Journal of Machine Learning Research 26 (2025) 1-77 Submitted 11/23; Revised 1/25; Published 2/25

Efficient and Robust Semi-supervised Estimation of Average Treatment Effect with Partially Annotated Treatment and Response

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

Rajarshi Mukherjee RAM521@MAIL.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02120, USA

Tianxi Cai TCAI@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Department of Biomedical Informatics
Harvard Medical School
Boston, MA 02120, USA

Editor: Ilya Shpitser

Abstract

A notable challenge of leveraging Electronic Health Records (EHR) for treatment effect assessment is the lack of precise information on important clinical variables, including the treatment received and the response. Both treatment information and response cannot be accurately captured by readily available EHR features in many studies and require labor-intensive manual chart review to precisely annotate, which limits the number of available gold standard labels on these key variables. We considered average treatment effect (ATE) estimation when 1) exact treatment and outcome variables are only observed together in a small labeled subset and 2) noisy surrogates of treatment and outcome, such as relevant prescription and diagnosis codes, along with potential confounders are observed for all subjects. We derived the efficient influence function for ATE and used it to construct a semi-supervised multiple machine learning (SMMAL) estimator. We justified that our SMMAL ATE estimator is semi-parametric efficient with B-spline regression under low-dimensional smooth models. We developed the adaptive sparsity/model doubly robust estimation under high-dimensional logistic propensity score and outcome regression models. Results from simulation studies demonstrated the validity of our SMMAL method and its superiority over supervised and unsupervised benchmarks. We applied SMMAL to the assessment of targeted therapies for metastatic colorectal cancer in comparison to chemotherapy.

Keywords: semi-parametric efficiency, double robustness, high-dimensional regression, semi-supervised learning

1 Introduction

The 21st Century Cures Act and the Prescription Drug User Fee Act VII have shone a spotlight on the use of real-world evidence, generated from real-world data, to support regulatory-decision making on drug effectiveness. Large scale electronic health records (EHRs) data are being increasingly used for creating the real-world evidence on treatment effectiveness or efficacy (Franklin et al., 2021). In addition to the observational nature, another notable challenge in leveraging EHR for treatment effect assessment lies in the lack of readily available data for key clinical variables, including the treatment being investigated and the outcome of interest. Response variables such as disease progression may not be well represented by readily available EHR features (Bartlett et al., 2019). Treatment information can be partially captured but not always accurately reflected by procedure codes or medication prescription codes. New therapies may not be well coded in the introduction stage immediately after regulatory approval, and treatment initiation may be later than prescription date due to external factors such as insurance approval delay. For example in a real-world evidence study comparing chemotherapies and targeted therapies as first-line treatment for metastatic colorectal cancer, we discovered based on chart-review of 100 patients by a medical expert that 1) the progression-free-survival (PFS) outcomes were poorly structured in EHRs without clear indicators for progression or complete mortality data, and 2) the medication codes or natural language processing (NLP) identified mentions in notes could not accurately capture the use of targeted therapies (see Table 2).

Although it is possible to improve treatment or response definition by combining multiple EHR features through rule based or machine learning algorithms, these EHR derived features are at best good "surrogates" for approximating the true treatment or response information at patient level. Compared to the classic definition of surrogate, the notion of surrogate in retrospective EHR studies shares the availability trait but differs in the temporal order and causal pathway. In the advanced stage cancer trials or prospective studies, the progress-free-survival is often used as surrogate for overall survival because progression sometimes can be captured at an earlier time. In EHR studies, however, researchers do not have readily available progression data Y unless they perform the labor intensive manual chart review (Griffith et al., 2019), so it is natural to borrow information from the documentations about progression $\mathbf{S}$ like occurrence of diagnosis codes about secondary malignant neoplasm or NLP identified mention of metastasis at distant parts of body. These documentations $\mathbf{S}$ are considered as "surrogates" because 1) they can partially indicate progression, and 2) they are accessible earlier during the research process. Since the documentations $\mathbf{S}$ were recorded according to the true progression status Y, it is more reasonable to consider the true progression status temporally preceded and casually affected the documentation, $Y \to \mathbf{S}$ in a causal diagram. Directly using these surrogates as true treatment and outcome which would potentially induce bias in the subsequent analysis (Beaulieu-Jones et al., 2020). On the other hand, annotating exact treatment and response variables via manual chart review by domain expert is resource intensive, leading to limited sample size for gold standard labels on these key data. It is thus of great practical significance to leverage both the small number of gold standard labels and the vast unlabeled data to derive unbiased and efficient inference about the average treatment effect (ATE), fundamentally a nested problem with both missing data and causal inference components. When the labeling proportion is too small for standard (missing data) positivity assumption, the setting is often referred to as the semi-supervised learning (SSL).

Additional challenges arise from the high dimensionality of potential confounders. Unlike traditional cohort studies with a pre-specified number of clinical variables, EHRs provide rich data on a broader range and larger number of confounding factors (Hou et al., 2021c). Furthermore, multiple EHR features may be necessary to represent one specific clinical variable, further amplifying the dimensionality of features necessary to capture the underlying confounding factors. The complexity of the models from the high-dimensionality also increases the risk of model mis-specification for the propensity score (PS) and the outcome regression (OR). To the best of our knowledge, no method currently exists to estimate ATE under the SSL setting when both the treatment group, denoted by A, and the response, denoted by Y, are only observed in a small subset of the full data. We focus on the missing data patterns resulting from the lack of readily available data on the exact clinical information like treatment A and outcomes Y. For small subset, manual annotations can be created to recover the exact Y and A, but researchers have to rely on scalable yet imperfect computational tools to extract treatment outcome information over the majority of the vast EHR cohort, producing the surrogates $\mathbf{S}$ for Y and A. For conciseness, we refer to this specific SSL setting as *double missing SSL*. In this paper, we address the methodology gap by proposing **S**emi-supervised **M**ultiple **MA**chine **L**earning (SMMAL) estimators for ATE that leverage both the fully observed surrogates for Y and A, denoted by $\mathbf{S}$, and the partially observed gold standard labels on Y and A.

Under the supervised setting where both A and Y are observed, much progress has been made in recent years on estimation of ATE with confounding adjustment from machine-learning and/or high-dimensional regression. In the low-dimensional setting, the estimation of ATE is a well studied problem including procedures that achieve semi-parametric efficiency and double robustness (Robins et al., 1994; Bang and Robins, 2005). Extension to the high-dimensional setting, however, is not straightforward due to the slower convergence rates in the estimated model parameters and the difficulty posed not only by the bias and variance trade-off in the process of regularization but also by the inherent information theoretic barriers to obtaining fast enough estimation rates in high dimensional problems. Similar challenges arise when incorporating more flexible machine-learning models to overcome model mis-specifications. Following intuitions parallel to the low-dimensional setting, flexible approaches for confounder adjustments have been proposed via modeling of PS and OR, including L_1 regularized regression (Farrell, 2015), neural network (Farrell et al., 2021), and a general machine learning framework (Chernozhukov et al., 2018). Several methods accommodated the high dimensional confounder and achieved statistical inference on ATE based on consistent estimation for PS and OR, which translated to proper model specification and sparsity for high-dimensional regressions (Belloni et al., 2013; Liu et al., 2021; Hou et al., 2021a; Belloni et al., 2017, e.g.). Tan (2020) proposed a calibrated estimation that leads to valid inference for the average treatment effect even if one of the high-dimensional logistic PS or linear OR model is mis-specified. Smucler et al. (2019) formalized the concept of double robustness in high-dimensional setting by defining the sparsity double robustness and model double robustness properties; and also generalized the idea of Tan (2020) to a wide range of PS and OR models. For data with sample size n and dimension of covariate p, Smucler et al. (2019) defined the sparsity double robustness as producing $\sqrt{n}$-asymptotic normal estimator for ATE from consistently estimated PS and OR models as long as the *product* of sparsities for PS and OR models grow slower than $n\log(p)$. The model double robustness further allows the estimation of either PS or OR model to be inconsistent while still achieving $\sqrt{n}$-asymptotic normal estimation of ATE. Bradic et al. (2019) established a sharper sparsity double robustness property of the calibrated estimation. Unlike the two-model approaches (PS and OR) listed above, Wang and Shah (2020) considered a single model approach in which they debiased the regularized PS model in the inverse probability of treatment weighting estimator to achieve $\sqrt{n}$-inference.

Semi-supervised estimation for ATE is less studied. Existing literatures focused almost entirely on the setting where Y is observed for patients in the small labeled set of size n while A and surrogates/proxies of Y along with confounders $\mathbf{X}$ are observed for all subjects of size N. The semi-supervised learning (SSL) setting refers to the missing data proportion $(N-n)/N$ for Y tending to 1 along an asymptotic sequence where both number of labels and total sample size tend to infinity, $n, N \to \infty$. The SSL setting is distinguished from classical missing data problems as the standard (missing data) positivity assumption on observation rate is violated. SSL estimators for the ATE have been proposed by Cheng et al. (2021) when Y is missing-completely-at-random (MCAR) and by Zhang et al. (2023) and Kallus and Mao (2024) when Y is missing-at-random (MAR). However, these methods cannot be easily adapted to the setting where both Y and A are missing. The missingness in A is fundamentally different from the missingness in Y since treatment is an internal node in the causal pathway "confounder($\mathbf{X}$)$\to$treatment(A)$\to$outcome(Y)$\to$surrogates($\mathbf{S}$)" (Figure 1), introducing technical challenges on the projection by conditional expectation in the semi-parametric analysis.

In this paper, we propose an efficient and robust SSL estimator for ATE when both Y and A are only observed for a small labeled subset but the confounders $\mathbf{X}$ and surrogates $\mathbf{S}$ for Y and A are observed for all N patients. We derived the SMMAL estimator by first deriving the efficient influence function for the ATE under this double missing SSL setting and then constructing a cross-fitted multiple machine learning estimator. We subsequently provided a formal characterization of semi-parametric efficiency under the double missing SSL setting with the SMMAL estimator coupled to B-spline regressions over low-dimensional space. We also designed a doubly robust estimator with a two-layer cross-fitted calibrated estimation for high-dimensional logistic PS and OR models. Via cross-fitting and a truncation in initial OR/PS predictions, we relaxed the sparsity assumptions in the initial estimation for PS and OR, previously required for $\sqrt{n}$-inference of ATE (Tan, 2020; Smucler et al., 2019). We further showed that our doubly robust SMMAL estimator attains 1) the rate double robustness when both PS and OR models are correct and 2) model double robustness when one of them is correct, as defined by Smucler et al. (2019). The SMMAL estimator also does not require correct specifications of the imputation models for A or Y for proper inference under MCAR assumption. We summarize our key contributions herein:

1. We formalized the efficient estimation under a general SSL setting (including specifically the double missing SSL setting) with a decaying observation rate that violates the classical (missing data) positivity assumption. Our theory justified the efficiency claims of existing works and can provide benchmark for future work in this direction. A discussion regarding the subtleties and challenges involved in this formalization and subsequent analyses can be found in Remark 6.

2. We laid out a general approach for efficient SSL with a complex missing data structure. Our general approach is particularly convenient when the missing data and dependence patterns render typical projection approach difficult for the semi-parametric theory. A explanation of the challenge from missing treatment under double missing SSL setting can be found in Remark 4, and the general framework is given in Section 7.

3. We made progress in statistical inference on ATE based on the doubly robust estimation with high-dimensional confounders achieving the sparsity/model double robustness. We generalized to the SSL setting the techniques in the existing literatures on calibrated estimation of PS and OR models so that the final ATE estimator has weak dependence on estimation of these models, characterized by small derivatives, also known as the Neyman Orthogonality (Chernozhukov et al., 2018). Using a truncation of initial model prediction, we removed the sparsity requirement in initial estimation. In addition, we demonstrated that the SSL estimation derived from our modified semi-parametric theory contributed to robustness toward estimation of the imputation models. The comparison with related work can be found in Remark 9.

The paper is organized as follows. In Section 2, we introduce our causal inference structure under the double missing SSL setting along with the notations. In Section 3, we first present the efficient influence function, followed by the multiple machine learning estimator and the model multiply robust estimator derived from the efficient influence function. In Section 4, we state the theoretical guarantees of the $\sqrt{n}$-inference on the ATE from our methods, whose proofs are provided in the Supplementary Materials. We also provide the semi-parametric efficiency lower bound for average treatment effect under double missing SSL setting in low-dimensional space. In Section 5, we assess the finite sample performance of our SSL methods and compare them to supervised benchmarks. In Section 6, we apply SMMAL to the real-world evidence study on targeted therapies for metastatic colorectal cancer in comparison with chemotherapy. In Section 7, we offer the efficiency lower bound for general low-dimensional parameters under broader SSL settings with flexible missing data components. In Section 8, we conclude with a brief discussion.

2 Setting and notation

For the i-th observation in a study of N subjects, $Y_i \in \mathbb{R}$ denotes the outcome variable, $A_i \in \{0,1\}$ denotes the treatment group indicator, $R_i \in \{0,1\}$ indicates whether (Y_i, A_i) is annotated, $\mathbf{S}_i \in \mathbb{R}^q$ denotes the surrogates for Y_i and A_i, and $\mathbf{X}_i \in \mathbb{R}^{p+1}$ denotes the vector of potential confounders including 1 as the first element. We use the notations without the subscript indices $Y, A, R, \mathbf{S}, \mathbf{X}$ to denote the generic versions of these random variables. In EHR studies, routine documentations on treatments and outcomes in the form of digital codes and mentions in narrative notes are often prone to errors (Zhang et al., 2019) and hence can only serve as surrogates $\mathbf{S}$. To ascertain Y and A, researchers may design the sampling scheme for a representative labeled subset, $\{i \in [N] : R_i = 1\}$, over which the exact data (Y_i, A_i) are annotated by medical experts, where $[N] = \{1, ..., N\}$. For those with $R_i = 0$, exact values of the pair (Y_i, A_i) are not ascertained, creating the joint missingness

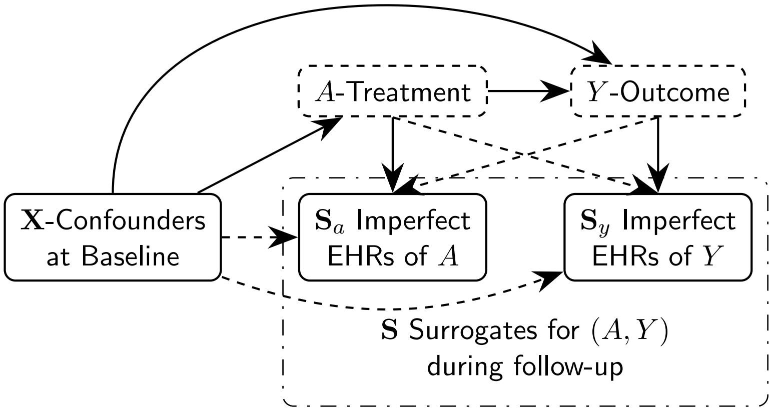

A-Treatment
Y-Outcome
$\mathbf{X}$-Confounders at Baseline
$\mathbf{S}_a$ Imperfect EHRs of A
$\mathbf{S}_y$ Imperfect EHRs of Y
$\mathbf{S}$ Surrogates for (A, Y) during follow-up

Figure 1: Causal diagrams of double missing SSL setting with missing treatment and outcome. The surrogates $\mathbf{S}$ represent the imprecise documentation of A and Y, which should be predictive for A and Y but not affecting the causal identification based on perfect data $(Y, A, \mathbf{X})$.

of (Y_i, A_i). The observed data consist of N independent and identically distributed (i.i.d.) random vectors, $\mathscr{D} = \{\mathbf{D}_i = (R_i, R_iY_i, R_iA_i, \mathbf{W}_i^\intercal)^\intercal, i = 1, ..., N\}$, where $\mathbf{W}_i = (\mathbf{X}_i^\intercal, \mathbf{S}_i^\intercal)^\intercal$.

We assume the MCAR mechanism for the sampling process with

$$R \perp\!\!\!\perp (Y, A, \mathbf{X}, \mathbf{S}), \tag{1}$$

and the number of labelled sample is $n = \sum_{i=1}^N R_i$ with the proportion of labeled observation being $\rho_\mathsf{N} = \mathbb{E}(R) \in (0, 1)$ with $\rho_\mathsf{N} \to 0$ as $N \to \infty$ while the expected number of labels also grow asymptotically to infinity $\rho_\mathsf{N} N \to \infty$. Under MCAR formulation, the size of labeled subset n is a random variable asymptotically equivalent to $\rho_\mathsf{N} N$, as $n/(\rho_\mathsf{N} N) = 1 + o_p(1)$. We use a simplified notation "$V_\mathsf{N} \asymp U_\mathsf{N}$", e.g. $n \asymp \rho_\mathsf{N} N$, to describe the equivalence in stochastic order, $V_\mathsf{N}/U_\mathsf{N} = O_p(1)$ and $U_\mathsf{N}/V_\mathsf{N} = O_p(1)$. To better reflect the dependence on labeled set and compare with supervised benchmarks, we use n instead of $\rho_\mathsf{N} N$ when describing the asymptotic orders. As the exception, we use $\rho_\mathsf{N} N$ in the derivation of efficiency lower bound. Extension to MAR is plausible through modeling and estimating the missing data pattern $\mathbb{P}(R = 1 \mid \mathbf{W})$ under classical semi-parametric theory, but a few technical and practical challenges exist as we listed in the Section 8. Thorough investigation of the MCAR setting would already provide methodological guidance to the rapidly growing real-world evidence studies in which random subsets are selected for gold-standard validation of intervention and outcome data (Hou et al., 2023). Our MCAR formulation, as opposed to the two sample formulations in existing statistical semi-supervised learning literatures (Chakrabortty et al., 2019; Zhang et al., 2023; Hou et al., 2021b), connects better with existing literature on semi-parametric estimation and missing data. The results under MCAR, with minor modification in theoretical derivations, are largely applicable to the other similar formulation like first n samples $R_i = \mathrm{I}(i \leq n)$ or sampling without replacement for a deterministic sequence n.

To properly define the causally interpretable ATE, we adopt the typical counterfactual outcome framework and its standard assumptions (Imbens and Rubin, 2015; Hernan and Robins, 2023). Let $Y^{(a)}$ be the counterfactual outcome with treatment set as a, for $a \in \{0, 1\}$. The ATE is defined as

$$\Delta_* = \mathbb{E}\left(Y^{(1)} - Y^{(0)}\right). \tag{2}$$

We make the following standard assumptions regarding the triplet $(Y, A, \mathbf{X})$,

Assumption 1 *(a) Consistency:* $Y = Y^{(A)}$*;*

(b) (Causal inference) Positivity of treatment assignment: $1/M \leq \mathbb{P}(A = 1 \mid \mathbf{X}) \leq 1 - 1/M$ *almost surely for an absolute constant* $M < \infty$*;*

(c) Ignorability: $\left(Y^{(1)}, Y^{(0)}\right) \perp\!\!\!\perp A \mid \mathbf{X}$.

The (causal inference) positivity in Assumption 1 is imposed on the treatment assignment A, which should be distinguished from the (missing data) positivity regarding the observation indicator R. Under the Assumption 1, the ATE can be alternatively expressed as

$$\Delta_* = \mathbb{E}\left\{\mathbb{E}(Y \mid \mathbf{X}, A = 1) - \mathbb{E}(Y \mid \mathbf{X}, A = 0)\right\}. \tag{3}$$

In the motivating EHR studies, $\mathbf{S}_i$ represents the documentations and retrospective data curation of (Y_i, A_i), such as the presence of diagnosis code in follow-up for outcomes and medication codes at baseline for treatments, that are conceivably determined by the underlying truth (Y_i, A_i). In Figure 1, we present a setting such that the surrogates can be classified into those for A_i and those for Y_i, $\mathbf{S}_i = (\mathbf{S}_{i,a}^\intercal, \mathbf{S}_{i,y}^\intercal)^\intercal$. The causal identification (3) still holds with the introduction of additional variable $\mathbf{S}_i$. Sometimes $\mathbf{S}_i$ may contain colliders that are affected by both treatment A_i and outcome Y_i, $A_i \to \mathbf{S}_i \leftarrow Y_i$, e.g. increased code counts from frequent healthcare visits as part of intense treatment or caused by poor outcome. Adjustment of colliders would distort causal relationship $A_i \to Y_i$ and should be excluded from causal identification (Hernan and Robins, 2023, Chapter 6.4). Throughout the paper, we assume MCAR (1) and Assumption 1.

Remark 1 *We herein summarize the setting of our study.*

- *Over a small randomly sampled labeled subset, we can causally identify ATE with outcome* Y_i*, binary treatment* A_i *and confounders* $\mathbf{X}_i$ *under standard consistency, (causal inference) positivity and ignorability assumptions.*
- *We seek to robustly enhance the efficiency of estimating ATE by incorporating the large unlabeled data containing confounders* $\mathbf{X}_i$ *and surrogates* $\mathbf{S}_i$ *without stringent model assumptions on* $\mathbf{S}_i$*. The method should adaptively achieve*
  - *better efficiency if* $\mathbf{S}_i \mid \mathbf{X}_i$ *can effectively inform* $(Y_i, A_i) \mid \mathbf{X}_i$*;*
  - *the same property as the ATE estimated from labeled subset if* $\mathbf{S}_i \mid \mathbf{X}_i$ *cannot inform* $(Y_i, A_i) \mid \mathbf{X}_i$*.*

3 SMMAL Estimation

We start by presenting in Section 3.1 the efficient influence function under the double missing SSL setting without assuming known model for unlabeled data. Deviating from the classical missing data setting, the derived efficient influence functions under double missing SSL setting have diverging variances, which requires a formal justification of its connection with efficiency lower bound in Sections 4.2. Our approach is hence distinguished from existing SSL literatures (Cheng et al., 2021; Kallus and Mao, 2024) that considered a simplified theoretical formulation to define the efficient influence function assuming known model for large unlabeled data. Then, we discuss the estimation of ATE with different ways of estimating the nuisance models involved in the efficient influence function in Section 3.2 for low-dimensional $\mathbf{X}$ and in Section 3.3 for high-dimensional $\mathbf{X}$. As the standard tool to control over-fitting from using estimated models in subsequent estimation procedures (Lin and Ying, 1994; Chernozhukov et al., 2018; Newey and Robins, 2018; Hou et al., 2021b), cross-fitting is adopted for both settings, where we split the data into K (e.g. $K = 5$) folds of approximately equal size. For $k = 1, ..., K$, we let $\mathcal{I}_k$ denote the index set for the kth fold of the data with size $N_k = |\mathcal{I}_k|$ and let $\mathcal{I}_k^c = \{1, \ldots, N\} \setminus \mathcal{I}_k$, where $|\mathcal{I}|$ denotes the cardinality of $\mathcal{I}$. Here, we do not split the folds separately for labeled and the unlabeled data because the label indicator R_i is random under the MCAR formulation (1).

3.1 The efficient influence function

We define the following nuisance models:

$$\begin{aligned} &\text{PS:} && \mathbb{P}(A = a \mid \mathbf{X}) = \pi(a, \mathbf{X}), && \text{OR:} && \mathbb{E}(Y \mid A = a, \mathbf{X}) = \mu(a, \mathbf{X}), \\ &\text{Imputations:} && \mathbb{P}(A = a \mid \mathbf{W}) = \Pi(a, \mathbf{W}), && && \mathbb{E}(Y \mid A = a, \mathbf{W}) = m(a, \mathbf{W}). \end{aligned}$$

We use the subscript star to indicate the true models, π_*, μ_*, Π_*, m_*. Starting from the efficient influence function with complete (cmp) observation of treatment and outcome (Robins et al., 1994; Kallus and Mao, 2024),

$$\begin{aligned} \phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) =& \mu_*(1, \mathbf{X}) - \mu_*(0, \mathbf{X}) + \frac{\mathrm{I}(A = 1)}{\pi_*(1, \mathbf{X})}\{Y - \mu_*(1, \mathbf{X})\} \\ &- \frac{\mathrm{I}(A = 0)}{\pi_*(0, \mathbf{X})}\{Y - \mu_*(0, \mathbf{X})\} - \Delta_*, \end{aligned}$$

we produced the efficient influence function through the following mapping

$$
\begin{aligned}
&\phi_{\mathsf{SSL}}(RY, RA, \mathbf{W}, R)\\
=&\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\} + \frac{R}{\rho_{\mathsf{N}}}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}] \qquad (4)\\
=&\mu_*(1, \mathbf{X}) + \frac{\Pi_*(1, \mathbf{W})}{\pi_*(1, \mathbf{X})}\{m_*(1, \mathbf{W}) - \mu_*(1, \mathbf{X})\}\\
&- \mu_*(0, \mathbf{X}) - \frac{\Pi_*(0, \mathbf{W})}{\pi_*(0, \mathbf{X})}\{m_*(0, \mathbf{W}) - \mu_*(0, \mathbf{X})\} - \Delta_*\\
&+ \frac{R\{\mathrm{I}(A = 1)Y - \mathrm{I}(A = 1)\mu_*(1, \mathbf{X}) - \Pi_*(1, \mathbf{W})m_*(1, \mathbf{W}) + \Pi_*(1, \mathbf{W})\mu_*(1, \mathbf{X})\}}{\rho_{\mathsf{N}}\pi_*(1, \mathbf{X})}\\
&- \frac{R\{\mathrm{I}(A = 0)Y - \mathrm{I}(A = 0)\mu_*(0, \mathbf{X}) - \Pi_*(0, \mathbf{W})m_*(0, \mathbf{W}) + \Pi_*(0, \mathbf{W})\mu_*(0, \mathbf{X})\}}{\rho_{\mathsf{N}}\pi_*(0, \mathbf{X})}. \qquad (5)
\end{aligned}
$$

In the formula (4) that produces ϕ_{SSL} from ϕ_{cmp} , $\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}$ is the maximal information on ATE from the unlabeled data with a known imputation model, and the second term is the price for training the best imputation model over the labeled data. We provide the rigorous justification of this procedure in Section 4.

The efficient influence function in the missing data context is usually derived by projecting an arbitrary initial influence function to the nuisance tangent space (Tsiatis, 2007). The approach has been applied to the SSL setting with missing outcome by first deriving the efficient influence function under missing data setting and then setting $n/N \asymp \rho_{\mathsf{N}} = 0$ for the SSL setting with very large unlabeled data (Kallus and Mao, 2024). No formal justification of efficiency has been given in exiting literatures under the semi-supervised setting with $\rho_{\mathsf{N}} \to 0$ yet $\rho_{\mathsf{N}} > 0$. Moreover, such standard procedure for deriving efficient influence function under missing data or causal inference settings is usually specific for the assumed dependence structure among variables, reflected by the correspondent chain-rule decomposition of nuisance model tangent space (Robins et al., 1994; Tsiatis, 2007; Kallus and Mao, 2024; Cheng et al., 2021). For estimating of ATE under SSL setting, existing formulation focused on the surrogates $\mathbf{S}$ that are defined as short-term markers for long-term outcomes Y, represented by the $\mathbf{S} \to Y$ dependence pattern in causal diagram (Kallus and Mao, 2024; Cheng et al., 2021). The generalization to other types of surrogates $\mathbf{S}$ is currently absent. For example, surrogates $\mathbf{S}$ in our motivating EHR studies were imperfect documentations for treatment and outcome variables (A, Y), represented by the $(A, Y) \to \mathbf{S}$ dependence pattern (Figure 1). The shift from $\mathbf{S} \to Y$ to $(A, Y) \to \mathbf{S}$ also creates the technical challenges in deriving projections to nuisance model tangent space defined according to the dependence pattern (see Section E2 of the Supplementary Materials).

While our derivation of efficient ϕ_{SSL} also involved projecting an inefficient $(R/\rho_{\mathsf{N}})\phi_{\mathsf{cmp}}$, a common approach among existing literatures (Robins et al., 1994; Kallus and Mao, 2024), our approach did not impose stringent assumptions on surrogates $\mathbf{S}$. To provide a general theoretical basis consistent across various surrogate mechanism, we established the connection between efficiency lower bound of complete data setting and that of double missing SSL setting through asymptotic local minimax result similar to Begun et al. (1983) in Section 4.2. Our efficiency lower bound justified projecting complete data efficient influence function ϕ_{cmp} to derive the SSL efficient influence function ϕ_{SSL}. We further generalized the efficiency theory to other parameters with missing data under SSL setting in Section 7. Our alternative justification only requires 1) the target ATE parameter Δ can be identified by $(Y, A, \mathbf{X})$ through ϕ_{cmp} and 2) $\mathbf{S}$ can provide information on (Y, A) when they are not observed over the unlabeled set. Hence, our framework covered a broad range of surrogate mechanism including both the setting considered by Kallus and Mao (2024) and the causal diagram in Figure 1.

3.2 SMMAL Procedure

Inspired by the double machine learning estimation (Chernozhukov et al., 2018) based on ϕ_{cmp}, we propose the following SMMAL estimator for ATE:

1. For each labelled fold k, we estimate the nuisance models by the out-of-fold data $\mathcal{I}_k^c$, obtaining $\widehat{\pi}^{\mathsf{(k)}}$, $\widehat{\mu}^{\mathsf{(k)}}$, $\widehat{\Pi}^{\mathsf{(k)}}$, $\widehat{m}^{\mathsf{(k)}}$;

2. Construct the estimated influence functions

$$
\begin{aligned}
\widehat{\mathcal{V}}_{ik} =& \widehat{\mu}^{\mathsf{(k)}}(1, \mathbf{X}_i) + \frac{\widehat{\Pi}^{\mathsf{(k)}}(1, \mathbf{W}_i)}{\widehat{\pi}^{\mathsf{(k)}}(1, \mathbf{X}_i)} \{\widehat{m}^{\mathsf{(k)}}(1, \mathbf{W}_i) - \widehat{\mu}^{\mathsf{(k)}}(1, \mathbf{X}_i)\} \\
& - \widehat{\mu}^{\mathsf{(k)}}(0, \mathbf{X}_i) - \frac{\widehat{\Pi}^{\mathsf{(k)}}(0, \mathbf{W}_i)}{\widehat{\pi}^{\mathsf{(k)}}(0, \mathbf{X}_i)} \{\widehat{m}^{\mathsf{(k)}}(0, \mathbf{W}_i) - \widehat{\mu}^{\mathsf{(k)}}(0, \mathbf{X}_i)\} \\
& + \frac{R_i\{A_i Y_i - A_i \widehat{\mu}^{\mathsf{(k)}}(1, \mathbf{X}_i)\}}{\rho_{\mathsf{N}} \widehat{\pi}^{\mathsf{(k)}}(1, \mathbf{X}_i)} - \frac{R_i\{(1 - A_i) Y_i - (1 - A_i) \widehat{\mu}^{\mathsf{(k)}}(0, \mathbf{X}_i)\}}{\rho_{\mathsf{N}} \widehat{\pi}^{\mathsf{(k)}}(0, \mathbf{X}_i)} \\
& - \frac{R_i\{\widehat{\Pi}^{\mathsf{(k)}}(1, \mathbf{W}_i) \widehat{m}^{\mathsf{(k)}}(1, \mathbf{W}_i) - \widehat{\Pi}^{\mathsf{(k)}}(1, \mathbf{W}_i) \widehat{\mu}^{\mathsf{(k)}}(1, \mathbf{X}_i)\}}{\rho_{\mathsf{N}} \widehat{\pi}^{\mathsf{(k)}}(1, \mathbf{X}_i)} \\
& + \frac{R_i\{\widehat{\Pi}^{\mathsf{(k)}}(0, \mathbf{W}_i) \widehat{m}^{\mathsf{(k)}}(0, \mathbf{W}_i) - \widehat{\Pi}^{\mathsf{(k)}}(0, \mathbf{W}_i) \widehat{\mu}^{\mathsf{(k)}}(0, \mathbf{X}_i)\}}{\rho_{\mathsf{N}} \widehat{\pi}^{\mathsf{(k)}}(0, \mathbf{X}_i)}.
\end{aligned}
$$

   and estimate the ATE by

$$
\widehat{\Delta}_{\mathsf{SMMAL}} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \widehat{\mathcal{V}}_{ik}. \tag{6}
$$

3. Estimate the asymptotic variance of $\sqrt{n}(\widehat{\Delta}_{\mathsf{SMMAL}} - \Delta_*)$ by

$$
\widehat{\mathcal{V}}_{\mathsf{SMMAL}} = \frac{\rho_{\mathsf{N}}}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (\widehat{\mathcal{V}}_{ik} - \widehat{\Delta}_{\mathsf{SMMAL}})^2. \tag{7}
$$

Here we considered the $\sqrt{n}$ standardized estimation error $\sqrt{n}(\widehat{\Delta}_{\mathsf{SMMAL}} - \Delta_*)$ instead of the $\sqrt{N}$ standardized estimation error $\sqrt{N}(\widehat{\Delta}_{\mathsf{SMMAL}} - \Delta_*)$ because the latter is diverging at $\sqrt{N/n} \asymp \rho_{\mathsf{N}}^{-1/2}$ rate due to the unbounded variance of R_i/ρ_{N} as $\rho_{\mathsf{N}} \to 0$. The $(1-\alpha) \times 100\%$ confidence interval for ATE can be constructed with $\widehat{\Delta}_{\mathsf{SMMAL}}$ and $\widehat{\mathcal{V}}_{\mathsf{SMMAL}}$,

$$
\left[\widehat{\Delta}_{\mathsf{SMMAL}} - \mathcal{Z}_{\alpha/2} \sqrt{\widehat{\mathcal{V}}_{\mathsf{SMMAL}}/n},\ \widehat{\Delta}_{\mathsf{SMMAL}} + \mathcal{Z}_{\alpha/2} \sqrt{\widehat{\mathcal{V}}_{\mathsf{SMMAL}}/n}\right]
$$

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

Similar to existing results in double machine learning literature, any estimators for the nuisance models with suitable rates of consistency can be used in our proposal as well. For low-dimensional $\mathbf{W}$ and smooth nuisance models, we can choose B-spline regression with proper order and degrees. Precise discussions on these rates, related conditions for general estimators and relevant smoothness classes for B-spline regression are listed in Section 4.2.

3.3 Doubly Robust SMMAL Construction in high-dimensions

We next discuss a specific construction of the SMMAL estimator when the dimensions p and q grow with n and p may be larger than n. In real-world evidence studies using EHRs, confounding adjustment often involves selection of the few determinants of treatment and risk factors for outcomes from a large number of candidate variables (Hou et al., 2021c, 2022). We focus on the binary Y and put the high-dimensional logistic regression models with link function $g(x) = 1/(1+e^{-x})$ on the nuisance models

$$\begin{aligned} \pi(1,\mathbf{X}) = g(\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{X});\ \mu(a,\mathbf{X}) = g(\boldsymbol{\beta}_{\mathsf{a}}^{\mathsf{T}}\mathbf{X}),\ a=0,1; \\ \Pi(1,\mathbf{W}) = g(\boldsymbol{\xi}^{\mathsf{T}}\mathbf{W});\ m(a,\mathbf{W}) = g(\boldsymbol{\zeta}_{\mathsf{a}}^{\mathsf{T}}\mathbf{W}),\ a=0,1. \end{aligned} \tag{8}$$

We denote the derivatives of the link g as $\dot{g}(x) = e^x/(1+e^x)^2$ and the corresponding loss function as $\ell(y,x) = \log(1+e^x) - yx$. Other types of generalized linear models for OR model $\mu(a,\mathbf{X})$ may also be considered and derived similarly. To enhance the robustness against model mis-specification in π and μ, we propose a bias-reducing calibration after an initial estimation (Smucler et al., 2019). We added another layer of cross-fitting to reduce the overfitting bias when using initial estimators in the bias-reducing calibration. Compare with the general SMMAL algorithm in Section 3.2, the generic estimation process for nuisance models (Step 1 in Section 3.2) is expanded into the Step 1-4 of the following SMMAL algorithm for high-dimensional logistic regression. To ensure that estimated PS and OR are bounded away from zero and one, we propose to truncate linear predictors according to a predetermined constant M corresponding to a reasonable range for PS and OR probabilities, e.g. $M = 2.2$ for range $[0.1, 0.9]$. Our algorithm for doubly robust SMMAL estimator $\widehat{\Delta}_{\mathsf{DR}}$ has the following steps:

1. For each labelled fold k, we estimate the imputation models by the Lasso over out-of-fold data $\mathcal{I}_k^c$,

$$\begin{aligned} \widehat{\boldsymbol{\xi}}^{(\mathsf{k})} &= \underset{\boldsymbol{\xi}\in\mathbb{R}^{p+q+1}}{\operatorname{argmin}} \frac{\sum_{i\in\mathcal{I}_k^c} R_i \ell(A_i, \boldsymbol{\xi}^{\mathsf{T}}\mathbf{W}_i)}{\sum_{i\in\mathcal{I}_k^c} R_i} + \lambda_\eta \|\boldsymbol{\xi}\|_1, \qquad \lambda_\eta \asymp \sqrt{\log(p+q)/n}, \\ \widehat{\boldsymbol{\zeta}}_{\mathsf{a}}^{(\mathsf{k})} &= \underset{\boldsymbol{\zeta}\in\mathbb{R}^{p+q+1}}{\operatorname{argmin}} \frac{\sum_{i\in\mathcal{I}_k^c} \mathrm{I}(A_i = a) R_i \ell(Y_i, \boldsymbol{\zeta}^{\mathsf{T}}\mathbf{W}_i)}{\sum_{i\in\mathcal{I}_k^c} \mathrm{I}(A_i = a) R_i} + \lambda_\zeta \|\boldsymbol{\zeta}\|_1, \lambda_\zeta \asymp \sqrt{\log(p+q)/n}; \end{aligned} \tag{9}$$

Choice of the imputation method is flexible (See Remark 10).

