Journal of Machine Learning Research 27 (2026) 1-23 Submitted 9/23; Revised 7/26; Published 7/26

Efficient Modeling of Surrogates to Improve Multi-source High-dimensional Integrative Regression

Yue Liu* YUELIU@FAS.HARVARD.EDU
Department of Statistics
Harvard University
Cambridge, MA 02138, USA

Molei Liu* MOLEILIU@BJMU.EDU.CN
Peking University Health Science Center;
Beijing International Center for Mathematical Research;
Center for Statistical Science
Peking University
Haidian District, Beijing 100084, China

Zijian Guo ZIJGUO@ZJU.EDU.CN
Center for Data Science
Zhejiang University
Xixi District, Hangzhou 310058, China

Tianxi Cai TCAI.HSPH@GMAIL.COM
Department of Biostatistics
Harvard Chan School of Public Health
Boston, MA 02115, USA

Editor: Zaid Harchaoui

Abstract

Surrogate variables play an important role in various fields due to the scarcity or absence of gold-standard labels. We develop a novel approach named SASH for **S**urrogate-**A**ssisted and data-**S**hielding **H**igh-dimensional integrative regression. It is a semi-supervised approach that efficiently leverages sizable unlabeled samples with error-prone surrogate outcomes from multiple local sites to improve model estimation using the small gold-labeled sample. To facilitate stable and efficient knowledge extraction from the surrogates, our method first obtains a preliminary supervised estimator, and then uses it to assist in training a regularized single-index model (SIM) for the surrogates. Interestingly, through a chain of convex and properly penalized sparse regressions that approximate the SIM loss using bias correction, our method avoids the problem of local minima in the SIM, and fully eliminates the impact of the preliminary estimator's excessive error. In addition, it protects individual-level information through the aggregation of summary statistics from local sites, leveraging a similar idea of bias-corrected approximation. Through simulation studies, we demonstrate that our method outperforms existing approaches. Finally, we apply our method to develop a genetic risk model for type 2 diabetes using large-scale data sets from UK and Mass General Brigham biobanks, where only a small fraction of subjects in one site are labeled through manual chart review.

1. Yue Liu and Molei Liu contributed equally. Correspondence to: Molei Liu (moleiliu@bjmu.edu.cn).

©2026 Yue Liu, Molei Liu, 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/v27/23-1235.html.

Keywords: Electronic health records, Surrogate-assisted semi-supervised learning, Single index model, Federated learning, Sparse regression, One-step bias correction.

1. Introduction

1.1 Background

Electronic health records (EHRs) have become a crucial resource for a growing number of data-driven biomedical studies including disease diagnosis (Beaulieu-Jones et al., 2020; Zhang et al., 2020a), clinical knowledge extraction (Harnoune et al., 2021; Hong et al., 2021), treatment response evaluation (Franklin et al., 2021), and genotype-phenotype association characterization when linked with biobank data sets (Kohane, 2011; Cai et al., 2018; Maiorino and Loscalzo, 2023). However, realizing the potential of EHRs in data-driven research is challenging. This is because surrogate variables, such as diagnostic codes, are widely available but often noisy, failing to accurately reflect the true disease status and introducing bias (Zhang et al., 2019; Geva et al., 2021). Manual chart review, while accurate, is labor-intensive and yields only a small number of gold-standard labels, exacerbating challenges in analyzing high-dimensional data. This creates a trade-off: surrogates reduce variance due to their large sample sizes but introduce bias, whereas gold labels are unbiased but inflate variance due to their limited availability. To balance this trade-off, robust semi-supervised methods are needed to integrate large, noisy surrogate data sets with small samples with gold-standard labels efficiently.

In addition to EHR studies, surrogate variables also play an important role in diverse application fields. For example, in time-series forecasting, surrogates often represent intermediate data points collected between the observed present and the predicted future. They are useful for improving predictions in social and economic contexts by bridging the gap between initial and final observations (Box et al., 2015). In addition, various types of surrogates have been introduced for long-term user experience characterization in recommendation systems (Wang et al., 2022), automated vehicle safety modeling (Wang et al., 2021), and purchasing behavior prediction on e-commerce platforms (Wu et al., 2018).

When combining data across institutions, privacy constraints further complicate matters, as individual-level data cannot be directly shared. Frameworks like DataSHIELD (Wolfson et al., 2010) allow only the transfer of summary statistics, which, along with high-dimensional data, intensifies the issue of noisiness and label scarcity (e.g., Jones et al., 2012; Doiron et al., 2013; Gaye et al., 2014). Our real-world study aims to address the combined challenges of label scarcity, high-dimensionality, and privacy constraints.

1.2 Problem Setup

Suppose individual-level data are stored at M local sites indexed by $1, 2, \ldots, M$. In each site $m \in \{1, 2, \ldots, M\}$, there are N_m independent and identically distributed subjects with underlying full data $\mathscr{F}^{(\mathsf{m})} = \{(Y_i^{(\mathsf{m})}, S_i^{(\mathsf{m})}, \mathbf{X}_i^{(\mathsf{m})\mathsf{T}})^\mathsf{T} : i = 1, 2, \ldots, N_m\}$, where $\mathbf{X}_i^{(\mathsf{m})}$ is the p-dimensional covariate vector of subject i in site m, $Y_i^{(\mathsf{m})}$ denotes the binary true outcome of interest, and $S_i^{(\mathsf{m})}$ is some imperfect surrogate or proxy of $Y_i^{(\mathsf{m})}$.

We assume that the true label $Y_i^{(\mathsf{m})}$ is only observed for $m = 1$ and $i = 1, 2, ..., n$ with $n \ll \min(N_1, p)$ and denote the labeled subset of subjects in site 1 by $\mathscr{L}^{(1)} =$

$\{(Y_i^{(1)}, S_i^{(1)}, \mathbf{X}_i^{(1)\mathsf{T}})^\mathsf{T} : i = 1, 2, \ldots, n\}$, while for the remaining subjects in site 1 and all subjects from sites $2, \ldots, M$, we only observe their covariates and surrogates, and let $\mathscr{U}^{(\mathsf{m})} = \{(S_i^{(\mathsf{m})}, \mathbf{X}_i^{(\mathsf{m})\mathsf{T}})^\mathsf{T} : i = 1, 2, \ldots, N_m\}$ denote the observed unlabeled data for $m = 1, 2, \ldots, M$. For the true outcome $Y_i^{(\mathsf{m})}$, we are interested in estimating the coefficients in the logistic risk model

$$\mathbb{P}(Y_i^{(\mathsf{m})} = 1 \mid \mathbf{X}_i^{(\mathsf{m})}) = g(\alpha_0^{(\mathsf{m})} + \boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}) \text{ with } g(a) = e^a/(1+e^a), \quad \text{for } 1 \le m \le M, \tag{1}$$

where $\alpha_0^{(\mathsf{m})}$ is the intercept for site m and $\boldsymbol{\beta}_0$ represents the high-dimensional sparse model coefficients shared by the local sites. One could simply use ℓ_1-regularized logistic regression on the labeled data $\mathscr{L}^{(1)}$ to obtain a supervised estimator of $\boldsymbol{\beta}_0$. However, this may not perform well due to the small size of $\mathscr{L}^{(1)}$ and the high-dimensionality.

To estimate and infer $\boldsymbol{\beta}_0$ more efficiently, we leverage the large unlabeled samples from all sites with the surrogate $S_i^{(\mathsf{m})}$ that is predictive of $Y_i^{(\mathsf{m})}$ but imperfect and error-prone. Our main model assumption for the surrogate outcome $S_i^{(\mathsf{m})}$ is that it relates to $\mathbf{X}_i^{(\mathsf{m})}$ through a single index model (SIM):

$$\mathbb{E}(S_i^{(\mathsf{m})} \mid \mathbf{X}_i^{(\mathsf{m})}) = \breve{f}_m(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}) = f_m(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}) \quad \mathsf{for} \quad 1 \le m \le M, \tag{2}$$

where $\boldsymbol{\gamma}_0$ is a scaled vector with the *same direction* as $\boldsymbol{\beta}_0$ and $\breve{f}_m$ and f_m are some unknown and nuisance link functions varying across different sites.

The SIM assumption (2) holds in various practical scenarios such as the post hoc surrogate illustrated in diagram (i) on the right panel of Figure 1, which is frequently used in practice (e.g., Hong et al., 2019). In this case, $S_i^{(\mathsf{m})}$ is a proxy coming directly from the disease outcome $Y_i^{(\mathsf{m})}$ appearing between $S_i^{(\mathsf{m})}$ and the baseline risk factors $\mathbf{X}_i^{(\mathsf{m})}$. For example, in our real-world study, $\mathbf{X}_i^{(\mathsf{m})}$ is genetic variants and $S_i^{(\mathsf{m})}$ is the total count of diagnostic codes for the disease $Y_i^{(\mathsf{m})}$. The diagnostic counts $S_i^{(\mathsf{m})}$ are related to the disease-causing genetic variants only through the disease status, leading to the conditional independence:

$$S_i^{(\mathsf{m})} \perp\!\!\!\perp \mathbf{X}_i^{(\mathsf{m})} \mid Y_i^{(\mathsf{m})} \quad \text{for} \quad m = 1, 2, \ldots, M. \tag{3}$$

In Proposition 1 (justified in Appendix B.1), we show that our SIM assumption (2) holds under the logistic model assumption (1) and the conditional independence condition (3) that is reasonable for post hoc surrogates.

Proposition 1 *The surrogate SIM assumption (2) holds under conditions (1) and (3).*

Assumption (2) also holds in other contexts such as the early-endpoint surrogate described by diagram (ii) in Figure 1 and the example (iii). Early-endpoint surrogates in (ii) are frequently encountered in contemporary clinical and biomedical studies (Prentice, 1989; VanderWeele, 2013). They encompass factors that can be rapidly assessed and used as an indicator of the long-term outcome Y. For instance, the tumor response rate could be used as a surrogate for overall survival in clinical studies. Similar to Proposition 1, it is not hard to justify that such types of surrogates in (ii) and (iii) also follow the SIM (2) under the outcome model (1).

Note that $\boldsymbol{\beta}_0$ can be identified from (2) only up to scalar multiples. To ensure identifiability, we assume that $\beta_{01} \neq 0$ and take $\boldsymbol{\gamma}_0 = \boldsymbol{\beta}_0/\beta_{01}$, consequently $\gamma_{01} = 1$. For training

stability, one may set the first covariate in $\mathbf{X}$ as the risk factor believed to have a strong association with the outcome so that β_{01} is away from zero. This requires moderate prior knowledge of the model. For example, in our T2D genetic risk study, we set X_1 as the ethnicity variable since it has been found and confirmed that the risk of T2D is significantly associated with ethnicity (Abate and Chandalia, 2003).

Under the framework introduced above, we aim to derive an efficient estimator for $\boldsymbol{\beta}_0$ through semi-supervised learning assisted by the surrogate while overcoming the DataSHIELD constraint (Wolfson et al., 2010) that only summary statistics like mean vectors and covariance matrices are allowed to be transferred across the local sites. In our method, we will only transfer summary-level data between site 1 (with a few samples of Y) and the other sites. We illustrate this data-sharing scheme as well as the structure of our observed data sets in the left panel of Figure 1.

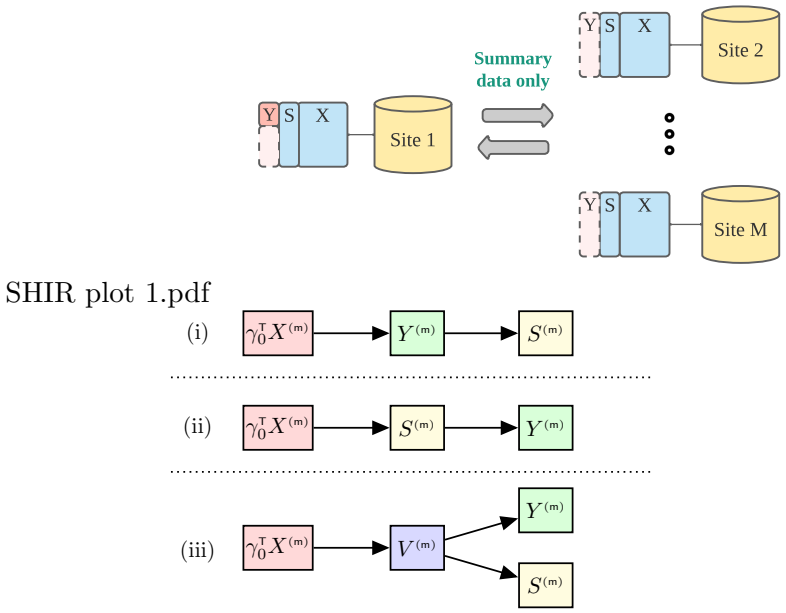

Figure 1: Left panel: an illustration of the structure of the observed data and the scheme of communication under DataSHIELD. Right panel: directed acyclic graphs (DAGs) of data generation examples. (i) Post hoc surrogates; (ii) Early-endpoint surrogates; (iii) An illustrative example in which the surrogate is neither post hoc nor early-endpoint but our SIM assumption (2) still holds. Here $V^{(m)}$ is some key latent factor mediating the effects of $X^{(m)}$ on both $S^{(m)}$ and $Y^{(m)}$.

1.3 Related Literature

Surrogates play an important role in various application fields where the collection of the primary or true outcome of interest is expensive, e.g., requiring intensive human efforts or long-term follow-up. With a partially observed true outcome Y and fully observed surrogate S, surrogate-assisted semi-supervised (SAS) learning methods have been proposed to

leverage unlabeled data to improve the learning accuracy of the true outcome model in the paucity or even absence of the gold-standard label Y. The key is to model and incorporate S in a way that could reduce the variance without introducing bias. For example, Huang et al. (2018), Zhang et al. (2020a), and Hong et al. (2019) used a maximum likelihood framework to address error-prone or anchor-positive labels. Kallus and Mao (2020) studied the semiparametric efficiency bounds for estimating the average treatment effects (ATE) in the presence of surrogates, to demonstrate the usefulness of surrogates in causal inference. Athey et al. (2019) and Athey et al. (2020) proposed methods that leverage observational data and secondary surrogates to mitigate the missingness of the long-term primary outcome Y in experimental data. Lin and Bradic (2021) addressed the noisiness of surrogate outcomes in deep learning through a novel algorithm curbing memorization of the noisy instances in neural networks.

For high-dimensional data, Zhang et al. (2022) proposed an adaptive SAS sparse regression approach that effectively leverages valid and informative surrogates while maintaining robustness to the surrogates of poor quality. They directly modeled S against a linear function of $\mathbf{X}$, which is valid for the SIM only when $\mathbf{X}$ satisfies a restrictive Gaussian condition. They also aim to model Y against both S and $\mathbf{X}$ for the purpose of EHR phenotyping. In contrast, we are interested in $Y \sim \mathbf{X}$ only, which is for risk prediction with the baseline factors $\mathbf{X}$, as well as identifying key risk factors. Recently, Hou et al. (2021) introduced a working imputation model for Y including both $\mathbf{X}$ and S as predictors and deployed it on unlabeled data to estimate the high-dimensional sparse $Y \sim \mathbf{X}$. Their method is guaranteed to improve statistical accuracy when the imputation model is sparser than $Y \sim \mathbf{X}$. To the best of our knowledge, there is no existing SAS method that is effective under a flexible SIM in (2) with a non-Gaussian design $\mathbf{X}$, let alone one that integrates multi-source SIMs under the DataSHIELD constraint. As will be discussed later, this methodological gap is due to the statistical and computational challenges of sparse SIM under non-Gaussian designs.

With the promise of more generalizable research findings through integrative analysis of multi-institutional data, much research in recent years has focused on federated learning that achieves information integration without sharing individual-level data. For example, He et al. (2016) utilized this strategy for sparse meta-analysis assuming that the model coefficients from local sites have similar supports. Under high-dimensional settings, Lee et al. (2017) and Battey et al. (2018) conducted communication-efficient distributed learning for the high-dimensional model by averaging and truncating the local debiased Lasso estimators (e.g., van de Geer et al., 2014). Cai et al. (2022b) proposed a federated learning method that accommodates model heterogeneity across local sites and complies with the DataSHIELD constraint by aggregating bias-corrected summary statistics. Jordan et al. (2019) proposed an approximate likelihood approach for communication-efficient aggregation of homogeneous data under the DataSHIELD constraint and Duan et al. (2022) extended their framework to address distributional heterogeneity. Nevertheless, all these methods require observing the outcome Y in all local sites. In our current set-up with only a few gold labels available in one site and abundant samples with the error-prone surrogate S in all sites, directly implementing existing methods for Y on the error-prone S can result in severe bias due to the error of S and the heterogeneity of $Y \mid S$ across the sites.

Technically, our work is also relevant to the estimation and inference of high-dimensional sparse SIM. Due to the unknown link function, it is much more challenging to study sparse

SIM under high-dimensional settings. A common strategy is to employ surrogate loss functions. Specifically, Neykov et al. (2016) and Eftekhari et al. (2021) used ℓ_1-regularized linear regression to learn the sparse coefficients in SIM. The validity of such a surrogate-loss approach heavily relies on the Gaussian (elliptical) distribution assumption of the design, which is often violated in practice. Under SIM with a non-Gaussian design, Radchenko (2015) used splines to model the unknown link function and gradient-based procedures to estimate coefficients. To the best of our knowledge, no existing work ensures convergence of the optimization algorithm for sparse SIM with a non-Gaussian design matrix.

Semiparametric efficient score functions have been frequently used to correct regularization bias in various inference problems such as those of high-dimensional generalized linear models (GLM) (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014), treatment and structural effects (e.g., Chernozhukov et al., 2018), and deep neural networks (e.g., Farrell et al., 2021). Recently, Wu et al. (2023) developed a debiasing method to infer the coefficient of the high-dimensional sparse SIM, which is motivated by the semiparametric efficient score for SIM coefficients derived in Liang et al. (2010).

More complicated than earlier work, the debiasing form of Wu et al. (2023) can simultaneously remove the first-order regularization bias of the parametric part and the smoothing bias of the nonparametric link function. This is because the unknown link in SIM is a nuisance model for the target coefficients. Our work essentially leverages this property to construct the one-step approximation for the non-convex SIM loss, which enables both the iterative updates towards the global solution and efficient data aggregation under the privacy constraints. A similar intrinsic relationship between debiased inference and efficient one-step approximation has been discussed in the context of high-dimensional GLMs (Cai et al., 2022b), but not studied for the SIM with a larger technical complication.

1.4 Our Contribution

To address limitations of existing approaches for sparse SIM and effectively leverage the surrogates to improve the learning efficiency on the outcome model in a multi-source setting, we propose a novel **S**urrogate-**A**ssisted and data-**S**hielding **H**igh-dimensional integrative regression (SASH) method. First, it extracts a preliminary estimator with the small labeled samples only, to initiate the training of high-dimensional SIM (2) at each local site. Interestingly, our use of the preliminary supervised estimator plays an important role in ensuring the desirable convergence of the sparse SIM regression while the impact of its error can be fully removed through proper iteration and regularization. Then SASH derives proper bias-corrected summary statistics to approximate the SIM loss in local sites, and aggregates them to obtain an accurate estimator for the direction coefficients of $\boldsymbol{\beta}$. Finally, it refits to the labeled data with the estimated direction to obtain the estimator for $\boldsymbol{\beta}$.

We show that under mild and reasonable assumptions, our SASH estimator attains the optimal convergence rate, outperforms the supervised and local analyses, and incurs no first-order convergence-rate loss relative to the ideal IPD estimator obtained by pooling the individual data. We also demonstrate that our method achieves better performance than recent SAS methods in simulation and real-world studies. This is because the above-reviewed SAS learning methods (e.g., Hou et al., 2021; Zhang et al., 2022) incorporate the surrogate S through fully parametric regression or unstable SIM-training procedures

that fail to characterize the semiparametric SIM (2) effectively and consequently cause efficiency loss. Technically, this limitation is mainly due to the convergence problem caused by local minima in single-index regression under non-Gaussian designs, which is even more pronounced in high-dimensional settings.

To overcome the convergence issue of the sparse SIM, we introduce a preliminary supervised estimator to guide the training, followed by two key innovative components: an efficient (bias-corrected) one-step approximation for the loss of SIM and an iterative updating algorithm with carefully tuned parameters. The resulting semi-supervised estimator achieves an error rate independent of the small labeled sample size n, instead leveraging the large unlabeled data size. However, the preliminary estimator driven by the small n plays a central role in guiding our iterative algorithm and protecting it from falling into local minima at the very beginning. Importantly, we justify that an ℓ_2-consistent preliminary estimator, guaranteed by the mild sparsity and regularity assumptions, is sufficient for our method to achieve the same convergence rate as the global optimum without reliance on the small n. At first glance, this seems to be just a simple improvement of the Newton-type optimization. However, our strategy of addressing the nonparametric estimator in SIM via orthogonality, as well as adapting the regularization to control the excessive approximate errors, is actually more sophisticated than approaches in the existing literature and essential for eliminating the impact of the high-error preliminary estimator and ensuring the optimal convergence rate. Our work is also the first one to achieve an efficient multi-site aggregation for high-dimensional sparse SIM under the DataSHIELD constraint. This advancement is also based on our novel bias-corrected approximation procedure for SIM.

2. Method

2.1 Outline of SASH

Let $\widehat{\mathbb{P}}^{(\mathsf{m})}$ denote the empirical probability measure with observed data in $\mathscr{U}^{(\mathsf{m})}$ and $\widehat{\mathbb{P}}_n^{(1)}$ denote the empirical probability measure on the labeled data set $\mathscr{L}^{(1)}$. For a vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_d)^\intercal \in \mathbb{R}^d$ and $q > 0$, we denote the ℓ_q norm of $\boldsymbol{v}$ as $\|\boldsymbol{v}\|_q = \left(\sum_{i=1}^d |v_i|^q\right)^{1/q}$. For ease of notation, we drop the superscript in $\alpha^{(1)}$ and denote it as α throughout the remainder of the paper since the intercept terms in sites $2, 3, \ldots, M$ do not appear in our formulation. We write $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some constant $C > 0$, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if $C \leq a_n/b_n \leq C'$ for some constants $C, C' > 0$. Also, we write $a_n = O_p(b_n)$ and $a_n = o_p(b_n)$ respectively for $a_n = O(b_n)$ and $a_n = o(b_n)$ with probability approaching 1.

Algorithm 1 Outline of SASH.

Input: $\mathscr{L}^{(1)}$ and $\{\mathscr{U}^{(\mathsf{m})}, m = 1, ..., M\}$.

Step I: Obtain an initial estimator of the direction vector $\boldsymbol{\gamma}_0$ using $\mathscr{L}^{(1)}$ through supervised learning (4), and broadcast the estimator $\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}$ to all local sites.

Step II: In each local site, fit the SIM to estimate the direction $\boldsymbol{\gamma}_0$ using the preliminary direction estimator from site 1 as shown in Algorithm 2. Then derive summary statistics to approximate the individual-level loss function, and transfer them to site 1.

Step III: Aggregate in site 1 the summary statistics from the local sites to obtain the final estimator of $\boldsymbol{\beta}_0$.

In Algorithm 1, we outline SASH, a *three-step* learning approach that effectively leverages the unlabeled data $\{\mathscr{U}^{(\mathsf{m})}, m = 1, ..., M\}$ and the labeled subset $\mathscr{L}^{(1)}$. Our identification strategy is to decompose the target $\boldsymbol{\beta}_0$ into two parts, its direction $\boldsymbol{\gamma}_0$ and magnitude β_{01}. We first estimate $\boldsymbol{\gamma}_0$ precisely through fitting model (2), the SIM of $S_i^{(\mathsf{m})} \sim \mathbf{X}_i^{(\mathsf{m})}$, with the large samples $\{\mathscr{U}^{(\mathsf{m})}, m = 1, ..., M\}$ via federated SIM learning. Crucially, this procedure is supported by a preliminary supervised estimator derived with the labeled data $\mathscr{L}^{(1)}$. Then for the one-dimensional magnitude β_{01}, we estimate it using $\mathscr{L}^{(1)}$ and the estimate of $\boldsymbol{\gamma}$. The details of the main steps are described in Section 2.2.

