Journal of Machine Learning Research 27 (2026) 1-65 Submitted 7/25; Revised 6/26; Published 6/26

Domain Adaptation Targeting Heterogeneous and Imbalanced Subgroups

Doudou Zhou DDZHOU@NUS.EDU.SG
Department of Statistics and Data Science
National University of Singapore, Singapore

Mengyan Li MENGYANLI@BENTLEY.EDU
Department of Mathematical Sciences
Bentley University, Waltham, MA, USA

Yun Wang* YUNWANG@STU.PKU.EDU.CN
Department of Biostatistics
Peking University Health Science Center, Beijing, China

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

Molei Liu† MOLEILIU@BJMU.EDU.CN
Department of Biostatistics, Peking University Health Science Center;
Beijing International Center for Mathematical Research;
Center for Statistical Science, Peking University, Beijing, China

Editor: Brian Kulis

Abstract

Domain adaptation enables generalizable and efficient data-driven research. However, existing work has largely focused on domain adaptation for some intrinsically homogeneous target cohort, overlooking inherent heterogeneity within the target, which can exacerbate biases and unfairness in the presence of subgroups with imbalanced sample sizes. We develop a novel domain adaptation framework that addresses a more complicated target dataset that consists of heterogeneous and data-sparse subgroups and lacks gold-standard label observations. Our method simultaneously handles high-dimensionality, covariate shift, and outcome model heterogeneity by combining a model-assisted debiasing step used for covariate shift correction with an adaptive knowledge-guided sparsification procedure used to mitigate the issue of sample disparity. We also introduce a new model selection strategy to avoid negative knowledge transfer in the absence of labels in the target data. Our method is theoretically justified for being robust to nuisance model misspecification and adaptive to heterogeneity between the subgroups. Numerical experiments and two real-world applications, including genetic risk modeling of type 2 diabetes and prediction of mutation-induced protein stability changes, demonstrate the practical advantages of our method.

Keywords: Covariate shift; sampling disparity; model heterogeneity; double robustness; model calibration; knowledge transfer.

∗. Zhou, Li, and Wang are equal contributors.
†. Corresponding author.

1. Introduction

1.1 Background

Domain adaptation has been increasingly used when accurate labels are scarce or unavailable in a target population, but related labeled data are available from a source population. This setting arises broadly in scientific prediction problems where gold-standard target outcomes are expensive, delayed, or infeasible to collect. Existing work often focuses on adapting to an intrinsically homogeneous target cohort (e.g., Liu et al., 2023; Li et al., 2022). In many real-world application scenarios, however, the target population contains heterogeneous subgroups with imbalanced sample sizes, creating a harder problem: one must correct source–target distributional shift while protecting underrepresented subgroups from bias and negative transfer. Here, negative transfer means that borrowing information from a larger or related subgroup worsens estimation or prediction relative to the underrepresented-subgroup-only procedure.

Biobank studies linked to electronic health records (EHR) serve as a representative motivating example. Genetic disease risk modeling is critical for decision-making and knowledge discovery, especially in precision medicine (Moons et al., 2012). In these studies, the scientific target is often a risk model based on prespecified genetic variants or baseline clinical factors, while gold-standard disease labels are available only for a limited subset of samples because manual chart review is costly. Meanwhile, the target cohort may contain highly imbalanced ancestry or demographic subgroups, so a model transferred from the labeled source cohort can perform unevenly across these subgroups.

The prediction of protein stability changes $\Delta\Delta G$ provides another example. In our real-world study, S461 (Hernández et al., 2023) is used as the labeled source collection and S8754 (Xu et al., 2024) as the target collection whose labels are withheld during training. The two collections differ in data source, curation workflow, and sequence filtering, making source–target distributional shift plausible. We define majority and minority subgroups by assay environment: near-neutral-pH assays (pH 6–7) are substantially more represented than acidic-pH assays (pH 1–5). This distinction is important because the measured $\Delta\Delta G$ can depend on experimental conditions. Hence, a model dominated by near-neutral assays may not transfer reliably to the acidic-pH subgroup.

The same structure appears not just in biomedical applications. In remote sensing, labels may be sufficient for well-surveyed regions or common land-cover types, but scarce for rare habitats or imagery from new sensors. In industrial reliability assessment, failure labels are often available under standard laboratory protocols, but limited under extreme operating conditions or new production lines. These examples share the same structure: labeled outcomes are available in a related source population, target labels are unavailable during training, and the target population contains subgroups with both sample-size imbalance and different outcome models. Therefore, model heterogeneity and sampling disparity are not simply descriptive features of the data. If they are ignored, a transferred model or decision can be dominated by the majority subgroup, leading to biased and poor minority-subgroup prediction. In addition, naive borrowing from the majority subgroup can harm the minority subgroup when the subgroup-specific outcome models differ. These challenges motivate domain adaptation methods that explicitly account for target-subgroup heterogeneity, sample disparity, and negative transfer protection.

1.2 Related literature

Recent work relevant to our setting has studied two distinct issues: source–target covariate shift and outcome-model heterogeneity across related domains or subgroups. Covariate shift is a central challenge in domain adaptation and is often addressed by importance weighting, including kernel mean matching (Huang et al., 2007) and robust weighting schemes under strong shift (Reddi et al., 2015). In the broader machine learning literature, feature alignment methods provide another important class of unsupervised domain adaptation approaches. For example, CORAL (Sun et al., 2017) reduces the domain discrepancy by aligning second-order feature statistics without using target labels. Pseudo-labeling methods instead exploit unlabeled target samples by generating surrogate labels for target domain learning. Recent theory has studied this idea for kernel ridge regression (Wang, 2026; Kim et al., 2025). A related statistical line combines importance weighting with outcome imputation in doubly robust estimating equations (Chakrabortty et al., 2019; Liu et al., 2023; Qiu et al., 2024; Tian et al., 2024; Zhou et al., 2025). These methods improve robustness to nuisance model misspecification, but are mainly designed for homogeneous target cohorts and low-dimensional target models, whereas our setting involves high-dimensional sparse target models and imbalanced heterogeneous target subgroups.

Beyond covariate shift, the heterogeneity of the outcome model poses challenges to domain adaptation. Knowledge-guided transfer methods address this issue by assuming structural similarity between outcome models across related domains or tasks. Examples include source-guided shrinkage (Bastani, 2021), Trans-Lasso and its extensions (Li et al., 2022; Tian and Feng, 2023; Li et al., 2023), and semi-supervised transfer learning with covariate and model shift (Cai et al., 2024). In a related high-dimensional setting, Trans-Fusion (He et al., 2024) addresses covariate and model shifts through a fused-regularization strategy. These methods are effective for model shift problems, but generally require labeled target data or treat the target population as homogeneous. Thus, they do not directly address our label-free target setting with subgroup imbalance.

Several recent works study domain adaptation problems with multiple sources. Specifically, Zhan et al. (2024) and Guha et al. (2025) combine heterogeneous sources using individual probabilistic weights. Li and Zhang (2025) formulate domain adaptation with complex structures as a problem of tensor completion. These works are conceptually related because the four source/target–subgroup pairs in our setting can be viewed as distinct distributions. Our problem setup is more structured and asymmetric than in the existing work. Specifically, our target is the high-dimensional sparse model for the underrepresented target subgroup; the corresponding labeled source data are used for domain adaptation without labels in the target; and the majority subgroup is used only as auxiliary knowledge whose transferability must be assessed. Existing multi-source domain adaptation methods do not directly address subgroup imbalance, target labels unavailable for training, high-dimensionality, and potential negative transfer simultaneously.

Our work is also related to source-free domain adaptation, where one adapts a source model to an unlabeled target domain without direct access to the source data (Wang et al., 2024). This literature is motivated in part by storage, privacy, and deployment constraints. Our setting differs in both data access and inferential target: we use labeled source data together with unlabeled target data, and our goal is to estimate a subgroup-specific target

parameter rather than only to adapt a black-box source predictor. Nevertheless, source-free and multi-source domain adaptation raise related questions about when and how information should be transferred, and we discuss possible extensions of our setting in Section 7.

Finally, our theoretical framework is related to model-assisted semiparametric estimation with complex nuisance models, particularly in the context of conditional models and heterogeneous treatment effects (HTE). Recent works have extended the double machine learning (DML) framework (Chernozhukov et al., 2018) to this setting, including Fan et al. (2022) and Semenova and Chernozhukov (2021) for low-dimensional effect modifiers with high-dimensional adjustment covariates, Kennedy (2023) for smooth or sparse HTE structures, and calibrated or debiased inference under model misspecification (e.g., Kato, 2024; Baybutt and Navjeevan, 2026; Tan, 2020). However, these methods are not applicable to our setting involving a high-dimensional sparse target model under covariate shift, where we seek both robustness to nuisance model misspecification and tolerance to slow nuisance convergence. Simultaneously achieving these is technically challenging due to the difficulty in controlling regularization bias and establishing asymptotic linearity.

1.3 Our contribution

To mitigate the methodological gap in domain adaptation for targets with heterogeneous and imbalanced subgroups, we study a practically important and technically challenging problem that involves three interrelated difficulties: (1) covariate shift between the source and target domains, (2) sampling disparity and model heterogeneity across subgroups within the target domain, and (3) high-dimensionality of the outcome model. These challenges are further compounded by the lack of target labels available for training. To address them, we propose a new framework for domain Adaptation to targets with heterogeneous and IMbalanced Subgroups (AIMS). AIMS simultaneously delivers doubly robust covariate shift correction, knowledge transfer between target subgroups, and negative transfer protection within a single unified approach for high-dimensional target domains without outcome labels available for training.

Methodological contributions. AIMS consists of three components, each addressing a specific challenge that prior work cannot handle jointly:

(i) *Doubly robust covariate shift correction with bifold debiasing (Algorithm 1).* AIMS combines importance weighting and imputation in a doubly robust estimating equation, then uses bifold bias correction, consisting of nodewise Lasso debiasing and moment calibration, to remove regularization and nuisance estimation bias. A subsequent thresholding step recovers ℓ_2-consistency. Existing doubly robust covariate shift methods (Liu et al., 2023; Zhou et al., 2025) are restricted to low-dimensional outcome models.

(ii) *Knowledge transfer from majority to minority through thresholded contrast estimation (Algorithm 2).* AIMS transfers knowledge from the majority to the minority subgroup without using target labels for training. It uses the majority model as an "offset" and estimates only the subgroup contrast. When this contrast is sparse, the minority estimator gains accuracy by borrowing information from the majority subgroup. This

differs from existing knowledge transfer methods in the model shift setting (Cai et al., 2024; He et al., 2024), which require labeled target samples.

(iii) *Label-free negative transfer protection via model ensemble (Algorithm 3).* AIMS protects against negative transfer when the subgroup contrast is dense and the majority cannot help in learning the minority model. It constructs a dense and debiased reference vector and uses cross-fitting and exponential weighting to aggregate the minority-only and knowledge transfer estimators. Existing negative transfer protection methods (Tian and Feng, 2023; Cai et al., 2024) compare candidate estimators using labeled target data, which is unavailable in our setting.

Theoretical contributions. We establish three main results that demonstrate the theoretical advantages of AIMS over the existing work. First, our calibrated estimator is model doubly robust. The closest comparator, Zhou et al. (2025), establishes a similar property but only under a low-dimensional outcome model, and their analysis does not extend to our error rates when the dimension diverges. Second, for the estimation of the minority model, AIMS matches, under a mild nuisance error condition, the rate of supervised knowledge transfer methods that use target labels for training (Tian and Feng, 2023; He et al., 2024; Cai et al., 2024). Third, our adaptive estimator achieves the faster of the two convergence rates of the minority-only and knowledge transfer estimators, up to a negative transfer protection cost. This is comparable to existing work that requires labeled target samples for model selection (e.g., Cai et al., 2024).

Practical relevance. By jointly addressing covariate shift, model shift, and sample imbalance when target outcome labels are unavailable for training, AIMS provides a novel and practical solution for fair and equitable domain adaptation. This is particularly relevant in biomedical research, where minority subgroups often lack labeled outcomes, yet demand accurate and fair risk prediction. Similarly, in protein engineering, available stability measurements may be collected under heterogeneous and imbalanced experimental conditions. In this scenario, variants relevant to the setting of the target design can be underrepresented in training data and may lack reliable target-setting labels, despite the need for accurate predictions of stability in the downstream design.

2. Formulation of the Problem

Let $Y \in \mathbb{R}$ be the outcome of interest, $R \in \{0, 1\}$ be the subgroup indicator ($R = 1$ for the majority, $R = 0$ for the minority), and $\mathbf{Z} = (\mathbf{X}^\intercal, \mathbf{W}^\intercal)^\intercal \in \mathbb{R}^{q+p}$ be the covariate vector, where $\mathbf{X} = (X_1, \ldots, X_q)^\intercal$ is the vector of risk factors for predicting Y, with $X_1 = 1$ representing the intercept, and $\mathbf{W} = (W_1, \ldots, W_p)^\intercal$ is the vector of auxiliary covariates that may be informative about Y and the source–target shift but are not included in the final target risk model. Both $\mathbf{X}$ and $\mathbf{W}$ can be high-dimensional. Throughout the paper, the terms majority and minority refer to subgroup sample sizes rather than to the class distribution of Y. The subgroup indicator R is assumed to be observed and may be determined by an observed attribute, such as ancestry, sex, age group, or experimental condition. In our formulation, R indexes subgroup-specific distributions and target parameters rather than being treated as an ordinary component of $\mathbf{X}$ in a pooled model.

Our primary goal is to construct a prediction model of $Y \sim \mathbf{X}$ for the minority ($R = 0$) in the target population $\mathcal{T}$:

$$\mathbb{E}_{\mathcal{T}}[Y \mid \mathbf{X}, R = 0] = g(\mathbf{X}^\intercal \bar{\boldsymbol{\beta}}^{[0]}), \tag{1}$$

where $\mathbb{E}_{\mathcal{T}}[\cdot]$ denotes the expectation operator on $\mathcal{T}$, $g(\cdot)$ is a known link function, and $\bar{\boldsymbol{\beta}}^{[0]}$ is a high-dimensional and sparse coefficient vector. We do not require the *working* model (1) to hold and define the population parameter $\bar{\boldsymbol{\beta}}^{[0]}$ as the solution to

$$\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{Y - g(\mathbf{X}^\intercal \boldsymbol{\beta})\} \mid R = 0] = \mathbf{0}. \tag{2}$$

This solution corresponds to the ordinary least squares regression when $g(x) = x$ and logistic regression when $g(x) = 1/(1+e^{-x})$. The generalized linear form is used as a working target model rather than a strict assumption on the true conditional mean. By (2), $\bar{\boldsymbol{\beta}}^{[0]}$ represents the projection of the conditional mean onto the risk factors $\mathbf{X}$. This choice is motivated by biomedical risk modeling, where GLMs provide interpretable adjusted associations and deployable risk scores. In high-dimensional and imbalanced settings, sparsity further yields a tractable structure.

We assume Y is only observed in the source cohort $\mathcal{S}$, not in the target $\mathcal{T}$, and observe a dataset $\mathscr{D} = \{\mathbf{D}_i = (S_i Y_i, \mathbf{Z}_i^\intercal, S_i, R_i)^\intercal : i = 1, 2, \ldots, n\}$. Here, $R_i \in \{0, 1\}$ indicates the group and S_i is the source/target indicator ($S_i = 1$ for $\mathcal{S}$, $S_i = 0$ for $\mathcal{T}$). The sample size of each stratum is $n_{\mathcal{S},r} := \sum_{i=1}^n \mathbb{I}(S_i = 1, R_i = r)$ and $n_{\mathcal{T},r} := \sum_{i=1}^n \mathbb{I}(S_i = 0, R_i = r)$ for $r \in \{0, 1\}$, where $\mathbb{I}(\cdot)$ is the indicator function. In addition, we use $n_{\mathcal{S}} := n_{\mathcal{S},0} + n_{\mathcal{S},1}$ and $n_{\mathcal{T}} := n_{\mathcal{T},0} + n_{\mathcal{T},1}$ to represent the sample sizes of $\mathcal{S}$ and $\mathcal{T}$ so we have $n = n_{\mathcal{S}} + n_{\mathcal{T}}$. We consider an imbalanced sampling scenario on both $\mathcal{S}$ and $\mathcal{T}$ where $n_{\mathcal{S},1} \gg n_{\mathcal{S},0}$ and $n_{\mathcal{T},1} \gg n_{\mathcal{T},0}$. A common example is that $R = 1$ represents the European majority ancestry group and $R = 0$ represents the underrepresented African group. Other examples include better-represented versus underrepresented protein families, mutation classes, or assay platforms in protein mutation studies.

We assume that for $(s, r) \in \{0, 1\}^2$, the samples of $(Y, \mathbf{Z})$ are generated by

$$Y, \mathbf{Z} \mid S = s, R = r \sim p_{\mathbf{Z}|S=s,R=r}(\mathbf{z}) \cdot p_{Y|\mathbf{Z},R=r}(y), \tag{3}$$

where $p_{\mathbf{Z}|S=s,R=r}(\cdot)$ and $p_{Y|\mathbf{Z},R=r}(\cdot)$ denote the conditional densities (or probability mass functions, as appropriate) of $\mathbf{Z}$ given $S = s, R = r$ and of Y given $\mathbf{Z}, R = r$, respectively. This defines a domain adaptation problem where, given $R = r$, the distribution of covariates $\mathbf{Z} = (\mathbf{X}^\intercal, \mathbf{W}^\intercal)^\intercal$ is assumed to be different between $\mathcal{S}$ (source) and $\mathcal{T}$ (target), while the conditional distribution of $Y \mid \mathbf{Z}, R = r$ is assumed to be the same. Thus, within each subgroup, the source and target samples may have different covariate distributions, but share the same conditional outcome mechanism given the full covariate vector $\mathbf{Z}$. Conditional on $R = r$, this is a typical covariate shift scenario frequently encountered in practice. For example, EHR cohorts can vary in their coding systems and versions (Guo et al., 2021), and, notably, MGB underwent a change in its coding system around 2015, resulting in distributional shifts of EHR features. Covariate shifts can also be caused by the variation of demographic characteristics between different institutions or time intervals and the evolving practice patterns of healthcare professionals (Braithwaite, 2018). In protein applications,

covariate shift may arise from differences between assay platforms, experimental protocols, protein families, or mutation distributions.

We introduce the density ratio $h_r^\star(\mathbf{z}) := p_{\mathbf{Z}|S=0,R=r}(\mathbf{z})/p_{\mathbf{Z}|S=1,R=r}(\mathbf{z})$ to characterize the covariate shift between the source and target within subgroup $r \in \{0,1\}$. Under such an assumption in (3), the conditional mean function $m_r^\star(\mathbf{z}) := \mathbb{E}[Y \mid \mathbf{Z} = \mathbf{z}, R = r]$ remains the same between $\mathcal{S}$ and $\mathcal{T}$. Nevertheless, we have $\mathbb{E}_{\mathcal{T}}[Y \mid \mathbf{X}, R = 0] \neq \mathbb{E}_{\mathcal{S}}[Y \mid \mathbf{X}, R = 0]$ due to the joint distributional shift of $\mathbf{X}$ and the auxiliary $\mathbf{W}$. Consequently, directly regressing Y against $\mathbf{X}$ on $\mathcal{S}$ will typically produce a biased estimator for the target $\bar{\boldsymbol{\beta}}^{[0]}$, regardless of whether model (1) is correctly specified. Our first technical challenge is to correct this bias when using labeled observations from $\mathcal{S}$ to estimate the model parameters defined on $\mathcal{T}$. Standard domain adaptation via importance weighting or imputation is challenged by high-dimensionality and subgroup disparities.

Remark 1 *In EHR studies such as the application in Section 6.1,* $\mathbf{X}$ *includes genetic variants and demographic variables to model disease risk* Y*, while* $\mathbf{W}$ *contains EHR proxies such as diagnostic codes and laboratory tests. The focus is on predicting* Y *using* $\mathbf{X}$ *rather than all covariates in* $\mathbf{Z}$ *because the auxiliary* $\mathbf{W}$ *is not available at baseline, whereas biomedical discoveries and deployment decisions are often meant to be based on baseline risk factors. More generally, the distinction between* $\mathbf{X}$ *and* $\mathbf{W}$ *is determined by the scientific or deployment target, not by their predictive strength. Auxiliary variables* $\mathbf{W}$ *may be strongly associated with* Y *but excluded from the target model because they are derived from post-baseline information, unavailable at deployment, or difficult to interpret. Such* $\mathbf{W}$ *can appear in diverse domains. For example, in longitudinal clinical studies,* $\mathbf{W}$ *can include an early response or endpoint, such as short-term biomarker changes or tumor response, which may predict the long-term endpoint* Y *but may not be available at the decision time (Prentice, 1989). In the industrial reliability examination,* $\mathbf{W}$ *may include accelerated stress-test readouts or early degradation signals.*

Usually, the distributional shift between the source and target populations, such as the temporal shift in the clinical profile of the biobank participants, cannot be fully captured by the difference in $\mathbf{X}$*. Thus, it is important to include auxiliary variables* $\mathbf{W}$ *to adjust for their covariate shift and transfer the knowledge of* $Y \mid \mathbf{X}, \mathbf{W}$ *from source to target, since* $Y \mid \mathbf{X}, R = 0$ *can be inherently different between the source and target populations. For this purpose,* $\mathbf{W}$ *is used directly in nuisance models for the density ratio and conditional mean, while the final target model remains defined on* $\mathbf{X}$*. Importantly, our framework also allows for the case without auxiliary variables: if no such* $\mathbf{W}$ *is available or needed, we set* $\mathbf{W} = \emptyset$ *and apply the same framework with* $\mathbf{Z} = \mathbf{X}$*.*

The preceding discussion addresses source–target covariate shift within a fixed subgroup. We next describe the second component of the problem: borrowing information from the larger subgroup to improve estimation for the smaller subgroup while avoiding negative transfer. Analogously to (2), we define the model coefficient in the majority subgroup of $\mathcal{T}$, $\bar{\boldsymbol{\beta}}^{[1]}$ as the solution to the equation:

$$\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{Y - g(\mathbf{X}^\intercal\boldsymbol{\beta})\} \mid R = 1] = \mathbf{0}.$$

Due to the distributional heterogeneity between the two subgroups, $\bar{\boldsymbol{\beta}}^{[1]}$ and $\bar{\boldsymbol{\beta}}^{[0]}$ tend to be different. The larger $n_{\mathcal{S},1}$ and $n_{\mathcal{T},1}$ can lead to a more precise estimate of $\bar{\boldsymbol{\beta}}^{[1]}$ compared

to $\bar{\boldsymbol{\beta}}^{[0]}$ learned with small minority samples. It is important to mitigate this disparity and a tentative strategy is to leverage the estimate of $\bar{\boldsymbol{\beta}}^{[1]}$ as external knowledge to guide the learning of $\bar{\boldsymbol{\beta}}^{[0]}$. This is motivated by the observation that genome-wide associations or models of a wide range of diseases and traits tend to have similar patterns between different ancestry groups, as revealed in recent studies (e.g., Lam et al., 2019; Verma et al., 2024). More generally, this motivates modeling the difference between subgroup-specific target coefficients as sparse: the two subgroups may share many active effects, while only a smaller set of effects differs meaningfully. For these purposes, we introduce the assumption

$$(\bar{\boldsymbol{\beta}}^{[0]}, \bar{\boldsymbol{\beta}}^{[1]}) \in \left\{ \bar{\boldsymbol{\beta}}^{[0]}, \bar{\boldsymbol{\beta}}^{[1]} : \|\bar{\boldsymbol{\beta}}^{[0]}\|_0 \leq s_\beta^{[0]},\ \|\bar{\boldsymbol{\beta}}^{[0]} - \bar{\boldsymbol{\beta}}^{[1]}\|_v \leq R_{\delta,v}, v \in [0,1] \right\}, \tag{4}$$

where $\|\boldsymbol{\beta}\|_0$ represents the number of non-zero entries in the vector $\boldsymbol{\beta}$ and $\|\boldsymbol{\beta}\|_v = \sum_j |\beta_j|^v$, when $v \in (0,1]$. Here, $\|\cdot\|_v$ corresponds to the exact ($v = 0$) or approximate ($v \in (0,1]$) sparsity norm. $s_\beta^{[0]}$ and $R_{\delta,v}$ represent the sparsity levels of the coefficients $\bar{\boldsymbol{\beta}}^{[0]}$ and the difference between the two subgroups $\bar{\boldsymbol{\beta}}^{[0]} - \bar{\boldsymbol{\beta}}^{[1]}$.

Since outcome models across sub-populations are presumed to be similar for most risk factors, we expect $R_{\delta,v} \ll s_\beta^{[0]}$ when taking $v = 0$ as a special case. This promises an efficiency gain by using $\bar{\boldsymbol{\beta}}^{[1]}$ to assist in learning $\bar{\boldsymbol{\beta}}^{[0]}$ in the underrepresented group, as their difference is sparser and easier to estimate compared to $\bar{\boldsymbol{\beta}}^{[0]}$ itself. On the other hand, given that $s_\beta^{[0]}$ and $R_{\delta,v}$ are unknown in practice, our objective is to maintain adaptivity and robustness to excessive model heterogeneity, i.e., the case where $R_{\delta,v}$ is large compared to $s_\beta^{[0]}$. In this scenario, the knowledge of $\bar{\boldsymbol{\beta}}^{[1]}$ can be non-informative or misleading to the target parameters $\bar{\boldsymbol{\beta}}^{[0]}$, and it is desirable to avoid the potential negative transfer caused by this. To facilitate understanding, the problem setup described above is illustrated in Figure 1.