2. For each labelled fold pair (k_1, k_2), we estimate the initial PS and OR models by the Lasso over out-of-two-folds data $\mathcal{I}^c_{k_1,k_2} = (\mathcal{I}_{k_1} \cup \mathcal{I}_{k_2})^c$,
$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}}^{(\mathsf{k}_1,\mathsf{k}_2)}_{\mathsf{init}} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{\sum_{i\in\mathcal{I}^c_{k_1,k_2}} R_i \ell(A_i, \boldsymbol{\alpha}^\intercal \mathbf{X}_i)}{\sum_{i\in\mathcal{I}^c_{k_1,k_2}} R_i} + \lambda_{\alpha,\mathsf{init}} \|\boldsymbol{\alpha}\|_1, \\
\widehat{\boldsymbol{\beta}}^{(\mathsf{k}_1,\mathsf{k}_2)}_{\mathsf{a},\mathsf{init}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{\sum_{i\in\mathcal{I}^c_{k_1,k_2}} \mathrm{I}(A_i = a) R_i \ell(Y_i, \boldsymbol{\beta}^\intercal \mathbf{X}_i)}{\sum_{i\in\mathcal{I}^c_{k_1,k_2}} \mathrm{I}(A_i = a) R_i} + \lambda_{\beta,\mathsf{a},\mathsf{init}} \|\boldsymbol{\beta}\|_1,
\end{aligned} \tag{10}
$$
with $\lambda_{\alpha,\mathsf{init}}, \lambda_{\beta,\mathsf{a},\mathsf{init}} \asymp \sqrt{\log(p)/n}$;

3. Define the truncation at $2M$, $\tau(x)=\operatorname{sign}(x)\min\{|x|, 2M\}$, and its composition with functions $\dot{g}_\tau(x) = \dot{g}(\tau(x))$ and $\exp_\tau(x) = \exp(\tau(x))$. For each labelled fold k_1, we construct the calibrated losses,
$$
\begin{aligned}
\ell_{\alpha,\mathsf{a}}(A, \boldsymbol{\alpha}^\intercal \mathbf{X}; \boldsymbol{\beta}) &= \dot{g}_\tau\left(\mathbf{X}^\intercal \boldsymbol{\beta}\right) \{(a - A)\boldsymbol{\alpha}^\intercal \mathbf{X} + I(A = a) e^{(-1)^a \boldsymbol{\alpha}^\intercal \mathbf{X}}\}, \\
\ell_{\beta,\mathsf{a}}(Y, \boldsymbol{\beta}^\intercal \mathbf{X}; \boldsymbol{\alpha}) &= \exp_\tau\{(-1)^a \boldsymbol{\alpha}^\intercal \mathbf{X}\}\, \ell(Y_i, \boldsymbol{\beta}^\intercal \mathbf{X}_i),
\end{aligned} \tag{11}
$$
and estimate the PS and OR models by cross-fitting within out-of-fold data $\mathcal{I}^c_{k_1}$,
$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}}^{(\mathsf{k}_1)}_{\mathsf{a}} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{k_2\neq k_1} \sum_{i\in\mathcal{I}_{k_2}} \frac{R_i}{n} \ell_{\alpha,\mathsf{a}}(A, \boldsymbol{\alpha}^\intercal \mathbf{X}_i; \widehat{\boldsymbol{\beta}}^{(\mathsf{k}_1,\mathsf{k}_2)}_{\mathsf{a},\mathsf{init}}) + \lambda_{\alpha,\mathsf{a}} \|\boldsymbol{\alpha}\|_1, \\
\widehat{\boldsymbol{\beta}}^{(\mathsf{k}_1)}_{\mathsf{a}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{\sum_{k_2\neq k_1} \sum_{i\in\mathcal{I}_{k_2}} I(A_i = a) R_i \ell_{\beta,\mathsf{a}}(Y_i, \boldsymbol{\beta}^\intercal \mathbf{X}_i; \widehat{\boldsymbol{\alpha}}^{(\mathsf{k}_1,\mathsf{k}_2)}_{\mathsf{init}})}{\sum_{i\in\mathcal{I}^c_{k_1}} I(A_i = a) R_i} + \lambda_{\beta,\mathsf{a}} \|\boldsymbol{\beta}\|_1,
\end{aligned} \tag{12}
$$
with $\lambda_{\alpha,\mathsf{a}}, \lambda_{\beta,\mathsf{a}} \asymp \sqrt{\log(p)/n}$.

4. Construct the nuisance model estimators:
$$
\begin{aligned}
\widehat{\pi}^{(\mathsf{k})}(1, \mathbf{X}_i) = g_\tau(\mathbf{X}_i^\intercal \widehat{\boldsymbol{\alpha}}^{(\mathsf{k})}_1),\ \widehat{\pi}^{(\mathsf{k})}(0, \mathbf{X}_i) = g_\tau(-\mathbf{X}_i^\intercal \widehat{\boldsymbol{\alpha}}^{(\mathsf{k})}_0),\ \widehat{\mu}^{(\mathsf{k})}(a, \mathbf{X}_i) = g(\mathbf{X}_i^\intercal \widehat{\boldsymbol{\beta}}^{(\mathsf{k})}_{\mathsf{a}}), \\
\widehat{\Pi}^{(\mathsf{k})}(a, \mathbf{W}_i) = g(\mathbf{W}_i^\intercal \widehat{\boldsymbol{\xi}}^{(\mathsf{k})}),\ \widehat{m}^{(\mathsf{k})}(a, \mathbf{W}_i) = g(\mathbf{W}_i^\intercal \widehat{\boldsymbol{\zeta}}^{(\mathsf{k})}_{\mathsf{a}});
\end{aligned} \tag{13}
$$

5. Estimate the ATE by sending (13) to (6), producing $\widehat{\Delta}_{\mathsf{DR}}$.

6. Estimate the variance by sending (13) and $\widehat{\Delta}_{\mathsf{DR}}$ to (7), producing $\widehat{\mathcal{V}}_{\mathsf{DR}}$.

The $(1-\alpha)\times 100\%$ confidence interval for ATE can be constructed with $\widehat{\Delta}_{\mathsf{DR}}$ and $\widehat{\mathcal{V}}_{\mathsf{DR}}$,
$$
\left[\widehat{\Delta}_{\mathsf{DR}} - \mathcal{Z}_{\alpha/2}\sqrt{\widehat{\mathcal{V}}_{\mathsf{DR}}/n},\ \widehat{\Delta}_{\mathsf{DR}} + \mathcal{Z}_{\alpha/2}\sqrt{\widehat{\mathcal{V}}_{\mathsf{DR}}/n}\right]
$$
where $\mathcal{Z}_{\alpha/2}$ is the $1-\alpha/2$ quantile of standard normal distribution.

The calibrated losses (11) aim to estimate OR and PS models by approximately solving the equations of the partial derivatives of $\widehat{\Delta}_{\mathsf{DR}}$ with respect to PS and OR models being zero (Tan, 2020; Smucler et al., 2019). The correctly specified model will be recovered as it can be identified by the same equation. Even with mis-specified model, Δ will be insensitive to estimation errors in OR and PS models, guaranteed by the small partial derivatives. The property is referred to as the *Neyman orthogonality* (Chernozhukov et al., 2018), which produces $\sqrt{n}$ asymptotic normal estimator with sub $\sqrt{n}$ rate nuisance model estimations. We didn't use imputations to improve estimation of OR and PS models because there is no asymptotic efficiency gain due to Neyman orthogonality but potential risk of introducing bias.

To control the overfitting bias from the sequential estimation process with 3 steps $(\widehat{\boldsymbol{\alpha}}_{\mathsf{init}}, \widehat{\boldsymbol{\beta}}_{\mathsf{a,init}}) \to (\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}, \widehat{\boldsymbol{\beta}}_{\mathsf{a}}) \to \widehat{\Delta}_{\mathsf{DR}}$, we propose the two-level cross-fitting for learning ATE in (10) and (12), previously considered for semi-supervised learning of high-dimensional regression in (Hou et al., 2021b). The two-level cross-fitting has the advantage of having larger training set for each Lasso (using $k-2$ folds) compared to the averaging after data splitting (using $(k-1)/2$ folds) in Smucler et al. (2019). If we choose $K = 10$, we are able to use at least 80% data while the data splitting in Smucler et al. (2019) may only use 45% data. Larger training sample typically allows the choice of smaller penalty factor thus reducing the bias. Taking averaging after data splitting, however, cannot reduce bias.

The truncation τ at M in (12) secured the (causal inference) positivity property of the initially estimated models with no compromise in estimation accuracy. Truncation of PS has been commonly invoked in practice when (causal inference) positivity holds in principle but is violated practically by estimated PS (Petersen et al., 2012; Ju et al., 2019). Our method generalized the truncation to OR prediction for binary outcome and identified a novel theoretical property of relaxing sparsity requirement for initial Lasso with the truncation. When the initial estimated model $\mathbf{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\alpha}}_{\mathsf{init}}$ is consistent for true models satisfying the (causal inference) positivity conditions, i. e. $\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\alpha}_*$ such that $0 < g(-M) \le g(\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\alpha}_*) \le g(M) < \infty$, truncation at M brings the $\mathbf{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\alpha}}_{\mathsf{init}}$ closer to $\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\alpha}_*$ (See Lemma A20 in Supplementary Materials). Otherwise, the truncation always ensure $\exp(-2M) \le \exp_\tau(-\mathbf{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\alpha}}_{\mathsf{init}}) \le \exp(2M)$. Then, estimating calibrated OR coefficients $\widehat{\boldsymbol{\beta}}_{\mathsf{a}}^{(k_1)}$ is a weighted L_1 penalized regression with bound weights independent of the responses, which we have shown to be consistent under mild assumptions (see Section D1 in Supplementary Materials). Same argument applies to truncation of $\mathbf{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{a,\mathsf{init}}$. Besides numerical stability, we can remove the sparsity condition associated with the initial estimator of the mis-specified model.

4 Theoretical Properties of the SMMAL

We established the $\sqrt{n}$-consistency of $\widehat{\Delta}_{\mathsf{SMMAL}}$ and the honest asymptotic coverage of the confidence intervals with consistent estimation of PS and OR models in Section 4.1. In Section 4.2, we derived the asymptotic distribution of the SMMAL estimator $\widehat{\Delta}_{\mathsf{SMMAL}}$ and the subsequent matching lower bound to show its semi-parametric efficiency in the low-dimensional $\mathbf{W}$ case while using B-spline series estimators for nuisance regression models. For high-dimensional sub-Gaussian $\mathbf{X}$ and $\mathbf{W}$ and sparse nuisance models, we demonstrated in Section 4.3 that $\widehat{\Delta}_{\mathsf{DR}}$ is *adaptively* sparsity/model doubly robust with sparse nuisance models (Rotnitzky et al., 2020; Smucler et al., 2019): sparsity doubly robust when both OR and PS are correctly specified; model doubly robust when one of OR or PS is correctly specified.

4.1 $\sqrt{n}$-inference

We require the following assumptions for nuisance models and the machine-learning estimators. We denote the true propensity score as $\pi_*(\mathbf{X}) = \mathbb{E}(A \mid \mathbf{X})$ and outcome regression as $\mu_*(a, \mathbf{X}) = \mathbb{E}(Y \mid \mathbf{X}, A = a)$. As we do not require consistency of imputation models, we denote $\bar{\Pi}$ and $\bar{m}$ as the potentially biased asymptotic limits of the estimated imputation models.

Assumption 2 *For a fixed constant M, we assume*

(a) (Bounded response) almost surely $\sup_{i=1,\ldots,N} |Y_i| \leq M$*;*

(b) (Causal Inference Positivity) almost surely $\sup_{i=1,\ldots,N} \sup_{a=0,1} 1/\pi_*(a, \mathbf{X}_i) \leq M$*;*

(c) (Bounded estimators) almost surely

$$\sup_{k=1,\ldots,K} \sup_{i \in \mathcal{I}_k} \sup_{a=0,1} \max \left\{ |1/\widehat{\pi}^{(\mathsf{k})}(a, \mathbf{X}_i)|, |\widehat{\mu}^{(\mathsf{k})}(a, \mathbf{X}_i)|, |\widehat{\Pi}^{(\mathsf{k})}(a, \mathbf{W}_i)|, |\widehat{m}^{(\mathsf{k})}(a, \mathbf{W}_i)| \right\} \leq M;$$

(d) (Rate of estimation)

$$\sup_{k=1,\ldots,K} \|\widehat{\pi}^{(\mathsf{k})} - \pi_*\|_2 + \|\widehat{\mu}^{(\mathsf{k})} - \mu_*\|_2 + \|\widehat{\Pi}^{(\mathsf{k})} - \bar{\Pi}\|_2 + \|\widehat{m}^{(\mathsf{k})} - \bar{m}\|_2 \\ + \sqrt{n}\|\widehat{\pi}^{(\mathsf{k})} - \pi_*\|_2 \|\widehat{\mu}^{(\mathsf{k})} - \mu_*\|_2 = o_p(1)$$

for some $\bar{\Pi}$ and $\bar{m}$ satisfying $\sup_{i=1,\ldots,N} \sup_{a=0,1} \max \left\{ \bar{\Pi}(a, \mathbf{W}_i), |\bar{m}(a, \mathbf{W}_i)| \right\} \leq M$*, where for two models $h_1(a, \mathbf{W})$ and $h_2(a, \mathbf{W})$, we define*

$$\|h_1 - h_2\|_2 = \max_{a \in \{0,1\}} \sqrt{\mathbb{E}[\{h_1(a, \mathbf{W}) - h_2(a, \mathbf{W})\}^2]}. \tag{14}$$

Here we use the ℓ_2-norm notation because the mean squared error (MSE) thus defined correspond to the ℓ_2-estimation error for model coefficients under parametric models.

(e) (Stable variance)

$$\begin{aligned} \mathcal{V}_* = \mathrm{Var} \Bigg[& \frac{AY - A\mu_*(1, \mathbf{X})}{\pi_*(1, \mathbf{X})} - \frac{(1 - A)Y - (1 - A)\mu_*(0, \mathbf{X})}{\pi_*(0, \mathbf{X})} \\ & - \frac{\{\bar{\Pi}(1, \mathbf{W})\bar{m}(1, \mathbf{W}) - \bar{\Pi}(1, \mathbf{W})\mu_*(1, \mathbf{X})\}}{\pi_*(1, \mathbf{X})} \\ & + \frac{\{\bar{\Pi}(0, \mathbf{W})\bar{m}(0, \mathbf{W}) - \bar{\Pi}(0, \mathbf{W})\mu_*(0, \mathbf{W})\}}{\pi_*(0, \mathbf{X})} \Bigg] \in [1/M, M]. \end{aligned}$$

We established the validity and asymptotic distribution of $\widehat{\Delta}_{\mathsf{SMMAL}}$ in the following theorem.

Theorem 2 *Under Assumption 2,*

$$\sqrt{n/\widehat{\mathcal{V}}_{\mathsf{SMMAL}}}(\widehat{\Delta}_{\mathsf{SMMAL}} - \Delta_*) \rightsquigarrow N(0, 1),$$

where "$\rightsquigarrow$" denotes convergence in distribution.

Assumption 2a guarantees the boundedness of all nuisance models. When Y is binary, the models π_*, μ_*, Π_* and m_* are all bounded by one. Assumption 2b is equivalent to the standard (causal inference) positivity condition $\pi_*(1, \mathbf{X}_i) \in [1/M, 1-1/M]$ as in Assumption 1b. Assumption 2c can be guaranteed by truncation of nuisance model estimators at M, which would not compromise the estimation accuracy under Assumptions 2a and 2b. Assumption 2e ensures the proper scaling of the asymptotic variance of $\widehat{\Delta}_{\mathsf{SMMAL}}$. As noted following (7), the term with the R/ρ_{N} factor from labeled data in ϕ_{SSL} dominates its variance if $\rho_{\mathsf{N}} \to 0$. The rate condition for the PS and OR models in Assumption 2d matches those for the double machine-learning estimator proposed in Chernozhukov et al. (2018) if applied to the complete data subset of size n. Under MCAR by design, the missing data mechanism is known a priori, which we utilized to accommodate the mis-specified imputation models estimated at an arbitrarily slow rate.

Remark 3 *Compared to existing work on semi-supervised estimation of ATE (Cheng et al., 2021; Kallus and Mao, 2024) approximating the $\rho_{\mathsf{N}} \to 0$ setting by the $\rho_{\mathsf{N}} = 0$ setting, our SMMAL incorporates additionally the uncertainty from large yet finite unlabeled data through (5)-(7). As the result, the inference from SMMAL has two methodological advantages. First, by harmonizing the $n \ll N$ and $n \asymp N$ settings, users may use the same SMMAL procedure without choosing from two setting-specific approaches (Kallus and Mao, 2024). Especially, it seems implausible to decide the asymptotic limit of ρ_{N} by a single realization of the data. Second, the uncertainty of $\widehat{\Delta}_{\mathsf{SMMAL}}$ consists of the uncertainty from labeled data and the uncertainty from large but finite unlabeled data,*

$$\mathcal{V}_{\mathsf{SMMAL}} = \underbrace{\mathrm{Var}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]}_{\mathcal{V}_L \text{ from labeled set}} + \underbrace{\rho_{\mathsf{N}} \mathrm{Var}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}}_{\mathcal{V}_U \text{ from unlabeled set}}.$$

Unlike existing work (Cheng et al., 2021; Kallus and Mao, 2024) that only considered $\mathcal{V}_L$ from labeled set, our SMMAL variance estimation captures both $\mathcal{V}_L$ and $\mathcal{V}_U$ by involving estimated influence functions $\widehat{\mathcal{V}}_{ik}$ for all observations so that SMMAL is expected to have less issues in underestimation of uncertainty particularly from unlabeled data with a moderately small ρ_{N} in practice.

4.2 Semi-parametric efficiency with low-dimensional confounder

We next formally establish the semi-parametric efficiency lower bound under the double missing SSL setting. Consider the non-parametric model for $(\mathbf{W}, RA, RY, R)$

$$\mathcal{S}_{\mathsf{SSL}} = \Big\{ d\mathrm{P}_f(\mathbf{w}, a, y, r) = \{\rho_{\mathsf{N}} f(\mathbf{w}, a, y)\}^r \{(1-\rho_{\mathsf{N}}) f_{\mathbf{W}}(\mathbf{w})\}^{(1-r)} d\nu_{\mathsf{SSL}}(\mathbf{w}, a, y, r) :$$

$$f \text{ is density over } \mathcal{W} \otimes \{0,1\} \otimes \mathcal{Y}, \text{ and } f_{\mathbf{W}}(\mathbf{w}) = \sum_{a \in \{0,1\}} \int_{y \in \mathcal{Y}} f(\mathbf{w}, a, y) d\nu_y(y) \Bigg\}$$

for some measures ν_y over $\mathcal{Y}$, ν_w over $\mathcal{W}$ and

$$\nu_{\mathsf{SSL}}(\mathbf{w}, a, y, r) = (\nu_w \times \delta_{\{0,1\}} \times \nu_y)(\mathbf{w}, a, y) \times \delta_1(r) + \nu_w(w) \times \delta_0(r)$$

where $\delta_{\mathcal{A}}$ is the counting/Dirac measure over the set $\mathcal{A}$. Elements in $\mathcal{S}_{\mathsf{SSL}}$ can be indexed by the density f, and we denote the true density as f_* and the true model P_{f_*}.

Remark 4 *In existing work on ATE (Robins et al., 1994; Kallus and Mao, 2024), the model $A \mid \mathbf{X}$ provides no information on the $Y \mid A, \mathbf{X}$, and thus would not be included in the nuisance tangent space. In our setting, however, the surrogates $\mathbf{S}$ induced a correlation between subspaces corresponding to $A \mid \mathbf{X}$ and $Y \mid A, \mathbf{X}$ in the nuisance tangent space, which indicates that $A \mid \mathbf{X}$ provides information on the $Y \mid A, \mathbf{X}$ through the unlabelled data. As the result, the geometry of the model tangent space is more complex, and the projection can no longer be obtained through simple conditional expectation. See Section E2 of the Supplementary Materials for details.*

We denote the total variation norm as $\|\cdot\|_{\mathsf{TV}}$. In the following theorem, we establish the semi-parametric efficiency lower bound for Δ under $\mathcal{S}_{\mathsf{SSL}}$ in the form of a local minimax theorem obtained in the spirit of Begun et al. (1983).

Theorem 5 *Under Assumptions 2a, 2c and 2e, we have*

$$\liminf_{c\to\infty} \liminf_{N\to\infty} \inf_{\widehat{\Delta}} \sup_{\|f-f_*\|_{\mathsf{TV}}\le c/\sqrt{\rho_{\mathsf{N}} N}} \frac{\int N(\widehat{\Delta}-\Delta_*)^2 d\prod_{i=1}^N \mathrm{P}_f(\mathbf{w}_i, a_i, y_i, r_i)}{\mathrm{Var}\{\phi_{\mathsf{SSL}}(RY, RA, \mathbf{W}, R)\}} \ge 1.$$

Remark 6 *Theorem 5 offers one example that the semi-parametric efficiency bound (SEB) derived under the classical missing data setting can be generalized to the double missing SSL setting with $\rho_{\mathsf{N}} \to 0$ while $\rho_{\mathsf{N}} N \to \infty$. Later in Section 7, we present Theorem 13 for general SSL setting (including specifically the double missing SSL setting). Previous attempts to formalize semi-parametric efficiency in a the SSL settings have assumed that the entire distribution of $\mathbf{W}$ is known, i.e. $N = \infty$ and $\rho_{\mathsf{N}} = 0$. Under the simplified SSL setting with $N = \infty$, the SEB can be derived by straightforward applications of standard results in classical semiparametric literature – see e.g. van der Vaart (1998). Indeed, another possible consideration for choosing this simplified formulation version is the ambiguity of defining regular estimators without (missing data) positivity assumption and thereby formalizing efficiency through the calibration of the best regular estimator. We bypassed this conceptual difficulty by providing the alternative characterization based on local asymptotic minimax theory – which may operate on all possible estimators instead of restricting to the class of regular procedures.*

Utilizing the correlation structure induced by the projection

$$\mathrm{Cov}(\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}, R/\rho_{\mathsf{N}}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]) = 0,$$
$$\mathrm{Cov}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}), \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}] = \mathrm{Var}[\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}],$$

we obtain the limiting lower bound in Theorem 5 when $\rho_{\mathsf{N}} \to 0$ that matches the asymptotic variance of the labeled data component in ϕ_{SSL} ,

$$
\begin{aligned}
&\lim_{\rho_{\mathsf{N}} \to 0} \rho_{\mathsf{N}} \operatorname{Var}\{\phi_{\mathsf{SSL}}(RY, RA, \mathbf{W}, R)\} \\
&= \lim_{\rho_{\mathsf{N}} \to 0} \rho_{\mathsf{N}} \operatorname{Var}(\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\} + R/\rho_{\mathsf{N}}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]) \\
&= \lim_{\rho_{\mathsf{N}} \to 0} \underbrace{\rho_{\mathsf{N}} \operatorname{Var}[\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]}_{\to 0} + \rho_{\mathsf{N}} \operatorname{Var}(R/\rho_{\mathsf{N}}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]) \\
&\quad + 2\rho_{\mathsf{N}} \underbrace{\operatorname{Cov}(\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}, R/\rho_{\mathsf{N}}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}])}_{=0} \\
&= \operatorname{Var}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}] \\
&= \operatorname{Var}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X})\} + \operatorname{Var}[\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}] \\
&\quad - 2 \underbrace{\operatorname{Cov}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}), \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]}_{= \operatorname{Var}[\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}]} \\
&= \operatorname{Var}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X})\} - \operatorname{Var}[\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}].
\end{aligned}
$$

From the representation above, we showed that the efficiency gain from the unlabeled data with surrogates is given by the variance of the ϕ_{cmp} explained by the surrogates and confounders. The efficiency gain based on semi-parametric efficiency theory typically requires consistent estimation of nuisance models. Under mis-specified imputation models, there is no general guarantee on efficiency gain. In Discussion (Section 8), we offered efficient linear combination as the backup plan when quality of estimated nuisance models is in doubt.

The key idea of the proof is to construct the two-dimensional least favorable perturbation in an asymmetric neighborhood with different size in two directions. The first direction is proportional to $\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}$ and of size $\asymp 1/\sqrt{N\rho_{\mathsf{N}}} \asymp 1/\sqrt{n}$, which reflects the level of the information on Δ_* from the labels and should naturally scale with the number of expected labels. The second direction is proportional to $\mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}$ and of size $\asymp 1/\sqrt{N}$, which reflects the level of the information on Δ_* from the unlabelled data and should scale with the total sample size. The design of the different scales ensured the tightness of log-likelihood ratio between the perturbed and the true models, which would otherwise be degenerating or diverging.

We next show that the lower bound is attained under low-dimensional smoothness class models for the nuisance functions and can be operationalized by feeding B-spline regressions to $\widehat{\Delta}_{\mathsf{SMMAL}}$. Suppose the confounders and surrogates are bounded continuous variables of fixed dimension, $\mathbf{W} \in [-M, M]^{p+q}, \quad p + q < d \asymp 1$. We measure the smoothness of the models by $\mathcal{H}(f(\cdot))$ the Hölder class defined in Definition A22, Section E of the Supplementary Materials.

Assumption 3 *For a fixed constant M, we assume*

(a) (Bounded density) the density functions for $\mathbf{X}$ and $\mathbf{W}$, $f_{\mathbf{X}}(\mathbf{x})$ and $f_{\mathbf{W}}(\mathbf{w})$, are bounded and bounded from zero,

$$
f_{\mathbf{X}}(\mathbf{x}) \in [1/M, M], \forall \mathbf{x} \in [-M, M]^p, \; f_{\mathbf{W}}(\mathbf{w}) \in [1/M, M], \forall \mathbf{w} \in [-M, M]^{p+q};
$$

(b) (Smooth models) the smoothness of the nuisance models observe

$$\frac{1}{1+\mathcal{H}(\pi_*(a,\cdot))/p}+\frac{1}{1+\mathcal{H}(\mu_*(a,\cdot))/p}<1,\ \mathcal{H}(\Pi_*(a,\cdot))>0,\ \mathcal{H}(m_*(a,\cdot))>0,$$

for $a = 0, 1$.

Corollary 7 *Under Assumptions 2a, 2b, 2e and 3, we may choose B-spline regressions with order*

$$\kappa \geq \max\{\mathcal{H}(\pi_*(a,\cdot)), \mathcal{H}(\mu_*(a,\cdot)), \mathcal{H}(\Pi_*(a,\cdot)), \mathcal{H}(m_*(a,\cdot)) : a = 0, 1\} - 1,$$

degrees

$$\begin{aligned} &1/\pi(a,\cdot) : n^{\frac{1}{1+\mathcal{H}(\pi_*(a,\cdot))/p}}, \quad \mu(a,\cdot) : n^{\frac{1}{1+\mathcal{H}(\mu_*(a,\cdot))/p}}, \\ &\Pi(a,\cdot) : n^{\frac{1}{1+\mathcal{H}(\Pi_*(a,\cdot))/(p+q)}}, m(a,\cdot) : n^{\frac{1}{1+\mathcal{H}(m_*(a,\cdot))/(p+q)}} \end{aligned}$$

and truncation at M for $\widehat{\Delta}_{\mathsf{SMMAL}}$ to achieve

$$\sqrt{n}(\widehat{\Delta}_{\mathsf{SMMAL}} - \Delta_*)/\sqrt{\rho_{\mathsf{N}} \operatorname{Var}\{\phi_{\mathsf{SSL}}(RY, RA, \mathbf{W}, R)\}} \rightsquigarrow N(0, 1).$$

Corollary 7 is special case of Theorem 2 under smooth models estimated by standard nonparametric estimation. By Corollary 7, the asymptotic MSE of $\widehat{\Delta}_{\mathsf{SMMAL}}$ is $\rho_{\mathsf{N}} \operatorname{Var}\{\phi_{\mathsf{SSL}}\}/n \asymp \operatorname{Var}\{\phi_{\mathsf{SSL}}\}/N$, matching the lower bound established in Theorem 5. Therefore, we have justified the semi-parametric efficiency of $\widehat{\Delta}_{\mathsf{SMMAL}}$. At the same time, the lower bound in Theorem 5 is the sharp semi-parametric efficiency bound for Δ_* under double missing SSL setting $\mathcal{S}_{\mathsf{SSL}}$.

4.3 Doubly robustness with high-dimensional confounder

To describe the sparsity/model double robustness of $\widehat{\Delta}_{\mathsf{DR}}$, we define the asymptotic limits for Lasso estimators in (9)-(12) under potentially mis-specified models.

$$\begin{aligned} \bar{\boldsymbol{\xi}} &= \underset{\boldsymbol{\xi}\in\mathbb{R}^{p+q+1}}{\operatorname{argmin}}\ \mathbb{E}\{\ell(A_i, \boldsymbol{\xi}^{\mathsf{T}}\mathbf{W}_i)\},\ \bar{\boldsymbol{\zeta}}_{\mathsf{a}} = \underset{\boldsymbol{\zeta}\in\mathbb{R}^{p+q+1}}{\operatorname{argmin}}\ \mathbb{E}\{\mathrm{I}(A_i = a)\ell(Y_i, \boldsymbol{\zeta}^{\mathsf{T}}\mathbf{W}_i)\}, \\ \bar{\boldsymbol{\alpha}}_{\mathsf{init}} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\,\mathbb{E}\{\ell(A_i, \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{X}_i)\},\ \bar{\boldsymbol{\beta}}_{\mathsf{a,init}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\,\mathbb{E}\{\mathrm{I}(A_i = a)\ell(Y_i, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\}, \\ \bar{\boldsymbol{\alpha}}_a &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\,\mathbb{E}\left[\dot{g}\left(\mathbf{X}_i^{\mathsf{T}}\bar{\boldsymbol{\beta}}_{\mathsf{a,init}}\right)\{(a - A_i)\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{X}_i + I(A_i = a)e^{(-1)^a\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{X}_i}\}\right], \\ \bar{\boldsymbol{\beta}}_{\mathsf{a}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\,\mathbb{E}\left[\exp\{(-1)^a\mathbf{X}_i^{\mathsf{T}}\bar{\boldsymbol{\alpha}}_{\mathsf{init}}\} I(A_i = a)\ell(Y_i, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\right], \end{aligned} \tag{15}$$

We use $\|\cdot\|_0$ to denote the sparsity of a vector and $\|\cdot\|_{\psi_2}$ denote the sub-Gaussian norm for random variables or vectors. The sparsities of coefficients for OR $\|\boldsymbol{\beta}_{\mathsf{a}}\|_0$ and PS $\|\boldsymbol{\alpha}\|_0$ models reflect the numbers of true determinants for the treatment and outcomes, including the true confounders that must be adjusted for. The detailed definition is given in Definition A23, Section E of the Supplementary Materials.

Assumption 4 *For constant M independent of dimensions n, N, p, q,*

(a) *(Sub-Gaussian and bounded covariates) the vector of confounders and surrogates is sub-Gaussian,* $\sup_{\|\mathbf{v}\|_2=1}\|\mathbf{v}^\mathsf{T}\mathbf{W}\|_{\psi_2} \leq M$*, and coordinate-wisely bounded* $\|\mathbf{W}\|_\infty \leq M$ *almost surely;*

(b) *(Identifiability) the variance of* $\mathbf{W}$ *is invertible* $\inf_{\|\mathbf{v}\|_2=1}\mathbf{v}^\mathsf{T}\operatorname{Var}(\mathbf{W})\mathbf{v} \geq 1/M$*;*

(c) *(Causal Inference positivity) the true propensity scores and the asymptotic predictions of all models are bounded away from zero and one, almost surely,*

$$\pi_*(a,\mathbf{X}) \in [1/M, 1-1/M],\ \max\left\{|\bar{\boldsymbol{\xi}}^\mathsf{T}\mathbf{W}|, |\bar{\boldsymbol{\zeta}}_\mathsf{a}^\mathsf{T}\mathbf{W}|, |\bar{\boldsymbol{\alpha}}_\mathsf{a}^\mathsf{T}\mathbf{X}|, |\bar{\boldsymbol{\beta}}_\mathsf{a}^\mathsf{T}\mathbf{X}| : a = 0,1\right\} \leq M;$$

(d) *and one of the following:*

(i) *(PS correct) the propensity model is correct,* $\mathbb{E}(A \mid \mathbf{X}) = g(\boldsymbol{\alpha}_*^\mathsf{T}\mathbf{X})$ *and the dimensions satisfy*

$$\begin{aligned}&\frac{\left(\|\bar{\boldsymbol{\beta}}_\mathsf{1}\|_0 + \|\bar{\boldsymbol{\beta}}_\mathsf{0}\|_0\right)\log(p) + \left(\|\bar{\boldsymbol{\xi}}\|_0 + \|\bar{\boldsymbol{\zeta}}_\mathsf{1}\|_0 + \|\bar{\boldsymbol{\zeta}}_\mathsf{0}\|_0\right)\log(p+q)}{n}\\ &+ \|\boldsymbol{\alpha}_*\|_0\left(\|\boldsymbol{\alpha}_*\|_0 + \|\bar{\boldsymbol{\beta}}_\mathsf{1}\|_0 + \|\bar{\boldsymbol{\beta}}_\mathsf{0}\|_0\right)\log(p)^2/n = o_p(1); \end{aligned} \tag{16}$$

(ii) *(OR correct) the OR model is correct,* $\mathbb{E}(Y \mid A = a, \mathbf{X}) = g(\boldsymbol{\beta}_{*,\mathsf{a}}^\mathsf{T}\mathbf{X})$ *and the dimensions satisfy*

$$\begin{aligned}&\frac{\left(\|\bar{\boldsymbol{\alpha}}_\mathsf{1}\|_0 + \|\bar{\boldsymbol{\alpha}}_\mathsf{0}\|_0\right)\log(p) + \left(\|\bar{\boldsymbol{\xi}}\|_0 + \|\bar{\boldsymbol{\zeta}}_\mathsf{1}\|_0 + \|\bar{\boldsymbol{\zeta}}_\mathsf{0}\|_0\right)\log(p+q)}{n}\\ &+ \sum_{a=0,1}\|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0\left(\|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0 + \|\bar{\boldsymbol{\alpha}}_\mathsf{a}\|_0\right)\log(p)^2/n = o_p(1); \end{aligned} \tag{17}$$

(iii) *(both correct) both models are correct,* $\mathbb{E}(A \mid \mathbf{X}) = g(\boldsymbol{\alpha}_*^\mathsf{T}\mathbf{X})$ *and* $\mathbb{E}(Y \mid A = a, \mathbf{X}) = g(\boldsymbol{\beta}_{*,\mathsf{a}}^\mathsf{T}\mathbf{X})$ *and the dimensions satisfy*

$$\begin{aligned}&\frac{\left(\|\boldsymbol{\alpha}_*\|_0 + \|\boldsymbol{\beta}_{*,\mathsf{1}}\|_0 + \|\boldsymbol{\beta}_{*,\mathsf{0}}\|_0\right)\log(p) + \left(\|\bar{\boldsymbol{\xi}}\|_0 + \|\bar{\boldsymbol{\zeta}}_\mathsf{1}\|_0 + \|\bar{\boldsymbol{\zeta}}_\mathsf{0}\|_0\right)\log(p+q)}{n}\\ &+ \|\boldsymbol{\alpha}_*\|_0\left(\|\boldsymbol{\beta}_{*,\mathsf{1}}\|_0 + \|\boldsymbol{\beta}_{*,\mathsf{0}}\|_0\right)\log(p)^2/n = o_p(1); \end{aligned} \tag{18}$$

Theorem 8 *Under Assumption 4,* $\widehat{\Delta}_\mathsf{DR}$ *converges in distribution to a normal random variable at* $\sqrt{n}$*-rate,*

$$\sqrt{n/\widehat{\mathcal{V}}_\mathsf{DR}}(\widehat{\Delta}_\mathsf{DR} - \Delta_*) \rightsquigarrow N(0,1),$$

where "$\rightsquigarrow$*" denotes convergence in the distribution.*

Besides the double robustness toward PS and OR, $\widehat{\Delta}_\mathsf{DR}$ is additionally robust to the imputation models. Similar to the general Theorem 2, we utilized the known missing data mechanism under MCAR by design to allow model mis-specifications on the imputation.