2.2 Details of the main steps

Step I. We first obtain an initial estimator by fitting a supervised penalized logistic model using the small labeled samples in site 1:

$$\{\widetilde{\alpha}_{\mathsf{sup}}, \widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}\} = \underset{\alpha\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \widehat{\mathbb{P}}_n^{(1)} \ell\{Y, g(\alpha + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\} + \lambda_{\mathsf{sup}} \|\boldsymbol{\beta}_{\text{-}1}\|_1, \tag{4}$$

where $\ell(y, x) = -\{y \log x + (1 - y) \log(1 - x)\}$ and $\lambda_{\mathsf{sup}} > 0$ is a tuning parameter. Since we assume that $\beta_{01} \neq 0$, we impose the Lasso penalty only on $\boldsymbol{\beta}_{\text{-}1} = (\beta_2, \ldots, \beta_p)^{\mathsf{T}}$ in (4), thereby avoiding direct shrinkage of the normalizing coefficient toward zero. We then broadcast $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} = \widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}/\widetilde{\beta}_{\mathsf{sup},1}$ to all sites for local training of the sparse SIM in the next step. Here, the normalization of $\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}$ with $\widetilde{\beta}_{\mathsf{sup},1}$ is made corresponding to the identification condition $\gamma_1 = 1$.

Remark 1 *We assume that at least one of the risk factors is* known *to be strongly associated with Y and specify it as X_1. Without such prior knowledge, it is plausible to modify our proposed algorithm by introducing a screening procedure to select a strong risk factor of Y at the beginning. Existing variable selection methods such as Lasso (Tibshirani, 1996) and sure independence screening (Fan and Lv, 2008) could be applied to the labeled sample $\mathscr{L}^{(1)}$. Alternatively, surrogate-based screening could be applied to the combined $(S, \mathbf{X})$ data from the labeled sample and the unlabeled samples $\{\mathscr{U}^{(\mathsf{m})}, m = 1, \ldots, M\}$.*

Step II. Based on Step I, we next model $S_i^{(\mathsf{m})} \sim \mathbf{X}_i^{(\mathsf{m})}$ in each local site, to obtain a more accurate estimator for $\boldsymbol{\gamma}$ and use it to derive appropriate summary statistics for integrative analyses. For this purpose, we first define the loss of the SIM (2). Given any $\boldsymbol{\gamma}$ and

subject $\mathbf{X}_i^{(\mathsf{m})}$, the unknown link function $f_m(\cdot)$ can be naturally estimated through a kernel smoothing estimator:

$$\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}) = \frac{\sum_{j=1,j\neq i}^{N_m} K_{h_m}(\boldsymbol{\gamma}^\intercal \mathbf{X}_j^{(\mathsf{m})} - \boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})}) S_j^{(\mathsf{m})}}{\sum_{j=1,j\neq i}^{N_m} K_{h_m}(\boldsymbol{\gamma}^\intercal \mathbf{X}_j^{(\mathsf{m})} - \boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})})} \tag{5}$$

with $K_{h_m}(u) = K(u/h_m)/h_m$, where $K(u)$ represents a kernel function and $h_m > 0$ is a bandwidth parameter specific for each m. Then the individual-level loss function for $S^{(\mathsf{m})}$ can be formulated as

$$L^{(\mathsf{m})}(\boldsymbol{\gamma}) = \widehat{\mathbb{P}}^{(\mathsf{m})} \widehat{\phi}_m(\mathbf{X};\boldsymbol{\gamma}) \{S - \widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})\}^2, \tag{6}$$

where $\widehat{\phi}_m(\mathbf{X};\boldsymbol{\gamma})$ is a weighting function introduced to improve estimation efficiency and it may be data-driven. Typically, improper specification of ϕ_m has no impact on the convergence rate of the estimator but the finite-sample performance may be affected. In Appendix D.1, we provide practical suggestions on ϕ_m for binary, count, and continuous surrogates.

The objective (6) is challenging to optimize since the presence of $\widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})$ makes it non-convex and, thus, prone to the problem of local minima. Our solution is to initiate at the supervised estimator $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ and approximate (6) with some quadratic function of $\boldsymbol{\gamma}^{(\mathsf{m})}$. For the quadratic approximation with sufficiently small bias, the key procedure is to linearize $\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. Heuristically, with some preliminary $\widetilde{\boldsymbol{\gamma}}$, e.g., $\widetilde{\boldsymbol{\gamma}} = \widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$, we approximate $\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})$ by

$$\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}) \approx \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widetilde{\boldsymbol{\gamma}}) + (\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}})^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widetilde{\boldsymbol{\gamma}}) + O(\|\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}\|_2^2), \tag{7}$$

where $\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})$ represents the partial derivative function of $\widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})$ on $\boldsymbol{\gamma}$ with its explicit form given by equation (D.35) in Appendix. To make the explicit form of $\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})$ extractable, we further require $K(u)$ to be differentiable with derivative $K'(u)$ and $K_h'(u) = h^{-2}K'(u/h)$, as satisfied, for example, by a suitably normalized triweight kernel $K(u) \propto (1-u^2)^3 \mathbf{1}(|u| \leq 1)$. Based on (7), we define the one-step approximation of the SIM loss $L^{(\mathsf{m})}(\boldsymbol{\gamma})$ around $\widetilde{\boldsymbol{\gamma}}$ as

$$Q^{(\mathsf{m})}(\boldsymbol{\gamma};\widetilde{\boldsymbol{\gamma}}) = \widehat{\mathbb{P}}^{(\mathsf{m})} \left\{ S - \widehat{f}_m(\mathbf{X};\widetilde{\boldsymbol{\gamma}}) - (\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}})^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X};\widetilde{\boldsymbol{\gamma}}) \right\}^2. \tag{8}$$

As will be justified in Section 3, impact of the preliminary $\widetilde{\boldsymbol{\gamma}}$ on the solution of (8) is up to the second order, in a similar spirit to the classic one-step approximation theory (e.g., Bickel, 1975). However, when n is much smaller than N_m, the second-order impact of $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ can still exceed the error of the desirable global solution to $L^{(\mathsf{m})}(\boldsymbol{\gamma})$ and, thus, dominate the error rate. Thus, instead of one-shot updating, we carry out multiple rounds of updates on $\boldsymbol{\gamma}$ with $Q^{(\mathsf{m})}(\boldsymbol{\gamma};\widetilde{\boldsymbol{\gamma}})$ in our proposed Algorithm 2.

Algorithm 2 is an iterative algorithm that starts at $\widehat{\boldsymbol{\gamma}}_{[0]}^{(\mathsf{m})} = \widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$, and updates from $\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}$ to $\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}$ with (9), an ℓ_1-regularized quadratic objective with loss $Q^{(\mathsf{m})}(\cdot)$. In (9), we force the first coefficient of $\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}$ to be 1 corresponding to our identification assumption. The penalty parameter $\lambda_t^{(\mathsf{m})}$ in (9) is set to gradually decrease with t so that the excessive error of the preliminary estimator can be properly controlled in early iterations, and the regularization

Algorithm 2 Sparse SIM regression with multi-round convex approximation.

Input: The preliminary estimator $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$, samples $\{\mathscr{U}^{(\mathsf{m})}, m = 1, ..., M\}$ with surrogates, and the number of iterations T_m. **For** site $m = 1, 2, \ldots, M$, initiate with $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[0]} = \widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$,

For $t = 1, 2, \ldots, T_m$,

$$\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^p}{\operatorname{argmin}}\, Q^{(\mathsf{m})}(\boldsymbol{\gamma}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]}) + \lambda^{(\mathsf{m})}_t \|\boldsymbol{\gamma}\|_1, \quad \text{s.t.} \quad \gamma_1 = 1. \tag{9}$$

Return in site m: $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})} = \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[T_m]}$.

error caused by $\lambda^{(\mathsf{m})}_t$ can be controlled at its optimal rate around the final rounds; see Theorem 1. This indicates an essential difference between our strategy and a standard Newton-type procedure for the ℓ_1-penalized SIM with a fixed tuning parameter.

Remark 2 *Due to the non-convexity of the original SIM loss in (6), it is necessary to initiate with a sufficiently good $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ (i.e., being close enough to the true parameters) to ensure the convergence of Algorithm 2. Intuitively, this can protect our iteratively updated optimizer from falling into some local minima at the very beginning. In Section 3, we prove that $\|\boldsymbol{\gamma}_0 - \widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}\|_2 = o_p(1)$ is a sufficient condition for convergence. This ℓ_2-consistency property is justified in Lemma 1 under mild assumptions, in which the labeled sample size n could be substantially smaller than the unlabeled size N_m. In addition, we show that the impact of the initial supervised estimator $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ will decay exponentially fast along the iteration of (9) and vanish in the end of Algorithm 2. As a consequence, the number of iterations T_m does not need to be larger than $\log\log N_m$, indicating that Algorithm 2 is not costly in computation; see Theorem 1.*

Remark 3 *In (7) and (8), we use $\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \boldsymbol{\gamma})$, the derivative of the kernel estimator $\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \boldsymbol{\gamma})$ on $\boldsymbol{\gamma}$ as the gradient to construct the linear expansion of $\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \boldsymbol{\gamma})$. Unlike the GLM, this term is not an estimator of $\mathbf{X}f'_m(\boldsymbol{\gamma}^\mathsf{T}\mathbf{X}^{(\mathsf{m})}_i)$ where f'_m is the derivative of the true link f_m. Actually, there is a subtle but essential difference between using $\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \boldsymbol{\gamma})$ and an estimate of $\mathbf{X}f'_m(\boldsymbol{\gamma}^\mathsf{T}\mathbf{X}^{(\mathsf{m})}_i)$ in that the former is taken on the kernel estimator $\widehat{f}_m$ while the latter treats f_m as known. Consequently, using the latter would cause excessive bias from the kernel part. The term $\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \boldsymbol{\gamma})$ serves a similar role as the efficient score function of SIM $(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \boldsymbol{\gamma}^\mathsf{T}\mathbf{X}^{(\mathsf{m})}_i])f'_m(\boldsymbol{\gamma}^\mathsf{T}\mathbf{X}^{(\mathsf{m})}_i)$ used for bias correction with respect to the nuisance kernel estimator in the inference of SIM (Liang et al., 2010; Wu et al., 2023).*

Based on $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}$, we derive appropriate summary statistics to approximate the individual data loss function $L^{(\mathsf{m})}(\boldsymbol{\gamma})$ in each site m. Again, inspired by the bias-corrected approximation used in (7) and (8), we consider the quadratic expansion

$$L^{(\mathsf{m})}(\boldsymbol{\gamma}) \approx Q^{(\mathsf{m})}(\boldsymbol{\gamma}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) = C^{(\mathsf{m})} - 2\boldsymbol{\gamma}^\mathsf{T}\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,s} + \boldsymbol{\gamma}^\mathsf{T}\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}\boldsymbol{\gamma}, \tag{10}$$

where the second-order summary statistics $\widehat{\mathbf{\Omega}}_{x,x}^{(\mathsf{m})} \in \mathbb{R}^{p\times p}$ and the first-order summary statistics $\widehat{\mathbf{\Omega}}_{x,s}^{(\mathsf{m})} \in \mathbb{R}^{p}$ are formulated and computed as:

$$\begin{aligned}
\widehat{\mathbf{\Omega}}_{x,x}^{(\mathsf{m})} &= \widehat{\mathbb{P}}^{(\mathsf{m})}\widehat{\phi}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\}^{\mathsf{T}},\\
\widehat{\mathbf{\Omega}}_{x,s}^{(\mathsf{m})} &= \widehat{\mathbb{P}}^{(\mathsf{m})}\widehat{\phi}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{S-\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})+\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}),
\end{aligned} \tag{11}$$

and $C^{(\mathsf{m})} = \widehat{\mathbb{P}}^{(\mathsf{m})}\widehat{\phi}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\big\{S-\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})+\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\big\}^2$ is a constant not depending on $\boldsymbol{\gamma}$. Motivated by the above discussion, $\widehat{\mathbf{\Omega}}_{x,x}^{(\mathsf{m})}$ and $\widehat{\mathbf{\Omega}}_{x,s}^{(\mathsf{m})}$ are sufficient for an effective quadratic approximation to the local individual loss $L^{(\mathsf{m})}(\boldsymbol{\gamma})$, while being free of individual-level information ruled out by DataSHIELD (Wolfson et al., 2010). These two statistics are then transferred to site 1 for an integrative sparse SIM regression.

Step III. Finally, we aggregate summary statistics $\widehat{\mathbf{\Omega}}_{x,s}^{(\mathsf{m})}, \widehat{\mathbf{\Omega}}_{x,x}^{(\mathsf{m})}$ from local sites and the small labeled samples at site 1, to derive the final estimator for $\boldsymbol{\beta}$. For simplicity, we drop the useless constant $C^{(\mathsf{m})}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})$ in (10) and still denote by $Q^{(\mathsf{m})}(\boldsymbol{\gamma};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) = -2\boldsymbol{\gamma}^{\mathsf{T}}\widehat{\mathbf{\Omega}}_{x,s}^{(\mathsf{m})} + \boldsymbol{\gamma}^{\mathsf{T}}\widehat{\mathbf{\Omega}}_{x,x}^{(\mathsf{m})}\boldsymbol{\gamma}$. We first derive a temporary estimator for the direction coefficients $\boldsymbol{\gamma}$ through:

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{N}\sum_{m=1}^{M} N_m Q^{(\mathsf{m})}(\boldsymbol{\gamma};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) + \lambda\|\boldsymbol{\gamma}\|_1 \quad \text{s.t.} \quad \gamma_1 = 1, \tag{12}$$

where $N = \sum_{m=1}^{M} N_m$. Then, we plug $\widehat{\boldsymbol{\gamma}}$ in a low-dimensional logistic regression on the labeled data, and estimate the intercept α and the scalar β_1:

$$(\widehat{\alpha},\widehat{\beta}_1) = \underset{\alpha,\beta_1}{\operatorname{argmin}}\, L_Y(\alpha,\beta_1,\widehat{\boldsymbol{\gamma}}), \quad \text{where} \quad L_Y(\alpha,\beta_1,\boldsymbol{\gamma}) = \widehat{\mathbb{P}}_n^{(1)}\ell\big\{Y, g(\alpha+\beta_1\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{X})\big\}. \tag{13}$$

Finally, we aggregate the summary data and the labeled samples based on $(\widehat{\alpha},\widehat{\beta}_1)$ to derive:

$$\widehat{\boldsymbol{\gamma}}^{\dagger} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^p}{\operatorname{argmin}}\frac{1}{N^+}\Big[\sum_{m=1}^{M} N_m Q^{(\mathsf{m})}(\boldsymbol{\gamma};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) + nL_Y(\widehat{\alpha},\widehat{\beta}_1,\boldsymbol{\gamma})\Big] + \lambda^{\dagger}\|\boldsymbol{\gamma}\|_1 \quad \text{s.t.} \quad \gamma_1 = 1, \tag{14}$$

where $N^+ = N + n$. The coefficient n multiplying L_Y is used consistently in the SASH objective, the corresponding BIC, and the IPD objective below. Note that $\widehat{\boldsymbol{\gamma}}^{\dagger}$ solved from (14) could be slightly more accurate than $\widehat{\boldsymbol{\gamma}}$ solved from (12) due to the use of the small labeled sample. Our final SASH estimator is taken as $\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}} = \widehat{\beta}_1\widehat{\boldsymbol{\gamma}}^{\dagger}$.

2.3 Tuning strategy

Theoretically optimal rates for all tuning parameters will be given in Section 3. In this section, we outline our tuning strategy in the numerical implementation that may be slightly different from the theory. Details of our tuning procedures can be found in Appendix D. For the penalty parameter λ_{sup} used in Step I, we select it using cross-validation (CV). For the bandwidth h_m, we fix $h_m = \{s\log(p\vee N_m)/N_m\}^{1/5}$ as suggested by Theorem 1. Through simulation, we demonstrate that a moderate variation of h_m does not substantially change the overall results in Section 4.3.4. Thus, for simplicity and computation efficiency, we

suggest using the fixed value for h_m introduced above. For the unknown sparsity level s, since h_m is proportional to $s^{1/5}$ that is typically small in practice, the value of h_m is not sensitive to the choice of this unknown parameter.

By Theorem 1, setting the number of iterations in Algorithm 2 as $\log\log N_m$ is sufficient for desirable convergence. Since $\log\log N_m$ is typically small in practice, we fix $T_m = 4$ in our numerical studies. Considering the unavailability of individual-level samples from sites $2, 3, \ldots, M$ in Step III, we select the tuning parameters λ and $\lambda^\dagger$ using the Bayesian information criterion (BIC) inspired by Cai et al. (2022b). To this end, we additionally calculate $\{\widehat{\sigma}^{(\mathsf{m})}\}^2 = L^{(\mathsf{m})}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})$ at site m and transfer it to site 1 together with $\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}$ and $\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,s}$. Then in Step III, we propose to choose the λ that minimizes:

$$\mathrm{BIC}_1(\lambda) = \frac{1}{N}\left[\sum_{m=1}^{M} \frac{N_m Q^{(\mathsf{m})}(\widehat{\boldsymbol{\gamma}}(\lambda); \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})}{\{\widehat{\sigma}^{(\mathsf{m})}\}^2} + \mathrm{df}\{\widehat{\boldsymbol{\gamma}}(\lambda)\}\log(N)\right], \tag{15}$$

where $\widehat{\boldsymbol{\gamma}}(\lambda)$ is the solution of (12) using λ as the penalty parameter, and $\mathrm{df}\{\widehat{\boldsymbol{\gamma}}(\lambda)\}$ is the degrees of freedom, i.e., the number of nonzero coefficients in $\widehat{\boldsymbol{\gamma}}(\lambda)$. $\mathrm{BIC}_1(\lambda)$ is a commonly used tuning criterion and it depends on the data only through the summary statistics $\{\widehat{\sigma}^{(\mathsf{m})}\}^2$, $\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}$, and $\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,s}$. Similarly, for $\lambda^\dagger$ in (14) with its corresponding solution $\widehat{\boldsymbol{\gamma}}^\dagger(\lambda^\dagger)$, we choose the one minimizing:

$$\mathrm{BIC}_2(\lambda^\dagger) = \frac{1}{N^+}\left[\sum_{m=1}^{M} \frac{N_m Q^{(\mathsf{m})}(\widehat{\boldsymbol{\gamma}}^\dagger(\lambda^\dagger); \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})}{\{\widehat{\sigma}^{(\mathsf{m})}\}^2} + nL_Y(\widehat{\alpha}, \widehat{\beta}_1, \widehat{\boldsymbol{\gamma}}^\dagger(\lambda^\dagger)) + \mathrm{df}\{\widehat{\boldsymbol{\gamma}}^\dagger(\lambda^\dagger)\}\log(N^+)\right].$$

Finally, the penalty parameter $\lambda_t^{(\mathsf{m})}$ used in Algorithm 2 is also selected using the BIC similar to (15), with the loss function calculated based on $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]}$ derived in the previous iteration; see details in Appendix D.

2.4 Interval estimation

We introduce a method to construct a confidence interval (CI) for $\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_0$ with any given $\boldsymbol{x}$ based on the SASH estimator and the labeled samples at site 1. Because $(\widehat{\alpha}, \widehat{\beta}_1)$ in (13) is fitted using $\widehat{\boldsymbol{\gamma}}$, the low-dimensional likelihood expansion additionally requires $\sqrt{n}\,\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 = o_p(1)$. To replace $\boldsymbol{\gamma}_0$ by $\widehat{\boldsymbol{\gamma}}^\dagger$ in the final target, we also require $\sqrt{n}\,\boldsymbol{x}^\mathsf{T}(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0) = o_p(1)$. For loadings with bounded ℓ_2 norm, the rate condition $ns\log p/N \to 0$ is sufficient for both requirements by Theorem 2; $N/n \to \infty$ alone is not sufficient. Under these conditions, $n^{1/2}\widehat{\sigma}_1^{-1}(\widehat{\beta}_1 - \beta_{01})$ converges weakly to the standard normal distribution, where $\widehat{\sigma}_1^2$ is the $(2,2)$ entry of $n\left\{(\mathbf{1}_n, \mathbf{X}^{(1)}_{1:n}\widehat{\boldsymbol{\gamma}})^\mathsf{T}\widehat{\mathbb{W}}(\mathbf{1}_n, \mathbf{X}^{(1)}_{1:n}\widehat{\boldsymbol{\gamma}})\right\}^{-1}$, the second column corresponding to β_1, $\mathbf{X}^{(1)}_{1:n} = (\mathbf{X}^{(1)}_1, \ldots, \mathbf{X}^{(1)}_n)^\mathsf{T}$, and

$$\widehat{\mathbb{W}} = \mathrm{diag}\left\{\dot{g}\big(\widehat{\alpha} + \widehat{\beta}_1\mathbf{X}^{(1)\mathsf{T}}_1\widehat{\boldsymbol{\gamma}}\big), \ldots, \dot{g}\big(\widehat{\alpha} + \widehat{\beta}_1\mathbf{X}^{(1)\mathsf{T}}_n\widehat{\boldsymbol{\gamma}}\big)\right\}.$$

Thus, the $100(1-\delta)\%$ CI for $\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_0$ is

$$\left[\widehat{\beta}_1\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\gamma}}^\dagger - \frac{\widehat{\sigma}_1|\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\gamma}}^\dagger|}{\sqrt{n}}\Phi^{-1}(1-\delta/2), \quad \widehat{\beta}_1\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\gamma}}^\dagger + \frac{\widehat{\sigma}_1|\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\gamma}}^\dagger|}{\sqrt{n}}\Phi^{-1}(1-\delta/2)\right],$$

where Φ is the standard normal distribution function.

When either direction-estimation condition above fails, debiasing is needed to obtain a valid interval for $\boldsymbol{x}^\intercal\boldsymbol{\beta}_0$. Compared with supervised debiased inference based solely on the labeled sample (Cai et al., 2021; Guo et al., 2021), the unlabeled covariates and the more precise direction estimator $\widehat{\boldsymbol{\gamma}}^\dagger$ may reduce the cost of debiasing. When $\boldsymbol{x}$ is dense, the supervised debiased estimator of Cai et al. (2021) can have asymptotic variance increasing with $\|\boldsymbol{x}\|_2^2$. If $\widehat{\boldsymbol{\gamma}}^\dagger$ accurately identifies the nonzero set S_β of $\boldsymbol{\beta}_0$, the variance may potentially be reduced from order $\|\boldsymbol{x}\|_2^2$ to $\|\boldsymbol{x}_{S_\beta}\|_2^2$ when $|S_\beta| \ll p$. A robust implementation of this extension is left for future research.

3. Theoretical Results

3.1 Notations and Assumptions

Let $\|\boldsymbol{v}\|_0$ be the total number of nonzero elements in $\boldsymbol{v}$ and $\|\boldsymbol{v}\|_\infty = \max_{1\le i\le d}|v_i|$. For a matrix $A = [A_{ij}]$, let ℓ_1-norm $\|A\|_1 = \max_j \sum_i |A_{ij}|$, ℓ_∞-norm $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, and $\|A\|_2 = \lambda_{\max}(A)$ where $\lambda_{\max}(A)$ represents the largest singular value of matrix A. We say that a random variable X is sub-Gaussian if $\mathbb{P}(|X| > t) \le Ce^{-ct^2}$ for some constants C, c, and any $t > 0$, and a random vector $\mathbf{X}$ is sub-Gaussian if $\boldsymbol{v}^\intercal\mathbf{X}$ is sub-Gaussian for any $\boldsymbol{v} \in \mathbb{R}^d$ such that $\|\boldsymbol{v}\|_2 = 1$. Recall that we assume the labeled sample size $n \lesssim N_m$ for $m \in \{1, 2, \ldots, M\}$. For simplicity, we set $N = \sum_{m=1}^M N_m$. Define the set of nonzero coefficients as $\mathcal{B} = \{j : \boldsymbol{\beta}_{0j} \neq 0\} = \{j : \boldsymbol{\gamma}_{0j} \neq 0\}$ and its cardinality, i.e., the sparsity level as $s = |\mathcal{B}|$. We focus on the scenario that the weighting function $\widehat{\phi}_m(\cdot;\cdot) = 1$ in the theoretical analysis. Let $\widetilde{\mathbf{X}}_i^{(1)} = (1, \mathbf{X}_i^{(1)\intercal})^\intercal$ and denote the joint Fisher information for the unpenalized intercept and the regression coefficients by $\mathbf{H}_{\text{joint}} = \mathbb{E}\left[\dot{g}(\alpha_0 + \boldsymbol{\beta}_0^\intercal\mathbf{X}_i^{(1)})\widetilde{\mathbf{X}}_i^{(1)}\widetilde{\mathbf{X}}_i^{(1)\intercal}\right]$. To establish the theoretical properties of our SASH estimator, we introduce regularity, sparsity, and smoothness assumptions as below.