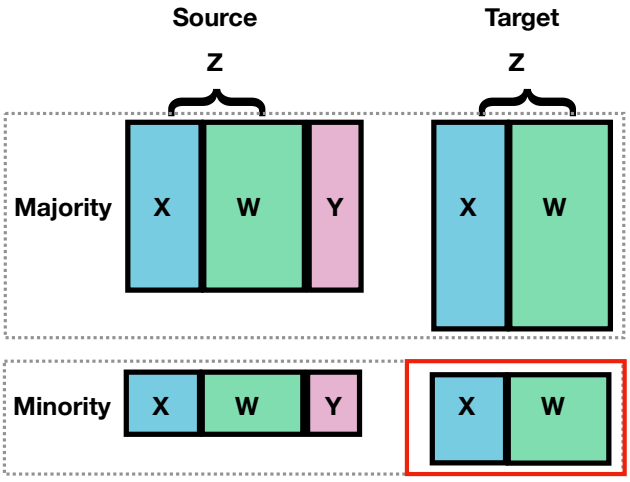

Figure 1: Structure of the observed data in our problem setup.

3. Method

We first give a brief overview of our method. AIMS consists of three main components. First, for each subgroup $r \in \{0,1\}$, we construct a doubly robust estimating equation to correct the source–target covariate shift using the density ratio and conditional mean nuisance models. In this step, we apply a bifold bias correction to reduce both the regularization bias in the target coefficient estimator and the first-order bias from regularized nuisance estimation. The resulting dense debiased vector is then thresholded to recover an ℓ_2-consistent sparse estimator. Second, for the underrepresented subgroup, we use the majority-subgroup estimator as auxiliary knowledge and estimate the sparse difference between the two subgroup-specific models. Third, a final surrogate-loss-based aggregation step protects against negative transfer when the majority model is insufficiently transferable.

3.1 Model-assisted domain adaptation for high-dimensional target models

Define the empirical mean operators on the source samples, target samples, and their union for group $r \in \{0,1\}$ as:

$$\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} f(\mathbf{D}) = n_{\mathcal{S},r}^{-1} \sum_{i=1}^{n} \mathbb{I}(S_i = 1, R_i = r) f(\mathbf{D}_i),$$

$$\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} f(\mathbf{D}) = n_{\mathcal{T},r}^{-1} \sum_{i=1}^{n} \mathbb{I}(S_i = 0, R_i = r) f(\mathbf{D}_i), \text{ and}$$

$$\widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} f(\mathbf{D}) = (n_{\mathcal{S},r} + n_{\mathcal{T},r})^{-1} \sum_{i=1}^{n} \mathbb{I}(R_i = r) f(\mathbf{D}_i).$$

We specify the nuisance models for density ratio and conditional mean imputation as: $h_r(\mathbf{z}) = \exp\big(\phi(\mathbf{z})^\intercal \boldsymbol{\alpha}^{[r]}\big)$ and $m_r(\mathbf{z}) = b\big(\phi(\mathbf{z})^\intercal \boldsymbol{\gamma}^{[r]}\big)$, where $\phi : \mathbb{R}^{p+q} \to \mathbb{R}^d$ is a flexible basis function of $\mathbf{Z}$. For example, one could simply take $\phi(\mathbf{z}) = \mathbf{z}$ or set $\phi(\mathbf{z})$ as a concatenation of non-linear bases of the components of $\mathbf{z}$. More generally, $\phi(\cdot)$ may include nonlinear transformations, interactions, splines, or fixed representations extracted from external models. The vectors $\boldsymbol{\alpha}^{[r]}$ and $\boldsymbol{\gamma}^{[r]}$ are the coefficients for the nuisance models that need to be estimated. The function $b(\cdot)$ is a prespecified, monotonically increasing link function, which may or may not be the same as the link function $g(\cdot)$. Denote by $\boldsymbol{\Phi} = \phi(\mathbf{Z})$.

The nuisance models above are working models used for covariate shift correction and outcome imputation. They are not intended to impose rigid linearity in the raw covariates, since the feature map $\phi(\cdot)$ can enlarge the nuisance feature space. At the same time, the sparse high-dimensional form keeps nuisance estimation feasible when the labeled source sample is limited, especially in the minority subgroup. As shown below, the doubly robust construction does not require both nuisance models to be correctly specified simultaneously: the target estimating equation remains valid when either the density ratio model or the conditional mean model is correctly specified within the working model class.

When $m_r(\mathbf{z}) = m_r^\star(\mathbf{z})$, the target parameter $\bar{\boldsymbol{\beta}}^{[r]}$ is the solution to the following imputation-based (IM) estimating equation: $\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{m_r(\mathbf{Z}) - g(\mathbf{X}^\intercal\boldsymbol{\beta})\} \mid R = r] = \mathbf{0}$. If $h_r(\mathbf{z}) = h_r^\star(\mathbf{z})$, $\bar{\boldsymbol{\beta}}^{[r]}$ solves the importance weighted (IW) estimating equation: $\mathbb{E}_{\mathcal{S}}[h_r(\mathbf{Z})\mathbf{X}\{Y - g(\mathbf{X}^\intercal\boldsymbol{\beta})\} \mid R = r] = \mathbf{0}$.

Empirically, $\bar{\boldsymbol{\beta}}^{[r]}$ can be estimated using sample versions of these equations. However, these simple strategies are prone to potential model misspecification. Specifically, the IW method leads to inconsistency if the density ratio model h_r is misspecified, and the IM method fails when m_r is misspecified. To overcome this challenge, we combine the two nuisance models to construct a doubly robust domain adaptation estimating equation for $\bar{\boldsymbol{\beta}}^{[r]}$ as

$$\mathbb{E}_{\mathcal{S}}[h_r(\mathbf{Z})\mathbf{X}\{m_r(\mathbf{Z}) - Y\} \mid R = r] - \mathbb{E}_{\mathcal{T}}[\mathbf{X}\{m_r(\mathbf{Z}) - g(\mathbf{X}^\intercal\boldsymbol{\beta})\} \mid R = r] = \mathbf{0}. \tag{5}$$

Proposition 2 shows that $\bar{\boldsymbol{\beta}}^{[r]}$ satisfies (5) when either the density ratio or imputation model is correctly specified.

Proposition 2 *When either* $h_r^\star(\mathbf{z}) = \exp\big(\phi(\mathbf{z})^\intercal\boldsymbol{\alpha}^{[r]}\big)$ *or* $m_r^\star(\mathbf{z}) = b\big(\phi(\mathbf{z})^\intercal\boldsymbol{\gamma}^{[r]}\big)$, *the target parameter* $\bar{\boldsymbol{\beta}}^{[r]}$ *is a solution to the population estimating equation* (5).

Motivated by this, we propose an empirical doubly robust loss function. Specifically, given the two nuisance parameters $\boldsymbol{\alpha}^{[r]}$ and $\boldsymbol{\gamma}^{[r]}$ whose estimation will be discussed later, we define the doubly robust covariate shift-corrected sparse regression for $\bar{\boldsymbol{\beta}}^{[r]}$ as:

$$\widehat{\boldsymbol{\beta}}^{[r]} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^q} \mathcal{L}_r(\boldsymbol{\beta}; \boldsymbol{\alpha}^{[r]}, \boldsymbol{\gamma}^{[r]}) + \lambda^{[r]}\|\boldsymbol{\beta}\|_1, \tag{6}$$

where the loss function is defined as

$$\mathcal{L}_r(\boldsymbol{\beta}; \boldsymbol{\alpha}^{[r]}, \boldsymbol{\gamma}^{[r]}) = \Big[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\mathbf{X}^\intercal\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}^{[r]})\{b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]}) - Y\} - \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\mathbf{X}^\intercal b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})\Big]\boldsymbol{\beta} + \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}G(\mathbf{X}^\intercal\boldsymbol{\beta}),$$

where $G(a) = \int_0^a g(u)\mathrm{d}u$, and $\lambda^{[r]}$ is a penalization parameter. We will present the optimal rate for all tuning parameters, including $\lambda^{[r]}$, in Section 4, with empirical tuning strategies in Supplementary S.2.3. A summary of the theoretical orders and practical tuning rules for all tuning parameters is also provided in Supplementary S.2.3. By construction, $\mathbb{E}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\boldsymbol{\beta}; \boldsymbol{\alpha}^{[r]}, \boldsymbol{\gamma}^{[r]})$ matches the left-hand side of (5). Thus, by Proposition 2, $\mathcal{L}_r(\boldsymbol{\beta}; \boldsymbol{\alpha}^{[r]}, \boldsymbol{\gamma}^{[r]})$ is doubly robust to the misspecification of the two nuisance models in the sense that $\bar{\boldsymbol{\beta}}^{[r]}$ satisfies its population first-order condition when either the density ratio model or the conditional mean imputation model is correctly specified.

Motivated by the above-introduced construction, a natural strategy for the estimation of $\bar{\boldsymbol{\beta}}^{[r]}$ is to first estimate $\boldsymbol{\alpha}^{[r]}$ and $\boldsymbol{\gamma}^{[r]}$ through regularized regression

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}^{[r]} =& \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]}\{\rho_{\mathcal{S},r}S\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}) - \rho_{\mathcal{T},r}(1-S)\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}\} + \lambda_\alpha^{[r]}\|\boldsymbol{\alpha}\|_1;\\ \widetilde{\boldsymbol{\gamma}}^{[r]} =& \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^d}\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{-Y\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma} + B(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma})\} + \lambda_\gamma^{[r]}\|\boldsymbol{\gamma}\|_1,\end{aligned} \tag{7}$$

where $\rho_{\mathcal{T},r} = (n_{\mathcal{S},r} + n_{\mathcal{T},r})/n_{\mathcal{T},r}$, $\rho_{\mathcal{S},r} = (n_{\mathcal{S},r} + n_{\mathcal{T},r})/n_{\mathcal{S},r}$, $B(a) = \int_0^a b(u)\mathrm{d}u$, and $\lambda_\alpha^{[r]}$, $\lambda_\gamma^{[r]}$ are two tuning parameters, then obtain the preliminary estimator $\widetilde{\boldsymbol{\beta}}^{[r]}$ as

$$\widetilde{\boldsymbol{\beta}}^{[r]} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^q}\mathcal{L}_r(\boldsymbol{\beta}, \widetilde{\boldsymbol{\alpha}}^{[r]}, \widetilde{\boldsymbol{\gamma}}^{[r]}) + \lambda^{[r]}\|\boldsymbol{\beta}\|_1\,. \tag{8}$$

According to our above discussion, $\widetilde{\boldsymbol{\beta}}^{[r]}$ tends to be consistent when either the density ratio or the conditional mean model is correctly specified. However, it is unlikely to achieve a desirable convergence rate due to the excessive biases in $\widetilde{\boldsymbol{\alpha}}^{[r]}$ and $\widetilde{\boldsymbol{\gamma}}^{[r]}$ introduced by regularization; see Remark 8 for more discussion. To address this issue, we propose a bifold bias correction approach. Denote by $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{[r]} = \mathbb{E}_{\mathcal{T}}[\dot{g}(\mathbf{X}^\intercal\boldsymbol{\beta})\mathbf{X}\mathbf{X}^\intercal \mid R = r]$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{[r]} = \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}[\dot{g}(\mathbf{X}^\intercal\boldsymbol{\beta})\mathbf{X}\mathbf{X}^\intercal]$ where $\dot{g}(\cdot)$ is the derivative of $g(\cdot)$. As the first fold of bias correction, we address the bias caused by the regularization term in (8) through a one-step debiased construction for each $\beta_j^{[r]} = \mathbf{e}_j^\intercal\boldsymbol{\beta}^{[r]}$ ($j \in \{1, \ldots, q\}$):

$$\widehat{\beta}_{\mathrm{Deb},j}^{[r]}(\boldsymbol{\alpha}^{[r]}, \boldsymbol{\gamma}^{[r]}) = \widetilde{\beta}_j^{[r]} - \widehat{\boldsymbol{\Omega}}_j^{[r]}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]}; \boldsymbol{\alpha}^{[r]}, \boldsymbol{\gamma}^{[r]}), \tag{9}$$

where $\mathbf{e}_j$ is the j-th unit vector in $\mathbb{R}^q$, $\widetilde{\beta}_j^{[r]}$ is the j-th element of $\widetilde{\boldsymbol{\beta}}^{[r]}$, $\widehat{\boldsymbol{\Omega}}_j^{[r]}$ is a regularized estimate of $\bar{\boldsymbol{\Omega}}_j^{[r]}$, the j-th row of $\bar{\boldsymbol{\Omega}}^{[r]} = \big[\bar{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\beta}}^{[r]}}^{[r]}\big]^{-1}$, and $\partial_{\boldsymbol{\beta}}$ is the partial derivative operator with respect to $\boldsymbol{\beta}$. To construct $\widehat{\boldsymbol{\Omega}}_j^{[r]}$, we use the nodewise lasso method proposed by Van de Geer et al. (2014). See Supplementary S.2.1 for the detailed expression of $\widehat{\boldsymbol{\Omega}}_j^{[r]}$.

This step only reduces the biases arising from the regularization on $\boldsymbol{\beta}$ in (8). In the second fold of bias correction, we further mitigate the biases due to the excessive regularization errors in $\widetilde{\boldsymbol{\alpha}}^{[r]}$ and $\widetilde{\boldsymbol{\gamma}}^{[r]}$. Our key idea is to further calibrate the two nuisance models and make them satisfy the moment conditions:

$$\bar{\boldsymbol{\Omega}}_j^{[r]}\mathbb{E}\big[\partial_{\boldsymbol{\alpha}}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]})\big] = \mathbf{0}; \quad \bar{\boldsymbol{\Omega}}_j^{[r]}\mathbb{E}\big[\partial_{\boldsymbol{\gamma}}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]})\big] = \mathbf{0}. \tag{10}$$

Under these two conditions, the first-order errors in $\boldsymbol{\alpha}^{[r]}$ and $\boldsymbol{\gamma}^{[r]}$ can be removed through concentration in the same spirit as Neyman orthogonality (Chernozhukov et al., 2018). This motivates us to obtain the calibrated nuisance estimators as $\widehat{\boldsymbol{\alpha}}_j^{[r]} = \widetilde{\boldsymbol{\alpha}}^{[r]} + \widehat{\boldsymbol{\xi}}_j^{[r]}$ and $\widehat{\boldsymbol{\gamma}}_j^{[r]} = \widetilde{\boldsymbol{\gamma}}^{[r]} + \widehat{\boldsymbol{\zeta}}_j^{[r]}$ with

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_j^{[r]} &= \operatorname*{argmin}_{\boldsymbol{\xi}\in\mathbb{R}^d} \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]}\widehat{w}_j^{[r]}\dot{b}(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[r]})\mathcal{F}^{[r]}(\boldsymbol{\xi}; \widetilde{\boldsymbol{\alpha}}^{[r]}) + \lambda_{\alpha_j}^{[r]}\|\boldsymbol{\xi}\|_1\,; \\
\widehat{\boldsymbol{\zeta}}_j^{[r]} &= \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathbb{R}^d} \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\widehat{w}_j^{[r]}\exp(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\alpha}}^{[r]})\mathcal{G}(\boldsymbol{\zeta}; \widetilde{\boldsymbol{\gamma}}^{[r]}) + \lambda_{\gamma_j}^{[r]}\|\boldsymbol{\zeta}\|_1\,,
\end{aligned} \tag{11}$$

where $\widehat{w}_j^{[r]} = \widehat{\boldsymbol{\Omega}}_j^{[r]}\mathbf{X}$ and $\mathcal{F}^{[r]}(\boldsymbol{\xi}; \boldsymbol{\alpha}) = \rho_{\mathcal{S},r}S\exp\{\boldsymbol{\Phi}^\intercal(\boldsymbol{\alpha}+\boldsymbol{\xi})\} - \rho_{\mathcal{T},r}(1-S)\boldsymbol{\Phi}^\intercal(\boldsymbol{\alpha}+\boldsymbol{\xi})$, $\mathcal{G}(\boldsymbol{\zeta}; \boldsymbol{\gamma}) = -Y\boldsymbol{\Phi}^\intercal(\boldsymbol{\gamma}+\boldsymbol{\zeta}) + B\{\boldsymbol{\Phi}^\intercal(\boldsymbol{\gamma}+\boldsymbol{\zeta})\}$. The Lasso problems in (11) are designed such that their Karush–Kuhn–Tucker (gradient) conditions empirically match our desirable moment conditions in (10). We then plug the calibrated $\widehat{\boldsymbol{\alpha}}_j^{[r]}$ and $\widehat{\boldsymbol{\gamma}}_j^{[r]}$ into (9) for a bifold bias-corrected estimator of each $\beta_j^{[r]}$, denoted as $\widehat{\beta}_{\mathrm{Deb},j}^{[r]} = \widehat{\beta}_{\mathrm{Deb},j}^{[r]}(\widehat{\boldsymbol{\alpha}}_j^{[r]}, \widehat{\boldsymbol{\gamma}}_j^{[r]})$. We also denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]} = \big(\widehat{\beta}_{\mathrm{Deb},1}^{[r]}, \ldots, \widehat{\beta}_{\mathrm{Deb},q}^{[r]}\big)^\intercal$. The above-introduced covariate shift correction approach is summarized in Algorithm 1.

Remark 3 *Note that the weight $\widehat{w}_j^{[r]}$ used in (11) may take negative values, which could make the loss function in (11) irregular and ill-posed. To handle this, one can divide the*

Algorithm 1 Covariate shift correction with bifold debiasing

Input: $\mathscr{D}^{[r]} = \{\mathbf{D}_i = (S_iY_i, \mathbf{X}_i^\intercal, \mathbf{W}_i^\intercal, S_i, R_i)^\intercal : R_i = r,\ i \in [n]\}$;

1: Obtain the preliminary estimators $\widetilde{\boldsymbol{\alpha}}^{[r]}$ and $\widetilde{\boldsymbol{\gamma}}^{[r]}$ by (7) and $\widetilde{\boldsymbol{\beta}}^{[r]}$ by (8). This step fits the initial density ratio model, imputation model, and target model coefficients.

2: For $j = 1, \ldots, q$: derive the form of the first fold bias correction for $\beta_j^{[r]}$ by (9). This step removes the leading regularization bias in the sparse target estimator.

3: For $j = 1, \ldots, q$: conduct the second fold calibration and obtain $\widehat{\boldsymbol{\gamma}}_j^{[r]}$ and $\widehat{\boldsymbol{\alpha}}_j^{[r]}$ by (11). This step calibrates the nuisance estimators to reduce first-order nuisance-estimation bias.

4: For $j = 1, \ldots, q$: plug $\widehat{\boldsymbol{\gamma}}_j^{[r]}$ and $\widehat{\boldsymbol{\alpha}}_j^{[r]}$ into (9) to obtain $\widehat{\beta}_{\mathrm{Deb},j}^{[r]} = \widehat{\beta}_{\mathrm{Deb},j}^{[r]}(\widehat{\boldsymbol{\alpha}}_j^{[r]}, \widehat{\boldsymbol{\gamma}}_j^{[r]})$.

Output: The debiased coefficient vector $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]} = (\widehat{\beta}_{\mathrm{Deb},1}^{[r]}, \ldots, \widehat{\beta}_{\mathrm{Deb},q}^{[r]})^\intercal$.

samples according to whether $\widehat{w}_{ij}^{[r]}$ is positive or negative and solve (11) *separately on the two subsets. Details of this stratification strategy are presented in Supplementary S.2.2.*

Although the bias-corrected $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]}$ is elementwise consistent for $\bar{\boldsymbol{\beta}}^{[r]}$ with a desirable convergence rate, its overall ℓ_2-error $\big\|\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\big\|_2$ does not converge to zero because $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]}$ is dense whereas $\bar{\boldsymbol{\beta}}^{[r]}$ is sparse, so its ℓ_2 error can accumulate with the dimension q. For the purpose of risk prediction, we further construct an ℓ_2-consistent estimator for $\bar{\boldsymbol{\beta}}^{[r]}$ through a thresholding and sparsifying approach:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]} = \big\{\mathrm{Thre}\big(\widehat{\beta}_{\mathrm{Deb},j}^{[r]}, \tau^{[r]}\big)\big\}_{j=1,\ldots,q}, \tag{12}$$

where $\mathrm{Thre}(z, c) = z\mathbb{I}(|z| \geq c)$, and $\tau^{[r]}$ is a tuning parameter. In Theorem 6, we prove the ℓ_2-consistency of $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]}$ and its convergence rate. As discussed above, we also comment in Remarks 7 and 8 that $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]}$ is substantially more robust to regularization errors in the nuisance estimators compared with the preliminary estimator $\widetilde{\boldsymbol{\beta}}^{[r]}$. For practical implementation, $\tau^{[r]}$ is chosen on the theoretical scale derived in Theorem 6; practical tuning details are provided in Supplementary S.2.3, and sensitivity analyses are reported in Supplementary S.3.3.

3.2 Adaptive knowledge transfer between the subgroups

The construction of $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[0]}$ enables us to borrow information from the source minority population, which may still suffer from high variability due to small sample size. Thus, we propose a knowledge-guided transfer learning procedure using the majority group estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[1]}$ to aid the estimation of $\bar{\boldsymbol{\beta}}^{[0]}$. Similar to recent work such as Li et al. (2022), our approach relies on the presumption that $\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{\beta}}^{[0]} - \bar{\boldsymbol{\beta}}^{[1]}$ is sparser and easier to estimate well than $\bar{\boldsymbol{\beta}}^{[0]}$, e.g., $R_{\delta,v} \ll s_\beta^{[0]}$ when $v = 0$ in (4). The idea is to take $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[1]}$ as a baseline (offset) and estimate the difference term $\bar{\boldsymbol{\beta}}^{[0]} - \widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[1]}$ instead of $\bar{\boldsymbol{\beta}}^{[0]}$ through thresholding and sparsifying. This strategy is beneficial when the majority and minority target models share most relevant predictors, but it may be harmful when their difference is not sufficiently sparse, i.e., when

it is denser than the target coefficient vector itself. The details are described in Algorithm 2.

Algorithm 2 Majority-knowledge-guided thresholding estimation

Input: The debiased coefficient vector $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}$ obtained by implementing Algorithm 1 on the minority group, and the sparsified estimator $\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Thr}}$ obtained on the majority group using (12).

Output: Knowledge transfer estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}} = \widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Thr}} + \widehat{\boldsymbol{\delta}}$ where $\widehat{\boldsymbol{\delta}} = (\widehat{\delta}_1, \ldots, \widehat{\delta}_q)^{\intercal}$, $\widehat{\delta}_j = \mathrm{Thre}\big(\widehat{\beta}^{[0]}_{\mathrm{Deb},j} - \widehat{\beta}^{[1]}_{\mathrm{Thr},j}, \tau_{\mathrm{KTr}}\big)$ and τ_{KTr} is a tuning parameter.

As shown in Section 4, when $\bar{\boldsymbol{\delta}}$ is sparser than $\bar{\boldsymbol{\beta}}^{[0]}$, Algorithm 2 effectively leverages knowledge from the majority group to improve the estimation of $\bar{\boldsymbol{\beta}}^{[0]}$. However, if $\bar{\boldsymbol{\delta}}$ is denser than $\bar{\boldsymbol{\beta}}^{[0]}$, i.e., $\bar{\boldsymbol{\beta}}^{[1]}$ differs significantly from $\bar{\boldsymbol{\beta}}^{[0]}$, $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}}$ given by Algorithm 2 could be less efficient than the minority-only estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr}}$, suffering from negative knowledge transfer. In this sense, negative transfer refers to the case where majority-guided borrowing worsens estimation relative to the minority-only estimator.

To prevent this, we propose a novel negative transfer protection approach that ensures the final estimator for $\bar{\boldsymbol{\beta}}^{[0]}$ is no worse than $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr}}$ up to an additive detection error, and is adaptive to the unknown transferability level between the two sub-populations (i.e., sparsity level of $\bar{\boldsymbol{\delta}}$). For similar purposes, recent work such as Tian and Feng (2023) uses held-out samples from the minority group to select between or ensemble the minority-only and knowledge transfer estimators. However, such a strategy is not directly applicable to our case due to the absence of labeled data on the target site $\mathcal{T}$. To handle this challenge, we introduce Algorithm 3 that relies on a surrogate loss $\widehat{\mathcal{Q}}(\boldsymbol{\beta}^{[0]}; \widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}) := \big\|\boldsymbol{\beta}^{[0]} - \widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}\big\|_2^2$ to evaluate the accuracy of $\boldsymbol{\beta}^{[0]}$. In this formulation, the dense vector $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}$ can essentially be viewed as a coordinatewise asymptotically normal and approximately unbiased reference vector and thus serves as an appropriate criterion for model selection. The data splitting in Algorithm 3 ensures that the dense debiased vector used for evaluation is constructed using samples independent from the candidate estimators being compared. We demonstrate this point more rigorously in Lemma 12.