Remark 9 *Regarding the PS model and OR model estimated over the labeled data of size n, our $\widehat{\Delta}_{\mathsf{DR}}$ is both rate doubly robust Rotnitzky et al. (2020) and model doubly robust (Smucler et al., 2019). When both models are correct, the dimension condition* (18) *for the PS model and OR model in Assumption 4d-iii satisfies the condition for rate doubly robust, i.e. each sparsity obeying $\|\boldsymbol{\alpha}_*\|_0 \ll n/\log(p), \|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0 \ll n/\log(p)$ and their product satisfying $\|\boldsymbol{\alpha}_*\|_0\|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0 \ll n/\log(p)^2$. In the case of only one model is correct, our $\widehat{\Delta}_{\mathsf{DR}}$ can still provide $\sqrt{n}$-inference, thus being model doubly robust. By the truncation τ in* (12)*, we are able to completely remove the sparsity requirement of the mis-specified initial model under the (causal inference) positivity condition of Assumption 4c. The general framework of Smucler et al. (2019) would require all models in* (15) *being sparse.*

Remark 10 *As the correct model specification is only required for OR or PS in Assumption 4d, the validity of Theorem 8 does not rely on the consistency of imputation models based on the MCAR missing data mechanism by design. Therefore, the choice on imputation methods* (9) *can be flexible. If preliminary evidence suggests that certain element in* $\mathbf{S}$ *contains the most information such as* $\mathbf{S}_a$ *for A and* $\mathbf{S}_y$ *for Y, we can remove penalty for the associated coefficients or simply run the low-dimensional regressions* $A \sim \mathbf{S}_a$ *and* $Y \sim \mathbf{S}_y$*.*

Remark 11 *We presented the theory according to the exact sparsity in Assumption 4d-iii for two considerations. First, the exact sparsity has a clear interpretation that classifies the covariates into relevant signals and irrelevant noises, about which domain experts may have a preliminary evaluation in applications. Second, the exact sparsity facilitates direct comparison with many related literatures have used exact sparsity to measure the local efficiency or robustness of their proposed methods (Farrell, 2015; Tan, 2020; Smucler et al., 2019; Zhang et al., 2023). The exact sparsity in Assumption 4d-iii can be substituted by other conditions that produce the appropriate estimation rate in the more general Assumption 2d. For example, estimation rates of L_1 penalized high-dimensional generalized linear models have been established for approximately sparse models (Negahban et al., 2012; Smucler et al., 2019).*

5 Simulation

We conducted extensive simulation studies to evaluate the finite sample performance of the SMMAL methods. Throughout the simulations, we set the total sample size $N = 10000$, the number of labels $n = 500$, the number of repeats as 1000, $q = 2$ with one surrogate S_A for A and another S_Y for Y. We focused on the situation that Y is also binary. Let Φ be cumulative distribution function for standard normal distribution. The surrogates for binary A and Y were generated from mixture Beta distribution of the form:

$$\begin{aligned}
&S_A = AS_{A,1} + (1-A)S_{A,0},\ S_Y = YS_{Y,1} + (1-Y)S_{Y,0},\\
\text{Low-dimensional model: }&S_{A,1} \sim \mathit{Beta}(\alpha_A + X, 1),\ S_{A,0} \sim \mathit{Beta}(1, \alpha_A + X),\\
&S_{Y,1} \sim \mathit{Beta}(\alpha_Y + X, 1),\ S_{Y,0} \sim \mathit{Beta}(1, \alpha_Y + X);\\
\text{High-dimensional model: }&S_{A,1} \sim \mathit{Beta}(\alpha_A + \Phi(X_1), 1),\ S_{A,0} \sim \mathit{Beta}(1, \alpha_A + \Phi(X_1)),\\
&S_{Y,1} \sim \mathit{Beta}(\alpha_Y + \Phi(X_1), 1),\ S_{Y,0} \sim \mathit{Beta}(1, \alpha_Y + \Phi(X_1)).
\end{aligned}$$

Table 1: List of parameters used in the mixture Beta distribution for the surrogates.

Setting	OK	Reasonable	Good	Great	Perfect
AUC	0.80	0.90	0.95	0.99	0.999
Low-dimensional smooth model					
α_A	1.39	1.99	2.54	3.86	5.49
α_Y	1.39	1.96	2.57	3.80	5.70
High-dimensional logistic regression					
α_A	1.36	1.99	2.54	3.80	5.64
α_Y	1.33	1.96	2.51	3.89	5.55
High-dimensional regression: mis-specified PS					
α_A	1.36	1.96	2.54	3.80	5.55
α_Y	1.39	1.93	2.54	3.74	5.52
High-dimensional regression: mis-specified OR					
α_A	1.36	1.96	2.54	3.80	5.55
α_Y	1.39	1.93	2.54	3.74	5.52

The mixture Beta distribution mimicked the outputs from phenotyping algorithms, which typically take value between zero and one (Liao et al., 2019). We considered a list of values for α_A and α_Y (Table 1), corresponding to different level of prediction accuracy measured by area-under-curve (AUC) of the receiver operating characteristic (ROC). Five values were considered for α_A and α_Y, creating 25 two-way combinations for each simulation setting.

We considered two scenarios for generating the data, the low-dimensional smooth model and high-dimensional logistic regression.

Low-dimensional smooth model We generated the one dimensional $X \in \mathbb{R}$ from Uniform(0,1) and set the PS and OR to be the following smooth models (Figure 2):

$$\pi_*(1, X) = \mu_*(1, X) = 1 - 1.2/(3 - X^2),\ \mu_*(0, X) = 1 - 1.2/\{3 - (1 - X)^2\}.$$

We used tensor product first order B-spline (piece-wise linear splines) regression to estimate the nuisance models. The splines were constructed from *bs* function of the *splines* R package. The degrees were selected by 10 fold cross-validation among integers less than $\sqrt{n} \approx 22$ according to the out-of-fold entropy. Using the cross-fitted nuisance models from B-spline regression with $K = 10$, we obtained point and interval estimates for the ATE based on $\widehat{\Delta}_{\textsf{SMMAL}}$ and $\widehat{\mathcal{V}}_{\textsf{SMMAL}}$. As the benchmark, we also estimated the ATE using the labeled data only by the double machine learning method (Chernozhukov et al., 2018).

High-dimensional logistic regression We generated the high-dimensional $\mathbf{X} \in \mathbb{R}^p$ with $p = 500$ from the multivariate Gaussian distribution with auto-regressive correlation structure:

$$U_1, \dots, U_p \overset{i.i.d.}{\sim} N(0, 1),\ X_1 = U_1,\ X_j = 0.5X_{j-1} + \sqrt{0.75}U_j.$$

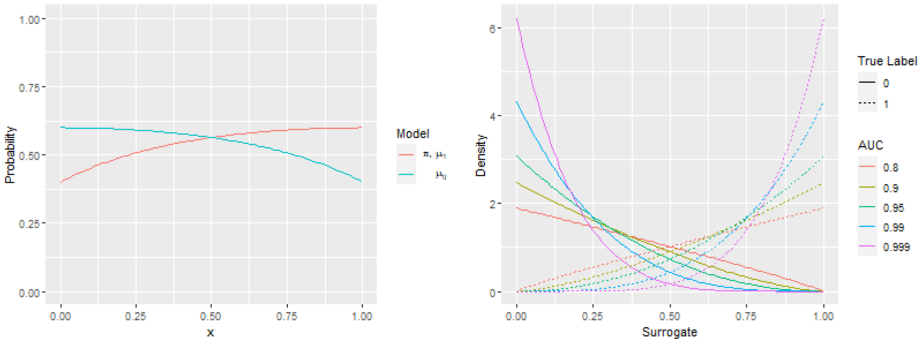


Figure 2: Visualized simulation settings. Left-the models for PS and OR under the low-dimensional setting. Right-the mixture Beta distribution for surrogates at different level of prediction accuracy (AUC 0.8, 0.9, 0.95, 0.99, 0.999) at the median covariate ($X = 0.5$ under low-dimensional smooth model and $X_1 = 0$ under high-dimensional logistic regression).

We generated A and Y from the high-dimensional logistic regression models

$$\begin{aligned}
\text{PS Linear :}&\pi_*(1,\mathbf{X}) = g(0.5X_1 + 0.25X_2 + 0.125X_3);\\
\text{PS Interaction: }&\pi_*(1,\mathbf{X}) = g\{(0.5X_1 + 0.25X_2 + 0.125X_3)(1 + 0.0625X_1 + 0.125X_2 - 0.5X_3)\};\\
\text{OR Linear: }&\mu_*(1,\mathbf{X}) = g(0.1 + 0.25X_1 + 0.125X_2 + 0.0625X_3),\\
&\mu_*(0,\mathbf{X}) = g(-0.1 - 0.25X_1 - 0.125X_2 - 0.0625X_3);\\
\text{OR Interaction: }&\mu_*(1,\mathbf{X}) = g\{(0.1 + 0.25X_1 + 0.125X_2 + 0.0625X_3)\\
&\qquad\qquad \times (1 + 0.0625X_1 + 0.125X_2 - 0.5X_3)\},\\
&\mu_*(0,\mathbf{X}) = g\{(-0.1 - 0.25X_1 - 0.125X_2 - 0.0625X_3)\\
&\qquad\qquad \times (1 + 0.0625X_1 + 0.125X_2 - 0.5X_3)\}.
\end{aligned}$$

As signal strength is known to impact variable selection in theory and practice (Fan and Peng, 2004; Fan and Lv, 2010), we set up the coefficients in the models to reflect different level of signal strength: 0.5-strong, 0.25-moderately strong, 0.125-moderately weak, 0.0625-weak. For PS/OR models with second order interactions, we still fitted high-dimensional logistic regression without interactions, creating the mis-specification scenarios. We considered 3 combinations corresponding to the three settings of Assumption 4d: correct models (PS Linear + OR Linear); mis-specified PS (PS Interaction + OR Linear); mis-specified OR (PS Linear + OR Interaction). We set the number of the folds as 10 and fitted the imputations (9) and initial estimators (10) using *glmnet* from R-package *glmnet*. We fitted the calibrated estimators (12) using *rcal* from from R-package *rcal*. The penalty parameters were selected by 10-fold cross validation with out-of-fold entropy. Using the cross-fitted nuisance models, we estimated the ATE using $\widehat{\Delta}_{\mathsf{DR}}$ and construct the 95% confidence interval based on the variance estimator $\widehat{\mathcal{V}}_{\mathsf{DR}}$. As the benchmark, we also estimated the ATE by the model doubly robust estimation (Smucler et al., 2019) using (1) the labeled data alone; (2)

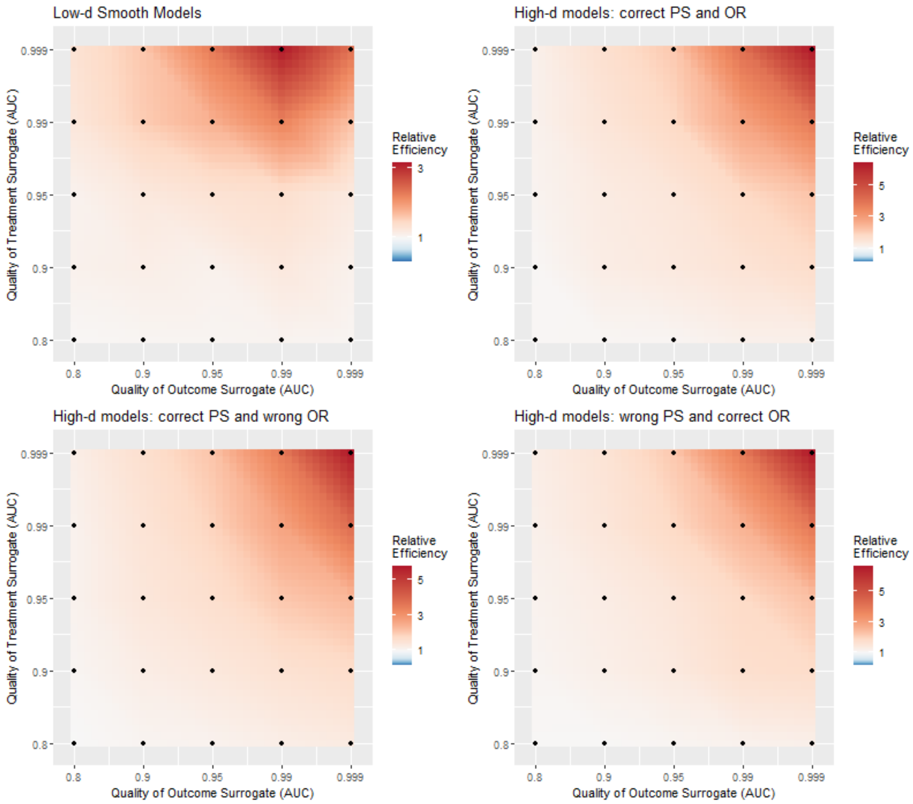


Figure 3: Heat map for relative efficiency of the SMMAL compared to the benchmark supervised learning in all four simulation settings. Deeper red indicates larger advantage of the semi-supervised estimation. We set relative efficiency one as white in all plots, but the scale varies between low-dimensional setting and high-dimensional settings.

the dichotomized surrogates defined by

$$\widetilde{Y}_i = \mathrm{I}\left(S_{Y,i} \geq 1 - n^{-1}\sum_{i=1}^{N} R_i Y_i\right), \ \widetilde{A}_i = \mathrm{I}\left(S_{A,i} \geq 1 - n^{-1}\sum_{i=1}^{N} R_i A_i\right).$$

We refer to the two benchmarks as supervised learning (SL) and unsupervised learning (UL).

Results Results generally followed a consistent pattern across low-d and high-d settings. Comparison between settings, however, is not meaningful due to the completely different data generating processes. In Figure 3, we visualized the relative efficiency of our semi-supervised $\widehat{\Delta}_{\mathsf{SMMAL}}$, $\widehat{\Delta}_{\mathsf{DR}}$ compared to their supervised benchmarks. In general, our semi-supervised approaches gained efficiency from the unlabeled data whose magnitude was increasing with the minimal prediction accuracy of the two surrogates. With good imputation (AUC .95) from both surrogates, the relative efficiency was about 1.32-1.64 across

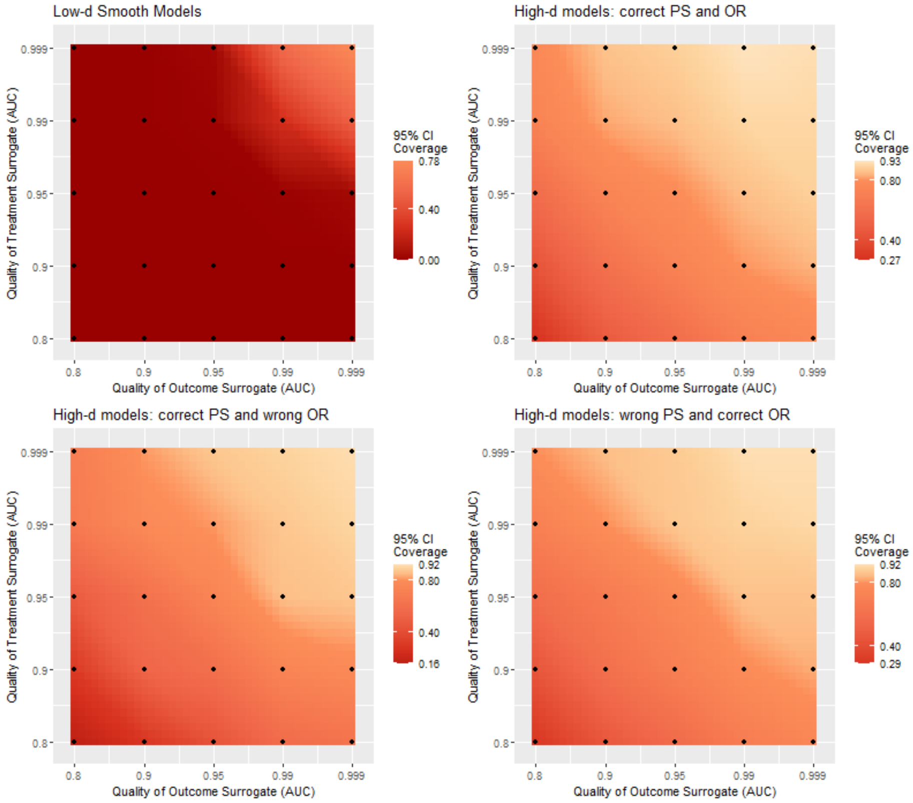


Figure 4: Heat map for coverage of 95 % confidence intervals by unsupervised learning. White marks 0.95 coverage rate. Orange marks 0.8 coverage rate. Deeper red indicates poorer coverage rate by unsupervised learning.

all settings. With great imputation (AUC .99) from both surrogates, the relative efficiency was about 2.23-2.89 across all settings. The result quantified the benefit from improving the quality of surrogates in terms of relative increase in labels. Since the algorithms to curate surrogates are often portable to other studies sharing the variables, effort put into high-quality labels is more cost-effective compared to the brutal expansion in labeling. The detailed simulation results containing the bias, standard deviation, average standard error, coverage of 95% confidence interval for our semi-supervised $\widehat{\Delta}_{\mathsf{SMMAL}}$, $\widehat{\Delta}_{\mathsf{DR}}$ along with those for the supervised benchmarks were presented in Tables A4-A7 in Section A of the Supplementary Materials. Our semi-supervised $\widehat{\Delta}_{\mathsf{SMMAL}}$, $\widehat{\Delta}_{\mathsf{DR}}$ achieved reasonably honest inference with coverage of 95% confidence interval close to the nominal level. In Figure 4, we visualized the coverage of 95 % confidence intervals by unsupervised learning. Using the dichotomized surrogates as if they were the true treatment and outcome led to under coverage of the confidence intervals even for nearly perfect surrogates, and the under coverage exacerbated with poorer surrogates. The detailed summaries on the bias, standard deviation and coverage of 95% confidence interval for the unsupervised benchmark were presented in Table A8 in Section A of the Supplementary Materials.

6 Real-world evidence on targeted cancer therapy

We applied the proposed SMMAL method to EHR data from Mass General Brigham healthcare to generate real-world evidence (RWE) on treatment effect of targeted therapy for metastatic colorectal cancer in comparison with conventional chemotherapy. Over the past two decades, a total of 9 targeted therapies have been approved for the treatment of colorectal cancer (Xie et al., 2020), the 4th most prevalent and lethal cancer (U.S. Cancer Statistics Working Group, 2022). While the targeted therapies have been reported as advantageous compared to conventional chemotherapy in clinical trials within specific trial populations, their effectiveness in real-world patient population has not been fully established. With increasing availability of EHR data, it is now plausible to generate RWE on targeted cancer therapy with respect to their efficacy in improving progression free survival via causal modeling treating EHR data as an observational cohort. Unfortunately, such a modeling task is highly challenging with EHR data due to the lack of readily available precise information on both treatments patient received and progression free survival. To overcome this challenge, we manually annotated treatment-response information for 100 randomly selected patients. We derived several potential surrogates for both S_A and S_Y from codified and narrative EHR data, which have varying degree of accuracy as shown in Table 2. Our goal was to leverage both the labeled observations on Y and S as well as the larger set of unlabeled EHR data to infer about ATE for targeted therapy based on SMMAL.

The full study cohort consisted of $N = 4147$ colorectal cancer patients who have available cancer stage information extracted via a natural language process tool (Yuan et al., 2021) and received chemotherapy and/or targeted therapy. We grouped therapies into chemotherapy alone and targeted therapy which includes those treated with any of the 9 treatments: Bevacizumab, Cetuximab, Ipilimumab, Regorafenib, Pembrolizumab, Nivolumab, and Tipiracil. We set the outcome as 1-year progression free survival, a binary outcome defined as: 1 – exit in terminal condition (death/terminal care) or development

Table 2: **Accuracy of extracted EHR feature counts** for targeted therapy and 1-year progression (defined as new metastasis site) free survival from EHR valided over 100 patients reviewed by abstractor. False positive rates (FPR) and false negative rates (FNR) were calculated by the dichotomized extractions: Benchmark features – count > 0; **Engineered features** – classification by the quantiles matching prevalence in gold-standard labels. Area under reception operating curve (AUC) were calculated using count/score as predictor (death encoded as a very large value 1000). Straightforward rule based extraction (indicated by *) failed to capture treatment and response. Two surrogates in **bold font** were chosen for SMMAL for their **reasonably good AUC**.

Surrogate	FPR	FNR	AUC
Targeted Therapy			
Medication Code	0.44	0.17	0.60
Mention in Note	0.35*	0.10*	**0.93**
1-year Progression Free Survival			
Death Registry	0.02	0.43	–
Death & New Site Code	0.34	0.20	0.84
Death & New Site in Note	0.31	0.20	0.85
Terminal-Progression Score	0.31*	0.10*	**0.93**

of new metastasis site with 1-year from the treatment initiation; 0 – otherwise. As the standard quality control (Hou et al., 2023), an abstractor randomly sampled $n = 100$ from the study cohort and annotated the gold-standard labels for prescription of targeted medication, terminal condition and new metastasis site by manually reviewing those patients' EHR. The treatment A and Y outcome were defined based on annotations over the labeled set, creating the MCAR data. We reported the treatment and outcome labels as well as their EHR proxies in Table A9 of Supplementary Materials Section B, where we also described the construction of the reasonably good surrogates shown in Table 2.

We extracted a comprehensive list of potential confounders (Table 3). From EHR near the colorectal cancer diagnosis date, we used location specific colorectal cancer diagnosis code to identify the initial tumor location and natural language process tool (Yuan et al., 2021) to extract the initial stage. We also extracted the code for secondary malignancy at lymph node and other distant organs. From EHR between cancer diagnosis and subsequent metastasis, we extracted the codes for common procedures (chemotherapy, radiotherapy, colon biopsy and colon rescission). From EHR near the metastasis date, we used location specific secondary malignancy code to identify the initial metastasis site(s). We also adjusted for the time gap between diagnosis and metastasis, healthcare utilization before metastasis or one year before metastasis measured by days with diagnosis codes and the high-dimensional general health status consisting of diagnosis code counts grouped by the PheWAS catalog (Hou et al., 2022). The targeted therapy arm was associated with factors for poor prognosis including higher proportion of stage IV at diagnosis (81 % vs 58 %), higher proportion of likely liver metastasis (57 % or 67 % vs 34 %). After merging rare lev-

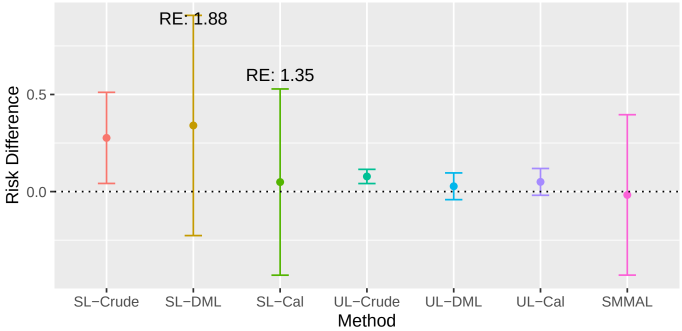
Table 3: Baseline characteristics of full study cohort and two arms in the labeled subset. The format is "count (percentage %)" for binary/categorical variables and "mean (standard deviation)" for numerical variables.

	Full data	Labeled set	
		Chemotherapy	Targeted Therapy
Size	4147	79	21
Demographics			
Age at Metastasis	62.5 (13.8)	65.2 (11.4)	62.7 (15.9)
Female	1926 (46%)	46 (58%)	11 (52%)
White	3470 (84%)	68 (86%)	20 (95%)
Cancer characteristics at diagnosis			
Left Colon Tumor	221 (5%)	5 (6%)	0 (0%)
Right Colon Tumor	890 (21%)	23 (29%)	8 (38%)
Transverse Colon Tumor	420 (10%)	9 (11%)	1 (5%)
Sigmoid Colon Tumor	2092 (50%)	44 (56%)	9 (43%)
Rectum Tumor	2002 (48%)	43 (54%)	5 (24%)
Metastasis Code	2674 (64%)	50 (63%)	16 (76%)
Lymph Node Tumor	364 (9%)	10 (13%)	0 (0%)
Stage I	55 (1%)	0 (0%)	0 (0%)
Stage II	159 (4%)	3 (4%)	0 (0%)
Stage II	992 (24%)	22 (28%)	2 (10%)
Stage IV	2546 (61%)	46 (58%)	17 (81%)
Stage Missing	395 (10%)	8 (10%)	2 (10%)
Cancer characteristics at metastasis			
Year since Diagnosis	0.7 (2)	0.5 (1)	0.8 (1.7)
Lung Metastasis Code	646 (16%)	10 (13%)	8 (38%)
Liver Metastasis Code	1694 (41%)	27 (34%)	14 (67%)
Liver Metastasis in Note	1422 (34%)	27 (34%)	12 (57%)
Treatments between diagnosis and metastasis			
Chemotherapy Code	1.4 (5.3)	1.3 (3.7)	0 (0)
Radiotherapy Code	10.1 (35.1)	10.5 (30.4)	7.2 (29.1)
Colon Biopsy Code	0.6 (1.7)	0.5 (1.5)	0 (0)
Colon Rescission Code	0.4 (0.8)	0.3 (0.7)	0.2 (0.5)
Healthcare utilization			
Before Metastasis	29.7 (59.3)	36.2 (74.5)	13 (21.2)
One Year Before Metastasis	9.8 (15.4)	10.2 (14.7)	4.2 (8.1)

els for cancer characteristics at initial diagnosis (tumor location, cancer stage) and deleting features with fewer than 10 occurrence in labeled subset, we obtained the $p = 55$ potential confounders.

We applied the doubly robust SMMAL in high-dimensions described in Section 3.3. Besides the crude analysis, we ran two benchmark analyses, the double machine learning (DML) (Chernozhukov et al., 2018) using initial estimators (10) and the calibrated estimation (Cal) (Tan, 2020; Smucler et al., 2019) using the calibrated estimators (12). Both supervised learning (SL) using labeled data only and the unsupervised learning (UL) deriving treatment and outcome from the dichotomized surrogates by matching observed prevalence in labeled data were considered. The number of fold was set as $K = 5$, and the penalties factors were selected by the minimal cross-validated entropy. In Figure 5, we displayed the point estimation and the 95 % confidence interval. The confounder adjusted analysis results suggested that on average, targeted therapy had comparable efficacy compared to traditional chemotherapy. Compared to the SL crude analysis which indicated worse outcomes for targeted therapy, our SMMAL accounted for substantial confounding

Figure 5: Point estimate and 95% confidence interval of average risk difference from crude, Double Machine-Learning (DML), calibrated (Cal) and SMMAL analyses. Supervised learning (SL) benchmark analysed only uses the labeled data. Unsupervised learning (UL) benchmark analyses used dichotomized surrogates by matching prevalence observed in labeled data. The RE value indicated the SMMAL's relative efficiency in comparison with the two supervised benchmark methods (ratio of estimated variances).



caused by association between target therapy and factors indicating poor prognosis. Except for the crude analysis that did not adjust for any confounding, our SMMAL had the shortest confidence interval, achieving 1.88 relative efficiency with respect to SL DML and 1.35 relative efficiency with respect to the SL cal. The results from UL methods were questionable as we observed a significant deviation of the UL crude estimation from the SL crude estimation, indicating substantial bias from imperfect data. Coupled with the short confidence intervals, researcher should take caution in the risk of misleading conclusions from the UL methods.

7 General Efficiency Lower Bound

While the paper focused on the method for ATE under double missing SSL setting, we established the theoretical efficient lower bound for general parameter and broader missing data pattern in this section. We considered a generic model for data $(R, R\mathbf{Z}, \mathbf{W})$ with always observed $\mathbf{W}$ and MCAR $\mathbf{Z}$. Specifically, consider

$$\mathcal{S}_{\mathsf{SSL}} = \Bigg\{ d\mathrm{P}_f(r, \mathbf{z}, \mathbf{w}, r) = [\rho_{\mathsf{N}} f(\mathbf{z}, \mathbf{w})]^r \left[(1-\rho_{\mathsf{N}}) \int_{\mathbf{z}\in\mathcal{Z}} f(\mathbf{z}, \mathbf{w}) d\nu_z(\mathbf{z})\right]^{(1-r)} d\nu_{\mathsf{SSL}}(r, \mathbf{z}, \mathbf{w}) : \\ f(\mathbf{z}, \mathbf{w}) d\nu_{\mathsf{cmp}}(\mathbf{z}, \mathbf{w}) \in \mathcal{S}_{\mathsf{cmp}} \Bigg\}$$

for a complete data model class $\mathcal{S}_{\mathsf{cmp}}$ over $\mathcal{Z} \otimes \mathcal{W}$ and measures ν_z over $\mathcal{Z}$, ν_w over $\mathcal{W}$ and

$$\nu_{\mathsf{cmp}} = \nu_z \times \nu_w, \ \nu_{\mathsf{SSL}}(r, \mathbf{z}, \mathbf{w}) = \delta_1(r) \times \nu_{\mathsf{cmp}}(\mathbf{z}, \mathbf{w}) + \delta_0(r) \times \nu_w(\mathbf{w}).$$

Let $\mathscr{H}$ be the nuisance tangent space of $\mathcal{S}_{\mathsf{SSL}}$ at the true model $d\mathrm{P}_{f_*}$ with $f = f_*$. Suppose $\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W})$ is the efficient influence function for parameter $\boldsymbol{\theta}$ under $\mathcal{S}_{\mathsf{cmp}}$. Here we use a different notation $\boldsymbol{\psi}$ for general parameter under missing data components to distinguish from the $\boldsymbol{\phi}$ used specifically for ATE under double missing SSL setting. Our theory was established under the following basic assumptions.

Assumption 5 *For absolute constant M,*

(a) (MCAR) $R \perp\!\!\!\perp (\mathbf{Z}, \mathbf{W})$;

(b) (Informative labels) $\inf_{\|\mathbf{v}\|_2=1} \mathbf{v}^{\mathsf{T}} \operatorname{Var}\left[\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W}) - \mathbb{E}\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W}) \mid \mathbf{W}\}\right] \mathbf{v} \geq 1/M$;

(c) (Model flexibility) $\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W}) \mid \mathbf{W}\} \in \mathscr{H}$;

(d) (Bounded influence function) $\|\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W})\|_2 \leq M$ almost surely.

We derived the SSL efficient influence function by the following proposition.

Proposition 12 *Let $\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W})$ be the efficient influence function for parameter $\boldsymbol{\theta}$ under complete data model $\mathcal{S}_{\mathsf{cmp}}$. Under Assumptions 5a and 5c, the efficient influence function for $\boldsymbol{\theta}$ under SSL model $\mathcal{S}_{\mathsf{SSL}}$ is*

$$\boldsymbol{\psi}_{\mathsf{SSL}}(R, \mathbf{Z}, \mathbf{W}) = \frac{R}{\rho_{\mathsf{N}}}\left[\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W}) - \mathbb{E}\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W}) \mid \mathbf{W}\}\right] + \mathbb{E}\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z}, \mathbf{W}) \mid \mathbf{W}\}. \tag{19}$$

The influence function $\boldsymbol{\psi}_{\mathsf{SSL}}$ leads to a semi-parametric efficiency lower bound.

Theorem 13 *Under Assumptions 5a-5d, we have the minimax semi-parametric efficiency for SSL of $\boldsymbol{\theta}$ under $\mathcal{S}_{\mathsf{SSL}}$,*

$$\inf_{\mathbf{a}:\|\mathbf{a}\|_2=1} \liminf_{c\to\infty} \liminf_{N\to\infty} \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\|f-f_*\|_{\mathsf{TV}}\leq c/\sqrt{\rho_{\mathsf{N}} N}} \frac{\int N\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}^2 d\prod_{i=1}^N \mathbb{P}_f(\mathbf{z}_i, \mathbf{w}_i, r_i)}{\mathbf{a}^{\mathsf{T}} \operatorname{Var}\{\boldsymbol{\psi}_{\mathsf{SSL}}(R, \mathbf{Z}, \mathbf{W})\}\mathbf{a}} \geq 1.$$

We offered the proof of Theorem 13 in Section C6 of the Supplementary Materials. Upper bound would depend on the context. Like Corollary 7, the bound can be attained if non-parametric estimation of nuisance models admit sufficiently fast rate of consistency, which has been thoroughly studied under classical low-dimensional settings by Stone (1977, 1982). While we focus on $\rho_{\mathsf{N}} \to 0$ and $n \ll N$ setting, the theory also applies to classical setting with $\rho_{\mathsf{N}} \in [1/M, 1 - 1/M]$ and $n \asymp N$ setting.

8 Discussion

Motivated by the increasing interest of generating real-world evidence on treatment effect with big yet noisy EHR data, we proposed a robust and efficient semi-supervised estimator for ATE under the double missing SSL setting. The SMMAL estimator gained efficiency by leveraging the large unlabelled data containing noisy yet predictive surrogates for Y and A with almost no additional requirement than those needed for the supervised analysis using the labeled set alone. We established semi-parametric efficiency bound for the ATE estimator under the low dimensional confounder setting and constructed a doubly robust SMMAL estimator for the high dimensional confounder setting.

Unlike the MCAR setting, the missing data propensity score $\mathbb{P}(R = 1 \mid \mathbf{W}) = \rho(\mathbf{W})$ must be modeled and estimated. We conjecture that the efficient influence function under MAR may take the form

$$\begin{aligned}\phi_{\mathsf{MAR}}(RY, RA, \mathbf{W}, R) =& \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\} \\ &+ \frac{R}{\rho(\mathbf{W})}[\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) - \mathbb{E}\{\phi_{\mathsf{cmp}}(Y, A, \mathbf{X}) \mid \mathbf{W}\}].\end{aligned}$$

The estimation of the decaying $\rho(\mathbf{W})$ has been studied in Zhang et al. (2023). When all nuisance models, (μ, π, Π, m, ρ), are consistently estimated at suitable rates, the efficiency lower bound should be attained under ideal conditions. However, extension of the SMMAL with high-dimensional regressions to MAR setting would require a more sophisticated calibration procedure for all 5 models (μ, π, Π, m, ρ), as the potential bias from mis-specified ρ now may impact the orthogonality of ATE estimator toward all 4 other estimated models. Moreover, caution must be taken when making MAR assumption for treatment and outcome data from linked observational data such as a disease registry. Enrollment in registry led by pioneering clinical experts may systematically impact the treatment pattern and care quality, which would put the MAR assumption in doubt.

The classical semi-parametric efficiency theory relies on the correct modeling and estimation of the nuisance models. When some nuisance models cannot be consistently estimated, there is no universal efficiency guarantee for estimation procedures derived from semi-parametric efficiency theory. To ensure efficiency improvement when both the supervised estimator

$$\begin{aligned}\widehat{\Delta}_{\mathsf{SL}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\rho_{\mathsf{N}}}&\left[\widehat{\mu}^{(\mathsf{k})}(1, \mathbf{X}_i) + \frac{A_i}{\widehat{\pi}^{(\mathsf{k})}(1, \mathbf{X}_i)}\{Y_i - \widehat{\mu}^{(\mathsf{k})}(1, \mathbf{X}_i)\}\right] \\ &- \frac{R_i}{\rho_{\mathsf{N}}}\left[\widehat{\mu}^{(\mathsf{k})}(0, \mathbf{X}_i) + \frac{1 - A_i}{\widehat{\pi}^{(\mathsf{k})}(0, \mathbf{X}_i)}\{Y_i - \widehat{\mu}^{(\mathsf{k})}(0, \mathbf{X}_i)\}\right]\end{aligned}$$

and the SMMAL estimator $\widehat{\Delta}_{\mathsf{SMMAL}}$ are consistent and asymptotically normal, we may consider the linear ensemble

$$\widehat{\Delta}_{\mathsf{comb}} = \widehat{\Delta}_{\mathsf{SMMAL}} + b(\widehat{\Delta}_{\mathsf{SL}} - \widehat{\Delta}_{\mathsf{SMMAL}}).$$

Suppose the influence functions for $\widehat{\Delta}_{\mathsf{SMMAL}}$ and $\widehat{\Delta}_{\mathsf{SL}}$ are ϕ_{SSL} and $R\phi_{\mathsf{cmp}}/\rho_{\mathsf{N}}$, respectively. The optimal linear ensemble is given by

$$b_{\mathsf{opt}} = \frac{\operatorname{Var}(\phi_{\mathsf{SSL}}) - \operatorname{Cov}(\phi_{\mathsf{SSL}}, R\phi_{\mathsf{cmp}}/\rho_{\mathsf{N}})}{\operatorname{Var}(\phi_{\mathsf{SSL}}) + \operatorname{Var}(R\phi_{\mathsf{cmp}}/\rho_{\mathsf{N}}) - 2\operatorname{Cov}(\phi_{\mathsf{SSL}}, R\phi_{\mathsf{cmp}}/\rho_{\mathsf{N}})},$$

which can be estimated by the empirical variances and covariance of estimated influence functions constructed with estimated nuisance models $(\widehat{\mu}, \widehat{\pi}, \widehat{m}, \widehat{\Pi})$.