Assumption 1 *It holds that*

$$\kappa^{-1} < \lambda_{\min}(\mathbf{H}_{\text{joint}}) \le \lambda_{\max}(\mathbf{H}_{\text{joint}}) < \kappa, \qquad \|\boldsymbol{\beta}_0\|_2 \le \kappa, \qquad \|\boldsymbol{\gamma}_0\|_2 \le \kappa_\gamma,$$

for fixed constants $\kappa, \kappa_\gamma > 0$. The joint-information condition accounts for the unpenalized intercept and the unpenalized normalizing coefficient β_1, and excludes collinearity involving these coordinates. The sparsity level satisfies $s = |\mathcal{B}| = o(n\beta_{01}^2/\log p)$ and $|\beta_{01}| = O(1)$.

Assumption 2 *Let $r_n = C_\Gamma\sqrt{s\log p/n\beta_{01}^2}$, and $R_n = 4C_\Gamma s\sqrt{\log p/n\beta_{01}^2}$, for a sufficiently large fixed constant C_Γ, and define*

$$\Gamma = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \gamma_1 = 1,\ \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2 \le r_n,\ \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_1 \le R_n\}.$$

The following conditions hold for every $m = 1, \ldots, M$.

(a) The link f_m has bounded first three derivatives on $\mathbb{R}$, and

$$\mathbb{E}\big[\{f_m'(\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(m)})\}^2\big] = a_m^2 > 0, \qquad \max_i |f_m'(\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(m)})| \le b,$$

for fixed constants $a_m, b > 0$.

(b) *Let $g_{m,\boldsymbol{\gamma}}$ be the density of $\boldsymbol{\gamma}^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}$. Uniformly over $\boldsymbol{\gamma}\in\Gamma$, $g_{m,\boldsymbol{\gamma}}$ is twice differentiable, $g_{m,\boldsymbol{\gamma}}$ and its first two derivatives are bounded, and $g_{m,\boldsymbol{\gamma}}$ is bounded away from zero.*

(c) *For $r,q\in\{0,1,2\}$, let $\mathcal{K}_{m,h}^{(r,q)}=\left\{(\mathbf{x},\mathbf{x}')\mapsto K_h^{(r)}\{\boldsymbol{\gamma}^\mathsf{T}(\mathbf{x}'-\mathbf{x})\}(\mathbf{x}'-\mathbf{x})^{\otimes q}:\boldsymbol{\gamma}\in\Gamma\right\}$. Let $\mathfrak{F}_{m,h}$ contain these classes and the corresponding coordinatewise score classes. There exist constants $A,C>0$ and envelopes $F_{\mathcal{F}}$ such that, uniformly over probability measures Q, $\mathcal{F}\in\mathfrak{F}_{m,h}$, and $0<\varepsilon\le 1$,*

$$\log N(\varepsilon\|F_{\mathcal{F}}\|_{Q,2},\mathcal{F},L_2(Q))\le Cs\log\left(\frac{Ap}{\varepsilon h}\right),$$

with the envelope moments required in the concentration bounds.

Assumption 3 *The kernel function $K(\cdot)$ is nonnegative, bounded, symmetric around zero, and twice differentiable. $K(\cdot)$ and its derivatives $K'(\cdot),K''(\cdot)$ are all Lipschitz on $\mathbb{R}$ and satisfy that*

$$\lim_{|\nu|\to\infty}K(\nu)=0,\quad \int_{-\infty}^{\infty}K(\nu)d\nu=1,\quad \textit{and}\quad \int_{-\infty}^{\infty}\nu K'(\nu)d\nu=-1.$$

In addition, for $i=0,1,\ldots,4$, $\int\left|\nu^iK(\nu)\right|d\nu<\infty$, $\int\left|\nu^iK'(\nu)\right|d\nu<\infty$; and for $i=0,1,2$, $\int\left|\nu^iK''(\nu)\right|d\nu<\infty$, $\int\left|\nu^iK'^2(\nu)\right|d\nu<\infty$.

Assumption 4 *For each sample i from site m, $\mathbf{X}_i^{(\mathsf{m})}$ and $\epsilon_i^{(\mathsf{m})}=S_i^{(\mathsf{m})}-f_m(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})})$ are both sub-Gaussian.*

Remark 4 *Assumption 1 is a standard joint-curvature condition for logistic regression with an unpenalized intercept and an unpenalized normalizing coefficient. Combined with Assumption 4, it yields the ℓ_1- and ℓ_2-rates for the preliminary estimator in Lemma 1. Assumption 2(a) imposes standard smoothness and informativeness conditions on f_m, while Assumption 2(b)–(c) imposes regularity conditions on the projected-index density and the associated kernel classes. Assumption 3 is satisfied by standard smooth kernels such as the Gaussian kernel. Assumption 4 controls the tail behavior of the covariates and residuals. The present theory assumes a continuously distributed index and sub-Gaussian residuals; the count-surrogate and discrete-genetic-index applications are evaluated empirically and extensions to sub-exponential residuals or discrete indices are left for future work.*

Assumption 5 *Define the normalized tangent space $\mathbb{V}_0=\left\{\boldsymbol{v}\in\mathbb{R}^p:v_1=0,\ \|\boldsymbol{v}\|_2=1\right\}$. There exist positive constants c and C such that the following conditions hold for each sample i from site m:*

$$\begin{aligned}
c&\le\inf_{\boldsymbol{v}\in\mathbb{V}_0}\boldsymbol{v}^\mathsf{T}\mathbb{E}\left[\left\{f_m'(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})})\right\}^2\mathrm{Cov}\left(\mathbf{X}_i^{(\mathsf{m})}\mid\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}\right)\right]\boldsymbol{v}\\
&\le\sup_{\|\boldsymbol{v}\|_2=1}\boldsymbol{v}^\mathsf{T}\mathbb{E}\left[\left\{f_m'(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})})\right\}^2\mathrm{Cov}\left(\mathbf{X}_i^{(\mathsf{m})}\mid\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}\right)\right]\boldsymbol{v}\le C,\\
&\sup_{\boldsymbol{\gamma}\in\Gamma}\frac{1}{N_m}\sum_{i=1}^{N_m}\left[\lambda_{\max}\big(\mathbb{E}(\mathbf{X}_i^{(\mathsf{m})}\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\mid\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\gamma})\big)\right]^2\le C,\\
&\max_{1\le i\le N_m}\sup_{\boldsymbol{\gamma}\in\Gamma}\lambda_{\max}\big(\mathbb{E}(\mathbf{X}_i^{(\mathsf{m})}\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\mid\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\gamma})\big)\le C\log(p\vee N_m).
\end{aligned}$$

Assumption 6 *For each site m, every $\boldsymbol{\gamma} \in \Gamma$, and $t \in \mathbb{R}$, the conditional mean $\mathbb{E}(\mathbf{X}_i^{(\mathsf{m})} \mid \boldsymbol{\gamma}^\top \mathbf{X}_i^{(\mathsf{m})} = t)$ is twice differentiable in t. Let $\mathbb{E}^{(1)}$ and $\mathbb{E}^{(2)}$ denote its first two derivatives. For every $\boldsymbol{v} \in \mathbb{R}^p$,*

$$\begin{aligned}
\max_i \sup_{\boldsymbol{\gamma}\in\Gamma} \left|\mathbb{E}(\boldsymbol{v}^\top \mathbf{X}_i^{(\mathsf{m})} \mid \boldsymbol{\gamma}^\top \mathbf{X}_i^{(\mathsf{m})})\right| &\le C\|\boldsymbol{v}\|_2, \\
\max_i \sup_{\boldsymbol{\gamma}\in\Gamma} \left|\mathbb{E}^{(1)}(\boldsymbol{v}^\top \mathbf{X}_i^{(\mathsf{m})} \mid \boldsymbol{\gamma}^\top \mathbf{X}_i^{(\mathsf{m})})\right| &\le C\|\boldsymbol{v}\|_2, \\
\sup_{|t|\le cs\sqrt{\log(p\vee N_m)}} \sup_{\boldsymbol{\gamma}\in\Gamma} \left|\mathbb{E}^{(2)}(\boldsymbol{v}^\top \mathbf{X}_i^{(\mathsf{m})} \mid \boldsymbol{\gamma}^\top \mathbf{X}_i^{(\mathsf{m})} = t)\right| &\le C\|\boldsymbol{v}\|_2, \\
\max_i \sup_{\boldsymbol{\gamma}\in\Gamma} \left|\mathbb{E}^{(1)}\{(\boldsymbol{v}^\top \mathbf{X}_i^{(\mathsf{m})})^2 \mid \boldsymbol{\gamma}^\top \mathbf{X}_i^{(\mathsf{m})}\}\right| &\le C\|\boldsymbol{v}\|_2^2\sqrt{\log(p\vee N_m)}.
\end{aligned}$$

Remark 5 *Assumption 5 imposes restricted curvature on the derivative-weighted conditional covariance, which is the relevant population Hessian of the SIM risk. The restriction $v_1 = 0$ reflects the normalization $\gamma_1 = 1$ and removes the intrinsically non-identifiable direction of the single-index model. Assumption 6 imposes smoothness conditions on the conditional moments of $\mathbf{X}_i^{(\mathsf{m})}$ given $\boldsymbol{\gamma}^\top \mathbf{X}_i^{(\mathsf{m})}$, which are used to control the kernel-based approximation uniformly over local neighborhoods of $\boldsymbol{\gamma}_0$. We provide a justification of Assumption 5 under a Gaussian design in Appendix C; analogous conditions for more general sub-Gaussian designs may be justified using results such as Zhou (2009) and Rudelson and Zhou (2012). Similar assumptions have also been used to analyze high-dimensional sparse SIM by Wu et al. (2023). Unlike that work, we do not impose sparsity assumptions on the inverse Hessian matrix of the SIM because SASH does not rely on local debiased estimators for aggregation.*

3.2 Convergence Rates

In this section, we derive the estimation error rate for the SASH estimator. We first present the estimation error of the local direction estimators obtained via regularized SIM regression in Step II, with its proof provided in Appendix A.2.

Lemma 1 (Consistency of the initial supervised estimator) *Under the logistic model assumption (1) and Assumptions 1 and 4, there exists $\lambda_{\mathsf{sup}} \asymp (\log p/n)^{1/2}$ such that, with probability approaching* 1*,*

$$|\widetilde{\beta}_{\mathsf{sup},1}| \ge \frac{1}{2}|\beta_{01}|, \qquad \|\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} - \boldsymbol{\gamma}_0\|_2 \lesssim \sqrt{\frac{s\log p}{n\beta_{01}^2}}, \qquad \|\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} - \boldsymbol{\gamma}_0\|_1 \lesssim s\sqrt{\frac{\log p}{n\beta_{01}^2}}.$$

Consequently, $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} \in \Gamma$ for a sufficiently large C_Γ.

Based on the consistency of $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ shown in Lemma 1, we can justify the convergence of our iterative Algorithm 2 initialized with $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ by characterizing and controlling the impact of $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ along the iterations, and then establish the convergence of the local-site estimators $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}$ in Theorem 1. It is important to note the complication of justifying Theorem 1 due to the excessive bias from the kernel estimator of f_m. As a key step in the proof, we

showed that this bias can be reduced leveraging the orthogonality between $\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma})$ and $\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\gamma}$, i.e., $\psi(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\gamma})\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma})$ concentrate around zero for arbitrary $\boldsymbol{\gamma}$ and $\psi(\cdot)$. Remark 3 is intrinsically related to this technical property.

Theorem 1 (Convergence rate for the local estimators) *Under the logistic model (1), the SIM assumption (2), Assumptions 1 – 6 and the sparsity assumption $s = O(N_m^{1/6}/\log(p \vee N_m))$ for all $m \in [M]$, there exist $\lambda_{\mathsf{sup}} \asymp (\log p/n)^{1/2}$, h_m satisfying $\{s \log(p \vee N_m)/N_m\}^{1/5} \lesssim h_m \lesssim N_m^{-1/6}$ and $h_m < \delta$, $T_m \asymp \log(-\log h_m) - \log\{\log(n\beta_{01}^2/(s \log p)\}$, $\lambda_{T_m}^{(\mathsf{m})} \asymp (\log p/N_m)^{1/2}$, and $\{\lambda_t^{(\mathsf{m})} : t < T_m\}$ as specified in* (A.6) *of Appendix, such that*

$$\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})} - \boldsymbol{\gamma}_0\|_2 \lesssim \sqrt{\frac{s \log p}{N_m}}, \quad \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})} - \boldsymbol{\gamma}_0\|_1 \lesssim s\sqrt{\frac{\log p}{N_m}}$$

with probability approaching 1.

Remark 6 *In terms of the sparsity conditions in Theorem 1, we impose them regarding both the small labeled sample size n and the large N_m. For n, we require $s = o(n\beta_{01}^2/\log p)$ which just warrants ℓ_2-consistency of the GLM Lasso estimator when $\beta_{01} \asymp 1$ and, thus, is regarded to be mild in existing literature (e.g., Bühlmann and van de Geer, 2011; Negahban et al., 2012). This condition also implies a rate constraint on the strong signal β_{01} that $|\beta_{01}|$ needs to be larger than $(s \log p/n)^{1/2}$ in rate. For N_m, we assume $s = O(N_m^{1/6}/\log(p \vee N_m))$, which is more restrictive than the GLM setting with the true outcome Y of all the N_m samples. Nevertheless, this assumption is still reasonable and mild in our EHR applications where N_m is typically much larger than n and p.*

Remark 7 *It is well-known that under Assumption 2 and a low-dimensional setting, the optimal bandwidth for the kernel estimator of $f_m(\cdot)$ is $N_m^{-1/5}$ for a twice differentiable function (e.g., Wand and Jones, 1994). The rate condition $\{s \log(p \vee N_m)/N_m\}^{1/5} \lesssim h_m \lesssim N_m^{-1/6}$ in Theorem 1 is inconsistent with this choice due to the presence of high-dimensional sparse $\boldsymbol{\gamma}$. In addition, the number of iterations T_m is around the order $\log\{\log(N_m)\}$, indicating that Algorithm 2 can converge fast and be computationally efficient in practice.*

The error rate of $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}$ provided in Theorem 1 is nearly minimax optimal in the sense that even the sparse GLM estimator extracted under known link functions cannot attain any better convergence rates than this up to a $\log N_m$ scale (Raskutti et al., 2011). To the best of our knowledge, except for Gaussian or elliptical design under which one could reduce the SIM fitting to a linear regression (e.g., Neykov et al., 2016; Eftekhari et al., 2021), no existing sparse SIM approach has a similar optimal convergence guarantee for the non-Gaussian (general) design due to the non-convexity issue. From the methodological perspective, we leverage the supervised estimator and the one-step approximation of SIM introduced in Section 2 to fill this gap. In theory, justifying this convergence property is more challenging than analyzing the standard parametric sparse regression, because first, the impact of the nuisance kernel estimator $\widehat{f}_m$ needs to be properly reduced through orthogonality; second, an essential effort is needed to analyze our iterative Algorithm 2 with the error rates changing successively.

Based on Theorem 1, we derive the convergence rate of our SASH estimator in Theorem 2 that is proved in Appendix A.3.

Theorem 2 (Convergence rate of SASH) *Under all assumptions stated in Theorem 1, there exists* $\lambda, \lambda^\dagger \asymp \sqrt{\log p/N}$ *where* $N = \sum_{m=1}^{M} N_m$*, such that the estimators introduced in Step III satisfy*

$$\|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_2 \lesssim \sqrt{\frac{s\log p}{N}}, \quad \|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_1 \lesssim s\sqrt{\frac{\log p}{N}}, \quad |\widehat{\beta}_1 - \beta_{01}| \lesssim \sqrt{\frac{1}{n}} + \sqrt{\frac{s\log p}{N}},$$

and, thus,

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}} - \boldsymbol{\beta}_0\|_2 \lesssim \sqrt{\frac{1}{n}} + \sqrt{\frac{s\log p}{N}}, \quad \|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}} - \boldsymbol{\beta}_0\|_1 \lesssim \sqrt{\frac{s}{n}} + s\sqrt{\frac{\log p}{N}}.$$

with probability approaching 1.

One can see from Theorem 2 that the ℓ_2 estimation error of SASH has a parametric component $n^{-1/2}$ from estimating (α, β_1) and a direction-estimation component $\sqrt{s\log p/N}$. For the ℓ_1 error, multiplication of the scalar error $|\widehat{\beta}_1 - \beta_{01}|$ by $\|\boldsymbol{\gamma}_0\|_1 \lesssim \sqrt{s}\|\boldsymbol{\gamma}_0\|_2$ yields the term $\sqrt{s/n}$; the direction component is $s\sqrt{\log p/N}$. Notably, although Algorithm 2 initiates at the supervised estimator $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ with its ℓ_2-error being $\{s\log p/(n\beta_{01}^2)\}^{1/2}$, the error rate of our final estimator $\widehat{\boldsymbol{\gamma}}^\dagger$, as well as that of $\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}$, turns out to be free of the small labeled sample size n and dominated by the large N and N_m. This is highly desirable but not readily achieved in the current literature of semi-supervised learning (e.g., Zhang et al., 2020b; Hou et al., 2021) and sparse SIM (e.g., Radchenko, 2015; Wu et al., 2023). To achieve this result in Theorems 1 and 2, our main technical novelty can be summarized as (1) developing the one-step approximation of the SIM loss introduced in Section 2.2; (2) designing and justifying Algorithm 2 with a proper number of iteration T_m and sequence of tuning parameters $\{\lambda_t^{(\mathsf{m})} : t = 1, 2, \ldots, T_m\}$ to ensure the optimal convergence rate of our SIM estimator.

3.3 Comparison with the IPD Estimator

In this section, we compare SASH with the IPD pooling estimator obtained using the same samples $\mathscr{L}^{(1)}$ and $\{\mathscr{U}^{(\mathsf{m})}, m = 1, ..., M\}$ but without the DataSHIELD constraint. The detailed implementation of IPD is given in Appendix A.3. In short, we replace the local training in Algorithm 2 by an integrative sparse SIM regression that pools individual-level samples from all sites, and then use labeled data as in (13) and (14). Theorem 3 establishes equality of the first-order convergence rates and a one-sided no-rate-loss comparison. It does not claim that $\|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}} - \widehat{\boldsymbol{\beta}}_{\mathsf{IPD}}\|_q$ is of smaller order than the common estimation rate.

Theorem 3 (First-order rate comparison with IPD) *Under all assumptions in Theorem 1, there exists* $\lambda_{\mathsf{IPD}} \asymp \sqrt{\log p/N}$ *such that the IPD estimator satisfies*

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{IPD}} - \boldsymbol{\beta}_0\|_2 \lesssim \sqrt{\frac{1}{n}} + \sqrt{\frac{s\log p}{N}}, \quad \|\widehat{\boldsymbol{\beta}}_{\mathsf{IPD}} - \boldsymbol{\beta}_0\|_1 \lesssim \sqrt{\frac{s}{n}} + s\sqrt{\frac{\log p}{N}}.$$

with probability approaching 1. *Further, for some* $\lambda_\delta = o(\lambda_{\mathsf{IPD}})$*, SASH with the penalty parameter* $\lambda^\dagger = \lambda_{\mathsf{IPD}} + \lambda_\delta$ *in Step III incurs no first-order rate loss relative to IPD in the*

following one-sided sense:

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}}-\boldsymbol{\beta}_0\|_2 &\le \|\widehat{\boldsymbol{\beta}}_{\mathsf{IPD}}-\boldsymbol{\beta}_0\|_2+o_p\Big(\sqrt{\frac{1}{n}}+\sqrt{\frac{s\log p}{N}}\Big),\\ \|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}}-\boldsymbol{\beta}_0\|_1 &\le \|\widehat{\boldsymbol{\beta}}_{\mathsf{IPD}}-\boldsymbol{\beta}_0\|_1+o_p\Big(\sqrt{\frac{s}{n}}+s\sqrt{\frac{\log p}{N}}\Big).\end{aligned} \tag{16}$$

Theorem 3 shows that SASH matches the first-order ℓ_1 and ℓ_2 convergence rates of IPD and that its estimation error is no larger than the IPD error plus a smaller-order remainder. This is a rate comparison, not a direct estimator-difference result. Similar to Cai et al. (2022b), the sparsity condition $s = O\{N_m^{1/6}/\log(p \vee N_m)\}$ is used to make the higher-order error from the summary-level approximation negligible.

4. Simulation

4.1 Data Generation Setup

We conduct simulation studies to evaluate our proposed SASH estimator and compare it with existing methods. **R** code for implementation can be downloaded from `https://github.com/moleibobliu/SASH`. Throughout, we let the number of sites $M = 4$, the labeled sample (in site 1) size $n = 200$, dimension $p = 300$, and the size of unlabeled sample $N_m = 8000$ for each site $1 \le m \le M$. To mimic our real example of EHR data which consists of count variables, we generated $\widetilde{\mathbf{X}}_i^{(\mathsf{m})}$ from a multivariate Poisson distribution with mean 5 and equal correlation being 0.25, and we take covariates $\mathbf{X}_i^{(\mathsf{m})}$ as $\log(\widetilde{\mathbf{X}}_i^{(\mathsf{m})}+1)$. The binary outcome $Y^{(\mathsf{m})} \in \mathbb{R}^{N_m}$ is only observed in site 1 and it follows the logistic model: $\mathbb{P}(Y_i^{(\mathsf{m})}=1 \mid \mathbf{X}_i^{(\mathsf{m})}) = g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})$, with a sparse regression vector $\boldsymbol{\beta}_0 = (1,-1,0.5,-0.5,0.25,-0.25,0.125,-0.125,\mathbf{0}_{p-8})^{\mathsf{T}}$. Since the EHR study consists of surrogates of different forms, we generate two categories of surrogate $S^{(\mathsf{m})}$ as follows:

$$\begin{aligned} &S_i^{(\mathsf{m})} \mid Y_i^{(\mathsf{m})}=1 \sim \mathrm{Pois}(\mu_1^{(\mathsf{m})}) \quad \text{and} \quad S_i^{(\mathsf{m})} \mid Y_i^{(\mathsf{m})}=0 \sim \mathrm{Pois}(\mu_0^{(\mathsf{m})}) \quad \text{for} \quad 1 \le m \le [M/2],\\ &S_i^{(\mathsf{m})} \mid Y_i^{(\mathsf{m})}=1 \sim \mathrm{Bern}(v_1^{(\mathsf{m})}) \quad \text{and} \quad S_i^{(\mathsf{m})} \mid Y_i^{(\mathsf{m})}=0 \sim \mathrm{Bern}(v_0^{(\mathsf{m})}) \quad \text{for} \quad [M/2] < m \le M.\end{aligned}$$

We consider two hyperparameter settings for the surrogate: (i) weak surrogates with $\mu_1^{(\mathsf{m})} = 3$, $\mu_0^{(\mathsf{m})} = 1$, $v_1^{(\mathsf{m})} = 0.75$, and $v_0^{(\mathsf{m})} = 0.25$; and (ii) strong surrogates with $\mu_1^{(\mathsf{m})} = 5$, $\mu_0^{(\mathsf{m})} = 1$, $v_1^{(\mathsf{m})} = 0.85$, and $v_0^{(\mathsf{m})} = 0.15$. Because Poisson residuals are sub-exponential rather than sub-Gaussian, the Poisson component of this simulation evaluates empirical robustness beyond the scope of Assumption 4; the Bernoulli component is covered by the bounded-residual case. Throughout, we summarize results based on 200 replications in each configuration.

4.2 Benchmark

For each simulated data set, we obtain the estimator for $\boldsymbol{\beta}_0$ using SASH as well as the following benchmark methods:

IPD The ideal IPD version of our framework introduced in Section 3.2, with its complete details included in Appendix A.3. Although it is not realistic to pool all individual data together under the privacy constraints, IPD is still an important benchmark method helping

evaluate how SASH loses efficiency due to aggregation based on summary statistics, which has been theoretically studied in Section 3.2.

SL The supervised learning estimator obtained by (4) only using labeled data $\mathscr{L}^{(1)}$.

SS-uLasso The semi-supervised (SS) version of unsupervised Lasso (**uLasso**) proposed by Chakrabortty et al. (2017). They developed an unsupervised learning method for the SIM coefficient that applies Lasso to the subset of unlabeled data restricted to some extreme quantile of the surrogate. To adapt their method, we first select subjects with the top 2% (or 5%) large or small $S^{(\mathsf{m})}$ at each site m, and define $\widetilde{Y}_i^{(\mathsf{m})} = 1$ if $S_i^{(\mathsf{m})}$ is extremely large, and $\widetilde{Y}_i^{(\mathsf{m})} = 0$ if $S_i^{(\mathsf{m})}$ is extremely small. Then we perform distributed sparse logistic regression (Cai et al., 2022b) for $\widetilde{Y}_i^{(\mathsf{m})} \sim \mathbf{X}_i^{(\mathsf{m})}$ on the selected samples, to obtain an integrative uLasso estimator for $\boldsymbol{\gamma}_0$. At last, we combine the estimated $\boldsymbol{\gamma}$ with the labeled samples in the same way as Step III of SASH to derive the final SS-uLasso estimator of $\boldsymbol{\beta}_0$.

