in $N = 20216$ patients who are free of T2DM at baseline and have both EHR and genomics features for risk modeling. Among those, we have a total of $n = 271$ patients whose T2DM status during follow up, Y, has been obtained via manual chart review. The prevalence of T2DM was about 14% based on labeled data.

We aim to develop a risk prediction model for Y by fitting a working model $P(Y = 1 \mid \mathbf{X}) = g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X})$, where the baseline covariate vector $\mathbf{X}$ includes age, gender, indicator for occurrence of diagnosis code and note counts for obesity, hypertension, coronary artery disease (CAD), hyperlipidemia during the first year window, as well as a total of 49 single nucleotide polymorphism previously reported as associated with T2DM in Mahajan et al. (2018) with odds ratio greater than 1.1. We additionally adjust for follow up by including $\log(C)$ and allow for non-linear effects by including two-way interactions between the SNPs and other baseline covariates. All variables with less than 10 nonzero values within the labelled set are removed, resulting the final covariates to be of dimension $p = 260$. We standardize the covariates to have mean 0 and variance 1. To impute the outcome, we used the predicted probability of T2DM derived from the unsupervised phenotyping method MAP (Liao et al., 2019), which achieves an AUC of 0.98, indicating a strong surrogate. In addition to the proposed SAS procedure, we derive risk prediction models based on the supervised LASSO with both the same set of covariates. We let $K = 5$ in cross-fitting and use 5-fold cross-validation for tuning parameter selection. To compare the performance of different risk prediction models, we use 10-fold cross-validation to estimate the out-of-sample AUC. We repeated the process 10 times and took average of predicted probabilities across the repeats for each labelled sample and method in comparison.

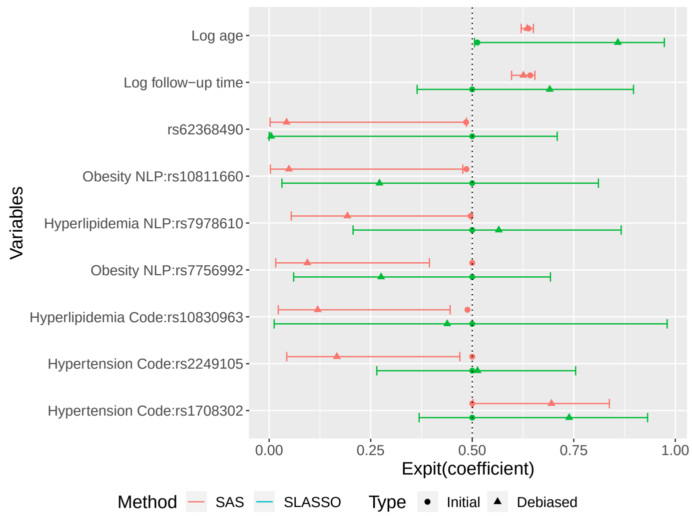

Figure 2: Point and 95% confidence interval estimates for the coefficients with nominal p-value < 0.05 from SAS inference. The horizontal bars indicate the estimated 95% confidence intervals. The solid points indicate the (initial) estimates, and the triangles indicate debiased estimates. Colors red and green indicate different methods, SAS and SLASSO, respectively.

Table 4: The cross-validated (CV) AUC the estimated risk prediction models with high dimensional EHR and genetic features based on SAS and supervised LASSO. Shown also are the AUC of the imputation model derived for the SAS procedure.

Method	Imputation	SAS	SLASSO
CV AUC	0.928	0.763	0.488

In Figure 2, we present the estimated $\boldsymbol{\beta}$ coefficients for the covariates that received p-value less than 0.05 from the SAS inference. The confidence intervals are generally narrower from the SAS inference. For the coefficients of baseline age and follow-up time, the SAS inference produced much narrower confidence interval than debiased SLASSO, which are expected to have a positive effect on the T2DM onset status during the observation. In addition, the SAS inference identified one global genetic risk factor and 6 other subgroup genetic risk factors while SLASSO identified none of these.

In Table 4, we present the AUCs of the estimated risk prediction models using the high dimensional $\mathbf{X}$. It is important to note that AUC is a measurement of prediction accuracy, so debiasing might lead to worse AUC by accepting larger variability for reduced bias. The AUC from SLASSO is very poor, probably due to the over-fitting bias with the small sample sizes of the labeled set. With the information from a large unlabeled data, SAS produced the significantly higher AUC than the SLASSO.

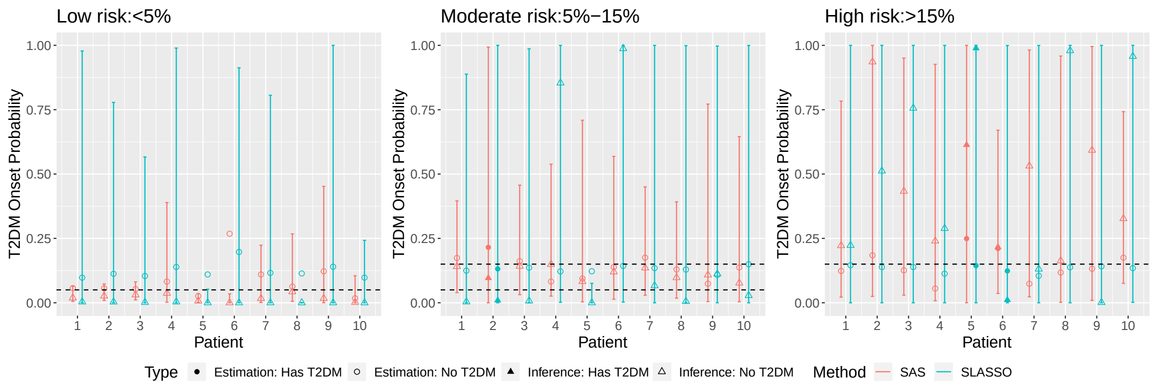

Figure 3: Point and 95% confidence interval estimates for the predicted risks of 30 randomly selected patients. The vertical bars indicate the estimated 95% confidence intervals. The circle and the triangle shapes correspond to (initial) estimation and debiased estimation, correspondingly. Solid points indicate the observed T2DM cases. Colors red and green indicate different methods, SAS and SLASSO.

For illustration, we present in Figure 3 the individual risk predictions with 95% confidence intervals for three sets of 10 patients with each set randomly selected from low ($< 5\%$), medium ($5\% \sim 15\%$) or high risk ($> 15\%$) subgroups. These risk groups are constructed for illustration purposes and a patient with $\mathbf{x}_{\text{new}}$ classified to low, medium and high risk if $\text{expit}(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}_{\text{new}})$ belongs to the low, medium and high tertiles of $\{\text{expit}(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i), i = 1, ..., N\}$. We observe that the confidence intervals for patients with predicted The debiased SLASSO inference is not very informative with most error bars stretching from zero to one. The contrast between SAS CIs and SLASSO CIs demonstrates the improved efficiency as the result of leveraging information from the unlabeled data through predictive surrogates.

7. Discussion

We proposed the SAS estimation and inference method for high-dimensional risk prediction model with diminishing proportion of observed outcomes. With a sparse imputation model based on predictive surrogates, the SAS can recover a dense risk prediction model impossible to learn from supervised method, as well as achieve better efficiency than supervised method when the latter is applicable. We show that the theoretical advantages lead to better prediction accuracy and shorter confidence intervals in simulations and real data example.

While the SAS procedure is a powerful tool with minimal requirements, caution should be given to the inclusion of highly informative surrogates so that the imputation model is sparse (or approximately sparse). If all surrogates poorly predicts Y with a dense imputation model, the SAS procedure can lead to a compromised convergence rate in estimation. While the current study is motivated by the existence of easy-to-learn imputation model with highly predictive surrogates, the SAS framework can be extended to settings where the imputation model is not easier to learn than the model for $Y \mid \mathbf{X}$. When the imputation model is estimable but more dense than the risk prediction model (i.e. $s_{\boldsymbol{\beta}} < s_{\boldsymbol{\gamma}}$), we can following similar strategies as in our SAS inference procedure to reduce the bias incurred during the imputation step from $\widehat{\boldsymbol{\gamma}}$. Specifically, we may consider a debiased estimator for $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}}_{debias} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{k=1}^{K} \left[\frac{1}{N} \sum_{i\in\mathcal{J}_k} \ell(g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i), \boldsymbol{\beta}^\intercal\mathbf{X}_i) + \frac{1}{n} \sum_{i\in\mathcal{I}_k} \boldsymbol{\beta}^\intercal\mathbf{X}_i\{g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i) - Y_i\} \right] + \lambda\|\boldsymbol{\beta}_{-1}\|_1.$$

This debiased SAS estimation will attain the optimal rate $\sqrt{s_{\boldsymbol{\beta}}\log(p)/n}$ and we also expect an efficiency gain in the resulting variance compared to the supervised estimator, in analog to the efficiency gain observed in SAS inference. Adaptive approaches to infer whether a given dataset falls into the setting with $s_{\boldsymbol{\beta}} > s_{\boldsymbol{\gamma}}$ or $s_{\boldsymbol{\beta}} < s_{\boldsymbol{\gamma}}$ is straightforward in simpler settings when $s_{\boldsymbol{\beta}}$ and $s_{\boldsymbol{\gamma}}$ can be estimated but warrants future research in general. In the extremely dense imputation model setting when $s_{\boldsymbol{\gamma}} > n$, information theoretical bound has indicated that the imputation model will be inestimable, invalidating any subsequent steps involving $\widehat{\boldsymbol{\gamma}}$. A possible solution is to redefine the imputation model as the sparser model between the risk prediction model and the original imputation model. A potential approach to identifying such a sparser imputation model is through the under-identified Dantzig Selector

$$\begin{aligned} \widehat{\boldsymbol{\gamma}}_{ada} &= \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p+q}}{\operatorname{argmin}} \|\boldsymbol{\gamma}\|_1, \\ \text{Subject to } & \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\{Y_i - g(\boldsymbol{\gamma}^\intercal\mathbf{W}_i)\} \right\|_\infty \le \lambda. \end{aligned}$$

Both $\boldsymbol{\gamma}_0$ and $(\boldsymbol{\beta}_0^\intercal, \mathbf{0}_q^\intercal)^\intercal$ should fall in the feasible region with suitable λ, and the minimization over L_1 norm may pick the sparsest element from the feasible class. Using $\widehat{\boldsymbol{\gamma}}_{ada}$ in SAS estimation may attain uniform optimal rate for any $s_{\boldsymbol{\beta}}$ and $s_{\boldsymbol{\gamma}}$. Theoretical studies of the above proposals warrant future research.

Supplementary Material

We present the simulation Scenario II in which the imputation model is correctly specified and exactly sparse in Appendix A. We also compared the SAS estimation and inference with unsupervised regression that uses only $\mathbf{S}$ to derive a proxy outcome in the simulation Scenario III very similar to Scenario I. The proofs of Theorems 1, 7, Corollary 5 and Propositions 9 and 10 are given in Appendix B. The technical details are put in Appendix C. Definitions and existing results are stated in Appendix D.

Appendix A. Additional Simulation

Scenario II: The imputation model is correctly specified and exactly sparse. In the second scenario, we first generate $\mathbf{S}_i$ from

$$S_{i,1} = \left[\varsigma\{\nu Z_{i,1}^s + \boldsymbol{\alpha}^\intercal(\mathbf{Z}_i^x\sqrt{1-p^{-1}} + Z_i^u/\sqrt{p})\} - \mu_{\mathbf{S}}\right]/\sigma_{\mathbf{S}}.$$

and $S_{i,j} = \{\varsigma(Z_{i,j}^s) - \mu_{\mathbf{X}}\}/\sigma_{\mathbf{X}}$ for $j = 2, \dots, p$, and then generate Y_i from a sparse model

$$\mathbb{P}(Y_i = 1|\mathbf{X}_i) = \text{expit}(\theta S_{i,1}).$$

We chose $\mu_{\mathbf{S}} \approx 0.66$ and $\sigma_{\mathbf{S}} \approx 1$ such that $S_{i,1}$ is roughly mean 0 and variance 1. Under this setting, the imputation model holds with $s_{\boldsymbol{\gamma}} = 1$. The factor ν and the coefficients $\boldsymbol{\alpha}$ control the predictiveness of $\mathbf{X}$ for S_1 and Y while θ controls the predictiveness of $\mathbf{W}$ for Y. We consider two $\boldsymbol{\alpha}$ of different sparsity patterns,

$$\begin{aligned} &\text{Sparse } (s_{\boldsymbol{\alpha}} = 3): &&\boldsymbol{\alpha} = (0.3, 0.212, 0.212, \mathbf{0}_{497\times 1}^\intercal)^\intercal \\ &\text{Dense } (s_{\boldsymbol{\alpha}} = 500): &&\boldsymbol{\alpha} = (0.211, \mathbf{0.039}_{29\times 1}^\intercal, \mathbf{0.004}_{470\times 1}^\intercal)^\intercal, \end{aligned}$$

where $\mathbf{a}_{k\times 1} = (a, ..., a)_{k\times 1}^\intercal$ for any a. Similar to Scenario I, the sparsity of $\boldsymbol{\alpha}$ regulates the approximate sparsity of $\boldsymbol{\beta}$ measured by (33) (See Table 1). We consider two sets of (ν, θ) to allow $\mathbf{W}$ to be either moderately or strongly predictive of Y:

$$\text{Moderate: } \nu = 0.4, \theta = 2; \qquad \text{and} \qquad \text{Strong: } \nu = 0.6, \theta = 3.7.$$

The layouts of Tables A2 and A3 are different from those of 2 and Table 3 because of the different data generating mechanism. The distribution of $Y_i \mid \mathbf{X}_i$ is not affected by the distribution of $\mathbf{S}_i$ in Scenario I, while the property does not hold in Scenario II.

Table A1: AUC Table for simulations with 500 labels under Scenario II. The AUCs are evaluated on an independent testing set of size 100. We approximately measure the sparsity by $\mathcal{S}(\mathbf{v}) = \|\mathbf{v}\|_1^2/\|\mathbf{v}\|_2^2$.

Scenario			Prediction Accuracy (AUC)		
Surrogate	$\mathcal{S}(\boldsymbol{\beta}_0)$	$\mathcal{S}(\boldsymbol{\gamma}_0)$	Oracle	SLASSO	SAS
Strong	159	1.10	0.715	0.660	0.702
Moderate	128	1.06	0.715	0.665	0.704
Strong	26.4	1.09	0.710	0.691	0.708
Moderate	18.4	1.03	0.709	0.693	0.707

Table A2: Comparison of SAS Estimation to the supervised LASSO (SLASSO) with Bias, Empirical standard error (ESE) and root mean-squared error (rMSE) of the linear predictions $\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta}_0$ under Scenario II with 500 labels, moderate or large $\mathcal{S}(\boldsymbol{\beta}_0)$ and strong or moderate surrogates.

	Moderate Surrogates						Strong Surrogates					
	SLASSO			SAS			SLASSO			SAS		
Type	Bias	ESE	rMSE	Bias	ESE	rMSE	Bias	ESE	rMSE	Bias	ESE	rMSE
Risk prediction model approximatedly sparse												
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.505	0.378	0.631	0.163	0.283	0.327	0.349	0.278	0.446	0.085	0.222	0.238
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.140	0.331	0.359	-0.047	0.272	0.276	-0.113	0.282	0.304	-0.058	0.217	0.225
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-0.713	0.512	0.878	-0.262	0.313	0.408	-0.678	0.469	0.825	-0.210	0.241	0.320
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-0.111	0.143	0.181	-0.058	0.081	0.100	-0.190	0.142	0.237	-0.037	0.063	0.072
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.437	0.098	0.448	-0.119	0.076	0.141	-0.155	0.098	0.183	-0.065	0.061	0.089
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.349	0.093	0.361	-0.138	0.078	0.158	-0.150	0.093	0.176	-0.112	0.063	0.129
Large $\mathcal{S}(\boldsymbol{\beta}_0)$												
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.366	0.266	0.453	0.142	0.224	0.265	0.482	0.398	0.625	0.117	0.300	0.322
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.060	0.275	0.282	-0.035	0.213	0.216	-0.199	0.337	0.391	-0.082	0.299	0.310
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-0.656	0.475	0.810	-0.272	0.257	0.374	-0.749	0.503	0.903	-0.214	0.325	0.389
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-0.236	0.139	0.274	-0.087	0.079	0.117	-0.054	0.138	0.148	0.003	0.063	0.063
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.173	0.096	0.197	-0.078	0.077	0.109	-0.409	0.097	0.420	-0.094	0.057	0.110
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.144	0.092	0.171	-0.101	0.080	0.129	-0.359	0.093	0.371	-0.154	0.060	0.166

Scenario III: Similar to Scenario I. In the third scenario, we repeat the data generation process of Scenario I except for difference values for $\boldsymbol{\xi}$:

$$
\begin{aligned}
s_{\boldsymbol{\alpha}} = 3,\, \theta = 0.6: &\quad \boldsymbol{\xi} = (-0.593, 0.330, 0.330, \mathbf{0.005}_{497\times 1}^{\mathsf{T}})^{\mathsf{T}},\\
s_{\boldsymbol{\alpha}} = 3,\, \theta = 1: &\quad \boldsymbol{\xi} = (-0.801, 0.163, 0.163, \mathbf{0.002}_{497\times 1}^{\mathsf{T}})^{\mathsf{T}},\\
s_{\boldsymbol{\alpha}} = 500,\, \theta = 0.6: &\quad \boldsymbol{\xi} = (-0.650, \mathbf{0.064}_{29\times 1}^{\mathsf{T}}, \mathbf{0.011}_{470\times 1}^{\mathsf{T}})^{\mathsf{T}},\\
s_{\boldsymbol{\alpha}} = 500,\, \theta = 1: &\quad \boldsymbol{\xi} = (-0.831, \mathbf{0.032}_{29\times 1}^{\mathsf{T}}, \mathbf{0.005}_{470\times 1}^{\mathsf{T}})^{\mathsf{T}}.
\end{aligned}
$$

We focus on the comparison between SAS estimation and inference and an unsupervised learning (UL) approach. For the UL approach, a proxy outcome $\widetilde{Y}$ is derived directly from the dichotomized informative surrogate S_1

$$
\widetilde{Y} = \mathbb{1}(S_1 \geq s_*),
$$

where the threshold s_* is chose in order to match the prevalence $\mathbb{E}(\widetilde{Y}) \approx \mathbb{E}(Y)$. Then, the UL estimation of $\boldsymbol{\beta}$ is obtained by regression $\widetilde{Y}$ to $\mathbf{X}$ under the logistic regression model over all N observations. Classical inference is used for construction of UL confidence intervals.