Our doubly robust estimation can be generalized to other models if the calibrated estimation for the model is available. For example, we can directly adopt the estimators from Tan (2020) for linear outcome model. The calibrated estimation is, however, limited to M-estimator in high-dimensional regression due to the paucity of works on Z-estimators in high-dimensional setting. It would be interesting to study if the Z-estimator approach (Vermeulen and Vansteelandt, 2015) can be generalized to high-dimensional setting.

Acknowledgment

This publication is partially supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award [FAIN] totaling $367,807 with 50 percent funded by FDA/HHS. The project is also supported by grants R01 LM013614 and R01 AR080193 from the National Institute of Health. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Supplementary Materials

We present the detailed summaries of simulation results in Appendix A and additional information on the treatment and outcome in the data example in Appendix B. The proofs of Theorems 2-13, Corollary 7 and Proposition 12 are given in Appendix C. The technical details in these proofs are put in Appendix D. Definitions and additional details are stated in Appendix E.

Appendix A. Simulation Tables

The detailed simulation results containing the bias, standard deviation, average standard error, coverage of 95% confidence interval for our semi-supervised $\widehat{\Delta}_{\mathsf{SMMAL}}$, $\widehat{\Delta}_{\mathsf{DR}}$ along with those for the supervised benchmarks were presented in Tables A4-A7. Our semi-supervised $\widehat{\Delta}_{\mathsf{SMMAL}}$, $\widehat{\Delta}_{\mathsf{DR}}$ achieved reasonably honest inference with coverage of 95% confidence interval close to the nominal level and a better efficiency than the supervised benchmark. The detailed simulation results containing the bias, standard deviation, coverage of 95% confidence interval for the unsupervised benchmarks were presented in Table A8.

Table A4: Detailed simulation results on bias, standard deviation (SD), average standard error (ASE), coverage of 95% confidence interval (Cov) for SMMAL and supervised benchmark (SL) under low-dimensional smooth models. The 25 rows correspond to 5×5 points in **top left plot in Figure 3**, indexed by the designed AUC of surrogates for A (column 1, vertical axis in Figure 3) and Y (column 2, horizontal axis in Figure 3). Bias, standard deviation (SD) and average standard error (ASE) were multiplied by 100.

AUC		SL				SMMAL				RE
A	Y	Bias	SD	ASE	Cov	Bias	SD	ASE	Cov	
0.80	0.80	-0.81	4.08	4.07	0.94	-0.79	4.04	3.98	0.94	1.02
0.90	0.80	-0.85	4.14	4.08	0.94	-0.79	3.97	3.95	0.94	1.09
0.95	0.80	-0.75	3.87	4.09	0.96	-0.71	3.72	3.84	0.95	1.08
0.99	0.80	-0.90	4.16	4.11	0.94	-0.92	3.69	3.66	0.94	1.25
0.999	0.80	-0.56	4.21	4.11	0.94	-0.61	3.64	3.59	0.94	1.33
0.80	0.90	-1.01	4.10	4.07	0.95	-1.00	4.01	3.92	0.94	1.04
0.90	0.90	-0.69	4.08	4.09	0.95	-0.66	3.83	3.85	0.95	1.14
0.95	0.90	-0.70	4.11	4.09	0.95	-0.72	3.70	3.63	0.94	1.22
0.99	0.90	-0.66	4.21	4.10	0.94	-0.61	3.34	3.29	0.94	1.58
0.999	0.90	-0.35	4.02	4.10	0.95	-0.43	3.14	3.13	0.94	1.62
0.80	0.95	-1.14	4.09	4.06	0.94	-1.07	4.01	3.87	0.94	1.05
0.90	0.95	-0.71	3.98	4.08	0.95	-0.70	3.81	3.76	0.94	1.09
0.95	0.95	-0.69	4.01	4.09	0.96	-0.59	3.49	3.51	0.96	1.32
0.99	0.95	-0.27	4.12	4.09	0.94	-0.23	3.06	3.01	0.94	1.80
0.999	0.95	-0.22	4.08	4.10	0.95	-0.20	2.69	2.78	0.96	2.29
0.80	0.99	-0.92	4.07	4.07	0.94	-0.89	3.94	3.86	0.94	1.07
0.90	0.99	-0.71	3.99	4.09	0.95	-0.52	3.63	3.71	0.95	1.22
0.95	0.99	-0.10	4.12	4.09	0.94	-0.30	3.50	3.32	0.93	1.38
0.99	0.99	-0.37	4.05	4.10	0.95	-0.07	2.73	2.61	0.94	2.23
0.999	0.99	-0.22	4.12	4.10	0.95	-0.05	2.31	2.21	0.94	3.17
0.80	0.999	-0.95	3.87	4.08	0.95	-0.94	3.79	4.00	0.95	1.04
0.90	0.999	-0.60	4.10	4.08	0.94	-0.69	3.89	3.87	0.94	1.10
0.95	0.999	-0.39	4.02	4.08	0.94	-0.37	3.67	3.59	0.94	1.20
0.99	0.999	0.09	4.05	4.09	0.94	0.02	3.13	3.04	0.94	1.67
0.999	0.999	0.03	4.08	4.11	0.95	-0.12	2.64	2.64	0.95	2.38

Table A5: Detailed simulation results on bias, standard deviation (SD), average standard error (ASE), coverage of 95% confidence interval (Cov) for SMMAL and supervised benchmark (SL) under high-dimensional models with logistic regression PS and OR. The 25 rows correspond to 5 × 5 points in **top right plot in Figure 3**, indexed by the designed AUC of surrogates for A (column 1, vertical axis in Figure 3) and Y (column 2, horizontal axis in Figure 3). Bias, standard deviation (SD) and average standard error (ASE) were multiplied by 100.

AUC		SL				SMMAL				RE
A	Y	Bias	SD	ASE	Cov	Bias	SD	ASE	Cov	
0.80	0.80	-0.28	4.71	4.49	0.94	-0.22	4.62	4.47	0.95	1.04
0.90	0.80	-0.24	4.70	4.48	0.95	-0.19	4.63	4.36	0.95	1.03
0.95	0.80	-0.24	4.66	4.49	0.96	-0.20	4.43	4.29	0.95	1.11
0.99	0.80	-0.22	4.65	4.49	0.94	-0.17	4.33	4.18	0.95	1.15
0.999	0.80	-0.32	4.70	4.48	0.95	-0.33	4.33	4.11	0.94	1.18
0.80	0.90	-0.26	4.68	4.49	0.94	-0.16	4.56	4.34	0.94	1.06
0.90	0.90	-0.26	4.64	4.49	0.94	-0.22	4.13	4.09	0.94	1.27
0.95	0.90	-0.33	4.72	4.49	0.94	-0.18	4.13	3.92	0.94	1.31
0.99	0.90	-0.31	4.63	4.49	0.95	0.03	3.79	3.65	0.94	1.50
0.999	0.90	-0.24	4.68	4.48	0.95	-0.07	3.63	3.50	0.94	1.67
0.80	0.95	-0.32	4.69	4.49	0.95	-0.26	4.53	4.27	0.95	1.07
0.90	0.95	-0.31	4.62	4.49	0.95	-0.36	3.94	3.91	0.94	1.37
0.95	0.95	-0.27	4.71	4.49	0.95	-0.27	3.89	3.67	0.94	1.46
0.99	0.95	-0.21	4.71	4.48	0.95	-0.11	3.44	3.28	0.93	1.87
0.999	0.95	-0.33	4.71	4.48	0.93	-0.19	3.21	3.05	0.93	2.15
0.80	0.99	-0.31	4.67	4.49	0.95	-0.25	4.29	4.15	0.95	1.19
0.90	0.99	-0.31	4.68	4.49	0.94	-0.09	3.76	3.64	0.94	1.55
0.95	0.99	-0.28	4.71	4.49	0.94	-0.11	3.46	3.27	0.94	1.86
0.99	0.99	-0.30	4.73	4.49	0.94	-0.09	2.79	2.67	0.94	2.89
0.999	0.99	-0.28	4.70	4.48	0.95	-0.12	2.32	2.27	0.94	4.11
0.80	0.999	-0.30	4.65	4.48	0.96	-0.22	4.26	4.09	0.95	1.19
0.90	0.999	-0.26	4.70	4.49	0.95	-0.04	3.55	3.50	0.96	1.76
0.95	0.999	-0.24	4.68	4.48	0.95	-0.09	3.13	3.06	0.94	2.23
0.99	0.999	-0.27	4.69	4.49	0.95	-0.08	2.39	2.33	0.95	3.86
0.999	0.999	-0.32	4.69	4.49	0.95	-0.05	1.85	1.78	0.94	6.49

Table A6: Detailed simulation results on bias, standard deviation (SD), average standard error (ASE), coverage of 95% confidence interval (Cov) for SMMAL and supervised benchmark (SL) under high-dimensional models with miss-specified PS and correct OR models. The 25 rows correspond to 5 × 5 points in **bottom right plot in Figure 3**, indexed by the designed AUC of surrogates for A (column 1, vertical axis in Figure 3) and Y (column 2, horizontal axis in Figure 3). Bias, standard deviation (SD) and average standard error (ASE) were multiplied by 100.

AUC		SL				SMMAL				RE
A	Y	Bias	SD	ASE	Cov	Bias	SD	ASE	Cov	
0.80	0.80	-0.28	4.67	4.46	0.94	-0.21	4.61	4.43	0.94	1.03
0.90	0.80	-0.31	4.72	4.45	0.94	-0.32	4.55	4.31	0.94	1.08
0.95	0.80	-0.26	4.74	4.45	0.94	-0.17	4.46	4.22	0.94	1.13
0.99	0.80	-0.29	4.67	4.45	0.95	-0.19	4.28	4.09	0.94	1.20
0.999	0.80	-0.29	4.72	4.45	0.95	-0.29	4.21	4.03	0.96	1.25
0.80	0.90	-0.28	4.70	4.45	0.94	-0.26	4.64	4.32	0.94	1.02
0.90	0.90	-0.29	4.71	4.45	0.94	-0.26	4.27	4.08	0.95	1.22
0.95	0.90	-0.30	4.73	4.46	0.95	-0.25	4.18	3.90	0.93	1.28
0.99	0.90	-0.31	4.78	4.46	0.94	0.04	3.95	3.65	0.93	1.47
0.999	0.90	-0.27	4.76	4.46	0.95	-0.04	3.83	3.51	0.94	1.55
0.80	0.95	-0.29	4.70	4.45	0.95	-0.25	4.54	4.23	0.93	1.08
0.90	0.95	-0.32	4.72	4.46	0.94	-0.13	4.03	3.89	0.95	1.37
0.95	0.95	-0.27	4.69	4.45	0.95	-0.10	3.85	3.63	0.93	1.49
0.99	0.95	-0.26	4.72	4.46	0.94	-0.07	3.43	3.24	0.94	1.90
0.999	0.95	-0.31	4.65	4.45	0.94	-0.01	3.16	3.01	0.94	2.18
0.80	0.99	-0.28	4.70	4.45	0.94	-0.22	4.39	4.12	0.94	1.15
0.90	0.99	-0.31	4.72	4.45	0.94	-0.15	3.61	3.65	0.96	1.71
0.95	0.99	-0.23	4.71	4.45	0.94	-0.05	3.45	3.27	0.92	1.87
0.99	0.99	-0.37	4.72	4.46	0.94	-0.12	2.92	2.71	0.95	2.62
0.999	0.99	-0.25	4.75	4.45	0.94	-0.04	2.40	2.32	0.94	3.91
0.80	0.999	-0.32	4.72	4.46	0.94	-0.25	4.31	4.05	0.94	1.20
0.90	0.999	-0.26	4.70	4.45	0.94	-0.09	3.51	3.50	0.96	1.79
0.95	0.999	-0.28	4.69	4.45	0.94	-0.08	3.07	3.04	0.94	2.33
0.99	0.999	-0.32	4.76	4.45	0.94	-0.07	2.38	2.32	0.95	4.03
0.999	0.999	-0.30	4.68	4.46	0.94	0.11	1.81	1.79	0.95	6.68

Table A7: Detailed simulation results on bias, standard deviation (SD), average standard error (ASE), coverage of 95% confidence interval (Cov) for SMMAL and supervised benchmark (SL) under high-dimensional models with correct PS and miss-specified OR models. The 25 rows correspond to 5×5 points in **bottom left plot in Figure 3**, indexed by the designed AUC of surrogates for A (column 1, vertical axis in Figure 3) and Y (column 2, horizontal axis in Figure 3). Bias, standard deviation (SD) and average standard error (ASE) were multiplied by 100.

AUC		SL				SMMAL				RE
A	Y	Bias	SD	ASE	Cov	Bias	SD	ASE	Cov	
0.80	0.80	-0.21	4.70	4.49	0.95	-0.19	4.70	4.46	0.94	1.00
0.90	0.80	-0.20	4.75	4.49	0.95	-0.22	4.60	4.35	0.94	1.07
0.95	0.80	-0.16	4.64	4.49	0.96	-0.18	4.35	4.27	0.94	1.13
0.99	0.80	-0.22	4.64	4.49	0.95	-0.14	4.36	4.15	0.94	1.13
0.999	0.80	-0.17	4.74	4.49	0.95	-0.19	4.30	4.09	0.95	1.21
0.80	0.90	-0.20	4.69	4.48	0.95	-0.23	4.50	4.33	0.95	1.08
0.90	0.90	-0.23	4.70	4.48	0.95	-0.24	4.21	4.10	0.94	1.24
0.95	0.90	-0.25	4.75	4.48	0.95	-0.16	3.99	3.92	0.95	1.42
0.99	0.90	-0.22	4.67	4.49	0.95	0.01	3.83	3.66	0.95	1.49
0.999	0.90	-0.26	4.68	4.48	0.95	-0.15	3.68	3.51	0.94	1.62
0.80	0.95	-0.18	4.67	4.48	0.96	-0.21	4.41	4.25	0.93	1.12
0.90	0.95	-0.19	4.69	4.48	0.95	-0.19	3.99	3.93	0.95	1.38
0.95	0.95	-0.21	4.67	4.48	0.96	-0.03	3.65	3.68	0.95	1.64
0.99	0.95	-0.21	4.67	4.48	0.95	-0.32	3.38	3.29	0.94	1.90
0.999	0.95	-0.17	4.65	4.49	0.96	0.09	3.20	3.07	0.94	2.12
0.80	0.99	-0.18	4.64	4.49	0.96	-0.21	4.20	4.13	0.95	1.22
0.90	0.99	-0.23	4.65	4.49	0.95	-0.25	3.72	3.68	0.95	1.56
0.95	0.99	-0.16	4.71	4.48	0.95	-0.05	3.27	3.30	0.95	2.08
0.99	0.99	-0.22	4.61	4.49	0.95	-0.08	2.81	2.70	0.94	2.70
0.999	0.99	-0.25	4.69	4.48	0.96	-0.01	2.51	2.32	0.92	3.51
0.80	0.999	-0.20	4.67	4.48	0.95	-0.08	4.09	4.06	0.96	1.31
0.90	0.999	-0.17	4.71	4.48	0.95	-0.09	3.60	3.53	0.95	1.71
0.95	0.999	-0.21	4.60	4.48	0.95	-0.08	3.02	3.09	0.95	2.33
0.99	0.999	-0.24	4.68	4.48	0.95	-0.05	2.39	2.35	0.95	3.85
0.999	0.999	-0.17	4.67	4.48	0.95	0.11	1.94	1.84	0.94	5.81

Table A8: Detailed simulation results on bias, standard deviation (SD) and coverage of 95 % confidence interval (Cov) for unsupervised analyses under **settings in Figure 4** low-dimensional smooth model (Low-d, top left plot in Figure 4), high-dimensional logistic model (High-d, top right plot), high-dimensional model with mis-specified propensity scores (High-d misPS, bottom right plot in Figure 4), high-dimensional model with mis-specified outcome regression (High-d misOR, bottom left plot in Figure 4). The 25 rows correspond to 5 × 5 points in each plot in Figure 4, indexed by the designed AUC of surrogates for A (column 1, vertical axis in Figure 4) and Y (column 2, horizontal axis in Figure 4). Bias and standard deviation (SD) were multiplied by 100.

AUC		Low-d			High-d			High-d misPS			High-d misOR		
A	Y	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
0.80	0.80	-9.08	0.92	0.00	-2.55	1.04	0.27	-2.57	1.04	0.29	-2.87	0.97	0.16
0.90	0.80	-8.37	0.92	0.00	-2.14	0.98	0.44	-2.12	0.98	0.45	-2.47	0.95	0.31
0.95	0.80	-7.85	0.99	0.00	-1.77	1.04	0.58	-1.75	1.04	0.59	-2.10	1.07	0.48
0.99	0.80	-7.15	0.99	0.00	-1.36	1.06	0.73	-1.38	1.09	0.72	-1.59	1.03	0.68
0.999	0.80	-6.94	1.05	0.00	-1.25	1.06	0.77	-1.13	1.09	0.80	-1.47	1.09	0.69
0.80	0.90	-8.47	0.94	0.00	-2.04	1.00	0.48	-2.04	1.05	0.46	-2.46	0.94	0.29
0.90	0.90	-7.18	0.98	0.00	-1.66	1.04	0.61	-1.73	0.92	0.61	-1.92	1.01	0.54
0.95	0.90	-6.31	0.98	0.00	-1.35	0.96	0.74	-1.42	1.11	0.68	-1.61	1.07	0.61
0.99	0.90	-5.27	0.99	0.00	-1.01	1.09	0.83	-0.96	1.11	0.80	-1.19	1.08	0.78
0.999	0.90	-4.75	1.01	0.00	-0.80	1.12	0.87	-0.82	1.10	0.86	-1.03	1.15	0.81
0.80	0.95	-8.09	0.95	0.00	-1.77	1.01	0.59	-1.77	1.03	0.58	-2.15	0.99	0.43
0.90	0.95	-6.42	0.99	0.00	-1.37	1.02	0.73	-1.46	1.00	0.70	-1.74	1.04	0.60
0.95	0.95	-5.29	1.00	0.00	-1.15	1.07	0.79	-1.21	1.09	0.77	-1.34	1.02	0.74
0.99	0.95	-4.11	1.02	0.01	-0.75	1.16	0.86	-0.75	1.12	0.87	-0.96	1.13	0.81
0.999	0.95	-3.50	0.96	0.03	-0.60	1.14	0.89	-0.59	1.17	0.88	-0.74	1.11	0.87
0.80	0.99	-7.53	0.93	0.00	-1.39	1.05	0.72	-1.49	1.00	0.69	-1.74	0.96	0.58
0.90	0.99	-5.40	0.99	0.00	-1.05	1.02	0.81	-1.16	0.97	0.79	-1.38	0.98	0.74
0.95	0.99	-3.95	0.97	0.01	-0.82	1.09	0.86	-0.86	1.06	0.86	-0.97	1.04	0.86
0.99	0.99	-2.46	1.02	0.25	-0.47	1.14	0.91	-0.51	1.11	0.90	-0.58	1.16	0.87
0.999	0.99	-1.65	1.04	0.57	-0.31	1.11	0.93	-0.33	1.14	0.92	-0.44	1.14	0.90
0.80	0.999	-7.29	0.96	0.00	-1.23	1.00	0.76	-1.28	1.01	0.75	-1.60	0.98	0.63
0.90	0.999	-5.01	1.01	0.00	-0.93	1.02	0.84	-1.01	0.95	0.83	-1.20	0.96	0.78
0.95	0.999	-3.53	1.06	0.04	-0.71	1.06	0.89	-0.73	1.06	0.86	-0.84	1.06	0.87
0.99	0.999	-1.82	1.03	0.51	-0.38	1.15	0.91	-0.40	1.14	0.91	-0.48	1.13	0.90
0.999	0.999	-1.03	1.00	0.77	-0.23	1.15	0.92	-0.20	1.15	0.92	-0.29	1.17	0.92

Table A9: Treatment, outcomes and their surrogates in full study cohort and two arms in the labeled subset. The format is "count (percentage %)" for binary variables and "mean (standard deviation)" for numerical variables.

	Full data	Labeled set	
Size	4147	100	
Treatment Arms		Chemotherapy	Targeted Therapy
Gold-standard labels	–	79	21
EHR proxies for treatment between metastasis and treatment			
Targeted Medication Code Count	0.2 (0.7)	0.1 (0.4)	0.6 (1.2)
Targeted Therapy Mention Count in Note	4.6 (13.6)	1.4 (3.9)	14.4 (24.6)
Endpoints up to 1 year after treatment:			
Gold-standard labels			
Terminal Condition	–	18 (23%)	11 (52%)
New Metastasis	–	7 (9%)	2 (10%)
EHR proxies for outcome during 1 year follow-up			
Occurrence of death record	781 (19%)	12 (15%)	9 (43%)
Diagnosis Code Count in Last Month	33.5 (50.8)	34.1 (49)	53.4 (52.6)
Procedure Code Count in Last Month	1.9 (4.8)	2 (5)	2.8 (4.9)
New Metastasis Code Counts	3.2 (9.3)	4.9 (10.7)	1.9 (4.9)

Appendix B. Treatment and Outcomes in Data Example

We report the treatment and outcome labels as well as their EHR proxies in Table A9. The targeted therapy arm was marginally associated with poorer outcomes in terms of terminal conditions (52 % vs 23 %). Tables A9 shows that occurrence of EHR proxies cannot accurately indicate prescription information. We used the log counts of targeted medication mention in note from metastasis to treatment as the surrogate for treatment indicator due to its large contrast between the two labeled arms. Progress-free survival is poorly structured in EHR with no clear indicator. We construct a terminal-progression score with reasonably good prediction power (see Table 2) using death records, activity (diagnosis and procedure codes) in last EHR month and metastasis code for a new site during 1 year follow-up.

Appendix C. Proofs of Main Text Theorems

To analyze the cross-fitting, we adopt the following notations for the conditional expectations given different part of the data.

Definition A14 *Recall that we denote the full data, fold-k data, out-of-fold-k data and out-of-folds-k_1-k_2 data as $\mathscr{D}$, $\mathscr{D}_k$, $\mathscr{D}_k^c$ and $\mathscr{D}_{k_1,k_2}^c$, respectively. The conditional expectation for samples with index in set $\mathcal{I}$ conditionally on subset of the data $\mathscr{D}'$ is denoted as*

$$\mathbb{E}_{i\in\mathcal{I}}\{f(\mathbf{D}_i) \mid \mathscr{D}\},\ \mathcal{I} \subseteq \{1,\dots,n+N\}, \mathscr{D}' \subseteq \mathscr{D}.$$

C1 Proof of Theorem 2

We first analyze the bias terms in $\widehat{\Delta}_{\mathsf{SMMAL}}$ and $\widehat{\mathcal{V}}_{\mathsf{SMMAL}}$ from the cross-fitted estimators through a lemma. We group the limiting nuisance models as $\bar{\boldsymbol{\eta}} = (\pi_*, \mu_*, \bar{\Pi}, \bar{m}_*)$. In the proof, we repeat the same analyze on two components for treatment arms in ϕ_{SSL},

$$\begin{aligned}\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}) =&\mu(a,\mathbf{X}) + \frac{\mathrm{I}(A=a)}{\pi(a,\mathbf{X})}\{Y-\mu(a,\mathbf{X})\} &\text{(A.20)}\\ &+\frac{1}{\pi(a,\mathbf{X})}\left\{\frac{R}{\rho_{\mathsf{N}}}-1\right\}\{\mathrm{I}(A=a)Y-\Pi(a,\mathbf{W})m(a,\mathbf{W})\}\\ &-\frac{\mu(a,\mathbf{X})}{\pi(a,\mathbf{X})}\left\{\frac{R}{\rho_{\mathsf{N}}}-1\right\}\{\mathrm{I}(A=a)-\Pi(a,\mathbf{W})\}. &\text{(A.21)}\end{aligned}$$

Notice the connection of $\phi_{\mathsf{SSL,a}}$ to the efficient influence function when $\boldsymbol{\eta}$ equals the true models $\boldsymbol{\eta}_*$ $\phi_{\mathsf{SSL}}(RY, RA, \mathbf{W}, R) = \phi_{\mathsf{SSL,1}}(\mathbf{D};\boldsymbol{\eta}*) - \phi_{\mathsf{SSL,0}}(\mathbf{D};\boldsymbol{\eta}*) - \Delta_*$. We may identify the average treatment by $\Delta_* = \mathbb{E}\{\mu(1,\mathbf{X})-\mu(0,\mathbf{X})\} = \mathbb{E}\{\phi_{\mathsf{SSL,1}}(\mathbf{D};\bar{\boldsymbol{\eta}}) - \phi_{\mathsf{SSL,0}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}$.

Lemma A15 *Let $\boldsymbol{\eta}_n = (\pi_n, \mu_n, \Pi_n, m_n)$ be a (deterministic) sequence of nuisance models satisfying almost surely*

$$\sup_{a=0,1} \max\{|1/\pi_n(a,\mathbf{X}_i)|, |\mu_n(a,\mathbf{X}_i)|, |\Pi_n(a,\mathbf{W}_i)|, |m_n(a,\mathbf{W}_i)|\} \le M. \tag{A.22}$$

Under Assumptions 2a and 2b, we have

1. *For bias:*
$$|\mathbb{E}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n) - \phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}| \lesssim \|\pi_n-\pi_*\|_2\|\mu_n-\mu_*\|_2,$$

2. *For variance:*
$$\begin{aligned}&\rho_{\mathsf{N}}\operatorname{Var}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n) - \phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}\\ \lesssim&\|\pi_n-\pi_*\|_2^2 + \|\mu_n-\mu_*\|_2^2 + \|m_n-\bar{m}\|_2^2 + \|\Pi_n-\bar{\Pi}\|_2^2,\end{aligned}$$

3. *For variance estimation:*
$$\begin{aligned}&\mathbb{E}\left\{\rho_{\mathsf{N}}\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)^2\right\} - \mathbb{E}\left\{\rho_{\mathsf{N}}\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})^2\right\}\\ \lesssim&\|\pi_n-\pi_*\|_2 + \|\mu_n-\mu_*\|_2 + \|m_n-\bar{m}\|_2 + \|\Pi_n-\bar{\Pi}\|_2\\ &+\|\pi_n-\pi_*\|_2^2 + \|\mu_n-\mu_*\|_2^2 + \|m_n-\bar{m}\|_2^2 + \|\Pi_n-\bar{\Pi}\|_2^2.\end{aligned}$$

Proof [Proof of Lemma A15] We prove the lemma through a calculation of the expectations and variances of the quantities of interest. To simplify our notation, we denote

$$\triangle\mu = \mu_n(a,\mathbf{X}) - \mu_*(a,\mathbf{X}),\ \triangle\pi = \pi_n(a,\mathbf{X}) - \pi_*(a,\mathbf{X}), \tag{A.23}$$

$$\triangle m = m_n(a,\mathbf{W}) - \bar{m}(a,\mathbf{W}),\ \triangle\Pi = \Pi_n(a,\mathbf{W}) - \bar{\Pi}(a,\mathbf{W}). \tag{A.24}$$

We substitute Y and A in analysis by the model definition

$$\mathbb{E}\{\mathrm{I}(A=a) \mid \mathbf{X}\} = \pi_*(a,\mathbf{X}),\ \mathbb{E}\{\mathrm{I}(A=a)Y \mid \mathbf{X}\} = \pi_*(a,\mathbf{X})\mu_*(a,\mathbf{X}),$$
$$\mathbb{E}\{\mathrm{I}(A=a) \mid \mathbf{W}\} = \bar{\Pi}(a,\mathbf{W}),\ \mathbb{E}\{\mathrm{I}(A=a)Y \mid \mathbf{W}\} = \bar{\Pi}(a,\mathbf{W})\bar{m}(a,\mathbf{W}).$$

1. **For bias:**

   We decompose the expectation into

$$\begin{aligned}
&\mathbb{E}\left\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n) - \phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\right\}\\
=&\underbrace{\mathbb{E}\left\{\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_n,\mu_n,\Pi_n,m_n)) - \phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_n,\mu_*,\Pi_n,m_n))\right\}}_{T_1}\\
&+\underbrace{\mathbb{E}\left\{\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_n,\mu_*,\Pi_n,m_n)) - \phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_*,\mu_*,\Pi_n,m_n))\right\}}_{T_2}\\
&+\underbrace{\mathbb{E}\left\{\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_*,\mu_*,\Pi_n,m_n)) - \phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_*,\mu_*,\bar{\Pi},\bar{m}))\right\}}_{T_3},
\end{aligned}$$

   which we shall analyze separately.

   First, note that T_1 can be written as

$$\begin{aligned}
T_1 =&\mathbb{E}\left(\triangle\mu\left[1 - \frac{I(A=a)}{\pi_n(a,\mathbf{X})} - \frac{1}{\pi_n(a,\mathbf{X})}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a) - \Pi_n(a,\mathbf{W})\}\right]\right)\\
=&\mathbb{E}\left\{-\triangle\mu\triangle\pi/\pi_n(a,\mathbf{X})\right\}.
\end{aligned}$$

   As a result, we have a bound for T_1 by (A.22) and the Cauchy-Schwartz inequality to obtain $|T_1| \le M\|\triangle\mu\|_2\|\triangle\pi\|_2$. Second, we calculate T_2,

$$\begin{aligned}
T_2 =&\mathbb{E}\left[\frac{\triangle\pi}{\pi_n(a,\mathbf{X})\pi_*(a,\mathbf{X})}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)Y - \Pi_n(a,\mathbf{W})m_n(a,\mathbf{W})\}\right]\\
&-\mathbb{E}\left[\frac{\triangle\pi}{\pi_n(a,\mathbf{X})\pi_*(a,\mathbf{X})}\mu_*(a,\mathbf{X})\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a) - \Pi_n(a,\mathbf{W})\}\right]\\
=&0.
\end{aligned}$$

   Third, we calculate T_3,

$$\begin{aligned}
T_3 =&\mathbb{E}\left[\frac{1}{\pi_*(a,\mathbf{X})}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)Y - \Pi_n(a,\mathbf{W})m_n(a,\mathbf{W})\}\right]\\
&-\mathbb{E}\left[\frac{1}{\pi_*(a,\mathbf{X})}\mu_*(a,\mathbf{X})\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a) - \Pi_n(a,\mathbf{W})\}\right]
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}\left[\frac{1}{\pi_*(a,\mathbf{X})}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)Y-\bar{\Pi}(a,\mathbf{W})\bar{m}(a,\mathbf{W})\}\right]\\
&+ \mathbb{E}\left[\frac{1}{\pi_*(a,\mathbf{X})}\mu_*(a,\mathbf{X})\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)-\bar{\Pi}(a,\mathbf{W})\}\right]\\
=&0.
\end{aligned}
$$

Putting the bounds for T_1-T_3 together, we have therefore have that

$$
|\mathbb{E}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}| \lesssim \|\pi_n-\pi_*\|_2\|\mu_n-\mu_*\|_2.
$$

2. **For variance:**

   Here, we first establish the order for the second moments of terms with $(R/\rho_{\mathsf{N}}-1)$ in them as follows

$$
\begin{aligned}
&\mathbb{E}[\{h_1(Y,A,\mathbf{W})+(R/\rho_{\mathsf{N}}-1)h_2(Y,A,\mathbf{W})\}^2]\\
=&\mathbb{E}\{h_1(Y,A,\mathbf{W})^2\}+\mathbb{E}\{(R/\rho_{\mathsf{N}}-1)^2\}\mathbb{E}\{h_2(Y,A,\mathbf{W})^2\}\\
=&\|h_1\|_2^2+(1/\rho_{\mathsf{N}}-1)\|h_2\|_2^2\\
\leq&\|h_1\|_2^2+\|h_2^2\|_2/\rho_{\mathsf{N}}.
\end{aligned}
\tag{A.25}
$$

The bound for the variance is derived from the bound for the second moment

$$
\mathrm{Var}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\} \leq \mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}^2\right].
$$

By the inequality $(a+b+c)^2\leq 4(a^2+b^2+c^2)$ for any $a,b,c\in\mathbb{R}$, we can control the bound in the decomposition:

$$
\begin{aligned}
&\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}^2\right]\\
=&4\underbrace{\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_n,\mu_n,\Pi_n,m_n))-\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_n,\mu_*,\Pi_n,m_n))\}^2\right]}_{T_1'}\\
&+4\underbrace{\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_n,\mu_*,\Pi_n,m_n))-\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_*,\mu_*,\Pi_n,m_n))\}^2\right]}_{T_2'}\\
&+4\underbrace{\mathbb{E}\left[\Bigl\{\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_*,\mu_*,\Pi_n,m_n))-\phi_{\mathsf{SSL,a}}(\mathbf{D};(\pi_*,\mu_*,\bar{\Pi},\bar{m}))\Bigr\}^2\right]}_{T_3'}.
\end{aligned}
$$

Under Assumptions 2a, 2b and (A.22), since we have everything except for R/ρ_{N} bounded in T_1'-T_3', we can derive upper bounds on their rates as follows:

$$
\begin{aligned}
T_1'=&\mathbb{E}\left(\triangle\mu^2\left[1-\frac{I(A=a)}{\pi_n(a,\mathbf{X}_i)}-\frac{1}{\pi_n(a,\mathbf{X}_i)}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)-\Pi_n(a,\mathbf{W})\}\right]^2\right)\\
\lesssim&\|\mu_n-\mu_*\|_2^2/\rho_{\mathsf{N}},
\end{aligned}
$$

$$\begin{aligned}
T_2' =&\mathbb{E}\left(\left(\frac{\triangle\pi}{\pi_n(a,\mathbf{X})\pi_*(a,\mathbf{X})}\right)^2\left[\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)Y-\Pi_n(a,\mathbf{W})m_n(a,\mathbf{W})\}\right.\right.\\
&\left.\left.-\mu_*(a,\mathbf{X})\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\{I(A=a)-\Pi_n(a,\mathbf{W})\}\right]^2\right)\\
\lesssim&\|\pi_n-\pi_*\|_2^2/\rho_{\mathsf{N}},\\
T_3' \leq&2\mathbb{E}\left[\triangle\Pi^2\left\{\frac{\mu_*(a,\mathbf{X})-\bar{m}(a,\mathbf{W})}{\pi_*(a,\mathbf{X})}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\right\}^2\right]+2\mathbb{E}\left(\triangle m^2\left[\frac{\Pi_n(a,\mathbf{W})}{\pi_*(a,\mathbf{X})}\left(\frac{R}{\rho_{\mathsf{N}}}-1\right)\right]^2\right)\\
\lesssim&\|\Pi_n-\bar{\Pi}\|_2^2/\rho_{\mathsf{N}}+\|m_n-\bar{m}\|_2^2/\rho_{\mathsf{N}}.
\end{aligned}$$

Putting the rates for T_1'-T_3' together, we therefore obtain

$$\begin{aligned}
&\rho_{\mathsf{N}}\operatorname{Var}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}\\
\leq&\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}^2\right]\\
\lesssim&\|\pi_n-\pi_*\|_2^2+\|\mu_n-\mu_*\|_2^2+\|m_n-\bar{m}\|_2^2+\|\Pi_n-\bar{\Pi}\|_2^2.
\end{aligned}$$

3. **For variance estimator:**

   We establish the last result by connecting it to the second bound above,

$$\begin{aligned}
&\mathbb{E}\left\{\rho_{\mathsf{N}}\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)^2\right\}-\mathbb{E}\left\{\rho_{\mathsf{N}}\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})^2\right\}\\
\leq&\rho_{\mathsf{N}}\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}^2\right]\\
&+2\rho_{\mathsf{N}}\mathbb{E}\left[\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}\right]\\
\leq&\rho_{\mathsf{N}}\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}^2\right]\\
&+\rho_{\mathsf{N}}\sqrt{\mathbb{E}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})^2\}\mathbb{E}\left[\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)-\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}^2\right]}\\
\lesssim&\|\pi_n-\pi_*\|_2+\|\mu_n-\mu_*\|_2+\|m_n-\bar{m}\|_2+\|\Pi_n-\bar{\Pi}\|_2\\
&+\|\pi_n-\pi_*\|_2^2+\|\mu_n-\mu_*\|_2^2+\|m_n-\bar{m}\|_2^2+\|\Pi_n-\bar{\Pi}\|_2^2.
\end{aligned}$$

Thus, we have obtained all three rates for bias, variance and variance estimation. ■

Using Lemma A15, we can now proceed to prove Theorem 2. Denote the out-of-k-fold estimators for nuisance models as $\widehat{\boldsymbol{\eta}}^{(\mathsf{k})}=(\widehat{\pi}^{(\mathsf{k})},\widehat{\mu}^{(\mathsf{k})},\widehat{\Pi}^{(\mathsf{k})},\widehat{m}^{(\mathsf{k})})$. We shall first establish the asymptotic approximation for the fold-k estimator

$$\begin{aligned}
\widehat{\Delta}_{\mathsf{SMMAL}}^{(\mathsf{k})}=&\frac{K}{N}\sum_{i\in\mathcal{I}_k}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\\
=&\frac{K}{N}\sum_{i\in\mathcal{I}_k}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})+o_p\left(n^{-1/2}\right).
\end{aligned}$$

Then the asymptotic normally follows from the central limit theorem regarding the empirical mean term of the i.i.d. random variables. We will thereafter conclude the proof by showing the variance estimator is indeed consistent.