SS-RMRCE The SS version of the regularized maximum rank correlation estimator (**RMRCE**) (Han et al., 2017) that maximizes the rank correlation between $S_i^{(\mathsf{m})}$ and $\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\gamma}$. We aggregate the local RMRCE by simple averaging, and, again, use Step III of SASH to obtain the SS version. Due to the prohibitively high computational cost of RMRCE under large p, we consider two versions: (i) the "oracle" version only including $\{X_{ij}^{(\mathsf{m})} : j = 1, \ldots, 8\}$ to fit the SIM, as if we knew the true set of active predictors; (ii) the "screening" version first selecting top 8 predictors of $S^{(\mathsf{m})}$ at each local site according to the absolute correlation between each $X_{ij}^{(\mathsf{m})}$ and $S_i^{(\mathsf{m})}$, then combining the M selected sets by majority voting, and finally fitting RMRCE on the selected predictors.

pLasso The prior Lasso method (Jiang et al., 2016) that leverages some external knowledge of $\boldsymbol{\beta}$ by adding to the logistic loss on the labeled data $\mathscr{L}^{(1)}$ a penalty of the discrepancy between the external knowledge and the target model. In our case, this external knowledge is naturally taken as the SS estimator obtained by calibrating some SIM estimator of $\boldsymbol{\gamma}$ with Step III of SASH. The SIM estimator is obtained using the adaptive Lasso (Zou, 2006) version of ℓ_1-regularized least squares regression (**L1LS**) (Neykov et al., 2016) prone to bias under non-Gaussian designs. Thanks to its shrinkage strategy, pLasso achieves some adaptivity to this bias but this strategy is still not optimal in terms of the SIM estimation. Note that one can directly implement L1LS under the privacy constraints since summary data $\widehat{\mathbb{P}}^{(\mathsf{m})}\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\widehat{\mathbb{P}}^{(\mathsf{m})}S\mathbf{X}$ are sufficient for L1LS.

PASS The prior adaptive semi-supervised learning method (Zhang et al., 2022) incorporating the SIM estimator by adaptively shrinking $\boldsymbol{\beta}$ to its direction with the ℓ_1-penalty when fitting logistic regression of $\mathscr{L}^{(1)}$. It uses the adaptive Lasso version of **L1LS** to estimate $\boldsymbol{\gamma}$.

The RMRCE method is subject to the non-convexity issue of SIM and both uLasso and L1LS encounter bias for non-Gaussian designs. One could note for our descriptions that to ensure fair comparisons, all methods are adapted to the SS setting and utilize unlabeled data from all sources. Additional implementation details (e.g., tuning) are included in Appendix D. Since the direction parameter $\boldsymbol{\gamma}$ is an important by-product, we also study SASH's estimation performance of $\boldsymbol{\gamma}$ and compare it with **IPD**, **L1LS**, **RMRCE**, and **uLasso**. As an ablation study, we include **SASH (w/o Step I)**, the SASH estimator starting by guessing $\boldsymbol{\gamma}_1 = 1, \boldsymbol{\gamma}_{-1} = 0$ in Algorithm 2 instead of using the supervised $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$

obtained in Step I. This procedure does not take advantage of the labeled samples to fit the regularized SIM. By Remark 3 and Section 3, using $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$ to initialize the SIM training is essential for the stability and efficiency of SASH. We compare SASH with SASH (w/o Step I) to further demonstrate this point.

4.3 Main Results

4.3.1 SASH shows desirable estimation performance

In Table 1 and Figure 2, we present the average and boxplot of the ℓ_2 estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ of SASH and other competing methods. Both IPD and SASH show better performance under strong surrogacy than weak surrogacy since the former has more informative surrogates to Y. Among all the privacy-preserving methods, SASH achieves the closest accuracy to the ideal IPD, with around 20% larger average estimation error than IPD under the weak surrogate setting and 10% larger error under the strong surrogacy. Compared to other semi-supervised methods, SASH attains around 10%–15% smaller error than PASS and SS-uLasso (two versions using 2% or 5% extreme $S^{(\mathsf{m})}$), 19%–25% smaller error than SS-RMRCE (the oracle and screening versions), and 40% smaller error than pLasso under weak surrogacy. Meanwhile, SASH shows even better performance than these existing methods under strong surrogacy. Lastly, since SL does not utilize the unlabeled data with surrogates, it performs the worst among all methods, with around 120% and 200% higher average estimation errors than SASH under the weak and strong surrogate settings, respectively.

	Weak surrogate	Strong surrogate
IPD	0.735	0.600
PASS	1.106	1.058
pLasso	1.560	1.640
SASH	0.949	0.661
SL	1.834	1.834
SS-RMRCE (screening)	1.299	1.276
SS-RMRCE (oracle)	1.168	1.122
SS-uLasso (2%)	1.096	1.060
SS-uLasso (5%)	1.038	0.841

Table 1: Average ℓ_2 estimation errors of $\boldsymbol{\beta}$ under the weak surrogate and strong surrogate settings with $N = 8000$, $M = 4$, $n = 200$, and $p = 300$. The results are based on 200 simulation replications.

SASH outperforms existing semi-supervised methods in terms of estimating $\boldsymbol{\beta}$ because our approach to learning $\boldsymbol{\gamma}$ utilizes the information from SIM of S more effectively than existing methods. To further demonstrate this point, we evaluate and compare the estimation performance on $\boldsymbol{\gamma}$ of the SIM methods introduced in Section 4.2; see Figure 3. SASH attains a substantially smaller estimation error than L1LS, uLasso, and RMRCE, as well as a fairly close performance to the ideal IPD estimator. For example, in the strong surrogacy scenario, the estimator of $\boldsymbol{\gamma}$ by SASH has around 30% smaller median error than uLasso and 40% smaller than RMRCE. This confirms that our way of fitting the sparse SIM is more effective than the existing methods. In addition, compared with SASH utilizing $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}}$, SASH

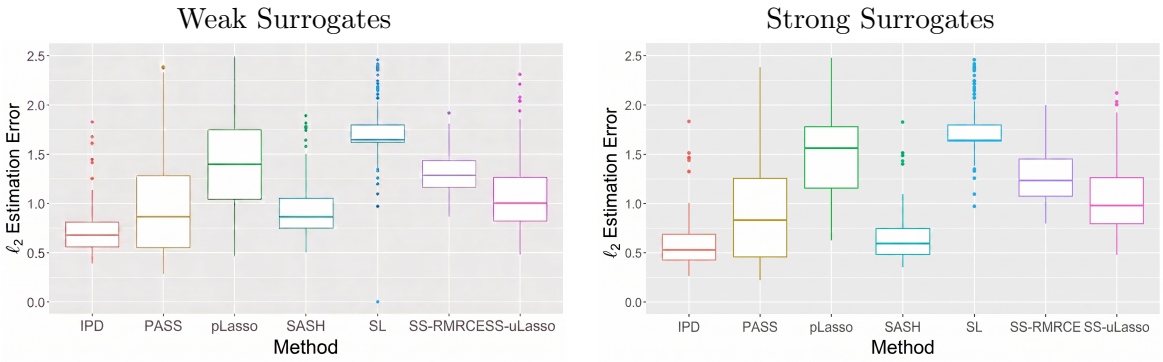

Figure 2: Boxplots of the ℓ_2 estimation errors of $\boldsymbol{\beta}$ under the strong surrogate and weak surrogate settings with $N = 8000$, $M = 4$, $n = 200$, and $p = 300$. For simplicity, we present the version of SS-uLasso with 2% extreme $S^{(\mathrm{m})}$ and SS-RMRCE with empirically screened predictors, with the other versions of them presented in Table 1. The methods under comparison are introduced in Section 4.2. The results are based on 200 simulation replications. SASH achieves the closest accuracy to the ideal IPD, with around 20% larger average estimation error than IPD under the weak surrogate setting and 10% larger error under the strong surrogate settings.

w/o Step I has much higher estimation errors on $\boldsymbol{\gamma}$, i.e., around 55% higher under weak surrogacy and 125% higher under strong surrogacy, implying that the supervised estimator obtained in Step I does play an important role in improving the performance of SASH.

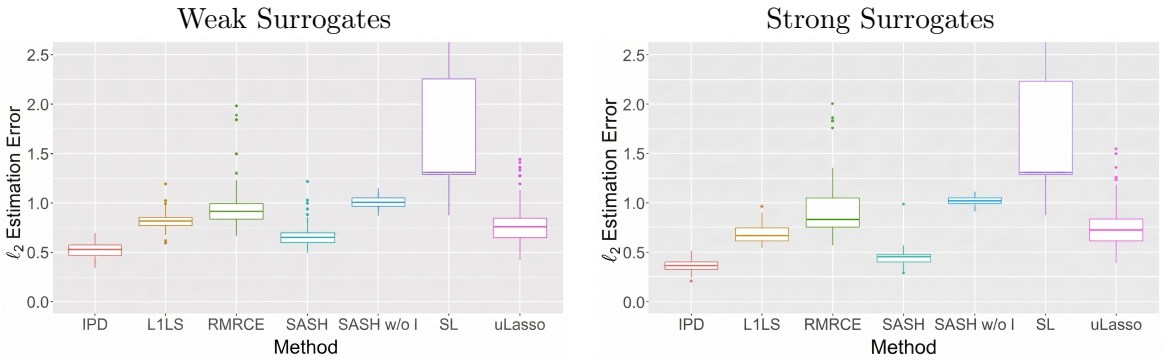

Figure 3: Boxplots of the ℓ_2 estimation errors of $\boldsymbol{\gamma}$ under the strong surrogate and weak surrogate settings with $N = 8000$, $M = 4$, $n = 200$, and $p = 300$. We present the version of uLasso with 2% extreme $S^{(\mathrm{m})}$ and RMRCE with empirically screened predictors. The methods under comparison are introduced in Section 4.2. The results are based on 200 simulation replications. SASH attains a fairly close performance to the ideal IPD estimator. In addition, compared with SASH, SASH w/o Step I has much higher estimation errors on $\boldsymbol{\gamma}$, implying that the supervised estimator obtained in Step I does play an important role in improving the performance of SASH.

4.3.2 Additional communication can improve efficiency

SASH is moderately outperformed by the ideal IPD estimator under strong surrogacy and the gap between their performance is even more pronounced with a weaker surrogate. This is due to the error caused by approximating the IPD loss with summary statistics and analyzed in Theorems 2 and 3. To further mitigate this error in finite-sample studies, we investigate a modified SASH estimator, denoted as SASH+, obtained by performing one more round of communication between site 1 and other sites. Specifically, we transfer the current SASH estimator of $\boldsymbol{\gamma}$ obtained in site 1 to all sites, use it to derive the summary statistics $\widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})}$ and $\widehat{\boldsymbol{\Omega}}_{x,s}^{(\mathsf{m})}$ as introduced in Step II, and transfer the updated statistics back to site 1 for the integrative regression in Step III. This strategy has been studied in existing distributed and federated learning methods (Jordan et al., 2019; Duan et al., 2022) to reduce the error occurring in data aggregation.

In Figure D.1 of Appendix D.3, we present the simulated ℓ_2 errors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of IPD, SASH, and SASH+, under both the strong and weak surrogate settings. Interestingly, we find that benefiting from the one additional round of communication, SASH+ reduces the error gap on $\boldsymbol{\beta}$ between SASH and IPD from 20% to 10% under weak surrogacy, and from 10% to 7% under strong surrogacy. These results demonstrate that more rounds of communication are promising to further improve SASH and make it closer to the ideal IPD.

4.3.3 CI estimation could be further improved

We also present CI coverage probabilities (CPs) for all nine truly nonzero coefficients produced based on IPD, SASH and SL in Table 2, constructed using the procedure introduced in Section 2.4. SASH attains a close coverage performance to IPD and is significantly better than plugging the SL estimator into the method introduced in Section 2.4. Specifically, both SASH and IPD maintain desirable coverage rate (around 95%) on more than half of the parameters in both scenarios while SL only shows good performance on two or three of them. This is due to the excessive regularization error of SL.

However, SASH and IPD attain lower coverage rates than the nominal level for the coefficients indexed by $\{1, 3, 4\}$. This is because the error of the SIM estimator $\widehat{\boldsymbol{\gamma}}$ brings about some non-ignorable bias in finite-sample studies. Our proposed inference approach in Section 2.4 is easy to implement but overlooks the regularization bias in $\widehat{\boldsymbol{\gamma}}$, subject to the condition that the unlabeled sample size N is much larger than n. More accurately, for valid asymptotic inference, our method requires the error rate of the sparse SIM, i.e., $\{s \log p/N\}^{1/2}$ by Theorem 2, to be much smaller than $n^{-1/2}$. To overcome this issue and enhance our inference procedure in both the asymptotic regime and finite-sample study, one could incorporate the debiasing method of sparse SIMs (Wu et al., 2023) to adjust for the regularization bias of $\widehat{\boldsymbol{\gamma}}$. This warrants future studies.

4.3.4 Performance of SASH is not sensitive to choices of h

We also conduct an additional simulation study to investigate the sensitivity of SASH to different choices of the kernel bandwidth parameter h fixed as $h_0 = \{s \log(p \vee N_m)/N_m\}^{1/5}$ in our main simulation studies. The descriptions and results of this study are presented in Appendix D.4. Figure D.2 demonstrates that different versions of SASH with $h \in \{0.5h_0, h_0, 2h_0\}$ have similar estimation performance. For example, the mean estimation

Setting	Method	β_j								
		1	2	3	4	5	6	7	8	9
Weak Surrogates	IPD	0.85	0.95	0.89	0.91	0.94	0.94	0.95	0.96	0.92
	SASH	0.84	0.97	0.87	0.91	0.94	0.94	0.95	0.96	0.92
	SL	0.66	0.84	0.79	0.84	0.89	0.94	0.92	0.97	0.91
Strong Surrogates	IPD	0.86	0.95	0.88	0.91	0.94	0.94	0.95	0.96	0.94
	SASH	0.86	0.97	0.87	0.93	0.96	0.95	0.95	0.96	0.93
	SL	0.66	0.84	0.78	0.84	0.88	0.95	0.92	0.97	0.91

Table 2: Coverage probability of the CIs for $\boldsymbol{\beta}$ under the weak surrogate and strong surrogate settings with $N = 8000$, $M = 4$, $n = 200$, and $p = 300$. CIs are constructed using the procedure introduced in Section 2.4. SASH and IPD maintain desirable coverage rate (around 95%) on more than half of the parameters in both scenarios, and are significantly better than plugging the SL estimator into the method introduced in Section 2.4.

error of SASH with $h = 2h_0$ is 4.1% smaller than that of h_0. Compared to the relative efficiency between SASH and other benchmark methods, this discrepancy caused by the changes of h is substantially smaller, even with $h \in \{0.5h_0, h_0, 2h_0\}$ varying in a wide enough range. This means that our method is not sensitive to the specification of h.

5. Real example

Due to the increasing linkage of EHR with bio-repositories, large-scale biobank data have emerged as an important and rich resource for risk prediction modeling (Kho et al., 2011). For example, both UK Biobank (UKB) and Massachusetts General and Brigham (MGB) Healthcare Biobank contain not only a wealth of medical and clinical records but also their linked genomic data, which enables us to derive genetic risk prediction models useful for personalized medicine (Lewis and Vassos, 2020). However, there exist statistical challenges impeding the effective use of EHR-linked biobank data in this application. Among them, the most prominent obstacle is the lack of accurate or gold-standard labels on disease status in EHR. For example, patients may receive diagnostic codes for type 2 diabetes (T2D) even if they do not actually have the disease. Precise information on T2D status requires human curation via manual chart review, which is not scalable and often infeasible due to the lack of access to clinical notes. Together with high-dimensional genetic features, it is challenging if not infeasible to derive reliable genetic risk prediction models based on labeled data sets. On the other hand, diagnostic codes and other patient-reported outcomes are readily available yet inaccurate. With data from multiple institutions and the availability of gold-standard labels for a small subset of patients, surrogate-assisted and federated approaches like SASH have the potential to address these challenges and produce robust and efficient genetic risk models using biobank data.

To illustrate the potential of these surrogate-assisted federated methods, we develop a T2D genetic risk prediction model using four sets of biobank data from two healthcare systems, the UK Biobank (UKB) and the Mass General Brigham (MGB). Since the preva-

lence of T2D differs greatly by age and we only have gold-standard labels on T2D status on older patients, we further partitioned UKB and MGB data into ages ≥ 60 and < 60. This results in $N_1 = 10{,}685$ and $N_2 = 7{,}653$ for $\text{MGB}_{\geq 60}$ and $\text{MGB}_{<60}$; and $N_3 = 211{,}061$ and $N_4 = 275{,}730$ for $\text{UKB}_{\geq 60}$ and $\text{UKB}_{<60}$. Our target model includes gender, ethnicity, and 336 T2D-associated single nucleotide polymorphisms (SNPs) reported in previous studies (Kozak and Anunciado-Koza, 2009; Rodrigues et al., 2013) in covariates $\mathbf{X}$. For the surrogate S, we took it as the log-transformed total count of the ICD codes associated with T2D for the MGB subjects, and the binary response to the T2D screening question for those in UKB. The surrogate's distribution and its informativeness to Y could differ greatly between MGB and UKB due to the difference in their definition. Within $\text{MGB}_{\geq 60}$, we have an additional $n = 277$ patients with gold labels Y on the true T2D status obtained via manual chart review.

The area under the receiver operating characteristic curve (AUC) of the surrogate S for classifying Y on $\text{MGB}_{\geq 60}$ is around 0.85. This indicates that the surrogate S is error-prone and needs to be properly modeled although $P(Y = 0 \mid S = 0)$ is almost one, indicating that S is a reasonable negative predictor. Although the sample sizes of the unlabeled data differ greatly among those four data sets, the number of potential cases with $S > 0$ does not differ as substantially, with 3,081, 875, 14,992, and 9,393 in $\text{MGB}_{\geq 60}$, $\text{MGB}_{<60}$, $\text{UKB}_{\geq 60}$ and $\text{UKB}_{<60}$ respectively. Similar to Section 4, we implement SL, SASH, pLasso, PASS, SS-uLasso, and SS-RMRCE to construct the T2D genetic risk model (1) while the IPD estimator is not feasible to construct due to the DataSHIELD constraint. For model evaluation, we consider several commonly used accuracy measures including the AUC, the Brier score (BS) defined as the mean squared prediction error of Y, and the logistic deviance of the estimator. These accuracy quantities are computed through 5-fold CV on the labeled samples and presented in Table 3. As a result of effectively using the unlabeled surrogate samples, SASH performs better than all benchmark methods on all of our evaluation metrics. For example, compared with pLasso, our method attains a 20% larger AUC, a 12% smaller BS, as well as a 13% smaller deviance.

	PASS	pLasso	SASH	SL	SS-RMRCE	SS-uLasso
AUC	0.533	0.563	**0.673**	0.556	0.592	0.638
BS	0.146	0.142	**0.125**	0.130	0.133	0.128
deviance	0.982	0.940	**0.822**	0.852	0.870	1.054

Table 3: AUC, Brier score (BS), and deviance of the approaches under comparison in our real-world application evaluated via 5-fold CV.

We also present features assigned nonzero estimated coefficients by at least one of the compared methods in Table D.1 of Appendix D.5. Interestingly, SASH returns 46 nonzero coefficients while the other methods select at most 16 of them. Meanwhile, SASH shows less shrinkage to zero compared to the benchmark methods. Both the larger detection set and the less shrinkage of SASH are also benefits of using the unlabeled surrogate samples to learn the direction of $\boldsymbol{\beta}$ more effectively and accurately.

6. Conclusion and Discussion

We developed SASH, a novel surrogate-assisted federated learning approach that leverages large unlabeled data with surrogates from multiple local sites to improve the efficiency of statistical learning with labeled data. Our method is efficient in terms of communication and protects individual-level information by aggregating appropriate summary statistics. It is shown to match the first-order convergence rates of, and to be numerically close to, the IPD estimator, which is ideal but only available without the DataSHIELD constraints. In particular, our newly proposed Algorithm 2 ensures the convergence of the local sparse SIM estimator by using a sufficiently accurate supervised estimator and varying the penalty parameter adaptively during the iteration. In the existing literature on SIM, such a convergence property is not readily achieved for non-Gaussian design due to the non-convexity of the SIM loss function. This novel algorithm, as well as the private data aggregation (11), is based on our efficient one-step approximation theory of high-dimensional sparse SIM. As we remarked, this development is intrinsically relevant to the efficient score and debiased inference of SIM (Liang et al., 2010; Wu et al., 2023).

More generally speaking, any preliminary estimator consistent with $\boldsymbol{\gamma}_0$, e.g., prior knowledge of $\boldsymbol{\gamma}$ derived from some external source data, can be used as a warm start for Algorithm 2 to enable the desirable convergence. With the availability of such (decent) prior knowledge, our strategy of iteratively solving the (debiased) one-step approximation could be generalized to overcome the non-convex optimization issue of other semiparametric regression problems such as maximum rank correlation (Han, 1987), and sliced inverse regression (Li, 1991). Inspired by Remark 3, this extension requires one to derive the forms of the bias-corrected quadratic approximation for those more complicated regression problems.

By Remark 1, our method requires prior knowledge of at least one strong enough risk factor with coefficient β_{01} to ensure that $\boldsymbol{\gamma}_0 = \boldsymbol{\beta}_0/\beta_{01}$ can be well-estimated. Specifically, $|\beta_{01}|$ needs to be larger than $(s \log p/n)^{1/2}$ to guarantee the convergence of Algorithm 2. Although data-driven feature screening methods (e.g., Tibshirani, 1996; Fan and Lv, 2008) could be used to address the absence of such prior knowledge, they may not guarantee a correct solution without additional regularity and signal strength assumptions. We also note that some previous work uses $\|\boldsymbol{\beta}_0\|_2$ instead of β_{01} to standardize $\boldsymbol{\beta}_0$, i.e., taking $\boldsymbol{\gamma}_0 = \boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$. This strategy can work without our assumption on β_{01} but its corresponding constraint $\|\boldsymbol{\gamma}\|_2 = 1$ will make the optimization of SIM non-convex. Thus, we could modify our identifying condition on $\boldsymbol{\gamma}$ to remove the assumption on β_{01} but we need to be more careful about the quality of SIM training. More instance-dependent theoretical analysis could refine the convergence rate of SASH and relax some of our technical assumptions in certain cases. For example, we now assume in Assumption 2 that the unknown link function f_m and the probability density function of $\boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})}$ are two or three times differentiable. When those functions have higher-order smoothness properties, the kernel approximation bias will become smaller and our other assumptions like the sparsity $s = O(N_m^{1/6}/\log(p \vee N_m))$ could be potentially refined and relaxed. Also, the approximate sparsity (i.e., ℓ_r-sparse for some $r \in (0, 1)$) regime and the bounded design scenario can be studied to further extend or refine our convergence rate analyses.

Our method accommodates heterogeneous link functions f_m for $m = 1, 2, \ldots, M$, which allows the aggregation of different types of surrogates across local sites. This is in a similar

spirit to some recent work of angle-based federated or transfer learning (e.g., Gu et al., 2022; Tian et al., 2023). However, we do not consider the scenario with heterogeneous model coefficients. This problem has been addressed in the federated learning of GLM by Cai et al. (2022b) and others. It would be interesting to explore surrogate-assisted federated learning in this direction. Also, it warrants future research to enhance robustness to the violation of the SIM assumption (2), or more generally speaking, misleading surrogates. Some recent work of robust SAS (Hou et al., 2021; Zhang et al., 2022) could be inspiring and useful. In EHR and e-commerce applications, there is often a need to handle multiple or even high-dimensional surrogates to ensure statistical efficiency (e.g., Hou et al., 2021; Cai et al., 2022a). SASH can be extended to the multi-surrogate problem through the ensemble of surrogates' SIM objective functions. When the surrogates have dependence structures and different strengths, finding the optimal ensemble strategy is important and warrants future studies.

Acknowledgment

The authors declare no funding.

References

Nicola Abate and Manisha Chandalia. The impact of ethnicity on type 2 diabetes. *Journal of Diabetes and its Complications*, 17(1):39–58, 2003.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Susan Athey, Raj Chetty, and Guido Imbens. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, Ziwei Zhu, et al. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

Brett K Beaulieu-Jones, Samuel G Finlayson, William Yuan, Russ B Altman, Isaac S Kohane, Vinay Prasad, and Kun-Hsing Yu. Examining the use of real-world evidence in the regulatory process. *Clinical Pharmacology & Therapeutics*, 107(4):843–852, 2020.