In Algorithm 3, we split the minority data and obtain two independent bias-corrected dense vectors, $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},1}$ and $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},2}$. One of them, say $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},1}$, is used to derive the sparse minority-only estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},1}$ and the knowledge transfer estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},1}$. Then, the other one $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},2}$ is used to ensemble these two estimators through the surrogate loss function $\widehat{\mathcal{Q}}(\cdot; \widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},2})$. We use cross-fitting to leverage the two folds more effectively. For ensembling, we use the exponential weighting strategy (e.g., Dalalyan and Tsybakov, 2007) with some prespecified temperature parameter $a > 0$ in (13). When $a = \infty$, $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{AIMS},k}$ will be the one chosen from $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},k}$ and $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},k}$ that has a smaller loss $\widehat{\mathcal{Q}}$ evaluated against the debiased vector from the other fold of data. For finite $a > 0$, the exponential weights provide a smooth aggregation between the minority-only and knowledge transfer estimators. In

Algorithm 3 Protect against negative knowledge transfer

Input: $\mathscr{D}^{[r]} = \{\mathbf{D}_i = (S_iY_i, \mathbf{X}_i^\intercal, \mathbf{W}_i^\intercal, S_i, R_i)^\intercal : R_i = r,\ i \in [n]\}, r = 0, 1;$

1: Randomly split the minority data $\mathscr{D}^{[0]}$ (including both the source and target samples) into two disjoint sets $\mathscr{D}_1^{[0]}$ and $\mathscr{D}_2^{[0]}$ of equal size.

2: Implement Algorithm 1 on $\mathscr{D}_k^{[0]}$ to derive $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb},k}^{[0]}$ for $k = 1, 2$ and on $\mathscr{D}^{[1]}$ for $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[1]}$.

3: Implement thresholding on the dense vectors to obtain $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr},k}^{[0]}$ for $k = 1, 2$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[1]}$.

4: Implement Algorithm 2 with $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[1]}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb},k}^{[0]}$ to obtain $\widehat{\boldsymbol{\beta}}_{\mathrm{KTr},k}^{[0]}$ for $k = 1, 2$.

5: For $k = 1, 2$, aggregate the estimators as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{AIMS},k}^{[0]} = w_k \widehat{\boldsymbol{\beta}}_{\mathrm{Thr},k}^{[0]} + (1 - w_k)\widehat{\boldsymbol{\beta}}_{\mathrm{KTr},k}^{[0]},$$
$$\text{where } w_k = \frac{e^{-a\widehat{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}_{\mathrm{Thr},k}^{[0]}; \widehat{\boldsymbol{\beta}}_{\mathrm{Deb},3-k}^{[0]})}}{e^{-a\widehat{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}_{\mathrm{Thr},k}^{[0]}; \widehat{\boldsymbol{\beta}}_{\mathrm{Deb},3-k}^{[0]})} + e^{-a\widehat{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}_{\mathrm{KTr},k}^{[0]}; \widehat{\boldsymbol{\beta}}_{\mathrm{Deb},3-k}^{[0]})}}. \tag{13}$$

Output: The final AIMS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{AIMS}}^{[0]} = \big(\widehat{\boldsymbol{\beta}}_{\mathrm{AIMS},1}^{[0]} + \widehat{\boldsymbol{\beta}}_{\mathrm{AIMS},2}^{[0]}\big)/2.$

practice, a is fixed at a prespecified value; tuning details are provided in Supplementary S.2.3, and sensitivity analyses are reported in Supplementary S.3.3. Theorem 13 establishes the adaptivity guarantee for the hard-selection limit $a = \infty$; finite-a variants are smooth practical aggregators evaluated empirically.

4. Theoretical Justification

To introduce the key sparsity assumptions and the convergence results, we define the population-level model parameters as follows. Let $\bar{\boldsymbol{\alpha}}^{[r]} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^d}{\arg\min} \mathbb{E}_{\mathcal{S}}\{\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha})|R = r\} - \mathbb{E}_{\mathcal{T}}\{\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}|R = r\}$ and $\bar{\boldsymbol{\gamma}}^{[r]} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^d}{\arg\min} \mathbb{E}_{\mathcal{S}}\{-Y\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma} + B(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma})|R = r\}$ be the population-level preliminary nuisance model coefficients. For $j = 1, \ldots, q$, let $\bar{\boldsymbol{\alpha}}_j^{[r]} = \bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]}$ and $\bar{\boldsymbol{\gamma}}_j^{[r]} = \bar{\boldsymbol{\gamma}}^{[r]} + \bar{\boldsymbol{\zeta}}_j^{[r]}$ be the population-level calibrated coefficients corresponding to equation (11), where

$$\begin{aligned}
\bar{\boldsymbol{\xi}}_j^{[r]} = \underset{\boldsymbol{\xi}\in\mathbb{R}^d}{\operatorname{argmin}} \Bigg\{ & \mathbb{E}_{\mathcal{S}}\left[\bar{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}}^{[r]}) \exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \boldsymbol{\xi})\} \mid R = r\right] \\
& - \mathbb{E}_{\mathcal{T}}\left[\bar{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}}^{[r]})\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \boldsymbol{\xi})\} \mid R = r\right] \Bigg\}, \qquad (14) \\
\bar{\boldsymbol{\zeta}}_j^{[r]} = \underset{\boldsymbol{\zeta}\in\mathbb{R}^d}{\operatorname{argmin}} & \, \mathbb{E}_{\mathcal{S}}\left\{\bar{w}_j^{[r]} \exp(\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\alpha}}^{[r]}) \mathcal{G}(\boldsymbol{\zeta}; \bar{\boldsymbol{\gamma}}^{[r]}) \mid R = r\right\}.
\end{aligned}$$

and $\bar{w}_j^{[r]} = \bar{\boldsymbol{\Omega}}_j^{[r]}\mathbf{X}$. If the density ratio model is correct, we can show that $\bar{\boldsymbol{\xi}}_j^{[r]} = \mathbf{0}$ and $\bar{\boldsymbol{\alpha}}^{[r]} = \bar{\boldsymbol{\alpha}}_j^{[r]}$ for all j. Similarly, if the imputation model is correct, we can show that $\bar{\boldsymbol{\zeta}}_j^{[r]} = \mathbf{0}$ and $\bar{\boldsymbol{\gamma}}^{[r]} = \bar{\boldsymbol{\gamma}}_j^{[r]}$ for all j. Define $S_\beta^{[r]} = \mathrm{supp}(\bar{\boldsymbol{\beta}}^{[r]})$ and $s_\beta^{[r]} = |S_\beta^{[r]}| = \big\|\bar{\boldsymbol{\beta}}^{[r]}\big\|_0$, and let

$s_{\text{nui}}^{[r]} = \max\left\{\left\|\bar{\boldsymbol{\alpha}}^{[r]}\right\|_0, \left\|\bar{\boldsymbol{\gamma}}^{[r]}\right\|_0, \max_{j\in\{1,\ldots,q\}}\left\|\bar{\boldsymbol{\Omega}}_j^{[r]}\right\|_0\right\}$ to represent the (exact) sparsity levels of the target and nuisance model coefficients, respectively. Since the nuisance parameters $\bar{\boldsymbol{\xi}}_j^{[r]}$ and $\bar{\boldsymbol{\zeta}}_j^{[r]}$ are defined through the reweighted regression in (14), we introduce $R_{\text{nui},v}^{[r]} = \max_{1\le j\le q}\max\left\{\|\bar{\boldsymbol{\xi}}_j^{[r]}\|_v, \|\bar{\boldsymbol{\zeta}}_j^{[r]}\|_v\right\}$ for an arbitrary $v \in [0,1]$ to accommodate both the exact (i.e., $v = 0$) and approximate sparsity (i.e., $v \in (0,1]$) regimes. In addition, we define $R_{\delta,v} = \|\bar{\boldsymbol{\beta}}^{[0]} - \bar{\boldsymbol{\beta}}^{[1]}\|_v = \|\bar{\boldsymbol{\delta}}\|_v$ for any $v \in [0,1]$ to characterize the model heterogeneity between the majority and minority sub-populations on the target. Again, our analyses cover both the exact and approximate sparse $\bar{\boldsymbol{\delta}}$ scenarios.

We use $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, $a_n = O(b_n)$ if $\limsup_{n\to\infty}|a_n/b_n| \le C$ for some constant C, and $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We denote convergence in probability by $\xrightarrow{\mathbb{P}}$ and convergence in distribution by $\xrightarrow{\mathcal{D}}$. For a sequence of random variables Z_n, we use $Z_n = o_p(a_n)$ if $|Z_n|/a_n \xrightarrow{\mathbb{P}} 0$ and $Z_n = O_p(a_n)$ if $\lim_{C\to\infty}\limsup_{n\to\infty}\mathbb{P}(|Z_n/a_n| > C) = 0$. Let $n_r = n_{\mathcal{S},r} \wedge n_{\mathcal{T},r} = \min\{n_{\mathcal{S},r}, n_{\mathcal{T},r}\}$, and assume $n_{\mathcal{S},0} = o(n_{\mathcal{S},1})$ and $n_{\mathcal{T},0} = o(n_{\mathcal{T},1})$, i.e., the minority sample sizes are much smaller than the corresponding majority sample sizes.

We begin by justifying the robustness and effectiveness of the model-assisted covariate shift correction step in Algorithm 1 and its subsequent thresholding estimator $\widehat{\boldsymbol{\beta}}_{\text{Thr}}^{[r]}$. First, we introduce Assumption 1, which states that at least one nuisance model is correctly specified. This assumption is referred to as the model-doubly-robust assumption in the semiparametric literature (Smucler et al., 2019).

Assumption 1 *For $r \in \{0,1\}$, either $h_r^\star(\mathbf{z}) := p_{\mathbf{Z}|S=0,R=r}(\mathbf{z})/p_{\mathbf{Z}|S=1,R=r}(\mathbf{z}) = \exp\{\phi(\mathbf{z})^\intercal\bar{\boldsymbol{\alpha}}^{[r]}\}$ or $m_r^\star(\mathbf{z}) := \mathbb{E}[Y \mid \mathbf{Z} = \mathbf{z}, R = r] = b\{\phi(\mathbf{z})^\intercal\bar{\boldsymbol{\gamma}}^{[r]}\}$ holds.*

All technical assumptions are presented in Supplementary S.1.1 and are standard in the literature on high-dimensional inference (e.g., Van de Geer et al., 2014) and doubly robust inference (e.g., Tan, 2020). In summary, Assumption S.1 rules out heavy tails and singularity in the predictors $\mathbf{X}$ and nuisance covariates $\boldsymbol{\Phi}$. In this assumption, we introduce $K > 0$ such that $\|\mathbf{X}_i\|_\infty \le K$ for all $i \in \{1,\ldots,n\}$ with probability approaching 1, where $\|(x_1,\ldots,x_q)\|_\infty := \max_{j\in\{1,\ldots,q\}}|x_j|$. For bounded designs, we have $K = O(1)$ and for Gaussian or sub-Gaussian design, $K = O[\{\log(q)\}^{1/2}]$. Assumption S.2 imposes sparsity constraints on $s_\beta^{[r]}$, $s_{\text{nui}}^{[r]}$, $R_{\text{nui},v}^{[r]}$ and $R_{\delta,v}$, ensuring they are not excessively large relative to sample sizes. These constraints are crucial for the consistency of the nuisance and target model estimators, as in the sparse GLM literature (Wainwright, 2019). Assumption S.3 imposes regularity conditions on link functions $g(\cdot)$ and $b(\cdot)$, which are satisfied in broad GLM settings. Assumption S.4 specifies the optimal tuning parameter rates.

Under Assumption 1 and Assumptions S.1–S.4, the convergence results for $\widehat{\boldsymbol{\Omega}}_j^{[r]}$, $\widetilde{\boldsymbol{\beta}}^{[r]}$, and $\{\widehat{\boldsymbol{\alpha}}_j^{[r]}, \widehat{\boldsymbol{\gamma}}_j^{[r]}\}$ are established in Lemmas S.1, S.2, and S.3; see Supplementary Section S.1.1. We then use these results to establish the properties of each bias-corrected estimator $\widehat{\beta}_{\text{Deb},j}^{[r]}$ in Lemma 4.

Lemma 4 *Under Assumption 1 and Assumptions S.1–S.4, for $r \in \{0,1\}$, we have*

$$\widehat{\beta}_{\text{Deb},j}^{[r]} - \bar{\beta}_j^{[r]} = -\bar{\boldsymbol{\Omega}}_j^{[r]}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]}) + r_j^{[r]},$$

where $-n_r^{1/2}\sigma_{r,j,n}^{-1}\bar{\boldsymbol{\Omega}}_j^{[r]}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]}) \xrightarrow{\mathcal{D}} N(0,1)$, $\sigma_{r,j,n}^2 = n_r\mathrm{Var}\{\bar{\boldsymbol{\Omega}}_j^{[r]}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]})\} \asymp 1$, $r_j^{[r]} = O_p(\Delta_r)$, $\boldsymbol{r}^{[r]} = (r_1^{[r]},\ldots,r_q^{[r]})^\intercal$ *denotes the vector of coordinatewise remainders, and*

$$\Delta_r = \frac{K^5(s_{\mathrm{nui}}^{[r]} + s_\beta^{[r]})\log(q)}{n_r} + KR_{\mathrm{nui},v}^{[r]}\left\{\frac{\log(q)}{n_r}\right\}^{1-v/2}. \tag{15}$$

In Lemma 4, we derive the asymptotic expansion $\widehat{\beta}_{\mathrm{Deb},j}^{[r]} = \bar{\beta}_j^{[r]} + \epsilon_j + \mathrm{bias}_j$, where ϵ_j is a mean-zero stochastic term with variance of order $1/n_r$, and bias_j stands for a bias term free of any first-order errors from the nuisance estimators $\widehat{\boldsymbol{\alpha}}_j^{[r]}$ and $\widehat{\boldsymbol{\gamma}}_j^{[r]}$. Thus, $\widehat{\beta}_{\mathrm{Deb},j}^{[r]}$ is insensitive to the nuisance errors. This point will be further discussed in Remark 7.

Remark 5 *Assumption S.2 alone is not sufficient to guarantee the* $n_r^{-1/2}$ *convergence rate and asymptotic normality of* $\widehat{\beta}_{\mathrm{Deb},j}^{[r]}$ *in Lemma 4. This rate is also not required for the subsequent convergence theorems. Nonetheless, under moderately stronger sparsity conditions* $s_{\mathrm{nui}}^{[r]} + s_\beta^{[r]} = o\big[n_r^{1/2}/\{K^5\log(q)\}\big]$ *and* $R_{\mathrm{nui,v}}^{[r]} = o\big(n_r^{(1-v)/2}/[K\{\log(q)\}^{1-v/2}]\big)$, $n_r^{1/2}(\widehat{\beta}_{\mathrm{Deb},j}^{[r]} - \bar{\beta}_j^{[r]})$ *will converge to a zero-mean normal distribution, enabling the interval estimation for* $\bar{\beta}_j^{[r]}$.

Although Lemma 4 establishes elementwise convergence for $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]} = (\widehat{\beta}_{\mathrm{Deb},1}^{[r]},\ldots,\widehat{\beta}_{\mathrm{Deb},q}^{[r]})^\intercal$, it can still fail to achieve ℓ_2-consistency, i.e., $\big\|\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\big\|_2 = o_p(1)$. As pointed out at the end of Section 3.1, this is because $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]}$ is dense, so its estimation errors can accumulate as q increases. This motivates us to employ the thresholding procedure to obtain $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]}$. The ℓ_1- and ℓ_2-convergence rates of $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]}$ are given in Theorem 6.

Theorem 6 *Under Assumption 1 and Assumptions S.1–S.4 with* $\tau^{[r]} = C\Big\{\sqrt{\log(q)}\big(n_r^{-1/2} + \Delta_r\big)\Big\}$, *where* $C > 0$ *is a sufficiently large constant, we have* $\|\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 = O_p\big(s_\beta^{[r]}\tau^{[r]}\big)$ *and* $\|\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2 = O_p\big(\sqrt{s_\beta^{[r]}}\tau^{[r]}\big)$.

Remark 7 *The error term* Δ_r *defined in (15) encodes the impact of the nuisance estimators on the convergence rate of* $\widehat{\beta}_{\mathrm{Deb},j}^{[r]}$ *and* $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]}$. *If* $\Delta_r = O(n_r^{-1/2})$, *then the rate in Theorem 6 matches the oracle rate for sparse estimation of* $\bar{\boldsymbol{\beta}}^{[r]}$. *Particularly, when* $v = 0$ *and* $K = O(1)$,

$$\Delta_r = O\left\{\frac{(s_{\mathrm{nui}}^{[r]} + s_\beta^{[r]} + R_{\mathrm{nui},0}^{[r]})\log(q)}{n_r}\right\},$$

which has an n_r^{-1} *factor rather than an* $n_r^{-1/2}$ *factor. Hence,* $\Delta_r = o(n_r^{-1/2})$ *provided that the parameters are sufficiently sparse, i.e.,* $(s_{\mathrm{nui}}^{[r]} + s_\beta^{[r]} + R_{\mathrm{nui},0}^{[r]})\log(q) = o(n_r^{1/2})$.

Remark 8 *Recall that our preliminary estimator* $\widetilde{\boldsymbol{\beta}}^{[r]}$ *is also based on a doubly robust formulation but does not include the calibration steps 2 – 4 in Algorithm 1. Unlike* $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[r]}$, *it*

does not remove the first-order nuisance-estimation error $\{s_{\mathrm{nui}}^{[r]}\log(q)/n_r\}^{1/2}$*; see Lemma S.2. This implies that our proposed calibration steps can effectively reduce the sensitivity to errors in the nuisance estimators, which is in a similar spirit to the rate-double-robustness property studied in Kennedy (2023). Nevertheless, Kennedy (2023) does not establish any model-double-robustness analogous to our Assumption 1.*

Zhou et al. (2025) also established model-double-robustness while correcting the bias from the two nuisance functions. However, their analysis is restricted to a low-dimensional outcome model (i.e., q is fixed) and does not extend to the error rates in Lemma 4 and Theorem 6 when q diverges. For fixed q, the rate in Lemma 4 is comparable to that in Theorem 1 of Zhou et al. (2025); in this low-dimensional setting, the lasso penalty used to estimate $\boldsymbol{\beta}^{[r]}$ *and the subsequent lasso-based debiasing step are no longer necessary.*

In the following results, we establish the convergence rates of the knowledge transfer estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{KTr}}^{[0]}$ and the final estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{AIMS}}^{[0]}$, showing their potential improvement over the minority-only estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[0]}$.

Lemma 9 *Under Assumption 1 and Assumptions S.1–S.4, suppose that* $\tau^{[r]}$ *is of the order given in Theorem 6 for* $r \in \{0,1\}$ *and* $\tau_{\mathrm{KTr}} = C\{\sqrt{\log(q)}(n_0^{-1/2}+\Delta)\}$*, where* $\Delta = \max_{r\in\{0,1\}}\Delta_r$*, and* $C > 0$ *is a sufficiently large constant. Then we have* $\|\widehat{\bar{\boldsymbol{\delta}}} - \bar{\boldsymbol{\delta}}\|_1 = O_p\Big[\{\log(q)/n_0\}^{-v/2}R_{\delta,v}\tau_{\mathrm{KTr}}\Big]$ *and* $\|\widehat{\bar{\boldsymbol{\delta}}} - \bar{\boldsymbol{\delta}}\|_2 = O_p\Big[\{\log(q)/n_0\}^{-v/4}R_{\delta,v}^{1/2}\tau_{\mathrm{KTr}}\Big]$.

Lemma 9 is an important intermediate result for the estimation error of $\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{\beta}}^{[0]} - \bar{\boldsymbol{\beta}}^{[1]}$. Based on this lemma, we derive the ℓ_2-convergence rate of the knowledge transfer estimator in Algorithm 2 in Theorem 10.

Theorem 10 *Under Assumption 1 and Assumptions S.1–S.4, with* $\tau^{[r]}$ *for* $r \in \{0,1\}$ *and* τ_{KTr} *of the orders given in Lemma 9, we have*

$$\big\|\widehat{\boldsymbol{\beta}}_{\mathrm{KTr}}^{[0]} - \bar{\boldsymbol{\beta}}^{[0]}\big\|_2 = O_p\left[\big\{s_\beta^{[1]}\log(q)\big\}^{1/2}\big(n_1^{-1/2}+\Delta_1\big) + \frac{\{\log(q)\}^{1/2-v/4}}{n_0^{-v/4}}R_{\delta,v}^{1/2}\big(n_0^{-1/2}+\Delta\big)\right]. \tag{16}$$

The first error term in (16) arises from the majority estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[1]}$ and the second term corresponds to the error of $\widehat{\bar{\boldsymbol{\delta}}}$ given in Lemma 9. We discuss the potential efficiency gain from the knowledge transfer step in Algorithm 2 in Remark 11.

Remark 11 *For simplicity, we consider a scenario where* $\bar{\boldsymbol{\delta}}$ *is exactly sparse (*$v = 0$*),* $s_\beta^{[1]} \asymp s_\beta^{[0]}$*,* $s_{\mathrm{nui}}^{[1]} \asymp s_{\mathrm{nui}}^{[0]}$*, and* $R_{\mathrm{nui},v}^{[1]} \asymp R_{\mathrm{nui},v}^{[0]}$*, which implies that* $\Delta_1 = o(\Delta_0)$*. When* $R_{\delta,0} = o(s_\beta^{[0]})$*, i.e., the difference* $\bar{\boldsymbol{\delta}}$ *is sparser than the target model coefficients, Theorem 6 and Lemma 9 imply that the ratio of the* ℓ_2*-convergence rate of* $\widehat{\bar{\boldsymbol{\delta}}}$ *to that of* $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[0]}$ *is* $(R_{\delta,0}/s_\beta^{[0]})^{1/2} = o(1)$*, indicating efficiency improvement of the knowledge transfer estimator* $\widehat{\boldsymbol{\beta}}_{\mathrm{KTr}}^{[0]}$ *over the minority-data-only estimator* $\widehat{\boldsymbol{\beta}}_{\mathrm{Thr}}^{[0]}$.

Our knowledge transfer between the majority and minority subgroups does not rely on target labels. Despite this, under the additional assumption $\Delta_r = O(n_r^{-1/2})$ *for* $r \in \{0,1\}$*,*

the error rate of $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}}$ *is comparable to the standard transfer learning rate obtained when* n_0 *labeled target observations are available. This comparison applies directly to supervised transfer learning methods (Li et al., 2022; Tian and Feng, 2023; He et al., 2024), and is also aligned with the intermediate result in Theorem 1 of Cai et al. (2024) after accounting for the additional density ratio estimation error in their analysis.*

Nevertheless, in a slightly simplified setting where $\Delta_r = O(n_r^{-1/2})$ for $r \in \{0,1\}$, Theorem 10 shows that the rate upper bound for $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}}$ can be larger than that for $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr}}$ whenever $s^{[0]}_\beta / s^{[1]}_\beta = o(n_0/n_1)$ or $s^{[0]}_\beta = o\big(R_{\delta,v}\{\log(q)/n_0\}^{-v/2}\big)$. This phenomenon is referred to as negative knowledge transfer. To avoid this issue, we propose the model selection Algorithm 3 to obtain the final estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{AIMS}}$ and justify its effectiveness below. We begin with Lemma 12, which shows that the $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}$-based loss function $\widehat{\mathcal{Q}}\big(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}\big)$ offers a good approximation of the ideal model selection criterion $\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$.

Lemma 12 *Under Assumption 1 and Assumptions S.1–S.4, we have* $\widehat{\mathcal{Q}}\big(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb}}\big) = \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2 + 2\big(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{[0]}\big)^{\intercal}\big(\mathcal{E} + \mathrm{Rem}\big) + C_2$*, where*

$$\mathcal{E} = \big\{\bar{\boldsymbol{\Omega}}^{[0]}_1 \partial_{\boldsymbol{\beta}} \mathcal{L}_0(\bar{\boldsymbol{\beta}}^{[0]}; \bar{\boldsymbol{\alpha}}^{[0]}_1, \bar{\gamma}^{[0]}_1), \ldots, \bar{\boldsymbol{\Omega}}^{[0]}_q \partial_{\boldsymbol{\beta}} \mathcal{L}_0(\bar{\boldsymbol{\beta}}^{[0]}; \bar{\boldsymbol{\alpha}}^{[0]}_q, \bar{\gamma}^{[0]}_q)\big\}^{\intercal}$$

is a mean-zero empirical process satisfying $\|\mathcal{E}\|_\infty = O_p[\{\log(q)/n_0\}^{1/2}]$, Rem *denotes the remainder approximation-error term satisfying*

$$\|\mathrm{Rem}\|_\infty = O_p[\{\log(q)\}^{1/2}\Delta_0],$$

and C_2 *is independent of* $\boldsymbol{\beta}$*.*

As discussed in Remark 7, the term Rem, whose ℓ_∞ norm is controlled by $\{\log(q)\}^{1/2}\Delta_0$, captures the second-order influence of nuisance estimation errors. Meanwhile, the term $2(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{[0]})^{\top}\mathcal{E}$ can be controlled by the concentration of $\mathcal{E}$. Combining this lemma with Theorems 6 and 10, we derive the convergence rate of the final estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{AIMS}}$ in Theorem 13.