The layouts of Tables A5 and A6 are different from those of 2 and Table 3 because of the different benchmark method. The supervised learning methods SLASSO and Debiased SLASSO are not affected by the distribution of $\mathbf{S}_i$ in Scenario I, while UL considered in Scenario III uses $\mathbf{S}_i$ to construct its proxy outcome. UL obviously failed with ineffective classification (AUC <0.50 in Table A4), large bias in Table A5 and severe under-covering confidence intervals in Table A6. The performance of SAS is solid as in Scenario I.

Table A3: Bias, Empirical standard error (ESE) along with empirical coverage of the 95% confidence intervals (CP) for the debiased supervised LASSO (SLASSO) and debiased SAS estimator of linear predictions $\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta}_0$ under Scenario II with 500 labels, moderate or large $\mathcal{S}(\boldsymbol{\beta}_0)$ and strong or moderate surrogates.

	Debiased SLASSO				Debiased SAS			
Type	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
Risk prediction model approximatedly sparse, moderate surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	-0.236	1.936	1.915	0.947	0.014	1.786	1.771	0.950
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.044	2.031	1.997	0.944	-0.028	1.873	1.853	0.947
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	0.364	2.110	2.084	0.944	-0.045	1.943	1.924	0.947
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	0.133	0.156	0.127	0.784	-0.028	0.133	0.130	0.944
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	0.004	0.124	0.126	0.942	-0.014	0.102	0.100	0.936
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	0.149	0.121	0.139	0.848	-0.014	0.104	0.105	0.948
Risk prediction model approximatedly sparse, strong surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	-0.070	1.953	1.935	0.947	0.021	1.371	1.366	0.949
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.031	2.019	1.986	0.946	-0.026	1.408	1.401	0.949
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	0.148	2.073	2.055	0.948	-0.010	1.458	1.444	0.949
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	0.029	0.153	0.127	0.894	-0.016	0.103	0.096	0.928
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	0.018	0.134	0.126	0.938	-0.004	0.081	0.079	0.944
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	0.134	0.128	0.141	0.842	-0.007	0.081	0.083	0.956
Large $\mathcal{S}(\boldsymbol{\beta}_0)$, moderate surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	-0.092	1.942	1.925	0.950	0.004	1.796	1.792	0.951
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.034	1.995	1.969	0.947	-0.018	1.852	1.835	0.951
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	0.082	2.061	2.036	0.946	-0.027	1.912	1.890	0.948
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-0.009	0.155	0.125	0.876	-0.027	0.131	0.125	0.922
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	0.000	0.126	0.125	0.952	-0.009	0.104	0.103	0.950
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	0.119	0.126	0.139	0.894	-0.012	0.108	0.108	0.940
Large $\mathcal{S}(\boldsymbol{\beta}_0)$, strong surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	-0.221	1.929	1.926	0.949	0.022	1.353	1.349	0.951
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	0.032	2.047	2.017	0.947	-0.003	1.427	1.414	0.950
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	0.442	2.137	2.104	0.940	-0.039	1.479	1.469	0.951
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	0.176	0.150	0.128	0.698	-0.018	0.094	0.099	0.946
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	0.030	0.128	0.129	0.936	-0.002	0.079	0.077	0.952
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	0.167	0.125	0.142	0.804	-0.008	0.082	0.080	0.954

Table A4: AUC Table for simulations with 500 labels under Scenario III. The AUCs are evaluated on an independent testing set of size 100. We approximately measure the sparsity by $S(\mathbf{v}) = \|\mathbf{v}\|_1/\|\mathbf{v}\|_2$.

Scenario			Prediction Accuracy (AUC)		
Surrogate	$\mathcal{S}(\boldsymbol{\beta}_0)$	$\mathcal{S}(\boldsymbol{\gamma}_0)$	Oracle	UL	SAS
Strong	174.08	2.31	0.72	0.43	0.69
Moderate	174.08	2.25	0.72	0.46	0.69
Strong	28.27	2.32	0.72	0.45	0.69
Moderate	28.27	2.24	0.72	0.47	0.69

Table A5: Bias, Empirical standard error (ESE) and mean-squared error (MSE) for the unsupervised learning (UL) and SAS estimator of linear predictions $\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta}_0$ with under Scenario III 500 labels, approximately sparse or dense $\boldsymbol{\beta}_0$ and strong or moderate surrogates.

	UL			SAS		
Type	Bias	ESE	MSE	Bias	ESE	MSE
Moderate $\mathcal{S}(\boldsymbol{\beta}_0)$, moderate surrogates						
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.42	0.92	1.01	0.02	0.42	0.42
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	0.20	2.31	2.32	-0.03	0.51	0.51
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-3.39	3.90	5.16	0.09	0.59	0.60
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.73	0.07	3.73	-0.09	0.14	0.17
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.42	0.14	0.44	-0.19	0.10	0.22
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.03	0.15	0.16	-0.18	0.09	0.20
Large $\mathcal{S}(\boldsymbol{\beta}_0)$,moderate surrogates						
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.61	0.70	0.93	0.04	0.37	0.37
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.04	2.54	2.54	-0.03	0.49	0.49
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-2.53	4.66	5.30	0.07	0.59	0.59
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.56	0.11	3.56	-0.14	0.14	0.19
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.70	0.14	0.72	-0.11	0.09	0.15
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.30	0.13	0.33	-0.12	0.09	0.15
Moderate $\mathcal{S}(\boldsymbol{\beta}_0)$, strong surrogates						
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.89	0.67	1.11	-0.05	0.38	0.38
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.15	2.51	2.51	-0.04	0.54	0.54
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-3.28	4.57	5.62	0.22	0.60	0.64
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.83	0.07	3.83	-0.09	0.13	0.15
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.99	0.12	1.00	-0.08	0.08	0.11
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.57	0.12	0.58	-0.10	0.08	0.13
Large $\mathcal{S}(\boldsymbol{\beta}_0)$, strong surrogates						
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.76	0.82	1.12	-0.09	0.45	0.46
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	0.09	2.35	2.35	-0.03	0.57	0.58
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-4.10	3.97	5.71	0.23	0.63	0.67
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.99	0.09	3.99	-0.05	0.13	0.14
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.82	0.12	0.83	-0.12	0.08	0.15
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.37	0.14	0.40	-0.13	0.08	0.15

Table A6: Bias, Empirical standard error (ESE), average of the estimated standard error (ASE) along with empirical coverage of the 95% confidence intervals (CP) for the unsupervised learning (UL) and debiased SAS estimator of linear predictions $\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\beta}_0$ under Scenario III with 500 labels, approximately sparse or dense $\boldsymbol{\beta}_0$ and strong or moderate surrogates.

	UL				SAS			
Type	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
Moderate $\mathcal{S}(\boldsymbol{\beta}_0)$, moderate surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.42	0.92	0.40	0.53	-0.27	1.95	1.92	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	0.20	2.31	0.42	0.26	-0.04	2.09	2.03	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-3.39	3.90	0.45	0.12	0.43	2.18	2.14	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.73	0.07	0.06	0.00	-0.04	0.16	0.15	0.93
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.42	0.14	0.03	0.00	-0.03	0.13	0.12	0.92
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.03	0.15	0.03	0.28	-0.02	0.12	0.12	0.95
Large $\mathcal{S}(\boldsymbol{\beta}_0)$, moderate surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.61	0.70	0.41	0.56	-0.22	1.98	1.96	0.95
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.04	2.54	0.43	0.29	-0.05	2.06	2.02	0.95
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-2.53	4.66	0.45	0.11	0.47	2.16	2.10	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.56	0.11	0.06	0.00	-0.03	0.16	0.15	0.92
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.70	0.14	0.03	0.00	-0.02	0.13	0.12	0.91
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.30	0.13	0.03	0.02	-0.02	0.13	0.12	0.93
Moderate $\mathcal{S}(\boldsymbol{\beta}_0)$, strong surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.89	0.67	0.41	0.37	-0.22	1.69	1.67	0.95
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	-0.15	2.51	0.42	0.27	-0.04	1.78	1.71	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-3.28	4.57	0.44	0.10	0.48	1.86	1.79	0.93
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.83	0.07	0.07	0.00	-0.02	0.14	0.13	0.93
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.99	0.12	0.03	0.00	-0.00	0.11	0.10	0.92
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.57	0.12	0.03	0.00	-0.01	0.11	0.10	0.93
Large $\mathcal{S}(\boldsymbol{\beta}_0)$, strong surrogates								
$\mathbf{x}_{\text{new}}^{\mathsf{L}}$	0.76	0.82	0.40	0.41	-0.27	1.67	1.65	0.95
$\mathbf{x}_{\text{new}}^{\mathsf{M}}$	0.09	2.35	0.42	0.26	-0.03	1.81	1.73	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{H}}$	-4.10	3.97	0.45	0.10	0.45	1.88	1.83	0.94
$\mathbf{x}_{\text{new}}^{\mathsf{S}}$	-3.99	0.09	0.07	0.00	-0.03	0.14	0.14	0.93
$\mathbf{x}_{\text{new}}^{\mathsf{I}}$	-0.82	0.12	0.03	0.00	-0.01	0.11	0.10	0.93
$\mathbf{x}_{\text{new}}^{\mathsf{D}}$	-0.37	0.14	0.03	0.01	-0.01	0.11	0.11	0.97

Appendix B. Proofs of Main Results

We first summarize below notations used Section 3 for the conditional expectations given different part of the data.

Definition 11 *The conditional expectation for samples with index in set $\mathcal{S}$ conditionally on subset of the data $\mathscr{D}$ is denoted as*

$$\mathbb{E}_{i\in\mathcal{S}}\{f(Y_i,\mathbf{X}_i,\mathbf{S}_i)\mid\mathscr{D}\},\ \mathcal{S}\subseteq\{1,\dots,n+N\},\mathscr{D}\subset\mathscr{L}\cup\mathscr{U}.$$

We denote the conditional expectation of unlabeled data given labelled data by $\mathbb{E}_{i>n}\{f(\mathbf{W}_i)\mid\mathscr{L}\}$ and the conditional probability of new copy of data given current data by $\mathbb{P}_{\text{new}}\{f(\mathbf{W}_i)\mid\mathscr{D}\}$. With $\mathscr{L}$ and $\mathscr{U}$ partitioned into K folds indexed respectively by $\{\mathcal{I}_k,k=1,...,K\}$ and $\{\mathcal{J}_k,k=1,...,K\}$, we denote the conditional expectation of fold-k labelled data and unlabeled data given the out-of-fold data respectively by

$$\begin{aligned}&\mathbb{E}_{i\in\mathcal{I}_k}\{f(Y_i,\mathbf{X}_i,\mathbf{S}_i)\mid\mathscr{D}_k^c\}\quad\textit{and}\quad\mathbb{E}_{i\in\mathcal{J}_k}\{f(\mathbf{W}_i)\mid\mathscr{D}_k^c\},\\ &\textit{where}\quad\mathscr{D}_k^c=\{\mathbf{S}_i,\mathbf{X}_i,i\in\mathcal{J}_k^c\}\cup\{Y_i,\mathbf{S}_i,\mathbf{X}_i,i\in\mathcal{I}_k^c\}.\end{aligned}$$

B1 Proof of Theorem 1

Our proof shares the general steps with the the restricted strong convexity framework laid down in Negahban et al. (2010) while we have a delicate analysis of the symmetrized Bregman divergence to establish the improved rate of estimation under semi-supervised learning setting. To bound $\widehat{\boldsymbol\beta}$ through the symmetrized Bregman divergence $(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)^\intercal\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})$, instead of directly applying the Hölder's bound, we first split it into two parts,

$$\begin{aligned}(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)^\intercal\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})=&\underbrace{(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)^\intercal\left[\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})-\mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})\mid\mathscr{L}\}+\mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\boldsymbol\gamma_0)\mid\mathscr{L}\}\right]}_{\text{variance from unlabeled data}}\\&+\underbrace{(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)^\intercal\mathbb{E}\left\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})-\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\boldsymbol\gamma_0)\mid\mathscr{L}\right\}}_{\text{bias from }\widehat{\boldsymbol\gamma}}\qquad\text{(A.1)}\end{aligned}$$

and discuss which part dominates the estimation error. When the first variance term in (A.1) is dominant, the bias from $\widehat{\boldsymbol\gamma}$ becomes eligible. Then, we should recover the usual error bound for LASSO as if $\boldsymbol\gamma_0$ is used. When the second bias term in (A.1) is dominant, the error bound of $\widehat{\boldsymbol\beta}$ can be controlled by the error bound of $\widehat{\boldsymbol\gamma}$. Combining the error bounds in the two cases, we obtain the oracle inequalities.

Lemma 12 *On event*

$$\begin{aligned}\Omega=\Big\{&\ell_{\mathsf{PL}}(\boldsymbol\beta_0+\Delta)-\ell_{\mathsf{PL}}(\boldsymbol\beta_0)-\Delta^\intercal\dot{\boldsymbol\ell}_{\mathsf{PL}}(\boldsymbol\beta_0)\\&\ge\kappa_{\mathsf{rsc},1}\|\Delta\|_2\{\|\Delta\|_2-\kappa_{\mathsf{rsc},2}\sqrt{\log(p)/N}\|\Delta\|_1\},\forall\|\Delta\|_2\le1\Big\},\end{aligned}$$

setting $\lambda_\beta\gtrsim\sqrt{\log(p)/N}$ such that

$$\lambda_\beta\ge3\left\|\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})-\mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})\mid\mathscr{L}\}+\mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\boldsymbol\gamma_0)\mid\mathscr{L}\}\right\|_\infty+\kappa_{\mathsf{rsc},1}\kappa_{\mathsf{rsc},2}\sqrt{\frac{\log(p)}{N}},$$

we have the oracle inequalities for estimation error of $\widehat{\boldsymbol\beta}$,

$$\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_0\|_2 \le \max\left\{14\sqrt{s_{\boldsymbol\beta}}\lambda_{\boldsymbol\beta}/\kappa_{\mathsf{rsc},1}, (1-\rho)7M\sigma^2_{\mathsf{max}}\|\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}\|_2/\kappa_{\mathsf{rsc},1}\right\},$$
$$\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_0\|_1 \le \max\left\{84 s_{\boldsymbol\beta}\lambda_{\boldsymbol\beta}/\kappa_{\mathsf{rsc},1}, (1-\rho)^2 21M^2\sigma^4_{\mathsf{max}}\|\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}\|_2^2/(\kappa_{\mathsf{rsc},1}\lambda_{\boldsymbol\beta})\right\}.$$

The constants $\kappa_{\mathsf{rsc},1},\kappa_{\mathsf{rsc},2}$ are the restrictive strong convexity parameters specified in Lemma 19.

We next prove the oracle inequalities. First, we note that by the definition of $\widehat{\boldsymbol\beta}$,

$$\ell^\dagger(\widehat{\boldsymbol\beta};\widehat{\boldsymbol\gamma}) + \lambda_{\boldsymbol\beta}\|\widehat{\boldsymbol\beta}\|_1 \le \ell^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) + \lambda_{\boldsymbol\beta}\|\boldsymbol\beta_0\|_1. \tag{A.2}$$

Denote the standardized estimation error as $\boldsymbol\delta = (\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)/\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_0\|_2$. Due to convexity of the loss function, we have for $t=\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_0\|_2\wedge 1$

$$\ell^\dagger(\boldsymbol\beta_0+t\boldsymbol\delta;\widehat{\boldsymbol\gamma}) + \lambda_{\boldsymbol\beta}\|\boldsymbol\beta_0+t\boldsymbol\delta\|_1 \le \ell^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) + \lambda_{\boldsymbol\beta}\|\boldsymbol\beta_0\|_1. \tag{A.3}$$

By the triangle inequality $\|\boldsymbol\beta_0\|_1 - \|\boldsymbol\beta_0+t\boldsymbol\delta\|_1 \le t\|\boldsymbol\delta\|_1$, we have from (A.3)

$$\ell^\dagger(\boldsymbol\beta_0+t\boldsymbol\delta;\widehat{\boldsymbol\gamma}) - \ell^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) \le t\lambda_{\boldsymbol\beta}\|\boldsymbol\delta\|_1 \tag{A.4}$$

To apply the restricted strong convexity of the complete data loss (11) established in Lemma 19, we show that the second order approximation error of the imputed loss is equivalent to that of the complete data loss,

$$\ell^\dagger(\boldsymbol\beta_0+t\boldsymbol\delta;\widehat{\boldsymbol\gamma}) - \ell^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) - t\boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) = \ell_{\mathsf{PL}}(\boldsymbol\beta_0+t\boldsymbol\delta) - \ell_{\mathsf{PL}}(\boldsymbol\beta_0) - t\boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}_{\mathsf{PL}}(\boldsymbol\beta_0).$$

Then by applying the restricted strong convexity event Ω, we obtain

$$\ell^\dagger(\boldsymbol\beta_0+t\boldsymbol\delta;\widehat{\boldsymbol\gamma}) - \ell^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) - t\boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) \ge t^2\kappa_{\mathsf{rsc},1} - t\kappa_{\mathsf{rsc},1}\kappa_{\mathsf{rsc},2}\sqrt{\log(p)/N}\|\boldsymbol\delta\|_1. \tag{A.5}$$

Applying (A.5) to (A.4), we have with large probability

$$t\boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) + t^2\kappa_{\mathsf{rsc},1} - t\kappa_{\mathsf{rsc},1}\kappa_{\mathsf{rsc},2}\sqrt{\log(p)/N}\|\boldsymbol\delta\|_1 \le t\lambda_{\boldsymbol\beta}\|\boldsymbol\delta\|_1$$

where $\|\boldsymbol\delta\|_2=1$ from definition. Thus, we have reach

$$t\kappa_{\mathsf{rsc},1} \le \lambda_{\boldsymbol\beta}\|\boldsymbol\delta\|_1 - \boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) + \kappa_{\mathsf{rsc},1}\kappa_{\mathsf{rsc},2}\sqrt{\log(p)/N}\|\boldsymbol\delta\|_1. \tag{A.6}$$

Next, we analyze $\boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})$ by the decomposition

$$\begin{aligned}\left|\boldsymbol\delta^{\mathsf T}\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma})\right| =& \boldsymbol\delta^{\mathsf T}\left[\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) - \mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) \mid \mathscr{L}\} + \mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\boldsymbol\gamma_0) \mid \mathscr{L}\}\right]\\ &+ \boldsymbol\delta^{\mathsf T}\mathbb{E}\left\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) - \dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\boldsymbol\gamma_0) \mid \mathscr{L}\right\}\\ \le& \|\boldsymbol\delta\|_1\left\|\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) - \mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\widehat{\boldsymbol\gamma}) \mid \mathscr{L}\} + \mathbb{E}\{\dot{\boldsymbol\ell}^\dagger(\boldsymbol\beta_0;\boldsymbol\gamma_0) \mid \mathscr{L}\}\right\|_\infty\\ &+ (1-\rho)\left\|\mathbb{E}_{i>n}[\mathbf{X}_i\{g(\boldsymbol\beta_0^{\mathsf T}\mathbf{X}_i) - g(\widehat{\boldsymbol\gamma}^{\mathsf T}\mathbf{W}_i)\} \mid \mathscr{L}]\right\|_2.\end{aligned} \tag{A.7}$$