To establish the asymptotic expansion, we consider the decomposition

$$\begin{aligned}
&\widehat{\Delta}^{(\mathsf{k})}_{\mathsf{SMMAL}} \\
=&\frac{K}{N}\sum_{i\in\mathcal{I}_k}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}}) \\
&+\frac{K}{N}\sum_{i\in\mathcal{I}_k}\left[\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\mathbb{E}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\}\right] \\
&+\mathbb{E}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\} \\
&-\frac{K}{N}\sum_{i\in\mathcal{I}_k}\left[\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\mathbb{E}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\}\right] \\
&-\mathbb{E}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\}. \qquad (A.26)
\end{aligned}$$

We can now apply Lemma A15 along with Assumption 2d to get

$$\begin{aligned}
&\mathrm{Var}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\}=o_p\left(1/\rho_{\mathsf{N}}\right), \\
&\mathbb{E}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\}=o_p\left(n^{-1/2}\right). \qquad (A.27)
\end{aligned}$$

Therefore, by the Tchebychev's inequality and the fact that $\rho_{\mathsf{N}}N=n$, we have

$$\begin{aligned}
&\frac{K}{N}\sum_{i\in\mathcal{I}_k}\left[\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\mathbb{E}_{i\in\mathcal{I}_k}\{\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL,a}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\mid\mathscr{D}_k^c\}\right] \\
=&o_p\left(n^{-1/2}\right). \qquad (A.28)
\end{aligned}$$

Thereafter combining (A.27) and (A.28) to (A.26), we have

$$\widehat{\Delta}^{(\mathsf{k})}_{\mathsf{SMMAL}}=\frac{K}{N}\sum_{i\in\mathcal{I}_k}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})+o_p\left(n^{-1/2}\right).$$

Summing over all the folds, we obtain

$$\widehat{\Delta}_{\mathsf{SMMAL}}=\frac{1}{K}\sum_{k=1}^{K}\widehat{\Delta}^{(\mathsf{k})}_{\mathsf{SMMAL}}=\frac{1}{N}\sum_{i=1}^{N}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})+o_p\left(n^{-1/2}\right). \qquad (A.29)$$

Using (A.25) along with Assumptions 2a and 2b, we therefore obtain that the variance of each summand in (A.29) scales with $1/\rho_{\mathsf{N}}$ as,

$$\mathrm{Var}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\}=\mathcal{V}_*/\rho_{\mathsf{N}}+O(1).$$

Under Assumption 2e, we can then scale it to obtain a stable variance

$$\mathcal{V}_{\mathsf{SMMAL}}=\mathrm{Var}\{\sqrt{\rho_{\mathsf{N}}}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\sqrt{\rho_{\mathsf{N}}}\phi_{\mathsf{SSL,0}}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\}=\mathcal{V}_*+O(\rho_{\mathsf{N}})\in[1/2M,2M] \qquad (A.30)$$

for sufficiently small ρ_{N}. Applying the central limit theorem at $\sqrt{n}$-scale, we have

$$\sqrt{n}(\widehat{\Delta}_{\mathsf{SMMAL}} - \Delta_*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sqrt{\rho_{\mathsf{N}}}\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \bar{\boldsymbol{\eta}}) - \sqrt{\rho_{\mathsf{N}}}\phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \bar{\boldsymbol{\eta}}) + o_p(1) \rightsquigarrow N(0, \mathcal{V}_{\mathsf{SMMAL}}) \tag{A.31}$$

Finally, we can use Lemma A15 again to show the consistency of the variance estimator. To this end, we decompose the variance estimator as

$$\begin{aligned}
\widehat{\mathcal{V}}_{\mathsf{SMMAL}} =& \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2 - \rho_{\mathsf{N}}\widehat{\Delta}^2_{\mathsf{SMMAL}} \\
=& \mathcal{V}_{\mathsf{SMMAL}} + \rho_{\mathsf{N}}(\Delta_*^2 - \widehat{\Delta}^2_{\mathsf{SMMAL}}) \\
&+ \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2 \\
&\quad - \mathbb{E}_{i \in \mathcal{I}_k}[\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2 \mid \mathscr{D}_k^c] \\
&+ \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{i \in \mathcal{I}_k}[\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2 \mid \mathscr{D}_k^c] \\
&\quad - \mathbb{E}[\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \bar{\boldsymbol{\eta}}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \bar{\boldsymbol{\eta}})\}^2 \mid \mathscr{D}_k^c].
\end{aligned} \tag{A.32}$$

By the asymptotic normality of $\widehat{\Delta}_{\mathsf{SMMAL}}$ and the boundedness of Δ_* from Assumption 2a, we have

$$\rho_{\mathsf{N}}(\Delta_*^2 - \widehat{\Delta}^2_{\mathsf{SMMAL}}) = O_p\left(\rho_{\mathsf{N}} n^{-1/2}\right). \tag{A.33}$$

We denote the labeled data component and unlabeled data component of $\phi_{\mathsf{SSL,1}} - \phi_{\mathsf{SSL,0}}$ as

$$\begin{aligned}
\psi_1(\mathbf{D}; \boldsymbol{\eta}) =& \mu(1, \mathbf{X}) + \frac{\Pi(1, \mathbf{X})}{\pi(1, \mathbf{X})} \{m(1, \mathbf{W}) - \mu(1, \mathbf{X})\} \\
&- \mu(0, \mathbf{X}) - \frac{\Pi(0, \mathbf{X})}{\pi(0, \mathbf{X})} \{m(0, \mathbf{W}) - \mu(0, \mathbf{X})\} \\
\psi_2(\mathbf{D}; \boldsymbol{\eta}) =& \frac{\mathrm{I}(A=1)}{\pi(1, \mathbf{X})} \{Y - \mu(1, \mathbf{X})\} - \frac{\Pi(1, \mathbf{X})}{\pi(1, \mathbf{X})} \{m(1, \mathbf{W}) - \mu(1, \mathbf{X})\} \\
&- \frac{\mathrm{I}(A=0)}{\pi(0, \mathbf{X})} \{Y - \mu(0, \mathbf{X})\} + \frac{\Pi(0, \mathbf{X})}{\pi(0, \mathbf{X})} \{m(0, \mathbf{W}) - \mu(0, \mathbf{X})\}.
\end{aligned}$$

Using the identity $R_i^2 = R_i$, we express the term $\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2$ in the generic form analyzed in Lemma A19

$$\begin{aligned}
&\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2 \\
=& \rho_{\mathsf{N}} \left\{ \psi_1(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})}) + \frac{R_i}{\rho_{\mathsf{N}}} \psi_2(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})} \right\}^2 \\
=& \underbrace{\rho_{\mathsf{N}}\psi_1(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})^2 + 2R_i\psi_1(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\psi_2(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})}_{h_1 \text{ for Lemma A19}} + \frac{R_i}{\rho_{\mathsf{N}}} \underbrace{\psi_2(\mathbf{D}_i; \widehat{\boldsymbol{\eta}}^{(\mathsf{k})})^2}_{h_2 \text{ for Lemma A19}}.
\end{aligned}$$

Under Assumptions 2a and 2c, the h_1 and h_2 components are all bounded, so we can apply the concentration result in Lemma A19 to get

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}&\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL},1}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL},0}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2\\&-\mathbb{E}_{i\in\mathcal{I}_k}[\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL},1}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL},0}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2\mid\mathscr{D}_k^c]=O_p\left(n^{-1/2}\right).\end{aligned}\tag{A.34}$$

Also, applying Lemma A15 with Assumption 2d, we have

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}&\mathbb{E}_{i\in\mathcal{I}_k}[\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL},1}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})-\phi_{\mathsf{SSL},0}(\mathbf{D}_i;\widehat{\boldsymbol{\eta}}^{(\mathsf{k})})\}^2\mid\mathscr{D}_k^c]\\&-\mathbb{E}[\rho_{\mathsf{N}}\{\phi_{\mathsf{SSL},1}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})-\phi_{\mathsf{SSL},0}(\mathbf{D}_i;\bar{\boldsymbol{\eta}})\}^2\mid\mathscr{D}_k^c]=o_p(1).\end{aligned}\tag{A.35}$$

Putting the rates (A.33)-(A.35) to (A.32), we have

$$\widehat{\mathcal{V}}_{\mathsf{SMMAL}}=\mathcal{V}_{\mathsf{SMMAL}}+o_p(1).\tag{A.36}$$

With the asymptotic normality (A.31), stable variance (A.30) and consistent variance estimator (A.36), we apply the continuous mapping theorem to get

$$\sqrt{n/\widehat{\mathcal{V}}_{\mathsf{SMMAL}}}(\widehat{\Delta}_{\mathsf{SMMAL}}-\Delta_*)\rightsquigarrow N(0,1).$$

C2 Proof of Theorem 5

Here we directly apply the conclusion of Theorem 13 by verifying Assumption 5. See the proof of Theorem 13 for details. By the Theorem 2.2 Setting IV of Kallus and Mao (2024), the efficient influence function for Δ_* under complete data is ϕ_{cmp}. Now, we verify each item in Assumption 5.

(a) We assume missing completely at random (1) throughout;

(b) We assume the stable variance from labeled portion as Assumption 2e;

(c) We are considering nonparametric model, so $\mathscr{H}$ contains all the mean zero square integrable random variables;

(d) The complete data efficient influence function ϕ_{cmp} is bounded under Assumptions 2a and 2b.

C3 Proof of Corollary 7

We verify the Assumption 2 using existing results for B-spline regression summarized in Lemma A24 (see Newey and Robins, 2018, for example). Under (1), Assumptions 3a and 2b, the densities of $\mathbf{X}\mid R=1$, $\mathbf{X}\mid R=1,A=1$, $\mathbf{X}\mid R=1,A=0$, $\mathbf{W}\mid R=1$, $\mathbf{W}\mid R=1,A=1$ and $\mathbf{W}\mid R=1,A=0$ are all bounded and bounded away from zero. We have the boundedness of Y from Assumption 2a, and A is naturally bounded by one. Under Assumption 3b, all nuisance models are Hölder smooth. Choosing tensor B-splines with equally spaced knots and proper normalization, B-spline regressions for π, μ, Π and m all satisfy the conditions of Lemma A24.

Apply Lemma A24 with the order κ and degrees from the statement of Corollary 7, we have

$$\|\widehat{\pi}^{(\mathsf{k})}(a,\cdot)-\pi_*(a,\cdot)\|_2=O_p\left(n^{-\frac{\mathcal{H}(\pi_*(a,\cdot))/p}{1+\mathcal{H}(\pi_*(a,\cdot))/p}}\right),\ \|\widehat{\mu}^{(\mathsf{k})}(a,\cdot)-\mu_*(a,\cdot)\|_2=O_p\left(n^{-\frac{\mathcal{H}(\mu_*(a,\cdot))/p}{1+\mathcal{H}(\mu_*(a,\cdot))/p}}\right),$$
$$\|\widehat{\Pi}^{(\mathsf{k})}(a,\cdot)-\pi_*(a,\cdot)\|_2=O_p\left(n^{-\frac{\mathcal{H}(\pi_*(a,\cdot))/p}{1+\mathcal{H}(\Pi_*(a,\cdot))/p}}\right),\ \|\widehat{m}^{(\mathsf{k})}(a,\cdot)-\mu_*(a,\cdot)\|_2=O_p\left(n^{-\frac{\mathcal{H}(\mu_*(a,\cdot))/p}{1+\mathcal{H}(m_*(a,\cdot))/p}}\right).$$

Under Assumptions 2a and 2b, truncation at M does not increase estimation error,

$$\begin{aligned}\|\min\{M,\widehat{\pi}^{(\mathsf{k})}(a,\cdot)\}-\pi_*(a,\cdot)\|_2&=O_p\left(n^{-\frac{\mathcal{H}(\pi_*(a,\cdot))/p}{1+\mathcal{H}(\pi_*(a,\cdot))/p}}\right),\\ \|\min\{M,\widehat{\mu}^{(\mathsf{k})}(a,\cdot)\}-\mu_*(a,\cdot)\|_2&=O_p\left(n^{-\frac{\mathcal{H}(\mu_*(a,\cdot))/p}{1+\mathcal{H}(\mu_*(a,\cdot))/p}}\right),\\ \|\min\{M,\widehat{\Pi}^{(\mathsf{k})}(a,\cdot)\}-\pi_*(a,\cdot)\|_2&=O_p\left(n^{-\frac{\mathcal{H}(\pi_*(a,\cdot))/p}{1+\mathcal{H}(\Pi_*(a,\cdot))/p}}\right),\\ \|\min\{M,\widehat{m}^{(\mathsf{k})}(a,\cdot)\}-\mu_*(a,\cdot)\|_2&=O_p\left(n^{-\frac{\mathcal{H}(\mu_*(a,\cdot))/p}{1+\mathcal{H}(m_*(a,\cdot))/p}}\right).\end{aligned}$$

The truncation at M secures Assumption 2c. Under Assumptions 2b and 3b, the rates for the truncated estimators above satisfy Assumption 2d with $\bar{\Pi}=\Pi_*$ and $\bar{m}=m_*$. Therefore, we can apply the asymptotic normality (A.31) from Theorem 2 to get

$$\sqrt{n}(\widehat{\Delta}_{\mathsf{SSL}}-\Delta_*)\rightsquigarrow N(0,\rho_{\mathsf{N}}\operatorname{Var}\{\phi_{\mathsf{SSL}}(RY,RA,\mathbf{W},R)\}).$$

C4 Proof of Theorem 8

In Section D1, we developed the estimation rates for Lasso estimators defined in (9)-(12) in Lemma A21. Two estimation rates from Lemma A21 corresponds to the two situation: 1) the general case in which the cross-fitted parameter must be consistent to identify the target parameter; 2) the special case in which the cross-fitted parameter does not exist or is not needed for identifying the target parameter. The special case applies to the imputation Lasso (9), the initial Lasso (10) and the calibrated Lasso (12) if the underlying model is correct. The general case applies to the calibrated Lasso (12) if the underlying model is wrong. We do not require any concentration of the initial estimator for the mis-specified model following the truncation of the weights in the calibrated Lasso (12). The proofs for Lemma A21 are based on the technique developed in Hou et al. (2021b). In summary, we obtain

1. Both models correct: $\boldsymbol{\alpha}_*=\bar{\boldsymbol{\alpha}}_{\mathsf{init}}=\bar{\boldsymbol{\alpha}}_{\mathsf{a}}$ and $\boldsymbol{\beta}_*=\bar{\boldsymbol{\beta}}_{\mathsf{a,init}}=\bar{\boldsymbol{\beta}}_{\mathsf{a}}$,

$$\|\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})}-\boldsymbol{\alpha}_*\|_2=O_p\left(\sqrt{\|\boldsymbol{\alpha}_*\|_0\log(p)/n}\right),\ \|\widehat{\boldsymbol{\beta}}_{\mathsf{a}}^{(\mathsf{k})}-\boldsymbol{\beta}_{*,\mathsf{a}}\|_2=O_p\left(\sqrt{\|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0\log(p)/n}\right),$$
$$\|\widehat{\boldsymbol{\xi}}^{(\mathsf{k})}-\bar{\boldsymbol{\xi}}\|_2=O_p\left(\sqrt{\|\bar{\boldsymbol{\xi}}\|_0\log(p+q)/n}\right),\ \|\widehat{\boldsymbol{\zeta}}_{\mathsf{a}}^{(\mathsf{k})}-\bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_2=O_p\left(\sqrt{\|\bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_0\log(p+q)/n}\right).$$

2. PS model correct: $\boldsymbol{\alpha}_* = \bar{\boldsymbol{\alpha}}_{\mathsf{init}} = \bar{\boldsymbol{\alpha}}_{\mathsf{a}}$,

$$\|\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} - \boldsymbol{\alpha}_*\|_2 = O_p\left(\sqrt{\|\boldsymbol{\alpha}_*\|_0 \log(p)/n}\right),$$
$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\beta}}_{\mathsf{a}}\|_2 = O_p\left(\sqrt{\left(\|\bar{\boldsymbol{\beta}}_{\mathsf{a}}\|_0 + \|\boldsymbol{\alpha}_*\|_0\right)\log(p)/n}\right),$$
$$\|\widehat{\boldsymbol{\xi}}^{(\mathsf{k})} - \bar{\boldsymbol{\xi}}\|_2 = O_p\left(\sqrt{\|\bar{\boldsymbol{\xi}}\|_0 \log(p+q)/n}\right),\ \|\widehat{\boldsymbol{\zeta}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_2 = O_p\left(\sqrt{\|\bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_0 \log(p+q)/n}\right).$$

3. OR model correct: $\boldsymbol{\beta}_* = \bar{\boldsymbol{\beta}}_{\mathsf{a,init}} = \bar{\boldsymbol{\beta}}_{\mathsf{a}}$,

$$\|\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\alpha}}_{\mathsf{a}}\|_2 = O_p\left(\sqrt{\left(\|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0 + \|\bar{\boldsymbol{\alpha}}_{\mathsf{a}}\|_0\right)\log(p)/n}\right),$$
$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{a}}^{(\mathsf{k})} - \boldsymbol{\beta}_{*,\mathsf{a}}\|_2 = O_p\left(\sqrt{\|\boldsymbol{\beta}_{*,\mathsf{a}}\|_0 \log(p)/n}\right),$$
$$\|\widehat{\boldsymbol{\xi}}^{(\mathsf{k})} - \bar{\boldsymbol{\xi}}\|_2 = O_p\left(\sqrt{\|\bar{\boldsymbol{\xi}}\|_0 \log(p+q)/n}\right),\ \|\widehat{\boldsymbol{\zeta}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_2 = O_p\left(\sqrt{\|\bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_0 \log(p+q)/n}\right).$$

Similar to the proof of Theorem 2, we group the limiting nuisance models (15) as

$$\begin{aligned}\bar{\boldsymbol{\eta}} = (\bar{\pi}, \bar{\mu}, \bar{\Pi}, \bar{m}),\ \bar{\pi}(a, \mathbf{X}) = a g(\bar{\boldsymbol{\alpha}}_1^{\mathsf{T}}\mathbf{X}) + (1-a) g(-\bar{\boldsymbol{\alpha}}_0^{\mathsf{T}}\mathbf{X}),\\ \bar{\mu}(a, \mathbf{X}) = g(\bar{\boldsymbol{\alpha}}_{\mathsf{a}}^{\mathsf{T}}\mathbf{X}),\ \bar{\Pi}(a, \mathbf{W}) = g(\bar{\boldsymbol{\xi}}^{\mathsf{T}}\mathbf{W}),\ \bar{m}(a, \mathbf{W}) = g(-\bar{\boldsymbol{\zeta}}_{\mathsf{a}}^{\mathsf{T}}\mathbf{W}).\end{aligned} \tag{A.37}$$

The L_2 estimation rates translate to the mean square error rate of the model estimators (13) by Lemma A20,

$$\begin{aligned}\|\widehat{\pi}^{(\mathsf{k})}(a, \cdot) - \bar{\pi}(a, \cdot)\|_2 \lesssim \|\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\alpha}}_{\mathsf{a}}\|_2,\ \|\widehat{\mu}^{(\mathsf{k})}(a, \cdot) - \bar{\mu}(a, \cdot)\|_2 \lesssim \|\widehat{\boldsymbol{\beta}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\beta}}_{\mathsf{a}}\|_2,\\ \|\widehat{\Pi}^{(\mathsf{k})}(a, \cdot) - \bar{\Pi}(a, \cdot)\|_2 \lesssim \|\widehat{\boldsymbol{\xi}}^{(\mathsf{k})} - \bar{\boldsymbol{\xi}}\|_2,\ \|\widehat{m}^{(\mathsf{k})}(a, \cdot) - \bar{m}(a, \cdot)\|_2 \lesssim \|\widehat{\boldsymbol{\zeta}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\zeta}}_{\mathsf{a}}\|_2.\end{aligned}$$

For the case of both models being correct in Assumption 4d-iii, we can directly apply Theorem 2. We study the other cases in which one of the PS or OR is correct in the rest of the proof.

We use the $\phi_{\mathsf{SSL,a}}(\mathbf{D}; \boldsymbol{\eta})$ notation defined in (A.21). Notice that

$$\begin{aligned}\mathbb{E}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}; \bar{\boldsymbol{\eta}}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}; \bar{\boldsymbol{\eta}})\} =& \mathbb{E}\left[g(\mathbf{X}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_1) + \frac{A}{g(\mathbf{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}_1)}\{Y - g(\mathbf{X}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_1)\}\right]\\ &- \mathbb{E}\left[g(\mathbf{X}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_0) + \frac{1-A}{g(-\mathbf{X}^{\mathsf{T}}\bar{\boldsymbol{\alpha}}_0)}\{Y - g(\mathbf{X}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_0)\}\right]\end{aligned}$$

so $\Delta_* = \mathbb{E}\{\phi_{\mathsf{SSL,1}}(\mathbf{D}; \bar{\boldsymbol{\eta}}) - \phi_{\mathsf{SSL,0}}(\mathbf{D}; \bar{\boldsymbol{\eta}})\}$ if either the PS or OR is correct (Bang and Robins, 2005). We state a modified version of Lemma A15.

Lemma A16 *Let $\boldsymbol{\eta}_n = (\pi_n, \mu_n, \Pi_n, m_n)$ be a (deterministic) sequence of nuisance models satisfying almost surely*

$$\sup_{a=0,1} \max\left\{|1/\pi_n(a, \mathbf{X}_i)|, |\mu_n(a, \mathbf{X}_i)|, |\Pi_n(a, \mathbf{W}_i)|, |m_n(a, \mathbf{W}_i)|\right\} \le M. \tag{A.38}$$

Define

$$\Psi_1(\mu_n) = \mathbb{E}\left[\{\mu_n(a,\mathbf{X}) - \bar{\mu}(a,\mathbf{X})\}\left\{\frac{\mathrm{I}(A=a)}{\bar{\pi}(a,\mathbf{X})} - 1\right\}\right],$$

$$\Psi_2(\pi_n) = \mathbb{E}\left[\left\{\frac{1}{\pi_n(a,\mathbf{X})} - \frac{1}{\bar{\pi}(a,\mathbf{X})}\right\}\mathrm{I}(A=a)\{Y - \bar{\mu}(a,\mathbf{X})\}\right]. \tag{A.39}$$

Under Assumptions 2a and 2b, we have

1. *For bias:*

$$|\mathbb{E}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n) - \phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}| \lesssim \|\pi_n - \bar{\pi}\|_2\|\mu_n - \bar{\mu}\|_2 + \Psi_1(\mu_n) + \Psi_2(\pi_n),$$

2. *For variance:*

$$\begin{aligned}&\rho_{\mathsf{N}}\operatorname{Var}\{\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n) - \phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})\}\\ \lesssim&\|\pi_n - \bar{\pi}\|_2^2 + \|\mu_n - \bar{\mu}\|_2^2 + \|m_n - \bar{m}\|_2^2 + \|\Pi_n - \bar{\Pi}\|_2^2,\end{aligned}$$

3. *For variance estimation:*

$$\begin{aligned}&\mathbb{E}\left\{\rho_{\mathsf{N}}\phi_{\mathsf{SSL,a}}(\mathbf{D};\boldsymbol{\eta}_n)^2\right\} - \mathbb{E}\left\{\rho_{\mathsf{N}}\phi_{\mathsf{SSL,a}}(\mathbf{D};\bar{\boldsymbol{\eta}})^2\right\}\\ \lesssim&\|\pi_n - \bar{\pi}\|_2 + \|\mu_n - \bar{\mu}\|_2 + \|m_n - \bar{m}\|_2 + \|\Pi_n - \bar{\Pi}\|_2\\ &+ \|\pi_n - \bar{\pi}\|_2^2 + \|\mu_n - \bar{\mu}\|_2^2 + \|m_n - \bar{m}\|_2^2 + \|\Pi_n - \bar{\Pi}\|_2^2.\end{aligned}$$

We omit the proof of Lemma A16 as it merely repeats that of Lemma A15. The only difference is that limiting models $\bar{\pi}$ and $\bar{\mu}$ can deviate from the truth π_* and μ_*, so we have the extra terms (A.39) in the bias representation. In the next lemma, we study (A.39) under Assumption 4d-i or 4d-ii.

Lemma A17 *Let π_n and μ_n be the logistic regression predictions*

$$\pi_n(a,\mathbf{X}) = g_\tau((-1)^{a+1}\mathbf{X}^\intercal\boldsymbol{\alpha}_n),\ \mu_n(a,\mathbf{X}) = g((-1)^{a+1}\mathbf{X}^\intercal\boldsymbol{\beta}_n).$$

We have

1. *under Assumption 4d-i: for $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2 \leq 1/(2M)$,*

$$\Psi_1(\mu_n) = 0,\ \Psi_2(\pi_n) \lesssim \|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2^2 + e^{-1/(2\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2^2)};$$

2. *under Assumption 4d-ii:*

$$\Psi_2(\pi_n) = 0,\ \Psi_1(\mu_n) \lesssim \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{*,\mathsf{a}}\|_2^2.$$

Proof [Proof of Lemma A17]
<u>PS correct:</u> We have $\bar{\boldsymbol{\alpha}}_{\mathsf{a}} = \boldsymbol{\alpha}_*$ and $\bar{\pi} = \pi_*$. As the result, we have

$$\Psi_1(\mu_n) = \mathbb{E}\left[\{\mu_n(a,\mathbf{X}) - \bar{\mu}(a,\mathbf{X})\}\frac{\mathbb{E}\{\mathrm{I}(A=a) - \pi_*(a,\mathbf{X}) \mid \mathbf{X}\}}{\pi_*(a,\mathbf{X})}\right] = 0.$$

In the following, we set $a = 1$, while the $a = 0$ case can be obtained by the same steps. To analyze $\Psi_2(\pi_n)$, we consider the following decomposition

$$\begin{aligned}\Psi_2(\pi_n) =& \mathbb{E}\left[\{\exp_\tau(-\mathbf{X}^\intercal\boldsymbol{\alpha}_n) - \exp(-\mathbf{X}^\intercal\boldsymbol{\alpha}_n)\} A\{Y - \bar{\mu}(1,\mathbf{X})\}\right] \\ & - (\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)^\intercal \mathbb{E}\left[\exp(-\mathbf{X}^\intercal\boldsymbol{\alpha}_*)\mathbf{X}A\{Y - \bar{\mu}(1,\mathbf{X})\}\right] \\ & + \frac{1}{2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)^\intercal \mathbb{E}\left[\exp(-\mathbf{X}^\intercal\widetilde{\boldsymbol{\alpha}})\mathbf{X}\mathbf{X}^\intercal A\{Y - \bar{\mu}(1,\mathbf{X})\}\right](\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*) \end{aligned} \tag{A.40}$$

for some $\widetilde{\boldsymbol{\alpha}}$ between $\boldsymbol{\alpha}_n$ and $\boldsymbol{\alpha}_*$. Under Assumption 4c, $|\mathbf{X}^\intercal\boldsymbol{\alpha}_*| \le M$ the truncation at $2M$ for $\mathbf{X}^\intercal\boldsymbol{\alpha}_n$ would only be triggered if $|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)| \ge M$. Thus, we may derive the following upper bound for the truncation error,

$$\begin{aligned} & \left|\mathbb{E}\left[\{\exp_\tau(-\mathbf{X}^\intercal\boldsymbol{\alpha}_n) - \exp(-\mathbf{X}^\intercal\boldsymbol{\alpha}_n)\} A\{Y - \bar{\mu}(1,\mathbf{X})\}\right]\right| \\ \le& \mathbb{E}\left[\exp(|\mathbf{X}^\intercal\boldsymbol{\alpha}_n|)\mathrm{I}(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)| \ge M)\right] \\ \le& \mathbb{E}\left[e^M \exp(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)|)\mathrm{I}(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)| \ge M)\right] \\ \le& e^M\sqrt{\mathbb{E}\left\{\exp(2|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)|)\right\}\mathbb{P}(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)| \ge M)} \\ \le& e^M\sqrt{\mathbb{E}\left\{\exp\left(\frac{|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)|}{M\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2}\right)\right\}\mathbb{P}(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)| \ge M)}. \end{aligned} \tag{A.41}$$

The last inequality above follows from the assumption that $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2 \le 1/(2M)$. Under Assumption 4a, $\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)$ is sub-Gaussian thus sub-exponential

$$\|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)\|_{\psi_1} \le \|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)\|_{\psi_2} \le M\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2,$$

so we may apply the definition of sub-Gaussian/sub-exponential random variable and its property in tail probability to get

$$\mathbb{E}\left\{\exp\left(\frac{|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)|}{M\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2}\right)\right\} \le 2,\ \mathbb{P}(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)| \ge M) \le 2e^{-1/\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2^2},. \tag{A.42}$$

Applying (A.42) to (A.41), we have

$$\left|\mathbb{E}\left[\{\exp_\tau(-\mathbf{X}^\intercal\boldsymbol{\alpha}_n) - \exp(-\mathbf{X}^\intercal\boldsymbol{\alpha}_n)\} A\{Y - \bar{\mu}(1,\mathbf{X})\}\right]\right| \lesssim e^{-1/(2\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2^2)}. \tag{A.43}$$

By the definition $\bar{\boldsymbol{\beta}}_{\mathsf{a}}$ (15), they must satisfy the first order condition of optimality

$$\mathbb{E}\left[\exp(-\mathbf{X}^\intercal\bar{\boldsymbol{\alpha}}_{\mathsf{init}})A\mathbf{X}\{Y - g(\mathbf{X}^\intercal\bar{\boldsymbol{\beta}}_{\mathsf{1}})\right] = 0. \tag{A.44}$$

By the definition of $\bar{\mu}$ and the fact $\bar{\boldsymbol{\alpha}}_{\mathsf{init}} = \boldsymbol{\alpha}_*$ under correct OR model, we infer from (A.44),

$$\mathbb{E}\left[\exp(-\mathbf{X}^\intercal\boldsymbol{\alpha}_*)\mathbf{X}A\{Y - \bar{\mu}(1,\mathbf{X})\}\right] = 0. \tag{A.45}$$

We bound the quadratic term in (A.40) with Assumptions 4a, 4c,

$$\begin{aligned} & \left|\frac{1}{2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)^\intercal \mathbb{E}\left[\exp(-\mathbf{X}^\intercal\widetilde{\boldsymbol{\alpha}})\mathbf{X}\mathbf{X}^\intercal A\{Y - \bar{\mu}(1,\mathbf{X})\}\right](\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)\right| \qquad \text{(A.46)} \\ \lesssim& \mathbb{E}[e^M \exp(|\mathbf{X}^\intercal(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)|)\{(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*)^\intercal\mathbf{X}\}^2] \end{aligned}$$

$$\begin{aligned}
&\lesssim e^{M}\sqrt{\mathbb{E}\left\{\exp\left(\frac{|\mathbf{X}^{\intercal}(\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_*)|^2}{M^2\|\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_*\|_2^2}\right)\right\}\mathbb{E}[\{(\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_*)^{\intercal}\mathbf{X}\}^4]}\\
&\lesssim\|\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_*\|_2^2. \qquad\qquad (A.47)
\end{aligned}$$

We have used again the sub-Gaussian property of (A.42) in the last inequality. Applying A.43, (A.45) and (A.47) to (A.40), we have shown

$$\Psi_2(\pi_n)\lesssim\|\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_*\|_2^2+e^{-1/(2\|\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_*\|_2^2)}.$$

<u>OR correct:</u> We have $\bar{\boldsymbol{\beta}}_{\mathsf{a}}=\boldsymbol{\beta}_{*,\mathsf{a}}$ and $\bar{\mu}=\mu_*$. As the result, we have

$$\Psi_2(\pi_n)=\mathbb{E}\left[\left\{\frac{1}{\pi_n(a,\mathbf{X})}-\frac{1}{\bar{\pi}(a,\mathbf{X})}\right\}\mathrm{I}(A=a)\mathbb{E}\{Y-\mu_*(a,\mathbf{X})\mid A,\mathbf{X}\}\right]=0.$$

Using the Mean Value Theorem on $\Psi_1(\mu_n)$, we have

$$\begin{aligned}
\Psi_1(\mu_n)=&(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}})^{\intercal}\mathbb{E}\left[\dot{g}(\mathbf{X}^{\intercal}\boldsymbol{\beta}_{*,\mathsf{a}})\mathbf{X}\left\{\frac{\mathrm{I}(A=a)}{\bar{\pi}(a,\mathbf{X})}-1\right\}\right]\\
&+\frac{1}{2}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}})^{\intercal}\mathbb{E}\left[g''(\mathbf{X}^{\intercal}\widetilde{\boldsymbol{\beta}})\mathbf{X}\mathbf{X}^{\intercal}\left\{\frac{\mathrm{I}(A=a)}{\bar{\pi}(a,\mathbf{X})}-1\right\}\right](\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}}) \qquad (A.48)
\end{aligned}$$

for some $\widetilde{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_{*,\mathsf{a}}$. By the definition $\bar{\boldsymbol{\alpha}}_{\mathsf{a}}$ (15), they must satisfy the first order condition of optimality

$$\begin{aligned}
\mathbb{E}\left[\dot{g}(\mathbf{X}^{\intercal}\bar{\boldsymbol{\beta}}_{1,\mathsf{init}})\{1-A(1+e^{-\mathbf{X}^{\intercal}\bar{\boldsymbol{\alpha}}_1})\}\right]&=0,\\
\mathbb{E}\left[\dot{g}(\mathbf{X}^{\intercal}\bar{\boldsymbol{\beta}}_{0,\mathsf{init}})\{1-(1-A)(1+e^{\mathbf{X}^{\intercal}\bar{\boldsymbol{\alpha}}_0})\}\right]&=0. \qquad (A.49)
\end{aligned}$$

By the definition of $\bar{\pi}$ and the fact $\bar{\boldsymbol{\beta}}_{\mathsf{a,init}}=\boldsymbol{\beta}_{*,\mathsf{a}}$ under correct OR model, we infer from (A.49),

$$\mathbb{E}\left[\dot{g}(\mathbf{X}^{\intercal}\boldsymbol{\beta}_{*,\mathsf{a}})\mathbf{X}\left\{\frac{\mathrm{I}(A=a)}{\bar{\pi}(a,\mathbf{X})}-1\right\}\right]=0. \qquad (A.50)$$

We bound the quadratic term in (A.48) with Assumptions 4a, 4c and bounds for $\|g''\|_\infty\leq 1/(6\sqrt{3})$,

$$\begin{aligned}
\left|\frac{1}{2}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}})^{\intercal}\mathbb{E}\left[g''(\mathbf{X}^{\intercal}\widetilde{\boldsymbol{\beta}})\mathbf{X}\mathbf{X}^{\intercal}\left\{\frac{\mathrm{I}(A=a)}{\bar{\pi}(a,\mathbf{X})}-1\right\}\right](\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}})\right|&\lesssim\mathbb{E}[\{(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}})^{\intercal}\mathbf{X}\}^2]\\
&\lesssim\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}}\|_2^2. \qquad (A.51)
\end{aligned}$$

Applying (A.50) and (A.51) to (A.48), we have shown

$$\Psi_1(\mu_n)\lesssim\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{*,\mathsf{a}}\|_2^2.$$

■

As Lemma A17 require $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_*\|_2$ to be sufficiently small, we create the hypothetical error truncated estimators for $\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})}$,

$$\check{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} = \bar{\boldsymbol{\alpha}}_{\mathsf{a}} + (\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\alpha}}_{\mathsf{a}}) \min\left\{1, \frac{1}{M\|\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\alpha}}_{\mathsf{a}}\|_2}\right\}. \tag{A.52}$$

Under Assumption 4d, we have $\|\widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} - \bar{\boldsymbol{\alpha}}_{\mathsf{a}}\|_2 = o_p(1)$, so $\check{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})} = \widehat{\boldsymbol{\alpha}}_{\mathsf{a}}^{(\mathsf{k})}$ with large probability. From here, we repeat the proof of Theorem 2 after Lemma A15 and obtain

$$\sqrt{n/\widehat{\mathcal{V}}_{\mathsf{DR}}}(\widehat{\Delta}_{\mathsf{DR}} - \Delta_*) \rightsquigarrow N(0,1).$$

C5 Proof of Proposition 12

The proof runs in two steps. First, we show that $\boldsymbol{\psi}_{\mathsf{SSL}}$ is an influence function for $\boldsymbol{\theta}$ under $\mathcal{S}_{\mathsf{SSL}}$ using the property of $\boldsymbol{\psi}_{\mathsf{cmp}}$ being the influence function for $\boldsymbol{\theta}$ under $\mathcal{S}_{\mathsf{cmp}}$. Second, we show that $\boldsymbol{\psi}_{\mathsf{SSL}}$ is efficient in the sense that it belongs to the tangent space of model $\mathcal{S}_{\mathsf{SSL}}$.