Theorem 13 *For the hard-selection version of Algorithm 3 with* $a = \infty$*, suppose that Assumption 1 and Assumptions S.1–S.4 hold, with* $\tau^{[r]}$ *for* $r \in \{0,1\}$ *and* τ_{KTr} *of the orders given in Lemma 9. Suppose additionally that the remainder vector from Lemma 4, computed on each fold, satisfies* $\|\boldsymbol{r}^{[0]}_{(k)}\|_\infty = O_p(\Delta_0)$ *for* $k \in \{1,2\}$*. Then*

$$\big\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{AIMS}} - \bar{\boldsymbol{\beta}}^{[0]}\big\|_2 = O_p\Big\{\Big(\big\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr}} - \bar{\boldsymbol{\beta}}^{[0]}\big\|_2 \wedge \big\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}} - \bar{\boldsymbol{\beta}}^{[0]}\big\|_2\Big) + \mathrm{DetectErr}\Big\},$$

with the detection error term $\mathrm{DetectErr} = O_p\Big[\{s_d^{1/2}(n_0^{-1/2} + \Delta_0)(\big\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}} - \bar{\boldsymbol{\beta}}^{[0]}\big\|_2 + \big\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr}} - \bar{\boldsymbol{\beta}}^{[0]}\big\|_2)\}^{1/2}\Big]$*, where* $s_d = O_p\{s^{[0]}_\beta + R_{\delta,v}(\tau^{[1]} \wedge \tau_{\mathrm{KTr}})^{-v}\}$*. When* $\Delta_r = o(n_r^{-1/2})$ *for* $r \in \{0,1\}$*, we have*

$$\mathrm{DetectErr} = O_p\left(\Big[s_d^{1/2} n_0^{-1/2} \max_{r\in\{0,1\}} \Big\{\frac{s^{[r]}_\beta \log(q)}{n_r}\Big\}^{1/2} + s_d^{1/2} n_0^{-1/2}\Big\{\frac{\log(q)}{n_0}\Big\}^{1/2-v/4} R_{\delta,v}^{1/2}\Big]^{1/2}\right).$$

If, in addition, $v = 0$ *and* $R_{\delta,0} = o\big[\{n_0/\log(q)\}^{1/2}\big]$, $\mathrm{DetectErr} = o_p(n_0^{-1/4})$.

Theorem 13 shows that the ℓ_2 error of $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{AIMS}}$ is controlled, up to an additive detection error incurred by model selection in Algorithm 3, by the better of the two candidate error rates: the error rate of the knowledge transfer estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr}}$ and the error rate of the minority-only estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr}}$. The detection error is governed by the effective support size s_d of the difference between the two candidate estimators, the minority estimation scale $n_0^{-1/2} + \Delta_0$, and the distance between the two candidates. When $v = 0$, $\Delta_r = o(n_r^{-1/2})$, and $R_{\delta,0} = o[\{n_0/\log(q)\}^{1/2}]$, this additional cost is $o_p(n_0^{-1/4})$.

Similar target-data-splitting-based procedures are employed in Tian and Feng (2023), Cai et al. (2024), and related work to protect against negative transfer. These procedures use target labels, whereas our detection step is implemented without observed labels in the target sample. In particular, the procedure of Cai et al. (2024) also incurs an additional selection cost of order $n_0^{-1/4}$ on the ℓ_2 scale for adaptive selection between the knowledge transfer estimator and the target-only estimator.

5. Simulation Studies

We conduct simulation studies under three related data-generating mechanisms. Across the three settings, the sparse linear components are fixed, and the majority-group coefficients are generated as local perturbations of the minority-group coefficients. This creates a regime in which majority-to-minority transfer is plausible, but not guaranteed to outperform the minority-only estimator. For $r \in \{0, 1\}$, we consider the following three settings:

$$
\begin{aligned}
\text{Setting I:}\quad & \mathbb{P}(Y = 1 \mid \mathbf{Z}, R = r) = \mathrm{expit}\{\mathbf{Z}^\intercal \boldsymbol{\gamma}^{[r]}\},\\
& \mathbb{P}(S = 1 \mid \mathbf{Z}, R = r) = \mathrm{expit}\{\mathbf{Z}^\intercal \boldsymbol{\alpha}^{[r]}\};\\
\text{Setting II:}\quad & \mathbb{P}(Y = 1 \mid \mathbf{Z}, R = r) = \mathrm{expit}\{\mathbf{Z}^\intercal \boldsymbol{\gamma}^{[r]} + u_Y(\mathbf{Z})\},\\
& \mathbb{P}(S = 1 \mid \mathbf{Z}, R = r) = \mathrm{expit}\{\mathbf{Z}^\intercal \boldsymbol{\alpha}^{[r]}\};\\
\text{Setting III:}\quad & \mathbb{P}(Y = 1 \mid \mathbf{Z}, R = r) = \mathrm{expit}\{\mathbf{Z}^\intercal \boldsymbol{\gamma}^{[r]}\},\\
& \mathbb{P}(S = 1 \mid \mathbf{Z}, R = r) = \mathrm{expit}\{\mathbf{Z}^\intercal \boldsymbol{\alpha}^{[r]} + u_S(\mathbf{Z})\}.
\end{aligned}
$$

Setting I is the correctly specified reference case. Setting II tests robustness when the outcome working model is misspecified but the density ratio model remains correctly specified, whereas Setting III reverses this pattern by misspecifying the density ratio model while keeping the imputation model correctly specified. Exact coefficient vectors, perturbation functions, and full data-generating formulas are provided in Supplementary S.3.2.

We set $q = 100$, $n_{\mathcal{S},1} = 2000$, and $n_{\mathcal{T},1} = 3000$. The primary target-unlabeled minority sample size is $n_{\mathcal{T},0} = 2000$; the additional cases $n_{\mathcal{T},0} \in \{1000, 3000\}$ are reported in Supplementary Table S.4. The main results are organized around two one-dimensional grids. Along the sample-size axis, $p = 400$ and $n_{\mathcal{S},0} \in \{300, 400, 500, 600\}$. Along the dimensionality axis, $n_{\mathcal{S},0} = 400$ and $p \in \{50, 100, 200, 400, 800\}$. Unless otherwise stated, the AIMS implementation follows Algorithm 3 with $C_\tau = 2$ in $\tau^{[r]} = C_\tau\{\log(q)/n_{\mathcal{S},r}\}^{1/2}$, bootstrap quantile $q_\tau = 0.8$ with $B = 500$ Gaussian multiplier bootstrap draws for the calibrated nui-

sance and transfer thresholds, and aggregation temperature $a = 5$. Performance is measured by the squared ℓ_2 error $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$.

These choices define a moderately imbalanced high-dimensional reference design. The majority strata are larger than the minority source stratum, reflecting the sampling regime for which transfer from the majority group can be useful but must be protected against bias. The two main grids vary the minority labeled-source sample size $n_{\mathcal{S},0}$ and the auxiliary-covariate dimension p separately, while keeping the target-model dimension q, the minority target-unlabeled sample size $n_{\mathcal{T},0}$, and the majority sample sizes fixed. The algorithmic constants C_τ, q_τ, and a are used as default tuning values for the main comparison; Supplementary Tables S.7 and S.8 report sensitivity analyses for q_τ and C_τ, Supplementary Table S.5 reports ablations of the minority-side, majority-guided, and aggregation-temperature variants of AIMS, and Supplementary Table S.6 reports the majority-side estimator performance.

We compare AIMS with three groups of benchmarks. The first group consists of single-nuisance covariate shift estimators derived from the estimating equations in Section 3.1. Importance weighting strategy (IW; Supplementary S.3.1, Algorithm S.1) estimates the minority source-to-target density ratio and then fits the working target model using an ℓ_1-penalized weighted logistic loss on labeled minority-source observations, following the classical covariate shift correction idea (Huang et al., 2007). Imputation strategy (IM, also referred to as pseudo-labeling; Supplementary S.3.1, Algorithm S.2) estimates the minority conditional mean, evaluates it on minority-target covariates as pseudo-outcomes, and then fits the same working target model on the target covariate distribution. IW and IM use only one of the two nuisance models in the doubly robust estimating equation (Chakrabortty et al., 2019; Liu et al., 2023; Tian et al., 2024). $\text{IW}_{\text{aLasso}}$ and $\text{IM}_{\text{aLasso}}$ replace the ordinary ℓ_1 penalty in the final target-model fit by an adaptive weighted ℓ_1 penalty, while IM_{RF} and IM_{XGB} replace the parametric conditional mean fit in IM by random forests and XGBoost.

The second group contains transfer and distribution-alignment baselines. Since TransGLM (Tian and Feng, 2023) requires labeled observations from the recipient population, it is not directly applicable to our unlabeled-target setting. For comparison, we use the source-labeled adaptation in Supplementary S.3.1, Algorithm S.3: the labeled minority-source sample is treated as the recipient sample, and the labeled majority-source sample is treated as the auxiliary source sample. $\text{TransGLM}_{\text{iw}}$ further weights these two labeled samples by the estimated source-to-target density ratios to account for covariate shift. We adapt TransFusion (He et al., 2024) in the same transfer format, using labeled minority-source data as the pseudo-target task and labeled majority-source data as the auxiliary task. CORAL (Sun et al., 2017) is used as a minority-only distribution-alignment baseline: it aligns the second-order covariance structure of the labeled minority-source covariates to that of the unlabeled minority-target covariates before fitting the sparse logistic target model. The third group contains diagnostic AIMS variants: $\text{AIMS}_{\text{min-o}}$ is the minority-only thresholded estimator from the covariate shift correction step in Algorithm 1, $\text{AIMS}_{\text{maj-g}}$ is the majority-guided transfer estimator from Algorithm 2, and the aggregation-temperature variants correspond to the final negative-transfer-protected aggregation in Algorithm 3. For readability, the main figures focus on AIMS, $\text{IM}_{\text{aLasso}}$, $\text{TransGLM}_{\text{iw}}$, and TransFusion; complete numerical comparisons are reported in Supplementary Tables S.2 and S.3.

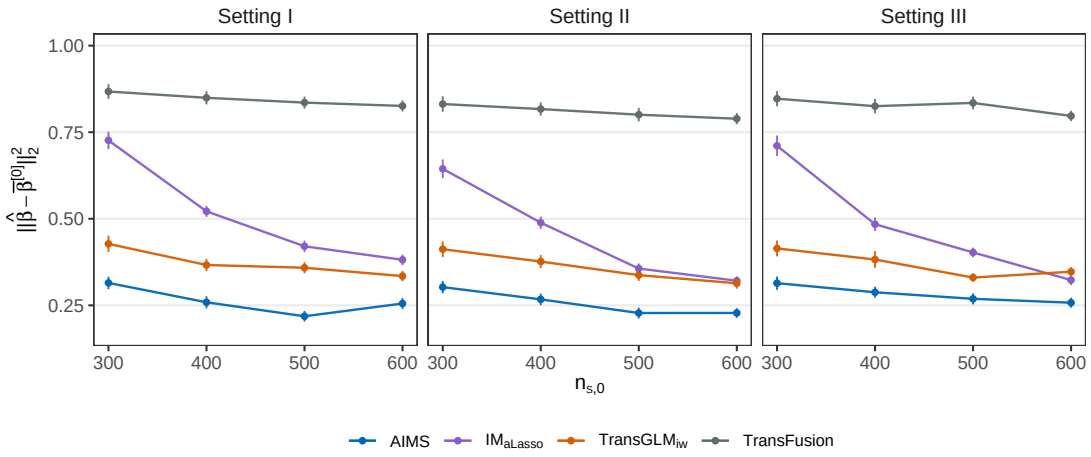

Figure 2: Simulation performance along the minority source-sample-size axis, varying $n_{\mathcal{S},0}$ with $p = 400$ and fixed $n_{\mathcal{T},0} = 2000$. Points show empirical mean squared ℓ_2 errors and error bars show one standard error across Monte Carlo repetitions. Across the displayed methods and grid points, the empirical standard deviations are all below 0.30 (standard errors below 0.03).

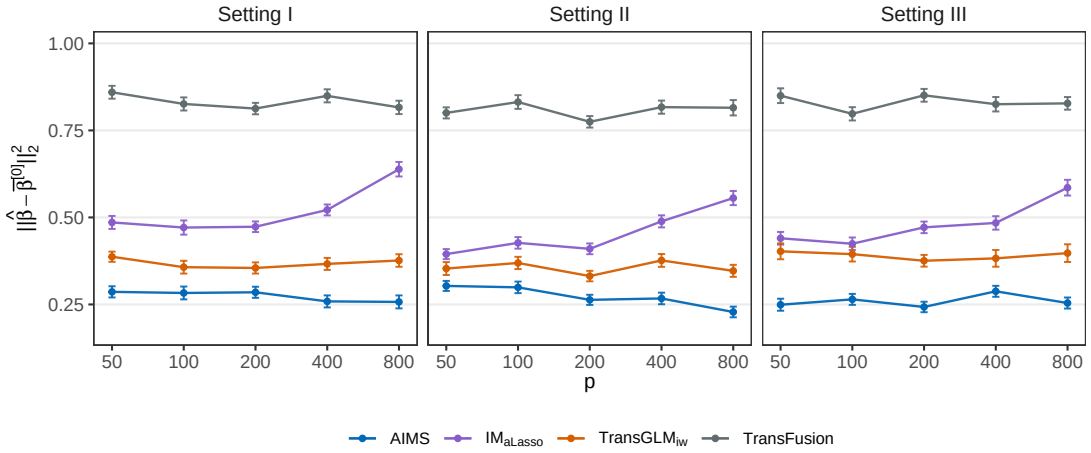

Figure 3: Simulation performance along the dimensionality axis, varying p with fixed $n_{\mathcal{S},0} = 400$ and $n_{\mathcal{T},0} = 2000$. The horizontal axis is displayed on a log scale with tick labels at the original p values. Points show empirical mean squared ℓ_2 errors and error bars show one standard error across Monte Carlo repetitions. Across the displayed methods and grid points, the empirical standard deviations are all below 0.30 (standard errors below 0.03).

Figures 2 and 3 show that AIMS is consistently the leading or near-leading method among the plotted competitors across the three nuisance-model regimes. The gap relative to $\text{IM}_{\text{aLasso}}$ illustrates the gain from using both the imputation and weighting information rather than relying on the imputation component alone. The improvement over TransFusion is most visible in settings where direct transfer is sensitive to the source–target covariate shift. $\text{TransGLM}_{\text{iw}}$ is the closest external competitor in several cases, but AIMS remains more stable across the sample-size and dimensionality axes. The overall pattern supports the role of the calibrated nuisance correction and the adaptive transfer-protection step in high-dimensional covariate shift settings. To check that the comparison is not specific to binary outcomes, we also repeat both grids under a continuous Gaussian outcome model; the results are reported in Supplementary Tables S.9 and S.10.

6. Real Data Analysis

6.1 Genetic risk model for type 2 diabetes

Type 2 diabetes (T2D), a common chronic disease characterized by insulin resistance and relative insulin deficiency, places a substantial economic and social burden on society. In the United States alone, the total estimated cost of diagnosed diabetes reached $412.9 billion in 2022 (Parker et al., 2024). Accurately predicting the risk of developing type 2 diabetes is crucial for early intervention, prevention, and reducing the long-term health and economic impacts associated with the disease. Extensive genetic studies have suggested that genetic factors play a significant role in the risk of T2D and that some non-White racial groups including African Americans and Hispanic and Latin Americans have a higher genetic predisposition to T2D (e.g., Mercader and Florez, 2017; Mahajan et al., 2018; Armstrong et al., 2024).

In our study, we aim to use EHR-linked biobank data from Mass General Brigham (MGB) biobank (Castro et al., 2022) to derive genetic risk prediction models optimized for underrepresented populations by transferring knowledge from the majority population. On the one hand, both the demographic variables and single nucleotide polymorphisms (SNPs) $\mathbf{X}$ and the EHR features $\mathbf{W}$ are readily available for all MGB biobank patients. On the other hand, the gold-standard label for the T2D status, $Y \in \{0, 1\}$, has only been collected for a subset of $n_{\mathcal{S},1} = 375$ for the majority (White race) and $n_{\mathcal{S},0} = 77$ for the minority (non-White race) patients, whose medical records were manually reviewed in a biomedical study in 2014. In this study, to obtain gold-standard labels for multiple phenotypes more efficiently, patients with International Classification of Diseases (ICD) codes for several phenotypes, such as rheumatoid arthritis and coronary artery disease, were sampled and reviewed simultaneously, leading to sampling that was not completely at random. The target data consist of $n_{\mathcal{T},1} = 5000$ and $n_{\mathcal{T},0} = 1000$ patients drawn from the MGB biobank participants with their EHR features updated in 2021. Importantly, the EHR system at MGB, as well as its ICD version, changed around 2015. Thus, the shift in EHR and genomic features between the source and target samples can be attributed to both variation in the sampling method and the difference in the time window of data collection.

Our goal is to construct a genetic risk model for the T2D status Y using the high-dimensional demographic and genetic variants $\mathbf{X} \in \mathbb{R}^{272}$. To adjust for the distributional shift between the source and target samples, we include 66 EHR features denoted as $\mathbf{W}$, con-

sisting of T2D-relevant diagnostic and procedure codes provided by Hong et al. (2021) and a measure of total health utilization. The two nuisance models are fitted on $\mathbf{Z} = (\mathbf{X}^\intercal, \mathbf{W}^\intercal)^\intercal$, which is clearly high-dimensional given the labeled sample sizes $n_{\mathcal{S},0}$ and $n_{\mathcal{S},1}$. A validation dataset $\mathcal{V}$ is created by randomly selecting and labeling 47 individuals from the target minority population solely for evaluation; these validation labels are not used for model fitting, nuisance estimation, tuning, or transfer learning. To measure predictive performance, we calculate the following metrics on $\mathcal{V}$: (1) Brier skill score (BSS), defined as $1-\widehat{\mathbb{E}}_{\mathcal{V}}\{Y-g(\mathbf{X}^\intercal\boldsymbol{\beta})\}^2/\widehat{\mathrm{Var}}_{\mathcal{V}}(Y)$ for some estimator $\boldsymbol{\beta}$, where $\widehat{\mathbb{E}}_{\mathcal{V}}$ and $\widehat{\mathrm{Var}}_{\mathcal{V}}$ respectively denote the empirical mean and variance operators on the validation samples; (2) a deviance-based goodness-of-fit score (GOF), defined as one minus the empirical mean Bernoulli deviance on the validation data: $1-2\widehat{\mathbb{E}}_{\mathcal{V}}\{-Y\mathbf{X}^\intercal\boldsymbol{\beta}+G(\mathbf{X}^\intercal\boldsymbol{\beta})\}$; and (3) area under the receiver operating characteristic curve (AUC) of the predictor $g(\mathbf{X}^\intercal\boldsymbol{\beta})$ on $\mathcal{V}$. We report BSS and GOF together with AUC because these metrics capture complementary aspects of probabilistic risk prediction: AUC measures discrimination, whereas BSS and GOF assess calibration and deviance-based likelihood fit. In this binary-outcome analysis, all methods are evaluated through predicted probabilities $g(\mathbf{X}^\top\boldsymbol{\beta}) \in (0,1)$ on the same validation set, so larger BSS and GOF values indicate better probabilistic predictive fit rather than ranking performance alone.

Following our simulation studies, we applied AIMS together with several representative baselines to construct the T2D risk model for the minority target subgroup, including IW, IM, $\mathrm{IM}_{\mathrm{XGB}}$, CORAL, adapted TransFusion, and the phenotyping-based benchmark $\mathrm{IM}_{\mathrm{PheNorm}}$. Here $\mathrm{IM}_{\mathrm{XGB}}$ uses XGBoost-generated pseudo-outcomes in the imputation step, CORAL aligns labeled source covariates with unlabeled target covariates before sparse logistic regression, and TransFusion is adapted from He et al. (2024). Additionally, we incorporated two unsupervised phenotyping algorithms, PheNorm (Yu et al., 2018) and MAP (Liao et al., 2019), for comparison. We implemented these two algorithms with the EHR features $\mathbf{W}$ on the unlabeled target-minority covariates and used their imputed phenotypes as outcomes to regress against $\mathbf{X}$. This strategy is similar to the IM approach, and we denote the two benchmarks as $\mathrm{IM}_{\mathrm{PheNorm}}$ and $\mathrm{IM}_{\mathrm{MAP}}$. Full results for additional benchmark variants, including adaptive-Lasso versions, RF-based imputation, TransGLM variants, and $\mathrm{IM}_{\mathrm{MAP}}$, are reported in Supplementary Table S.11.

We evaluated the performance of these methods on the validation set $\mathcal{V}$, with results presented in Table 1. AIMS achieved the best performance across all three evaluation metrics among the representative baselines. In particular, it substantially improved calibration-oriented performance relative to IM, increasing BSS from 0.23 to 0.36 and GOF from -0.02 to 0.16, while also attaining the highest AUC of 0.89. Among the representative baselines, $\mathrm{IM}_{\mathrm{PheNorm}}$, IM, $\mathrm{IM}_{\mathrm{XGB}}$, and TransFusion achieved relatively strong discrimination, but each yielded clearly worse BSS and GOF than AIMS. The full supplementary comparison in Table S.11 leads to the same overall conclusion.

These results also provide diagnostic evidence about the role of the proposed components in the T2D application. The source and target samples differ in chart-review design, calendar time, and EHR coding system, making source–target shift plausible; the EHR-derived variables $\mathbf{W}$ therefore provide adjustment information not captured by the genetic covariates $\mathbf{X}$ alone. The largest gains of AIMS over IM occur in BSS and GOF rather than only in AUC, suggesting improved calibration of target-minority risk probabilities. In con-

trast, CORAL and adapted TransFusion have weaker BSS and GOF, indicating that feature alignment or source-labeled transfer alone is insufficient for this high-dimensional, subgroup-imbalanced setting with target labels unavailable for training. Because gold-standard target-majority labels are not available in the current T2D evaluation set, direct majority-side predictive evaluation is not feasible for this application; the full target-minority benchmark comparison is therefore reported here and in Supplementary Table S.11.

	AIMS	IW	IM	IM_{XGB}	CORAL	TransFusion	$\text{IM}_{\text{PheNorm}}$
BSS	**0.36**	-0.11	0.23	-1.07	-0.11	0.01	-0.20
GOF	**0.16**	-0.85	-0.02	-1.46	-0.88	-0.25	-0.45
AUC	**0.89**	0.64	0.81	0.81	0.62	0.81	0.82

Table 1: Predictive performance of representative T2D risk models evaluated on the validation data. Full benchmark results are reported in Supplementary Table S.11.

6.2 Prediction of mutation-induced Gibbs free energy changes

Protein thermodynamic stability is commonly characterized by the Gibbs (un)folding free energy ΔG, while mutation-induced stability changes are correspondingly quantified by $\Delta\Delta G$, the difference in ΔG between the mutant and wild-type proteins. Accordingly, $\Delta\Delta G$ is directly relevant to protein engineering and functional variant assessment. Our application study uses two curated protein-stability datasets. S8754 (Xu et al., 2024) is a large-scale $\Delta\Delta G$ resource containing 8,754 single-point mutations across 301 proteins, compiled from ProThermDB (Nikam et al., 2021) and ThermoMutDB (Xavier et al., 2021) through extensive cleaning and manual verification. S461 (Hernández et al., 2023) is a cleaned benchmark subset retaining 461 single-point mutations with corrected annotations. Since S8754 and S461 were assembled under distinct curation pipelines and benchmark-oriented de-redundancy was used during dataset construction, nontrivial source–target distributional shift is to be expected in this setting.

Whereas standard supervised studies typically train on S8754 and evaluate on S461, the present analysis targets the labeled-scarce transfer learning regime studied by AIMS. We therefore treat S461 as the labeled source cohort and S8754 as the target cohort. The $\Delta\Delta G$ labels in S8754 are withheld throughout model fitting and used only for held-out evaluation. To construct the subgroup structure required by AIMS, we discretize assay conditions into coarse environment bins based on pH and temperature, as experimental assay conditions are known to significantly affect measured $\Delta\Delta G$ values (Vetriani et al., 1998; Tollinger et al., 2003; O'Brien et al., 2012; Leuenberger et al., 2017). In our analysis, the majority subgroup consists of mutations measured at pH 6–7, and the minority subgroup consists of mutations measured at pH 1–5. Both groups are restricted to the common temperature range 20–30°C. The resulting datasets contain 78 observations in the source minority subgroup, 237 in the source majority subgroup, 677 in the target minority subgroup, and 4,140 in the target majority subgroup.

For feature construction, both the wild-type and mutant sequences are encoded using the pretrained ESM-2-650M (Lin et al., 2023) model, and each mutation is represented by

the difference between the final-layer CLS token embeddings of the mutant and wild-type sequences. These representations are pooled across source and target samples, standardized, and reduced by principal component analysis. The 300 principal components are retained as downstream covariates. Since no separate auxiliary covariates are available in this application, we set $\mathbf{W} = \emptyset$ and $\mathbf{Z} = \mathbf{X}$, so the same feature vector serves both as the target risk factors and as the nuisance covariates for covariate shift correction.

We then apply AIMS to the minority subgroup using labeled minority-source data and unlabeled minority-target covariates, while borrowing information from the majority subgroup. We report the final AIMS estimator with built-in negative transfer protection, and compare its performance with several baselines including importance weighting, imputation-only estimators based on linear, random forest, and XGBoost pseudo-outcomes, CORAL, adapted TransFusion, and TransGLM with and without importance weighting. All methods are evaluated on the target minority subgroup using held-out ground-truth labels, and performance is summarized by root-mean-square error (RMSE), mean absolute error (MAE), Pearson correlation, and Spearman correlation, with results presented in Table 2. When a method collapses to essentially constant predictions on the target minority subgroup, the correlation measures are undefined and are therefore reported as "—".