To establish the rate for L_2-norm of $\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}}) \mid \mathscr{L}\}$, we note that

$$\mathbb{E}_{i>n}[\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)\} \mid \mathscr{L}] = \mathbb{E}_{i>n}\left[\mathbf{X}_i\left\{Y_i - g\left(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i\right)\right\} \mid \mathscr{L}\right]. \tag{A.8}$$

By the characterization of $\boldsymbol{\gamma}_0$ as in (6), we may rewrite (A.8) as

$$\mathbb{E}_{i>n}\left[\mathbf{X}_i\left\{Y_i - g\left(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i\right)\right\} \mid \mathscr{L}\right] = \mathbb{E}_{i>n}\left[g'(\boldsymbol{\gamma}_u^{\intercal}\mathbf{W}_i)\mathbf{X}_i\mathbf{W}_i^{\intercal} \mid \mathscr{L}\right](\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}), \tag{A.9}$$

where $\boldsymbol{\gamma}_u = u\widehat{\boldsymbol{\gamma}} + (1-u)\boldsymbol{\gamma}_0$ for some $u \in [0,1]$. Under Assumptions 1b and 2a, as well as the fact that $\mathbf{X}_i$ is a sub-vector of $\mathbf{W}_i$, we have

$$\begin{aligned}\left\|\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}}) \mid \mathscr{L}\}\right\|_2 &\leq \left\|\mathbb{E}_{i\in\mathcal{U}}\left[g'(\boldsymbol{\gamma}_u^{\intercal}\mathbf{W}_i)\mathbf{W}_i^{\otimes 2} \mid \mathscr{L}\right]\right\|_2 \|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2 \\ &\leq M\left\|\mathbb{E}\left(\mathbf{W}_i^{\otimes 2}\right)\right\|_2 \|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2 \leq M\sigma_{\max}^2\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2,\end{aligned} \tag{A.10}$$

where for any vector $\mathbf{x}$, $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^{\intercal}$. By the bound for (A.10) and the definition of $\lambda_{\boldsymbol{\beta}}$, we have the bound from (A.6)

$$t\kappa_{\mathsf{rsc},1} \leq 2\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}\|_1 + (1-\rho)M\sigma_{\max}^2\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2. \tag{A.11}$$

Hence, we can reach an immediate bound for estimation error from (A.11) without considering the sparsity of $\boldsymbol{\beta}_0$. We shall proceed to derive a sharper bound that involves the sparsity of $\boldsymbol{\beta}_0$. We separately analyze two cases.

$$\underline{\textbf{Case 1:}} \quad (1-\rho)M\sigma_{\max}^2\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2 \geq \|\boldsymbol{\delta}\|_1\lambda_{\boldsymbol{\beta}}/3.$$

In this case, the estimation error is dominated by $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$. We simply have from (A.11)

$$\begin{gathered} t\kappa_{\mathsf{rsc},1} \leq 7(1-\rho)M\sigma_{\max}^2\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2, \\ t\kappa_{\mathsf{rsc},1}\|\boldsymbol{\delta}\|_1\lambda_{\boldsymbol{\beta}}/3 \leq 7(1-\rho)^2M^2\sigma_{\max}^4\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2^2. \end{gathered}$$

Thus, we have

$$\begin{gathered} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq (1-\rho)7M\sigma_{\max}^2\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2/\kappa_{\mathsf{rsc},1}, \\ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq (1-\rho)^2 21M^2\sigma_{\max}^4\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2^2/(\kappa_{\mathsf{rsc},1}\lambda_{\boldsymbol{\beta}}). \end{gathered} \tag{A.12}$$

If case 1 does not hold, then instead

$$\underline{\textbf{Case 2:}} \quad (1-\rho)M\sigma_{\max}^2\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2 \leq \|\boldsymbol{\delta}\|_1\lambda_{\boldsymbol{\beta}}/3. \tag{A.13}$$

In this case, the estimation error is comparable to that when we have the true $\boldsymbol{\gamma}_0$ for the imputation. Thus, the sparsity of $\boldsymbol{\beta}_0$ may affect the estimation error.

Following the typical approach to establish the cone condition for $\boldsymbol{\delta}$, we analyze the symmetrized Bregman's divergence,

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\intercal}\left\{\dot{\boldsymbol{\ell}}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}) - \dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\right\} = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\boldsymbol{\delta}^{\intercal}\left\{\dot{\boldsymbol{\ell}}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}) - \dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\right\}. \tag{A.14}$$

Due to the convexity of the loss $\dot{\boldsymbol{\ell}}^{\dagger}(\cdot;\widehat{\boldsymbol{\gamma}})$ under Assumption 1b, the symmetrized Bregman's divergence (A.14) is nonnegative through a mean-value theorem,

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\intercal}\left\{\dot{\boldsymbol{\ell}}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})-\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\right\}\geq\inf_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\frac{1}{N}\sum_{i>n}g'(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i)\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\intercal}\mathbf{X}_i\}^2\geq 0.$$

Denote the indices set of nonzero coefficient in $\boldsymbol{\beta}_0$ as $\mathcal{O}_{\boldsymbol{\beta}}=\{j:\boldsymbol{\beta}_{0,j}\neq 0\}$. We denote the $\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}}$ and $\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}}$ as the sub-vectors for $\boldsymbol{\delta}$ at positions in $\mathcal{O}_{\boldsymbol{\beta}}$ and at positions not in $\mathcal{O}_{\boldsymbol{\beta}}$, respectively. The solution $\widehat{\boldsymbol{\beta}}$ satisfies the KKT condition

$$\|\dot{\boldsymbol{\ell}}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})\|_{\infty}\leq\lambda_{\boldsymbol{\beta}},\ \dot{\ell}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})_j=-\lambda_{\boldsymbol{\beta}}\operatorname{sign}(\widehat{\boldsymbol{\beta}}_j),\ j:\widehat{\beta}_j\neq 0.$$

From the KKT condition and the definitions of $\boldsymbol{\delta}$ and $\mathcal{O}_{\boldsymbol{\beta}}$, we have

$$\delta_j\dot{\ell}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})_j\leq|\delta_j|\lambda_{\boldsymbol{\beta}},\ j\in\mathcal{O}_{\boldsymbol{\beta}};\ \delta_j\dot{\ell}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})_j=\frac{-\widehat{\beta}_j\lambda_{\boldsymbol{\beta}}\operatorname{sign}(\widehat{\beta}_j)}{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2}=-\lambda_{\boldsymbol{\beta}}|\delta_j|,\ j\in\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}.\tag{A.15}$$

Applying the (A.15) to (A.14), we have the upper bound,

$$\begin{aligned}\boldsymbol{\delta}^{\intercal}\left\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})-\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\right\}=&\sum_{j\in\mathcal{O}_{\boldsymbol{\beta}}}\delta_j\dot{\ell}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})_j+\sum_{j\in\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}}\delta_j\dot{\ell}^{\dagger}(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}})_j+\boldsymbol{\delta}^{\intercal}\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\\ \leq&\lambda_{\boldsymbol{\beta}}\sum_{j\in\mathcal{O}_{\boldsymbol{\beta}}}|\delta_j|-\lambda_{\boldsymbol{\beta}}\sum_{j\in\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}}|\delta_j|+\boldsymbol{\delta}^{\intercal}\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\\ \leq&\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}}\|_1-\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}}\|_1+\left|\boldsymbol{\delta}^{\intercal}\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\right|.\end{aligned}$$

Then, we apply (A.10), the definition of $\lambda_{\boldsymbol{\beta}}$ and (A.13),

$$0\leq\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}}\|_1-\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}}\|_1+\frac{2}{3}\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}\|_1\quad\text{and}\quad\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}^{\mathsf{c}}}\|_1\leq 5\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}}\|_1.$$

Therefore, we can bound the L_1 norm of $\boldsymbol{\delta}$ by the cone property,

$$\|\boldsymbol{\delta}\|_1\leq 6\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\delta}_{\mathcal{O}_{\boldsymbol{\beta}}}\|_1\leq 6\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\delta}\|_2=6\sqrt{s_{\boldsymbol{\beta}}}.\tag{A.16}$$

We then apply the cone condition (A.16) and the case condition (A.13) to the bound (A.11),

$$t\kappa_{\mathsf{rsc},1}\leq 14\sqrt{s_{\boldsymbol{\beta}}}\lambda_{\boldsymbol{\beta}},\quad\text{and}\quad t\kappa_{\mathsf{rsc},1}\|\boldsymbol{\delta}\|_1\leq 84s_{\boldsymbol{\beta}}\lambda_{\boldsymbol{\beta}}$$

Thus, we obtain the rate for estimation error

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2\leq 14\sqrt{s_{\boldsymbol{\beta}}}\lambda_{\boldsymbol{\beta}}/\kappa_{\mathsf{rsc},1},\quad\text{and}\quad\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1\leq 84s_{\boldsymbol{\beta}}\lambda_{\boldsymbol{\beta}}/\kappa_{\mathsf{rsc},1}.\tag{A.17}$$

Since Case 1 and Case 2 are the complement of each other, one of them must occur. Thus, the bound of estimation error is controlled by the larger bound in the two cases,

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2&\leq\max\left\{14\sqrt{s_{\boldsymbol{\beta}}}\lambda_{\boldsymbol{\beta}}/\kappa_{\mathsf{rsc},1},(1-\rho)7M\sigma_{\mathsf{max}}^2\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}\|_2/\kappa_{\mathsf{rsc},1}\right\},\\ \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1&\leq\max\left\{84s_{\boldsymbol{\beta}}\lambda_{\boldsymbol{\beta}}/\kappa_{\mathsf{rsc},1},(1-\rho)^2 21M^2\sigma_{\mathsf{max}}^4\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}\|_2^2/(\kappa_{\mathsf{rsc},1}\lambda_{\boldsymbol{\beta}})\right\},\end{aligned}$$

which is our oracle inequality in Lemma 12.

Consistency We next show that the oracle inequality leads to the consistency under dimension condition (26). To show

$$\begin{aligned}&\left\|\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})-\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\mid\mathscr{L}\}+\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\mid\mathscr{L}\}\right\|_{\infty}\\ \le&\left\|\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})-\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\mid\mathscr{L}\}\right\|_{\infty}+\left\|\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\mid\mathscr{L}\}\right\|_{\infty}=O_p\left(\sqrt{\log(p)/N}\right),\end{aligned}$$

we express the term of interest as the sum of the following empirical processes

$$\begin{aligned}\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})-\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\mid\mathscr{L}\}=&\frac{1}{N}\sum_{i>n}\Big(\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)-Y_i+Y_i-g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)\}\\ &-\mathbb{E}_{i>n}\left[\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)-Y_i+Y_i-g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)\}\mid\mathscr{L}\right]\Big),\\ \mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\mid\mathscr{L}\}=&\frac{1}{N}\sum_{i=1}^{n}\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)-Y_i\}.\end{aligned}$$

Under Assumption 1a and 1a, $\mathbf{X}_i$ and $g(\boldsymbol{\beta}_0^{\intercal}\mathbf{X}_i)-Y_i$ are sub-Gaussian. According to Lemma 21, the event $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\le 1$ occurs with large probability, on which we have a bound for the sub-Gaussian norm of $Y_i-g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)$ by Lemma 14.

$$\|Y_i-g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)\|_{\psi_2}\le\max\{2\nu_1,M\sqrt{2}\sigma_{\mathsf{max}}\},\ i>n \tag{A.18}$$

Thus, we obtain from (A.18) that $Y_i-g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{W}_i)$ is sub-Gaussian with large probability. Thus by the properties of sub-Gaussian random variables in Lemma 17-d and 17-f, we have established that the elements in the summands of $\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})$ are all sub-exponential random variables conditionally on the labelled data. We apply the Bernstein's inequality (Lemma 17-h) conditionally on the labelled data to obtain

$$\begin{aligned}\left\|\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})-\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\widehat{\boldsymbol{\gamma}})\mid\mathscr{L}\}\right\|_{\infty}&=O_p\left(\sqrt{(1-\rho)\log(p)/N}\right),\\ \left\|\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\mid\mathscr{L}\}\right\|_{\infty}&=O_p\left(\sqrt{\rho\log(p)/N}\right).\end{aligned}$$

This establishes the order for $\lambda_{\boldsymbol{\beta}}$,

$$\lambda_{\boldsymbol{\beta}}\gtrsim\sqrt{(1-\rho)\log(p)/N}+\sqrt{\rho\log(p)/N}\asymp\sqrt{\log(p)/N}. \tag{A.19}$$

By Lemma 19 from Negahban et al. (2010), we have that the probability of restricted strong convexity event converges to one,

$$\mathbb{P}(\Omega)\ge 1-\kappa_{\mathsf{rsc},3}e^{\kappa_{\mathsf{rsc},4}N}\to 1.$$

Setting $\lambda_{\boldsymbol{\beta}}\asymp\sqrt{\log(p)/N}$ for optimal L_2 estimation, we achieve the stated conclusion

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2&=O_p\left(\sqrt{s_{\boldsymbol{\beta}}\log(p)/N}+(1-\rho)\sqrt{s_{\gamma}\log(p+q)/n}\right),\\ \sqrt{\log(p)/N}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1&=O_p\left(s_{\boldsymbol{\beta}}\log(p)/N+(1-\rho)^2\frac{s_{\gamma}\log(p+q)}{n}\right),\end{aligned}$$

by applying the rates from Lemma 21 and (A.19). For optimal L_1 estimation, we set a larger penalty $\lambda'_{\boldsymbol{\beta}}\asymp\sqrt{\log(p)/N}\vee\sqrt{s_{\gamma}\log(p+q)/(s_{\boldsymbol{\beta}}n)}\gtrsim\lambda_{\boldsymbol{\beta}}$ to achieve

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1=O_p\left(s_{\boldsymbol{\beta}}\sqrt{\log(p)/N}+(1-\rho)^2\sqrt{\frac{s_{\gamma}s_{\boldsymbol{\beta}}\log(p+q)}{n}}\right).$$

B2 Proof of Corollary 5

Under Assumption 1b, we have

$$|g(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}_{\text{new}}) - g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{\text{new}})| \le M|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}}\mathbf{x}_{\text{new}}|. \tag{A.20}$$

Since $\mathbf{x}_{\text{new}}$ satisfies Assumption 2a, we have

$$\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}}\mathbf{x}_{\text{new}}\|_{\psi_2} \le \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}.$$

The tail distribution is regulated by the sub-Gaussian norm by Lemma 17-a,

$$\mathbb{P}_{\text{new}}\left(|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}}\mathbf{x}_{\text{new}}| \ge t \mid \mathscr{D}\right) \le 2\exp\left(-t\sqrt{2}/\left\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}\right\}\right). \tag{A.21}$$

Combining (A.20) and (A.21), we obtain

$$\mathbb{P}_{\text{new}}\left(|g(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}_{\text{new}}) - g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{\text{new}})| \ge tM\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2} \mid \mathscr{D}\right) \le 2e^{-t}.$$

Thus,

$$\left|g(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}_{\text{new}}) - g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{\text{new}})\right| = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\right).$$

B3 Proof of Theorem 7

Our proof is organized in five parts. In Part 1, we establish the consistency of the cross-fitting estimator for precision matrix, namely $\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 = o_p(1)$ with $\widehat{\mathbf{u}}^{(k)}$ and $\mathbf{u}_0$ defined in (20) and (19), respectively. In Part 2, we show that the debiased estimator can be approximated by the empirical process

$$\begin{aligned}\sqrt{n}\left(\widehat{\mathbf{x}_{\text{std}}^{\mathsf{T}}\boldsymbol{\beta}} - \mathbf{x}_{\text{std}}^{\mathsf{T}}\boldsymbol{\beta}_0\right) &= -(1-\rho)\sqrt{n}\mathbf{u}_0^{\mathsf{T}}\dot{\boldsymbol{\ell}}_{\mathsf{imp}}(\boldsymbol{\gamma}_0) - \sqrt{n}\mathbf{u}_0^{\mathsf{T}}\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0) + o_p(1)\\ &= -n^{-\frac{1}{2}}\left[\sum_{i=1}^{n}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i) + \rho\cdot g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i) - Y_i\}\right.\\ &\qquad\left.+\rho\sum_{i>n}\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)\}\right] + o_p(1)\end{aligned}$$

As long as the asymptotic variance V_{SAS} defined in (29) is bounded and bounded away from zero, we have the asymptotic normality of the leading term from the Central Limit Theorem

$$\begin{aligned}&-n^{-\frac{1}{2}}V_{\mathsf{SAS}}^{-1/2}\left[\sum_{i=1}^{n}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i) + \rho\cdot g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i) - Y_i\}\right.\\ &\qquad\left.+\rho\sum_{i>n}\mathbf{X}_i\{g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)\}\right] \rightsquigarrow N(0,1).\end{aligned}$$

In Part 3, we deal with the asymptotic variance V_{SAS} and the consistency of the variance estimator $\widehat{V}_{\mathsf{SAS}}$ defined in (23). In Part 4, we reach the conclusion of the theorem based on,

$$\begin{aligned}&(1-\rho)\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2 + \|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2\\ &+\sqrt{n}\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2\left(\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 + \|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2\right) = o_p(1),\end{aligned} \tag{A.22}$$

for all $1 \le k \le K$. Following Part 4, we show in Part 5 that (28) implies (A.22).

Part 1: Consistency of estimated precision matrix

The definitions of $\mathbf{u}_0$ and $\widehat{\mathbf{u}}^{(k)}$ are given in (19) and (20). In this part, we show

$$\begin{aligned}\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2 =&O_p\left(\sqrt{(s_{\mathsf{u}}+s_{\boldsymbol{\beta}})\log(p)/(N-N_k)}+(1-\rho)\sqrt{s_{\gamma}\log(p+q)/(n-n_k)}\right)\\ =&O_p\left(\sqrt{(s_{\mathsf{u}}+s_{\boldsymbol{\beta}})\log(p)/N}+(1-\rho)\sqrt{s_{\gamma}\log(p+q)/n}\right).\end{aligned}$$

Since we set the number of folds $K\le 10$ to be finite, the estimation rate applies for $\widehat{\mathbf{u}}^{(k)}$ for all $k=1,\dots,K$.