Peter J Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

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

Tianxi Cai, Yichi Zhang, Yuk-Lam Ho, Nicholas Link, Jiehuan Sun, Jie Huang, Tianrun A Cai, Scott Damrauer, Yuri Ahuja, Jacqueline Honerlaw, et al. Association of interleukin 6 receptor variant with cardiovascular disease effects of interleukin 6 receptor blocking therapy: a phenome-wide association study. *JAMA cardiology*, 3(9):849–857, 2018.

Tianxi Cai, T Tony Cai, and Zijian Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):669–719, 2021.

Tianxi Cai, Mengyan Li, and Molei Liu. Semi-supervised triply robust inductive transfer learning. *arXiv preprint arXiv:2209.04977*, 2022a.

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, 117(540):2105–2119, 2022b.

Abhishek Chakrabortty, Matey Neykov, Raymond Carroll, and Tianxi Cai. Surrogate aided unsupervised recovery of sparse signals in single index models for binary outcomes. *arXiv preprint arXiv:1701.05230*, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. Technical Report 1, 01 2018. URL `https://doi.org/10.1111/ectj.12097`.

Dany Doiron, Paul Burton, Yannick Marcon, Amadou Gaye, Bruce HR Wolffenbuttel, Markus Perola, Ronald P Stolk, Luisa Foco, Cosetta Minelli, Melanie Waldenberger, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerging themes in epidemiology*, 10(1):12, 2013.

Rui Duan, Yang Ning, and Yong Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83, 2022.

Hamid Eftekhari, Moulinath Banerjee, and Yaacov Ritov. Inference in high-dimensional single-index models under symmetric designs. *J. Mach. Learn. Res.*, 22(27):1–63, 2021.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Jessica M Franklin, Kai-Li Liaw, Solomon Iyasu, Cathy W Critchlow, and Nancy A Dreyer. Real-world evidence to support regulatory decision making: new or expanded medical product indications. *Pharmacoepidemiology and Drug Safety*, 30(6):685–693, 2021.

Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6):1929–1944, 2014.

Alon Geva, Molei Liu, Vidul A Panickan, Paul Avillach, Tianxi Cai, and Kenneth D Mandl. A high-throughput phenotyping algorithm is portable from adult to pediatric populations. *Journal of the American Medical Informatics Association*, 28(6):1265–1269, 2021.

Tian Gu, Yi Han, and Rui Duan. Robust angle-based transfer learning in high dimensions. *arXiv preprint arXiv:2210.12759*, 2022.

Zijian Guo, Prabrisha Rakshit, Daniel S Herman, and Jinbo Chen. Inference for the case probability in high-dimensional logistic regression. *The Journal of Machine Learning Research*, 22(1):11480–11533, 2021.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

Fang Han, Hongkai Ji, Zhicheng Ji, and Honglang Wang. A provable smoothing approach for high dimensional generalized regression with applications in genomics. *Electronic Journal of Statistics*, 11(2):4347–4403, 2017.

Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and Bouchra El Asri. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042, 2021.

Qianchuan He, Hao Helen Zhang, Christy L Avery, and DY Lin. Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2):205–220, 2016.

Chuan Hong, Katherine P Liao, and Tianxi Cai. Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics*, 75(1):78–89, 2019.

Chuan Hong, Everett Rush, Molei Liu, Doudou Zhou, Jiehuan Sun, Aaron Sonabend, Victor M Castro, Petra Schubert, Vidul A Panickan, Tianrun Cai, et al. Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data. *NPJ digital medicine*, 4(1):1–11, 2021.

Jue Hou, Rajarshi Mukherjee, and Tianxi Cai. Efficient and robust semi-supervised estimation of ate with partially annotated treatment and response. *arXiv preprint arXiv:2110.12336*, 2021.

Jing Huang, Rui Duan, Rebecca A Hubbard, Yonghui Wu, Jason H Moore, Hua Xu, and Yong Chen. Pie: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, 25(3):345–352, 2018.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Yuan Jiang, Yunxiao He, and Heping Zhang. Variable selection with prior information for generalized linear models via the prior lasso method. *Journal of the American Statistical Association*, 111(513):355–376, 2016.

EM Jones, NA Sheehan, N Masca, SE Wallace, MJ Murtagh, and PR Burton. DataSHIELD–shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk epidemiologi*, 21(2), 2012.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

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

Abel N Kho, Jennifer A Pacheco, Peggy L Peissig, Luke Rasmussen, Katherine M Newton, Noah Weston, Paul K Crane, Jyotishman Pathak, Christopher G Chute, Suzette J Bielinski, et al. Electronic medical records for genetic research: results of the emerge consortium. *Science translational medicine*, 3(79):79re1–79re1, 2011.

Isaac S Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417–428, 2011.

LP Kozak and R Anunciado-Koza. Ucp1: its involvement and utility in obesity. *International journal of obesity*, 32(S7):S32, 2009.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):1–11, 2020.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Hua Liang, Xiang Liu, Runze Li, and Chih-Ling Tsai. Estimation and testing for partially linear single-index models. *Annals of statistics*, 38(6):3811, 2010.

Jason Z Lin and Jelena Bradic. Learning to combat noisy labels via classification margins. *arXiv preprint arXiv:2102.00751*, 2021.

Enrico Maiorino and Joseph Loscalzo. Phenomics and robust multiomics data for cardiovascular disease subtyping. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 43(7): 1111–1123, 2023.

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

Matey Neykov, Jun S Liu, and Tianxi Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *The Journal of Machine Learning Research*, 17(1):2976–3012, 2016.

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

Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

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.

AC Rodrigues, B Sobrino, FDV Genvigir, MAV Willrich, SS Arazi, EL Dorea, MMS Bernik, M Bertolami, AA Faludi, MJ Brion, et al. Genetic variants in genes related to lipid metabolism and atherosclerosis, dyslipidemia and atorvastatin response. *Clinica Chimica Acta*, 417:8–11, 2013.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

Ye Tian, Yuqi Gu, and Yang Feng. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*, 2023.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

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

Tyler J VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69(3): 561–565, 2013.

Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

Chen Wang, Yuanchang Xie, Helai Huang, and Pan Liu. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. *Accident Analysis & Prevention*, 157:106157, 2021.

Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H Chi, and Minmin Chen. Surrogate for long-term user experience in recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4100–4109, 2022.

Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. DataSHIELD: resolving a conflict in contemporary bioscience performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382, 2010.

Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. Turning clicks into purchases: Revenue optimization for product search in e-commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 365–374, 2018.

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

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.

Lingjiao Zhang, Xiruo Ding, Yanyuan Ma, Naveen Muthu, Imran Ajmal, Jason H Moore, Daniel S Herman, and Jinbo Chen. A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. *Journal of the American Medical Informatics Association*, 27(1):119–126, 2020a.

Yichi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk-Lam Ho, Ashwin N Ananthakrishnan, Zongqi Xia, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature protocols*, 14(12):3426–3444, 2019.

Yichi Zhang, Molei Liu, Matey Neykov, and Tianxi Cai. Prior adaptive semi-supervised learning with application to ehr phenotyping. *arXiv preprint arXiv:2003.11744*, 2020b.

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.

Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendix A. Proof of the Main Theorems

A.1 Proof Outline

Before proving the main theorems, we first give an outline of the main steps in the proofs.

S1 We use Lemmas B.9 and B.10 to control kernel estimation error.

S2 We use Lemmas B.3 and B.4 bound the additional higher-order error terms introduced by kernel approximation.

S3 We establish the restricted curvature condition in Lemma B.8 on the normalized tangent space $\mathbb{V}_0$ specified in Assumption 5.

S4 To prove Theorem 1, since $\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}$ minimizes the loss function $Q(\cdot)$ defined in (A.1), we start from the basic inequality $Q(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}) \leq Q(\boldsymbol{\gamma}_0)$, and use the results of S1, S2 and S3.

S5 To prove Theorem 2, we still utilize the basic inequality $Q(\widehat{\boldsymbol{\gamma}}) \leq Q(\boldsymbol{\gamma}_0)$ to bound $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$, where the quadratic loss function $Q(\cdot)$ in aggregation step is defined in (A.7).

S6 To prove the rate comparison in Theorem 3, following a strategy similar to that used in Cai et al. (2022b), we start from the basic inequality $Q(\widehat{\boldsymbol{\gamma}}) \leq Q(\widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}})$ to compare $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}$, as $\widehat{\boldsymbol{\gamma}}$ minimizes quadratic loss function $Q(\cdot)$ in aggregation, and leverage the fact that $\widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}$ minimizes the individual-level loss function to simplify the basic inequality. Since both $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}$ satisfy the normalization $\gamma_1 = 1$, their difference has first coordinate zero, so the curvature condition in Assumption 5 applies directly to the comparison direction.

A.2 Proof of Theorem 1

Proof In Step II, we update the estimator for $\boldsymbol{\gamma}_0$ (direction of $\boldsymbol{\beta}_0$) iteratively in Algorithm 2. Without loss of generality, we consider the m-th site ($1 \le m \le M$) and the t-th iteration. For the theoretical analysis, we take $\widehat{\phi}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) = 1$. Lemma 1 gives the initial membership $\widehat{\boldsymbol{\gamma}}_{[0]}^{(\mathsf{m})} \in \Gamma$, and the recursion below keeps all iterates in Γ on the same high-probability event.

Recall that we defined $Q^{(\mathsf{m})}(\boldsymbol{\gamma}; \widetilde{\boldsymbol{\gamma}}) = \widehat{\mathbb{P}}^{(\mathsf{m})} \left\{ S - \widehat{f}_m(\mathbf{X}; \widetilde{\boldsymbol{\gamma}}) - (\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}})^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}; \widetilde{\boldsymbol{\gamma}}) \right\}^2$. Here we use $Q^{(\mathsf{m})}_{\lambda_t^{(\mathsf{m})}}(\boldsymbol{\gamma}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})$ to denote the loss function in Site m at iteration t with penalty that

$$
\begin{aligned}
&Q^{(\mathsf{m})}_{\lambda_t^{(\mathsf{m})}}(\boldsymbol{\gamma}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \\
=:& N_m^{-1} \sum_{i=1}^{N_m} \left\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) - (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \right\}^2 + \lambda_t^{(\mathsf{m})} \|\boldsymbol{\gamma}\|_1 \\
=& N_m^{-1} \sum_{i=1}^{N_m} \left\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) + \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) - (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \right\}^2 + \lambda_t^{(\mathsf{m})} \|\boldsymbol{\gamma}\|_1 \\
=& N_m^{-1} \sum_{i=1}^{N_m} \left\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) + (\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \int_0^1 \partial_{\boldsymbol{\gamma}} \widehat{f}_m\big(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})} + \tau(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\big) \mathrm{d}\tau \right. \\
& \left. - (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \right\}^2 + \lambda_t^{(\mathsf{m})} \|\boldsymbol{\gamma}\|_1 \\
=& N_m^{-1} \sum_{i=1}^{N_m} \left\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) + (\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \right. \\
& \left. \cdot \int_0^1 \Big(\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})} + \tau(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})) - \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \Big) \mathrm{d}\tau - (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \right\}^2 + \lambda_t^{(\mathsf{m})} \|\boldsymbol{\gamma}\|_1,
\end{aligned}
\tag{A.1}
$$

where the third equality comes from Taylor expansion that

$$
\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) = (\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \int_0^1 \partial_{\boldsymbol{\gamma}} \widehat{f}_m\big(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})} + \tau(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\big) \mathrm{d}\tau.
$$

From the optimization procedure in Step II at iteration t that $\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}$ minimize the loss function $Q^{(\mathsf{m})}_{\lambda_t^{(\mathsf{m})}}(\cdot; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})$, we have the basic inequality $Q^{(\mathsf{m})}_{\lambda_t^{(\mathsf{m})}}(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \le Q^{(\mathsf{m})}_{\lambda_t^{(\mathsf{m})}}(\boldsymbol{\gamma}_0; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})$, i.e.,

$$
\begin{aligned}
& N_m^{-1} \sum_{i=1}^{N_m} \left\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) + (\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \right. \\
& \left. \cdot \int_0^1 \Big(\partial_{\boldsymbol{\gamma}} \widehat{f}_m\big(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})} + \tau(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\big) - \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \Big) \mathrm{d}\tau - (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \right\}^2 + \lambda_t^{(\mathsf{m})} \|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}\|_1 \\
\le & N_m^{-1} \sum_{i=1}^{N_m} \left\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) + (\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^\intercal \cdot \int_0^1 \Big(\partial_{\boldsymbol{\gamma}} \widehat{f}_m\big(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})} + \tau(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\big) - \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \Big) \mathrm{d}\tau \right\}^2 + \lambda_t^{(\mathsf{m})} \|\boldsymbol{\gamma}_0\|_1,
\end{aligned}
$$

which further gives

$$
\begin{aligned}
&\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\Big\}^2\\
\le&\frac{2}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\cdot\Big\{S_i^{(\mathsf{m})}-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol\gamma_0)\\
&\quad+(\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})^\intercal\int_0^1\Big(\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})}))-\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}+\lambda_t^{(\mathsf{m})}(\|\boldsymbol\gamma_0\|_1-\|\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}\|_1)\\
=&\frac{2}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\cdot\Big\{S_i^{(\mathsf{m})}-f_m(\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})+f_m(\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol\gamma_0)\\
&\quad+(\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})^\intercal\int_0^1\Big(\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})}))-\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}+\lambda_t^{(\mathsf{m})}(\|\boldsymbol\gamma_0\|_1-\|\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}\|_1)\\
=&\frac{2}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal f_m'(\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\Big)\epsilon_i^{(\mathsf{m})}\\
&\quad+\frac{2}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal\Big\{\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})})-f_m'(\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\Big)\Big\}\epsilon_i^{(\mathsf{m})}\\
&\quad+\frac{2}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\cdot\Big\{f_m(\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol\gamma_0)\Big\}\\
&\quad+\frac{2}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0)^\intercal\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\cdot\Big\{(\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})^\intercal\int_0^1\Big(\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol\gamma_0-\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})-\partial_{\boldsymbol\gamma}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol\gamma}_{[t-1]}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}\\
&\quad+\lambda_t^{(\mathsf{m})}(\|\boldsymbol\gamma_0\|_1-\|\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}\|_1):=A_1+A_2+A_3+A_4+\lambda_t^{(\mathsf{m})}(\|\boldsymbol\gamma_0\|_1-\|\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}\|_1),
\end{aligned}
\tag{A.2}
$$

where $\epsilon_i^{(\mathsf{m})}:=S_i^{(\mathsf{m})}-f_m(\boldsymbol\gamma_0^\intercal\mathbf{X}_i^{(\mathsf{m})})$. Denote $\Delta_t^{(\mathsf{m})}:=\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}-\boldsymbol\gamma_0$. By Lemmas B.1 - B.4, we have

$$
\begin{aligned}
|A_1|&\lesssim\|\Delta_t^{(\mathsf{m})}\|_1\sqrt{\frac{\log p}{N_m}},\\
|A_2|&\lesssim\|\Delta_t^{(\mathsf{m})}\|_1\sqrt{\frac{\log p}{N_m}},\\
|A_3|&\lesssim\|\Delta_t^{(\mathsf{m})}\|_1\sqrt{\frac{1}{N_m}}+\|\Delta_t^{(\mathsf{m})}\|_1h_m^3+\|\Delta_t^{(\mathsf{m})}\|_2\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m^2,\\
|A_4|&\lesssim\|\Delta_t^{(\mathsf{m})}\|_2\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m+\|\Delta_t^{(\mathsf{m})}\|_2\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2.
\end{aligned}
$$

Combining the above bounds with Lemma B.8, we have

$$
\begin{aligned}
\|\Delta_t^{(\mathsf{m})}\|_2^2&\lesssim\|\Delta_t^{(\mathsf{m})}\|_1\Big(\sqrt{\frac{\log p}{N_m}}+\sqrt{\frac{1}{N_m}}+h_m^3\Big)+\|\Delta_t^{(\mathsf{m})}\|_2\Big(\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m^2+\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m+\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2\Big)+\lambda_t^{(\mathsf{m})}(\|\boldsymbol\gamma_0\|_1-\|\widehat{\boldsymbol\gamma}_{[t]}^{(\mathsf{m})}\|_1)\\
&\lesssim\sqrt{s}\|\Delta_t^{(\mathsf{m})}\|_2\sqrt{\frac{\log p}{N_m}}+\|\Delta_t^{(\mathsf{m})}\|_2\Big(\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m+\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2\Big)+\lambda_t^{(\mathsf{m})}\sqrt{s}\|\Delta_t^{(\mathsf{m})}\|_2
\end{aligned}
\tag{A.3}
$$

by Lemmas B.6 and B.7, which gives

$$
\|\Delta_t^{(\mathsf{m})}\|_2\lesssim\sqrt{\frac{s\log p}{N_m}}+\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m+\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2+\lambda_t^{(\mathsf{m})}\sqrt{s},
\tag{A.4}
$$

under the assumption that $h_m\lesssim N_m^{-1/6}$.

By Lemma 1, the initial estimator satisfies $\|\Delta_0^{(\mathsf{m})}\|_2=\|\widehat{\boldsymbol\gamma}_{[0]}^{(\mathsf{m})}-\boldsymbol\gamma_0\|_2\lesssim\sqrt{s\log p/n\beta_{01}^2}$. We choose $\Big(\frac{s\log(p\vee N_m)}{N_m}\Big)^{1/5}\lesssim h_m\lesssim N_m^{-1/6}$ assumed in Theorem 1, and choose $\lambda_t^{(\mathsf{m})}$ satisfying

the assumption of Lemma B.5 that $\lambda_t^{(\mathsf{m})} \asymp \sqrt{\frac{\log p}{N_m}} + \|\Delta_{t-1}^{(\mathsf{m})}\|_2 h_m + \|\Delta_{t-1}^{(\mathsf{m})}\|_2^2$. Then by (A.4) we have

$$\|\Delta_t^{(\mathsf{m})}\|_2 \lesssim \begin{cases} \left(\sqrt{\frac{s\log p}{n\beta_{01}^2}}\right)^{2^t}, & \text{if } t < t'. \\ \sqrt{\frac{s\log p}{N_m}} + \left(\sqrt{\frac{s\log p}{n\beta_{01}^2}}\right)^{2^{t'}} h_m^{t-t'}, & \text{otherwise,} \end{cases} \tag{A.5}$$

where $t' = \log_2\Big(\log h_m / \log(\sqrt{\frac{s\log p}{n\beta_{01}^2}})\Big) + 1$. When $t < t'$, $\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2$ dominates the error in (A.4). And when $t \geq t'$, $\|\Delta_{t-1}^{(\mathsf{m})}\|_2 h_m$ dominates the error in (A.4) and we have $\|\Delta_t^{(\mathsf{m})}\|_2 \lesssim \sqrt{\frac{s\log p}{N_m}} + \left(\sqrt{\frac{s\log p}{n\beta_{01}^2}}\right)^{2^{t'}} h_m^{t-t'} \lesssim \sqrt{\frac{s\log p}{N_m}} + h_m^{t-t'+1}$ by the definition of t'. By the assumption that $\left(\frac{s\log(p\vee N_m)}{N_m}\right)^{1/5} \lesssim h_m \lesssim N_m^{-1/6}$, after the iteration t', we only need to iterate constant times to achieve $\|\Delta_t^{(\mathsf{m})}\|_2 \lesssim \sqrt{\frac{s\log p}{N_m}}$. So we have $T_m \asymp t' \asymp \log(-\log h_m) - \log\{\log(n/(s\log p)\}$ such that

$$\|\Delta_{T_m}^{(\mathsf{m})}\|_2 \lesssim \sqrt{\frac{s\log p}{N_m}},$$

and thus $\|\Delta_{T_m}^{(\mathsf{m})}\|_1 \lesssim s\sqrt{\frac{\log p}{N_m}}$ from Lemma B.6.

Finally to be specific, by (A.5), we choose

$$\lambda_t^{(\mathsf{m})} = \begin{cases} \left(\sqrt{\frac{s\log p}{n\beta_{01}^2}}\right)^{2^t}, & \text{if } t < t'. \\ \sqrt{\frac{\log p}{N_m}} + \left(\sqrt{\frac{s\log p}{n\beta_{01}^2}}\right)^{2^{t'}} h_m^{t-t'}, & \text{otherwise.} \end{cases} \tag{A.6}$$

■

A.3 Proof of Theorem 2

Proof In Step III, we first aggregate summary statistics from local sites to derive a temporary estimator for the direction $\boldsymbol{\gamma}_0$. Let the loss function be

$$J(\boldsymbol{\gamma}) = \frac{1}{N}\Big\{ -2\boldsymbol{\gamma}^\intercal \sum_{m=1}^{M} N_m \widehat{\boldsymbol{\Omega}}_{x,s}^{(\mathsf{m})} + \boldsymbol{\gamma}^\intercal \sum_{m=1}^{M} N_m \widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})} \boldsymbol{\gamma} \Big\} + \lambda\|\boldsymbol{\gamma}\|_1, \tag{A.7}$$

where the summary statistics are defined as

$$\widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})} = N_m^{-1} \sum_{i=1}^{N_m} \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) \Big\{ \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) \Big\}^\intercal,$$

$$\widehat{\boldsymbol{\Omega}}_{x,s}^{(\mathsf{m})} = N_m^{-1} \sum_{i=1}^{N_m} \Big\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) + \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})\intercal} \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) \Big\} \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}).$$

$\widehat{\boldsymbol{\Omega}}_{x,s}^{(\mathsf{m})}$ can be decomposed as

$$\widehat{\boldsymbol{\Omega}}_{x,s}^{(\mathsf{m})} = -\widehat{\Xi}^{(\mathsf{m})} + \widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})} \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}, \text{ where } \widehat{\Xi}^{(\mathsf{m})} = -N_m^{-1} \sum_{i=1}^{N_m} \Big\{ S_i^{(\mathsf{m})} - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}) \Big\} \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}). \tag{A.8}$$

By the basic inequality $J(\widehat{\boldsymbol{\gamma}}) \le J(\boldsymbol{\gamma}_0)$, we have

$$\frac{1}{N}\Big\{\sum_{m=1}^{M} N_m(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\Big\}+\lambda\|\widehat{\boldsymbol{\gamma}}\|_1 \le \frac{2}{N}\sum_{m=1}^{M} N_m(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\Big\{-\widehat{\Xi}^{(\mathsf{m})}+\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)\Big\}+\lambda\|\boldsymbol{\gamma}_0\|_1. \tag{A.9}$$

We first bound $\frac{1}{N}\sum_{m=1}^{M} N_m(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\Big\{-\widehat{\Xi}^{(\mathsf{m})}+\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)\Big\}$. We have

$$\begin{aligned}
&\frac{1}{N}\sum_{m=1}^{M} N_m(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\Big\{-\widehat{\Xi}^{(\mathsf{m})}+\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)\Big\}\\
=&\frac{1}{N}\sum_{m=1}^{M}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\sum_{i=1}^{N_m}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{S_i^{(\mathsf{m})}-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})+\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)\Big\}\\
=&\frac{1}{N}\sum_{m=1}^{M}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\sum_{i=1}^{N_m}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{S_i^{(\mathsf{m})}-f_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})+f_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\\
&\qquad+\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})+\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)\Big\}\\
=&\frac{1}{N}\sum_{m=1}^{M}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\sum_{i=1}^{N_m}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{\epsilon_i^{(\mathsf{m})}+f_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\\
&\qquad+(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})^{\mathsf{T}}\int_0^1\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}+\tau(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}))-\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}.
\end{aligned}$$

By similar argument in the proof of Theorem 1 and the assumption $h_m \lesssim N_m^{-1/6}$, we have

$$\begin{aligned}
&\Big|\frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\epsilon_i^{(\mathsf{m})}\Big| \lesssim \|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1\sqrt{\frac{\log p}{N}},\\
&\Big|\frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{f_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big\}\Big|\\
\lesssim&\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1\sqrt{\frac{1}{N}}+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\frac{1}{N}\sum_{m=1}^{M}N_m h_m^3+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\frac{1}{N}\sum_{m=1}^{M}N_m\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2 h_m^2 \lesssim \|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1\sqrt{\frac{1}{N}},
\end{aligned}$$

and

$$\begin{aligned}
&\Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^{\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big\{(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})^{\mathsf{T}}\int_0^1\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0+\tau(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}))-\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}\Big|\\
\lesssim&\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\frac{1}{N}\sum_{m=1}^{M}N_m\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2 h_m+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\frac{1}{N}\sum_{m=1}^{M}N_m\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2\\
\lesssim&\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\frac{1}{N}\sum_{m=1}^{M}N_m\sqrt{\frac{s\log p}{N_m}}h_m+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\frac{1}{N}\sum_{m=1}^{M}N_m\frac{s\log p}{N_m}\\
\lesssim&\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2\sqrt{\frac{s\log p}{N}},
\end{aligned}$$

where the first bound follows from Lemmas B.1 and B.2, the second bound follows from Lemma B.3, and the third bound follows from Lemma B.4. Let

$$\mathcal{R}_N=\frac{1}{N}\sum_{m=1}^{M}N_m\Big\{-\widehat{\Xi}^{(\mathsf{m})}+\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)\Big\}.$$

The same termwise bounds, before taking the inner product with $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$, imply that the leading stochastic part of $\mathcal{R}_N$ is $O_p\{\sqrt{\log p/N}\}$ in the ℓ_∞ norm, while the remaining higher-order terms are $o_p(\lambda)$ under the bandwidth and sparsity conditions of Theorem 1. Taking the constant in $\lambda \asymp \sqrt{\log p/N}$ sufficiently large and applying the basic inequality (A.9) yield

$$\|(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S^c}\|_1 \leq 3\|(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_S\|_1.$$

Because $\widehat{\gamma}_1 = \gamma_{01} = 1$, the aggregate error also has first coordinate zero. Consequently,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \leq 4\sqrt{s}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2.$$

Thus we have

$$\left|\frac{1}{N}\sum_{m=1}^{M} N_m(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal\Big\{ -\widehat{\Xi}^{(\mathsf{m})} + \widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)\Big\}\right| \lesssim \sqrt{s}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2\sqrt{\frac{\log p}{N}}, \tag{A.10}$$

which follows from similar argument in Lemma B.6.