The results in Table 2 show that AIMS achieves the best performance on all four metrics. For point prediction accuracy, AIMS achieves the lowest RMSE (2.67) and MAE (1.90), with CORAL as the closest competitor (RMSE 2.69, MAE 1.92). The advantage is more pronounced for rank-order prediction: AIMS achieves a Spearman correlation of 0.27, compared to 0.19 for TransGLM, 0.16 for TransFusion, and 0.14 for CORAL. Methods including IM, IM_{RF}, IM_{XGB}, and $\text{TransGLM}_{\text{iw}}$ collapse to essentially constant predictions and yield undefined correlations. This indicates that naive imputation-based and heavily regularized transfer estimators lose predictive signal entirely on the small and distributionally shifted minority target. AIMS retains meaningful predictive structure even in this challenging regime, supporting the generalizability of the framework beyond the T2D biobank application.

For predictive performance on the target majority subgroup, Supplementary Table S.12 reports three versions of AIMS: the preliminary estimator $\widetilde{\boldsymbol{\beta}}^{[1]}$ ($\text{AIMS}_{\text{Init}}$), the dense debiased estimator $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Dense\_Deb}}$ ($\text{AIMS}_{\text{Dense\_Deb}}$), and the final thresholded estimator $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Thr}}$ (AIMS_{Thr}). Within the target majority subgroup, $\text{AIMS}_{\text{Init}}$ yields the smallest RMSE and MAE, whereas AIMS_{Thr} attains the strongest Pearson and Spearman correlations. $\text{AIMS}_{\text{Dense\_Deb}}$ performs worst under the metrics RMSE and MAE, indicating that the debiased coefficient vector is better interpreted as an intermediate estimator than as a final prediction model. Relative to the minority-target results in Table 2, prediction on the majority subgroup is empirically easier, with smaller errors and stronger correlations. This is consistent with the substantially larger labeled majority source sample.

7. Discussion

Unlike recent literature on doubly robust estimation with high-dimensional nuisance models and low-dimensional target parameters (e.g., Tan, 2020; Zhou et al., 2025), our target estimator is sparse-regularized and lacks asymptotic linearity, which makes calibrating the nuisance models more challenging. Our debiasing step (9) addresses this issue and enables

	AIMS	IW	IM	IM_{RF}	IM_{XGB}	CORAL	TransFusion	TransGLM	$\text{TransGLM}_{\text{iw}}$
RMSE	**2.67**	2.83	2.87	2.76	2.75	2.69	3.00	2.79	2.82
MAE	**1.90**	2.06	2.09	2.00	2.00	1.92	2.19	2.00	2.05
Pearson	**0.14**	0.09	—	—	—	0.03	0.09	0.04	—
Spearman	**0.27**	0.04	—	—	—	0.14	0.16	0.19	—

Table 2: Predictive performance of the $\Delta\Delta G$ models evaluated on the target minority subgroup using held-out ground-truth labels.

proper calibration, while the downstream sparsifying step removes error accumulated in the dense debiased estimator. This approach can be generalized to other important settings, such as estimating the HTE with high-dimensional effect modifiers considered in Kato (2024) and others. Also, compared with existing work on knowledge transfer (e.g., Li et al., 2022), we tackle a more challenging setup where target labels are unavailable for training, and the debiased coefficient vector obtained in the previous step is used as a substitute for actual labeled samples in sparse regression and negative transfer protection. This makes our method and theory in this part more technically involved.

In biomedical research and other fields, such as the social sciences, both covariate shift and sampling disparity are prominent challenges that impede generalizable and responsible statistical learning. For example, leveraging a clinical trial to infer treatment effects on an observational cohort also entails covariate shift between the two cohorts (e.g., Colnet et al., 2024), and age and gender disparities have been frequently considered in recent studies (e.g., Ting et al., 2017; Ludmir et al., 2019). Though these two challenges often co-occur in practice, existing analytical tools typically address only one at a time. Our work fills this methodological gap by simultaneously addressing both challenges in a coherent and complete framework, demonstrating potential for wide application. Finally, we note several challenges in applying AIMS, including high computational costs, potential privacy constraints, and extensions to multi-source settings. These warrant future research.

Acknowledgments

The authors thank the action editor and anonymous reviewers for their constructive comments and suggestions. Doudou Zhou was supported by the MOE AcRF Tier 1 Grant A-8003569-00-00 and the NUS Startup Grant A-0009985-00-00. The authors declare no competing interests.

Technical supplementary material to

Domain Adaptation Targeting Heterogeneous and Imbalanced Subgroups

Appendix S.1. Theoretical Justification

S.1.1 Technical Assumptions and Lemmas

We use $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| \le C$ for some constant C, and $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We denote convergence in probability by $\xrightarrow{\mathbb{P}}$ and convergence in distribution by $\xrightarrow{\mathcal{D}}$. For a sequence of random variables Z_n, we use $Z_n = o_p(a_n)$ if $|Z_n|/a_n \xrightarrow{\mathbb{P}} 0$ and $Z_n = O_p(a_n)$ if $\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(|Z_n/a_n| > C) = 0$. For sequences of random variables a_n and b_n, we write $a_n \le_p b_n$, or equivalently $b_n \ge_p a_n$, if $\mathbb{P}(a_n \le b_n) \to 1$. Let $n_r = n_{\mathcal{S},r} \wedge n_{\mathcal{T},r} = \min\{n_{\mathcal{S},r}, n_{\mathcal{T},r}\}$, and assume $n_{\mathcal{S},0} = o(n_{\mathcal{S},1})$ and $n_{\mathcal{T},0} = o(n_{\mathcal{T},1})$, i.e., the minority sample sizes are much smaller than the majority sample sizes. The sub-exponential norm of a random variable X is defined as $\|X\|_{\psi_1} = \sup_{t\ge 1} t^{-1}\{E(|X|^t)\}^{1/t}$. Note that $\|X\|_{\psi_1} < C_1$ for some constant C_1, if X is sub-exponential. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{t\ge 1} t^{-1/2}\{E(|X|^t)\}^{1/t}$. Note that $\|X\|_{\psi_2} < C_2$ for some constant C_2, if X is sub-Gaussian. Let $\mathcal{I}_{\mathcal{S}}^{[r]} = \{i : R_i = r,\ S_i = 1\}$, $\mathcal{I}_{\mathcal{T}}^{[r]} = \{i : R_i = r,\ S_i = 0\}$, $r \in \{0, 1\}$.

Assumption S.1 *We assume that*

(i) $\mathbf{X}_i$ *follows a sub-Gaussian distribution with* $\max_{i,j}(|X_{ij}|) = O_p(K)$, *and* $\max_i(\|\mathbf{\Phi}_i\|_\infty)$ *is finite.*

(ii) For $r \in \{0, 1\}$, *all matrices listed below have eigenvalues that are bounded above and bounded away from zero:* $\mathbb{E}_{\mathcal{T}}^{[r]}\{\mathbf{X}\mathbf{X}^\intercal\}$, $\mathbb{E}_{\mathcal{S}}^{[r]}\{\mathbf{X}\mathbf{X}^\intercal\}$, $\bar{\mathbf{\Sigma}}^{[r]} = \bar{\mathbf{\Sigma}}_{\bar{\boldsymbol{\beta}}^{[r]}}^{[r]}$, $\mathbb{E}_{\mathcal{T}}^{[r]}\{\mathbf{\Phi}\mathbf{\Phi}^\intercal\}$, $\mathbb{E}_{\mathcal{S}}^{[r]}\{\exp(\mathbf{\Phi}^\intercal\bar{\boldsymbol{\alpha}}^{[r]})\mathbf{\Phi}\mathbf{\Phi}^\intercal\}$, $\mathbb{E}_{\mathcal{S}}^{[r]}\{\dot{b}(\mathbf{\Phi}^\intercal\bar{\boldsymbol{\gamma}}^{[r]})\mathbf{\Phi}\mathbf{\Phi}^\intercal\}$,

$$\mathbb{E}_{\mathcal{S}}^{[r]}\left[\bar{w}_j^{[r]}\dot{b}(\mathbf{\Phi}^\intercal\bar{\boldsymbol{\gamma}}^{[r]})\exp\{\mathbf{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\mathbf{\Phi}\mathbf{\Phi}^\intercal\right],$$

and $$\mathbb{E}_{\mathcal{S}}^{[r]}\left[\bar{w}_j^{[r]}\dot{b}\{\mathbf{\Phi}^\intercal(\bar{\boldsymbol{\gamma}}^{[r]} + \bar{\boldsymbol{\zeta}}_j^{[r]})\}\exp(\mathbf{\Phi}^\intercal\bar{\boldsymbol{\alpha}}^{[r]})\mathbf{\Phi}\mathbf{\Phi}^\intercal\right].$$

(iii) There exists a constant $C_w > 0$ *such that, for* $r \in \{0, 1\}$,

$$\min_{1\le j\le q} \min_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}} \bar{w}_{ij}^{[r]} \ge C_w$$

with probability tending to one.

Assumption S.2 *We assume that for* $r \in \{0, 1\}$

(i) $\log(d) \asymp \log(q)$, $n_{\iota,0} = O(n_{\iota,1})$, $\iota \in \{\mathcal{S}, \mathcal{T}\}$, *and* $n_{\mathcal{S},r} \asymp n_{\mathcal{T},r}$.

(ii) $K^4(s_{\text{nui}}^{[r]}+s_\beta^{[r]})\sqrt{\log(q)/n_r}=o(1)$, $\|\bar{\boldsymbol{\beta}}^{[r]}\|_1=o\{\sqrt{n_r/\log(q)}\}$, *and* $\max_{1\le j\le q}\|\bar{\boldsymbol{\Omega}}_j^{[r]}\|_1=O(1)$, *for* $r=0,1$.

(iii) $\bar{\boldsymbol{\xi}}_j^{[r]}$ *and* $\bar{\boldsymbol{\zeta}}_j^{[r]}$ *are* ℓ_v *sparse, i.e.,* $\|\bar{\boldsymbol{\xi}}_j^{[r]}\|_v\le R_{\text{nui},v}^{[r]}$ *and* $\|\bar{\boldsymbol{\zeta}}_j^{[r]}\|_v\le R_{\text{nui},v}^{[r]}$, *for some* $v\in[0,1]$ *satisfying* $R_{\text{nui},v}^{[r]}\{\log(q)/n_r\}^{1/2-v/2}=o(1)$.

(iv) $\bar{\boldsymbol{\delta}}$ *is* ℓ_v *sparse for some* $v\in[0,1]$ *and* $\|\bar{\boldsymbol{\delta}}\|_v\le R_{\delta,v}$, *where* $0<R_{\delta,v}=O[\{n_0/\log(q)\}^{1-v/2}]$.

Assumption S.3 *We assume that*

(i) Functions $b(\cdot)$ *and* $g(\cdot)$ *are non-decreasing, and for any* u *and* $v\in\mathbb{R}$, *there exists a positive constant* L *such that* $|\dot b(u)-\dot b(v)|\le L|u-v|$ *and* $|\dot g(u)-\dot g(v)|\le L|u-v|$.

(ii) $\max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}}\exp(|\boldsymbol{\Phi}_i^\intercal\bar{\boldsymbol{\alpha}}^{[r]}|)=O_p(1)$ *and* $\max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]},1\le j\le q}\exp\{|\boldsymbol{\Phi}_i^\intercal(\bar{\boldsymbol{\alpha}}^{[r]}+\bar{\boldsymbol{\xi}}_j^{[r]})|\}=O_p(1)$, *for* $r\in\{0,1\}$.

(iii) There exists a constant $C_Y<\infty$, *independent of* n *and* q, *such that*

$$\begin{aligned}&\max_{r,s\in\{0,1\}}\|Y-m_r^*(\mathbf{Z})\|_{\psi_2|R=r,S=s}\\&\quad\vee\max_{\substack{r,s\in\{0,1\}\\0\le j\le q}}\Big\|Y-b(\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}}_j^{[r]})\Big\|_{\psi_2|R=r,S=s}\\&\quad\vee\max_{\substack{r,s\in\{0,1\}\\0\le j\le q}}\Big\|b(\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}}_j^{[r]})-g(\mathbf{X}^\intercal\bar{\boldsymbol{\beta}}^{[r]})\Big\|_{\psi_2|R=r,S=s}\le C_Y,\end{aligned}$$

where $\bar{\boldsymbol{\gamma}}_0^{[r]}=\bar{\boldsymbol{\gamma}}^{[r]}$.

(iv) For each $r\in\{0,1\}$, *with probability tending to one, the empirical loss* $\mathcal{L}_r(\cdot)$ *satisfies the restricted strong convexity conditions (Loh and Wainwright, 2015), relative to the population target* $\bar\beta^{[r]}$. *The corresponding curvature parameters are uniformly bounded away from zero, and the tolerance parameters are uniformly bounded, over* $r\in\{0,1\}$.

Assumption S.4 *We assume that for* $r\in\{0,1\}$, *the tuning parameters satisfy*

(i) $\lambda_{\theta_j}^{[r]}\asymp K\sqrt{\log(q)/n_{\mathcal{T},r}}$, *for* $j=1,\ldots,q$, *where* $\lambda_{\theta_j}^{[r]}$ *is the tuning parameter in the nodewise lasso defined in Supplementary Section S.2.1;*

(ii) $\lambda_\alpha^{[r]}$, $\lambda_\gamma^{[r]}$, $\lambda^{[r]}$, $\lambda_{\alpha_j}^{[r]}$, $\lambda_{\gamma_j}^{[r]}\asymp\sqrt{\log(q)/n_r}$, *for* $j=1,\ldots,q$.

Lemma S.1 *Under Assumptions S.1, S.2 and S.4, we have*

$$\|\widehat{\boldsymbol{\Omega}}_j^{[r]}-\bar{\boldsymbol{\Omega}}_j^{[r]}\|_1=O_p\left\{K^2(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\sqrt{\log(q)/n_r}\right\},$$

$$\|\widehat{\boldsymbol{\Omega}}_j^{[r]}-\bar{\boldsymbol{\Omega}}_j^{[r]}\|_2^2=O_p\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\},$$

and

$$\left\{\hat\varsigma_j^{[r]}-(\bar\Omega_{j,j}^{[r]})^{-1}\right\}^2=O_p\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\},$$

where $\widehat{\varsigma}_j^{[r]}$ *is the nodewise residual variance estimator defined in Supplementary Section S.2.1. Moreover,*

$$\left|\widehat{\boldsymbol{\Omega}}_j^{[r]}\bar{\boldsymbol{\Sigma}}^{[r]}\big(\widehat{\boldsymbol{\Omega}}_j^{[r]}\big)^\intercal-\bar{\Omega}_{j,j}^{[r]}\right| \leq \|\bar{\boldsymbol{\Sigma}}^{[r]}\|_{\max}\|\widehat{\boldsymbol{\Omega}}_j^{[r]}-\bar{\boldsymbol{\Omega}}_j^{[r]}\|_1^2\wedge\lambda_{\max}(\bar{\boldsymbol{\Sigma}}^{[r]})\|\widehat{\boldsymbol{\Omega}}_j^{[r]}-\bar{\boldsymbol{\Omega}}_j^{[r]}\|_2^2+2\big|(\hat{\varsigma}_j^{[r]})^{-1}-\bar{\Omega}_{j,j}^{[r]}\big|.$$

Proof Lemma S.1 can be proved using Lemma S.2 and Theorem 3.2 in Van de Geer et al. (2014). ■

Lemma S.2 *Under Assumption 1 and Assumptions S.1–S.4, we have*

$$\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_2^2 = O_p\left\{(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\},$$

and

$$\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1 = O_p\left\{(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\sqrt{\log(q)/n_r}\right\}.$$

Lemma S.3 *Under Assumptions S.1–S.4, we have*

$$\|\widehat{\boldsymbol{\alpha}}_j^{[r]}-\bar{\boldsymbol{\alpha}}_j^{[r]}\|_2^2+\|\widehat{\boldsymbol{\gamma}}_j^{[r]}-\bar{\boldsymbol{\gamma}}_j^{[r]}\|_2^2 = O_p\left[\frac{K^4(s_{\text{nui}}^{[r]}+s_\beta^{[r]})\log(q)}{n_r}+R_{\text{nui},v}^{[r]}\left\{\frac{\log(q)}{n_r}\right\}^{1-\frac{v}{2}}\right],$$

$$\|\widehat{\boldsymbol{\alpha}}_j^{[r]}-\bar{\boldsymbol{\alpha}}_j^{[r]}\|_1+\|\widehat{\boldsymbol{\gamma}}_j^{[r]}-\bar{\boldsymbol{\gamma}}_j^{[r]}\|_1 = O_p\left[K^4(s_{\text{nui}}^{[r]}+s_\beta^{[r]})\left\{\frac{\log(q)}{n_r}\right\}^{\frac{1}{2}}+R_{\text{nui},v}^{[r]}\left\{\frac{\log(q)}{n_r}\right\}^{\frac{1-v}{2}}\right].$$

S.1.2 Proof of Proposition 2

Proof We have

$$\begin{aligned}\partial_{\boldsymbol{\beta}}\mathbb{E}[\mathcal{L}_r(\boldsymbol{\beta};\boldsymbol{\alpha}^{[r]},\boldsymbol{\gamma}^{[r]})] =&\mathbb{E}_{\mathcal{S}}[\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}^{[r]})\mathbf{X}\{b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})-Y\}\mid R=r]\\ &+\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})\}\mid R=r].\end{aligned} \tag{S.1}$$

When $m_r^\star(\mathbf{z})$ is correctly specified, i.e., $m_r^\star(\mathbf{z})=b\big(\phi(\mathbf{z})^\intercal\boldsymbol{\gamma}^{[r]}\big)$, we have that

$$\mathbb{E}_{\mathcal{S}}[\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}^{[r]})\mathbf{X}\{b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})-Y\}\mid R=r]=\mathbf{0}$$

and $\bar{\boldsymbol{\beta}}^{[r]}$ is a solution to $\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})\}\mid R=r]=\mathbf{0}$ according to its definition and the equality

$$\begin{aligned}\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})\}\mid R=r] =&\mathbb{E}_{\mathcal{T}}\big[\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})\}\mid\boldsymbol{\Phi},R=r]\mid R=r\big]\\ =&\mathbb{E}_{\mathcal{T}}\big[\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-Y\}\mid\boldsymbol{\Phi},R=r]\mid R=r\big]\\ =&\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-Y\}\mid R=r].\end{aligned}$$

When $h_r^\star(\mathbf{z})=\exp\big(\phi(\mathbf{z})^\intercal\boldsymbol{\alpha}^{[r]}\big)$, we have

$$\mathbb{E}_{\mathcal{S}}[\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}^{[r]})\mathbf{X}\{b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})-g(\mathbf{X}^\intercal\boldsymbol{\beta})\}\mid R=r]+\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\boldsymbol{\beta})-b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^{[r]})\}\mid R=r]=\mathbf{0},$$

and, at $\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}^{[r]}$, the remaining term $\mathbb{E}_{\mathcal{S}}[\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}^{[r]})\mathbf{X}\{g(\mathbf{X}^\intercal\bar{\boldsymbol{\beta}}^{[r]})-Y\}\mid R=r]=\mathbb{E}_{\mathcal{T}}[\mathbf{X}\{g(\mathbf{X}^\intercal\bar{\boldsymbol{\beta}}^{[r]})-Y\}\mid R=r]=\mathbf{0}$, showing that $\bar{\boldsymbol{\beta}}^{[r]}$ satisfies (5). As a result, (5) is a doubly robust estimating equation for $\bar{\boldsymbol{\beta}}^{[r]}$. ■

S.1.3 Proof of Lemma S.2

Proof By the convexity of $\mathcal{L}_r(\boldsymbol{\beta};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})+\lambda^{[r]}\|\boldsymbol{\beta}\|_1$, we have

$$\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})+\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}\|_1\leq\lambda^{[r]}\|\bar{\boldsymbol{\beta}}^{[r]}\|_1,$$

where

$$\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\boldsymbol{\beta};\boldsymbol{\alpha},\boldsymbol{\gamma})=\frac{\mathrm{d}\mathcal{L}_r(\boldsymbol{\beta};\boldsymbol{\alpha},\boldsymbol{\gamma})}{\mathrm{d}\boldsymbol{\beta}}=-\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\alpha})\mathbf{X}_i\{Y_i-b(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\gamma})\}-\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\mathbf{X}_i\{b(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\gamma})-g(\mathbf{X}_i^{\intercal}\boldsymbol{\beta})\}.$$

We further have

$$\begin{aligned}&\langle\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})-\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]}),(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})\rangle+\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}\|_1\\ \leq&-\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})+\lambda^{[r]}\|\bar{\boldsymbol{\beta}}^{[r]}\|_1,\end{aligned}\tag{S.2}$$

where

$$\langle\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})-\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]}),(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})\rangle=\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{g(\mathbf{X}_i^{\intercal}\widetilde{\boldsymbol{\beta}}^{[r]})-g(\mathbf{X}_i^{\intercal}\bar{\boldsymbol{\beta}}^{[r]})\}\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}).$$

We now deal with $\widetilde{\boldsymbol{\alpha}}^{[r]}$ and $\widetilde{\boldsymbol{\gamma}}^{[r]}$ in $-\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})$. We have

$$-\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]})^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})=-\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]})^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})+V_1+V_2+V_3$$

where

$$V_1=\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{b(\boldsymbol{\Phi}_i^{\intercal}\widetilde{\boldsymbol{\gamma}}^{[r]})-b(\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\}\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})=\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\dot{b}(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\gamma}^*)\boldsymbol{\Phi}_i^{\intercal}(\widetilde{\boldsymbol{\gamma}}^{[r]}-\bar{\boldsymbol{\gamma}}^{[r]})\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}),$$

$$\begin{aligned}V_2&=\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\exp(\boldsymbol{\Phi}_i^{\intercal}\widetilde{\boldsymbol{\alpha}}^{[r]})-\exp(\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\alpha}}^{[r]})\}\{Y_i-b(\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\}\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})\\&=\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\alpha}^*)\{Y_i-b(\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\}\boldsymbol{\Phi}_i^{\intercal}(\widetilde{\boldsymbol{\alpha}}^{[r]}-\bar{\boldsymbol{\alpha}}^{[r]})\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}),\end{aligned}$$

$$\begin{aligned}V_3&=\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\boldsymbol{\Phi}_i^{\intercal}\widetilde{\boldsymbol{\alpha}}^{[r]})\{b(\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})-b(\boldsymbol{\Phi}_i^{\intercal}\widetilde{\boldsymbol{\gamma}}^{[r]})\}\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})\\&=\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\boldsymbol{\Phi}_i^{\intercal}\widetilde{\boldsymbol{\alpha}}^{[r]})\dot{b}(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\gamma}^*)\boldsymbol{\Phi}_i^{\intercal}(\bar{\boldsymbol{\gamma}}^{[r]}-\widetilde{\boldsymbol{\gamma}}^{[r]})\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}),\end{aligned}$$

and $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ are on the line connecting $\widetilde{\boldsymbol{\alpha}}^{[r]}$ and $\bar{\boldsymbol{\alpha}}^{[r]}$, and $\widetilde{\boldsymbol{\gamma}}^{[r]}$ and $\bar{\boldsymbol{\gamma}}^{[r]}$, respectively.

We then upper bound V_1, V_2, and V_3 with high probability. For V_1, we have

$$\begin{aligned}V_1&\leq\left[\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\dot{b}(\boldsymbol{\Phi}_i^{\intercal}\boldsymbol{\gamma}^*)^2\{\boldsymbol{\Phi}_i^{\intercal}(\widetilde{\boldsymbol{\gamma}}^{[r]}-\bar{\boldsymbol{\gamma}}^{[r]})\}^2\right]^{1/2}\left[\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2}\\&\leq_p C_{\gamma,1}\sqrt{s_{\mathrm{nui}}^{[r]}}\lambda_{\gamma}^{[r]}\left[\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{\mathbf{X}_i^{\intercal}(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2},\end{aligned}$$

where $C_{\gamma,1}$ is a positive constant.

Since $\max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}}\exp(|\mathbf{\Phi}_i^\intercal\bar{\boldsymbol{\alpha}}^{[r]}|) = O_p(1)$ and $\max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}}\|\mathbf{\Phi}_i\|_\infty$ is bounded, and $\|\widetilde{\boldsymbol{\alpha}}^{[r]} - \bar{\boldsymbol{\alpha}}^{[r]}\|_1 = O(s_{\text{nui}}^{[r]}\lambda_\alpha^{[r]}) = o_p(1)$, then we have $\max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}}\exp(\mathbf{\Phi}_i^\intercal\boldsymbol{\alpha}^*) = O_p(1)$. Hence, for V_2, we have

$$\begin{aligned}V_2 &\le \left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(2\mathbf{\Phi}_i^\intercal\boldsymbol{\alpha}^*)\{Y_i - b(\mathbf{\Phi}_i^\intercal\bar{\boldsymbol{\gamma}}^{[r]})\}^2\{\mathbf{\Phi}_i^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} - \bar{\boldsymbol{\alpha}}^{[r]})\}^2\right]^{1/2}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2}\\ &\le_p C_{\alpha,1}\sqrt{s_{\text{nui}}^{[r]}}\lambda_\alpha^{[r]}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2},\end{aligned}$$

where $C_{\alpha,1}$ is a positive constant.