We denote the components in the quadratic loss function of (20) and their derivatives as

$$\begin{aligned} m^{(k,k')}(\mathbf{u};\boldsymbol{\beta}) &= \frac{1}{N_{k'}}\sum_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\frac{1}{2}g'(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i)(\mathbf{X}_i^{\intercal}\mathbf{u})^2-\mathbf{u}^{\intercal}\mathbf{x}_{\mathrm{std}},\\ \dot{\mathbf{m}}^{(k,k')}(\mathbf{u};\boldsymbol{\beta}) &= \frac{\partial}{\partial\mathbf{u}}m^{(k,k')}(\mathbf{u};\boldsymbol{\beta}),\ \ddot{\mathbb{m}}^{(k,k')}(\boldsymbol{\beta})=\frac{\partial}{\partial\mathbf{u}}\dot{\mathbf{m}}^{(k,k')}(\mathbf{u};\boldsymbol{\beta}) \end{aligned} \tag{A.23}$$

for $k'\in\{1,\dots,K\}\setminus\{k\}$. We may express (20) as

$$\widehat{\mathbf{u}}^{(k)}=\operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^p}\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}m^{(k,k')}\left(\mathbf{u};\widehat{\boldsymbol{\beta}}^{(k,k')}\right)+\lambda_{\mathsf{u}}\|\mathbf{u}\|_1,$$

Similar to the proof of Theorem 1, we establish the estimation rate for $\widehat{\mathbf{u}}$ through an oracle inequality,

Lemma 13 *Under Assumptions 1, 2, we establish On event*

$$\Omega^{(k)}=\bigcap_{k'\neq k}\left\{\Delta^{\intercal}\ddot{\mathbb{m}}^{(k,k')}\left(\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\Delta\ge\kappa^*_{\mathsf{rsc},1}\|\Delta\|_2\{\|\Delta\|_2-\kappa^*_{\mathsf{rsc},2}\sqrt{\log(p)/N_{k'}}\|\Delta\|_1\},\forall\|\Delta\|_2\le 1\right\},$$

setting $\lambda_{\mathsf{u}}\asymp\sqrt{\log(p)/N}$ *such that*

$$\begin{aligned}\lambda_{\mathsf{u}}\ge 3\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}\Bigg\{&\left\|\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)-\mathbb{E}\left\{\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\mid\mathscr{D}^c_{k'}\right\}\right\|_\infty\\ &+\kappa^*_{\mathsf{rsc},1}\kappa^*_{\mathsf{rsc},2}\sqrt{\log(p)/N_{k'}}\Bigg\},\end{aligned}$$

we have the oracle inequality for estimation error of $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned}\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2&\le\max\left\{14\sqrt{s_{\mathsf{u}}}\lambda_{\mathsf{u}}/\kappa^*_{\mathsf{rsc},1},7M\sigma^3_{\mathsf{max}}\|\mathbf{u}_0\|_2\sup_{k'\neq k}\left\|\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2/\kappa^*_{\mathsf{rsc},1}\right\},\\ \|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_1&\le\max\left\{84s_{\mathsf{u}}\lambda_{\mathsf{u}}/\kappa^*_{\mathsf{rsc},1},21M\sigma^6_{\mathsf{max}}\|\mathbf{u}_0\|_2^2\sup_{k'\neq k}\left\|\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2^2/(\kappa^*_{\mathsf{rsc},1}\lambda_{\mathsf{u}})\right\}.\end{aligned}$$

The constants $\kappa^*_{\mathsf{rsc},1},\kappa^*_{\mathsf{rsc},2}$ *are the restrictive strong convexity parameters specified in Lemma 20.*

The proof of Lemma 13 repeats the proof of the oracle inequality for Theorem 1, so we put the detail to Section C.

To use Lemma 13 for the estimation rate of $\widehat{\mathbf{u}}$, we only need to verify two conditions. First, the event $\Omega^{(k)}$ occurs with probability tending to one. Second, the oracle choice of λ_{u} is of order $\sqrt{\log(p)/N}$.

Repeating Theorem 1 for each $\widehat{\boldsymbol{\beta}}^{(k,k')}$, we have under (28)

$$\left\|\widehat{\boldsymbol{\beta}}^{(k,k')} - \boldsymbol{\beta}_{0}\right\|_2 = o_p(1).$$

Then by Lemma 20, the sets whose intersection forms $\Omega^{(k)}$ each occurs with probability tending to one. Since we set the number of fold finite $K \leq 10$, we can take union bound to obtain that $\Omega^{(k)}$ occurs with probability tending to one.

We may write

$$\begin{aligned}
&\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_{0}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \mathbb{E}\left\{\dot{\mathbf{m}}^{(k)}\left(\mathbf{u}_{0}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \mid \mathscr{D}_{k'}^{c}\right\} \\
=&\frac{1}{N_{k'}} \sum_{i \in \mathcal{I}_{k'} \cup \mathcal{J}_{k'}} g'(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\intercal}\mathbf{u}_{0} - \mathbb{E}_{i \in \mathcal{I}_{k'} \cup \mathcal{J}_{k'}}\left\{g'(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\intercal}\mathbf{u}_{0} \mid \mathscr{D}_{k'}^{c}\right\}. \qquad \text{(A.24)}
\end{aligned}$$

Each element in (A.24) is an empirical process. Under Assumptions 1b and 2a, we can show that each summand is a sub-exponential random variable by Lemma 17-e, 17-f,

$$\left\|g'(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)X_{i,j}\mathbf{X}_i^{\intercal}\mathbf{u}_{0}\right\|_{\psi_1} \leq M\|X_{i,j}\|_{\psi_2}\|\mathbf{X}_i^{\intercal}\mathbf{u}_{0}\|_{\psi_2} \leq M\sigma_{\mathsf{max}}^2\|\mathbf{u}_{0}\|_2/2.$$

Hence, we can apply the Bernstein's inequality to show that

$$\left\|\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_{0}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \mathbb{E}\left\{\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_{0}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \mid \mathscr{D}_{k'}^{c}\right\}\right\|_{\infty} = O_p\left(\sqrt{\log(p)/N_{k'}}\right).$$

Using the fact that $N_{k'} \asymp N$, we obtain that the oracle λ_{u} is of order $O_p\left(\sqrt{\log(p)/N}\right)$.

Therefore, we can apply Lemma 13 to obtain

$$\begin{aligned}
\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_{0}\|_2 =&O_p\left(\sqrt{s_{\mathsf{u}}\log(p)/N} + \sup_{k' \neq k}\left\|\widehat{\boldsymbol{\beta}}^{(k,k')} - \boldsymbol{\beta}_{0}\right\|_2\right) \\
=&O_p\left(\sqrt{(s_{\boldsymbol{\beta}} + s_{\mathsf{u}})\log(p)/N} + (1-\rho)\sqrt{s_{\boldsymbol{\gamma}}\log(p+q)/n}\right).
\end{aligned}$$

Part 2: Asymptotic approximation

Under Assumption 2b-i, we also have the tightness of $\|\widehat{\mathbf{u}}^{(k)}\|_2$ from the bound of $\|\mathbf{u}_{0}\|_2$

$$\|\mathbf{u}_{0}\|_2 \leq \|\boldsymbol{\Sigma}_{0}^{-1}\|_2\|\mathbf{x}_{\mathrm{std}}\|_2 \leq \sigma_{\mathsf{min}}^{-2},\ \|\widehat{\mathbf{u}}^{(k)}\|_2 \leq \|\mathbf{u}_{0}\|_2 + \|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_{0}\|_2 = O_p(1). \qquad \text{(A.25)}$$

Define the scores of in-fold data as

$$\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}; \boldsymbol{\gamma}) = \frac{1}{N_k}\left[\sum_{i \in \mathcal{J}_k}\mathbf{X}_i\{g(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i) - g(\boldsymbol{\gamma}^{\intercal}\mathbf{W}_i)\} + \sum_{i \in \mathcal{I}_k}\mathbf{X}_i\{g(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i) - Y_i\}\right],$$

$$\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}) = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\mathbf{X}_i\{g(\boldsymbol{\gamma}^{\intercal}\mathbf{W}_i) - Y_i\}. \tag{A.26}$$

Since $\widehat{\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}}$ is the average over K (at most 10) cross-fitted estimators, it suffices to study one of the cross-fitted estimators,

$$\widehat{\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}}^{(k)} = \mathbf{x}_{\rm std}^{\intercal}\widehat{\boldsymbol{\beta}}^{(k)} - \dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)}) - (1-\rho)\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\widehat{\boldsymbol{\gamma}}^{(k)}),\ \widehat{\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}}^{(k)}. \tag{A.27}$$

We denote the expected Hessian matrices of losses in (A.26) as

$$\begin{aligned}\mathbb{H}(\boldsymbol{\beta}) = \mathbb{E}\left\{g'(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\intercal}\right\},\ \boldsymbol{\Sigma}_0 = \mathbb{H}(\boldsymbol{\beta}_0),\\ \mathbb{H}_{\mathsf{imp}}(\boldsymbol{\gamma}) = \mathbb{E}\left\{g'(\boldsymbol{\gamma}^{\intercal}\mathbf{W}_i)\mathbf{X}_i\mathbf{W}_i^{\intercal}\right\},\ \boldsymbol{\Sigma}_{\mathsf{imp}} = \mathbb{H}_{\mathsf{imp}}(\boldsymbol{\gamma}_0).\end{aligned} \tag{A.28}$$

Our analysis of the approximation error is based on the first order Mean Value Theorem identity,

$$\begin{aligned}&\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\} + (1-\rho)\mathbb{E}\{\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\}\\ =&\underbrace{\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\}}_{=0} + \mathbb{H}(\widetilde{\boldsymbol{\beta}})\{\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\} - (1-\rho)\mathbb{H}_{\mathsf{imp}}(\widetilde{\boldsymbol{\gamma}})\{\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\}\\ &+ (1-\rho)\underbrace{\mathbb{E}\{\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)\}}_{=0} + (1-\rho)\mathbb{H}_{\mathsf{imp}}(\widetilde{\boldsymbol{\gamma}})\{\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\}\\ =&\mathbb{H}(\widetilde{\boldsymbol{\beta}})\{\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\}\end{aligned} \tag{A.29}$$

for some $\widetilde{\boldsymbol{\beta}}$ on the path from $\bar{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}})$ to $\boldsymbol{\beta}_0$ and some $\widetilde{\boldsymbol{\gamma}}$ on the path from $\widehat{\boldsymbol{\gamma}}$ to $\boldsymbol{\gamma}_0$. The conditional expectation notation is declared at Definition 11. Based on (A.29), we analyze the approximation error for $\sqrt{n}\left(\widehat{\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}}^{(k)} - \mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}_0\right)$ through the following decomposition,

$$\begin{aligned}&\sqrt{n}\left(\widehat{\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}}^{(k)} - \sqrt{n}\mathbf{x}_{\rm std}^{\intercal}\boldsymbol{\beta}_0\right) + \sqrt{n}\mathbf{u}_0^{\intercal}\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0) + \sqrt{n}(1-\rho)\mathbf{u}_0^{\intercal}\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)\\ =&\sqrt{n}\mathbf{x}_{\rm std}^{\intercal}(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0) - \sqrt{n}\widehat{\mathbf{u}}^{(k)\intercal}\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)}) - \sqrt{n}(1-\rho)\widehat{\mathbf{u}}^{(k)\intercal}\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\widehat{\boldsymbol{\gamma}}^{(k)})\\ &+ \sqrt{n}\mathbf{u}_0^{\intercal}\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0) + (1-\rho)\mathbf{u}_0^{\intercal}\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0) + \sqrt{n}\widehat{\mathbf{u}}^{(k)\intercal}\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\}\\ &+ \sqrt{n}(1-\rho)\widehat{\mathbf{u}}^{(k)\intercal}\mathbb{E}\{\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\} - \sqrt{n}\widehat{\mathbf{u}}^{(k)\intercal}\mathbb{H}(\widetilde{\boldsymbol{\beta}})\{\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\}\\ =&\underbrace{\sqrt{n}\left\{\mathbf{x}_{\rm std} - \mathbb{H}(\widetilde{\boldsymbol{\beta}})\mathbf{u}_0\right\}^{\intercal}(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0)}_{T_1} + \underbrace{\sqrt{n}\left(\mathbf{u}_0 - \widehat{\mathbf{u}}^{(k)}\right)^{\intercal}\mathbb{H}(\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0)}_{T_2}\\ &+ \underbrace{\sqrt{n}\widehat{\mathbf{u}}^{(k)\intercal}\left[\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\} - \left\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)}) - \dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\right\}\right]}_{T_3}\\ &+ \underbrace{\sqrt{n}(1-\rho)\widehat{\mathbf{u}}^{(k)\intercal}\left[\mathbb{E}\{\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\} - \left\{\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\widehat{\boldsymbol{\gamma}}^{(k)}) - \dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)\right\}\right]}_{T_4}\end{aligned}$$

$$+\underbrace{\sqrt{n}\left(\mathbf{u}_0-\widehat{\mathbf{u}}^{(k)}\right)^{\mathsf{T}}\left\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)+(1-\rho)\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)\right\}}_{T_5} \tag{A.30}$$

Here we state the rates for T_1-T_5,

$$T_1=O_p\left(\sqrt{n}\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2^2\right),\ T_2=O_p\left(\sqrt{n}\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2\right),$$
$$T_3=O_p\left(\rho\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2+\sqrt{\rho(1-\rho)}\|\widehat{\boldsymbol{\gamma}}^{(k)}-\boldsymbol{\gamma}_0\|_2\right),$$
$$T_4=O_p\left((1-\rho)\|\widehat{\boldsymbol{\gamma}}^{(k)}-\boldsymbol{\gamma}_0\|_2\right),\ T_5=O_p\left(\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\right).$$

With the assumed estimation rate in (A.22), we have

$$T_1+T_2+T_3+T_4+T_5=o_p(1).$$

Thus, we have shown

$$\begin{aligned}\sqrt{n}\left(\widehat{\mathbf{x}_{\mathrm{std}}^{\mathsf{T}}\boldsymbol{\beta}}-\mathbf{x}_{\mathrm{std}}^{\mathsf{T}}\boldsymbol{\beta}_0\right)=&\frac{1}{K}\sum_{k=1}^{K}\sqrt{n}\left(\widehat{\mathbf{x}_{\mathrm{std}}^{\mathsf{T}}\boldsymbol{\beta}}-\mathbf{x}_{\mathrm{std}}^{\mathsf{T}}\boldsymbol{\beta}_0\right)\\ =&\frac{1}{K}\sum_{k=1}^{K}-\sqrt{n}\mathbf{u}_0^{\mathsf{T}}\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)-\sqrt{n}(1-\rho)\mathbf{u}_0^{\mathsf{T}}\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)+o_p(1)\\ =&-\sqrt{n}\mathbf{u}_0^{\mathsf{T}}\dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)-\sqrt{n}(1-\rho)\mathbf{u}_0^{\mathsf{T}}\dot{\boldsymbol{\ell}}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)+o_p(1).\end{aligned}$$

Using the indicator $R_i=I(i\le n)$, we can alternatively write

$$\widehat{\mathbf{x}_{\mathrm{std}}^{\mathsf{T}}\boldsymbol{\beta}}-\mathbf{x}_{\mathrm{std}}^{\mathsf{T}}\boldsymbol{\beta}_0=\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i\{Y_i-g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)\}-\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i\{g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)-g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\}+o_p\left((\rho n)^{-1/2}\right). \tag{A.31}$$

We provide the details of T_1-T_5 in Section C2.

Part 3: Variance estimation

Finally, we show that asymptotic variance V_{SAS} defined in (29) is bounded from infinity and zero with the consistent estimator $\widehat{V}_{\mathsf{SAS}}$ defined in (23).

By the Cauchy-Schwartz inequality, we have a bound for the variance

$$\begin{aligned}V_{\mathsf{SAS}}=&\mathbb{E}\left[(\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)+\rho\cdot g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)-Y_i\}^2\right]\\ &+\rho(1-\rho)\mathbb{E}[(\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i)^2\{g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)-g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\}^2]\\ \le&\sqrt{\mathbb{E}\left[(\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i)^4\right]\mathbb{E}\left[\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)+\rho\cdot g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)-Y_i\}^4\right]}\\ &+\rho(1-\rho)\sqrt{\mathbb{E}\left[(\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i)^4\right]\mathbb{E}\left[\{g(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{W}_i)-g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i)\}^4\right]}\end{aligned}$$

Under Assumptions 1a, 2a, we have the sub-Gaussian and sub-exponential variables

$$\|\mathbf{u}_0^{\mathsf{T}}\mathbf{X}_i\|_{\psi_2}\le\|\mathbf{u}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\le\sigma_{\mathsf{min}}^{-2}\sigma_{\mathsf{max}}/\sqrt{2},$$

$$\|(1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)+\rho\cdot g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)-Y_i\|_{\psi_2}\le 2(\nu_1\vee\nu_2),$$
$$\|g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)-g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)\|_{\psi_2}\le 2\{\|g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)-Y_i\|_{\psi_2}\vee\|g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)-Y_i\|_{\psi_2}\}\le 2(\nu_1\vee\nu_2).$$

By the bound for the moments of sub-Gaussian and sub-exponential random variables stated in Lemma 17-b, we have

$$V_{\mathsf{SAS}}\le 8\sqrt{2}\sigma_{\mathsf{min}}^{-4}\sigma_{\mathsf{max}}^{2}(\nu_1\vee\nu_2)^2.$$

Under Assumptions 1b, 2a, 2b-i and 2c, we have a lower bound for V_{SAS},

$$\begin{aligned}V_{\mathsf{SAS}}\ge&\mathbf{u}_0^\intercal\mathbb{E}[\mathbf{X}_i\mathbf{X}_i^\intercal\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)+\rho\cdot g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)-Y_i\}^2]\mathbf{u}_0\\ \ge&\|\mathbf{u}_0\|_2^2\sigma_{\mathsf{min}}^4\\ \ge&M^{-2}\sigma_{\mathsf{max}}^{-4}\sigma_{\mathsf{min}}^{4}\nu_3\|\mathbf{x}_{\mathrm{std}}\|_2^2,\end{aligned}$$

which is bounded away from zero.