Since $(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_1 = 0$, the normalized aggregate error belongs to $\mathbb{V}_0$ in Assumption 5. Therefore, by an argument identical to that in Lemma B.8, we have

$$\frac{1}{N}\left\{\sum_{m=1}^{M} N_m(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal\widehat{\boldsymbol{\Omega}}^{(\mathsf{m})}_{x,x}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\right\} \geq \kappa\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2. \tag{A.11}$$

Combining (A.9), (A.10) and (A.11), we have

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2 \lesssim \sqrt{s}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2\sqrt{\frac{\log p}{N}} + \lambda\sqrt{s}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2$$

since $\|\boldsymbol{\gamma}_0\|_1 - \|\widehat{\boldsymbol{\gamma}}\|_1 \lesssim \sqrt{s}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2$ by similar argument in Lemma B.7. So we finally have

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 \lesssim \sqrt{\frac{s\log p}{N}}$$

by taking $\lambda \asymp \sqrt{\log p/N}$.

Then plug $\widehat{\boldsymbol{\gamma}}$ in a low-dimensional logistic regression on labeled data to estimate intercept α and scalar β_1. Denoting $\boldsymbol{\theta} = (\alpha, \beta_1)$ and $\ell_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \ell\{Y_i^{(1)}, g(\alpha + \beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})\}$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\nabla^2_\theta\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\\
\leq &-\frac{2}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})\\
&-\frac{2}{n}\sum_{i=1}^{n}\int_0^1(1-t)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\Big\{\nabla^2_\theta\ell_i\big(\boldsymbol{\theta}_0 + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \widehat{\boldsymbol{\gamma}}\big) - \nabla^2_\theta\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})\Big\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\,\mathrm{d}t.
\end{aligned} \tag{A.12}$$

from basic inequality that $\frac{1}{n}\sum_{i=1}^n \ell_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) \leq \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})$ and the decomposition that

$$\begin{aligned}
\ell_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) &= \ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})\\
&\quad + \int_0^1(1-t)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\nabla^2_{\boldsymbol{\theta}}\ell_i\big(\boldsymbol{\theta}_0 + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \widehat{\boldsymbol{\gamma}}\big)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\,\mathrm{d}t\\
&= \ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\nabla^2_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\\
&\quad + \int_0^1(1-t)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\Big\{\nabla^2_{\boldsymbol{\theta}}\ell_i\big(\boldsymbol{\theta}_0 + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \widehat{\boldsymbol{\gamma}}\big) - \nabla^2_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}})\Big\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\,\mathrm{d}t.
\end{aligned}$$

where

$$
\begin{aligned}
&\ell_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \log(1+\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})) - Y_i^{(1)}(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)}),\\
&\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \Big(-Y_i + \frac{\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})}{1+\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})}\Big)\begin{pmatrix}1\\ \boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)}\end{pmatrix},\\
&\nabla^2_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})}{\big(1+\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})\big)^2}\begin{pmatrix}1\\ \boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)}\end{pmatrix}(1,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)}),\\
&\nabla^2_{\boldsymbol{\theta}\boldsymbol{\gamma}}\ell_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})}{\big(1+\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})\big)^2}\begin{pmatrix}1\\ \boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)}\end{pmatrix}\beta_1\mathbf{X}_i^{(1)\intercal} + \Big(-Y_i + \frac{\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})}{1+\exp(\alpha+\beta_1\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(1)})}\Big)\begin{pmatrix}\mathbf{0}\\ \mathbf{X}_i^{(1)\intercal}\end{pmatrix}.
\end{aligned}
\tag{A.13}
$$

First, by Lemma J.3 in Ning and Liu (2017), we have RE condition holds for some constant κ that

$$
\frac{1}{n}\sum_{i=1}^n(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\nabla^2_\theta\ell_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \geq \kappa\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2. \tag{A.14}
$$

Second, we bound $\frac{1}{n}\sum_{i=1}^n(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\gamma}})$. We have

$$
\Big|\frac{1}{n}\sum_{i=1}^n(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\Big| \leq \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1\Big\|\frac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\Big\|_\infty \lesssim \sqrt{\frac{1}{n}}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1 \tag{A.15}
$$

holds by Lemma E.2 in Ning and Liu (2017), and

$$
\begin{aligned}
&\Big|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\frac{1}{n}\sum_{i=1}^n\big(\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\gamma}})-\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\big)\Big|\\
=&\Big|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\frac{1}{n}\sum_{i=1}^n\int_0^1\nabla^2_{\boldsymbol{\theta}\boldsymbol{\gamma}}\ell_i(\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0+t(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0))\mathrm{d}t(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\Big|\\
\lesssim&\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2
\end{aligned}
\tag{A.16}
$$

follows from the expression of $\nabla^2_{\boldsymbol{\theta}\boldsymbol{\gamma}}\ell_i(\cdot)$ in (A.13). So

$$
\Big|\frac{1}{n}\sum_{i=1}^n(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\gamma}})\Big| \lesssim \sqrt{\frac{1}{n}}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1 + \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2
$$

Third, we have

$$
\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n\int_0^1(1-t)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\Big\{\nabla^2_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0+t(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),\widehat{\boldsymbol{\gamma}})-\nabla^2_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\gamma}})\Big\}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\mathrm{d}t\Big|\\
=&\Big|\frac{1}{n}\sum_{i=1}^n\int_0^1(1-t)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\intercal\Big\{\frac{\exp(a_t+b_t\widehat{\boldsymbol{\gamma}}^\intercal\mathbf{X}_i^{(1)})}{\big(1+\exp(a_t+b_t\widehat{\boldsymbol{\gamma}}^\intercal\mathbf{X}_i^{(1)})\big)^2}-\frac{\exp(a+b\widehat{\boldsymbol{\gamma}}^\intercal\mathbf{X}_i^{(1)})}{\big(1+\exp(a_0+b_0\widehat{\boldsymbol{\gamma}}^\intercal\mathbf{X}_i^{(1)})\big)^2}\Big\}\begin{pmatrix}1\\ \widehat{\boldsymbol{\gamma}}^\intercal\mathbf{X}_i^{(1)}\end{pmatrix}(1,\widehat{\boldsymbol{\gamma}}^\intercal\mathbf{X}_i^{(1)})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\mathrm{d}t\Big|\\
\lesssim&\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^3,
\end{aligned}
\tag{A.17}
$$

where $(a_t,b_t)^\intercal := \boldsymbol{\theta}_0+t(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$. Combining (A.12), (A.14), (A.15), (A.16) and (A.17), we have

$$
\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2 \lesssim \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1\sqrt{\frac{1}{n}} + \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2 + \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^3 \lesssim \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\Big(\sqrt{\frac{1}{n}}+\sqrt{\frac{s\log p}{N}}\Big) + \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^3,
$$

which immediately gives

$$
\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2 \lesssim \sqrt{\frac{1}{n}}+\sqrt{\frac{s\log p}{N}}. \tag{A.18}
$$

Finally, we aggregate the summary statistics from local sites and the labeled sample loss based on $(\widehat{\alpha}, \widehat{\beta}_1)$ to derive the final estimator for $\boldsymbol{\gamma}_0$. We have that the loss function

$$\mathcal{L}(\boldsymbol{\gamma}) = \frac{1}{N+n}\Big\{ -2\boldsymbol{\gamma}^\intercal \sum_{m=1}^{M} N_m \widehat{\boldsymbol{\Omega}}_{x,s}^{(\mathsf{m})} + \boldsymbol{\gamma}^\intercal \sum_{m=1}^{M} N_m \widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})} \boldsymbol{\gamma} + \sum_{i=1}^{n} \ell_i(\widehat{\alpha}, \widehat{\beta}_1, \boldsymbol{\gamma}) \Big\} + \lambda^\dagger \|\boldsymbol{\gamma}\|_1$$

and the basic inequality $\mathcal{L}(\widehat{\boldsymbol{\gamma}}^\dagger) \leq \mathcal{L}(\boldsymbol{\gamma}_0)$ gives

$$\begin{aligned}
&\frac{1}{N+n}\Big\{ \sum_{m=1}^{M} N_m (\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)^\intercal \widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})} (\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0) \\
&\qquad + \frac{1}{2} \sum_{i=1}^{n} (\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)^\intercal \nabla_{\boldsymbol{\gamma}}^2 \ell_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\gamma}_0)(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0) \Big\} + \lambda^\dagger \|\widehat{\boldsymbol{\gamma}}^\dagger\|_1 \\
\leq &\frac{1}{N+n}\Big\{ 2 \sum_{m=1}^{M} N_m (\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)^\intercal \Big\{ -\widehat{\Xi}^{(\mathsf{m})} + \widehat{\boldsymbol{\Omega}}_{x,x}^{(\mathsf{m})} (\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})} - \boldsymbol{\gamma}_0) \Big\} \\
&\quad - \sum_{i=1}^{n} (\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)^\intercal \nabla_{\boldsymbol{\gamma}} \ell_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\gamma}_0) \\
&\quad - \sum_{i=1}^{n} \int_0^1 (1-t)(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)^\intercal \Big\{ \nabla_{\boldsymbol{\gamma}}^2 \ell_i\big(\widehat{\boldsymbol{\theta}}, \boldsymbol{\gamma}_0 + t(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)\big) - \nabla_{\boldsymbol{\gamma}}^2 \ell_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\gamma}_0) \Big\} (\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)\, \mathrm{d}t \Big\} + \lambda^\dagger \|\boldsymbol{\gamma}_0\|_1.
\end{aligned}$$

The same score and penalty argument as above gives

$$\|(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)_{S^c}\|_1 \leq 3\|(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)_S\|_1.$$

Moreover, $(\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0)_1 = 0$, so the normalized error belongs to $\mathbb{V}_0$ in Assumption 5, and $\|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_1 \leq 4\sqrt{s}\|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_2$. which gives

$$\|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_2^2 \lesssim \frac{\sqrt{s}}{N+n}\|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_2 \Big\{ \sqrt{N \log p} + \sqrt{n \log p} \Big\} + \lambda^\dagger \sqrt{s} \|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_2$$

by an argument similar to the one used above. So we have

$$\|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_2 \lesssim \sqrt{s}\frac{\sqrt{N \log p} + \sqrt{n \log p}}{N+n} + \lambda^\dagger \sqrt{s} \lesssim \sqrt{\frac{s \log p}{N+n}} \lesssim \sqrt{\frac{s \log p}{N}} \tag{A.19}$$

by taking $\lambda^\dagger \asymp \sqrt{\frac{\log p}{N+n}} \asymp \sqrt{\frac{\log p}{N}}$. Combining (A.18) and (A.19), finally we have

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}} - \boldsymbol{\beta}_0\|_2 \lesssim \sqrt{\frac{1}{n}} + \sqrt{\frac{s \log p}{N}}.$$

Moreover,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_{\mathsf{SASH}} - \boldsymbol{\beta}_0\|_1 &\leq |\widehat{\beta}_1 - \beta_{01}|\, \|\boldsymbol{\gamma}_0\|_1 + |\widehat{\beta}_1|\, \|\widehat{\boldsymbol{\gamma}}^\dagger - \boldsymbol{\gamma}_0\|_1 \\
&\lesssim \sqrt{\frac{s}{n}} + s\sqrt{\frac{\log p}{N}},
\end{aligned}$$

where $\|\boldsymbol{\gamma}_0\|_1 \leq \sqrt{s}\|\boldsymbol{\gamma}_0\|_2$ and $\|\boldsymbol{\gamma}_0\|_2$ is bounded by Assumption 1. Both bounds hold with probability at least $1 - O(1/p)$.

■

A.4 Proof outline for Theorem 3

We present the IPD estimation procedure and the comparison argument in Section A.1. The proof uses the same score, cone, and derivative-weighted curvature bounds as Theorems 1 and 2. The displayed conclusion is deliberately stated as a one-sided first-order rate comparison rather than direct asymptotic equivalence of the two estimators.

Step I The first step is the same as our method, where we fit a supervised penalized logistic regression using labeled samples to obtain the initial estimator $\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}$.

Step II We still initialize with $\widehat{\boldsymbol{\gamma}}_{[0]} = \widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}/\widetilde{\beta}_{\mathsf{sup},1}$, then obtain the estimator $\widehat{\boldsymbol{\gamma}}_{[t]}$ iteratively for $t = 1, \ldots, T$ as:

$$\underset{\boldsymbol{\gamma}\in\mathbb{R}^p,\gamma_1=1}{\operatorname{argmin}} \sum_{m=1}^{M} \frac{N_m}{N} \widehat{\mathbb{P}}^{(\mathsf{m})} \left\{ S - \widehat{f}_m(\mathbf{X}; \widehat{\boldsymbol{\gamma}}_{[t-1]}) - (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{[t-1]})^{\mathsf{T}} \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}; \widehat{\boldsymbol{\gamma}}_{[t-1]}) \right\}^2 + \lambda_{[t]} \|\boldsymbol{\gamma}\|_1, \tag{A.20}$$

and take $\widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}} = \widehat{\boldsymbol{\gamma}}_{[T]}$.

Step III We leverage labeled data to obtain a final estimator for direction and magnitude. First, we plug $\widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}$ in a low-dimensional logistic regression on the labeled data, and estimate the intercept α and the scalar β_1:

$$(\widehat{\alpha}_{\mathsf{IPD}}, \widehat{\beta}_{\mathsf{IPD},1}) = \underset{\alpha,\beta_1}{\operatorname{argmin}} \widehat{\mathbb{P}}_n^{(1)} \ell\{Y, g(\alpha + \beta_1 \widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}^{\mathsf{T}} \mathbf{X})\}. \tag{A.21}$$

Finally, we aggregate all the labeled and unlabeled samples based on $(\widehat{\alpha}_{\mathsf{IPD}}, \widehat{\beta}_{\mathsf{IPD},1})$, to derive the final estimator for $\boldsymbol{\gamma}$:

$$\begin{aligned} \widehat{\boldsymbol{\gamma}}^{\dagger}_{\mathsf{IPD}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^p,\gamma_1=1}{\operatorname{argmin}} \frac{1}{N+n} \Bigg[& \sum_{m=1}^{M} N_m \widehat{\mathbb{P}}^{(\mathsf{m})} \Big\{ S - \widehat{f}_m(\mathbf{X}; \widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}) - (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}})^{\mathsf{T}} \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}; \widehat{\boldsymbol{\gamma}}_{\mathsf{IPD}}) \Big\}^2 \\ & + n \widehat{\mathbb{P}}_n^{(1)} \ell\{Y, g(\widehat{\alpha}_{\mathsf{IPD}} + \widehat{\beta}_{\mathsf{IPD},1} \boldsymbol{\gamma}^{\mathsf{T}} \mathbf{X})\} \Bigg] + \lambda_{\mathsf{IPD}} \|\boldsymbol{\gamma}\|_1. \end{aligned} \tag{A.22}$$

The IPD estimator is taken as $\widehat{\boldsymbol{\beta}}_{\mathsf{IPD}} = \widehat{\beta}_{\mathsf{IPD},1} \cdot \widehat{\boldsymbol{\gamma}}^{\dagger}_{\mathsf{IPD}}$.

Appendix B. Auxiliary Proofs

B.1 Proof of Proposition 1

Proof Under the outcome model (1) the conditional independence condition (3), for any $m \in \{1, \ldots, M\}$ and set $\mathcal{A}$ included by $S_i^{(\mathsf{m})}$'s domain, we have

$$\begin{aligned} \mathbb{P}(S_i^{(\mathsf{m})} \in \mathcal{A} \mid \mathbf{X}_i^{(\mathsf{m})}) &= \sum_{y=0,1} \mathbb{P}(S_i^{(\mathsf{m})} \in \mathcal{A}, Y_i^{(\mathsf{m})} = y \mid \mathbf{X}_i^{(\mathsf{m})}) \\ &= \sum_{y=0,1} \mathbb{P}(S_i^{(\mathsf{m})} \in \mathcal{A} \mid Y_i^{(\mathsf{m})} = y, \mathbf{X}_i^{(\mathsf{m})}) \mathbb{P}(Y_i^{(\mathsf{m})} = y \mid \mathbf{X}_i^{(\mathsf{m})}) \\ &= \sum_{y=0,1} \mathbb{P}(S_i^{(\mathsf{m})} \in \mathcal{A} \mid Y_i^{(\mathsf{m})} = y) \mathbb{P}(Y_i^{(\mathsf{m})} = y \mid \mathbf{X}_i^{(\mathsf{m})}) \\ &= \sum_{y=0,1} \mathbb{P}(S_i^{(\mathsf{m})} \in \mathcal{A} \mid Y_i^{(\mathsf{m})} = y) g_y(\alpha_0^{(\mathsf{m})} + \boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}_i^{(\mathsf{m})}), \end{aligned}$$

where $g_y(a) = yg(a) + (1-y)(1-g(a))$. This directly leads to the SIM condition:

$$\mathbb{E}(S_i^{(\mathsf{m})} \mid \mathbf{X}_i^{(\mathsf{m})}) = \breve{f}_m(\boldsymbol{\beta}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}) = f_m(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})}),$$

for some $\breve{f}_m$ and f_m.

■

B.2 Proof of Lemma 1

Proof Let $\widetilde{\boldsymbol{\theta}}_{\mathsf{sup}} = (\widetilde{\alpha}_{\mathsf{sup}}, \widetilde{\boldsymbol{\beta}}_{\mathsf{sup}}^\mathsf{T})^\mathsf{T}$ and $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0^\mathsf{T})^\mathsf{T}$. The joint-information condition in Assumption 1, together with the sub-Gaussian design, gives restricted strong convexity for the logistic loss after accounting for the unpenalized intercept and the unpenalized first coefficient. Standard decomposable-regularizer arguments (Negahban et al., 2012) therefore give

$$\|\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}} - \boldsymbol{\beta}_0\|_2 = O_p\left(\sqrt{\frac{s\log p}{n}}\right), \qquad \|\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}} - \boldsymbol{\beta}_0\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right),$$

and $|\widetilde{\alpha}_{\mathsf{sup}} - \alpha_0| = O_p\{\sqrt{s\log p/n}\}$. Since $s\log p/(n\beta_{01}^2) = o(1)$,

$$|\widetilde{\beta}_{\mathsf{sup},1} - \beta_{01}| \le \|\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}} - \boldsymbol{\beta}_0\|_2 = o_p(|\beta_{01}|),$$

so $|\widetilde{\beta}_{\mathsf{sup},1}| \ge |\beta_{01}|/2$ with probability approaching one. On this event,

$$\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} - \boldsymbol{\gamma}_0 = \frac{\widetilde{\boldsymbol{\beta}}_{\mathsf{sup}} - \boldsymbol{\beta}_0}{\widetilde{\beta}_{\mathsf{sup},1}} + \boldsymbol{\beta}_0\left(\frac{1}{\widetilde{\beta}_{\mathsf{sup},1}} - \frac{1}{\beta_{01}}\right).$$

Using $\|\boldsymbol{\beta}_0\|_2 \le \kappa$, $\|\boldsymbol{\beta}_0\|_1 \le \sqrt{s}\|\boldsymbol{\beta}_0\|_2$, and the two rates above yields

$$\|\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} - \boldsymbol{\gamma}_0\|_2 \lesssim_p \sqrt{\frac{s\log p}{n\beta_{01}^2}}, \qquad \|\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} - \boldsymbol{\gamma}_0\|_1 \lesssim_p s\sqrt{\frac{\log p}{n\beta_{01}^2}}.$$

The first coordinate equals zero after normalization. Choosing C_Γ sufficiently large gives $\widetilde{\boldsymbol{\gamma}}_{\mathsf{sup}} \in \Gamma$ with probability approaching one. ■

B.3 Auxiliary Lemmas Used in Section A

Lemma B.1 *Under the same conditions as Theorem 1,*

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\left\|\frac{1}{N_m}\sum_{i=1}^{N_m} f_m'(\boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})})\left\{\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X} \mid \boldsymbol{\gamma}_0^\mathsf{T}\mathbf{X}_i^{(\mathsf{m})})\right\}\epsilon_i^{(\mathsf{m})}\right\|_\infty \le C\sqrt{\frac{\log p}{N_m}}$$

with probability exceeding $1 - \exp(-c_1\log p)$.

Proof Conditional on $\mathbf{X}_i^{(\mathsf{m})}$, the summand has mean zero because $\mathbb{E}(\epsilon_i^{(\mathsf{m})} \mid \mathbf{X}_i^{(\mathsf{m})}) = 0$. The bounded derivative f_m', together with Assumption 4, gives a uniformly sub-exponential coordinate envelope. Bernstein's inequality and a union bound over the p coordinates yield the stated rate. ■

Lemma B.2 *Under the same conditions as Theorem 1, we have*

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\left\|\frac{1}{N_m}\sum_{i=1}^{N_m}\left\{\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}) - f_m'(\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)\big)\right\}\epsilon_i^{(\mathsf{m})}\right\|_\infty \le C\sqrt{\frac{\log p}{N_m}}$$

with probability exceeding $1-\exp(-c_1\log p)$*. The uniformity over* Γ *follows from the kernel-class entropy condition in Assumption 2(c).*

Proof *Define*

$$\widehat{G}^{(1)}(\boldsymbol{\gamma}^\intercal\mathbf{X}\mid\boldsymbol{\gamma}) := \frac{\sum_{i=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_i^{(\mathsf{m})}-\mathbf{X}\})S_i^{(\mathsf{m})}}{\sum_{i=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_i^{(\mathsf{m})}-\mathbf{X}\})} - \frac{\sum_{i=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_i^{(\mathsf{m})}-\mathbf{X}\})S_i^{(\mathsf{m})}\sum_{i=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_i^{(\mathsf{m})}-\mathbf{X}\})}{\left[\sum_{i=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_i^{(\mathsf{m})}-\mathbf{X}\})\right]^2}.$$

First, we have

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\left\|\frac{1}{N_m}\sum_{i=1}^{N_m}\left\{\left(\widehat{G}^{(1)}(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\mid\boldsymbol{\gamma}) - f_m'(\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\right)\left(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})})\right)\right\}\epsilon_i^{(\mathsf{m})}\right\|_\infty \lesssim \frac{[h_m^2\log(p\vee N_m)]^{1/4}}{\sqrt{N_m}} \quad \text{(B.23)}$$

directly from Lemma A8 in Wu et al. (2023) under the assumption that $\left(\frac{s\log p}{N_m}\right)^{1/5} \lesssim h_m \lesssim N_m^{-1/6}$*. Note that from Page 114 in Wu et al. (2023), even though we have* $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2 \lesssim \sqrt{\frac{s\log p}{n\beta_{01}^2}}$ *instead, we still have the same bound for* (B.23) *since* $\sqrt{\frac{s\log p}{n}}/2^{-l}h_m \lesssim \sqrt{\frac{s\log p}{n}}\sqrt{N_m}h_m \lesssim N_m$ *still holds.*

Then it suffices to bound

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\left\|\frac{1}{N_m}\sum_{i=1}^{N_m}\left\{\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}) - \widehat{G}^{(1)}(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\mid\boldsymbol{\gamma})\big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)\big)\right\}\epsilon_i^{(\mathsf{m})}\right\|_\infty.$$

We have

$$\begin{aligned}
&\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}) - \widehat{G}^{(1)}(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\mid\boldsymbol{\gamma})(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})}))\\
=&\frac{\sum_{j=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}(\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})})}{\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})} - \frac{\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}\sum_{j=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})(\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})})}{\left[\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})\right]^2}\\
&-\left(\frac{\sum_{j=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}}{\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})} - \frac{\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}\sum_{j=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})}{\left[\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})\right]^2}\right)\cdot(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})}))\\
=&\frac{\sum_{j=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}(\mathbf{X}_j^{(\mathsf{m})} - \mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})}))}{\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})}\\
&-\frac{\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}\sum_{j=1}^{N_m}K_h'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})(\mathbf{X}_j^{(\mathsf{m})} - \mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^\intercal\mathbf{X}_i^{(\mathsf{m})}))}{\left[\sum_{j=1}^{N_m}K_h(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})\right]^2}
\end{aligned}$$

depends on $\mathbf{X}_i^{(\mathsf{m})}$ *only through* $\boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})}$ *and* $\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}$. *Then, by arguments similar to those in Lemmas A11 and A8 of Wu et al. (2023), we have*

$$\max_i \sup_{\boldsymbol{\gamma}\in\Gamma} \Big\| \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}) - \widehat{G}^{(1)}(\boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})} \mid \boldsymbol{\gamma}) \big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}\big(\mathbf{X} \,\big|\, \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}\big)\big) \Big\|_\infty \le C_0 h_m,$$

$$\sup_{\boldsymbol{\gamma}\in\Gamma} \Big\| \frac{1}{N_m} \sum_{i=1}^{N_m} \Big\{ \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}) - \widehat{G}^{(1)}(\boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})} \mid \boldsymbol{\gamma}) \Big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}\big(\mathbf{X} \,\big|\, \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}\big)\Big) \Big\} \epsilon_i^{(\mathsf{m})} \Big\|_\infty \lesssim \frac{\big[h_m^2 \log(p \vee N_m)\big]^{1/4}}{\sqrt{N_m}}. \tag{B.24}$$

Combining (B.23), (B.24) *and the assumption of Theorem 1, finally we have*

$$\sup_{\boldsymbol{\gamma}\in\Gamma} \Big\| \frac{1}{N_m} \sum_{i=1}^{N_m} \Big\{ \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}) - f_m'(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) \Big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}\big(\mathbf{X} \,\big|\, \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}\big)\Big) \Big\} \epsilon_i^{(\mathsf{m})} \Big\|_\infty \lesssim \sqrt{\frac{\log p}{N_m}}$$

with probability exceeding $1 - \exp(-c_1 \log p)$.