For V_3, we further decompose it as $V_{31} + V_{32}$, where

$$\begin{aligned}V_{31} &= \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\mathbf{\Phi}_i^\intercal\bar{\boldsymbol{\alpha}}^{[r]})\dot{b}(\mathbf{\Phi}_i^\intercal\boldsymbol{\gamma}^*)\mathbf{\Phi}_i^\intercal(\bar{\boldsymbol{\gamma}}^{[r]} - \widetilde{\boldsymbol{\gamma}}^{[r]})\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\\ &\le_p C_{\gamma,2}\sqrt{s_{\text{nui}}^{[r]}}\lambda_\gamma^{[r]}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2},\end{aligned}$$

$$\begin{aligned}V_{32} &= \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\exp(\mathbf{\Phi}_i^\intercal\widetilde{\boldsymbol{\alpha}}^{[r]}) - \exp(\mathbf{\Phi}_i^\intercal\bar{\boldsymbol{\alpha}}^{[r]})\}\dot{b}(\mathbf{\Phi}_i^\intercal\boldsymbol{\gamma}^*)\mathbf{\Phi}_i^\intercal(\bar{\boldsymbol{\gamma}}^{[r]} - \widetilde{\boldsymbol{\gamma}}^{[r]})\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\\ &\le \left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(2\mathbf{\Phi}_i^\intercal\boldsymbol{\alpha}^*)\dot{b}(\mathbf{\Phi}_i^\intercal\boldsymbol{\gamma}^*)^2\{\mathbf{\Phi}_i^\intercal(\bar{\boldsymbol{\gamma}}^{[r]} - \widetilde{\boldsymbol{\gamma}}^{[r]})\}^2\{\mathbf{\Phi}_i^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} - \bar{\boldsymbol{\alpha}}^{[r]})\}^2\right]^{1/2}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2},\end{aligned}$$

where $C_{\gamma,2}$ is a positive constant. Hence,

$$V_3 \le_p 2C_{\gamma,2}\sqrt{s_{\text{nui}}^{[r]}}\lambda_\gamma^{[r]}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})^2\right]^{1/2}.$$

Coming back to (S.2), we have

$$\begin{aligned}&\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{g(\mathbf{X}_i^\intercal\widetilde{\boldsymbol{\beta}}^{[r]}) - g(\mathbf{X}_i^\intercal\bar{\boldsymbol{\beta}}^{[r]})\}\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}) + \lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}\|_1\\ \le_p &\|\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]})\|_\infty\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 + \lambda^{[r]}\|\bar{\boldsymbol{\beta}}^{[r]}\|_1 + V_4, \qquad \text{(S.3)}\end{aligned}$$

where

$$V_4 = C_{\gamma,1}\sqrt{s_{\text{nui}}^{[r]}}\lambda_\gamma^{[r]}\left[\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2} + \left\{\sqrt{s_{\text{nui}}^{[r]}}(C_{\alpha,1}\lambda_\alpha^{[r]} + 2C_{\gamma,2}\lambda_\gamma^{[r]})\right\}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2\right]^{1/2}.$$

Furthermore, there exists a small enough positive constant $C_{\beta,1} < 1$, such that $\|\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]})\|_\infty \le_p C_{\beta,1}\lambda^{[r]}$ by Bernstein's inequality and a union bound. Then by (S.3), we have

$$D + \lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}\|_1 \le_p C_{\beta,1}\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 + \lambda^{[r]}\|\bar{\boldsymbol{\beta}}^{[r]}\|_1 + V_4, \qquad \text{(S.4)}$$

where

$$D = \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{\dot{g}(\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})^\intercal\mathbf{X}_i\mathbf{X}_i^\intercal(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\},$$

and $\boldsymbol{\beta}^*$ is on the line connecting $\bar{\boldsymbol{\beta}}^{[r]}$ and $\widetilde{\boldsymbol{\beta}}^{[r]}$.

By the triangle inequality,

$$\lambda^{[r]}\|(\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1 \le \lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1 + \lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1,$$

and

$$\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{(S_\beta^{[r]})^c}\|_1 \le \lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]})_{(S_\beta^{[r]})^c}\|_1+\lambda^{[r]}\|(\bar{\boldsymbol{\beta}}^{[r]})_{(S_\beta^{[r]})^c}\|_1,$$

we obtain

$$\begin{aligned}D+(1-C_{\beta,1})\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{(S_\beta^{[r]})^c}\|_1 \le_p & 2\lambda^{[r]}\|(\bar{\boldsymbol{\beta}}^{[r]})_{(S_\beta^{[r]})^c}\|_1+(1+C_{\beta,1})\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1+V_4\\ =&(1+C_{\beta,1})\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1+V_4.\end{aligned}$$

The last equality holds because $S_\beta^{[r]}=\mathrm{supp}(\bar{\boldsymbol{\beta}}^{[r]})$. Equivalently, we obtain

$$D+(1-C_{\beta,1})\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1 \le_p 2\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1+V_4.$$

We now have two possible cases by comparing $\chi 2\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1$ and $(1-\chi)V_4$, where constant $\chi\in(0,1)$. We either have

$$(1-\chi)\left\{D+(1-C_{\beta,1})\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1\right\} \le_p 2\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1 \tag{S.5}$$

or

$$\chi\left\{D+(1-C_{\beta,1})\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1\right\} \le_p V_4. \tag{S.6}$$

When (S.5) holds, since $D>0$, we have

$$(1-C_{\beta,1})\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1 \le_p (1-\chi)^{-1}2\lambda^{[r]}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1,$$

and consequently

$$\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1^2 \le_p C_{\beta,3}\|(\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]})_{S_\beta^{[r]}}\|_1^2 \le C_{\beta,3}s_\beta^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_2^2,$$

where $C_{\beta,3}$ is a positive constant. By Assumption S.3(iv), we have

$$\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_2^2=O_p\left\{s_\beta^{[r]}(\lambda^{[r]})^2\right\},\quad\text{and}\quad\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1=O_p\left(s_\beta^{[r]}\lambda^{[r]}\right).$$

When (S.6) holds, since $\lambda^{[r]}\asymp\lambda_\alpha^{[r]}\asymp\lambda_\gamma^{[r]}\asymp\sqrt{\log(q)/n_r}$, there exists a positive constant $C_{\beta,4}$ such that

$$1-C_{\beta,1}\le_p C_{\beta,4}\left\{\sqrt{s_{\mathrm{nui}}^{[r]}}\frac{\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_2}{\|\widetilde{\boldsymbol{\beta}}^{[r]}-\bar{\boldsymbol{\beta}}^{[r]}\|_1}+\sqrt{s_{\mathrm{nui}}^{[r]}}\left(\|\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal-\mathbb{E}_{\mathcal{S}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal\|_{\max}^{1/2}\vee\|\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal-\mathbb{E}_{\mathcal{T}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal\|_{\max}^{1/2}\right)\right\}.$$

Since $s_{\mathrm{nui}}^{[r]}=o\{\sqrt{n_r/\log(q)}\}$, Bernstein's inequality and a union bound show that there exists a small enough positive constant

$$C_{\beta,5}<(1-C_{\beta,1})/(2C_{\beta,4})\wedge C_{\beta,4}\sqrt{\lambda_{\min}(\bar{\boldsymbol{\Sigma}}^{[r]})/2}$$

such that, with probability tending to one,

$$C_{\beta,4}\sqrt{s_{\text{nui}}^{[r]}}\left(\|\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal - \mathbb{E}_{\mathcal{S}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal\|_{\max}^{1/2} \vee \|\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal - \mathbb{E}_{\mathcal{T}}^{[r]}\mathbf{X}_i\mathbf{X}_i^\intercal\|_{\max}^{1/2}\right) \leq C_{\beta,5}.$$

Hence,

$$\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1^2 = O_p(s_{\text{nui}}^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2^2).$$

Moreover, by the standard local restricted strong convexity condition, we have

$$\begin{aligned} D &\geq \lambda_{\min}[\mathbb{E}_{\mathcal{T}}^{[r]}\{\dot{g}(\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)\mathbf{X}_i\mathbf{X}_i^\intercal\}]\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2^2 \\ &\quad - \|\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{\dot{g}(\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)\mathbf{X}_i\mathbf{X}_i^\intercal\} - \mathbb{E}_{\mathcal{T}}^{[r]}\{\dot{g}(\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)\mathbf{X}_i\mathbf{X}_i^\intercal\}\|_{\max}\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1^2 \\ &\geq_p C_{\beta,6}\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2^2, \end{aligned}$$

where $C_{\beta,6}$ is a positive constant. Since $(1 - C_{\beta,1})\lambda^{[r]}\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 \geq 0$, we also have

$$D \leq_p C(\sqrt{s_{\text{nui}}^{[r]}}\lambda_\alpha^{[r]} + \sqrt{s_{\text{nui}}^{[r]}}\lambda_\gamma^{[r]})\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2,$$

where C is a positive constant. Consequently, we obtain

$$\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2 = O_p\left(\sqrt{s_{\text{nui}}^{[r]}}\lambda_\alpha^{[r]} + \sqrt{s_{\text{nui}}^{[r]}}\lambda_\gamma^{[r]}\right),$$

and

$$\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 = O_p\left\{s_{\text{nui}}^{[r]}\frac{(\lambda_\alpha^{[r]})^2 + (\lambda_\gamma^{[r]})^2}{\lambda^{[r]}}\right\}.$$

In conclusion, we have

$$\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_2^2 = O_p\{s_\beta^{[r]}(\lambda^{[r]})^2 + s_{\text{nui}}^{[r]}(\lambda_\alpha^{[r]})^2 + s_{\text{nui}}^{[r]}(\lambda_\gamma^{[r]})^2\},$$

and

$$\|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 = O_p\left(s_\beta^{[r]}\lambda^{[r]} + s_{\text{nui}}^{[r]}\frac{(\lambda_\alpha^{[r]})^2 + (\lambda_\gamma^{[r]})^2}{\lambda^{[r]}}\right).$$

■

S.1.4 Proof of Lemma S.3

Proof Recall $\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]} = \widetilde{\boldsymbol{\alpha}}^{[r]} - \bar{\boldsymbol{\alpha}}^{[r]} + \widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}$. When the density ratio model is correctly specified, then $\bar{\boldsymbol{\xi}}_j^{[r]} = \mathbf{0}$, and $\bar{\boldsymbol{\alpha}}_j^{[r]} = \bar{\boldsymbol{\alpha}}^{[r]}$, for $j = 1, \ldots, q$. We allow $\bar{\boldsymbol{\xi}}_j^{[r]}$ to be soft sparse, and it can be approximated well by a sparse vector. Let $\eta > 0$ be a threshold and define the threshold subset

$$S_\eta = \{k \in \{1, \ldots, d\} : |(\bar{\boldsymbol{\xi}}_j^{[r]})_k| > \eta\}.$$

In practice, samples are divided according to the sign of $\widehat{w}_{ij}^{[r]} = \widehat{\boldsymbol{\Omega}}_j^{[r]}\mathbf{X}_i$, and the two calibration problems are solved separately for $\widehat{\boldsymbol{\xi}}_{j+}^{[r]}$ and $\widehat{\boldsymbol{\xi}}_{j-}^{[r]}$. Under (iii) in Assumption S.1 and (S.9), $\widehat{w}_{ij}^{[r]} > 0$ and $\bar{w}_{ij}^{[r]} > C_w > 0$ uniformly over i and j with probability tending to one. The same error rates can be obtained when applying the stratification strategy described in Supplementary S.2.2, using similar proof techniques as those for Lemma S.7 in Zhou et al. (2025).

Since the loss function

$$\widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]}\widehat{w}_j^{[r]}\dot{b}(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[r]})\mathcal{F}^{[r]}(\boldsymbol{\xi};\widetilde{\boldsymbol{\alpha}}^{[r]}) + \lambda_{\alpha_j}^{[r]}\|\boldsymbol{\xi}\|_1$$

is convex in $\boldsymbol{\xi}$, we have

$$\widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]}\widehat{w}_j^{[r]}\dot{b}(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[r]})\left[\rho_{\mathcal{S},r}S\exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]}+\widehat{\boldsymbol{\xi}}_j^{[r]})\} - \rho_{\mathcal{T},r}(1-S)\right]\boldsymbol{\Phi}^\intercal(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]}) + \lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}\|_1 \le \lambda_{\alpha_j}^{[r]}\|\bar{\boldsymbol{\xi}}_j^{[r]}\|_1.$$

After a simple rearrangement, we have

$$\begin{aligned}
&D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]},\widetilde{\boldsymbol{\beta}}^{[r]}) + \lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}\|_1 \\
\le& -\widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]}\widehat{w}_j^{[r]}\dot{b}(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[r]})\left[\rho_{\mathcal{S},r}S\exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]}+\bar{\boldsymbol{\xi}}_j^{[r]})\} - \rho_{\mathcal{T},r}(1-S)\right]\boldsymbol{\Phi}^\intercal(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]}) + \lambda_{\alpha_j}^{[r]}\|\bar{\boldsymbol{\xi}}_j^{[r]}\|_1,
\end{aligned} \tag{S.7}$$

where

$$\begin{aligned}
&D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]},\widetilde{\boldsymbol{\beta}}^{[r]}) \\
=&\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\widehat{w}_j^{[r]}\dot{b}(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[r]})\left[\exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]}+\widehat{\boldsymbol{\xi}}_j^{[r]})\} - \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]}+\bar{\boldsymbol{\xi}}_j^{[r]})\}\right](\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})^\intercal\boldsymbol{\Phi}.
\end{aligned}$$

We first deal with $\widetilde{\boldsymbol{\alpha}}^{[r]}$, $\widetilde{\boldsymbol{\gamma}}^{[r]}$ and $\widetilde{\boldsymbol{\beta}}^{[r]}$ in $D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]},\widetilde{\boldsymbol{\beta}}^{[r]})$. Recall that under Assumption S.2, we have $\|\widetilde{\boldsymbol{\alpha}}^{[r]}-\bar{\boldsymbol{\alpha}}^{[r]}\|_1 = O_p(s_{\text{nui}}^{[r]}\lambda_\alpha^{[r]}) = o_p(1)$ and $\|\widetilde{\boldsymbol{\gamma}}^{[r]}-\bar{\boldsymbol{\gamma}}^{[r]}\|_1 = O_p(s_{\text{nui}}^{[r]}\lambda_\gamma^{[r]}) = o_p(1)$. Under (i) in Assumption S.1 and (i) in Assumption S.3, there exists a positive constant $c_{\alpha,1}$ such that

$$\max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}}\left|\boldsymbol{\Phi}_i^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]}-\bar{\boldsymbol{\alpha}}^{[r]})\right| \le_p c_{\alpha,1}, \quad \text{and} \quad \max_{i\in\mathcal{I}_{\mathcal{S}}^{[r]}}\left|\dot{b}(\boldsymbol{\Phi}_i^\intercal\bar{\boldsymbol{\gamma}}^{[r]}) - \dot{b}(\boldsymbol{\Phi}_i^\intercal\widetilde{\boldsymbol{\gamma}}^{[r]})\right| = o_p(1). \tag{S.8}$$

By Lemma S.1, under Assumption S.2, we have $\|\widehat{\boldsymbol{\Omega}}_j^{[r]}-\bar{\boldsymbol{\Omega}}_j^{[r]}\|_1 = O_p\left\{K^2(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\sqrt{\log(q)/n_r}\right\} = o_p(1)$. By (i) in Assumption S.1 and (ii) in Assumption S.2, we have

$$\max_i|\widehat{w}_{ij}^{[r]}-\bar{w}_{ij}^{[r]}| = O_p\left\{K^3(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\sqrt{\log(q)/n_r}\right\} = o_p(1). \tag{S.9}$$

By (S.8) and (S.9), under (iii) in Assumption S.1, we have

$$D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\widetilde{\boldsymbol{\alpha}}^{[r]},\widetilde{\boldsymbol{\gamma}}^{[r]},\widetilde{\boldsymbol{\beta}}^{[r]}) \ge_p c_{\xi,1}D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]},\bar{\boldsymbol{\beta}}^{[r]}), \tag{S.10}$$

where $c_{\xi,1}$ is a positive constant, and

$$D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]}, \bar{\boldsymbol{\xi}}_j^{[r]}; \bar{\boldsymbol{\alpha}}^{[r]}, \bar{\boldsymbol{\gamma}}^{[r]}, \bar{\boldsymbol{\beta}}^{[r]}) = \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \bar{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]}) \left[\exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \widehat{\boldsymbol{\xi}}_j^{[r]})\} - \exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right] \boldsymbol{\Phi}^\intercal(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}).$$

Second, we deal with $\widetilde{\boldsymbol{\alpha}}^{[r]}$, $\widetilde{\boldsymbol{\gamma}}^{[r]}$ and $\widetilde{\boldsymbol{\beta}}^{[r]}$ in $\widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \widehat{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \left[\rho_{\mathcal{S},r} S \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} - \rho_{\mathcal{T},r}(1-S)\right] \boldsymbol{\Phi}$. We decompose

$$\begin{aligned}
&- \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \widehat{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \left[\rho_{\mathcal{S},r} S \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} - \rho_{\mathcal{T},r}(1-S)\right] \boldsymbol{\Phi}^\intercal(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}) \\
=&\mathbf{T}_0^\intercal(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}) + T_1 + T_2 + T_3 + T_4, \qquad \text{(S.11)}
\end{aligned}$$

where

$$\mathbf{T}_0 = \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \bar{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]}) \left[\rho_{\mathcal{T},r}(1-S) - \rho_{\mathcal{S},r} S \exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right] \boldsymbol{\Phi}$$

$$T_1 = \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \bar{w}_j^{[r]} \dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]}) \left[\exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} - \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right] (\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^\intercal \boldsymbol{\Phi},$$

$$T_2 = \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \bar{w}_j^{[r]} \{\dot{b}(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) - \dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]})\} \left[\rho_{\mathcal{T},r}(1-S) - \rho_{\mathcal{S},r} S \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right] (\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^\intercal \boldsymbol{\Phi},$$

$$T_3 = \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \mathbf{e}_j{}^\intercal (\widehat{\boldsymbol{\Omega}}^{[r]} - \bar{\boldsymbol{\Omega}}^{[r]}) \mathbf{X} \dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]}) \left[\rho_{\mathcal{T},r}(1-S) - \rho_{\mathcal{S},r} S \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right] (\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^\intercal \boldsymbol{\Phi},$$

and

$$T_4 = \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \mathbf{e}_j{}^\intercal (\widehat{\boldsymbol{\Omega}}^{[r]} - \bar{\boldsymbol{\Omega}}^{[r]}) \mathbf{X} \{\dot{b}(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) - \dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]})\} \left[\rho_{\mathcal{T},r}(1-S) - \rho_{\mathcal{S},r} S \exp\{\boldsymbol{\Phi}^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right] (\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^\intercal \boldsymbol{\Phi}.$$

Note that by (S.8), for each i there exists $u_i \in (0,1)$ such that

$$\begin{aligned}
&\left|\exp\{\boldsymbol{\Phi}_i^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} - \exp\{\boldsymbol{\Phi}_i^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right| \\
=&\exp[\boldsymbol{\Phi}_i^\intercal\{u_i \bar{\boldsymbol{\alpha}}^{[r]} + (1-u_i)\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]}\}] \left|(\bar{\boldsymbol{\alpha}}^{[r]} - \widetilde{\boldsymbol{\alpha}}^{[r]})^\intercal \boldsymbol{\Phi}_i\right| \\
=&\exp\{\boldsymbol{\Phi}_i^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} \exp\{(1-u_i)\boldsymbol{\Phi}_i^\intercal(\widetilde{\boldsymbol{\alpha}}^{[r]} - \bar{\boldsymbol{\alpha}}^{[r]})\} \left|(\bar{\boldsymbol{\alpha}}^{[r]} - \widetilde{\boldsymbol{\alpha}}^{[r]})^\intercal \boldsymbol{\Phi}_i\right| \\
\leq_p &e^{c_{\alpha,1}} \exp\{\boldsymbol{\Phi}_i^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} \left|(\bar{\boldsymbol{\alpha}}^{[r]} - \widetilde{\boldsymbol{\alpha}}^{[r]})^\intercal \boldsymbol{\Phi}_i\right|. \qquad \text{(S.12)}
\end{aligned}$$

By (S.12) and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
T_1 &\leq e^{c_{\alpha,1}} \left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} (\bar{w}_j^{[r]})^2 \{\dot{b}(\boldsymbol{\Phi}^\intercal \bar{\boldsymbol{\gamma}}^{[r]})\}^2 \exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} \{(\widetilde{\boldsymbol{\alpha}}^{[r]} - \bar{\boldsymbol{\alpha}}^{[r]})^\intercal \boldsymbol{\Phi}\}^2\right]^{1/2} T_{1,1}^{1/2} \\
&\leq_p c_{\xi,2} \{K^2 s_{\text{nui}}^{[r]} (\lambda_\alpha^{[r]})^2\}^{1/2} T_{1,1}^{1/2}, \qquad \text{(S.13)}
\end{aligned}$$

where

$$T_{1,1} = \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \exp\{\boldsymbol{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} \{(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^\intercal \boldsymbol{\Phi}\}^2,$$

and $c_{\xi,2}$ is a positive constant. The last inequality holds since $\max_{i,j} |\bar{w}_{ij}^{[r]}| = O_p(K)$, $\max_i \dot{b}(\boldsymbol{\Phi}_i^\intercal \bar{\boldsymbol{\gamma}}^{[r]}) = O_p(1)$, $\max_{i,j} \exp\{\boldsymbol{\Phi}_i^\intercal(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} = O_p(1)$, which are implied by Assumptions S.1 and S.3.

For T_2, we decompose it as $\mathbf{T}_{2,1}^{\intercal}(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}) + T_{2,2}$, where

$$\mathbf{T}_{2,1} = \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\bar{w}_j^{[r]}\{\dot{b}(\boldsymbol{\Phi}^{\intercal}\widetilde{\boldsymbol{\gamma}}^{[r]}) - \dot{b}(\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\}\boldsymbol{\Phi} - \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\bar{w}_j^{[r]}\{\dot{b}(\boldsymbol{\Phi}^{\intercal}\widetilde{\boldsymbol{\gamma}}^{[r]}) - \dot{b}(\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\}\exp\{\boldsymbol{\Phi}^{\intercal}(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\boldsymbol{\Phi},$$

and

$$T_{2,2} = \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\bar{w}_j^{[r]}\{\dot{b}(\boldsymbol{\Phi}^{\intercal}\widetilde{\boldsymbol{\gamma}}^{[r]}) - \dot{b}(\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\}\left[\exp\{\boldsymbol{\Phi}^{\intercal}(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} - \exp\{\boldsymbol{\Phi}^{\intercal}(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right](\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^{\intercal}\boldsymbol{\Phi}.$$

Under (i) in Assumption S.3, by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\mathbf{T}_{2,1}^{\intercal}(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}) \\
\leq& T_{2,3}^{1/2}\left[\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}(\bar{w}_j^{[r]})^2 L^2\{(\widetilde{\boldsymbol{\gamma}}^{[r]} - \bar{\boldsymbol{\gamma}}^{[r]})^{\intercal}\boldsymbol{\Phi}\}^2\right]^{1/2} \\
&+ T_{1,1}^{1/2}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}(\bar{w}_j^{[r]})^2\exp\{\boldsymbol{\Phi}^{\intercal}(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}L^2\{(\widetilde{\boldsymbol{\gamma}}^{[r]} - \bar{\boldsymbol{\gamma}}^{[r]})^{\intercal}\boldsymbol{\Phi}\}^2\right]^{1/2} \\
\leq_p& \{c_{\xi,3}K^2 s_{\mathrm{nui}}^{[r]}(\lambda_{\gamma}^{[r]})^2\}^{1/2}(T_{2,3}^{1/2} + T_{1,1}^{1/2}),
\end{aligned} \tag{S.14}$$

where $T_{2,3} = \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^{\intercal}\boldsymbol{\Phi}\}^2$, and $c_{\xi,3}$ is a positive constant. By (S.12) and the Cauchy–Schwarz inequality, under (ii) in Assumption S.2, we further have

$$T_{2,2} = o_p\left[\{K^2 s_{\mathrm{nui}}^{[r]}(\lambda_{\alpha}^{[r]})^2\}^{1/2}T_{1,1}^{1/2}\right].$$

We then decompose T_3 as $\mathbf{T}_{3,1}^{\intercal}(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}) + T_{3,2}$, where

$$\mathbf{T}_{3,1} = \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\{\widehat{w}_j^{[r]} - \bar{w}_j^{[r]}\}\dot{b}(\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\boldsymbol{\Phi} - \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\widehat{w}_j^{[r]} - \bar{w}_j^{[r]}\}\dot{b}(\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\exp\{\boldsymbol{\Phi}^{\intercal}(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\boldsymbol{\Phi}$$

and

$$T_{3,2} = \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\{\widehat{w}_j^{[r]} - \bar{w}_j^{[r]}\}\dot{b}(\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}}^{[r]})\left[\exp\{\boldsymbol{\Phi}^{\intercal}(\bar{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\} - \exp\{\boldsymbol{\Phi}^{\intercal}(\widetilde{\boldsymbol{\alpha}}^{[r]} + \bar{\boldsymbol{\xi}}_j^{[r]})\}\right](\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]})^{\intercal}\boldsymbol{\Phi}.$$

By Lemma S.1, under (ii) in Assumption S.1, we have

$$\mathbf{T}_{3,1}^{\intercal}(\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}) \leq_p \left\{c_{\xi,4}K^4(s_{\beta}^{[r]} + s_{\mathrm{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2} + T_{1,1}^{1/2}), \tag{S.15}$$

where $c_{\xi,4}$ is a positive constant, and

$$T_{3,2} = o_p\left\{K^4(s_{\beta}^{[r]} + s_{\mathrm{nui}}^{[r]})\log(q)/n_r T_{1,1}\right\}^{1/2}.$$

Further, by (S.9), we can show that $|T_4|$ is bounded by a constant multiple of the upper bound of $|T_2|$ proved above, with probability tending to 1.