We analyze the estimation error of variance $\widehat{V}_{\mathsf{SAS}}-V_{\mathsf{SAS}}$ through the decomposition,

$$\begin{aligned}&\widehat{V}_{\mathsf{SAS}}-V_{\mathsf{SAS}}\\ =&\left.\begin{aligned}&\sum_{k=1}^K\frac{n_k}{n}\Bigg(\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^2\\ &\quad-\mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^2\mid\mathscr{D}_k^c\right]\Bigg)\end{aligned}\right\}T_1'\\ &\left.\begin{aligned}+&\sum_{k=1}^K\frac{n_k}{n}\Bigg(\mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^2\mid\mathscr{D}_k^c\right]\\ &\quad-\mathbb{E}\left[(\mathbf{u}_0^\intercal\mathbf{X}_i)^2\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal\mathbf{W}_i)+\rho\cdot g(\boldsymbol{\beta}_0^\intercal\mathbf{X}_i)-Y_i\}^2\right]\Bigg)\end{aligned}\right\}T_2'\\ &\left.\begin{aligned}+&\rho(1-\rho)\sum_{k=1}^K\frac{N_k-n_k}{N-n}\Bigg(\frac{K}{N_k-n_k}\sum_{i\in\mathcal{J}_k}(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)\}^2\\ &\quad-\mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)\}^2\mid\mathscr{D}_k^c\right]\Bigg)\end{aligned}\right\}T_3'\\ &\left.\begin{aligned}+&\rho(1-\rho)\sum_{k=1}^K\frac{N_k-n_k}{N-n}\Bigg(\mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)\}^2\mid\mathscr{D}_k^c\right]\\ &\quad-\mathbb{E}\left[(\mathbf{u}_0^\intercal\mathbf{X}_i)^2\{g(\boldsymbol{\beta}_0^\intercal\mathbf{X}_i)-g(\boldsymbol{\gamma}_0^\intercal\mathbf{W}_i)\}^2\right]\Bigg)\end{aligned}\right\}T_4'\end{aligned}$$

Here we state the rates for T_1'-T_4',

$$T_1'=O_p\left(n^{-1/2}\right),\ T_2'=O_p\left(\|\widehat{\mathbf{u}}-\mathbf{u}_0\|_2+(1-\rho)\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2+\rho\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2\right),$$
$$T_3'=O_p\left(\rho(1-\rho)N^{-1/2}\right),\ T_4'=O_p\left(\rho(1-\rho)\left\{\|\widehat{\mathbf{u}}-\mathbf{u}_0\|_2+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2+\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2\right\}\right).$$

With the assumed estimation rate in (A.22), we have

$$T_1' + T_2' + T_3' + T_4' = o_p(1).$$

We provide the details of T_1'-T_4' in Section C2.

PART 4: CONCLUSION WITH ESTIMATION RATES

From the approximation in Part 2 and the boundedness and non-degeneracy of V_{SAS} in Part 3, we have shown the asymptotic normality of the cross-fitted debiased estimator

$$\sqrt{n} V_{\mathsf{SAS}}^{-1/2} \left(\widehat{\mathbf{x}_{\mathrm{std}}^{\mathsf{T}} \boldsymbol{\beta}} - \mathbf{x}_{\mathrm{std}}^{\mathsf{T}} \boldsymbol{\beta}_0 \right) \rightsquigarrow N(0,1).$$

Together with the consistency of $\widehat{V}_{\mathsf{SAS}}$ in Part 3, we have

$$\sqrt{n} \widehat{V}_{\mathsf{SAS}}^{-1/2} \left(\widehat{\mathbf{x}_{\mathrm{std}}^{\mathsf{T}} \boldsymbol{\beta}} - \mathbf{x}_{\mathrm{std}}^{\mathsf{T}} \boldsymbol{\beta}_0 \right) \rightsquigarrow N(0,1).$$

PART 5: SUFFICIENT DIMENSION CONDITION

We have established the rate of estimation for $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ from Lemma 21, Theorem 1 and Part 4 of this proof above. Since we only keep one fold of the data away for the cross-fitted estimators, they follow the same rates of estimation,

$$\begin{aligned}
\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2 &= O_p\left(\sqrt{s_{\boldsymbol{\gamma}} \log(p+q)/n}\right), \\
\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 &= O_p\left(\sqrt{s_{\boldsymbol{\beta}} \log(p)/N} + (1-\rho)\sqrt{s_{\boldsymbol{\gamma}} \log(p+q)/n}\right), \\
\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 &= O_p\left(\sqrt{(s_{\boldsymbol{\beta}} + s_{\mathsf{u}}) \log(p)/N} + (1-\rho)\sqrt{s_{\boldsymbol{\gamma}} \log(p+q)/n}\right).
\end{aligned}$$

Applying the rates of estimation, we show dimension assumption (28) is sufficient for (A.22).

B4 Efficiency of SAS Inference

RELATIVE EFFICIENCY TO SUPERVISED LEARNING

Proof [Proof of Proposition 9] We prove the Proposition by direct calculation

$$\begin{aligned}
& V_{\mathsf{SL}} - V_{\mathsf{SAS}} \\
=& \mathbb{E}[(\mathbf{u}_0^{\mathsf{T}} \mathbf{X}_i)^2 \{Y - g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i)\}^2] - \mathbb{E}[(\mathbf{u}_0^{\mathsf{T}} \mathbf{X}_i)^2 \{Y - (1-\rho) \cdot \mathbb{E}(Y|\mathbf{S}_i, \mathbf{X}_i)\rho \cdot g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i)\}^2] \\
=& \mathbb{E}[(\mathbf{u}_0^{\mathsf{T}} \mathbf{X}_i)^2 \{(1-\rho^2) g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i)^2 - 2(1-\rho) g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i) \mathbb{E}(Y|\mathbf{S}_i, \mathbf{X}_i) + (1-\rho^2) \mathbb{E}(Y|\mathbf{S}_i, \mathbf{X}_i)^2\}] \\
=& (1-\rho)^2 \mathbb{E}[(\mathbf{u}_0^{\mathsf{T}} \mathbf{X}_i)^2 \{\mathbb{E}(Y|\mathbf{S}_i, \mathbf{X}_i) - g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i)\}^2] \\
& + 2\rho(1-\rho) \mathbb{E}[(\mathbf{u}_0^{\mathsf{T}} \mathbf{X}_i)^2 \{\mathbb{E}(Y|\mathbf{S}_i, \mathbf{X}_i)^2 + g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i)^2\}].
\end{aligned}$$

The last expression is the sum of expectations of complete squares, so it must be non-negative. Thus, we have shown that the SAS asymptotic variance is no greater than the supervised learning variance. The equality holds only if 1) $\rho = 1$ all samples are labelled; 2) or $\rho = 0$ and $\mathbf{u}_0^{\mathsf{T}} \mathbf{X}_i \{\mathbb{E}(Y|\mathbf{S}_i, \mathbf{X}_i) - g(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i)\} = 0$ almost surely. ■

Efficiency Bound among Semi-parametric RAL Estimators

Proof [Proof of Proposition 10] The proof follows the flow of Section D.2 in Kallus and Mao (2020). The semi-parametric model for the observed data is

$$\mathcal{M}_{\mathsf{obs}} = \Big\{ f_{\mathbf{X},Y,\mathbf{S},R}(\mathbf{x},y,\mathbf{s},r) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{S}|\mathbf{X}}(\mathbf{s}|\mathbf{x}) \left\{\rho f_{Y|\mathbf{S},\mathbf{X}}(y|\mathbf{s},\mathbf{x})\right\}^r (1-\rho)^{1-r} : \\ f_{\mathbf{X}}, f_{\mathbf{S}|\mathbf{X}}, f_{Y|\mathbf{S},\mathbf{X}} \text{ are arbitrary pdf/pmf, } \Big\}. \tag{A.32}$$

We consider the parametric sub-model

$$\mathcal{M}_{\mathsf{par}} = \Big\{ f_{\mathbf{X},Y,\mathbf{S},R}(\mathbf{x},y,\mathbf{s},r;\boldsymbol{\zeta}) = f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\zeta}) f_{\mathbf{S}|\mathbf{X}}(\mathbf{s}|\mathbf{x};\boldsymbol{\zeta}) \left\{\rho f_{Y|\mathbf{S},\mathbf{X}}(y|\mathbf{s},\mathbf{x};\boldsymbol{\zeta})\right\}^r \\ \times (1-\rho)^{1-r} : \ \boldsymbol{\zeta} \in \mathbb{R}^d \Big\}. \tag{A.33}$$

The score vector of the parametric sub-model is

$$\begin{aligned} &\boldsymbol{\Psi}(\mathbf{X},Y,\mathbf{S},R) \\ =& \frac{\partial \log\{f_{\mathbf{X},Y,\mathbf{S},R}(\mathbf{X},Y,\mathbf{S},R;\boldsymbol{\zeta})\}}{\partial \boldsymbol{\zeta}}\Big|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} \\ =& \frac{\partial \log\{f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\zeta})\}}{\partial \boldsymbol{\zeta}}\Big|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} + \frac{\partial \log\{f_{\mathbf{S}|\mathbf{X}}(\mathbf{S}|\mathbf{X};\boldsymbol{\zeta})\}}{\partial \boldsymbol{\zeta}}\Big|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} + R\, \frac{\partial \log\{f_{Y|\mathbf{S},\mathbf{X}}(Y \mid \mathbf{S},\mathbf{X};\boldsymbol{\zeta})\}}{\partial \boldsymbol{\zeta}}\Big|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} \\ =& \boldsymbol{\Psi}_{\mathbf{X}}(\mathbf{X}) + \boldsymbol{\Psi}_{\mathbf{S}}(\mathbf{S},\mathbf{X}) + R\boldsymbol{\Psi}_Y(Y,\mathbf{S},\mathbf{X}). \end{aligned} \tag{A.34}$$

Next, we decompose the the Hilbert space of mean zero finite variance random variables measurable to $\sigma\{\mathbf{X},\mathbf{S},R,YR\}$, denoted as $\mathcal{H}$. The model tangent space spanned by the score (A.34) is a linear sub-space of $\mathcal{H}$,

$$\begin{gathered} \Lambda = \Lambda_{\mathbf{X}} \oplus \Lambda_{\mathbf{S}} \oplus \Lambda_Y, \\ \Lambda_{\mathbf{X}} = \bigcup_{\mathcal{M}_{\mathsf{par}}} \mathrm{span}\{\boldsymbol{\Psi}_{\mathbf{X}}(\mathbf{X})\} = \{h(\mathbf{X}) \in \mathcal{H} : \mathbb{E}[h(\mathbf{X})] = 0\}, \\ \Lambda_{\mathbf{S}} = \bigcup_{\mathcal{M}_{\mathsf{par}}} \mathrm{span}\{\boldsymbol{\Psi}_{\mathbf{S}}(\mathbf{S},\mathbf{X})\} = \{h(\mathbf{S},\mathbf{X}) \in \mathcal{H} : \mathbb{E}[h(\mathbf{S},\mathbf{X}) \mid \mathbf{X}] = 0\}, \\ \Lambda_{\mathbf{Y}} = \bigcup_{\mathcal{M}_{\mathsf{par}}} \mathrm{span}\{R\boldsymbol{\Psi}_Y(Y,\mathbf{S},\mathbf{X})\} = \{Rh(Y,\mathbf{S},\mathbf{X}) \in \mathcal{H} : \mathbb{E}[h(Y,\mathbf{S},\mathbf{X}) \mid \mathbf{S},\mathbf{X}] = 0\}. \end{gathered} \tag{A.35}$$

The orthogonal space of model tangent space Λ is

$$\Lambda^{\perp} = \{h(R,\mathbf{S},\mathbf{X}) \in \mathcal{H} : \mathbb{E}[h(R,\mathbf{S},\mathbf{X}) \mid \mathbf{S},\mathbf{X}] = 0\},\ \mathcal{H} = \Lambda \oplus \Lambda^{\perp}. \tag{A.36}$$

Now, we verify that the supervised learning influence function

$$\phi_{\mathsf{SL}}(\theta;\boldsymbol{\beta}) = \frac{R}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{Y - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\}$$

is indeed an influence function for $\mathbf{x}_{\text{std}}^{\mathsf{T}}\boldsymbol{\beta}$ by showing

$$\mathbb{E}\{\phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0)\boldsymbol{\Psi}(\mathbf{X},Y,\mathbf{S},R)\} = \mathbf{x}_{\text{std}}\frac{d}{d\boldsymbol{\zeta}}\boldsymbol{\beta}(\boldsymbol{\zeta})\bigg|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0}.$$

Since $\boldsymbol{\beta}(\boldsymbol{\zeta})$ is an implicit function of $\boldsymbol{\zeta}$ through the moment condition

$$\mathbb{E}_{\boldsymbol{\zeta}}[\mathbf{X}\{g(\boldsymbol{\beta}(\boldsymbol{\zeta})^{\mathsf{T}}\mathbf{X}) - Y\} = 0,$$

we solve for its derivative by differentiating the moment condition

$$\begin{aligned}
\frac{d}{d\boldsymbol{\zeta}}\mathbb{E}_{\boldsymbol{\zeta}}[\mathbf{X}\{g(\boldsymbol{\beta}(\boldsymbol{\zeta})^{\mathsf{T}}\mathbf{X}) - Y\}]\bigg|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} =& 0 \\
\frac{d}{d\boldsymbol{\zeta}}\mathbb{E}_{\boldsymbol{\zeta}}[\mathbf{X}\{g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}) - Y\}]\bigg|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} + \mathbb{E}_{\boldsymbol{\zeta}_0}\{\mathbf{X}\mathbf{X}^{\mathsf{T}}g'(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X})\}\,\frac{d}{d\boldsymbol{\zeta}}\boldsymbol{\beta}(\boldsymbol{\zeta})\bigg|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} =& 0 \\
-\boldsymbol{\Theta}_0\mathbb{E}_{\boldsymbol{\zeta}_0}[\mathbf{X}\{g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}) - Y\}\{\boldsymbol{\Psi}_{\mathbf{X}}(\mathbf{X}) + \boldsymbol{\Psi}_{\mathbf{S}}(\mathbf{S},\mathbf{X}) + \boldsymbol{\Psi}_{Y}(Y,\mathbf{S},\mathbf{X})\}] =& \frac{d}{d\boldsymbol{\zeta}}\boldsymbol{\beta}(\boldsymbol{\zeta})\bigg|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0}.
\end{aligned}$$

Then, we verify that the supervised learning influence function is valid

$$\begin{aligned}
\frac{d}{d\boldsymbol{\zeta}}\mathbf{x}_{\text{std}}^{\mathsf{T}}\boldsymbol{\beta}(\boldsymbol{\zeta})\bigg|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} =& -\mathbb{E}_{\boldsymbol{\zeta}_0}\left[\frac{R}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}) - Y\}\boldsymbol{\Psi}(\mathbf{X},Y,\mathbf{S},R)\right] \\
=& \mathbb{E}\{\phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0)\boldsymbol{\Psi}(\mathbf{X},Y,\mathbf{S},R)\}.
\end{aligned}$$

Finally, we derive the efficient influence function by subtract from ϕ_{SL} its projection onto $\Lambda^{\perp} = \Lambda_R$. Let $\Pi[h(\mathbf{D}) \mid \Lambda]$ be the projection of $h(\mathbf{D}) \in \mathcal{H}$ to the space Λ. We can easily calculate the projection of ϕ_{SL} onto Λ_R,

$$\begin{aligned}
\Pi[\phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0) \mid \Lambda_R] =& \mathbb{E}\{\phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0) \mid R,\mathbf{S},\mathbf{X}\} - \mathbb{E}\{\phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0) \mid \mathbf{S},\mathbf{X}\} \\
=& \frac{R}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{\mathbb{E}(Y \mid \mathbf{S},\mathbf{X}) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\} - \mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{\mathbb{E}(Y \mid \mathbf{S},\mathbf{X}) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\}.
\end{aligned}$$

The efficient influence function is thus obtained

$$\begin{aligned}
\phi_{\mathsf{eff}}(\theta_0;\boldsymbol{\beta}_0) =& \phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0) - \Pi[\phi_{\mathsf{SL}}(\theta_0;\boldsymbol{\beta}_0) \mid \Lambda_R] \\
=& \frac{R}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{Y - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\} - \frac{R}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{\mathbb{E}(Y \mid \mathbf{S},\mathbf{X}) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\} \\
&+ \mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{\mathbb{E}(Y \mid \mathbf{S},\mathbf{X}) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\} \\
=& \frac{R}{\rho}\mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{Y - \mathbb{E}(Y \mid \mathbf{S},\mathbf{X})\} + \mathbf{u}_0^{\mathsf{T}}\mathbf{X}\{\mathbb{E}(Y \mid \mathbf{S},\mathbf{X}) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\}.
\end{aligned}$$

■

Appendix C. Auxiliary Results

C1 General

Lemma 14 *Under Assumptions 1a, 1b, 2a, the residuals of the imputed loss are sub-Gaussian random variables ,*

$$\|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i)\|_{\psi_2} \leq 4\max\{\nu_1, \nu_2, M\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}, M\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\}$$

Similarly,

$$\begin{aligned}
\|Y_i - g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i)\|_{\psi_2} &\leq 2\max\{\nu_1, M\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\},\\
\|g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\|_{\psi_2} &\leq M\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2},\\
\|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)\|_{\psi_2} &\leq M\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2},\\
\|\rho\cdot g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) + (1-\rho)\cdot g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) - Y\|_{\psi_2} &\leq 4\max\{(1-\rho)\nu_1, \rho\nu_2,\\
&\qquad \rho M\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}, (1-\rho)M\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\},\\
\|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) + g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\|_{\psi_2} &\leq \sqrt{2}M\sigma_{\mathsf{max}}\max\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\}.
\end{aligned}$$

Proof [Proof of Lemma 14] To establish the sub-exponential tail, we consider the following decomposition

$$\begin{aligned}
g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) =& \{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - Y_i\} - \{g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\}\\
&+ \{g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)\} - \{g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\}.
\end{aligned} \tag{A.37}$$

According to Assumption 1a, the first two terms on the right-hand side of (A.37) are sub-Gaussian,

$$\|g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - Y_i\|_{\psi_2} \leq \nu_1,\ \|g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\|_{\psi_2} \leq \nu_2.$$

According to Assumption 1b, the latter two terms on the right-hand side of (A.37) are bounded by

$$|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)| \leq M|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\intercal \mathbf{X}_i|,\ |g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)| \leq M|(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\intercal \mathbf{W}_i|.$$

Under Assumption 2a, $(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\intercal \mathbf{X}_i$ and $(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\intercal \mathbf{W}_i$ are sub-Gaussian random variables,

$$\begin{aligned}
\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\intercal \mathbf{X}_i\|_{\psi_2} &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\\
\|(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\intercal \mathbf{W}_i\|_{\psi_2} &\leq \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}.
\end{aligned}$$

By Lemma 17-e,

$$\begin{aligned}
\|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)\|_{\psi_2} &\leq M\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\intercal \mathbf{X}_i\|_{\psi_2} \leq M\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\\
\|g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\|_{\psi_2} &\leq \|(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\intercal \mathbf{W}_i\|_{\psi_2} \leq M\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}.
\end{aligned}$$

Finally, we apply Lemma 17-d

$$\begin{aligned}
\|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i)\|_{\psi_2} \leq& 4\max\big\{\|g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - Y_i\|_{\psi_2}, \|g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\|_{\psi_2},\\
&\|g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i)\|_{\psi_2}, \|g(\boldsymbol{\gamma}^\intercal \mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\|_{\psi_2}\big\}\\
\leq& 4\max\left\{\nu_1, \nu_2, M\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}, M\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2\sigma_{\mathsf{max}}/\sqrt{2}\right\}.
\end{aligned}$$

Therefore, we have reached the conclusion.