Consider a generic parametric sub-model for complete data and SSL data as follows

$$\begin{aligned}\mathcal{S}_{\mathsf{cmp}}(\boldsymbol{\eta}) &= \{f(\mathbf{z},\mathbf{w};\boldsymbol{\eta}) : \boldsymbol{\eta} \in \mathbb{R}^p\} \subset \mathcal{S}_{\mathsf{cmp}},\\ f_{\mathbf{W}}(\mathbf{w};\boldsymbol{\eta}) &= \int f_{\mathbf{Z},\mathbf{W}}(\mathbf{z},\mathbf{w};\boldsymbol{\eta}) d\nu_z(\mathbf{z}),\\ \mathcal{S}_{\mathsf{SSL}}(\boldsymbol{\eta}) &= \left\{[\rho_{\mathsf{N}} f(\mathbf{z},\mathbf{w};\boldsymbol{\eta})]^r \left[(1-\rho_{\mathsf{N}}) f_{\mathbf{W}}(\mathbf{w};\boldsymbol{\eta})\right]^{(1-r)} d\nu_{\mathsf{SSL}}(r,\mathbf{z},\mathbf{w}) : \boldsymbol{\eta} \in \mathbb{R}^p\right\} \subset \mathcal{S}_{\mathsf{SSL}}.\end{aligned} \tag{A.53}$$

The true model is attained at $f(\mathbf{z},\mathbf{w};\boldsymbol{\eta}^*) = f_*(\mathbf{z},\mathbf{w})$. We denote the score function from the complete data as

$$\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) = \frac{\partial}{\partial\boldsymbol{\eta}} \log\{f(\mathbf{Z},\mathbf{W};\boldsymbol{\eta})\}\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}. \tag{A.54}$$

The score function under the SSL model $\mathcal{S}_{\mathsf{SSL}}$ can be expressed as

$$\begin{aligned}\boldsymbol{\Psi}_{\mathsf{SSL}}(R,\mathbf{Z},\mathbf{W}) =& R\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) + (1-R)\frac{\partial}{\partial\boldsymbol{\eta}} \log\{f_{\mathbf{W}}(\mathbf{W};\boldsymbol{\eta})\}\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}\\ =& R\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) + (1-R)\mathbb{E}_*\{\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) \mid \mathbf{W}\}\end{aligned} \tag{A.55}$$

Since $\boldsymbol{\psi}_{\mathsf{cmp}}$ is the influence function for $\boldsymbol{\theta}$, we must have

$$\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\} = \frac{\partial}{\partial\boldsymbol{\eta}}\boldsymbol{\theta}(\boldsymbol{\eta})\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}. \tag{A.56}$$

Under Assumption 5a, we can calculate

$$\begin{aligned}&\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{SSL}}(R,\mathbf{Z},\mathbf{W})\boldsymbol{\Psi}_{\mathsf{SSL}}(R,\mathbf{Z},\mathbf{W})\}\\ =&\mathbb{E}_*\left\{\frac{R}{\rho_{\mathsf{N}}}\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\right\}\\ &+\mathbb{E}_*\left(\frac{R}{\rho_{\mathsf{N}}}\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) \mid \mathbf{W}\}\left[\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) - \mathbb{E}_*\{\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}) \mid \mathbf{W}\}\right]\right)\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}_*\left\{(R/\rho_{\mathsf{N}}-1)\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}\mathbb{E}_*\{\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}\right\}\\
=&\mathbb{E}_*\left\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\boldsymbol{\Psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\right\}\\
=&\frac{\partial}{\partial\boldsymbol{\eta}}\boldsymbol{\theta}(\boldsymbol{\eta})\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}. \qquad\qquad (A.57)
\end{aligned}
$$

Thus, we verify that $\boldsymbol{\psi}_{\mathsf{SSL}}$ is an influence function for $\boldsymbol{\theta}$ under model $\mathcal{S}_{\mathsf{SSL}}$.

Now, we prove that $\boldsymbol{\psi}_{\mathsf{SSL}}$ belongs to the maximal nonparametric tangent space under model $\mathcal{S}_{\mathsf{SSL}}$. From the Assumption 5c and the efficient influence function $\boldsymbol{\psi}_{\mathsf{cmp}}$, we know two elements in the tangent space under model $\mathcal{S}_{\mathsf{cmp}}$,

$$
\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}),\ \mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}. \qquad (A.58)
$$

According to the connection between the scores for $\mathcal{S}_{\mathsf{cmp}}$ and $\mathcal{S}_{\mathsf{SSL}}$ (A.55), we obtain two elements in the tangent space under model $\mathcal{S}_{\mathsf{SSL}}$,

$$
\begin{aligned}
\mathbf{U}&=R\left[\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})-\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}\right]+\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\},\\
\mathbf{V}&=\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}. \qquad (A.59)
\end{aligned}
$$

Notice that we can express $\boldsymbol{\psi}_{\mathsf{SSL}}$ as the linear combination of the two elements above

$$
\boldsymbol{\psi}_{\mathsf{SSL}}(R,\mathbf{Z},\mathbf{W})=\mathbf{U}/\rho_{\mathsf{N}}+\mathbf{V}(1-1/\rho_{\mathsf{N}}). \qquad (A.60)
$$

Since the tangent space is a linear space, $\boldsymbol{\psi}_{\mathsf{SSL}}$ must also be an element in the tangent space.

We have shown that $\boldsymbol{\psi}_{\mathsf{SSL}}$ is an element in the tangent space of $\mathcal{S}_{\mathsf{SSL}}$ satisfying (A.57). Therefore, $\boldsymbol{\psi}_{\mathsf{SSL}}$ is the efficient influence function for $\boldsymbol{\theta}$ under $\mathcal{S}_{\mathsf{SSL}}$.

C6 Proof of Theorem 13

Suppose the dimension of $\boldsymbol{\theta}$ is q. We start with the construction of the $2q$-dimensional least favorable model. From the Assumption 5c and the efficient influence function $\boldsymbol{\psi}_{\mathsf{cmp}}$, we know two elements in the nuisance parameter tangent space under model $\mathcal{S}_{\mathsf{SSL}}$,

$$
\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W}),\ \mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}. \qquad (A.61)
$$

We set the two tilt directions as

$$
\begin{aligned}
\mathbf{g}_1(\mathbf{Z},\mathbf{W})&=\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})-\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\},\\
\mathbf{g}_2(\mathbf{W})&=\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}. \qquad (A.62)
\end{aligned}
$$

We denote the variances of two directions as

$$
\begin{aligned}
\mathcal{V}_{\psi,1}&=\mathrm{Var}_*\{\mathbf{g}_1(\mathbf{Z},\mathbf{W})\}=\mathbb{E}_*[\mathrm{Var}_*\{\mathbf{g}_1(\mathbf{Z},\mathbf{W})\}],\\
\mathcal{V}_{\psi,2}&=\mathrm{Var}_*\{\mathbf{g}_2(\mathbf{W})\}=\mathrm{Var}_*[\mathbb{E}_*\{\mathbf{g}_1(\mathbf{Z},\mathbf{W})\}],\\
\mathcal{V}_{\psi}&=\begin{pmatrix}\mathcal{V}_{\psi,1} & \mathbb{O}_q\\ \mathbb{O}_q & \mathcal{V}_{\psi,2}\end{pmatrix}. \qquad (A.63)
\end{aligned}
$$

Denote

$$
\left[1+\frac{\mathbf{h}_1^{\intercal}\mathbf{g}_1(\mathbf{z}_i,\mathbf{w}_i)}{\sqrt{\rho_{\mathsf{N}}N}}+\frac{\mathbf{h}_2^{\intercal}\mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right]_+=\max\left\{0,1+\frac{\mathbf{h}_1^{\intercal}\mathbf{g}_1(\mathbf{z}_i,\mathbf{w}_i)}{\sqrt{\rho_{\mathsf{N}}N}}+\frac{\mathbf{h}_2^{\intercal}\mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right\},
$$

$$C_{\mathbf{h}} = \mathbb{E}_* \left(\left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}} \right]_+ \right),$$

we construct the two-way tilted density as

$$f_{\mathbf{h}}(\mathbf{z}_i, \mathbf{w}_i) = f_*(\mathbf{z}_i, \mathbf{w}_i) \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} \right]_+ / C_{\mathbf{h}}.$$

In Lemma A25, we proved that the two-way tilted density falls in the neighborhood in $\|\cdot\|_{\mathsf{TV}}$ under Assumption 5d.

$$\|f_* - f_{\mathbf{h}}\|_{\mathsf{TV}} \le M \sqrt{\|\mathbf{h}_1\|_2^2 / \rho_{\mathsf{N}} N + \|\mathbf{h}_2\|_2^2 / N} + o(\|\mathbf{h}\|_2 / \sqrt{\rho_{\mathsf{N}} N}). \tag{A.64}$$

Thus for $\|\mathbf{h}_1\|_2^2 + \|\mathbf{h}_2\|_2^2 \le c^2$ and sufficiently large $\rho_{\mathsf{N}} N$ (as the expected number of labels grows asymptotically to infinity, $\rho_{\mathsf{N}} N \to \infty$), we have

$$\left\| f_* - f_* \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2}{\sqrt{N}} \right]_+ \right\|_{\mathsf{TV}} \le 2Mc / \sqrt{\rho_{\mathsf{N}} N}.$$

We may consider the relaxed minimax problem to the $2q$-dimensional least favorable model:

$$aMSE = \liminf_{c \to \infty} \liminf_{N \to \infty} \sup_{\|\mathbf{h}_1\|_2^2 + \|\mathbf{h}_2\|_2^2 \le c^2} \int \rho_{\mathsf{N}} N \{\mathbf{a}^\intercal (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}^2 d \prod_{i=1}^N \mathbb{P}_{\mathbf{h}}(r_i, \mathbf{z}_i, \mathbf{w}_i) \tag{A.65}$$

where $d\mathbb{P}_{\mathbf{h}^\intercal \mathbf{g}}$ is the tilted distribution

$$\begin{aligned} &d\mathbb{P}_{\mathbf{h}}(r_i, \mathbf{z}_i, \mathbf{w}_i) \\ =&\rho_{\mathsf{N}} \left(f_* \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2}{\sqrt{N}} \right]_+ \right) (\mathbf{z}_i, \mathbf{w}_i) / C_{\mathbf{h}} d\nu_{\mathsf{cmp}}(\mathbf{z}_i, \mathbf{w}_i) \times \delta_1(r_i) \\ &+ \int_{\mathbf{z}} (1 - \rho_{\mathsf{N}}) \left(f_* \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2}{\sqrt{N}} \right]_+ \right) (\mathbf{z}_i, \mathbf{w}_i) / C_{\mathbf{h}} d\nu_{\mathsf{cmp}}(\mathbf{z}_i, \mathbf{w}_i) \times \delta_0(r_i). \end{aligned} \tag{A.66}$$

To simplify the notation, we invoke Assumption 5d and drop the truncation at zero for $\sqrt{\rho_{\mathsf{N}} N} > 2cM$,

$$\begin{aligned} &d\mathbb{P}_{\mathbf{h}}(r_i, \mathbf{z}_i, \mathbf{w}_i) \\ =&\rho_{\mathsf{N}} \left(f_* \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2}{\sqrt{N}} \right] \right) (\mathbf{z}_i, \mathbf{w}_i) d\nu_{\mathsf{cmp}}(\mathbf{z}_i, \mathbf{w}_i) \times \delta_1(r_i) \\ &+ \int_{\mathbf{z}} (1 - \rho_{\mathsf{N}}) \left(f_* \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1}{\sqrt{\rho_{\mathsf{N}} N}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2}{\sqrt{N}} \right] \right) (\mathbf{z}_i, \mathbf{w}_i) d\nu_{\mathsf{cmp}}(\mathbf{z}_i, \mathbf{w}_i) \times \delta_0(r_i). \end{aligned} \tag{A.67}$$

Notice that the tilted data distribution $d\mathbb{P}_{\mathbf{h}}$ has two components: 1) the model is restricted to the least favorable model; 2) the neighborhood along the direction of $\mathbf{g}_2$ is narrowed to $c/\sqrt{N}$. The representation would hold approximately with an error $2Mc/\sqrt{\rho_{\mathsf{N}} N}$ without Assumption 5d, following the approximation of total variation established in Lemma A25.

The design of our least favorable model leads to a factorization of the tilted model

$$
\begin{aligned}
& f_*(\mathbf{z},\mathbf{w})\left\{1+\frac{\mathbf{h}_1^{\top}\mathbf{g}_1(\mathbf{z},\mathbf{w})}{\sqrt{\rho_{\mathsf{N}}N}}+\frac{\mathbf{h}_2^{\top}\mathbf{g}_2(\mathbf{w})}{\sqrt{N}}\right\} \\
=& f^*_{\mathbf{Z}\mid\mathbf{W}}(\mathbf{z}\mid\mathbf{w})\left\{1+\frac{\mathbf{h}_1^{\top}\mathbf{g}_1(\mathbf{z},\mathbf{w})}{\sqrt{\rho_{\mathsf{N}}N}}\right\} f^*_{\mathbf{W}}(\mathbf{w})\left\{1+\frac{\mathbf{h}_2^{\top}\mathbf{g}_2(\mathbf{w})}{\sqrt{N}}\right\}+O\left(1/(N\sqrt{\rho_{\mathsf{N}}})\right), \\
& f^*_{\mathbf{W}}(\mathbf{w})=\int_{\mathbf{z}\in\mathcal{Z}} f_*(\mathbf{z},\mathbf{w})d\nu_z(\mathbf{z}),\ f^*_{\mathbf{Z}\mid\mathbf{W}}(\mathbf{z}\mid\mathbf{w})=f_*(\mathbf{z},\mathbf{w})/f^*_{\mathbf{W}}(\mathbf{w}).
\end{aligned}
$$

The factorization is also reflected in the decomposition of the log-likelihood ratio. By the definition of $\mathbf{g}_1$ and $\mathbf{g}_2$, we have the identities

$$
\begin{aligned}
\mathbb{E}_*\{\mathbf{g}_1(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\} &= \mathbb{E}_*[\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})-\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mid\mathbf{W}\}\mid\mathbf{W}]=\mathbf{0}, \\
\mathbb{E}_*\{\mathbf{g}_1(\mathbf{Z},\mathbf{W})\mathbf{g}_2(\mathbf{W})\} &= \mathbb{O}_{q\times q}.
\end{aligned} \tag{A.68}
$$

For sufficiently large n and N, we have

$$
\begin{aligned}
& \log\left(\prod_{i=1}^{N}\frac{d\mathbb{P}_{\mathbf{h}}(R_i,\mathbf{Z}_i,\mathbf{W}_i)}{d\mathbb{P}_0(R_i,\mathbf{Z}_i,\mathbf{W}_i)}\right) \\
=& \sum_{i=1}^{N} R_i\log\left(1+\frac{\mathbf{h}_1^{\top}\mathbf{g}_1(\mathbf{Z}_i,\mathbf{W}_i)}{\sqrt{\rho_{\mathsf{N}}N}}+\frac{\mathbf{h}_2^{\top}\mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right) \\
& +\sum_{i=1}^{N}(1-R_i)\log\left(1+\frac{\mathbf{h}_2^{\top}\mathbf{g}_2(\mathbf{w})}{\sqrt{N}}\right) \\
& +\sum_{i=1}^{N}(1-R_i)\log\left(\int\left\{1+\frac{\mathbf{h}_1^{\top}\mathbf{g}_1(\mathbf{z},\mathbf{w})}{\sqrt{\rho_{\mathsf{N}}N}}\right\}f_*(\mathbf{z},\mathbf{W}_i)d\nu_z(\mathbf{z})\right) \\
& -\sum_{i=1}^{N}(1-R_i)\log\left(\int f_*(\mathbf{z},\mathbf{W}_i)d\mu(\mathbf{z})\right) \\
=& (\rho_{\mathsf{N}}N)^{-1/2}\sum_{i=1}^{N}R_i\mathbf{h}_1^{\top}\mathbf{g}_1(\mathbf{Z}_i,\mathbf{W}_i)+N^{-1/2}\sum_{i=1}^{N}\mathbf{h}_2^{\top}\mathbf{g}_2(\mathbf{W}_i) \\
& +\frac{1}{2}\mathbf{h}_1^{\top}\mathcal{V}_{\psi,1}\mathbf{h}_1+\frac{1}{2}\mathbf{h}_2^{\top}\mathcal{V}_{\psi,2}\mathbf{h}_2+o_p(1),
\end{aligned} \tag{A.69}
$$

where $\mathcal{V}_{\psi,1}$ and $\mathcal{V}_{\psi,2}$ were variances of $\mathbf{g}_1$ and $\mathbf{g}_2$ defined in (A.63). This shows that the locally asymptotically normality of the least favorable model.

Based on the local asymptotic normality of the proposed two-dimensional least favorable model (A.69) , we apply the standard "Le Cam" method (Le Cam and Yang, 2000; Tsybakov, 2009) of the minimax efficiency lower bound for parametric sub-model. First, we relax the supremum over the local neighborhood by the Bayesian posterior average over the neighborhood according to the truncated Gaussian prior,

$$
(\mathbf{h}_1^{\top},\mathbf{h}_2^{\top})^{\top}\sim p(\mathbf{h},c,\mathbb{A})=\frac{\phi(\mathbf{h},\mathbf{0},\mathbb{A})I(\|\mathbf{h}\|_2\le c)}{\int_{\|\mathbf{h}\|_2\le c}\phi(\mathbf{h},\mathbf{0},\mathbb{A})d\mathbf{h}},\ \phi(\mathbf{v},\boldsymbol{\mu},\Sigma)=\frac{\exp\left(-(\mathbf{v}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{v}-\boldsymbol{\mu})/2\right)}{(2\pi)^{-q}\det(\Sigma)^{-1/2}},
$$

$$
\begin{aligned}
&\liminf_{c\to\infty}\liminf_{N\to\infty}\sup_{\|\mathbf{h}_1\|_2^2+\|\mathbf{h}_2\|_2^2\le c^2}\int n\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathbf{h}})\}^2 d\prod_{i=1}^{N}\mathbb{P}_{\mathbf{h}}(r_i,\mathbf{z}_i,\mathbf{w}_i)\\
\ge&\liminf_{c\to\infty}\liminf_{N\to\infty}\int\int n\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathbf{h}})\}^2 d\prod_{i=1}^{N}\mathbb{P}_{\mathbf{h}}(r_i,\mathbf{z}_i,\mathbf{w}_i)\times p(\mathbf{h},c,\mathbb{A})d\mathbf{h}.
\end{aligned}
$$

Following Chapter 6 of Le Cam and Yang (2000), we concluded in Lemma A26 that the posterior distribution $\mathbf{h}\mid\mathscr{D}_N$ with $\mathscr{D}_N=\{(R_i,R_i\mathbf{Z}_i,\mathbf{W}_i):i=1,\dots,N\}$ approaches the Gaussian posterior with (untruncated) Gaussian prior $\phi(\mathbf{h},\mathbf{0},\mathbb{A})$ and Gaussian data $\mathbf{V}\mid\mathbf{h}\to N(\mathbf{h},\mathcal{V}_\psi)$ whose variance $\mathcal{V}_\psi$ is defined in (A.63). The limiting Gaussian data $\mathbf{V}$ comes from the limits of empirical processes in the log likelihood of the LAN model $\mathbb{P}_{\mathbf{h}}$,

$$
\mathbf{V}=(\mathbf{V}_1^{\mathsf{T}},\mathbf{V}_2^{\mathsf{T}})^{\mathsf{T}},\ \mathbf{V}_1=-\sum_{i=1}^{N}R_i\mathbf{g}_1(\mathbf{Z}_i,\mathbf{W}_i)/\sqrt{\rho_{\mathsf{N}}N},\ \mathbf{V}_2=-\sum_{i=1}^{N}\mathbf{g}_2(\mathbf{W}_i)/\sqrt{N}.
$$

The limiting Gaussian posterior of $\mathbf{h}=(\mathbf{h}_1^{\mathsf{T}},\mathbf{h}_2^{\mathsf{T}})^{\mathsf{T}}$ is thus

$$
\mathbf{h}\mid\mathscr{D}_N\xrightarrow{TV}\widetilde{\mathbf{h}}\mid\mathbf{V}\sim N\left(\widetilde{\boldsymbol{\mu}},\widetilde{\mathcal{V}}\right),\widetilde{\mathcal{V}}=(\mathbb{A}+\mathcal{V}_\psi)^{-1},\widetilde{\boldsymbol{\mu}}=\widetilde{\mathcal{V}}\mathcal{V}_\psi\mathbf{V}. \tag{A.70}
$$

According to Lemma A26, the average aMSE over truncated Gaussian prior approaches the average aMSE over Gaussian posterior and marginal,

$$
\begin{aligned}
&\liminf_{c\to\infty}\liminf_{N\to\infty}\int\int\rho_{\mathsf{N}}N\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathbf{h}})\}^2 d\prod_{i=1}^{N}\mathbb{P}_{\mathbf{h}}(r_i,\mathbf{z}_i,\mathbf{w}_i)p(\mathbf{h},c,\mathbb{A})d\mathbf{h}\\
=&\liminf_{c\to\infty}\liminf_{N\to\infty}\int\int\rho_{\mathsf{N}}N\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\widetilde{\mathbf{h}}})\}^2\phi(\widetilde{\mathbf{h}},\widetilde{\boldsymbol{\mu}},\widetilde{\mathcal{V}})d\widetilde{\mathbf{h}}\times\phi(\mathbf{v},\mathbf{0},\mathcal{V}_\psi+\mathbb{A})d\mathbf{v}\\
=&\liminf_{c\to\infty}\liminf_{N\to\infty}\widetilde{\mathbb{E}}\left(\widetilde{\mathbb{E}}\left[\rho_{\mathsf{N}}N\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\widetilde{\mathbf{h}}})\}^2\mid\mathbf{V}\right]\right).
\end{aligned} \tag{A.71}
$$

The expectation $\widetilde{\mathbb{E}}$ is taken according to the limiting Gaussian models $\mathbf{V}\sim N(\mathbf{0},\mathcal{V}_\psi+\mathbb{A})$ and $\widetilde{\mathbf{h}}\mid\mathbf{V}\sim N(\widetilde{\mu},\widetilde{\mathcal{V}})$ as defined in (A.70). We are using expression in (A.71) as the ultimate characterization of the asymptotic mean squared estimation error initially defined in (A.65).

While the efficient influence function $\boldsymbol{\psi}_{\mathsf{cmp}}$ under complete data setting $\mathcal{S}_{\mathsf{cmp}}$ can be characterized in multiple ways, we specifically chose the following definition connecting to the least-favorable model (van der Vaart, 1998, Section 25.3).

Definition A18 *Under local exponential-tilt sub-model,*

$$
f_{\mathbf{h}}(\mathbf{z},\mathbf{w})=f_*(\mathbf{z},\mathbf{w})[1+\mathbf{h}^{\mathsf{T}}\mathbf{g}(\mathbf{z},\mathbf{w})]_+/C_{\mathbf{h}},\ C_{\mathbf{h}}=\mathbb{E}_*\left([1+\mathbf{h}^{\mathsf{T}}\mathbf{g}(\mathbf{z},\mathbf{w})]_+\right),
$$

the local shift of parameter $\boldsymbol{\theta}$ along $\mathbf{h}$ observe

$$
\boldsymbol{\theta}_{\mathbf{h}}=\boldsymbol{\theta}_*+\mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mathbf{g}(\mathbf{Z},\mathbf{W})^{\mathsf{T}}\}\mathbf{h}+o(\|\mathbf{h}\|_2).
$$

According to Definition A18, estimating $\boldsymbol{\theta}$ under our chosen local sub-model (A.61) is asymptotically equivalent to projecting the estimated local sub-model

$$\begin{aligned}\boldsymbol{\theta}_{\mathbf{h}} =&\boldsymbol{\theta}_* + \mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mathbf{g}_1(\mathbf{Z},\mathbf{W})^{\mathsf{T}}\}\frac{\mathbf{h}_1}{\sqrt{\rho_{\mathsf{N}}N}} + \mathbb{E}_*\{\boldsymbol{\psi}_{\mathsf{cmp}}(\mathbf{Z},\mathbf{W})\mathbf{g}_2(\mathbf{Z},\mathbf{W})^{\mathsf{T}}\}\frac{\mathbf{h}_2}{\sqrt{N}} + o\left(\frac{c}{\sqrt{\rho_{\mathsf{N}}N}}\right)\\
=&\boldsymbol{\theta}_* + \mathcal{V}_{\psi,1}\frac{\mathbf{h}_1}{\sqrt{\rho_{\mathsf{N}}N}} + \mathcal{V}_{\psi,2}\frac{\mathbf{h}_2}{\sqrt{N}} + o\left(\frac{c}{\sqrt{\rho_{\mathsf{N}}N}}\right),\\
=&\boldsymbol{\theta}_* + \frac{1}{\sqrt{\rho_{\mathsf{N}}N}}\left\{\mathcal{V}_{\psi,1}\mathbf{h}_1 + \sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2}\mathbf{h}_2 + o\left(c\right)\right\},\\
\widehat{\boldsymbol{\theta}} =&\boldsymbol{\theta}_* + \frac{1}{\sqrt{\rho_{\mathsf{N}}N}}\left\{\mathcal{V}_{\psi,1}\widehat{\mathbf{h}}_1 + \sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2}\widehat{\mathbf{h}}_2 + o\left(c\right)\right\}.\end{aligned}$$

The variances $\mathcal{V}_{\psi,1}$ and $\mathcal{V}_{\psi,2}$ of $\mathbf{g}_1$ and $\mathbf{g}_2$ above have been defined in (A.63). The asymptotic mean squared estimation error of $\widehat{\boldsymbol{\theta}}$ can be derived from that of $\widehat{\mathbf{h}}$,

$$\begin{aligned}&\widetilde{\mathbb{E}}\left[\rho_{\mathsf{N}}N\{\mathbf{a}^{\mathsf{T}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathbf{h}})\}^2 \mid \mathbf{V}\right]\\
=&\widetilde{\mathbb{E}}\left[\left\{\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})(\widehat{\mathbf{h}}-\mathbf{h})\right\}^2 \mid \mathbf{V}\right] + o(c^2)\\
=&\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})\widetilde{\mathbb{E}}\left\{(\widehat{\mathbf{h}}-\mathbf{h})(\widehat{\mathbf{h}}-\mathbf{h})^{\mathsf{T}} \mid \mathbf{V}\right\}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})^{\mathsf{T}}\mathbf{a} + o(c^2). \qquad \text{(A.72)}\end{aligned}$$

Conditioning on $\mathbf{V}_1, \mathbf{V}_2$, the asymptotically optimal $\widehat{\mathbf{h}}$ is given by the conditional mean $\widehat{\mathbf{h}} = \widetilde{\mathbb{E}}(\mathbf{h} \mid \mathbf{V}_1, \mathbf{V}_2)$ according to the Andersen's Lemma

$$\mathbf{u}^{\mathsf{T}}\widetilde{\mathbb{E}}\left\{(\widehat{\mathbf{h}}-\mathbf{h})(\widehat{\mathbf{h}}-\mathbf{h})^{\mathsf{T}} \mid \mathbf{V}\right\}\mathbf{u} = \widetilde{\mathbb{E}}\left[\{\mathbf{u}^{\mathsf{T}}(\widehat{\mathbf{h}}-\mathbf{h})\}^2 \mid \mathbf{V}\right] \geq \mathbf{u}^{\mathsf{T}}\widetilde{\mathcal{V}}\mathbf{u}, \qquad \text{(A.73)}$$

where $\widetilde{\mathcal{V}}$ is the posterior variance of $\mathbf{h} \mid \mathbf{V}$ defined in (A.70). Applying (A.73) and (A.72) to the characterization of asymptotic mean squared estimation error (A.71), we have established the lower bound

$$\begin{aligned}aMSE \geq& \liminf_{c\to\infty}\liminf_{N\to\infty}\widetilde{\mathbb{E}}\left\{\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})\widetilde{\mathcal{V}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})^{\mathsf{T}}\mathbf{a} + o(c^2)\right\}\\
\geq&\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})\left\{\mathbb{A}+\mathcal{V}_{\psi}\right\}^{-1}\begin{pmatrix}\mathcal{V}_{\psi,1}\\ \sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2}\end{pmatrix}\mathbf{a}. \qquad \text{(A.74)}\end{aligned}$$

The lower bound (A.74) holds for any prior of $\mathbf{h}$ with arbitrary positive definite $\mathbb{A}$, so the limit of lower bound when $\mathbb{A} \to \mathbb{O}$ is still a lower bound,

$$\begin{aligned}aMSE \geq& \liminf_{\|\mathbb{A}\|_2\to 0}\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})\left\{\mathbb{A}+\mathcal{V}_{\psi}\right\}^{-1}\begin{pmatrix}\mathcal{V}_{\psi,1}\\ \sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi}\end{pmatrix}\mathbf{a}\\
\geq&\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1},\sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2})\begin{pmatrix}\mathcal{V}_{\psi,1}^{-1}, & \mathbb{O}\\ \mathbb{O}, & \mathcal{V}_{\psi,2}^{-1}\end{pmatrix}\begin{pmatrix}\mathcal{V}_{\psi,1}\\ \sqrt{\rho_{\mathsf{N}}}\mathcal{V}_{\psi,2}\end{pmatrix}\mathbf{a}\\
=&\mathbf{a}^{\mathsf{T}}(\mathcal{V}_{\psi,1}+\rho_{\mathsf{N}}\mathcal{V}_{\psi,2})\mathbf{a}.\end{aligned}$$

The lower bound is proportional to the variance of proposed SMMAL influence function

$$\boldsymbol{\psi}_{\mathsf{SSL}}(R,R\mathbf{Z},\mathbf{W}) = R\mathbf{g}_1(\mathbf{Z},\mathbf{W})/\rho_N + \mathbf{g}_2(\mathbf{W}),\ \mathrm{Var}\{\sqrt{\rho_{\mathsf{N}}}\boldsymbol{\psi}_{\mathsf{SSL}}(R,R\mathbf{Z},\mathbf{W})\} = \mathcal{V}_{\psi,1}+\rho_{\mathsf{N}}\mathcal{V}_{\psi,2}.$$

The Assumption 5b is only needed in the end to show that the variance $\rho_{\mathsf{N}}\mathrm{Var}_*\{\boldsymbol{\psi}_{\mathsf{SSL}}(R,\mathbf{Z},\mathbf{W})\}$ is not degenerating when $\rho_{\mathsf{N}} \to 0$.

Appendix D. Auxiliary Lemmas

Lemma A19 *Let $h_1(Y, A, \mathbf{W})$ and $h_2(Y, A, \mathbf{W})$ be two uniformly bounded measurable functions. We have the concentration*

$$\frac{1}{N}\sum_{i=1}^{N} h_1(Y_i, A_i, \mathbf{W}_i) + \frac{R_i}{\rho_{\mathsf{N}}} h_2(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_1(Y, A, \mathbf{W}) + h_2(Y, A, \mathbf{W})\} = O_p\left(n^{-1/2}\right).$$

Proof [Proof of Lemma A19]

First establish the rate for $\frac{\sqrt{\rho_{\mathsf{N}}}}{N}\sum_{i=1}^{N}(R_i/\rho_{\mathsf{N}} - 1)$. By the variance expression

$$\mathrm{Var}\{\sqrt{\rho_{\mathsf{N}}}(R_i/\rho_{\mathsf{N}} - 1)\} = 1 - \rho_{\mathsf{N}},$$

we may apply the Tchebychev's inequality to obtain

$$\frac{\sqrt{\rho_{\mathsf{N}}}}{N}\sum_{i=1}^{N}(R_i/\rho_{\mathsf{N}} - 1) = O_p\left(N^{-1/2}\right).$$

Thus, we have two consequences

$$\begin{aligned} \frac{\sum_{i=1}^{N} R_i}{\rho_{\mathsf{N}} N} - 1 &= O_p\left((\rho_{\mathsf{N}} N)^{-1/2}\right) = O_p\left(n^{-1/2}\right), \\ \left(\sum_{i=1}^{N} R_i\right)^{-1/2} &= \left\{n + O_p\left(\sqrt{n}\right)\right\}^{-1/2} = O_p\left(n^{-1/2}\right). \end{aligned} \tag{A.75}$$

Now, we decompose the empirical process of interest

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} h_1(Y_i, A_i, \mathbf{W}_i) + \frac{R_i}{\rho_{\mathsf{N}}} h_2(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_1(Y, A, \mathbf{W}) + h_2(Y, A, \mathbf{W})\} \\ =&\frac{1}{N}\sum_{i=1}^{N}[h_1(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_1(Y, A, \mathbf{W})\}] \\ &+ \frac{\sum_{i:R_i=1} h_2(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_2(Y, A, \mathbf{W})\}}{\sum_{i=1}^{N} R_i} \frac{\sum_{i=1}^{N} R_i}{\rho_{\mathsf{N}} N}. \end{aligned} \tag{A.76}$$

Conditionally on $R_1, \ldots, R_N$, we apply the Hoeffding's inequality,

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}[h_1(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_1(Y, A, \mathbf{W})\}] &= O_p\left(N^{-1/2}\right), \\ \frac{\sum_{i:R_i=1} h_2(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_2(Y, A, \mathbf{W})\}}{\sum_{i=1}^{N} R_i} &= O_p\left(\left(\sum_{i=1}^{N} R_i\right)^{-1/2}\right). \end{aligned} \tag{A.77}$$

Applying the rates of (A.75) and (A.77) to (A.76), we have shown

$$\frac{1}{N}\sum_{i=1}^{N} h_1(Y_i, A_i, \mathbf{W}_i) + \frac{R_i}{\rho_{\mathsf{N}}} h_2(Y_i, A_i, \mathbf{W}_i) - \mathbb{E}\{h_1(Y, A, \mathbf{W}) + h_2(Y, A, \mathbf{W})\}$$

$$
\begin{aligned}
&= O_p\left(N^{-1/2}\right) + O_p\left(\left(\sum_{i=1}^{N} R_i\right)^{-1/2}\right)\left\{1 + O_p\left(n^{-1/2}\right)\right\} \\
&= O_p\left(n^{-1/2}\right).
\end{aligned}
$$

■

Lemma A20 *Let* $\mathbf{X}$ *be sub-Gaussian random vector satisfying* $\sup_{\|\mathbf{v}\|_2=1} \|\mathbf{v}^\intercal \mathbf{X}\|_{\psi_2} \leq M$ *,* $g(x)$ *be a continuously differentiable link function and* $\tau(x) = \text{sign}(x)\min\{2M, |x|\}$ *be a truncation at* $2M$*. For coefficient* $\bar{\boldsymbol{\beta}}$ *satisfying* $|\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}| \leq M$ *almost surely, we have the mean squared error bound*

$$
\begin{aligned}
\sqrt{\mathbb{E}[\{g(\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}) - g(\mathbf{X}^\intercal \boldsymbol{\beta})\}^2]} &\leq \sqrt{2}/4M\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2, \\
\sqrt{\mathbb{E}[\{g(\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}) - g_\tau(\mathbf{X}^\intercal \boldsymbol{\beta})\}^2]} &\leq \sqrt{2}/4M\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2.
\end{aligned}
$$

Proof [Proof of Lemma A20] We focus on the case with truncation. The case without truncation can be derived from the same steps. By the Mean Value Theorem, we have

$$
g(\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}) - g_\tau(\mathbf{X}^\intercal \boldsymbol{\beta}) = \dot{g}(t)\{\mathbf{X}^\intercal \bar{\boldsymbol{\beta}} - \tau(\mathbf{X}^\intercal \boldsymbol{\beta})\}
$$

for some t between $\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}$ and $\tau(\mathbf{X}^\intercal \boldsymbol{\beta})$. The link function for logistic regression has bounded derivative $|\dot{g}(t)| \leq 1/4$. The truncation at $2M$ never increases estimation error

$$
|\mathbf{X}^\intercal \bar{\boldsymbol{\beta}} - \tau(\mathbf{X}^\intercal \boldsymbol{\beta})| \left\{ \begin{array}{ll} = |\mathbf{X}^\intercal \bar{\boldsymbol{\beta}} - \mathbf{X}^\intercal \boldsymbol{\beta}|, & |\mathbf{X}^\intercal \boldsymbol{\beta}| \leq 2M \\ < |\mathbf{X}^\intercal \bar{\boldsymbol{\beta}} - \mathbf{X}^\intercal \boldsymbol{\beta}|, & |\mathbf{X}^\intercal \boldsymbol{\beta}| > 2M \end{array} \right. .
$$

Thus, we have

$$
\sqrt{\mathbb{E}[\{g(\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}) - g_\tau(\mathbf{X}^\intercal \boldsymbol{\beta})\}^2]} \leq 1/4\sqrt{\mathbb{E}[\{\mathbf{X}^\intercal (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2]}.
$$

Applying the sub-Gaussian property of $\mathbf{X}$, we have

$$
\|\mathbf{X}^\intercal (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_{\psi_2} \leq M\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2.
$$

Using the bound of moments for sub-Gaussian random variables, we have

$$
\mathbb{E}[\{\mathbf{X}^\intercal (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2] \leq 2M^2\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2.
$$

Putting everything together, we have the conclusion

$$
\sqrt{\mathbb{E}[\{g(\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}) - g_\tau(\mathbf{X}^\intercal \boldsymbol{\beta})\}^2]} \leq \sqrt{2}/4M\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2.
$$

■

D1 Lasso with Cross-fitted Parameters

We establish the estimation rates for a generic problems. There are two estimation rates in Lemma A21. In the general case, the asymptotic solution can not be identified by the population level first order condition due to a bias from the cross-fitted parameters. The general case applies to the Lasso of a mis-specified model with cross-fitted parameters. In the special case, the asymptotic solution can be identified by the population level first order condition. The special case applies to the Lasso with no cross-fitted parameters or the Lasso of a correctly specified model which might include cross-fitted parameters.