Hence, by (S.7),(S.10), (S.11), (S.13), (S.14), and (S.15), there exists a positive constant $c_{\xi,5}$ such that

$$\begin{aligned}
c_{\xi,1}D_{\alpha}^{j}(\widehat{\boldsymbol{\xi}}_j^{[r]}, \bar{\boldsymbol{\xi}}_j^{[r]}; \bar{\boldsymbol{\alpha}}^{[r]}, \bar{\boldsymbol{\gamma}}^{[r]}, \bar{\boldsymbol{\beta}}^{[r]}) + \lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}\|_1 \leq_p& \|\mathbf{T}_0\|_{\infty}\|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_1 + \lambda_{\alpha_j}^{[r]}\|\bar{\boldsymbol{\xi}}_j^{[r]}\|_1 \\
&+ c_{\xi,5}\left\{K^4(s_{\beta}^{[r]} + s_{\mathrm{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2} + T_{1,1}^{1/2}).
\end{aligned} \tag{S.16}$$

We then upper bound $\|\mathbf{T}_0\|_\infty$ with high probability. Since $\bar{w}_j^{[r]}\dot{b}(\mathbf{\Phi}^\intercal\bar{\boldsymbol{\gamma}}^{[r]})\mathbf{\Phi}$ from the target population and $\bar{w}_j^{[r]}\dot{b}(\mathbf{\Phi}^\intercal\bar{\boldsymbol{\gamma}}^{[r]})\exp\{\mathbf{\Phi}^\intercal(\bar{\boldsymbol{\alpha}}^{[r]}+\bar{\boldsymbol{\xi}}_j^{[r]})\}\mathbf{\Phi}$ from the source population are sub-Gaussian, by the definition of $\bar{\boldsymbol{\xi}}_j^{[r]}$, Bernstein's inequality, and a union bound, there exists a small enough positive constant $c_{\xi,6}$, such that $\|\mathbf{T}_0\|_\infty \leq_p c_{\xi,6}\lambda_{\alpha_j}^{[r]}$. Hence, by (S.16), we obtain

$$c_{\xi,1}D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]},\bar{\boldsymbol{\beta}}^{[r]})+\lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}\|_1 \leq_p c_{\xi,6}\lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]}\|_1+\lambda_{\alpha_j}^{[r]}\|\bar{\boldsymbol{\xi}}_j^{[r]}\|_1 + c_{\xi,5}\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2}+T_{1,1}^{1/2}). \tag{S.17}$$

By the triangle inequality,

$$\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1 \leq \lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1+\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1,$$

and

$$\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1 \leq \lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1+\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1,$$

we obtain

$$\begin{aligned}&c_{\xi,1}D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]},\bar{\boldsymbol{\beta}}^{[r]})+(1-c_{\xi,6})\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1\\ \leq_p&2\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1+(1+c_{\xi,6})\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1+c_{\xi,5}\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2}+T_{1,1}^{1/2}),\end{aligned}$$

or equivalently,

$$\begin{aligned}&c_{\xi,1}D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]},\bar{\boldsymbol{\beta}}^{[r]})+(1-c_{\xi,6})\lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]}\|_1\\ \leq_p&2\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1+2\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1+c_{\xi,5}\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2}+T_{1,1}^{1/2}).\end{aligned} \tag{S.18}$$

Note that since $\bar{\boldsymbol{\xi}}_j^{[r]}$ is allowed to be soft sparse, $\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1$ can be nonzero.

We have two possible cases by comparing $\chi\{2\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1+2\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1\}$ and $(1-\chi)c_{\xi,5}\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2}+T_{1,1}^{1/2})$, for a $\chi\in(0,1)$. We have either

$$\chi\widetilde{D}_\alpha^j \leq_p c_{\xi,5}\left\{K^4(s_\beta^{[r]}+s_{\text{nui}}^{[r]})\log(q)/n_r\right\}^{1/2}(T_{2,3}^{1/2}+T_{1,1}^{1/2}), \tag{S.19}$$

or

$$(1-\chi)\widetilde{D}_\alpha^j \leq_p 2\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1+2\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1, \tag{S.20}$$

where

$$\widetilde{D}_\alpha^j = c_{\xi,1}D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]},\bar{\boldsymbol{\beta}}^{[r]})+(1-c_{\xi,6})\lambda_{\alpha_j}^{[r]}\|\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]}\|_1$$

When equation (S.20) holds, since $c_{\xi,1}D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]},\bar{\boldsymbol{\xi}}_j^{[r]};\bar{\boldsymbol{\alpha}}^{[r]},\bar{\boldsymbol{\gamma}}^{[r]},\bar{\boldsymbol{\beta}}^{[r]})\geq 0$, we have

$$(1-c_{\xi,6})\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1 \leq 2(1-\chi)^{-1}\lambda_{\alpha_j}^{[r]}\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1+(1-\chi)^{-1}(1+c_{\xi,6})\lambda_{\alpha_j}^{[r]}\|(\widehat{\boldsymbol{\xi}}_j^{[r]}-\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta}\|_1.$$

By equation (43) in Negahban et al. (2012), we have

$$D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]}, \bar{\boldsymbol{\xi}}_j^{[r]}; \bar{\boldsymbol{\alpha}}^{[r]}, \bar{\boldsymbol{\gamma}}^{[r]}, \bar{\boldsymbol{\beta}}^{[r]}) \geq C_1 \|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_2^2 - C_2 \frac{\log(p+q)}{n_r} \|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1^2$$

for some positive constant C_1 and C_2. Then by Theorem 1 in Negahban et al. (2012), we have

$$\|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_2^2 \leq_p C_4(\lambda_{\alpha_j}^{[r]})^2 |S_\eta| + C_3 \lambda_{\alpha_j}^{[r]} \left\{ \frac{\log(d)}{n_r} \|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1^2 + \|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1 \right\}$$

for some positive constant C_3 and C_4. By the proof of Corollary 3 in Negahban et al. (2012), under (iii) in Assumption S.2, we have

$$|S_\eta| \leq R_{\text{nui},v}^{[r]} \eta^{-v},$$

and

$$\|(\bar{\boldsymbol{\xi}}_j^{[r]})_{S_\eta^c}\|_1 \leq R_{\text{nui},v}^{[r]} \eta^{1-v}.$$

Setting $\eta \asymp \sqrt{n_r^{-1}\log(d)} \asymp \sqrt{n_r^{-1}\log(q)}$, we have

$$\|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_2^2 = O_p\left[R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-v/2} \right]. \tag{S.21}$$

When (S.19) holds, since $(1 - c_{\xi,6})\lambda_{\alpha_j}^{[r]} \|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_1 > 0$, we have

$$c_{\xi,1} D_\alpha^j(\widehat{\boldsymbol{\xi}}_j^{[r]}, \bar{\boldsymbol{\xi}}_j^{[r]}; \bar{\boldsymbol{\alpha}}^{[r]}, \bar{\boldsymbol{\gamma}}^{[r]}, \bar{\boldsymbol{\beta}}^{[r]}) \leq \chi^{-1} c_{\xi,5} \left\{ K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)/n_r \right\}^{1/2} (T_{2,3}^{1/2} + T_{1,1}^{1/2}).$$

Under Assumptions S.1 and S.3, we have

$$\|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_2^2 = O_p\left\{ K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)/n_r \right\},$$

and

$$\|\widehat{\boldsymbol{\xi}}_j^{[r]} - \bar{\boldsymbol{\xi}}_j^{[r]}\|_1 = O_p\left\{ K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \sqrt{\log(q)/n_r} \right\}.$$

In conclusion, we have

$$\|\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]}\|_2^2 = O_p\left\{ K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)/n_r + R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-v/2} \right\},$$

and

$$\|\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]}\|_1 = O_p\left\{ K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \sqrt{\log(q)/n_r} + R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1/2-v/2} \right\}.$$

Following similar derivations, under Assumptions S.1–S.4, we have

$$\|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_2^2 = O_p\left\{ K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)/n_r + R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-v/2} \right\},$$

and

$$\|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_1 = O_p\left\{K^4(s_\beta^{[r]} + s_{\text{nui}}^{[r]})\sqrt{\log(q)/n_r} + R_{\text{nui},v}^{[r]}\left\{\frac{\log(q)}{n_r}\right\}^{1/2-v/2}\right\}.$$

■

S.1.5 Proof of Lemma 4

Proof Recall that

$$\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\boldsymbol{\beta};\boldsymbol{\alpha},\boldsymbol{\gamma}) = -\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha})\mathbf{X}\{Y - b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma})\} - \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\mathbf{X}\{b(\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}) - g(\mathbf{X}^\intercal\boldsymbol{\beta})\},$$

and

$$\partial_{\boldsymbol{\beta}}^2\mathcal{L}_r(\boldsymbol{\beta};\boldsymbol{\alpha},\boldsymbol{\gamma}) = \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\dot{g}(\mathbf{X}^\intercal\boldsymbol{\beta})\mathbf{X}\mathbf{X}^\intercal.$$

We decompose

$$\begin{aligned}\widehat{\beta}_{\text{Deb},j}^{[r]} - \bar{\beta}_j^{[r]} &= \widetilde{\beta}_j^{[r]} - \bar{\beta}_j^{[r]} + \mathbf{e}_j^\intercal\widehat{\boldsymbol{\Omega}}^{[r]}\left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]}\exp(\boldsymbol{\Phi}^\intercal\widehat{\boldsymbol{\alpha}}_j^{[r]})\mathbf{X}\{Y - b(\boldsymbol{\Phi}^\intercal\widehat{\boldsymbol{\gamma}}_j^{[r]})\} + \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]}\mathbf{X}\{b(\boldsymbol{\Phi}^\intercal\widehat{\boldsymbol{\gamma}}_j^{[r]}) - g(\mathbf{X}^\intercal\widetilde{\boldsymbol{\beta}}^{[r]})\}\right]\\ &= \widetilde{\beta}_j^{[r]} - \bar{\beta}_j^{[r]} - \mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]}) + T_1 + T_2 + T_3,\end{aligned}$$

where

$$T_1 = -\mathbf{e}_j^\intercal\left(\widehat{\boldsymbol{\Omega}}^{[r]} - \bar{\boldsymbol{\Omega}}^{[r]}\right)\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]}),$$

$$T_2 = -\mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}\left\{\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]}) - \partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]})\right\},$$

$$T_3 = -\mathbf{e}_j^\intercal\left(\widehat{\boldsymbol{\Omega}}^{[r]} - \bar{\boldsymbol{\Omega}}^{[r]}\right)\left\{\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]}) - \partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]})\right\}.$$

By Lemma S.1 and Proposition 2, there exists a constant $c > 0$, such that

$$\begin{aligned}|T_1| &\le \|\mathbf{e}_j^\intercal\left(\widehat{\boldsymbol{\Omega}}^{[r]} - \bar{\boldsymbol{\Omega}}^{[r]}\right)\|_1\|\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\bar{\boldsymbol{\alpha}}_j^{[r]},\bar{\boldsymbol{\gamma}}_j^{[r]})\|_\infty\\ &\le_p cK^2(s_\beta^{[r]} + s_{\text{nui}}^{[r]})\frac{\log(q)}{n_r}.\end{aligned}$$

We decompose T_2 as $T_{2,1} + T_{2,2}$, where

$$\begin{aligned}T_{2,1} =& -\mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}\left\{\partial_{\boldsymbol{\beta}}\mathcal{L}_r(\widetilde{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]}) - \partial_{\boldsymbol{\beta}}\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]})\right\}\\ =& -\mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}\partial_{\boldsymbol{\beta}}^2\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]})(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}) - \mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}\{\partial_{\boldsymbol{\beta}}^2\mathcal{L}_r(\boldsymbol{\beta}^*;\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]}) - \partial_{\boldsymbol{\beta}}^2\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]})\}(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\\ =& -\widetilde{\beta}_j^{[r]} + \bar{\beta}_j^{[r]} - \mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}(\widehat{\boldsymbol{\Sigma}}^{[r]} - \bar{\boldsymbol{\Sigma}}^{[r]})(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\\ &- \mathbf{e}_j^\intercal\bar{\boldsymbol{\Omega}}^{[r]}\{\partial_{\boldsymbol{\beta}}^2\mathcal{L}_r(\boldsymbol{\beta}^*;\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]}) - \partial_{\boldsymbol{\beta}}^2\mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]};\widehat{\boldsymbol{\alpha}}_j^{[r]},\widehat{\boldsymbol{\gamma}}_j^{[r]})\}(\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}),\end{aligned}$$

$T_{2,2} = -\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \left\{ \partial_{\boldsymbol{\beta}} \mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \widehat{\boldsymbol{\alpha}}_j^{[r]}, \widehat{\boldsymbol{\gamma}}_j^{[r]}) - \partial_{\boldsymbol{\beta}} \mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]}) \right\}$, and $\boldsymbol{\beta}^*$ is on the line connecting $\bar{\boldsymbol{\beta}}^{[r]}$ and $\widetilde{\boldsymbol{\beta}}^{[r]}$. We further have

$$
\begin{aligned}
|T_{2,1} - \bar{\beta}_j^{[r]} + \widetilde{\beta}_j^{[r]}| &\le \|\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]}\|_1 \|\widehat{\mathbf{\Sigma}}^{[r]} - \bar{\mathbf{\Sigma}}^{[r]}\|_{\max} \|\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_1 + \sup_i |\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X}_i| L |\widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} \{\mathbf{X}^\intercal (\widetilde{\boldsymbol{\beta}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]})\}^2| \\
&= O_p\left\{ (s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)/n_r \right\} + O_p\left\{ K(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)/n_r \right\}.
\end{aligned}
$$

The last inequality holds under (i) in Assumption S.1, (ii) in Assumption S.2, (i) in Assumption S.3, Lemma S.2, and $\|\widehat{\mathbf{\Sigma}}^{[r]} - \bar{\mathbf{\Sigma}}^{[r]}\|_{\max} = O_p\{\sqrt{\log(q)/n_r}\}$ given the proof of Theorem 3.3 in Van de Geer et al. (2014).

Using the moment condition (14), under Assumption 1, (ii) and (iii) in Assumption S.2, (i) in Assumption S.3, by Lemma S.3, we have

$$
\begin{aligned}
|T_{2,2}| \le & \left\| \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \exp(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\alpha}}_j^{[r]}) \dot{b}(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\gamma}}_j^{[r]}) \mathbf{\Phi} - \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \dot{b}(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\gamma}}_j^{[r]}) \mathbf{\Phi} \right\|_\infty \|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_1 \\
& + \|\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \exp(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\alpha}}_j^{[r]}) \{Y_i - b(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\gamma}}_j^{[r]})\} \mathbf{\Phi}\|_\infty \|\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]}\|_1 + |T_{2,3}| + |T_{2,4}| + |T_{2,5}| \\
= & O_p\left[K^4 (s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \frac{\log(q)}{n_r} + R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-\frac{v}{2}} \right] + |T_{2,3}| + |T_{2,4}| + |T_{2,5}|,
\end{aligned}
$$

where

$$
|T_{2,3}| \le L \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} |\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X}| \{\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]})\}^2 = O_p(K \|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_2^2),
$$

$$
\begin{aligned}
|T_{2,4}| \le & \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \exp\{|\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]})|\} \dot{b}(\mathbf{\Phi}^\intercal \widehat{\boldsymbol{\gamma}}_j^{[r]}) |\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]})| |\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]})| \\
& + L \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \exp(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\alpha}}_j^{[r]}) \{\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]})\}^2 \\
= & O_p(K \|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_2^2 + K \|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_2 \|\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]}\|_2),
\end{aligned}
$$

and

$$
\begin{aligned}
|T_{2,5}| \le & \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \exp\{|\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]})|\} |Y_i - b(\mathbf{\Phi}^\intercal \widehat{\boldsymbol{\gamma}}_j^{[r]})| \{\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]})\}^2 \\
& + L \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \mathbf{X} \exp(\mathbf{\Phi}^\intercal \bar{\boldsymbol{\alpha}}_j^{[r]}) |\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]})| |\mathbf{\Phi}^\intercal (\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]})| \\
= & O_p(K \|\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]}\|_2^2 + K \|\widehat{\boldsymbol{\gamma}}_j^{[r]} - \bar{\boldsymbol{\gamma}}_j^{[r]}\|_2 \|\widehat{\boldsymbol{\alpha}}_j^{[r]} - \bar{\boldsymbol{\alpha}}_j^{[r]}\|_2).
\end{aligned}
$$

Since $|T_3|$ is bounded by a constant multiple of the proved upper bound of $|T_1|$ and $|T_2|$ with probability tending to 1, Lemma S.3 gives

$$
\begin{aligned}
\widehat{\beta}_{\text{Deb},j}^{[r]} - \bar{\beta}_j^{[r]} = & -\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \partial_{\boldsymbol{\beta}} \mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]}) \\
& + O_p\left[\frac{K^5 (s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \log(q)}{n_r} + K R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-\frac{v}{2}} \right].
\end{aligned} \tag{S.22}
$$

■

S.1.6 Proof of Theorem 6

Proof Since $-\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \partial_{\boldsymbol{\beta}} \mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]})$ is mean-zero and sub-exponential with sub-exponential norms uniformly bounded for all j under Assumption 1, (i) in Assumption S.1, (ii) in S.2 and (ii) in S.3, Bernstein's inequality and a union bound give

$$\max_{1\le j\le q} \left| -\mathbf{e}_j^\intercal \bar{\mathbf{\Omega}}^{[r]} \partial_{\boldsymbol{\beta}} \mathcal{L}_r(\bar{\boldsymbol{\beta}}^{[r]}; \bar{\boldsymbol{\alpha}}_j^{[r]}, \bar{\boldsymbol{\gamma}}_j^{[r]}) \right| = O_p\left\{ \sqrt{\frac{\log(q)}{n_r}} \right\}.$$

By (S.22), we further have

$$\|\widehat{\boldsymbol{\beta}}_{\text{Deb}}^{[r]} - \bar{\boldsymbol{\beta}}^{[r]}\|_\infty = O_p\left(\sqrt{\log(q)} \left[n_r^{-1/2} + \frac{K^5(s_\beta^{[r]} + s_{\text{nui}}^{[r]})\log(q)}{n_r} + K R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-\frac{v}{2}} \right] \right).$$

Then by the proof of Theorem 4.3 in Battey et al. (2018), we obtain the results in Theorem 6.

S.1.7 Proof of Lemma 9

Here we consider a general scenario where $\bar{\boldsymbol{\beta}}^{[0]}$, $\bar{\boldsymbol{\beta}}^{[1]}$, and $\bar{\boldsymbol{\delta}}$ are all ℓ_v sparse, for a fixed $v \in [0,1]$. Note that there exist positive constants C and c such that

$$\|\widehat{\boldsymbol{\beta}}_{\text{Deb}}^{[0]} - \widehat{\boldsymbol{\beta}}_{\text{Thr}}^{[1]} - \bar{\boldsymbol{\delta}}\|_\infty \le \|\widehat{\boldsymbol{\beta}}_{\text{Deb}}^{[0]} - \bar{\boldsymbol{\beta}}^{[0]}\|_\infty + \|\widehat{\boldsymbol{\beta}}_{\text{Thr}}^{[1]} - \bar{\boldsymbol{\beta}}^{[1]}\|_\infty \le C\left\{ \sqrt{\log(q)}(n_0^{-1/2} + \Delta) \right\},$$

where

$$\Delta = \max_{r=0,1} \left\{ K^5(s_\beta^{[r]} + s_{\text{nui}}^{[r]}) \frac{\log(q)}{n_r} + K R_{\text{nui},v}^{[r]} \left\{ \frac{\log(q)}{n_r} \right\}^{1-v/2} \right\},$$

with probability at least $1 - c/q$. Define $S_{\delta,v} = \{j \in \{1,\dots,q\} : |\bar{\delta}_j| \ge \sqrt{\log(q)/n_0}\}$. Let $\tau_{\text{KTr}} = C_1\sqrt{\log(q)}(n_0^{-1/2} + \Delta) \ge C\{\sqrt{\log(q)}(n_0^{-1/2} + \Delta)\} + \sqrt{\log(q)/n_0}$. Following the proof of Theorem 4.3 in Battey et al. (2018), we can first show that $\|(\widehat{\boldsymbol{\beta}}_{\text{Deb}}^{[0]})_{S_{\delta,v}^c} - (\widehat{\boldsymbol{\beta}}_{\text{Thr}}^{[1]})_{S_{\delta,v}^c}\|_\infty \le \tau_{\text{KTr}}$, $\widehat{\boldsymbol{\delta}}_{S_{\delta,v}^c} = \mathbf{0}$ and consequently $\|\widehat{\boldsymbol{\delta}}_{S_{\delta,v}^c} - \bar{\boldsymbol{\delta}}_{S_{\delta,v}^c}\|_2 = \|\bar{\boldsymbol{\delta}}_{S_{\delta,v}^c}\|_2$. Since $\|\bar{\boldsymbol{\delta}}\|_v \le R_{\delta,v}$ and $R_{\delta,v} = O[\{n_0/\log(q)\}^{1-v/2}]$, we have

$$\|\bar{\boldsymbol{\delta}}_{S_{\delta,v}^c}\|_2^2 = \sum_{j\in S_{\delta,v}^c} |\bar{\delta}_j|^v |\bar{\delta}_j|^{2-v} = O\left[\left\{ \frac{\log(q)}{n_0} \right\}^{1-v/2} R_{\delta,v} \right],$$

$$\|\bar{\boldsymbol{\delta}}_{S_{\delta,v}^c}\|_1 = \sum_{j\in S_{\delta,v}^c} |\bar{\delta}_j|^v |\bar{\delta}_j|^{1-v} = O\left[\left\{ \frac{\log(q)}{n_0} \right\}^{(1-v)/2} R_{\delta,v} \right],$$

and

$$\|\widehat{\boldsymbol{\delta}}_{S_{\delta,v}^c} - \bar{\boldsymbol{\delta}}_{S_{\delta,v}^c}\|_\infty = \|\bar{\boldsymbol{\delta}}_{S_{\delta,v}^c}\|_\infty \le \sqrt{\log(q)/n_0}.$$

For $j \in S_{\delta,v}$, if $|\bar{\delta}_j| \geq 2\tau_{\text{KTr}}$, we have $|\widehat{\beta}^{[0]}_{\text{Deb},j} - \widehat{\beta}^{[1]}_{\text{Thr},j}| \geq |\bar{\delta}_j| - C\{\sqrt{\log(q)}(n_0^{-1/2} + \Delta)\} \geq \tau_{\text{KTr}}$. Hence, $|\widehat{\delta}_j - \bar{\delta}_j| = |\widehat{\beta}^{[0]}_{\text{Deb},j} - \widehat{\beta}^{[1]}_{\text{Thr},j} - \bar{\delta}_j| \leq \tau_{\text{KTr}}$. If instead $|\bar{\delta}_j| < 2\tau_{\text{KTr}}$, we have $|\widehat{\delta}_j - \bar{\delta}_j| \leq \max\{|\bar{\delta}_j|, |\widehat{\beta}^{[0]}_{\text{Deb},j} - \widehat{\beta}^{[1]}_{\text{Thr},j} - \bar{\delta}_j|\} \leq 2\tau_{\text{KTr}}$. Hence, we have

$$\|\widehat{\boldsymbol{\delta}}_{S_{\delta,v}} - \bar{\boldsymbol{\delta}}_{S_{\delta,v}}\|_2 \leq 2\sqrt{|S_{\delta,v}|}\tau_{\text{KTr}},$$

and

$$\|\widehat{\boldsymbol{\delta}}_{S_{\delta,v}} - \bar{\boldsymbol{\delta}}_{S_{\delta,v}}\|_1 \leq 2|S_{\delta,v}|\tau_{\text{KTr}}.$$

Following the proof of Corollary 3 in Negahban et al. (2012), we have

$$|S_{\delta,v}| \leq \left\{\sqrt{\frac{\log(q)}{n_0}}\right\}^{-v} R_{\delta,v}.$$

Hence, we obtain

$$\|\widehat{\boldsymbol{\delta}} - \bar{\boldsymbol{\delta}}\|_2 = O_p\left[\left\{\frac{\log(q)}{n_0}\right\}^{-v/4} R_{\delta,v}^{1/2}\left\{\sqrt{\frac{\log(q)}{n_0}} + \sqrt{\log(q)}\Delta\right\}\right],$$

$$\|\widehat{\boldsymbol{\delta}} - \bar{\boldsymbol{\delta}}\|_\infty = O_p\left\{\sqrt{\frac{\log(q)}{n_0}} + \sqrt{\log(q)}\Delta\right\},$$

and

$$\|\widehat{\boldsymbol{\delta}} - \bar{\boldsymbol{\delta}}\|_1 = O_p\left[\left\{\frac{\log(q)}{n_0}\right\}^{-v/2} R_{\delta,v}\left\{\sqrt{\frac{\log(q)}{n_0}} + \sqrt{\log(q)}\Delta\right\}\right].$$

■

S.1.8 Proof of Theorem 10

Proof By the triangle inequality, Theorem 10 follows directly from Theorem 6 and Lemma 9. ■

S.1.9 Proof of Lemma 12

Proof By Bernstein's inequality with a union bound over $j \in \{1, ..., q\}$, Lemma 12 directly follows from the proof of Lemma 4. ■

S.1.10 Proof of Theorem 13

Proof Note that by symmetry between the two folds, it suffices to establish the result for one fold-specific estimator; applying the same argument to the other fold and using the triangle inequality yields the same rate for their average.