We may obtain the rest of bounds following the same derivation. ■

C2 Inference

Analysis of Estimated Precision Matrix

Proof [Proof of Lemma 13]

The definition of the cross-fitted loss functions $m^{(k,k')}$ and their derivatives can be found at (A.23). By the definition of $\widehat{\mathbf{u}}^{(k)}$, we have

$$\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}m^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)};\widehat{\boldsymbol{\beta}}^{(k,k')}\right)+\lambda_{\mathsf{u}}\|\widehat{\mathbf{u}}^{(k)}\|_1\leq\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}m^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)+\lambda_{\mathsf{u}}\|\mathbf{u}_0\|_1.$$

Denote the standardized estimation error as $\boldsymbol{\delta}=(\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0)/\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2$. Due to convexity of the loss function, we have for $t=\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\wedge 1$

$$\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}m^{(k,k')}\left(\mathbf{u}_0+t\boldsymbol{\delta};\widehat{\boldsymbol{\beta}}^{(k,k')}\right)+\lambda_{\mathsf{u}}\|\mathbf{u}_0+t\boldsymbol{\delta}\|_1\leq\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}m^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)+\lambda_{\mathsf{u}}\|\mathbf{u}_0\|_1. \tag{A.38}$$

By the triangle inequality $\|\mathbf{u}_0\|_1-\|\mathbf{u}_0+t\boldsymbol{\delta}\|_1\leq t\|\boldsymbol{\delta}\|_1$, we have from (A.38)

$$\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}\left\{m^{(k,k')}\left(\mathbf{u}_0+t\boldsymbol{\delta};\widehat{\boldsymbol{\beta}}^{(k,k')}\right)-m^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\right\}\leq t\lambda_{\mathsf{u}}\|\boldsymbol{\delta}\|_1 \tag{A.39}$$

Because the loss functions $m^{(k)}$ are quadratic functions of $\mathbf{u}$, we can apply the restricted strong convexity event $\Omega^{(k)}$ to obtain

$$\begin{aligned}
&m^{(k,k')}\left(\mathbf{u}_0+t\boldsymbol{\delta};\widehat{\boldsymbol{\beta}}^{(k,k')}\right)-m^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)-t\boldsymbol{\delta}^{\mathsf{T}}\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\\
=&t^2\boldsymbol{\delta}^{\mathsf{T}}\ddot{\mathrm{m}}^{(k,k')}\left(\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\boldsymbol{\delta}\\
\geq&t^2\kappa^*_{\mathsf{rsc},1}-t\kappa^*_{\mathsf{rsc},1}\kappa^*_{\mathsf{rsc},2}\sqrt{\log(p)/N_{k'}}\|\boldsymbol{\delta}\|_1.
\end{aligned} \tag{A.40}$$

Applying (A.40) to (A.39), we have with large probability

$$\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}\left\{t\boldsymbol{\delta}^{\mathsf{T}}\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)+t^2\kappa^*_{\mathsf{rsc},1}-t\kappa^*_{\mathsf{rsc},1}\kappa^*_{\mathsf{rsc},2}\sqrt{\log(p)/N_{k'}}\|\boldsymbol{\delta}\|_1\right\}\leq t\lambda_{\mathsf{u}}\|\boldsymbol{\delta}\|_1$$

where $\|\boldsymbol{\delta}\|_2=1$ from definition. Thus, we have reach

$$t\kappa^*_{\mathsf{rsc},1}\leq\lambda_{\mathsf{u}}\|\boldsymbol{\delta}\|_1-\sum_{k'\neq k}\frac{N_{k'}}{N-N_k}\left\{\boldsymbol{\delta}^{\mathsf{T}}\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)-\kappa^*_{\mathsf{rsc},1}\kappa^*_{\mathsf{rsc},2}\sqrt{\log(p)/N_{k'}}\|\boldsymbol{\delta}\|_1\right\}. \tag{A.41}$$

The target parameter $\mathbf{u}_0$ can be identify by $\mathbb{E}\left\{\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\mid\mathscr{D}^c_{k'}\right\}=0$. We use the fact to do a careful analysis of $\boldsymbol{\delta}^{\mathsf{T}}\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)$ by the decomposition

$$\begin{aligned}
\left|\boldsymbol{\delta}^{\mathsf{T}}\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\right|=&\boldsymbol{\delta}^{\mathsf{T}}\left[\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)-\mathbb{E}\left\{\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\mid\mathscr{D}^c_{k'}\right\}\right]\\
&+\boldsymbol{\delta}^{\mathsf{T}}\left[\mathbb{E}\left\{\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\widehat{\boldsymbol{\beta}}^{(k,k')}\right)\mid\mathscr{D}^c_{k'}\right\}-\mathbb{E}\left\{\dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0;\boldsymbol{\beta}_0\right)\mid\mathscr{D}^c_{k'}\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
\leq & \|\boldsymbol{\delta}\|_1 \left\| \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \mathbb{E}\left\{ \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \mid \mathscr{D}_{k'}^c \right\} \right\|_\infty \\
& + \left\| \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}[\mathbf{X}_i\mathbf{X}_i^\intercal \mathbf{u}_0\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)\} \mid \mathscr{D}_{k'}^c] \right\|_2 .
\end{aligned} \tag{A.42}
$$

We establish the rate for L_2-norm of the population score at $\mathbf{u}_0$ through analyzing

$$
\sup_{\|\mathbf{v}\|_2=1} \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}[\mathbf{v}^\intercal\mathbf{X}_i\mathbf{X}_i^\intercal \mathbf{u}_0\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)\} \mid \mathscr{D}_{k'}^c],
$$

whose bound can be derived from Assumptions 1b, 2a, 2a, the Cauchy-Schwartz inequality and Lemma 17-b,

$$
\begin{aligned}
& \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}[\mathbf{v}^\intercal\mathbf{X}_i\mathbf{X}_i^\intercal \mathbf{u}_0\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)\} \mid \mathscr{D}_{k'}^c] \\
\leq & M\mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}[|\mathbf{v}^\intercal\mathbf{X}_i\mathbf{X}_i^\intercal \mathbf{u}_0\{\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right)^\intercal \mathbf{X}_i\}| \mid \mathscr{D}_{k'}^c] \\
\leq & M\left[\mathbb{E}\{(\mathbf{v}^\intercal\mathbf{X}_i)^4\}\mathbb{E}\{(\mathbf{v}^\intercal\mathbf{X}_i)^4\}\right]^{1/4} \sqrt{\mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left[\left\{\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right)^\intercal \mathbf{X}_i\right\}^2 \mid \mathscr{D}_{k'}^c\right]} \\
\leq & M\sigma_{\mathsf{max}}^3\|\mathbf{v}\|_2\|\mathbf{u}_0\|_2\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2 .
\end{aligned}
$$

Hence, we have shown

$$
\left\|\mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left[\mathbf{X}_i\mathbf{X}_i^\intercal \mathbf{u}_0\left\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)\right\} \mid \mathscr{D}_{k'}^c\right]\right\|_2 \leq M\sigma_{\mathsf{max}}^3\|\mathbf{u}_0\|_2\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2 . \tag{A.43}
$$

By the bound for (A.42) through (A.43) and the definition of $\lambda_{\boldsymbol{\beta}}$, we have the bound from (A.41)

$$
t\kappa_{\mathsf{rsc},1}^* \leq 2\lambda_{\mathsf{u}}\|\boldsymbol{\delta}\|_1 + M\sigma_{\mathsf{max}}^3\|\mathbf{u}_0\|_2\sup_{k'\neq k}\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2 . \tag{A.44}
$$

Hence, we can reach an immediate bound for estimation error from (A.44) without considering the sparsity of $\mathbf{u}_0$. We shall proceed to derive a sharper bound that involves the sparsity of $\mathbf{u}_0$. We separately analyze two cases.

Case 1:

$$
M\sigma_{\mathsf{max}}^3\|\mathbf{u}_0\|_2\sup_{k=1,\dots,K}\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2 \geq \|\boldsymbol{\delta}\|_1\lambda_{\mathsf{u}}/3
$$

In this case, the estimation error is dominated by $\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}$. We simply have from (A.44)

$$
\begin{aligned}
t\kappa_{\mathsf{rsc},1}^* &\leq 7M\sigma_{\mathsf{max}}^3\|\mathbf{u}_0\|_2\sup_{k=1,\dots,K}\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2 , \\
t\kappa_{\mathsf{rsc},1}^*\|\boldsymbol{\delta}\|_1\lambda_{\mathsf{u}}/3 &\leq 7M\sigma_{\mathsf{max}}^6\|\mathbf{u}_0\|_2^2\sup_{k=1,\dots,K}\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2^2 .
\end{aligned}
$$

Thus, we have

$$
\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 \leq 7M\sigma_{\mathsf{max}}^3\|\mathbf{u}_0\|_2\sup_{k'\neq k}\left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2/\kappa_{\mathsf{rsc},1}^*,
$$

$$\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_1 \le 21 M\sigma_{\mathsf{max}}^6 \|\mathbf{u}_0\|_2^2 \sup_{k'\neq k} \left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2^2 / (\kappa_{\mathsf{rsc},1}^* \lambda_{\mathsf{u}}). \tag{A.45}$$

Case 2:

$$M\sigma_{\mathsf{max}}^3 \|\mathbf{u}_0\|_2 \sup_{k'\neq k} \left\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')}\right\|_2 \le \|\boldsymbol{\delta}\|_1 \lambda_{\mathsf{u}}/3 \tag{A.46}$$

In this case, the estimation error is comparable to the situation that we have the true $\boldsymbol{\beta}_0$ for the Hessian. Thus, the sparsity of $\mathbf{u}_0$ may affect the estimation error.

Following the typical approach to establish the cone condition for $\boldsymbol{\delta}$, we analyze the symmetrized Bregman's divergence,

$$\begin{aligned}
&(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \left\{ \dot{\mathbf{m}}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \right\} \\
=&\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \boldsymbol{\delta}^\intercal \left\{ \dot{\mathbf{m}}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \right\}.
\end{aligned} \tag{A.47}$$

Due to the convexity of the quadratic loss $m^{(k,k')}\left(\cdot; \widehat{\boldsymbol{\beta}}^{(k,k')}\right)$, the symmetrized Bregman's divergence (A.47) is nonnegative through a mean-value theorem,

$$\begin{aligned}
&(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \left\{ \dot{\mathbf{m}}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \right\} \\
=&\sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \sum_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}} g'(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i)\{(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\}^2 \\
\ge&0.
\end{aligned}$$

Denote the indices set of nonzero coefficient in $\mathbf{u}_0$ as $\mathcal{O}_{\mathsf{u}} = \{j : \mathbf{u}_{0,j} \neq 0\}$. We denote the $\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}}$ and $\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}^{\mathsf{c}}}$ as the sub-vectors for $\boldsymbol{\delta}$ at positions in $\mathcal{O}_{\mathsf{u}}$ and at positions not in $\mathcal{O}_{\mathsf{u}}$, respectively. The solution $\widehat{\mathbf{u}}^{(k)}$ satisfies the KKT condition

$$\left\| \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \dot{\mathbf{m}}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \right\|_\infty \le \lambda_{\mathsf{u}},$$

$$\sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \dot{m}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right)_j = -\lambda_{\mathsf{u}} \operatorname{sign}(\widehat{u}_j^{(k)}),\ j : \widehat{u}_j^{(k)} \neq 0.$$

From the KKT condition and the definitions of $\boldsymbol{\delta}$ and $\mathcal{O}_{\mathsf{u}}$, we have

$$\begin{aligned}
\delta_j \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \dot{m}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right)_j &\le |\delta_j|\lambda_{\mathsf{u}},\ j \in \mathcal{O}_{\mathsf{u}}; \\
\delta_j \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \dot{m}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right)_j &= \frac{-\widehat{u}_j^{(k)} \lambda_{\mathsf{u}} \operatorname{sign}(\widehat{u}_j^{(k)})}{\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2} = -\lambda_{\mathsf{u}}|\delta_j|,\ j \in \mathcal{O}_{\mathsf{u}}^{\mathsf{c}}.
\end{aligned} \tag{A.48}$$

Applying the (A.48) to (A.47), we have the upper bound,

$$\boldsymbol{\delta}^\intercal \sum_{k'\neq k} \frac{N_{k'}}{N - N_k} \left\{ \dot{\mathbf{m}}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) - \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \right\}$$

$$
\begin{aligned}
&= \sum_{j\in\mathcal{O}_{\mathsf{u}}} \delta_j \sum_{k'\neq k} \frac{N_{k'}}{N-N_k} \dot{m}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right)_j + \sum_{j\in\mathcal{O}_{\mathsf{u}}^{\mathsf{c}}} \delta_j \sum_{k'\neq k} \frac{N_{k'}}{N-N_k} \dot{m}^{(k,k')}\left(\widehat{\mathbf{u}}^{(k)}; \widehat{\boldsymbol{\beta}}^{(k,k')}\right)_j \\
&\quad + \sum_{k'\neq k} \frac{N_{k'}}{N-N_k} \boldsymbol{\delta}^{\mathsf{T}} \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \\
&\leq \lambda_{\mathsf{u}} \sum_{j\in\mathcal{O}_{\mathsf{u}}} |\delta_j| - \lambda_{\mathsf{u}} \sum_{j\in\mathcal{O}_{\mathsf{u}}^{\mathsf{c}}} |\delta_j| + \boldsymbol{\delta}^{\mathsf{T}} \dot{\boldsymbol{\ell}}^{\dagger}(\boldsymbol{\beta}_0; \widehat{\boldsymbol{\gamma}}) \\
&\leq \lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}}\|_1 - \lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}^{\mathsf{c}}}\|_1 + \left| \sum_{k'\neq k} \frac{N_{k'}}{N-N_k} \boldsymbol{\delta}^{\mathsf{T}} \dot{\mathbf{m}}^{(k,k')}\left(\mathbf{u}_0; \widehat{\boldsymbol{\beta}}^{(k,k')}\right) \right|.
\end{aligned}
$$

Then, we apply (A.42), the definition of λ_{u} and (A.46),

$$
0 \leq \lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}}\|_1 - \lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}^{\mathsf{c}}}\|_1 + \frac{2}{3}\lambda_{\mathsf{u}} \|\boldsymbol{\delta}\|_1, \quad \text{and} \quad \lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}^{\mathsf{c}}}\|_1 \leq 5\lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}}\|_1.
$$

Therefore, we can bound the L_1 norm of $\boldsymbol{\delta}$ by the cone property,

$$
\|\boldsymbol{\delta}\|_1 \leq 6\lambda_{\mathsf{u}} \|\boldsymbol{\delta}_{\mathcal{O}_{\mathsf{u}}}\|_1 \leq 6\sqrt{s_{\mathsf{u}}} \|\boldsymbol{\delta}\|_2 = 6\sqrt{s_{\mathsf{u}}}. \tag{A.49}
$$

Now, we apply the cone condition (A.49) and the case condition (A.46) to the bound (A.44),

$$
t\kappa^*_{\mathsf{rsc},1} \leq 14\sqrt{s_{\mathsf{u}}}\lambda_{\mathsf{u}}, \; t\kappa^*_{\mathsf{rsc},1} \|\boldsymbol{\delta}\|_1 \leq 84 s_{\mathsf{u}} \lambda_{\mathsf{u}}
$$

Thus, we obtain the rate for estimation error

$$
\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 \leq 14\sqrt{s_{\mathsf{u}}}\lambda_{\mathsf{u}}/\kappa^*_{\mathsf{rsc},1}, \; \|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_1 \leq 84 s_{\mathsf{u}} \lambda_{\mathsf{u}}/\kappa^*_{\mathsf{rsc},1}. \tag{A.50}
$$

Conclusion:

Since Case 1 and Case 2 are the complement of each other, one of them must occur. Thus, the bound of estimation error is controlled by the larger bound in the two cases,

$$
\begin{aligned}
\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 &\leq \max\left\{ 14\sqrt{s_{\mathsf{u}}}\lambda_{\mathsf{u}}/\kappa^*_{\mathsf{rsc},1}, 7M\sigma^3_{\mathsf{max}} \|\mathbf{u}_0\|_2 \sup_{k'\neq k} \left\| \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')} \right\|_2 / \kappa^*_{\mathsf{rsc},1} \right\}, \\
\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_1 &\leq \max\left\{ 84 s_{\mathsf{u}}\lambda_{\mathsf{u}}/\kappa^*_{\mathsf{rsc},1}, 21M\sigma^6_{\mathsf{max}} \|\mathbf{u}_0\|_2^2 \sup_{k'\neq k} \left\| \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{(k,k')} \right\|_2^2 / (\kappa^*_{\mathsf{rsc},1}\lambda_{\mathsf{u}}) \right\},
\end{aligned}
$$

which is our oracle inequality. ■

Analysis for Terms T_1-T_5 in Part 1

To show

$$
T_1 = \sqrt{n}\left\{ \mathbf{x}_{\text{std}} - \mathbb{H}(\widetilde{\boldsymbol{\beta}})\mathbf{u}_0 \right\} (\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0) = O_p\left(\sqrt{n} \|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2^2 \right),
$$

we rewrite the term as a conditional expectation

$$
T_1 = \sqrt{n}\mathbf{u}_0^{\mathsf{T}} \left\{ \boldsymbol{\Sigma}_0 - \mathbb{H}(\widetilde{\boldsymbol{\beta}}) \right\} (\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0)
$$

$$=\sqrt{n}\mathbb{E}_{i\in\mathcal{J}_k}\left[\mathbf{u}_0^\intercal\mathbf{X}_i(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i\{g'(\widetilde{\boldsymbol{\beta}}^\intercal\mathbf{X}_i)-g'(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)\}\mid\mathscr{D}_k^c\right].$$

Under Assumptions 1b, 2a, we derive the bound for the expectation using the Cauchy-Schwartz inequality and Lemma 17-b,

$$\begin{aligned}
|T_1|\leq& M\sqrt{n}\mathbb{E}_{i\in\mathcal{J}_k}\left[\mathbf{u}_0^\intercal\mathbf{X}_i\{(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i\}^2\mid\mathscr{D}_k^c\right]\\
\leq& M\sqrt{n\mathbb{E}_{i\in\mathcal{J}_k}\left\{(\mathbf{u}_0^\intercal\mathbf{X}_i)^2\mid\mathscr{D}_k^c\right\}\mathbb{E}_{i\in\mathcal{J}_k}\left[\{(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i\}^4\mid\mathscr{D}_k^c\right]}\\
\leq& M\sqrt{n8\|\mathbf{u}_0^\intercal\mathbf{X}_i\|_{\psi_2}^2\|(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i\|_{\psi_2}^4}\\
\leq&\sqrt{n}M\|\mathbf{u}_0\|_2\left\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\right\|_2^2\sigma_{\mathsf{max}}^3.
\end{aligned}\tag{A.51}$$

Since $\|\mathbf{u}_0\|_2$ is bounded according to (A.25), we have established in

$$|T_1|=O_p\left(\sqrt{n}\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2^2\right)$$

as declared.