■

Lemma B.3 *Under the same conditions as Theorem 1, we have*

$$\begin{aligned} &\Big| \frac{1}{N_m} \sum_{i=1}^{N_m} (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) \cdot \Big\{ f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big\} \Big| \\ \le & C \Big\{ \|\Delta_t^{(\mathsf{m})}\|_1 \sqrt{\frac{1}{N_m}} + \|\Delta_t^{(\mathsf{m})}\|_1 h_m^3 + \|\Delta_t^{(\mathsf{m})}\|_2 \|\Delta_{t-1}^{(\mathsf{m})}\|_2 h_m^2 \Big\}, \end{aligned}$$

where $\Delta_t^{(\mathsf{m})} = \widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0$.

Proof *By the bound for* I_{n4} *in the proof of Lemma A1 in Wu et al. (2023), and* $h_m[\log(p \vee N_m)]^{1/4} \le c$ *implied by the assumption of Theorem 1, we immediately obtain*

$$\Big\| \frac{1}{N_m} \sum_{i=1}^{N_m} f'(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) \Big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}[\mathbf{X} | \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}]\Big) \cdot \Big\{ f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big\} \Big\|_\infty \lesssim \frac{h_m[\log(p \vee N_m)]^{1/4}}{\sqrt{N_m}} \lesssim \sqrt{\frac{1}{N_m}}$$

with probability at least $1 - \exp(-c_1 \log p)$ *for some constant* $c_1 > 0$. *Then it suffices to bound*

$$\begin{aligned} &\Big| \frac{1}{N_m} \sum_{i=1}^{N_m} (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal \Big\{ \partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}) - f_m'(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) \Big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X} \,|\, \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})})\Big) \Big\} \cdot \Big\{ f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big\} \Big| \\ \le & \max_{i:\,>0} \sup_{\boldsymbol{\gamma}\in\Gamma} \Big| (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal \Big(\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}) - f_m'(\boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})}) (\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X} \,|\, \boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})})) \Big) \Big| \cdot \max_{i:\,>0} \Big| f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big| \\ &+ \Big| \frac{1}{N_m} \sum_{i=1}^{N_m} (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal} \mathbf{X}_i^{(\mathsf{m})}) \Big(\mathbb{E}(\mathbf{X} \,|\, \widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal} \mathbf{X}_i^{(\mathsf{m})}) - \mathbb{E}(\mathbf{X} \,|\, \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) \Big) \cdot \Big(f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big) \Big| \\ &+ \Big| \frac{1}{N_m} \sum_{i=1}^{N_m} (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal \Big(f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal} \mathbf{X}_i^{(\mathsf{m})}) - f_m'(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) \Big) \Big(\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X} \,|\, \boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})})\Big) \cdot \Big(f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big) \Big| \\ :=& I_1 + I_2 + I_3. \end{aligned}$$

We have

$$I_1 = \max_{i:\,>0} \sup_{\boldsymbol{\gamma}\in\Gamma} \Big| (\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})} - \boldsymbol{\gamma}_0)^\intercal \Big(\partial_{\boldsymbol{\gamma}} \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}) - f_m'(\boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})}) (\mathbf{X}_i^{(\mathsf{m})} - \mathbb{E}(\mathbf{X} \,|\, \boldsymbol{\gamma}^\intercal \mathbf{X}_i^{(\mathsf{m})})) \Big) \Big| \cdot \max_{i:\,>0} \Big| f_m(\boldsymbol{\gamma}_0^\intercal \mathbf{X}_i^{(\mathsf{m})}) - \widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})}; \boldsymbol{\gamma}_0) \Big| \lesssim \|\Delta_t^{(\mathsf{m})}\|_1 h_m^3,$$

holds by Lemma B.10 and Lemma B.9.

$$
\begin{aligned}
I_2 &= \Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal} f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbb{E}(\mathbf{X}\,|\,\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big)\cdot\Big(f_m(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big)\Big| \\
&= \Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal} f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(f_m(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big)\cdot\Big(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})}-\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})}\Big) \\
&\quad\cdot\int_0^1\mathbb{E}^{(1)}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})}+t(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\mathrm{d}t\Big| \\
&= \Big|\frac{1}{N_m}\sum_{i=1}^{N_m} f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(f_m(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big)(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}_i^{(\mathsf{m})} \\
&\quad\cdot\int_0^1\mathbb{E}^{(1)}((\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})}+t(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\mathrm{d}t\Big| \\
&\lesssim \|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2\|\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2 h_m^2
\end{aligned}
$$

where the last inequality follows by Assumption 2 that $f_m'(\boldsymbol{\gamma}^{\intercal}\mathbf{X}_i^{(\mathsf{m})})$ is bounded uniformly, Assumption 6 that

$$\max_{i:>0}\sup_{\boldsymbol{\gamma}\in\Gamma}\big|\mathbb{E}^{(1)}(\boldsymbol{v}^{\intercal}\mathbf{X}\mid\boldsymbol{\gamma}^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\big|\le C\|\boldsymbol{v}\|_2,$$

Lemma B.9 that $\max_{i:>0}\sup_{\boldsymbol{\gamma}\in\Gamma}\Big|f_m(\boldsymbol{\gamma}^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})\Big|\le c_0h_m^2$, and Lemma A3 of Wu et al. (2023) that $\Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\boldsymbol{v}^{\intercal}\mathbf{X}_i^{(\mathsf{m})})^2\Big|\lesssim\|\boldsymbol{v}\|_2^2$. Similarly, we have

$$
\begin{aligned}
I_3 &= \Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\Big(f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})-f_m'(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big)\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big)\Big(f_m(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big)\Big| \\
&\le \Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}_i^{(\mathsf{m})}\int_0^1 f''(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})}+t(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\mathrm{d}t \\
&\quad\cdot(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big)\Big(f_m(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big)\Big| \\
&\le \|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2\|\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2 h_m^2.
\end{aligned}
$$

Thus, we achieve

$$
\begin{aligned}
&\Big|\frac{1}{N_m}\sum_{i=1}^{N_m}(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\cdot\Big\{f_m(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}_0)\Big\}\Big| \\
\le& C\Big\{\|\Delta_t^{(\mathsf{m})}\|_1\sqrt{\frac{1}{N_m}}+\|\Delta_t^{(\mathsf{m})}\|_1h_m^3+\|\Delta_t^{(\mathsf{m})}\|_2\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m^2\Big\}
\end{aligned}
$$

for some constant C, where $\Delta_{t-1}^{(\mathsf{m})}:=\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0$.

■

Lemma B.4 *Under the same conditions as Theorem 1, we have*

$$
\begin{aligned}
&\Big|(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\frac{1}{N_m}\sum_{i=1}^{N_m}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\Big\{(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^{\intercal}\int_0^1\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}))-\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}\Big| \\
\le& C\Big\{\|\Delta_t^{(\mathsf{m})}\|_2\|\Delta_{t-1}^{(\mathsf{m})}\|_2h_m+\|\Delta_t^{(\mathsf{m})}\|_2\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2\Big\}.
\end{aligned}
$$

Proof *First, we aim to bound*

$$\left|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big((\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\Big)^2\right|.$$

We have

$$\left|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big((\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}f_m'(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\Big)\Big)^2\right|\le C'\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2$$

by Lemma B2 of Wu et al. (2023),

$$\begin{aligned}&\Big|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big((\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\Big\{f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\Big)-f_m'(\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\Big)\Big\}\Big)^2\Big|\\ \le&C'\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2\|\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2\end{aligned}$$

holds by similar argument in the proof of Lemma B.3, and

$$\left|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big((\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\Big\{\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})-f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})\big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\big)\Big\}\Big)^2\right|\le C'\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2h_m^2$$

holds by Lemma B.10. By combining the above inequalities, we have

$$\begin{aligned}&\left|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big((\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\Big)^2\right|\\ \le&3C'\Big\{\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2+\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2\|\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2+\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2h_m^2\Big\}\le C\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2.\end{aligned}\tag{B.25}$$

Then, we bound

$$\begin{aligned}&\left|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^{\intercal}\int_0^1\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})-\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}^2\right|\\ \lesssim&\max_i\sup_{\boldsymbol{\gamma}\in\Gamma}\Big|(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^{\intercal}\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})-f_m'(\boldsymbol{\gamma}^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}^{\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\big)\Big)\Big|^2\\ &+\left|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})^{\intercal}\int_0^1\Big(f_m'(\boldsymbol{\gamma}_\tau^{\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_\tau^{\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\Big)-f_m'(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})})\Big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}\big(\mathbf{X}\,|\,\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})}\big)\Big)\Big)\mathrm{d}\tau\Big\}^2\right|\\ \lesssim&\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}\|_2^2h_m^2+\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}\|_2^4,\end{aligned}\tag{B.26}$$

by an argument similar to that used in the proof of Lemma B.3. Here $\boldsymbol{\gamma}_\tau:=\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})$.

In conclusion, by combining (B.25) *and* (B.26)*, we have*

$$\begin{aligned}&\left|(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0)^{\intercal}\frac{1}{N_m}\sum_{i=1}^{N_m}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{t-1}^{(\mathsf{m})})\cdot\Big\{(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{t-1}^{(\mathsf{m})})^{\intercal}\int_0^1\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}+\tau(\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})-\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\widehat{\boldsymbol{\gamma}}_{t-1}^{(\mathsf{m})})\Big)\mathrm{d}\tau\Big\}\right|\\ \le&C\sqrt{\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2^2}\cdot\sqrt{\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}\|_2^2h_m^2+\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}\|_2^4}\\ \le&C\Big\{\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}\|_2h_m+\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}-\boldsymbol{\gamma}_0\|_2\|\boldsymbol{\gamma}_0-\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})}\|_2^2\Big\}\end{aligned}$$

for some constant C.

■

Lemma B.5 *Under the same conditions as Theorem 1 and assuming* $\sqrt{\log p/N_m}+\|\Delta_{t-1}^{(\mathsf{m})}\|_2 h_m+\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2 < \lambda_t^{(\mathsf{m})}/2$, *we have* $\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1 \le 3\|\Delta_{t,S}^{(\mathsf{m})}\|_1$, *where* $S=\{j:\boldsymbol{\gamma}_{0,j}\neq 0\}$.

Proof *By the assumption that*

$$\sqrt{\frac{\log p}{N_m}}+\|\Delta_{t-1}^{(\mathsf{m})}\|_2 h_m+\|\Delta_{t-1}^{(\mathsf{m})}\|_2^2 < \frac{\lambda_t^{(\mathsf{m})}}{2}$$

and (A.3), *we have*

$$\begin{aligned}
0 \le & \frac{\lambda_t^{(\mathsf{m})}}{2}\|\Delta_t^{(\mathsf{m})}\|_1+\lambda_t^{(\mathsf{m})}\|\boldsymbol{\gamma}_0\|_1-\lambda_t^{(\mathsf{m})}\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}\|_1 \\
= & \lambda_t^{(\mathsf{m})}\ (\frac{1}{2}\|\Delta_{t,S}^{(\mathsf{m})}\|_1+\frac{1}{2}\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1+\|\boldsymbol{\gamma}_{0,S}\|_1-\|\widehat{\boldsymbol{\gamma}}_{t,S}^{(\mathsf{m})}\|_1-\|\widehat{\boldsymbol{\gamma}}_{t,S^c}^{(\mathsf{m})}\|_1) \\
\le & \lambda_t^{(\mathsf{m})}\ (\frac{1}{2}\|\Delta_{t,S}^{(\mathsf{m})}\|_1+\frac{1}{2}\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1+\|\Delta_{t,S}^{(\mathsf{m})}\|_1-\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1) \\
\le & \lambda_t^{(\mathsf{m})}\ (\frac{3}{2}\|\Delta_{t,S}^{(\mathsf{m})}\|_1-\frac{1}{2}\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1),
\end{aligned} \tag{B.27}$$

where the equality holds by $\|\boldsymbol{\gamma}_0\|_1=\|\boldsymbol{\gamma}_{0,S}\|_1+\|\boldsymbol{\gamma}_{0,S^c}\|_1=\|\boldsymbol{\gamma}_{0,S}\|_1$ *and* $\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}\|_1=\|\widehat{\boldsymbol{\gamma}}_{t,S}^{(\mathsf{m})}\|_1+\|\widehat{\boldsymbol{\gamma}}_{t,S^c}^{(\mathsf{m})}\|_1$. *The second inequality holds by* $\|\boldsymbol{\gamma}_{0,S}\|_1-\|\widehat{\boldsymbol{\gamma}}_{t,S}^{(\mathsf{m})}\|_1 \le \|\Delta_{t,S}^{(\mathsf{m})}\|_1$ *and* $\|\widehat{\boldsymbol{\gamma}}_{t,S^c}^{(\mathsf{m})}\|_1=\|\boldsymbol{\gamma}_{0,S^c}-\widehat{\boldsymbol{\gamma}}_{t,S^c}^{(\mathsf{m})}\|_1=\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1$. *Then we immediately have* $0\le 3\|\Delta_{t,S}^{(\mathsf{m})}\|_1-\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1$. ■

Lemma B.6 *Under the same conditions in Lemma B.5, we have*

$$\|\Delta_t^{(\mathsf{m})}\|_1 \le 4\sqrt{s}\|\Delta_t^{(\mathsf{m})}\|_2,$$

where $s=|S|$ *and* $S=\{j:\boldsymbol{\gamma}_{0,j}\neq 0\}$.

Proof From Lemma B.5, we have $\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1 \le 3\|\Delta_{t,S}^{(\mathsf{m})}\|_1$, so

$$\|\Delta_t^{(\mathsf{m})}\|_1=\|\Delta_{t,S}^{(\mathsf{m})}\|_1+\|\Delta_{t,S^c}^{(\mathsf{m})}\|_1 \le 4\|\Delta_{t,S}^{(\mathsf{m})}\|_1 \le 4\sqrt{s}\|\Delta_{t,S}^{(\mathsf{m})}\|_2 \le 4\sqrt{s}\|\Delta_t^{(\mathsf{m})}\|_2,$$

where $\|\Delta_{t,S}^{(\mathsf{m})}\|_1 \le \sqrt{s}\|\Delta_{t,S}^{(\mathsf{m})}\|_2$ because vector $\Delta_{t,S}^{(\mathsf{m})}$ has at most s nonzero entries.

■

Lemma B.7 *Under the same conditions in Lemma B.5, we have*

$$\|\boldsymbol{\gamma}_0\|_1-\|\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}\|_1 \le 4\sqrt{s}\|\Delta_t^{(\mathsf{m})}\|_2,$$

where $s=|S|$ *and* $S=\{j:\boldsymbol{\gamma}_{0,j}\neq 0\}$.

Proof We have

$$\|\boldsymbol{\gamma}_0\|_1 - \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]}\|_1 = \|\boldsymbol{\gamma}_{0,S}\|_1 - \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{t,S}\|_1 - \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{t,S^c}\|_1 \le \|\Delta^{(\mathsf{m})}_{t,S}\|_1 - \|\Delta^{(\mathsf{m})}_{t,S^c}\|_1 \le \|\Delta^{(\mathsf{m})}_t\|_1 \le 4\sqrt{s}\|\Delta^{(\mathsf{m})}_t\|_2$$

following the same argument in (B.27). ■

Lemma B.8 (RE condition) *Under the conditions of Theorem 1, there exists a positive constant κ such that*

$$\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]})\Big\}^2 \ge \kappa\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0\|_2^2.$$

Proof *Since $a^2 \ge \frac{1}{2}b^2 - (a-b)^2$, we have*

$$\begin{aligned}
&\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]})\Big\}^2 \\
\ge&\frac{1}{2N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}} f'_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i)\Big(\mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i\big)\Big)\Big\}^2 \\
&-\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]}) - f'_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i)\Big(\mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i\big)\Big)\Big)\Big\}^2 \\
:=&\Pi_1 - \Pi_2.
\end{aligned} \tag{B.28}$$

Write $T_i = \boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i$ and $R_i = \mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}(\mathbf{X}^{(\mathsf{m})}_i \mid T_i)$. The leading term is

$$\Delta_t^{(\mathsf{m})\mathsf{T}}\left\{\frac{1}{N_m}\sum_{i=1}^{N_m} f'_m(T_i)^2 R_i R_i^{\mathsf{T}}\right\}\Delta_t^{(\mathsf{m})}.$$

By Lemma B.6, $\Delta_t^{(\mathsf{m})}/\|\Delta_t^{(\mathsf{m})}\|_2$ lies in the normalized tangent cone and has first coordinate zero. Assumption 5 and restricted covariance concentration give the required lower bound with probability at least $1 - \exp(-c_1 s\log p)$.

For Π_2, we have

$$\begin{aligned}
|\Pi_2| =& \Big|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]}) - f'_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i)\Big(\mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i\big)\Big)\Big)\Big\}^2\Big| \\
\le& 2\Big|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\Big(\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]}) - f'_m(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]}{}^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i)\Big(\mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}\big(\mathbf{X}\,|\,\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})\mathsf{T}}_{[t-1]}\mathbf{X}^{(\mathsf{m})}_i\big)\Big)\Big)\Big\}^2\Big| \\
&+2\Big|\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\Big(f'_m(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})\mathsf{T}}_{[t-1]}\mathbf{X}^{(\mathsf{m})}_i)\Big(\mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}\big(\mathbf{X}\,|\,\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})\mathsf{T}}_{[t-1]}\mathbf{X}^{(\mathsf{m})}_i\big)\Big) - f'_m(\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i)\Big(\mathbf{X}^{(\mathsf{m})}_i - \mathbb{E}\big(\mathbf{X}\,|\,\boldsymbol{\gamma}_0^{\mathsf{T}}\mathbf{X}^{(\mathsf{m})}_i\big)\Big)\Big)\Big\}^2\Big| \\
\lesssim& \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0\|_2^2 h_m^2 + \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0\|_2^2\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]} - \boldsymbol{\gamma}_0\|_2^2 \\
=& o(\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0\|_2^2)
\end{aligned} \tag{B.29}$$

by an argument similar to that used in the proof of Lemma B.3.

Combining (B.28) *and* (B.29)*, we have*

$$\frac{1}{N_m}\sum_{i=1}^{N_m}\Big\{(\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0)^{\mathsf{T}}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}^{(\mathsf{m})}_i; \widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t-1]})\Big\}^2 \ge \kappa\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{m})}_{[t]} - \boldsymbol{\gamma}_0\|_2^2$$

with probability at least $1-\exp(-c_1 s\log p)$ *for some positive constant* c_1.

■

B.4 Auxiliary Lemmas in Section B.3

For brevity, the proofs of the lemmas in this section are available in the extended arXiv version (`https://arxiv.org/abs/2302.04970`).

Lemma B.9 *Suppose Assumptions 2 - 4 are satisfied and* $d_0 s\log(p\vee N_m)\le N_m h_m^5$ *for some constant* d_0. *There exists some universal constant* c_0, c_1 *such that*

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\max_{1\le i\le N_m:\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\in\mathcal{I}_m^\delta}\left|f_m(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})-\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})\right|\le c_0h_m^2$$

with probability exceeding $1-\exp(-c_1\log(p\vee N_m))$. *Uniformity over* Γ *follows from Assumption 2(c). The maximum is restricted to retained target points. Since* $\mathrm{supp}(K)\subset[-1,1]$ *and* $h_m<\delta$, *all kernel arguments used for such a target lie in* $\mathcal{I}_m^{2\delta}$, *where the density and its derivatives are uniformly regular. Hence this is an interior expansion and no boundary-bias term arises.*

Lemma B.10 *Suppose Assumptions 2 - 4 are satisfied and* $d_0 s\log(p\vee N_m)\le N_m {h_m}^5$ *for some constant* d_0. *There exists some universal constant* c_0, c_1 *such that*

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\max_{1\le i\le N_m:\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\in\mathcal{I}_m^\delta}\left\|\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})-f_m'(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})\big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})\big)\right\|_\infty\le c_0h_m,$$

$$\text{and}\sup_{\boldsymbol{\gamma}\in\Gamma,\,\boldsymbol{v}\in\mathbb{V}_c}\max_{1\le i\le N_m:\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\in\mathcal{I}_m^\delta}\left|\left\{\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})-f_m'(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})\big(\mathbf{X}_i^{(\mathsf{m})}-\mathbb{E}(\mathbf{X}\,|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})\big)\right\}^\intercal\boldsymbol{v}\right|\le c_0h_m$$

with probability exceeding $1-\exp(-c_1\log(p\vee N_m))$. *Here* $\mathbb{E}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big):=\mathbb{E}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}=\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)$ *for simplicity of notation, Here* Γ *is the local* ℓ_1*–*ℓ_2 *neighborhood in Assumption 2, and* $\mathbb{V}_c=\{\boldsymbol{v}:v_1=0,\ \|\boldsymbol{v}\|_2\le1,\ \|\boldsymbol{v}_{S^c}\|_1\le3\|\boldsymbol{v}_S\|_1\}$. *Uniformity in* $\boldsymbol{\gamma}$ *is supplied by Assumption 2(c); the standard sparse-block argument is used only for the linear directional supremum in* $\boldsymbol{v}$. *The target-point restriction, the projected-index localization in Assumption 2(b), and* $h_m<\delta$ *again ensure an interior kernel expansion throughout the Taylor path.*

Lemma B.11 *Define that* $D_{i52}=\frac{1}{N_m-1}\sum_{j=1,j\neq i}^{N_m}K_{h_m}'(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})\mathbf{X}_j^{(\mathsf{m})}$. *Under the assumptions of Lemma B.10, we have*

$$\sup_{\boldsymbol{\gamma}\in\Gamma,\,\boldsymbol{v}\in\mathbb{V}_c}\max_{i:\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\in\mathcal{I}_m^\delta}\left|\left\{D_{i52}+\mathbb{E}^{(1)}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)g_{m,\boldsymbol{\gamma}}(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})+\mathbb{E}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)g_{\boldsymbol{\gamma}}'(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})\right\}^\intercal\boldsymbol{v}\right|\le c_0h_m,\tag{B.30}$$

$$\sup_{\boldsymbol{\gamma}\in\Gamma}\max_{i:\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\in\mathcal{I}_m^\delta}\left\|D_{i52}+\mathbb{E}^{(1)}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)g_{m,\boldsymbol{\gamma}}(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})+\mathbb{E}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big)g_{\boldsymbol{\gamma}}'(\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})})\right\|_\infty\le c_0h_m\tag{B.31}$$

with probability exceeding $1-\exp(-c_1\log(p\vee N_m))$. *Here*

$$\mathbb{E}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big):=\mathbb{E}\big(\mathbf{X}\,\big|\,\boldsymbol{\gamma}^\intercal\mathbf{X}=\boldsymbol{\gamma}^\intercal\mathbf{X}_i^{(\mathsf{m})}\big),$$

The parameter-indexed uniformity follows from Assumption 2(c), and the maxima are taken only over retained target points. The set $\mathbb{V}_c$ *is the directional cone defined above.*

Appendix C. Justification of Assumption 5

We verify Assumption 5 under a Gaussian design. Suppose that $\mathbf{X}\sim N(0,I_p)$. Then

$$\mathrm{Cov}(\mathbf{X}\mid\boldsymbol{\gamma}_0^\intercal\mathbf{X})=I_p-\frac{\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^\intercal}{\|\boldsymbol{\gamma}_0\|_2^2}.$$

For every $\boldsymbol{v}\in\mathbb{V}_0$, because $v_1=0$, $\|\boldsymbol{v}\|_2=1$, and $\gamma_{01}=1$, define

$$a_m^2=\mathbb{E}\left[\{f_m'(\boldsymbol{\gamma}_0^\intercal\mathbf{X})\}^2\right].$$

Assumption 2(a) requires $a_m^2>0$. Since the Gaussian conditional covariance is constant in the index,

$$\begin{aligned}&\boldsymbol{v}^\intercal\mathbb{E}\left[\{f_m'(\boldsymbol{\gamma}_0^\intercal\mathbf{X})\}^2\,\mathrm{Cov}(\mathbf{X}\mid\boldsymbol{\gamma}_0^\intercal\mathbf{X})\right]\boldsymbol{v}\\&\quad=a_m^2\left\{1-\frac{(\boldsymbol{v}^\intercal\boldsymbol{\gamma}_0)^2}{\|\boldsymbol{\gamma}_0\|_2^2}\right\}\geq\frac{a_m^2}{\|\boldsymbol{\gamma}_0\|_2^2}\geq\frac{a_m^2}{\kappa_\gamma^2}.\end{aligned}\tag{C.32}$$

Thus the derivative-weighted restricted lower bound holds on the normalized tangent space with $c=\min_m a_m^2\kappa_\gamma^{-2}>0$. The corresponding upper bound is at most $\max_m a_m^2$.