Consider the cross-fitted Lasso estimator with folds $\{1,\ldots,n\} = \sqcup_{k=1}^{K}\mathcal{I}_k$,

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{K}\sum_{k=1}^{K} \ell_k(\boldsymbol{\beta};\widehat{\boldsymbol{\gamma}}^{(\mathrm{k})}) + \lambda\|\boldsymbol{\beta}\|_1 \tag{A.78}$$

whose loss has derivatives with respect to $\boldsymbol{\beta}$ of the following forms:

$$\begin{aligned}\dot{\boldsymbol{\ell}}_k(\boldsymbol{\beta};\boldsymbol{\gamma}) &= \frac{K}{n}\sum_{i\in\mathcal{I}_k} w_1(\boldsymbol{\gamma}^\intercal\mathbf{X}_i)w_2(\boldsymbol{\beta}^\intercal\mathbf{X}_i)\mathbf{X}_i\{Y_i - g(\boldsymbol{\beta}^\intercal\mathbf{X}_i)\},\\ \ddot{\mathbb{I}}_k(\boldsymbol{\beta};\boldsymbol{\gamma}) &= \frac{K}{n}\sum_{i\in\mathcal{I}_k} w_1(\boldsymbol{\gamma}^\intercal\mathbf{X}_i)w_3(\boldsymbol{\beta}^\intercal\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^\intercal\end{aligned} \tag{A.79}$$

for nonnegative weights w_1, w_2 and w_3. The solution is identified by the population minimum at a specific $\bar{\boldsymbol{\gamma}}$,

$$\bar{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}\{\ell_k(\boldsymbol{\beta};\bar{\boldsymbol{\gamma}})\},\ \|\bar{\boldsymbol{\beta}}\|_0 = s. \tag{A.80}$$

We make the following generic assumptions:

Assumption 6 *(a) (Sub-Gaussian and bounded Covariates)* $\sup_{\|\mathbf{v}\|_2=1}\|\mathbf{v}^\intercal\mathbf{X}\|_{\psi_2} \le M$ *and* $\|\mathbf{X}\|_\infty \le M$ *almost surely;*

(b) (Bounded responses) $|Y| \le M$ *almost surely and* $\|g\|_\infty \le M$*;*

(c) (Identifiability) $\inf_{\|\mathbf{v}\|_2=1}\mathbf{v}^\intercal\mathbb{E}(\mathbf{X}\mathbf{X}^\intercal)\mathbf{v} \ge 1/M$*;*

(d) (Bounds for weights) $w_1(x) \in [1/M, M]$ $\forall x\in\mathbb{R}$, $\|w_1'\|_\infty \le M$, $w_2(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}) \in [1/M, M]$ *and* $w_3(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}) \in [1/M, M]$ *almost surely;*

(e) (Restricted strong convexity) w_3 *is the derivative of some generalized linear model link satisfying* $\|w_3\|_\infty \le M$ *or* $\mathbb{E}\left[\sup_{|u|<1}\{w_3(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}+u)\}^\alpha\right] \le M$ *for some* $\alpha \ge 2$.

Assumption 6 covers all the estimators in (9)-(12) under Assumption 4. The truncations in (12) secure the requirement for w_1 in Assumption 6d. Two w_3 needed for (9)-(12) correspond to the link of logistic regression $g(x)$ and Poisson model e^x, both have been studied in Negahban et al. (2010).

Lemma A21 *Choose the penalty* $\lambda \asymp \sqrt{\log(p)/n}$ *such that*

$$\lambda \ge \frac{3}{K}\sum_{k=1}^{K}\left\|\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathrm{k})}) - \mathbb{E}_{i\in\mathcal{I}_k}\{\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathrm{k})}) \mid \mathscr{D}_k\}\right\|_\infty + 3\kappa_1\kappa_2/M\sqrt{\log(p)/n} \tag{A.81}$$

with large probability for a restricted strong convexity (Negahban et al., 2012) constants κ_1 *and* κ_2 *associated with the auxiliary loss*

$$\widetilde{\ell}_k(\boldsymbol{\beta}) = \frac{K}{n} \sum_{i \in \mathcal{I}_k} \widetilde{G}(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - \widetilde{Y}_i \boldsymbol{\beta}^\intercal \mathbf{X}_i, \, \widetilde{G}'' = w_3.$$

Under Assumption 6, we have

$$\textit{In general: } \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2 = O_p\left(\sqrt{s\log(p)/n}\right) + \sup_{k=1,\ldots,K} \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2,$$

$$\textit{Special case: } \mathbb{E}_{i \in \mathcal{I}_k}\{\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \mid \mathscr{D}_k\} = \mathbf{0}, \, k = 1, \ldots, K, \, \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2 = O_p\left(\sqrt{s\log(p)/n}\right).$$

Proof [Proof of Lemma A21] We focus on the proof of the "in general" case. The proof of the "special case" is simpler and can be made by dropping the steps regarding $\mathbb{E}_{i \in \mathcal{I}_k}\{\dot{\boldsymbol{\ell}}^{(\mathsf{k})}(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \mid \mathscr{D}_k\}$ from proof of the "in general" case. The proof of the "in general" case takes three steps. First, we justify that the oracle choice yields $\lambda \asymp \sqrt{\log(p)/n}$. Second, we obtain a preliminary bound for the estimation error $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2$ through the restricted strong convexity argument. Third and finally, we analyze the preliminary bound in two situations: 1) the error inherited from $\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}$ is dominant, which leads to an immediate bound for $\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2$; 2) the error from Lasso is dominant, which leads to the typical cone property analysis for Lasso.

We first validate the rate for oracle λ. Under Assumptions 6a, 6b and 6d, the summands in $\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})$ have bounded infinity norm

$$\left\|w_1(\boldsymbol{\gamma}^\intercal \mathbf{X}_i) w_2(\boldsymbol{\beta}^\intercal \mathbf{X}_i) \mathbf{X}_i \{Y_i - g(\boldsymbol{\beta}^\intercal \mathbf{X}_i)\}\right\|_\infty \leq 2M^4.$$

By the union bound of the element wise Hoeffding inequality, we have

$$\left\|\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \mathbb{E}_{i \in \mathcal{I}_k}\{\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \mid \mathscr{D}_k\}\right\|_\infty = O_p\left(\sqrt{\log(p)/n}\right).$$

Thus, we may choose $\lambda \asymp \sqrt{\log(p)/n}$ to satisfy (A.81) with large probability.

By the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\frac{1}{K} \sum_{k=1}^K \ell_k(\widehat{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) + \lambda \|\widehat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{K} \sum_{k=1}^K \ell_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) + \lambda \|\bar{\boldsymbol{\beta}}\|_1. \tag{A.82}$$

Denote the standardized estimation error as $\boldsymbol{\delta} = (\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})/\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2$. The Hessian of the loss in (A.79) is positive semi-definite under Assumptions 6d,

$$\mathbf{v}^\intercal \frac{1}{K} \sum_{k=1}^K \ddot{\mathbb{l}}_k(\boldsymbol{\beta}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \mathbf{v} = \frac{1}{K} \sum_{k=1}^K w_1(\boldsymbol{\gamma}^\intercal \mathbf{X}_i) w_3(\boldsymbol{\beta}^\intercal \mathbf{X}_i) (\mathbf{v}^\intercal \mathbf{X}_i)^2 \geq 0, \tag{A.83}$$

indicating that the loss is convex. Using the convexity of the loss function, we have for the truncated L_2-estimation error $t = \min\{\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2, 1\}$

$$\frac{1}{K} \sum_{k=1}^K \ell_k(\bar{\boldsymbol{\beta}} + t\boldsymbol{\delta}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) + \lambda \|\bar{\boldsymbol{\beta}} + t\boldsymbol{\delta}\|_1 \leq \frac{1}{K} \sum_{k=1}^K \ell_k(\bar{\boldsymbol{\gamma}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) + \lambda \|\bar{\boldsymbol{\beta}}\|_1. \tag{A.84}$$

By the triangle inequality $\|\bar{\boldsymbol{\beta}}\|_1 - \|\bar{\boldsymbol{\beta}} + t\boldsymbol{\delta}\|_1 \leq t\|\boldsymbol{\delta}\|_1$, we have from (A.84)

$$\frac{1}{K}\sum_{k=1}^{K} \ell_k(\bar{\boldsymbol{\beta}} + t\boldsymbol{\delta}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \ell_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \leq t\lambda_\gamma \|\boldsymbol{\delta}\|_1 \tag{A.85}$$

Now, we establish the restricted strong convexity property for each $\ell_k(\cdot; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})$. By Assumption 6e, we may a hypothetical generalized linear model loss for $\widetilde{Y}_i \sim \widetilde{G}'(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)$

$$\widetilde{\ell}_k(\boldsymbol{\beta}) = \frac{K}{n}\sum_{i \in \mathcal{I}_k} \widetilde{G}(\boldsymbol{\beta}^\intercal \mathbf{X}_i) - \widetilde{Y}_i \boldsymbol{\beta}^\intercal \mathbf{X}_i,\ \widetilde{G}'' = w_3. \tag{A.86}$$

The restricted strong convexity of (A.86) is established by analyzing the lower bound for

$$\widetilde{\ell}_k(\bar{\boldsymbol{\beta}} + \boldsymbol{\Delta}) - \widetilde{\ell}_k(\bar{\boldsymbol{\beta}}) - \boldsymbol{\Delta}^\intercal \dot{\widetilde{\boldsymbol{\ell}}}(\bar{\boldsymbol{\beta}}) = \frac{K}{n}\sum_{i \in \mathcal{I}_k} w_3(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i + \nu \boldsymbol{\Delta}^\intercal \mathbf{X}_i)(\boldsymbol{\Delta}^\intercal \mathbf{X}_i)^2 \tag{A.87}$$

uniformly for $\|\boldsymbol{\Delta}\|_2 \leq 1$ and $\nu \in [0,1]$ (Negahban et al., 2010, Proof of Proposition 2). Under Assumptions 6a and 6e, the lower bound is given by

$$\frac{K}{n}\sum_{i \in \mathcal{I}_k} w_3(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i + \nu \boldsymbol{\Delta}^\intercal \mathbf{X}_i)(\boldsymbol{\Delta}^\intercal \mathbf{X}_i)^2 \geq \kappa_1 \|\boldsymbol{\Delta}\|_2 \left\{ \|\boldsymbol{\Delta}\|_2 - \kappa_2 \sqrt{\log(p)/n} \|\boldsymbol{\Delta}\|_1 \right\} \tag{A.88}$$

for all $\|\boldsymbol{\Delta}\|_2 \leq 1$ and $\nu \in [0,1]$ with absolute constants κ_1 and κ_2. Under Assumption 6d, the restricted strong convexity for $\ell_k(\cdot; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})$ can be also established by analyzing the same quantity in (A.87)

$$\begin{aligned}\ell_k(\bar{\boldsymbol{\beta}} + \boldsymbol{\Delta}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \ell_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \boldsymbol{\Delta}^\intercal \dot{\boldsymbol{\ell}}(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) =& \frac{K}{n}\sum_{i \in \mathcal{I}_k} w_1(\mathbf{X}_i^\intercal \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) w_3(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i + \nu \boldsymbol{\Delta}^\intercal \mathbf{X}_i)(\boldsymbol{\Delta}^\intercal \mathbf{X}_i)^2 \\ \geq& \frac{K}{n}\sum_{i \in \mathcal{I}_k} M^{-1} w_3(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i + \nu \boldsymbol{\Delta}^\intercal \mathbf{X}_i)(\boldsymbol{\Delta}^\intercal \mathbf{X}_i)^2\end{aligned} \tag{A.89}$$

Applying the lower bound in (A.88) to (A.89) at $\boldsymbol{\Delta} = t\boldsymbol{\delta}$, we obtain

$$\ell_k(\bar{\boldsymbol{\beta}} + t\boldsymbol{\delta}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \ell_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - t\boldsymbol{\delta}^\intercal \dot{\boldsymbol{\ell}}(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \geq t^2\kappa_1/M - t^2\kappa_1\kappa_2/M\sqrt{\log(p)/n}\|\boldsymbol{\delta}\|_1. \tag{A.90}$$

Combining (A.85) and (A.90), we obtain

$$t\kappa_1/M \leq \lambda\|\boldsymbol{\delta}\|_1 - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\delta}^\intercal \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) + t\kappa_1\kappa_2/M\sqrt{\log(p)/n}\|\boldsymbol{\delta}\|_1. \tag{A.91}$$

We decompose $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\delta}^\intercal \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})$

$$\left|\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\delta}^\intercal \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\right| = \left|\frac{\boldsymbol{\delta}^\intercal}{K}\sum_{k=1}^{K} \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \mathbb{E}_{i \in \mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}}; \widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) \mid \mathscr{D}_k^c\right]\right.$$

$$
\begin{aligned}
&+\frac{\boldsymbol{\delta}^{\mathsf{T}}}{K}\sum_{k=1}^{K}\mathbb{E}_{i\in\mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\mid\mathscr{D}_k^c\right]\Bigg| \\
\leq&\|\boldsymbol{\delta}\|_1\left\|\frac{K}{N}\sum_{i\in\mathcal{I}_k}\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})-\mathbb{E}_{i\in\mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\mid\mathscr{D}_k^c\right]\right\|_\infty \\
&+\sup_{k=1,\ldots,K}\left\|\mathbb{E}_{i\in\mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\mid\mathscr{D}_k^c\right]\right\|_2. \qquad (A.92)
\end{aligned}
$$

For "special case", the second term in (A.92) is zero, so the proofs up to **Case 2** following (A.99) can be skipped. Using the first order condition of optimality for $\bar{\boldsymbol{\beta}}$, we have

$$\mathbb{E}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\bar{\gamma})\right]=0,$$

so we can bound the second term in (A.92) for "in general" by

$$
\begin{aligned}
&\left\|\mathbb{E}_{i\in\mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\mid\mathscr{D}_k^c\right]\right\|_2 \\
=&\left\|\mathbb{E}_{i\in\mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})-\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\bar{\gamma})\mid\mathscr{D}_k^c\right]\right\|_2 \\
=&\left\|\mathbb{E}_{i\in\mathcal{I}_k}[\{w_1(\mathbf{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})-w_1(\mathbf{X}_i^{\mathsf{T}}\bar{\gamma})\}w_2(\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i)\{g(\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i)-Y_i\}\mid\mathscr{D}_k^c]\right\|_2. \qquad (A.93)
\end{aligned}
$$

Using the Lipschitz condition for w_1 from Assumption 6d and other bounds from Assumptions 6b and 6d, we may bound (A.93) by

$$
\begin{aligned}
&\left\|\mathbb{E}_{i\in\mathcal{I}_k}[\{w_1(\mathbf{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})-w_1(\mathbf{X}_i^{\mathsf{T}}\bar{\gamma})\}w_2(\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i)\{g(\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i)-Y_i\}\mid\mathscr{D}_k^c]\right\|_2 \\
\leq&2M^3\mathbb{E}_{i\in\mathcal{I}_k}\{|\mathbf{X}_i^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma})|\mid\mathscr{D}_k^c\} \qquad (A.94)
\end{aligned}
$$

Applying the sub-Gaussian property in Assumption 6a, we have

$$\mathbb{E}_{i\in\mathcal{I}_k}\{|\mathbf{X}_i^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma})|\mid\mathscr{D}_k^c\}\leq\sqrt{\pi}\|\mathbf{X}_i^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma})\|_{\psi_2}\leq\sqrt{\pi}M\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma}\|_2. \qquad (A.95)$$

Collecting (A.92)-(A.95), we obtain

$$
\begin{aligned}
\left|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\delta}^{\mathsf{T}}\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\right|\leq&\|\boldsymbol{\delta}\|_1\left\|\frac{K}{N}\sum_{i\in\mathcal{I}_k}\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})-\mathbb{E}_{i\in\mathcal{I}_k}\left[\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\mid\mathscr{D}_k^c\right]\right\|_\infty \\
&+2\sqrt{\pi}M^4\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma}\|_2. \qquad (A.96)
\end{aligned}
$$

Applying (A.96) and the definition of λ to (A.91), we get the preliminary estimation bound

$$t\kappa_1/M\leq 4/3\lambda\|\boldsymbol{\delta}\|_1+2\sqrt{\pi}M^4\sup_{k=1,\ldots,K}\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma}\|_2. \qquad (A.97)$$

Then, we separately analyze two cases.

Case 1:

$$2\sqrt{\pi}M^4\sup_{k=1,\ldots,K}\|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}-\bar{\gamma}\|_2\geq\lambda\|\boldsymbol{\delta}\|_1/3.$$

In this case, the estimation error is dominated by $\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}$. We simply have from (A.97)

$$t\kappa_1/M \le 10\sqrt{\pi}M^4 \sup_{k=1,\dots,K} \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2.$$

Thus, we have

$$\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2 \le 10\sqrt{\pi}M^5/\kappa_1 \sup_{k=1,\dots,K} \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2. \tag{A.98}$$

Case 2:

$$2\sqrt{\pi}M^4 \sup_{k=1,\dots,K} \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2 \le \lambda\|\boldsymbol{\delta}\|_1/3. \tag{A.99}$$

In this case, the estimation error is comparable to the situation that we have the asymptotic weights $w_1(\bar{\boldsymbol{\gamma}}^{\mathsf{T}}\mathbf{X}_i)$. Thus, the sparsity of $\bar{\boldsymbol{\beta}}$ may affect the estimation error.

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

$$(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})^{\mathsf{T}}\frac{1}{K}\sum_{k=1}^{K}\{\dot{\boldsymbol{\ell}}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\} = \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2\boldsymbol{\delta}^{\mathsf{T}}\frac{1}{K}\sum_{k=1}^{K}\{\dot{\boldsymbol{\ell}}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\}. \tag{A.100}$$

Due to the convexity of the quadratic loss $\ell(\boldsymbol{\gamma})$ from (A.83), the symmetrized Bregman's divergence (A.100) is nonnegative. Denote the indices set of nonzero coefficient in $\bar{\boldsymbol{\beta}}$ as $\mathcal{O} = \{j : \bar{\beta}_j \ne 0\}$. We denote the $\boldsymbol{\delta}_{\mathcal{O}}$ and $\boldsymbol{\delta}_{\mathcal{O}^c}$ as the sub-vectors for $\boldsymbol{\delta}$ at positions in $\mathcal{O}$ and at positions not in $\mathcal{O}$, respectively. The solution $\widehat{\boldsymbol{\beta}}$ satisfies the KKT condition

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\dot{\boldsymbol{\ell}}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\right\|_\infty \le \lambda,\ \frac{1}{K}\sum_{k=1}^{K}\dot{\ell}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})_j = -\lambda\,\mathrm{sign}(\widehat{\beta}_j),\ j : \widehat{\beta}_j \ne 0.$$

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

$$\delta_j\frac{1}{K}\sum_{k=1}^{K}\dot{\ell}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})_j \le |\delta_j|\lambda,\ j \in \mathcal{O};\ \delta_j\frac{1}{K}\sum_{k=1}^{K}\dot{\ell}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})_j = \frac{-\widehat{\beta}_j\lambda\,\mathrm{sign}(\widehat{\beta}_j)}{\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2} = -\lambda|\delta_j|,\ j \in \mathcal{O}^c. \tag{A.101}$$

Applying the (A.101) to (A.100), we have the upper bound,

$$\begin{aligned}
&\boldsymbol{\delta}^{\mathsf{T}}\frac{1}{K}\sum_{k=1}^{K}\{\dot{\boldsymbol{\ell}}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})}) - \dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\}\\
=&\sum_{j\in\mathcal{O}}\delta_j\frac{1}{K}\sum_{k=1}^{K}\dot{\ell}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})_j + \sum_{j\in\mathcal{O}^c}\delta_j\frac{1}{K}\sum_{k=1}^{K}\dot{\ell}_k(\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})_j - \boldsymbol{\delta}^{\mathsf{T}}\sum_{k=1}^{K}\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\\
\le&\lambda\sum_{j\in\mathcal{O}}|\delta_j| - \lambda\sum_{j\in\mathcal{O}^c}|\delta_j| + \left|\boldsymbol{\delta}^{\mathsf{T}}\sum_{k=1}^{K}\dot{\boldsymbol{\ell}}_k(\bar{\boldsymbol{\beta}};\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})})\right|.
\end{aligned}$$

Then, we apply (A.96), the definition of λ and (A.99),

$$0 \le \lambda\|\boldsymbol{\delta}_{\mathcal{O}}\|_1 - \lambda\|\boldsymbol{\delta}_{\mathcal{O}^c}\|_1 + \frac{2}{3}\lambda\|\boldsymbol{\delta}\|_1.$$

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

$$\|\boldsymbol{\delta}\|_1 \le 6\lambda\|\boldsymbol{\delta}_{\mathcal{O}}\|_1 \le 6\sqrt{s}\|\boldsymbol{\delta}\|_2 = 6\sqrt{s}. \tag{A.102}$$

Applying (A.99) and (A.102) to (A.97), we have the other bound for the estimation error

$$t\kappa_1/M \le 5/3\lambda\|\boldsymbol{\delta}\|_1 \le 10M\sqrt{s}\lambda,\ \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2 \le 10M^2/\kappa_1\sqrt{s}\lambda. \tag{A.103}$$

The "special case" estimation error is directly given by (A.103)

$$\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2 = O_p\left(\sqrt{s\log(p)/n}\right).$$

For "in general", we combine the bounds from the two cases (A.98) and (A.103),

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2 \le& \max\left\{10\sqrt{\pi}M^5/\kappa_1 \sup_{k=1,\dots,K} \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2, 10M^2/\kappa_1\sqrt{s}\lambda\right\} \\ =&O_p\left(\sqrt{s\log(p)/n} + \sup_{k=1,\dots,K} \|\widehat{\boldsymbol{\gamma}}^{(\mathsf{k})} - \bar{\boldsymbol{\gamma}}\|_2\right).\end{aligned}$$

■

Appendix E. Additional Technical Details

E1 Definitions

Definition A22 (Hölder class) *A function $f(\mathbf{x})$ defined over $[-M, M]^d$ is Hölder class s if*

$$\sup_{\mathbf{x}_1,\mathbf{x}_2\in[-M,M]^d} \sup_{\substack{a_1,\dots,a_d\in\mathbb{N}\\ a_1+\dots+a_d=[s]}} \left|\frac{\partial^{[s]}}{\partial x_1^{a_1}\dots x_d^{a_d}}\{f(\mathbf{x}_1) - f(\mathbf{x}_2)\}\right| \|\mathbf{x}_1 - \mathbf{x}_2\|_2^{[s-1]-s} < \infty.$$

We note the maximal Hölder class as $\mathcal{H}(f) = \sup\{s : f \text{ is Hölder class } s\}$.

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

Definition A23 (Sub-Gaussian and Sub-Exponential Random Variables) *The sub-Gaussian parameter for a random variable V is defined as*

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

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

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

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

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

The random variable V is sub-exponential if $\|V\|_{\psi_1}$ is finite.

E2 Geometry of model tangent space

The nonparametric model for observed data is thus

$$\begin{aligned}\mathcal{M}_{obs} =\Big\{ & f_{\mathbf{X},A,Y,\mathbf{S},R}(\mathbf{x},a,t,\mathbf{s},r) = f_{\mathbf{X}}(\mathbf{x})[\pi(a,\mathbf{x})^a f_{Y|A,\mathbf{X}}(y|a,\mathbf{x}) \\ & \times f_{\mathbf{S}|Y,A,\mathbf{X}}(s|y,a,\mathbf{x})m(\mathbf{x})]^r \big(f_{\mathbf{S}|\mathbf{X}}(\mathbf{s}|\mathbf{x})\{1-m(x)\}\big)^{1-r} : \\ & f_{\mathbf{X}}, \pi, f_{Y|A,\mathbf{X}}, f_{\mathbf{S}|Y,A,\mathbf{X}}, m \text{ are arbitrary pdfs/pmfs}\Big\}. \end{aligned} \tag{A.104}$$

We consider the parametric sub-model indexed by parameter $\boldsymbol{\gamma}$

$$\begin{aligned}\mathcal{M}_{par} =\Big\{ & f_{\mathbf{X},A,Y,\mathbf{S},R}(\mathbf{x},a,t,\mathbf{s},r;\boldsymbol{\gamma}) = f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\gamma})[\pi(a,\mathbf{x};\boldsymbol{\gamma}) f_{Y|A,\mathbf{X}}(y|a,\mathbf{x};\boldsymbol{\gamma}) \\ & \times f_{\mathbf{S}|Y,A,\mathbf{X}}(s|y,a,\mathbf{x};\boldsymbol{\gamma})m^*(\mathbf{x})]^r \\ & \times \big[f_{\mathbf{S}|\mathbf{X}}(\mathbf{s}|\mathbf{x};\boldsymbol{\gamma})\{1-m^*(\mathbf{x})\}\big]^{1-r} : \boldsymbol{\gamma}\in\Gamma\Big\}, \end{aligned} \tag{A.105}$$

$$f_{\mathbf{S}|\mathbf{X}}(\mathbf{s}|\mathbf{x};\boldsymbol{\gamma}) = \sum_{a\in\mathbb{N}}\int_{y\in\mathbb{R}} \pi(a,\mathbf{x};\boldsymbol{\gamma}) f_{Y|A,\mathbf{X}}(y|a,\mathbf{x};\boldsymbol{\gamma}) f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{s}|y,a,\mathbf{x};\boldsymbol{\gamma})dy. \tag{A.106}$$

where $\boldsymbol{\gamma}=\boldsymbol{\gamma}^*$ indicates the true parameter.

Utilizing the identity

$$\begin{aligned} & \frac{\partial \log\{f_{\mathbf{S}|\mathbf{X}}(\mathbf{S}|\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \\ =& \sum_{a\in\mathbb{N}}\int_{y\in\mathbb{R}} \frac{\partial}{\partial\boldsymbol{\gamma}}\pi(a,\mathbf{X};\boldsymbol{\gamma})\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \frac{f_{Y|A,\mathbf{X}}(y|a,\mathbf{X};\boldsymbol{\gamma}^*) f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|y,a,\mathbf{X};\boldsymbol{\gamma}^*)}{f_{\mathbf{S}|\mathbf{X}}(\mathbf{X}|\mathbf{X};\boldsymbol{\gamma}^*)}dy \\ & \sum_{a\in\mathbb{N}}\int_{y\in\mathbb{R}} \frac{\partial}{\partial\boldsymbol{\gamma}} f_{Y|A,\mathbf{X}}(y|a,\mathbf{X};\boldsymbol{\gamma})\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \frac{\pi(a,\mathbf{X};\boldsymbol{\gamma}^*) f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|y,a,\mathbf{X};\boldsymbol{\gamma}^*)}{f_{\mathbf{S}|\mathbf{X}}(\mathbf{X}|\mathbf{X};\boldsymbol{\gamma}^*)}dy \\ & +\sum_{a\in\mathbb{N}}\int_{y\in\mathbb{R}} \frac{\partial}{\partial\boldsymbol{\gamma}} f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|y,a,\mathbf{X};\boldsymbol{\gamma})\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \frac{\pi(a,\mathbf{X};\boldsymbol{\gamma}^*) f_{Y|A,\mathbf{X}}(y|a,\mathbf{X};\boldsymbol{\gamma}^*)}{f_{\mathbf{S}|\mathbf{X}}(\mathbf{X}|\mathbf{X};\boldsymbol{\gamma}^*)}dy \\ =& \sum_{a\in\mathbb{N}}\int_{y\in\mathbb{R}} \frac{\partial}{\partial\boldsymbol{\gamma}}[\log\{\pi(a,\mathbf{X};\boldsymbol{\gamma})\} + \log\{f_{Y|A,\mathbf{X}}(y|a,\mathbf{X};\boldsymbol{\gamma})\} + \log\{f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|y,a,\mathbf{X};\boldsymbol{\gamma})\}]\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \\ & \quad\times \frac{\pi(a,\mathbf{X};\boldsymbol{\gamma}^*) f_{Y|A,\mathbf{X}}(y|a,\mathbf{X};\boldsymbol{\gamma}^*) f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|y,a,\mathbf{X};\boldsymbol{\gamma}^*)}{f_{\mathbf{S}|\mathbf{X}}(\mathbf{X}|\mathbf{X};\boldsymbol{\gamma}^*)}dy \\ =& \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\gamma}}\log\{\pi(A,\mathbf{X};\boldsymbol{\gamma})\}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \mid \mathbf{S},\mathbf{X}\right] + \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\gamma}}\log\{f_{Y|A,\mathbf{X}}(Y|A,\mathbf{X};\boldsymbol{\gamma})\}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \mid \mathbf{S},\mathbf{X}\right] \\ & + \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\gamma}}\log\{f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|Y,A,\mathbf{X};\boldsymbol{\gamma})\}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \mid \mathbf{S},\mathbf{X}\right], \end{aligned}$$

we express the score vector of the parametric sub-model as

$$\boldsymbol{\Psi}(\mathbf{X},A,\mathbf{S},Y,R) = \frac{\partial \log\{f_{\mathbf{X},A,Y,\mathbf{S},R}(\mathbf{X},A,Y,\mathbf{S},R;\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}$$

$$=\boldsymbol{\Psi}_{\mathbf{X}}(\mathbf{X};\boldsymbol{\gamma}^*)+\boldsymbol{\Psi}_A(R,\mathbf{X};\boldsymbol{\gamma}^*)+\boldsymbol{\Psi}_Y(R,A,\mathbf{X};\boldsymbol{\gamma}^*)+\boldsymbol{\Psi}_{\mathbf{S}}(R,Y,A,\mathbf{X};\boldsymbol{\gamma}^*) \tag{A.107}$$

where the components are

$$\begin{aligned}
\boldsymbol{\Psi}_{\mathbf{X}}(\mathbf{X};\boldsymbol{\gamma}^*) &= \frac{\partial\log\{f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*},\\
\boldsymbol{\Psi}_A(R,\mathbf{X};\boldsymbol{\gamma}^*) &= R\,\frac{\partial\log\{\pi(A|\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}+(1-R)\mathbb{E}\left[\frac{\partial\log\{\pi(A|\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\mid\mathbf{S},\mathbf{X}\right],\\
\boldsymbol{\Psi}_Y(R,A,\mathbf{X};\boldsymbol{\gamma}^*) &= R\,\frac{\partial\log\{f_{Y|A,\mathbf{X}}(Y|A,\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\\
&\quad+(1-R)\mathbb{E}\left[\frac{\partial\log\{f_{Y|A,\mathbf{X}}(Y|A,\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\mid\mathbf{S},\mathbf{X}\right],\\
\boldsymbol{\Psi}_{\mathbf{S}}(R,\mathbf{S},A,\mathbf{X};\boldsymbol{\gamma}^*) &= R\,\frac{\partial\log\{f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|Y,A,\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\\
&\quad+(1-R)\mathbb{E}\left[\frac{\partial\log\{f_{\mathbf{S}|Y,A,\mathbf{X}}(\mathbf{S}|Y,A,\mathbf{X};\boldsymbol{\gamma})\}}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\mid\mathbf{S},\mathbf{X}\right].
\end{aligned}\tag{A.108}$$

Let $\mathscr{H}$ be the Hilbert space of mean zero finite variance random variables measurable to $\sigma\{\mathbf{X},AR,\mathbf{S},YR,R\}$. The nuisance parameter tangent space is spanned by $\boldsymbol{\Psi}(\mathbf{X},A,\mathbf{S},Y,R)$,

$$\Lambda=\{\mathbf{v}^{\intercal}\boldsymbol{\Psi}(\mathbf{X},A,\mathbf{S},Y,R):\mathcal{M}_{par}\subset\mathcal{M}_{obs}\}. \tag{A.109}$$

According to the decomposition (A.107), we can decompose the nuisance parameter tangent space

$$\Lambda=\Lambda_{\mathbf{X}}+\Lambda_A+\Lambda_Y+\Lambda_{\mathbf{S}}.$$

We derive $\Lambda_{\mathbf{X}}$, Λ_A, Λ_Y and $\Lambda_{\mathbf{S}}$ as

$$\begin{aligned}
\Lambda_{\mathbf{X}} &= \{h(\mathbf{X})\in\mathscr{H}:\mathbb{E}[h(\mathbf{X})]=0\},\\
\Lambda_A &= \Big\{Rh(A,\mathbf{X})+(1-R)\mathbb{E}[h(A,\mathbf{X})\mid\mathbf{S},\mathbf{X}]\in\mathscr{H}:\mathbb{E}[h(A,\mathbf{X})\mid\mathbf{X}]=0\Big\},\\
\Lambda_Y &= \Big\{Rh(Y,A,\mathbf{X})+(1-R)\mathbb{E}[h(Y,A,\mathbf{X})\mid\mathbf{S},\mathbf{X}]\in\mathscr{H}:\mathbb{E}[h(Y,A,\mathbf{X})\mid A,\mathbf{X}]=0\Big\},\\
\Lambda_{\mathbf{S}} &= \Big\{Rh(\mathbf{S},Y,A,\mathbf{X})+(1-R)\mathbb{E}[h(\mathbf{S},Y,A,\mathbf{X})\mid\mathbf{S},\mathbf{X}]\in\mathscr{H}:\\
&\qquad\mathbb{E}[h(\mathbf{S},Y,A,\mathbf{X})\mid Y,A,\mathbf{X}]=0\Big\}.
\end{aligned}\tag{A.110}$$

Under the settings of Robins et al. (1994) and Kallus and Mao (2024), the scores for $f_{A|\mathbf{X}}$ and $f_{Y|A,\mathbf{X}}$ belong to two linear subspaces orthogonal to each other in $\mathscr{H}$, the Hilbert space of mean zero finite variance random variables. However, the two scores under $\mathcal{S}_{\mathsf{SSL}}$ belong to the linear subspaces Λ_A and Λ_Y which share a correlated component from the unlabeled data induced by the surrogates $\mathbf{S}$.