To show

$$T_2=\sqrt{n}\left(\mathbf{u}_0-\widehat{\mathbf{u}}^{(k)}\right)^\intercal\mathbb{H}(\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)=O_p\left(\sqrt{n}\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\right),$$

we rewrite the term as a conditional expectation

$$T_2=\sqrt{n}\mathbb{E}_{i\in\mathcal{J}_k}\left[\left(\mathbf{u}_0-\widehat{\mathbf{u}}^{(k)}\right)^\intercal\mathbf{X}_i(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_ig'(\widetilde{\boldsymbol{\beta}}^\intercal\mathbf{X}_i)\mid\mathscr{D}_k^c\right].$$

Similar to (A.51), we derive the bound for the expectation under Assumptions 1b, 2a through the Cauchy-Schwartz inequality and Lemma 17-b, 17-f,

$$\begin{aligned}
|T_2|\leq& M\sqrt{n}\mathbb{E}_{i\in\mathcal{J}_k}\left[|\left(\mathbf{u}_0-\widehat{\mathbf{u}}^{(k)}\right)^\intercal\mathbf{X}_i(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i|\mid\mathscr{D}_k^c\right]\\
\leq&2M\sqrt{n}\|\left(\mathbf{u}_0-\widehat{\mathbf{u}}^{(k)}\right)^\intercal\mathbf{X}_i(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i\|_{\psi_1}\\
\leq&2M\sqrt{n}\|\left(\mathbf{u}_0-\widehat{\mathbf{u}}^{(k)}\right)^\intercal\mathbf{X}_i\|_{\psi_2}\|(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0)^\intercal\mathbf{X}_i\|_{\psi_2}\\
\leq& M\sqrt{n}\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2\sigma_{\mathsf{max}}^2.
\end{aligned}$$

This bound immediately implies

$$T_2=O_p\left(\sqrt{n}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\right).$$

To show

$$\begin{aligned}
T_3=&\sqrt{n}\widehat{\mathbf{u}}^{(k)\intercal}\left[\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)})\mid\mathscr{D}_k^c\}-\left\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\widehat{\boldsymbol{\beta}}^{(k)};\widehat{\boldsymbol{\gamma}}^{(k)})-\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0)\right\}\right]\\
=&O_p\left(\rho\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2+\sqrt{\rho(1-\rho)}\|\widehat{\boldsymbol{\gamma}}^{(k)}-\boldsymbol{\gamma}_0\|_2\right),
\end{aligned}$$

we rewrite the term as two empirical processes with diminishing summands

$$
\begin{aligned}
T_3 = & -\sqrt{n}\frac{1}{N_k}\sum_{i\in\mathcal{J}_k}\Bigg(\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) + g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\} \\
& - \mathbb{E}_{i\in\mathcal{J}_k}\left[\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) + g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\} \mid \mathscr{D}_k^c\right]\Bigg) \\
& - \sqrt{n}\frac{1}{N_k}\sum_{i\in\mathcal{I}_k}\Bigg(\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i)\} \\
& - \mathbb{E}_{i\in\mathcal{J}_k}\left[\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i)\} \mid \mathscr{D}_k^c\right]\Bigg).
\end{aligned}
$$

We have used the identity $\mathbb{E}\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0) \mid \mathscr{D}_k^c\} = 0$ above. Using Lemmas 14, 17-h and Assumptions (2a) and (2a), we show that each summand is sub-exponential

$$
\begin{aligned}
&\|\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) + g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\}\|_{\psi_1} \\
\le & \|\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\|_{\psi_2}\|g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i) - g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) + g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\|_{\psi_2} \\
\le & M\sigma_{\mathsf{max}}^2\|\widehat{\mathbf{u}}^{(k)}\|_2\left(\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 + \|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right), \\
&\|\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\beta}}^{(k)\mathsf{T}}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i)\}\|_{\psi_1} \le M\sigma_{\mathsf{max}}^2\|\widehat{\mathbf{u}}^{(k)}\|_2\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2/2.
\end{aligned}
$$

Applying the Bernstein's inequality, we obtain

$$
T_3 = O_p\left(\|\widehat{\mathbf{u}}^{(k)}\|_2\left\{\sqrt{\rho}\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 + \sqrt{\rho(1-\rho)}\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right\}\right).
$$

We achieve the stated rate with the tightness of $\|\widehat{\mathbf{u}}^{(k)}\|_2$ from (A.25).

To show

$$
\begin{aligned}
T_4 = & \sqrt{n}(1-\rho)\widehat{\mathbf{u}}^{(k)\mathsf{T}}\left[\mathbb{E}\{\dot{\boldsymbol{\ell}}_{\mathsf{imp}}^{(k)}(\widehat{\boldsymbol{\gamma}}^{(k)}) \mid \mathscr{D}_k^c\} - \left\{\dot{\boldsymbol{\ell}}_{\mathsf{imp}}^{(k)}(\widehat{\boldsymbol{\gamma}}^{(k)}) - \dot{\boldsymbol{\ell}}_{\mathsf{imp}}^{(k)}(\boldsymbol{\gamma}_0)\right\}\right] \\
= & O_p\left((1-\rho)\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right),
\end{aligned}
$$

we rewrite the term as the empirical process with diminishing summands

$$
\begin{aligned}
T_4 = & -\sqrt{n}(1-\rho)\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\Bigg(\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\} \\
& - \mathbb{E}_{i\in\mathcal{J}_k}\left[\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\} \mid \mathscr{D}_k^c\right]\Bigg).
\end{aligned}
$$

We have used the identity $\mathbb{E}\{\dot{\boldsymbol{\ell}}_{\mathsf{imp}}^{(k)}(\boldsymbol{\gamma}_0) \mid \mathscr{D}_k^c\} = 0$ above. Similar to the analysis of T_3, we show that each summand is sub-exponential

$$
\|\widehat{\mathbf{u}}^{(k)\mathsf{T}}\mathbf{X}_i\{g(\widehat{\boldsymbol{\gamma}}^{(k)\mathsf{T}}\mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{W}_i)\}\|_{\psi_1} \le M\sigma_{\mathsf{max}}^2\|\widehat{\mathbf{u}}^{(k)}\|_2\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2/2.
$$

Applying the Bernstein's inequality, we obtain

$$
T_4 = O_p\left((1-\rho)\|\widehat{\mathbf{u}}^{(k)}\|_2\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right).
$$

We achieve the stated rate with the tightness of $\|\widehat{\mathbf{u}}^{(k)}\|_2$ from (A.25).

To show

$$T_5 = \sqrt{n}\left(\mathbf{u}_0 - \widehat{\mathbf{u}}^{(k)}\right)^\intercal \left\{\dot{\boldsymbol{\ell}}^{\dagger(k)}(\boldsymbol{\beta}_0;\boldsymbol{\gamma}_0) + (1-\rho)\dot{\boldsymbol{\ell}}^{(k)}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)\right\} = O_p\left(\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2\right),$$

we rewrite the term as the empirical process with diminishing summands

$$\begin{aligned} T_5 = &- \sqrt{n}\frac{1}{N_k}\sum_{i\in\mathcal{J}_k}(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\} \\ &- \sqrt{n}\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\{\rho\cdot g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) + (1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\}. \end{aligned}$$

The summands have zero mean because

$$\begin{aligned} &\mathbb{E}_{i\in\mathcal{J}_k}\left[(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\} \mid \mathscr{D}_k^c\right] \\ =&(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbb{E}\left[\mathbf{X}_i\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\}\right] \\ =&0, \\ &\mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\{\rho\cdot g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) + (1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\} \mid \mathscr{D}_k^c\right] \\ =&(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \left(\rho\mathbb{E}\left[\mathbf{X}_i\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - Y_i\}\right] + (1-\rho)\mathbb{E}\left[\mathbf{X}_i\{g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\}\right]\right) \\ =&0. \end{aligned}$$

Similar to the analysis of T_3, we show that each summand is sub-exponential

$$\begin{aligned} \left\|(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\{g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) - g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i)\}\right\|_{\psi_1} &\le \sqrt{2}\sigma_{\mathsf{max}}(\nu_1\vee\nu_2)\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 \\ \left\|(\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0)^\intercal \mathbf{X}_i\{\rho\cdot g(\boldsymbol{\beta}_0^\intercal \mathbf{X}_i) + (1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal \mathbf{W}_i) - Y_i\}\right\|_{\psi_1} &\le \sqrt{2}\sigma_{\mathsf{max}}(\nu_1\vee\nu_2)\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 \end{aligned}$$

Applying the Bernstein's inequality, we obtain

$$T_5 = O_p\left(\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2\left\{\sqrt{\rho(1-\rho)}+1\right\}\right) = O_p\left(\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2\right).$$

Analysis for Terms T_1'-T_4' in Part 2

Conditionally on the out-of-fold data, the term T_1' is the empirical average of i.i.d. mean zero random variables,

$$\begin{aligned} T_1 =&\sum_{k=1}^K \frac{n_k}{n}\left(\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i) + \rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - Y_i\}^2\right. \\ &\left. - \mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i) + \rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - Y_i\}^2 \mid \mathscr{D}_k^c\right]\right). \end{aligned}$$

We bound the variance of each summand by the Cauchy-Schwartz inequality and Lemmas 17-b, 17-d, 17-e,

$$\operatorname*{Var}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i) + \rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i) - Y_i\}^2 \mid \mathscr{D}_k^c\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^4\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^4 \mid \mathscr{D}_k^c\right]\\
&\leq \sqrt{\mathbb{E}_{i\in\mathcal{I}_k}\left\{(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^8 \mid \mathscr{L}_k^c\right\}\mathbb{E}_{i\in\mathcal{I}_k}\left[\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^8 \mid \mathscr{D}_k^c\right]}\\
&\leq \sqrt{48\|\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i\|_{\psi_2}^8 48\left(\rho\|g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\|_{\psi_2}\vee(1-\rho)\|g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{X}_i)-Y_i\|_{\psi_2}\right)^8}
\end{aligned}
\tag{A.52}
$$

Under Assumption 1a, 1b, 2a, we have

$$
\|\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i\|_{\psi_2}\leq(\|\mathbf{u}_0\|_2+\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2)\sigma_{\max}/\sqrt{2}=O_p\left(1+\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2\right).
$$

We apply Lemma 14 to obtain

$$
\begin{aligned}
\|g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\|_{\psi_2}&=O_p\left(1+\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2\right),\\
\|g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)-Y_i\|_{\psi_2}&=O_p\left(1+\|\widehat{\boldsymbol{\gamma}}^{(k)}-\boldsymbol{\gamma}_0\|_2\right).
\end{aligned}
$$

We have shown that the variance in (A.52) is of order

$$
O_p\left(1+\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2^4+\rho^4\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2^4+(1-\rho)^4\|\widehat{\boldsymbol{\gamma}}^{(k)}-\boldsymbol{\gamma}_0\|_2^4\right).
$$

Thus by the Tchebychev's inequality, we obtain

$$
T_1'=O_p\left(\left\{1+\|\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0\|_2+\rho\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\|_2+(1-\rho)\|\widehat{\boldsymbol{\gamma}}^{(k)}-\boldsymbol{\gamma}_0\|_2\right\}/\sqrt{n}\right)
$$

Applying the consistency of $\widehat{\boldsymbol{\gamma}}^{(k)}$, $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\mathbf{u}}^{(k)}$ from (A.22)

$$
T_1'=O_p\left(n^{-1/2}\right)=o_p(1).
$$

To analyze T_2', we consider the decomposition in which the estimators are replaced by the estimands one by one,

$$
\begin{aligned}
&T_2'\\
=&\sum_{k=1}^K\frac{n_k}{n}\Bigg(\mathbb{E}_{i\in\mathcal{I}_k}\left[(\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^2 \mid \mathscr{D}_k^c\right]\\
&\qquad-\mathbb{E}\left[(\mathbf{u}_0^\intercal\mathbf{X}_i)^2\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^\intercal\mathbf{W}_i)+\rho\cdot g(\boldsymbol{\beta}_0^\intercal\mathbf{X}_i)-Y_i\}^2\right]\Bigg)\\
=&\sum_{k=1}^K\frac{n_k}{n}\mathbb{E}_{i\in\mathcal{I}_k}\left[\{(\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0)^\intercal\mathbf{X}_i\}\widehat{\mathbf{u}}^{(k)\intercal}\mathbf{X}_i\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^2 \mid \mathscr{D}_k^c\right]\\
&+\sum_{k=1}^K\frac{n_k}{n}\mathbb{E}_{i\in\mathcal{I}_k}\left[\{(\widehat{\mathbf{u}}^{(k)}-\mathbf{u}_0)^\intercal\mathbf{X}_i\}\mathbf{u}_0^\intercal\mathbf{X}_i\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}^2 \mid \mathscr{D}_k^c\right]\\
&+\sum_{k=1}^K\frac{n_k}{n}\mathbb{E}_{i\in\mathcal{I}_k}\Big[(\mathbf{u}_0^\intercal\mathbf{X}_i)^2\{(1-\rho)\cdot g(\widehat{\boldsymbol{\gamma}}^{(k)\intercal}\mathbf{W}_i)+\rho\cdot g(\widehat{\boldsymbol{\beta}}^{(k)\intercal}\mathbf{X}_i)-Y_i\}
\end{aligned}
$$

$$
\begin{aligned}
&\quad \times [(1-\rho)\{g(\widehat{\boldsymbol{\gamma}}^{(k)\top}\mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\top\mathbf{W}_i)\} + \rho\{g(\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\top\mathbf{X}_i)\}] \mid \mathscr{D}_k^c\Big] \\
&+ \sum_{k=1}^K \frac{n_k}{n}\mathbb{E}_{i\in\mathcal{I}_k}\Big[(\mathbf{u}_0^\top\mathbf{X}_i)^2\{(1-\rho)\cdot g(\boldsymbol{\gamma}_0^\top\mathbf{W}_i) + \rho\cdot g(\boldsymbol{\beta}_0^\top\mathbf{X}_i) - Y_i\} \\
&\quad \times [(1-\rho)\{g(\widehat{\boldsymbol{\gamma}}^{(k)\top}\mathbf{W}_i) - g(\boldsymbol{\gamma}_0^\top\mathbf{W}_i)\} + \rho\{g(\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i) - g(\boldsymbol{\beta}_0^\top\mathbf{X}_i)\}] \mid \mathscr{D}_k^c\Big].
\end{aligned}
$$

Following the same calculation as in (A.52), we can bound the expectations

$$
T_2' = O_p\left(\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 + \rho\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 + (1-\rho)\|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right)
$$

Applying the consistency of $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ from Lemma 21, Theorem 1 and Part 1 in the proof of Theorem 7, we have established

$$
T_2' = o_p(1).
$$

Repeating the analyses for T_1' and T_2', we can show

$$
\begin{aligned}
T_3' &= O_p\left(\rho\sqrt{(1-\rho)/N}\left\{1 + \|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 + \|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 + \|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right\}\right) = o_p(1), \\
T_4' &= O_p\left(\rho(1-\rho)\left\{\|\widehat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 + \|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0\|_2 + \|\widehat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{\gamma}_0\|_2\right\}\right) = o_p(1)
\end{aligned}
$$

Appendix D. Additional Technical Details

D1 Definitions

We adopt the following definition of sub-Gaussian and sub-exponential random variables.

Definition 15 (Sub-Gaussian and Sub-Exponential Random Variables) *The sub-Gaussian parameter for a random variable V is defined as*

$$
\|V\|_{\psi_2} = \inf\left\{\sigma > 0 : \mathbb{E}(e^{V^2/\sigma^2}) \le 2\right\}.
$$

The random variable V is sub-Gaussian if $\|V\|_{\psi_2}$ is finite. The sub-Gaussian parameter for a random vector $\mathbf{U}$ is defined as

$$
\|\mathbf{U}\|_{\psi_2} = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{v}^\top\mathbf{U}\|_{\psi_2}.
$$

The sub-Gaussian parameter for a random variable V is defined as

$$
\|V\|_{\psi_1} = \inf\left\{\nu > 0 : \mathbb{E}(e^{|V|/\nu}) \le 2\right\}.
$$

The random variable V is sub-exponential if $\|V\|_{\psi_1}$ is finite. The more general Orlicz norm for $\alpha \in (0,1)$ is defined as

$$
\|V\|_{\psi_\alpha} = \inf\left\{\nu > 0 : \mathbb{E}\left[e^{(|V|/\nu)^\alpha}\right] \le 2\right\}.
$$

Mimicking the (minimal) Restricted Eigenvalue condition on the minimal eigenvalue of matrix over a cone (Bickel et al., 2009), we define the maximal Restricted Eigenvalue in Definition 16.

Definition 16 (Maximal Restricted Eigenvalue) *For a cone-set of the indices set $\mathcal{O} \subset \{1, \ldots, p\}$*

$$\mathcal{C}_\gamma(\xi, \mathcal{O}) := \left\{ v \in \mathbb{R}^{p+q+1} : \|v_{\mathcal{O}^c}\|_1 \leq \xi \|v_{\mathcal{O}}\|_1 \right\}, \tag{A.53}$$

we define the maximal Restricted Eigenvalue of a matrix Σ as

$$\mathrm{RE}_{\mathsf{max}}(\xi, \mathcal{O}; \Sigma) = \sup_{v \in \mathcal{C}_\gamma(\xi, \mathcal{O}) \backslash \{0\}} \frac{\sqrt{v^\intercal \Sigma v}}{\|v\|_2}. \tag{A.54}$$

D2 Statements of Existing Results

The properties in Lemmas 17 and 18 are covered in Vershynin (2018) Chapter 2 and 4.