Furthermore, we evaluate the squared maximum eigenvalues:

$$\begin{aligned}&\sup_{\boldsymbol{\gamma}\in\Gamma}\frac{1}{N_m}\sum_{i=1}^{N_m}\Big[\lambda_{\max}\big(\mathbb{E}(\mathbf{X}\mathbf{X}^\intercal\mid\mathbf{X}_i^\intercal\boldsymbol{\gamma})\big)\Big]^2=\sup_{\boldsymbol{\gamma}\in\Gamma}\frac{1}{N_m}\sum_{i=1}^{N_m}\sup_{\|\boldsymbol{v}\|_2=1}\Big[\mathbb{E}((\mathbf{X}^\intercal\boldsymbol{v})^2\mid\mathbf{X}_i^\intercal\boldsymbol{\gamma})\Big]^2\\&=\sup_{\boldsymbol{\gamma}\in\Gamma}\frac{1}{N_m}\sum_{i=1}^{N_m}\sup_{\|\boldsymbol{v}\|_2=1}\Big[\frac{(\mathbf{X}_i^\intercal\boldsymbol{\gamma})^2(\boldsymbol{v}^\intercal\boldsymbol{\gamma})^2}{\|\boldsymbol{\gamma}\|_2^4}+\|\boldsymbol{v}\|_2^2-\frac{(\boldsymbol{v}^\intercal\boldsymbol{\gamma})^2}{\|\boldsymbol{\gamma}\|_2^2}\Big]^2\leq\sup_{\boldsymbol{\gamma}\in\Gamma}\frac{1}{N_m}\sum_{i=1}^{N_m}\sup_{\|\boldsymbol{v}\|_2=1}\Big[\frac{2(\mathbf{X}_i^\intercal\boldsymbol{\gamma})^4(\boldsymbol{v}^\intercal\boldsymbol{\gamma})^4}{\|\boldsymbol{\gamma}\|_2^8}+2\Big]\\&\leq\sup_{\boldsymbol{\gamma}\in\Gamma}\frac{1}{N_m}\sum_{i=1}^{N_m}\Big[\frac{2(\mathbf{X}_i^\intercal\boldsymbol{\gamma})^4}{\|\boldsymbol{\gamma}\|_2^4}+2\Big]\leq\sup_{\boldsymbol{\gamma}\in\Gamma}\Big[\frac{3\mathbb{E}(\mathbf{X}_i^\intercal\boldsymbol{\gamma})^4}{\|\boldsymbol{\gamma}\|_2^4}+2\Big]=11.\end{aligned}\tag{C.33}$$

The last inequality follows from Lemma B3 in Wu et al. (2023).

Lastly, we have

$$\begin{aligned}&\max_{1\leq i\leq N_m}\sup_{\boldsymbol{\gamma}\in\Gamma}\lambda_{\max}\Big(\mathbb{E}(\mathbf{X}\mathbf{X}^\intercal\mid\mathbf{X}_i^\intercal\boldsymbol{\gamma})\Big)=\max_{1\leq i\leq N_m}\sup_{\boldsymbol{\gamma}\in\Gamma}\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}((\mathbf{X}^\intercal\boldsymbol{v})^2\mid\mathbf{X}_i^\intercal\boldsymbol{\gamma})\\&=\max_{1\leq i\leq N_m}\sup_{\boldsymbol{\gamma}\in\Gamma}\sup_{\|\boldsymbol{v}\|_2=1}\Big[\frac{(\mathbf{X}_i^\intercal\boldsymbol{\gamma})^2(\boldsymbol{v}^\intercal\boldsymbol{\gamma})^2}{\|\boldsymbol{\gamma}\|_2^4}+\|\boldsymbol{v}\|_2^2-\frac{(\boldsymbol{v}^\intercal\boldsymbol{\gamma})^2}{\|\boldsymbol{\gamma}\|_2^2}\Big]\leq\max_{1\leq i\leq N_m}\sup_{\boldsymbol{\gamma}\in\Gamma}\Big[\frac{(\mathbf{X}_i^\intercal\boldsymbol{\gamma})^2}{\|\boldsymbol{\gamma}\|_2^2}+1\Big]\leq C\log(p\vee N_m),\end{aligned}\tag{C.34}$$

with high probability.

Appendix D. Additional implementation details and results

D.1 Additional details of the method

First, we suggest and comment on how to specify the weighting function ϕ_m for different types of S, including binary, count, and continuous surrogates. In general, our strategy is to estimate $\mathrm{Var}(S\mid\mathbf{X})$ and use inverse-variance weighting.

(a) Take $\widehat{\phi}_m(\mathbf{X};\boldsymbol{\gamma}) = \widehat{f}_m^{-1}(\mathbf{X};\boldsymbol{\gamma})$ for Poisson (count) S satisfying $\mathrm{Var}(S \mid \mathbf{X}) = \mathbb{E}(S \mid \mathbf{X})$;

(b) Take $\widehat{\phi}_m(\mathbf{X};\boldsymbol{\gamma}) = \left[\widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})\{1-\widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})\}\right]^{-1}$ for Bernoulli (binary) S satisfying $\mathrm{Var}(S \mid \mathbf{X}) = \mathbb{E}(S \mid \mathbf{X})[1-\mathbb{E}(S \mid \mathbf{X})]$;

(c) Take $\widehat{\phi}_m(\mathbf{X};\boldsymbol{\gamma}) = 1$ for Gaussian (continuous) S satisfying that $\mathrm{Var}(S \mid \mathbf{X})$ is some constant. For S believed to have heteroscedastic $\mathrm{Var}(S \mid \mathbf{X})$, we could use kernel smoothing on $\{S - \widehat{f}_m(\mathbf{X};\boldsymbol{\gamma})\}^2$ against $\mathbf{X}^\intercal\boldsymbol{\gamma}$ to estimate $\mathrm{Var}(S \mid \mathbf{X})$, and take $\widehat{\phi}_m$ as the inverse of this estimator.

Second, we present the specific form of $\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma})$ used to construct the main Step II introduced in Section 2.2

$$\begin{aligned}\partial_{\boldsymbol{\gamma}}\widehat{f}_m(\mathbf{X}_i^{(\mathsf{m})};\boldsymbol{\gamma}) =& \frac{\sum_{j=1,j\neq i}^{N_m} K'_{h_m}(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}(\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})})}{\sum_{j=1,j\neq i}^{N_m} K_{h_m}(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})} \\ & - \frac{\sum_{j=1,j\neq i}^{N_m} K_{h_m}(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})S_j^{(\mathsf{m})}\sum_{j=1,j\neq i}^{N_m} K'_{h_m}(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})(\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})})}{\left[\sum_{j=1,j\neq i}^{N_m} K_{h_m}(\boldsymbol{\gamma}^\intercal\{\mathbf{X}_j^{(\mathsf{m})}-\mathbf{X}_i^{(\mathsf{m})}\})\right]^2}.\end{aligned} \tag{D.35}$$

D.2 Tuning strategies

- For SL, the regularization parameter λ_{sup} is tuned on the labeled data set $\mathscr{L}^{(1)}$ using 10-fold cross-validation (CV), specifically using the default `cv.glmnet` function in R-package `glmnet`. The candidate range for λ_{sup} is set as the default range of `cv.glmnet`.

- For SASH and IPD, we select $\lambda_t^{(\mathsf{m})}$ at each iteration t by minimizing the BIC:

$$\mathrm{BIC}_t^{(\mathsf{m})}(\lambda) = \frac{Q^{(\mathsf{m})}(\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}(\lambda);\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})}{\{\widehat{\sigma}_{[t-1]}^{(\mathsf{m})}\}^2} + \mathrm{df}\{\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}(\lambda)\}\log(N_m),$$

  with respect to λ, where $\{\widehat{\sigma}_{[t-1]}^{(\mathsf{m})}\}^2 = L^{(\mathsf{m})}(\widehat{\boldsymbol{\gamma}}_{[t-1]}^{(\mathsf{m})})$ and $\widehat{\boldsymbol{\gamma}}_{[t]}^{(\mathsf{m})}(\lambda)$ is the estimator obtained by (9) with the penalty parameter $\lambda_t^{(\mathsf{m})}$ set as λ. The candidate range for $\lambda_t^{(\mathsf{m})}$ is $0.005 \times \{1,2,\ldots,600\} \times (\log p/N_m)^{1/2}$. As suggested in Section 2.3, we fix $h_m = \{\log(p \vee N_m)/N_m\}^{1/5}$, where the unknown $s^{1/5}$ is set directly as 1. Noting that $30^{1/5} < 2$, this choice $s^{1/5} = 1$ is reasonable. As described in Section 2.3, λ and $\lambda^\dagger$ used in Step III are also selected using the BICs. Their candidate range is set as $0.003 \times \{1,2,\ldots,2000\} \times (\log p/N)^{1/2}$. For IPD, the tuning strategies and candidate sets are the same as those of SASH, except for the BIC in its Step III, which is naturally set as the IPD loss of the SIM without the privacy constraint.

- For pLasso and PASS, the parameters are tuned in the same way as Zhang et al. (2022). For the adaptive Lasso version of L1LS implemented in these two methods, they use the BIC estimated by 10-fold CV to select the penalty coefficient, with the candidate range set $[\zeta_m/N_m, 1.5\zeta_m]$, where ζ_m is the maximum absolute correlation with $S^{(\mathsf{m})}$ among all predictors. For the penalty parameters λ_1 and λ_2 used in the prior-adaptive regression of PASS (Zhang et al., 2022), they use 10-fold CV with the

range of λ_1 set as $[\nu/n, 1.5\nu]$, where n is the labeled size and ν is the correlation between $S^{(1)}$ and $Y^{(1)}$, and the candidate range of $\kappa = \lambda_1/\lambda_2$ set as $\{1,2,\ldots,8\}$.

- For SS-uLasso, the ℓ_1-penalty parameter of uLasso is selected via CV using `cv.glmnet` in R-package `glmnet`, with candidate ranges determined by its default settings.
- For SS-RMRCE, we use the R-package `RMRCE`, which performs cross-validation to select the optimal tuning parameters for RMRCE. For the candidate sets of the tuning parameters in the function `RMRCE_cv`, their ℓ_1-penalty coefficient λ is chosen from $\{0.001, 0.01, 0.1\}$ and the smoothness parameter α is chosen from $\{1, 2, 3, 4, 5\}$. This choice is the same as their running example provided in `https://rdrr.io/github/zji90/RMRCE/`.

D.3 Simulation on SASH+

See Figure D.1 for the results.

D.4 Sensitivity analysis on tuning parameter h

In this section, we perform a sensitivity analysis on the tuning parameter h. This tuning parameter plays a crucial role in our method and its selection is essential for obtaining reliable results. We compare the results obtained using three different values of h taken $0.5h_0$, h_0, and $2h_0$, where $h_0 := \{s\log(p \vee N_m)/N_m\}^{1/5}$ is a fixed value of the tuning parameter chosen based on Theorem 1. This is a wide enough range as the largest value is four times the smallest.

We generate the data following the strong surrogate setting introduced in Section 4, and implement our method with $h = 0.5h_0, h_0$ and $2h_0$. The resulting estimation errors are presented in Figure D.2. It indicates that there is no significant difference in the overall performance of the method under these different values of h. For example, the mean estimation error with $h = 2h_0$ is 4.1% smaller than that with $h = h_0$. In comparison, the mean error of IPD is 9.2% smaller than that of SASH with $h = h_0$. Recall that the estimation error of all other methods (SS-uLasso, PASS, SS-RMRCE, pLasso, SL) is over 60% larger than the estimation error of SASH as shown in Section 4. Therefore, the performance discrepancy due to different choices of h (among a wide range) is much smaller compared to the discrepancy between our methods and other benchmark methods.

In conclusion, this sensitivity analysis shows that SASH is not sensitive to the choice of h. Although the choice of h could affect the variation of estimation errors across different repetitions, it does not change the mean (overall) performance substantially. Therefore, for the main body of this paper, we use $h = h_0$ for simplicity and to ensure consistency across our analyses.

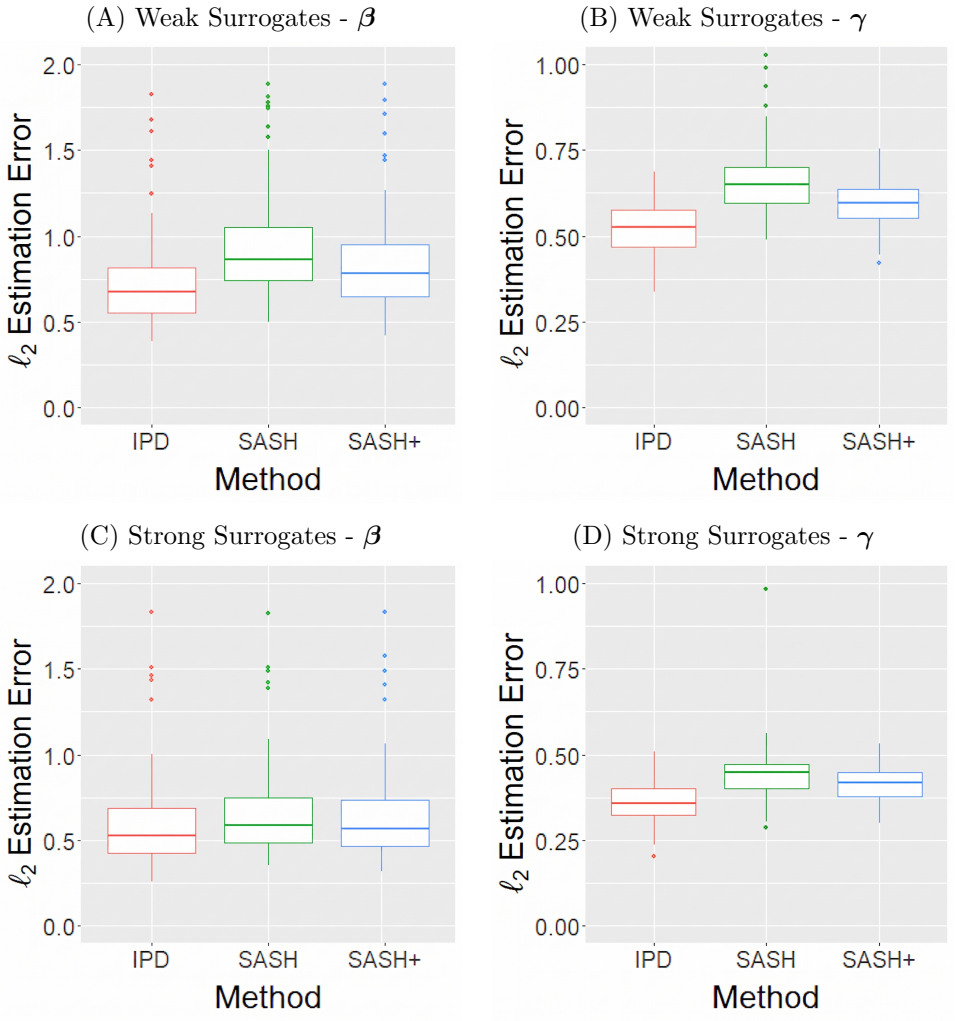

Figure D.1: Boxplots of the ℓ_2 estimation errors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ under the strong surrogate and weak surrogate settings with $N = 8000$, $M = 4$, $n = 200$, and $p = 300$. The results are based on 200 simulation replications.

D.5 Summary table for the selected features in our real-world study

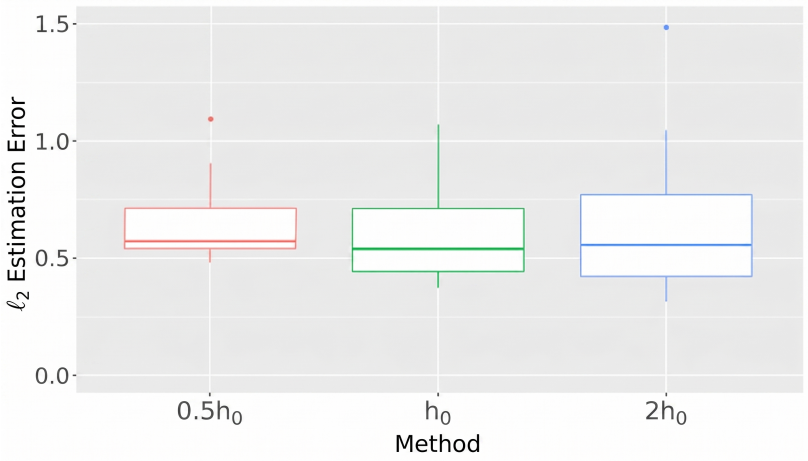

Figure D.2: Estimation error of SASH with different choices of h under the strong surrogate setting with $N = 8000$, $M = 4$, $n = 200$, and $p = 300$, where h_0 is as defined in Appendix D.4. It indicates that there is no significant difference in the overall performance of the method under these different values of h.

	SASH	SASH-CI	PASS	pLasso	SL	SS-RMRCE	SS-uLasso
Ethnicity (EUR)	-1.26	(-2.13, -0.39)	-1.72	0.00	-1.00	0.21	-1.25
Gender (Female)	-0.50	(-0.85, -0.15)	-0.99	-0.05	0.00	0.00	0.00
rs35011184_A	0.29	(0.09, 0.49)	0.67	0.00	0.00	0.00	0.50
rs10811662_A	-0.20	(-0.33, -0.06)	-2.58	-0.11	0.00	0.00	-0.30
rs1564348_C	0.11	(0.03, 0.19)	-0.52	-0.08	0.00	0.00	0.00
rs7756992_G	0.11	(0.03, 0.19)	0.34	0.00	0.00	0.00	0.00
rs11649653_G	-0.11	(-0.18, -0.03)	-0.23	0.00	0.00	0.00	-0.12
rs2902940_G	0.09	(0.03, 0.16)	0.00	-0.01	0.00	0.00	0.03
rs1077514_T	-0.09	(-0.15, -0.03)	-0.25	-0.02	0.00	0.00	0.00
rs10128711_C	-0.08	(-0.13, -0.02)	0.00	0.00	0.00	-0.19	0.00
rs3136441_C	-0.07	(-0.12, -0.02)	0.00	0.00	0.00	0.00	0.00
rs2972146_T	0.07	(0.02, 0.12)	0.29	0.00	0.00	0.00	0.00
rs11869286_C	-0.06	(-0.11, -0.02)	0.00	0.00	0.00	0.00	0.00
rs2131925_T	-0.06	(-0.10, -0.02)	0.00	0.02	0.00	0.00	0.00
rs12328675_C	-0.06	(-0.10, -0.02)	0.00	0.01	0.00	0.00	0.00
rs38855_G	-0.06	(-0.09, -0.02)	0.00	0.00	0.00	0.00	0.00
rs2328223_C	0.05	(0.02, 0.09)	0.00	0.00	0.00	0.00	0.07
rs11151789_C	-0.05	(-0.09, -0.02)	0.00	0.03	0.00	0.00	0.00
rs17299838_G	-0.05	(-0.09, -0.02)	0.00	-0.07	0.00	0.00	-0.18
rs4846914_A	-0.04	(-0.07, -0.01)	-0.25	-0.04	0.00	0.00	0.00
rs2871865_G	0.04	(0.01, 0.07)	0.20	0.00	0.00	0.00	0.22
rs10150332_C	0.04	(0.01, 0.07)	0.00	0.02	0.00	0.00	0.00
rs571312_A	0.04	(0.01, 0.07)	0.00	0.00	0.00	0.00	0.19
rs4771122_A	0.04	(0.01, 0.06)	0.00	0.00	0.00	0.00	-0.05
rs10187654_T	0.04	(0.01, 0.06)	0.07	0.00	0.00	0.00	0.26
rs622418_A	-0.03	(-0.06, -0.01)	0.00	0.00	0.00	0.00	0.00
rs2606736_T	-0.03	(-0.06, -0.01)	0.00	0.00	0.00	0.00	0.00
rs2890652_C	0.03	(0.01, 0.06)	0.00	0.00	0.00	0.00	0.00
rs4142995_T	0.03	(0.01, 0.05)	0.00	0.00	0.00	0.00	0.00
rs604109_T	0.03	(0.01, 0.05)	0.00	0.04	0.00	0.00	0.00
rs1260326_C	0.03	(0.01, 0.05)	-1.68	-0.07	0.00	0.00	0.05
rs2867125_C	0.03	(0.01, 0.05)	0.00	0.00	0.00	0.00	0.05
rs2922236_C	-0.03	(-0.05, -0.01)	0.00	-0.04	0.00	0.00	0.00
rs364585_G	0.03	(0.01, 0.05)	0.00	0.00	0.00	0.00	0.00
rs174546_T	-0.02	(-0.04, -0.01)	-0.22	0.00	0.00	0.00	0.00
rs2081687_C	0.02	(0, 0.03)	-0.13	-0.02	0.00	0.00	0.00
rs4650994_A	-0.02	(-0.03, 0)	0.00	0.02	0.00	0.00	0.00
rs12444979_T	-0.02	(-0.03, 0)	0.00	0.00	0.00	0.00	-0.08
rs10901371_A	-0.02	(-0.03, 0)	0.00	-0.02	0.00	0.00	-0.15
rs3184504_C	0.01	(0, 0.02)	0.00	0.00	0.00	0.00	0.10
rs698842_T	0.01	(0, 0.02)	0.00	0.03	0.00	0.00	0.08
rs7570971_A	0.01	(0, 0.02)	0.09	0.00	0.00	0.00	0.00
rs7134375_A	-0.01	(-0.02, 0)	0.00	0.00	0.00	0.00	0.00
rs217386_A	0.01	(0, 0.02)	0.00	0.00	0.00	0.00	0.00
rs6864049_G	0.01	(0, 0.02)	-0.93	-0.11	0.00	0.00	0.02
rs10904908_G	-0.01	(-0.01, 0)	0.00	0.00	0.00	0.00	-0.13

Table D.1: Features assigned nonzero estimated coefficients by SASH, PASS, pLasso, SL, SS-RMRCE, or SS-uLasso. We also display the confidence intervals of SASH constructed using the procedure described in Section 2.4.