E3 B-spline Regression

Lemma A24 *Under the assumptions:*

1. $\mathbf{X} \in [0,1]^p$ *with density* $f_{\mathbf{X}}(\mathbf{x}) \in [1/M, M]$, $\forall \mathbf{x} \in [0,1]^p$*;*
2. $\mathbf{q}(\mathbf{x}) \in \mathbb{R}^b$ *is the vector of tensor product B-splines of order* κ *with knot spacing approximately proportional to the number of knots;*
3. $\inf_{\|\mathbf{v}\|_2=1} \mathbf{v}^\intercal \mathbb{E}\{\mathbf{q}(\mathbf{X})\mathbf{q}(\mathbf{X})^\intercal\}\mathbf{v} \geq 1/M$*;*
4. $\sup_{\mathbf{x}\in[0,1]^p} \|\mathbf{q}(\mathbf{x})\|_2 \leq M\sqrt{b}$*;*
5. $|Y| \leq M$*;*
6. $\mu_*(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ *is Hölder of order* s.

Let $\bar{\mu}(\mathbf{x})$ *be the best linear approximation of* $\mu_*(\mathbf{x})$ *with basis* $\mathbf{q}(\mathbf{x})$

$$\bar{\mu}(\mathbf{x}) = \mathbf{q}(\mathbf{x})^\intercal \left[\mathbb{E}\{\mathbf{q}(\mathbf{X})\mathbf{q}(\mathbf{X})^\intercal\}\right]^{-1} \mathbb{E}\{\mathbf{q}(\mathbf{X})Y\},$$

and its estimator with n *samples*

$$\widehat{\mu}(\mathbf{x}) = \mathbf{q}(\mathbf{x})^\intercal \left\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{q}(\mathbf{X}_i)\mathbf{q}(\mathbf{X}_i)^\intercal\right\}^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{q}(\mathbf{X}_i)Y_i$$

The approximation error is

$$\|\bar{\mu} - \mu_*\|_2 = O(b^{-\min\{1+\kappa,s\}/p}).$$

The estimation error with sample n *is*

$$\|\bar{\mu} - \widehat{\mu}\|_2 = O_p\left(\sqrt{b/n}\right).$$

See in Newey and Robins (2018) for example.

E4 Minimax Lower Bound

Lemma A25 *Denote the truncation at zero* $[\cdot]_+$ *and the normalizing constant* $C_{\mathbf{h}}$

$$\left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right]_+ = \max\left\{0, 1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right\},$$
$$C_{\mathbf{h}} = \mathbb{E}_*\left(\left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right]_+\right)$$

and the two-way tilted density

$$f_{\mathbf{h}}(\mathbf{z}_i, \mathbf{w}_i) = f_*(\mathbf{z}_i, \mathbf{w}_i)\left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right]_+ / C_{\mathbf{h}}.$$

If $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *has bounded variance under* f_*,

$$\sup_{\|\mathbf{v}\|_2=1} \mathbb{E}_*[\{\mathbf{v}^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)\}^2] + \sup_{\|\mathbf{u}\|_2=1} \mathbb{E}_*[\{\mathbf{u}^\intercal \mathbf{g}_2(\mathbf{W}_i)\}^2] \leq M,$$

the tilted density falls in the neighborhood in $\|\cdot\|_{\mathsf{TV}}$,

$$\|f_* - f_{\mathbf{h}}\|_{\mathsf{TV}} \leq M\sqrt{\|\mathbf{h}_1\|_2^2/n + \|\mathbf{h}_2\|_2^2/N} + o(\|\mathbf{h}\|_2/\sqrt{n}). \tag{A.64}$$

Proof [Proof of Lemma A25] The proof extends Example 5 page 11 of Duchi (2021) to two-way tilted sub-models for characterizing semi-supervised learning setting. First we show that the normalizing constant approaches 1 at

$$C_{\mathbf{h}} = 1 + O(\|\mathbf{h}\|_2^2/n).$$

By definition of $C_{\mathbf{h}}$ and mean zero assumption for $\mathbf{g}_1$ and $\mathbf{g}_2$, we have the lower bound for $C_{\mathbf{h}}$

$$\begin{aligned}
C_{\mathbf{h}} =& \mathbb{E}_*\left(\left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right]_+\right) \\
\geq& \mathbb{E}_*\left(1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right) \\
=& 1 + \frac{\mathbf{h}_1^\intercal \mathbb{E}_*\{\mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)\}}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbb{E}_*\{\mathbf{g}_2(\mathbf{W}_i)\}}{\sqrt{N}} \\
=& 1.
\end{aligned} \tag{A.111}$$

Define the event of activated truncation

$$\Xi_i = \mathrm{I}\left\{1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}} < 0\right\}, \tag{A.112}$$

we may alternatively represent the tilt factor as

$$\begin{aligned}
&\left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right]_+ \\
=& 1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}} - \Xi_i\left\{1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right\}.
\end{aligned} \tag{A.113}$$

Using (A.113), we establish an upper bound of $C_{\mathbf{h}}$

$$\begin{aligned}
C_{\mathbf{h}} =& \mathbb{E}_*\left(1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right) \\
&- \mathbb{E}_*\left(\Xi_i\left\{1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right\}\right) \\
=& 1 + \mathbb{E}_*\left(\Xi_i\left|1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right|\right) \\
\leq& 1 + \mathbb{E}_*\left(\Xi_i\left|\frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}}\right|^2\right)
\end{aligned}$$

$$\leq 1 + 2\mathbb{E}_* \left(|\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)|^2/n + |\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)|^2/N \right). \tag{A.114}$$

Applying the bounded variance assumption for $\mathbf{g}_1$ and $\mathbf{g}_2$, we have the upper bound

$$\begin{aligned} C_{\mathbf{h}} \leq & 1 + 2\mathbb{E}_* \left(|\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)|^2/n + |\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)|^2/N \right) \\ \leq & 1 + 2\mathbb{E}_* \left(|\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)|^2 + |\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)|^2 \right)/n \\ \leq & 1 + 2M\|\mathbf{h}\|_2^2/n. \end{aligned} \tag{A.115}$$

Combining the lower bound (A.111) and upper bound (A.115) of C_h, we have shown

$$C_{\mathbf{h}} = 1 + O(\|\mathbf{h}\|_2^2/n) = 1 + o(\|\mathbf{h}\|_2/\sqrt{n}). \tag{A.116}$$

Then, we bound the distance in total variation

$$\begin{aligned} \|f_* - f_{\mathbf{h}}\|_{\mathsf{TV}} = & \int f_*(\mathbf{z}, \mathbf{w}) \left| 1 - \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} \right]_+ / C_{\mathbf{h}} \right| d\mathbf{z} d\mathbf{w} \\ = & \mathbb{E}_* \left\{ \left| 1 - \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} \right]_+ / C_{\mathbf{h}} \right| \right\}. \end{aligned}$$

We decompose the tilted factor into 3 parts

$$\begin{aligned} & 1 - \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} \right]_+ / C_{\mathbf{h}} \\ = & \underbrace{1 - 1/C_{\mathbf{h}}}_{T_1} + \underbrace{\left\{ \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} \right\} / C_{\mathbf{h}}}_{T_2} \\ & + \underbrace{\left(1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} - \left[1 + \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{z}_i, \mathbf{w}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}} \right]_+ \right) / C_{\mathbf{h}}}_{T_3} \end{aligned} \tag{A.117}$$

and evaluate their L_1-norm separately. Applying the order of $C_{\mathbf{h}}$ established in (A.116), we bound the L_1-norm of T_1

$$\mathbb{E}_*\{|T_1|\} = |1 - 1/C_{\mathbf{h}}| = 1 + o(\|\mathbf{h}\|_2/\sqrt{n}). \tag{A.118}$$

Applying the bounded variance assumption for $\mathbf{g}_1$ and $\mathbf{g}_2$ and the rate of $C_{\mathbf{h}}$, we bound the L_1-norm of T_2

$$\begin{aligned} \mathbb{E}_*\{|T_2|\} = & \mathbb{E}_* \left\{ \left| \frac{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)}{\sqrt{n}} + \frac{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)}{\sqrt{N}} \right| \right\} / C_{\mathbf{h}} \\ \leq & \sqrt{\mathbb{E}_* \left[\{\mathbf{h}_1^\intercal \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)\}^2 / n \right]} / C_{\mathbf{h}} + \sqrt{\mathbb{E}_* \left[\{\mathbf{h}_2^\intercal \mathbf{g}_2(\mathbf{W}_i)\}^2 / N \right]} / C_{\mathbf{h}} \\ \leq & M\sqrt{\|\mathbf{h}_1\|_2^2/n + \|\mathbf{h}_2\|_2^2/N / C_{\mathbf{h}}} \\ = & M\sqrt{\|\mathbf{h}_1\|_2^2/n + \|\mathbf{h}_2\|_2^2/N} + o(\|\mathbf{h}\|_2/\sqrt{n}). \end{aligned} \tag{A.119}$$

For T_3, we repeat the analysis of upper bound for $C_{\mathbf{h}}$ (A.114) and (A.115) through alternative representation (A.113) with truncation indicator Ξ defined in (A.112),

$$\begin{aligned}
&\mathbb{E}_*\{|T_3|\} \\
=&\mathbb{E}_*\left(\left|1+\frac{\mathbf{h}_1^\intercal\mathbf{g}_1(\mathbf{z}_i,\mathbf{w}_i)}{\sqrt{n}}+\frac{\mathbf{h}_2^\intercal\mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}-\left[1+\frac{\mathbf{h}_1^\intercal\mathbf{g}_1(\mathbf{z}_i,\mathbf{w}_i)}{\sqrt{n}}+\frac{\mathbf{h}_2^\intercal\mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right]_+\right|\right)/C_{\mathbf{h}} \\
=&\mathbb{E}_*\left\{\Xi_i\left|1+\frac{\mathbf{h}_1^\intercal\mathbf{g}_1(\mathbf{z}_i,\mathbf{w}_i)}{\sqrt{n}}+\frac{\mathbf{h}_2^\intercal\mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right|\right\}/C_{\mathbf{h}} \\
\leq&2M\|\mathbf{h}\|_2^2/n \\
=&o\left(\|\mathbf{h}\|_2/\sqrt{n}\right). \qquad\qquad (\text{A.120})
\end{aligned}$$

Combining the rates (A.118)-(A.120) and the decomposition (A.117), we have shown

$$\|f_*-f_{\mathbf{h}}\|_{\mathsf{TV}}\leq\mathbb{E}_*(|T_1|)+\mathbb{E}_*(|T_2|)+\mathbb{E}_*(|T_3|)\leq M\sqrt{\|\mathbf{h}_1\|_2^2/n+\|\mathbf{h}_2\|_2^2/N}+o(\|\mathbf{h}\|_2/\sqrt{n}).$$

$\blacksquare$

Lemma A26 *Consider the settings detailed in the proof of Theorem 13, i. e. the two-way tilted density for* $\mathbf{Z},\mathbf{W}\mid\mathbf{h}$

$$f_{\mathbf{h}}(\mathbf{z}_i,\mathbf{w}_i)=f_*(\mathbf{z}_i,\mathbf{w}_i)\left[1+\frac{\mathbf{h}_1^\intercal\mathbf{g}_1(\mathbf{z}_i,\mathbf{w}_i)}{\sqrt{n}}+\frac{\mathbf{h}_2^\intercal\mathbf{g}_2(\mathbf{w}_i)}{\sqrt{N}}\right]_+/C_{\mathbf{h}}$$

with the truncated Gaussian prior for $\mathbf{h}$,

$$(\mathbf{h}_1^\intercal,\mathbf{h}_2^\intercal)^\intercal\sim p(\mathbf{h};c,\mathbb{A})=\frac{\phi(\mathbf{h},\mathbf{0},\mathbb{A})I(\|\mathbf{h}\|_2\leq c)}{\int_{\|\mathbf{h}\|_2\leq c}\phi(\mathbf{h},\mathbf{0},\mathbb{A})d\mathbf{h}},\ \phi(\mathbf{v},\boldsymbol{\mu},\Sigma)=\frac{\exp\left(-(\mathbf{v}-\boldsymbol{\mu})^\intercal\Sigma^{-1}(\mathbf{v}-\boldsymbol{\mu})/2\right)}{(2\pi)^{-q}\det(\Sigma)^{-1/2}}.$$

Define the marginal distribution of i.i.d. data $\mathscr{D}_N=\{(R_i,R_i\mathbf{Z}_i,\mathbf{W}_i):i=1,\ldots,N\}$ *as*

$$\mathscr{P}(\mathscr{D}_N)=\int_{\mathbf{h}}\prod_{i=1}^N\{\rho f_{\mathbf{h}}(\mathbf{Z}_i,\mathbf{W}_i)\}^{R_i}\left\{(1-\rho)\int_{\mathbf{v}}f_{\mathbf{h}}(\mathbf{v},\mathbf{W}_i)d\mathbf{v}\right\}^{1-R_i}p(\mathbf{h};c,\mathbb{A})d\mathbf{h}$$

and posterior $\mathbf{h}\mid\mathscr{D}_N$ *as*

$$\mathscr{Q}(\mathbf{h}\mid\mathscr{D}_N)=\frac{\prod_{i=1}^N\{\rho f_{\mathbf{h}}(\mathbf{Z}_i,\mathbf{W}_i)\}^{R_i}\left\{(1-\rho)\int_{\mathbf{v}}f_{\mathbf{h}}(\mathbf{v},\mathbf{W}_i)d\mathbf{v}\right\}^{1-R_i}p(\mathbf{h};c,\mathbb{A})}{\mathscr{P}(\mathscr{D}_N)}.$$

With finite variances of $\mathbf{g}_1(\mathbf{Z},\mathbf{W})$ *and* $\mathbf{g}_2(\mathbf{W})$*, the posterior* $\mathscr{Q}(\mathbf{h}\mid\mathscr{D}_N)$ *is approximated by the Gaussian posterior* $\phi(\mathbf{h},\widetilde{\boldsymbol{\mu}},\widetilde{\mathcal{V}})$,

$$\lim_{c,N\to\infty}\int\left\|\mathscr{Q}(\mathbf{h}\mid\mathscr{D}_N)-\phi(\mathbf{h},\widetilde{\boldsymbol{\mu}},\widetilde{\mathcal{V}})\right\|_{\mathsf{TV}}\mathscr{P}(\mathscr{D}_N)d\mathscr{D}_N=0.$$

Proof [Proof of Lemma A26] The proof extends Theorem 2 page 15 of Duchi (2021) to two-way tilted sub-models for characterizing semi-supervised learning setting. it suffices to analyze the difference in conditional densities. In the proof of Theorem 13, we defined the empirical processes

$$\mathbf{V} = (\mathbf{V}_1^\intercal, \mathbf{V}_2^\intercal)^\intercal,\ \mathbf{V}_1 = -\sum_{i=1}^{N} R_i \mathbf{g}_1(\mathbf{Z}_i, \mathbf{W}_i)/\sqrt{n},\ \mathbf{V}_2 = -\sum_{i=1}^{N} \mathbf{g}_2(\mathbf{W}_i)/\sqrt{N} \to \mathbf{V}_2$$

characterizing the local asymptotic normality (LAN) property of the two-way tilted sub-model. Consider the event indicator

$$\mathscr{E}_{N,b} = \mathrm{I}\left\{\left\|\mathcal{V}_\psi^{-1}\mathbf{V}\right\|_2 \le b\right\}. \tag{A.121}$$

With finite variances of $\mathbf{g}_1(\mathbf{Z}, \mathbf{W})$ and $\mathbf{g}_2(\mathbf{W})$, we apply Le Cam and Yang (2000) Chapter 6.3 Proposition 2 to obtain

a) Event $\mathscr{E}_{N,b}$ occurs with large probability: there exists sufficiently large $b_{c,\varepsilon}$ and $N_{c,\varepsilon}$ such that

$$\mathbb{E}_{\mathbf{h}}(\mathscr{E}_{N,b}) \ge 1 - \varepsilon,\ \forall \|\mathbf{h}\|_2 \le c, N \ge N_{c,\varepsilon}, b \ge b_{c,\varepsilon}; \tag{A.122}$$

b) Approximation of tilted model

$$d\mathscr{M}_{\mathbf{h}}(\mathscr{D}_N) = \prod_{i=1}^{N} \{\rho f_{\mathbf{h}}(\mathbf{Z}_i, \mathbf{W}_i)\}^{R_i} \left\{(1-\rho)\int_{\mathbf{V}} f_{\mathbf{h}}(\mathbf{v}, \mathbf{W}_i) d\mathbf{v}\right\}^{1-R_i} d\mathscr{D}_N$$

by Gaussian model

$$d\mathscr{G}_{\mathbf{h}}(\mathscr{D}_N) = \exp\left\{-\frac{1}{2}(\mathbf{h} - \mathcal{V}_\psi^{-1}\mathbf{V})^\intercal \mathcal{V}_\psi (\mathbf{h} - \mathcal{V}_\psi^{-1}\mathbf{V})\right\} d\mathscr{M}_{\mathbf{0}}(\mathscr{D}_N)$$

$$\lim_{N\to\infty} \sup_{\|\mathbf{h}\|\le c} \int \mathscr{E}_{N,b} \left|d\mathscr{M}_{\mathbf{h}}(\mathscr{D}_N) - d\mathscr{G}_{\mathbf{h}}(\mathscr{D}_N)\right|. \tag{A.123}$$

In the following, we define a series models to link the exact posterior and its Gaussian approximation. First, we define the model restricted to the "good set" on which $\mathscr{E} = 1$ in the model,

$$\begin{aligned} d\mathscr{M}_{\mathbf{h}}^{\mathscr{E}}(\mathscr{D}_N) &= \mathscr{E} d\mathscr{M}_{\mathbf{h}}(\mathscr{D}_N), \\ d\mathscr{G}_{\mathbf{h}}^{\mathscr{E}}(\mathscr{D}_N) &= \exp\left\{-\frac{1}{2}(\mathbf{h} - \mathcal{V}_\psi^{-1}\mathbf{V})^\intercal \mathcal{V}_\psi (\mathbf{h} - \mathcal{V}_\psi^{-1}\mathbf{V})\right\} d\mathscr{M}_{\mathbf{0}}^{\mathscr{E}}(\mathscr{D}_N). \end{aligned} \tag{A.124}$$

Using the newly defined notations in (A.123), we have

$$\lim_{N\to\infty} \sup_{\|\mathbf{h}\|\le c} \|\mathscr{M}_{\mathbf{h}}^{\mathscr{E}} - \mathscr{G}_{\mathbf{h}}^{\mathscr{E}}\|_{\mathsf{TV}} = 0. \tag{A.125}$$

Next, we define the exact, approximate Gaussian and $\mathscr{E}$-restricted joint distributions with truncated prior and another approximate Gaussian joint distributions with (untruncated) Gaussian prior

$$d\mathscr{J}(\mathscr{D}_N, \mathbf{h}) = d\mathscr{M}_{\mathbf{h}}(\mathscr{D}_N) p(\mathbf{h}; c, \mathbb{A}) d\mathbf{h},$$

$$\begin{aligned}
d\mathscr{J}^{\mathscr{G}}(\mathscr{D}_N,\mathbf{h}) &= d\mathscr{G}^{\mathscr{E}}_{\mathbf{h}}(\mathscr{D}_N)p(\mathbf{h};c,\mathbb{A})d\mathbf{h},\\
d\mathscr{J}^{\mathscr{E}}(\mathscr{D}_N,\mathbf{h}) &= d\mathscr{M}^{\mathscr{E}}_{\mathbf{h}}(\mathscr{D}_N)p(\mathbf{h};c,\mathbb{A})d\mathbf{h},\\
d\mathscr{J}^{\mathscr{P}}(\mathscr{D}_N,\mathbf{h}) &= d\mathscr{G}^{\mathscr{E}}_{\mathbf{h}}(\mathscr{D}_N)\phi(\mathbf{h},\mathbf{0},\mathbb{A})d\mathbf{h}.
\end{aligned} \tag{A.126}$$

Since the truncated prior $p(\mathbf{h};c,\mathbb{A})$ restricts $\mathbf{h}$ to $\|\mathbf{h}\|_2 \leq c$, we may bound the difference between $d\mathscr{J}^{\mathscr{G}}$ and $d\mathscr{J}^{\mathscr{E}}$ by (A.125),

$$\lim_{N\to\infty}\|\mathscr{J}^{\mathscr{G}}_{\mathbf{h}} - \mathscr{J}^{\mathscr{E}}_{\mathbf{h}}\|_{\mathsf{TV}} \leq \lim_{N\to\infty}\int \|\mathscr{G}^{\mathscr{E}}_{\mathbf{h}} - \mathscr{M}^{\mathscr{E}}_{\mathbf{h}}\|_{\mathsf{TV}} p(\mathbf{h};c,\mathbb{A})d\mathbf{h} = 0. \tag{A.127}$$

Since $\mathbb{E}_{\mathbf{h}}\{\mathscr{E}\} \geq 1-\varepsilon$ uniformly in $\|\mathbf{h}\|_2 \leq c$, we can control the error from restricting measure in $\{\mathscr{E}=1\}$ for sufficiently large N,

$$\|\mathscr{J} - \mathscr{J}^{\mathscr{E}}\|_{\mathsf{TV}} \leq \int \|\mathscr{M}^{\mathscr{E}}_{\mathbf{h}} - \mathscr{M}^{\mathscr{E}}_{\mathbf{h}}\|_{\mathsf{TV}} p(\mathbf{h};c,\mathbb{A})d\mathbf{h} = \int (1-\mathscr{E})p(\mathbf{h};c,\mathbb{A})d\mathbf{h} \leq \varepsilon. \tag{A.128}$$

As the final link, we control the error from truncation in prior

$$\begin{aligned}
\|\mathscr{J}^{\mathscr{G}} - \mathscr{J}^{\mathscr{P}}\|_{\mathsf{TV}} &= \int |d\mathscr{G}^{\mathscr{E}}_{\mathbf{h}}(\mathscr{D}_N)p(\mathbf{h};c,\mathbb{A})d\mathbf{h} - d\mathscr{G}^{\mathscr{E}}_{\mathbf{h}}(\mathscr{D}_N)\phi(\mathbf{h},\mathbf{0},\mathbb{A})d\mathbf{h}|\\
&\leq \int \sup_{\mathbf{h}} d\mathscr{G}^{\mathscr{E}}_{\mathbf{h}}(\mathscr{D}_N)\|\phi(\mathbf{h},\mathbf{0},\mathbb{A}) - p(\mathbf{h};c,\mathbb{A})\|_{\mathsf{TV}}\\
&\leq \int \exp\left(\frac{1}{2}\mathbf{V}^{\intercal}\mathcal{V}_{\psi}^{-1}\mathbf{V}\right) d\mathscr{M}^{\mathscr{E}}_{\mathbf{0}}(\mathscr{D}_N)\|\phi(\mathbf{h},\mathbf{0},\mathbb{A}) - p(\mathbf{h};c,\mathbb{A})\|_{\mathsf{TV}}.
\end{aligned} \tag{A.129}$$

Over the restricted measure $\mathscr{M}^{\mathscr{E}}_{\mathbf{0}}$, $\|\mathcal{V}_{\psi}^{-1}\mathbf{V}\|_2 \leq b$ is bounded, so

$$\mathscr{E} = 1 : \exp\left(\frac{1}{2}\mathbf{V}^{\intercal}\mathcal{V}_{\psi}^{-1}\mathbf{V}\right) \leq \exp(\|\mathcal{V}_{\psi}\|_2 b^2/2) = O(1).$$

By choosing sufficiently large c such that

$$\int_{\|\mathbf{h}\|_2 > c} \phi(\mathbf{h},\mathbf{0},\mathbb{A})d\mathbf{h} \leq \varepsilon/\{2\exp(\|\mathcal{V}_{\psi}\|_2 b^2/2)\},$$

we can control the truncation error in prior

$$\|\phi(\mathbf{h},\mathbf{0},\mathbb{A}) - p(\mathbf{h};c,\mathbb{A})\|_{\mathsf{TV}} \leq \varepsilon/\exp(\|\mathcal{V}_{\psi}\|_2 b^2/2).$$

Apply the two bounds above to (A.129), we obtain

$$\|\mathscr{J}^{\mathscr{G}} - \mathscr{J}^{\mathscr{P}}\|_{\mathsf{TV}} \leq \varepsilon. \tag{A.130}$$

Through (A.127), (A.128), (A.130) and (A.130), we have established the approximation among joint measures defined in (A.126) and in particular

$$\limsup_{N\to\infty}\|\mathscr{J} - \mathscr{J}^{\mathscr{P}}\|_{\mathsf{TV}} \leq 2\varepsilon. \tag{A.131}$$

To derive the approximation in posterior distribution from the approximation of joint distribution, we invoke the following lemma,

Lemma A27 (Le Cam and Yang (2000) Chapter 6.4 Lemma 2 page 136) *Let*

$$\mathscr{M}_j(d\mathscr{D}, d\boldsymbol{\theta}) = \nu_j(d\mathscr{D})\mathscr{M}_j(d\boldsymbol{\theta} \mid \mathscr{D}),\ j = 1, 2,$$

be two joint measure for $(\mathscr{D}, \boldsymbol{\theta})$*. Then, the difference of conditional distributions in total variation is controlled by the difference of joint distributions in total variation*

$$\int \|\mathscr{M}_1(d\boldsymbol{\theta} \mid \mathscr{D}) - \mathscr{M}_2(d\boldsymbol{\theta} \mid \mathscr{D})\|_{\mathsf{TV}}\, |\nu_1(d\mathscr{D}) + \nu_2(d\mathscr{D})| \leq 4\|\mathscr{M}_1(d\mathscr{D}, d\boldsymbol{\theta}) - \mathscr{M}_2(d\mathscr{D}, d\boldsymbol{\theta})\|_{\mathsf{TV}}.$$

Notice that the joint measure $\mathscr{J}^{\mathscr{P}}$ has (untruncated) Gaussian prior and approximated Gaussian model, we can explicitly derive its posterior

$$d\mathscr{J}^{\mathscr{P}} = \phi(\mathbf{h}, \widetilde{\boldsymbol{\mu}}, \widetilde{\mathcal{V}})d\mathbf{h}\phi(\mathbf{V}, \mathbf{0}, \mathcal{V}_\varepsilon + \mathbb{A})d\mathscr{M}_{\mathbf{0}}(\mathscr{D}_N).$$

Applying Lemma A27 with (A.131), we have proven

$$\begin{aligned}
&\lim_{N\to\infty} \int \left\|\mathscr{Q}(\mathbf{h} \mid \mathscr{D}_N) - \phi(\mathbf{h}, \widetilde{\boldsymbol{\mu}}, \widetilde{\mathcal{V}})\right\|_{\mathsf{TV}} \mathscr{P}(\mathscr{D}_N)d\mathscr{D}_N \\
\leq &\lim_{N\to\infty} \int \left\|\mathscr{Q}(\mathbf{h} \mid \mathscr{D}_N) - \phi(\mathbf{h}, \widetilde{\boldsymbol{\mu}}, \widetilde{\mathcal{V}})\right\|_{\mathsf{TV}} |\mathscr{P}(\mathscr{D}_N)d\mathscr{D}_N + \phi(\mathbf{V}, \mathbf{0}, \mathcal{V}_\varepsilon + \mathbb{A})d\mathscr{M}_{\mathbf{0}}(\mathscr{D}_N)| \\
\leq &\lim_{N\to\infty} 4\|\mathscr{J} - \mathscr{J}^{\mathscr{P}}\|_{\mathsf{TV}} \\
\leq &2\varepsilon.
\end{aligned}$$

Setting $c \to \infty$ thus $\varepsilon \to 0$ yields

$$\lim_{c\to\infty} \lim_{N\to\infty} \int \left\|\mathscr{Q}(\mathbf{h} \mid \mathscr{D}_N) - \phi(\mathbf{h}, \widetilde{\boldsymbol{\mu}}, \widetilde{\mathcal{V}})\right\|_{\mathsf{TV}} \mathscr{P}(\mathscr{D}_N)d\mathscr{D}_N = 0.$$

We have proven that the asymptotic posterior follows the Gaussian distribution $N(\widetilde{\boldsymbol{\mu}}, \widetilde{\mathcal{V}})$.
■

References

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: https://doi.org/10.1111/j.1541-0420.2005.00377.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00377.x.

Victoria L. Bartlett, Sanket S. Dhruva, Nilay D. Shah, Patrick Ryan, and Joseph S. Ross. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. *JAMA Network Open*, 2(10):e1912869–e1912869, 10 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.12869. URL https://doi.org/10.1001/jamanetworkopen.2019.12869.

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. doi: https://doi.org/10.1002/cpt.1658. URL `https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.1658`.

Janet M. Begun, W. J. Hall, Wei-Min Huang, and Jon A. Wellner. Information and Asymptotic Efficiency in Parametric-Nonparametric Models. *The Annals of Statistics*, 11(2): 432 – 452, 1983. doi: 10.1214/aos/1176346151. URL `https://doi.org/10.1214/aos/1176346151`.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017. doi: https://doi.org/10.3982/ECTA12723. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12723`.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies*, 81(2):608–650, 11 2013. ISSN 0034-6527. doi: 10.1093/restud/rdt044. URL `https://doi.org/10.1093/restud/rdt044`.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity Double Robust Inference of Average Treatment Effects. *arXiv*, page 1905.00744, 2019.

Abhishek Chakrabortty, Jiarui Lu, T. Tony Cai, and Hongzhe Li. High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv*, page 1911.11345, 2019.

David Cheng, Ashwin N. Ananthakrishnan, and Tianxi Cai. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, 77(2):413–423, 2021. doi: https://doi.org/10.1111/biom.13298. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13298`.

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

John Duchi. A few notes on contiguity, asymptotics, and local asymptotic normality, March 2021. URL `https://web.stanford.edu/class/stats300b/Notes/contiguity-and-asymptotics.pdf`.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928 – 961, 2004. doi: 10.1214/009053604000000256. URL `https://doi.org/10.1214/009053604000000256`.

Max H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189:1–23, 2015.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021. doi: https://doi.org/10.3982/ECTA16901. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16901.

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. doi: https://doi.org/10.1002/pds.5222. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.5222.

Sandra D. Griffith, Melisa Tucker, Bryan Bowser, Geoffrey Calkins, Che-hsu (Joe) Chang, Ellie Guardino, Sean Khozin, Josh Kraut, Paul You, Deb Schrag, and Rebecca A. Miksad. Generating real-world tumor burden endpoints from electronic health record data: Comparison of recist, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. *Advances in Therapy*, 36(8):2122–2136, Aug 2019. ISSN 1865-8652. doi: 10.1007/s12325-019-00970-1. URL https://doi.org/10.1007/s12325-019-00970-1.

M.A. Hernan and J.M. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Taylor & Francis, 2023. ISBN 9781420076165. URL https://books.google.com/books?id=_KnHIAAACAAJ.

J. Hou, J. Bradic, and R. Xu. Treatment effect estimation under additive hazards models with high-dimensional confounding. *Journal of the American Statistical Association*, page In press, 2021a.

Jue Hou, Zijian Guo, and Tianxi Cai. Surrogate assisted semi-supervised inference for high dimensional risk prediction, 2021b.

Jue Hou, Nicole Kim, Tianrun Cai, Kumar Dahal, Howard Weiner, Tanuja Chitnis, Tianxi Cai, and Zongqi Xia. Comparison of dimethyl fumarate vs fingolimod and rituximab vsnatalizumab for treatment of multiple sclerosis. *JAMA Network Open*, page To appear, 2021c.

Jue Hou, Rachel Zhao, Tianrun Cai, Brett Beaulieu-Jones, Thany Seyok, Kumar Dahal, Qianyu Yuan, Xin Xiong, Clara-Lea Bonzel, Claire Fox, David C. Christiani, Thomas Jemielita, Katherine P. Liao, Kai-Li Liaw, and Tianxi Cai. Temporal Trends in Clinical Evidence of 5-Year Survival Within Electronic Health Records Among Patients With Early-Stage Colon Cancer Managed With Laparoscopy-Assisted Colectomy vs Open Colectomy. *JAMA Network Open*, 5(6):e2218371–e2218371, 06 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.18371. URL https://doi.org/10.1001/jamanetworkopen.2022.18371.

Jue Hou, Rachel Zhao, Jessica Gronsbell, Yucong Lin, Clara-Lea Bonzel, et al. Generate analysis-ready data for real-world evidence: Tutorial for harnessing electronic health records with advanced informatic technologies. *Journal of Medical Internet Research*, 25: e45662, 2023.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Cheng Ju, Joshua Schwab, and Mark J van der Laan. On adaptive propensity score truncation in causal inference. *Statistical Methods in Medical Research*, 28(6):1741–1760, 2019. doi: 10.1177/0962280218774817. URL https://doi.org/10.1177/0962280218774817. PMID: 29991330.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae099, 10 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae099. URL https://doi.org/10.1093/jrsssb/qkae099.

Lucien Le Cam and Grace Lo Yang. *Contiguity — Hellinger Transforms*, pages 34–49. Springer New York, New York, NY, 2000. ISBN 978-1-4612-1166-2. doi: 10.1007/978-1-4612-1166-2_3. URL https://doi.org/10.1007/978-1-4612-1166-2_3.

Katherine P Liao, Jiehuan Sun, Tianrun A Cai, Nicholas Link, Chuan Hong, Jie Huang, Jennifer E Huffman, Jessica Gronsbell, Yichi Zhang, Yuk-Lam Ho, et al. High-throughput multimodal automated phenotyping (map) with application to phewas. *Journal of the American Medical Informatics Association*, 26(11):1255–1262, 2019.

Dan-Yu Lin and Zhiliang Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994.

Molei Liu, Yi Zhang, and Doudou Zhou. Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3):559–588, 06 2021. ISSN 1368-4221. doi: 10.1093/ectj/utab019. URL https://doi.org/10.1093/ectj/utab019.

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

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, 2021/12/17/ 2012. ISSN 08834237. URL http://www.jstor.org/stable/41714783. Full publication date: November 2012.

Whitney K. Newey and James R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation, 2018.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012. doi: 10.1177/0962280210386207. URL https://doi.org/10.1177/0962280210386207. PMID: 21030422.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290910.

A Rotnitzky, E Smucler, and J M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 08 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa054. URL https://doi.org/10.1093/biomet/asaa054.

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

Charles J. Stone. Consistent Nonparametric Regression. *The Annals of Statistics*, 5(4): 595 – 620, 1977. doi: 10.1214/aos/1176343886. URL https://doi.org/10.1214/aos/1176343886.

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811 – 837, 2020. doi: 10.1214/19-AOS1824. URL https://doi.org/10.1214/19-AOS1824.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007. ISBN 9780387373454. URL https://books.google.com/books?id=xqZFi2EMB40C.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 2009. doi: 10.1007/b13794.

U.S. Cancer Statistics Working Group. U.s. cancer statistics data visualizations tool, based on 2021 submission data (1999-2019): U.s. department of health and human services, centers for disease control and prevention and national cancer institute, June 2022.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015. doi: 10.1080/01621459.2014.958155. URL https://doi.org/10.1080/01621459.2014.958155.

Yuhao Wang and Rajen D. Shah. Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders, 2020.

Yuan-Hong Xie, Ying-Xuan Chen, and Jing-Yuan Fang. Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduction and Targeted Therapy*, 5(1):22, Mar 2020. ISSN 2059-3635. doi: 10.1038/s41392-020-0116-z. URL https://doi.org/10.1038/s41392-020-0116-z.

Qianyu Yuan, Tianrun Cai, Chuan Hong, Mulong Du, Bruce E. Johnson, Michael Lanuti, Tianxi Cai, and David C. Christiani. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. *JAMA Network Open*, 4(7):e2114723–e2114723, 07 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.14723. URL https://doi.org/10.1001/jamanetworkopen.2021.14723.

Yinchi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk-Lam Ho, Ashwin N. Ananthakrishnan, Zonggi Xia, Stanley Y. Shaw, Vivian Gainer, Victor Castro, Nicholas Link, Jacqueline Honerlaw, Selena Huang, David Gagnon, Elizabeth W. Karlson, Robert M. Plenge, Peter Szolovits, Guergana Savova, Christopher O'Donnell, Shawn N. Murphy, J. Michael Gaziano, Isaac Kohane, Tianxi Cai, and Katherine P. Liao. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature Protocal*, page To appear, 2019.

Yuqian Zhang, Abhishek Chakrabortty, and Jelena Bradic. Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap. *Information and Inference: A Journal of the IMA*, 12(3):2066–2159, 07 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad021. URL https://doi.org/10.1093/imaiai/iaad021.