Lemma 17 (Properties of sub-Gaussian and sub-exponential random variables)

a) Tail-probability:

$$\mathbb{P}(|V| \geq x) \leq 2e^{-x/\|V\|_{\psi_1}},$$
$$\mathbb{P}(|V| \geq x) \leq 2e^{-x^2/\|V\|_{\psi_2}^2};$$

b) Moments: $\mathbb{E}(|V|^r) \leq \min\{\kappa_{\psi,1}\|V\|_{\psi_1}^r, \kappa_{\psi,2}\|V\|_{\psi_2}^r\}$ with $\kappa_{\psi,1} = r!2$ and $\kappa_{\psi,2} = \Gamma(r/2)r$, and $\mathbb{E}(|V|) \leq \sqrt{\pi}\|V\|_{\psi_2}$;

c) Hierarchy: $\|V\|_{\psi_1} \leq \|V\|_{\psi_2}$;

d) Arbitrary addition: $\|\sum_{i=1}^m V_i\|_{\psi_2} \leq m \max_{i=1,\ldots,m} \|V_i\|_{\psi_2}$ and $\|\sum_{i=1}^m V_i\|_{\psi_1} \leq m \max_{i=1,\ldots,m} \|V_i\|_{\psi_1}$;

e) Multiplication with bounded random variable: $\|V_1V_2\|_{\psi_2} \leq \|V_1\|_{\psi_2}K$, $\|V_1V_2\|_{\psi_1} \leq \|V_1\|_{\psi_1}K$ for $|V_2| \leq K$ almost surely;

f) Multiplication between sub-Gaussian random variables: $\|V_1V_2\|_{\psi_1} \leq \|V_1\|_{\psi_2}\|V_2\|_{\psi_2}$, in particular, $\|V_1\|_{\psi_1} \leq \|V_1\|_{\psi_2}/\sqrt{\log(2)}$;

g) Hoeffding's inequality: $V_1, \ldots, V_m$ are independent mean zero sub-Gaussian random variables. For $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^m V_i\right| \geq t\right) \leq 4\exp\left(-\frac{t^2}{\kappa_{\psi,3}\sum_{i=1}^m \|V_i\|_{\psi_2}^2}\right),\ \kappa_{\psi,3} = 8.$$

h) Bernstein's inequality: $V_1, \ldots, V_m$ are independent mean zero sub-exponential random variables. For $t > 0$, $\kappa_{\psi,4} = 16$ and $\kappa_{\psi,5} = 4$

$$\mathbb{P}\left(\left|\sum_{i=1}^m V_i\right| \geq t\right) \leq 2\exp\left[-\min\left\{t^2\left(\kappa_{\psi,4}\sum_{i=1}^m \|V_i\|_{\psi_1}^2\right)^{-1}, t\left(\kappa_{\psi,5}\max_{i=1,\ldots,m}\|V_i\|_{\psi_1}\right)^{-1}\right\}\right].$$

Lemma 18 *Let $V_1, \ldots, V_m$ be i.i.d sub-Gaussian vectors in $\mathbb{R}^p$ such that*

$$\|v^\intercal V\|_{\psi_2}^2 \le K^2 \mathbb{E}\{(v^\intercal V)^2\}$$

for some $1 \le K < \infty$. Then,

$$\left\| \frac{1}{m} \sum_{i=1}^m V_i V_i^\intercal - \mathbb{E}(VV^\intercal) \right\|_2 = O_p\left(p/m + \sqrt{p/m}\right).$$

From Negahban et al. (2010) and Huang and Zhang (2012) among other literatures, we have the following results concerning the LASSO under the generalized linear models.

Lemma 19 *Under Assumptions 1b, 2a and 2b,*

$$\begin{aligned}
&\mathbb{P}\Bigg(\ell_{\mathsf{imp}}(\boldsymbol{\gamma}_0 + \boldsymbol{\Delta}) - \ell_{\mathsf{imp}}(\boldsymbol{\gamma}_0) - \boldsymbol{\Delta}^\intercal \dot{\boldsymbol{\ell}}_{\mathsf{imp}}(\boldsymbol{\gamma}_0) \\
&\qquad \ge \kappa_{\mathsf{rsc},1} \|\boldsymbol{\Delta}\|_2 \{\|\boldsymbol{\Delta}\|_2 - \kappa_{\mathsf{rsc},2} \sqrt{\log(p+q)/n} \|\boldsymbol{\Delta}\|_1\}, \forall \|\boldsymbol{\Delta}\|_2 \le 1 \Bigg) \ge 1 - \kappa_{\mathsf{rsc},3} e^{-\kappa_{\mathsf{rsc},4} n}; \\
&\mathbb{P}\Bigg(\ell_{\mathsf{PL}}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \ell_{\mathsf{PL}}(\boldsymbol{\beta}_0) - \boldsymbol{\Delta}^\intercal \dot{\boldsymbol{\ell}}_{\mathsf{PL}}(\boldsymbol{\beta}_0) \\
&\qquad \ge \kappa_{\mathsf{rsc},1} \|\boldsymbol{\Delta}\|_2 \{\|\boldsymbol{\Delta}\|_2 - \kappa_{\mathsf{rsc},2} \sqrt{\log(p)/N} \|\boldsymbol{\Delta}\|_1\}, \forall \|\boldsymbol{\Delta}\|_2 \le 1 \Bigg) \ge 1 - \kappa_{\mathsf{rsc},3} e^{-\kappa_{\mathsf{rsc},4} N}.
\end{aligned}$$

The negative log-likelihoods are defined in (8) *and* (10)*, and their gradients defined in* (12)*. See Definition 11 for the definition of conditional expectation notation. The constants are all absolute.*

The two inequalities in Lemma 19 are direct application of Negahban et al. (2010) Proposition 2 page 22. We can construct an auxiliary loss function to prove the following lemma.

Lemma 20 *Under Assumptions 1b, 2a and 2b,*

$$\begin{aligned}
&\mathbb{P}\Bigg(\frac{1}{N_{k'}} \sum_{i \in \mathcal{I}_{k'} \cup \mathcal{J}_{k'}} g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal} \mathbf{X}_i\right) (\boldsymbol{\Delta}^\intercal \mathbf{X}_i)^2 \\
&\qquad \ge 2\kappa_{\mathsf{rsc},1}^* \|\boldsymbol{\Delta}\|_2^2 - \kappa_{\mathsf{rsc},1}^* \kappa_{\mathsf{rsc},2}^* \sqrt{\log(p)/N} \|\boldsymbol{\Delta}\|_2 \|\boldsymbol{\Delta}\|_1, \forall \|\boldsymbol{\Delta}\|_2 \le 1 \Bigg) \\
\ge &\mathbb{P}\left(\left\| \widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_0 \right\|_2 \le \frac{\sigma_{\mathsf{min}}^2}{2\sigma_{\mathsf{max}}^3} \right) - \kappa_{\mathsf{rsc},3}^* e^{-\kappa_{\mathsf{rsc},4}^* N}.
\end{aligned}$$

The constants are all absolute.

Proof [Proof of Lemma 20] First, we show $\sqrt{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal} \mathbf{X}_i\right)} \mathbf{X}_i$ is a sub-Gaussian random vector whose second moment has all eigenvalues bounded away from infinity and zero. Under Assumptions 1b and 2a, we may apply Lemma 17-e,

$$\left\| \mathbf{v}^\intercal \sqrt{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal} \mathbf{X}_i\right)} \mathbf{X}_i \right\|_{\psi_2} \le \sqrt{M} \|\mathbf{v}^\intercal \mathbf{X}_i\|_{\psi_2} \le \sqrt{M} \sigma_{\mathsf{max}} \|\mathbf{v}\|_2 / \sqrt{2}.$$

Thus, $\sqrt{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)}\mathbf{X}_i$ is a sub-Gaussian random vector. Under Assumptions 1b and 2a, we can bound the maximal eigenvalue of its second moment,

$$\mathbf{v}^\intercal \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left\{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)\mathbf{X}_i\mathbf{X}_i^\intercal \mid \mathscr{D}_{k'}^c\right\}\mathbf{v} \leq M\mathbb{E}\{(\mathbf{v}^\intercal\mathbf{X}_i)^2\} \leq M\|\mathbf{v}\|_2^2\sigma_{\mathsf{max}}^2.$$

We derive the lower bound for the minimal eigenvalue of its second moment from Assumptions 1b, 2a, 2a, 2b-i, the Cauchy-Schwartz inequality and Lemma 17-b,

$$\begin{aligned}
&\mathbf{v}^\intercal \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left\{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)\mathbf{X}_i\mathbf{X}_i^\intercal \mid \mathscr{D}_{k'}^c\right\}\mathbf{v}\\
\geq&\mathbf{v}^\intercal \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\{g'(\boldsymbol{\beta}_0^\intercal\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^\intercal \mid \mathscr{D}_{k'}^c\}\mathbf{v}\\
&-\mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left[(\mathbf{v}^\intercal\mathbf{X}_i)^2\left\{g(\boldsymbol{\beta}_0^\intercal\mathbf{X}_i)-g\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)\right\} \mid \mathscr{D}_{k'}^c\right]\\
\geq&\|\mathbf{v}\|_2^2\sigma_{\mathsf{min}}^2 - M\mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left[\left|(\mathbf{v}^\intercal\mathbf{X}_i)^2\left\{\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}^{(k,k')}\right)^\intercal\mathbf{X}_i\right\}\right| \mid \mathscr{D}_{k'}^c\right]\\
\geq&\|\mathbf{v}\|_2^2\sigma_{\mathsf{min}}^2 - M\sqrt{\mathbb{E}\{(\mathbf{v}^\intercal\mathbf{X}_i)^4\}\mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left[\left\{\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}^{(k,k')}\right)^\intercal\mathbf{X}_i\right\}^2 \mid \mathscr{D}_{k'}^c\right]}\\
\geq&\|\mathbf{v}\|_2^2\left(\sigma_{\mathsf{min}}^2 - M\sigma_{\mathsf{max}}^3\left\|\widehat{\boldsymbol{\beta}}^{(k,k')}-\boldsymbol{\beta}_0\right\|_2\right).
\end{aligned}$$

Whenever $\left\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_0\right\|_2 \leq \frac{\sigma_{\mathsf{min}}^2}{2\sigma_{\mathsf{max}}^3}$, we have

$$\mathbf{v}^\intercal \mathbb{E}_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left\{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)\mathbf{X}_i\mathbf{X}_i^\intercal \mid \mathscr{D}_{k'}^c\right\}\mathbf{v} \geq \|\mathbf{v}\|_2^2\sigma_{\mathsf{min}}^2/2.$$

Second, we construct an auxiliary least square loss to apply Negahban et al. (2010). Let ε_i be independent standard normal random variables. Construct the loss function

$$\mathcal{L}^{(k,k')}(\mathbf{v}) = \frac{1}{N_{k'}}\sum_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}}\left\{\varepsilon_i + (\mathbf{v}_0-\mathbf{v})^\intercal\sqrt{g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)}\mathbf{X}_i\right\}^2.$$

By the design, we have

$$\mathcal{L}^{(k,k')}(\mathbf{v}_0+\boldsymbol{\Delta}) - \mathcal{L}^{(k,k')}(\mathbf{v}_0) - \boldsymbol{\Delta}^\intercal\frac{\partial}{\partial\mathbf{v}}\mathcal{L}^{(k,k')}(\mathbf{v}_0+\boldsymbol{\Delta}) = \frac{1}{N_{k'}}\sum_{i\in\mathcal{I}_{k'}\cup\mathcal{J}_{k'}} g'\left(\widehat{\boldsymbol{\beta}}^{(k,k')\intercal}\mathbf{X}_i\right)(\boldsymbol{\Delta}^\intercal\mathbf{X}_i)^2.$$

We apply Proposition 2 in Negahban et al. (2010) for $\mathcal{L}^{(k,k')}(\mathbf{v})$ conditionally on out-of-fold data $\mathscr{D}_{k'}^c$ and the event $\left\{\left\|\widehat{\boldsymbol{\beta}}^{(k,k')}-\boldsymbol{\beta}_0\right\|_2 \leq \frac{\sigma_{\mathsf{min}}^2}{2\sigma_{\mathsf{max}}^3}\right\}$ to finish the proof.

■

Lemma 21 *For a constant* $\kappa_{\mathsf{cone}}(n,p,q,\varepsilon_r) \asymp \sqrt{s_\gamma\log(p+q)/n}$*, the event*

$$\Omega_{\mathsf{cone}} = \left\{\|\dot{\boldsymbol{\ell}}_{\mathsf{imp}}(\boldsymbol{\gamma}_0)\|_\infty = \left\|\frac{1}{n}\sum_{i=1}^n\mathbf{W}_i\{g(\boldsymbol{\gamma}_0^\intercal\mathbf{W}_i)-Y_i\}\right\|_\infty \leq \kappa_{\mathsf{cone}}(n,p,q,\varepsilon_r)\right\}$$

occur with probability greater than $1-\varepsilon_r$ *under Assumptions 1a and 2a. Setting* $\lambda_\gamma = 2$*, we have on event* Ω_{cone} *that*

$$\widehat{\gamma} - \gamma_0 \in \mathcal{C}_\gamma(3, \operatorname{supp}(\gamma_0)) = \left\{ v \in \mathbb{R}^{p+q+1} : \|v_{\mathcal{O}_\gamma^c}\|_1 \leq 3\|v_{\mathcal{O}_\gamma}\|_1 \right\},$$

where $\mathcal{O}_\gamma = \{j : \gamma_j \neq 0\}$ *is the indices set for nonzero coefficient in* γ_0*. Moreover, we have*

$$\|\widehat{\gamma} - \gamma_0\|_2 = O_p\left(\sqrt{s_\gamma \log(p+q)/n}\right).$$

The concentration on the event Ω_{cone} is established by the union bound of element wise concentration, which is in turn obtained by the Bernstein inequality for sub-exponential random variables (Lemma 17-h). The rest of Lemma 21 follows Huang and Zhang (2012) Lemma 1 page 5 (page 1843 of the issue) and Negahban et al. (2010) Corollary 5 page 23.

References

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973, 2005.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika, 98(4):791–806, 2011.

Alexandre Belloni, Abhishek Kaul, and Mathieu Rosenbaum. Pivotal estimation via self-normalization for high-dimensional linear models with error in variables, 2017.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. Ann. Statist., 37(4):1705–1732, 08 2009.

Peter Bühlmann and Sara Van De Geer. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. The Annals of statistics, 45(2):615–646, 2017.

T. Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(2):391–419, 2020.

Tianxi Cai, Tony Cai, and Zijian Guo. Optimal Statistical Inference for Individualized Treatment Effects in High-dimensional Models. arXiv e-prints:1904.12891, 2019.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. Ann. Statist., 46(4):1541–1572, 08 2018.

Abhishek Chakrabortty, Jiarui Lu, T. Tony Cai, and Hongzhe Li. High Dimensional M-Estimation with Missing Outcomes: A Semi-Parametric Framework. arXiv e-prints:1911.11345, 2019.

Kabir Aladin Chandrasekher, Ahmed El Alaoui, and Andrea Montanari. Imputation for high-dimensional linear regression, 2020.

David Cheng, Ashwin Ananthakrishnan, and Tianxi Cai. Efficient and Robust Semi-Supervised Estimation of Average Treatment Effects in Electronic Medical Records Data. arXiv e-prints:1804.00195, 2018.

Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal semi-supervised estimation and inference for high-dimensional linear regression. arXiv e-prints:2011.14185, 2020.

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

Kelly A Frazer, Sarah S Murray, Nicholas J Schork, and Eric J Topol. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics, 10(4):241–251, 2009.

Jessica Gronsbell, Jessica Minnier, Sheng Yu, Katherine Liao, and Tianxi Cai. Automated feature selection of predictors in electronic medical records data. Biometrics, 75(1):268–277, 2019.

Zijian Guo, Prabrisha Rakshit, Daniel S Herman, and Jinbo Chen. Inference for the case probability in high-dimensional logistic regression. arXiv preprint:2012.07133, 2020.

Jian Huang and Cun-Hui Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. J. Mach. Learn. Res., 13(1):1839–1864, 2012.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research, 15:2869–2909, 2014.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv e-prints:2003.12408, 2020.

Katherine P Liao, Jiehuan Sun, and 18 others. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. Journal of the American Medical Informatics Association, 26(11):1255–1262, 08 2019.

Po-ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 24. Curran Associates, Inc., 2011.

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, 2020.

Anubha Mahajan, Daniel Taliun, and 113 others. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics, 50(11):1505–1513, 2018.

Andrew P. Morris, Benjamin F. Voight, Tanya M. Teslovich, Teresa Ferreira, Ayellet V. Segrè, Valgerdur Steinthorsdottir, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics, (9):981–990, 2012.

Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Technical Report 797, University of California Berkeley, Department of Statistics, 2010.

Stephen Portnoy. Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. i. consistency. Ann. Statist., 12(4):1298–1309, 12 1984.

Stephen Portnoy. Asymptotic behavior of m estimators of p regression parameters when p^2/n is large; ii. normal approximation. Ann. Statist., 13(4):1403–1417, 12 1985.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. IEEE Transactions on Information Theory, 57(10):6976–6994, 2011.

Robert A. Scott, Laura J. Scott, Reedik Mägi, Letizia Marullo, Kyle J. Gaulton, et al. An expanded genome-wide association study of type 2 diabetes in europeans. Diabetes, 66 (11):2888–2902, 2017.

Ezequiel Smucler, Andrea Rotnitzky, and James M. Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. arXiv e-prints:1904.03737, 2019.

Caroline A Thompson, Allison W Kurian, and Harold S Luft. Linking electronic health records to better understand breast cancer patient pathways within and between two health systems. eGEMs, 3(1), 2015.

A. Tsiatis. Semiparametric Theory and Missing Data. Springer Series in Statistics. Springer New York, 2007.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist., 42(3): 1166–1202, 06 2014.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. Electron. J. Statist., 3:1360–1392, 2009.

Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Benjamin F. Voight, Laura J. Scott, Valgerdur Steinthorsdottir, Andrew P. Morris, Christian Dina, Ryan P. Welch, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature Genetics, 42(7):579–589, 2010.

Marijana Vujkovic, Jacob M Keaton, and 48 others. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nature genetics, 52(7):680–691, 2020.

Joan L Warren and K Robin Yabroff. Challenges and opportunities in measuring cancer recurrence in the united states. Journal of the National Cancer Institute, 107(8):djv134, 2015.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. Ann. Statist., 47(5):2538–2566, 10 2019.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):217–242, 2014.

Yichi Zhang, Molei Liu, Matey Neykov, and Tianxi Cai. Prior adaptive semi-supervised learning with application to ehr phenotyping. Journal of Machine Learning Research, 23 (83):1–25, 2022.

Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. Biometrika, 109(2):387–403, 09 2021. ISSN 1464-3510. doi: 10.1093/biomet/asab042. URL `https://doi.org/10.1093/biomet/asab042`.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. Journal of the American Statistical Association, 113(524):1583–1600, 2018a.

Yinchu Zhu and Jelena Bradic. Significance testing in non-sparse high-dimensional linear models. Electron. J. Statist., 12(2):3312–3364, 2018b.