First of all, we have

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2^2-\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2^2\\ =&2\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right)^{\mathsf{T}}\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\bar{\boldsymbol{\beta}}^{[0]}\right)+\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2^2-\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2^2.\end{aligned}$$

With binary selection, i.e., $a=\infty$ in (13), when $\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2^2-\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2^2\geq 0$, we have

$$\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2^2\leq 2\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right)^{\mathsf{T}}\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\bar{\boldsymbol{\beta}}^{[0]}\right)+\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2^2.$$

Similarly, when $\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2^2-\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2^2<0$, we have

$$\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2^2\leq -2\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right)^{\mathsf{T}}\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\bar{\boldsymbol{\beta}}^{[0]}\right)+\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2^2.$$

Hence,

$$\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{AIMS},(1)}-\bar{\boldsymbol{\beta}}^{[0]}\|_2\leq\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}\right\|_2\wedge\left\|\bar{\boldsymbol{\beta}}^{[0]}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right\|_2+\mathrm{DetectErr},\tag{S.23}$$

where

$$\mathrm{DetectErr}=\left\{2\left|\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\right)^{\mathsf{T}}\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Deb},(2)}-\bar{\boldsymbol{\beta}}^{[0]}\right)\right|\right\}^{1/2},$$

which is the stochastic price of transferability detection on the ℓ_2 scale.

By the proof of Lemma 4, we have

$$(\widehat{\beta}^{[0]}_{\mathrm{Deb},(2)})_j-\bar{\beta}^{[0]}_j=V_{0,j}+V_{1,j},$$

where

$$\begin{aligned}V_{0,j}=&-\mathbf{e}_j^{\mathsf{T}}\bar{\boldsymbol{\Omega}}^{[0]}\partial_{\boldsymbol{\beta}}\mathcal{L}_{0,(2)}(\bar{\boldsymbol{\beta}}^{[0]};\bar{\boldsymbol{\alpha}}^{[0]}_j,\bar{\boldsymbol{\gamma}}^{[0]}_j),\\ V_{1,j}=&r^{[0]}_{(2),j}=O_p(\Delta_0).\end{aligned}$$

Let $\mathbf{V}_k=(V_{k,1},\ldots,V_{k,q})^{\mathsf{T}}$, for $k=0,1$. Recall that $s_d=\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\|_0$. Then we have

$$(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)})^{\mathsf{T}}\mathbf{V}_0=O_p(s_d^{1/2}n_0^{-1/2}\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\|_2),$$

and

$$|(\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)})^{\mathsf{T}}\mathbf{V}_1|=O_p\{s_d^{1/2}\Delta_0\|\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{KTr},(1)}-\widehat{\boldsymbol{\beta}}^{[0]}_{\mathrm{Thr},(1)}\|_2\}.$$

Hence, we have

$$\text{DetectErr} = O_p[\{s_d^{1/2}(n_0^{-1/2} + \Delta_0)\|\widehat{\boldsymbol{\beta}}^{[0]}_{\text{KTr},(1)} - \widehat{\boldsymbol{\beta}}^{[0]}_{\text{Thr},(1)}\|_2\}^{1/2}]. \tag{S.24}$$

By Theorem 6 and Lemma 9, we also have

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}^{[0]}_{\text{KTr},(1)} - \widehat{\boldsymbol{\beta}}^{[0]}_{\text{Thr},(1)}\|_2 \\
=&O_p\left(\max_{r=0,1}\left[\sqrt{s_\beta^{[r]}\log(q)}\left\{n_r^{-1/2} + \Delta_r\right\}\right] + \left\{\frac{\log(q)}{n_0}\right\}^{-v/4} R_{\delta,v}^{1/2}\sqrt{\log(q)}\left\{n_0^{-1/2} + \Delta\right\}\right).
\end{aligned}$$

It remains to establish the corrected support bound regarding s_d. Choose the threshold constants sufficiently large so that, with probability tending to one,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}}^{[0]}_{\text{Deb},(1)} - \bar{\boldsymbol{\beta}}^{[0]}\right\|_\infty &\le \frac{1}{2}\tau^{[0]}, \\
\left\|\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Deb}} - \bar{\boldsymbol{\beta}}^{[1]}\right\|_\infty &\le \frac{1}{2}\tau^{[1]}, \\
\left\|(\widehat{\boldsymbol{\beta}}^{[0]}_{\text{Deb},(1)} - \widehat{\boldsymbol{\beta}}^{[1]}_{\text{Thr}}) - \bar{\delta}\right\|_\infty &\le \frac{1}{2}\tau_{\text{KTr}}.
\end{aligned}$$

On this event,

$$\text{supp}\left(\widehat{\boldsymbol{\beta}}^{[0]}_{\text{Thr},(1)}\right) \subseteq \text{supp}(\bar{\boldsymbol{\beta}}^{[0]}).$$

Moreover, if $j \notin \text{supp}(\bar{\boldsymbol{\beta}}^{[0]})$ and $\widehat{\beta}^{[1]}_{\text{Thr,j}} \neq 0$, then $|\bar{\delta}_j| = |\bar{\beta}^{[1]}_j| \ge \tau^{[1]}/2$. Likewise, if the thresholded contrast $\widehat{\delta}_{(k),j}$ is nonzero, then $|\bar{\delta}_j| \ge \tau_{\text{KTr}}/2$. Further, for every $t > 0$,

$$\left|\{j : |\bar{\delta}_j| \ge t\}\right| \le R_{\delta,v}t^{-v},$$

with the usual interpretation when $v = 0$. Counting the sets, with $n_1 \ge n_0$, we obtain

$$s_d = O_p\{s_\beta^{[0]} + R_{\delta,v}(\tau^{[1]} \wedge \tau_{\text{KTr}})^{-v}\}. \tag{S.25}$$

When $\Delta_r = O(n_r^{-1/2})$, we have

$$\text{DetectErr} = O_p\left[\sqrt{\frac{s_d}{n_0}}\max_{r\in\{0,1}\left\{\sqrt{\frac{s_\beta^{[r]}\log(q)}{n_r}}\right\} + \sqrt{\frac{s_d}{n_0}}\left\{\frac{\log(q)}{n_0}\right\}^{1/2-v/4} R_{\delta,v}^{1/2}\right]^{1/2}.$$

If we further set $v = 0$, we have

$$\text{DetectErr} = O_p\left[\sqrt{\frac{s_\beta^{[0]} + R_{\delta,0}}{n_0}}\max_{r\in\{0,1}\left\{\sqrt{\frac{s_\beta^{[r]}\log(q)}{n_r}}\right\} + \sqrt{\frac{s_\beta^{[0]} + R_{\delta,0}}{n_0}}\left\{\frac{\log(q)}{n_0}\right\}^{1/2} R_{\delta,0}^{1/2}\right]^{1/2}.$$

By (S.23), (S.24) and (S.25) we prove Theorem 13. ■

Appendix S.2. Additional Methodological Details

This section collects the methodological details of AIMS that are not expanded in the main text. We first give the nodewise precision-matrix estimator used in the debiasing step, then describe the weight-stratified calibration used when the calibration weights change sign, and finally summarize the tuning rules used by the algorithms.

S.2.1 Nodewise Precision-Matrix Estimator

The nodewise precision-matrix row $\widehat{\boldsymbol{\Omega}}_j^{[r]}$ is given by

$$\widehat{\boldsymbol{\Omega}}_j^{[r]} = \big(-\widehat{\boldsymbol{\theta}}_{j,1}^{[r]}, \ldots, -\widehat{\boldsymbol{\theta}}_{j,j-1}^{[r]}, 1, -\widehat{\boldsymbol{\theta}}_{j,j+1}^{[r]}, \ldots, -\widehat{\boldsymbol{\theta}}_{j,q}^{[r]} \big) / \widehat{\varsigma}_j^{[r]},$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_j^{[r]} =& \underset{\boldsymbol{\theta}\in\mathbb{R}^{q-1}}{\arg\min}\, \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} \dot{g}\big(\mathbf{X}^\intercal \widetilde{\boldsymbol{\beta}}^{[r]}\big)(X_j - \mathbf{X}_{\text{-}j}^\intercal \boldsymbol{\theta})^2 + 2\lambda_{\theta_j}^{[r]} \|\boldsymbol{\theta}\|_1, \\
\widehat{\varsigma}_j^{[r]} =& \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} \dot{g}\big(\mathbf{X}^\intercal \widetilde{\boldsymbol{\beta}}^{[r]}\big)\big(X_j - \mathbf{X}_{\text{-}j}^\intercal \widehat{\boldsymbol{\theta}}_j^{[r]}\big)^2 + \lambda_{\theta_j}^{[r]} \big\|\widehat{\boldsymbol{\theta}}_j^{[r]}\big\|_1,
\end{aligned}$$

and $\mathbf{X}_{\text{-}j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_q)^\intercal$. The tuning parameter $\lambda_{\theta_j}^{[r]}$ is chosen via five-fold cross-validation from the range $\big[0.1\sqrt{\frac{\log q}{n_{\mathcal{T},r}}}, 2\sqrt{\frac{\log q}{n_{\mathcal{T},r}}}\big]$.

S.2.2 Weight-Stratified Calibration

As mentioned in Remark 3, the weights $\widehat{w}_j^{[r]}$ may not always be nonnegative. For instance, if $\mathbf{X}$ has a mean-zero symmetric distribution (except for the intercept), the weights $\widehat{w}_j^{[r]}$ can have approximately equal probabilities of being positive or negative. As a result, the calibration loss may become irregular and ill-posed. To address this issue, we use a weight-stratified calibration strategy.

Instead of directly fitting (11), we define the following estimators:

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_{j+}^{[r]} &= \underset{\boldsymbol{\xi}\in\mathbb{R}^d}{\operatorname{argmin}} \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \mathbb{I}(\widehat{w}_j^{[r]} > 0) |\widehat{w}_j^{[r]}| \dot{b}(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \mathcal{F}^{[r]}(\boldsymbol{\xi}; \widetilde{\boldsymbol{\alpha}}^{[r]}) + \lambda_{\alpha_j+}^{[r]} \|\boldsymbol{\xi}\|_1 ; \\
\widehat{\boldsymbol{\xi}}_{j-}^{[r]} &= \underset{\boldsymbol{\xi}\in\mathbb{R}^d}{\operatorname{argmin}} \widehat{\mathbb{E}}_{\mathcal{S}\cup\mathcal{T}}^{[r]} \mathbb{I}(\widehat{w}_j^{[r]} \le 0) |\widehat{w}_j^{[r]}| \dot{b}(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \mathcal{F}^{[r]}(\boldsymbol{\xi}; \widetilde{\boldsymbol{\alpha}}^{[r]}) + \lambda_{\alpha_j-}^{[r]} \|\boldsymbol{\xi}\|_1 ; \\
\widehat{\boldsymbol{\zeta}}_{j+}^{[r]} &= \underset{\boldsymbol{\zeta}\in\mathbb{R}^d}{\operatorname{argmin}} \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbb{I}(\widehat{w}_j^{[r]} > 0) |\widehat{w}_j^{[r]}| \exp(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[r]}) \mathcal{G}(\boldsymbol{\zeta}; \widetilde{\boldsymbol{\gamma}}^{[r]}) + \lambda_{\gamma_j+}^{[r]} \|\boldsymbol{\zeta}\|_1 ; \\
\widehat{\boldsymbol{\zeta}}_{j-}^{[r]} &= \underset{\boldsymbol{\zeta}\in\mathbb{R}^d}{\operatorname{argmin}} \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \mathbb{I}(\widehat{w}_j^{[r]} \le 0) |\widehat{w}_j^{[r]}| \exp(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[r]}) \mathcal{G}(\boldsymbol{\zeta}; \widetilde{\boldsymbol{\gamma}}^{[r]}) + \lambda_{\gamma_j-}^{[r]} \|\boldsymbol{\zeta}\|_1.
\end{aligned} \tag{S.26}$$

We now compute $\widehat{\boldsymbol{\alpha}}_{j+}^{[r]} = \widetilde{\boldsymbol{\alpha}}^{[r]} + \widehat{\boldsymbol{\xi}}_{j+}^{[r]}$, $\widehat{\boldsymbol{\alpha}}_{j-}^{[r]} = \widetilde{\boldsymbol{\alpha}}^{[r]} + \widehat{\boldsymbol{\xi}}_{j-}^{[r]}$, $\widehat{\boldsymbol{\gamma}}_{j+}^{[r]} = \widetilde{\boldsymbol{\gamma}}^{[r]} + \widehat{\boldsymbol{\zeta}}_{j+}^{[r]}$, and $\widehat{\boldsymbol{\gamma}}_{j-}^{[r]} = \widetilde{\boldsymbol{\gamma}}^{[r]} + \widehat{\boldsymbol{\zeta}}_{j-}^{[r]}$. Using these, we derive the one-step debiased estimator $\widehat{\beta}_{\mathrm{Deb},j}^{[r]}$ as:

$$\widehat{\beta}_{\mathrm{Deb},j}^{[r]} = \mathbf{e}_j^\intercal \widetilde{\boldsymbol{\beta}}^{[r]} + \mathbf{e}_j^\intercal \widehat{\boldsymbol{\Omega}}^{[r]} \Big[\widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \widehat{h}^j(\boldsymbol{\Phi}) \mathbf{X} \{ Y - \widehat{r}^j(\boldsymbol{\Phi}) \} + \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} \mathbf{X} \{ \widehat{r}^j(\boldsymbol{\Phi}) - g(\mathbf{X}^\intercal \widetilde{\boldsymbol{\beta}}^{[r]}) \} \Big].$$

where $\widehat{h}^j(\boldsymbol{\Phi}) = \mathbb{I}(\widehat{w}_j^{[r]} > 0) \exp\{\boldsymbol{\Phi}^\intercal \widehat{\boldsymbol{\alpha}}_{j+}^{[r]}\} + \mathbb{I}(\widehat{w}_j^{[r]} \le 0) \exp\{\boldsymbol{\Phi}^\intercal \widehat{\boldsymbol{\alpha}}_{j-}^{[r]}\}$ and $\widehat{r}^j(\boldsymbol{\Phi}) = \mathbb{I}(\widehat{w}_j^{[r]} > 0) b(\boldsymbol{\Phi}^\intercal \widehat{\boldsymbol{\gamma}}_{j+}^{[r]}) + \mathbb{I}(\widehat{w}_j^{[r]} \le 0) b(\boldsymbol{\Phi}^\intercal \widehat{\boldsymbol{\gamma}}_{j-}^{[r]})$.

S.2.3 Tuning Rules for AIMS

We summarize the tuning parameters used in AIMS and describe their practical selection rules. The tuning parameters fall into three groups: regularization parameters for sparse nuisance and target estimation, bootstrap-calibrated thresholds for the bias-correction and transfer steps, and the aggregation temperature used for negative transfer protection. All tuning procedures are label-free with respect to the target population.

Parameter	Role	Theoretical order	Practical choice
$\lambda^{[r]}$	ℓ_1 penalty for the preliminary target estimator $\widetilde{\boldsymbol{\beta}}^{[r]}$	$\sqrt{\log(q)/n_r}$	Selected by five-fold cross-validation on a grid centered at the theoretical rate.
$\lambda_\alpha^{[r]}$	ℓ_1 penalty for the preliminary density ratio model	$\sqrt{\log(d)/n_r}$	Selected by five-fold cross-validation for the density ratio loss in (7).
$\lambda_\gamma^{[r]}$	ℓ_1 penalty for the preliminary conditional mean imputation model	$\sqrt{\log(d)/n_{\mathcal{S},r}}$	Selected by five-fold cross-validation using the labeled source samples in subgroup r.
$\lambda_{\theta_j}^{[r]}$	ℓ_1 penalty for the nodewise precision-matrix estimator	$\sqrt{\log(q)/n_{\mathcal{T},r}}$	Selected by five-fold cross-validation on the target covariates.
$\lambda_{\alpha_j}^{[r]}$	ℓ_1 penalty for the calibrated density ratio nuisance correction for coordinate j	Same order as $\lambda_\alpha^{[r]}$	Chosen by Gaussian multiplier bootstrap using quantile level q_τ.
$\lambda_{\gamma_j}^{[r]}$	ℓ_1 penalty for the calibrated imputation nuisance correction for coordinate j	Same order as $\lambda_\gamma^{[r]}$	Chosen by Gaussian multiplier bootstrap using quantile level q_τ.
$\tau^{[r]}$	Threshold for sparsifying the dense debiased estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb}}^{[r]}$	$\sqrt{\log(q)}\{n_r^{-1/2}+\Delta_r\}$	Implemented as $\tau^{[r]} = C_\tau\{\log(q)/n_{\mathcal{S},r}\}^{1/2}$.
τ_{KTr}	Threshold for estimating the sparse majority-minority difference in Algorithm 2	$\sqrt{\log(q)}\{n_0^{-1/2}+\Delta\}$	Chosen by Gaussian multiplier bootstrap using quantile level q_τ.
a	Temperature parameter in the exponential aggregation weight in Algorithm 3	$a = \infty$ in Theorem 13	Fixed at a positive value in the main comparison; $a = \infty$ gives hard selection and is used in the theory.

Table S.1: Summary of tuning parameters in AIMS. Here $n_r = n_{\mathcal{S},r} \wedge n_{\mathcal{T},r}$, d is the dimension of the nuisance feature map $\phi(\mathbf{Z})$, and Δ_r is the higher-order nuisance-estimation error term defined in Section 4.

For the bootstrap-calibrated quantities, we generate independent standard normal multipliers ϵ_i and compute

$$\mathcal{M}_{\alpha_j}^{[r]}(\epsilon) = \left\| \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \epsilon \widehat{w}_j^{[r]} \exp(\mathbf{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[r]}) \dot{b}(\mathbf{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \mathbf{\Phi} - \widehat{\mathbb{E}}_{\mathcal{T}}^{[r]} \epsilon \widehat{w}_j^{[r]} \dot{b}(\mathbf{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \mathbf{\Phi} \right\|_\infty ;$$

$$\mathcal{M}_{\gamma_j}^{[r]}(\epsilon) = \left\| \widehat{\mathbb{E}}_{\mathcal{S}}^{[r]} \epsilon \widehat{w}_j^{[r]} \exp(\mathbf{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[r]}) \mathbf{\Phi} \{ Y - b(\mathbf{\Phi}^\intercal \widetilde{\boldsymbol{\gamma}}^{[r]}) \} \right\|_\infty ;$$

$$\mathcal{M}_\kappa(\epsilon) = \max_{j=1,\ldots,q} \Bigg| \mathbf{e}_j^\intercal \widehat{\mathbf{\Omega}}^{[0]} \left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[0]} \epsilon \widehat{h}^j(\mathbf{\Phi}) \mathbf{X} \{ Y - \widehat{r}^j(\mathbf{\Phi}) \} + \widehat{\mathbb{E}}_{\mathcal{T}}^{[0]} \epsilon \mathbf{X} \{ \widehat{r}^j(\mathbf{\Phi}) - g(\mathbf{X}^\intercal \widetilde{\boldsymbol{\beta}}^{[0]}) \} \right]$$
$$- \mathbf{e}_j^\intercal \widehat{\mathbf{\Omega}}^{[1]} \left[\widehat{\mathbb{E}}_{\mathcal{S}}^{[1]} \epsilon \widehat{h}^j(\mathbf{\Phi}) \mathbf{X} \{ Y - \widehat{r}^j(\mathbf{\Phi}) \} + \widehat{\mathbb{E}}_{\mathcal{T}}^{[1]} \epsilon \mathbf{X} \{ \widehat{r}^j(\mathbf{\Phi}) - g(\mathbf{X}^\intercal \widetilde{\boldsymbol{\beta}}^{[1]}) \} \right] \Bigg|.$$

With $B = 500$ multiplier draws, for the non-weight-stratified estimators, we set $\lambda_{\alpha_j}^{[r]}$ and $\lambda_{\gamma_j}^{[r]}$ to the empirical q_τ quantiles of $\mathcal{M}_{\alpha_j}^{[r]}$ and $\mathcal{M}_{\gamma_j}^{[r]}$, respectively. For the weight-stratified estimators in (S.26), the same multiplier-bootstrap calculations are performed separately on the subsets with $\widehat{w}_j^{[r]} > 0$ and $\widehat{w}_j^{[r]} \leq 0$, using $|\widehat{w}_j^{[r]}|$ within each subset; the resulting penalties are denoted by $\lambda_{\alpha_j+}^{[r]}$, $\lambda_{\alpha_j-}^{[r]}$, $\lambda_{\gamma_j+}^{[r]}$, and $\lambda_{\gamma_j-}^{[r]}$. We set τ_{KTr} to the empirical q_τ quantile of $\mathcal{M}_\kappa$. In the main simulation comparison, we use the weight-stratified version with $C_\tau = 2$, $q_\tau = 0.8$, and $a = 5$. Supplementary S.3.3 reports sensitivity to q_τ and C_τ, as well as aggregation-temperature variants with $a \in \{1, 2, 5, 10, \infty\}$.

Appendix S.3. Benchmark Implementations and Additional Results

This section contains the benchmark algorithms used in the numerical comparisons and the supplementary result tables that are omitted from the main text for readability. We first describe the benchmark implementations, then give the additional simulation tables, and finally report the complete T2D benchmark table and supplementary $\Delta\Delta G$ results.

S.3.1 Benchmark Methods

In this section, we give the details of the benchmark methods. The importance weighting methods with ℓ_1 penalty (IW) and with weighted ℓ_1 penalty ($\mathrm{IW}_{\mathrm{aLasso}}$) are presented in Algorithm S.1. The imputation-only methods with ℓ_1 penalty (IM) and with weighted ℓ_1 penalty ($\mathrm{IM}_{\mathrm{aLasso}}$) are given in Algorithm S.2. The TransGLM method with two variants according to our setup is presented in Algorithm S.3. The tuning parameters λ for these benchmark methods are selected by five-fold cross-validation.

For CORAL, we use only the labeled minority-source sample and the unlabeled minority-target covariates. Let $\widehat{\Sigma}_{\mathcal{S},0}$ and $\widehat{\Sigma}_{\mathcal{T},0}$ be the empirical covariance matrices of the non-intercept target-model covariates in the minority source and target samples, respectively. After centering the minority-source covariates, CORAL applies the transformation $(\widehat{\Sigma}_{\mathcal{S},0} + \lambda_{\mathrm{coral}} I)^{-1/2} (\widehat{\Sigma}_{\mathcal{T},0} + \lambda_{\mathrm{coral}} I)^{1/2}$, with $\lambda_{\mathrm{coral}} = 1$, and then fits an ℓ_1-penalized logistic target model using the aligned labeled minority-source sample. The intercept is expressed on the target-centered covariate scale.

For TransFusion, we use a source-labeled adaptation because labeled target outcomes are unavailable. The labeled minority-source sample is used as the pseudo-target task,

and the labeled majority-source sample is used as the auxiliary task. We fit the TransFusion block-design penalized logistic regression of He et al. (2024), which parameterizes each auxiliary-task coefficient as a shared pseudo-target coefficient plus a sparse task contrast. The reported TransFusion estimator is the resulting task-averaged coefficient with its intercept calibrated on the labeled minority-source sample.

Algorithm S.1 Importance Weighting with ℓ_1 penalty (IW) and with weighted ℓ_1 penalty ($\text{IW}_{\text{aLasso}}$).

Input: $\mathscr{D}^{[0]} = \{\mathbf{D}_i = (S_iY_i, \mathbf{X}_i^\intercal, \mathbf{W}_i^\intercal, S_i, R_i)^\intercal : R_i = 0,\ i \in [n]\}$ and the preliminary estimator $\widetilde{\boldsymbol{\alpha}}^{[0]}$.

IW: Obtain $\widehat{\boldsymbol{\beta}}_{\text{iw}}$ as

$$\widehat{\boldsymbol{\beta}}_{\text{iw}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\, \widehat{\mathbb{E}}_{\mathcal{S}}^{[0]} \exp(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[0]})\{G(\mathbf{X}^\intercal\boldsymbol{\beta}) - Y\mathbf{X}^\intercal\boldsymbol{\beta}\} + \lambda\|\boldsymbol{\beta}\|_1\,.$$

$\text{IW}_{\text{aLasso}}$: Obtain $\widehat{\boldsymbol{\beta}}_{\text{adaiw}}$ as

$$\widehat{\boldsymbol{\beta}}_{\text{adaiw}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\, \widehat{\mathbb{E}}_{\mathcal{S}}^{[0]} \exp(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[0]})\{G(\mathbf{X}^\intercal\boldsymbol{\beta}) - Y\mathbf{X}^\intercal\boldsymbol{\beta}\} + \lambda\sum_{j=1}^{q} b_j|\beta_j|\,,$$

where $b_j = 1/|\hat{\beta}_{\text{r},j}|$ with

$$(\hat{\beta}_{\text{r},1}, \ldots, \hat{\beta}_{\text{r},q})^\intercal = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\, \widehat{\mathbb{E}}_{\mathcal{S}}^{[0]} \exp(\boldsymbol{\Phi}^\intercal \widetilde{\boldsymbol{\alpha}}^{[0]})\{G(\mathbf{X}^\intercal\boldsymbol{\beta}) - Y\mathbf{X}^\intercal\boldsymbol{\beta}\} + n^{-2/3}\|\boldsymbol{\beta}\|_2^2\,.$$

Output: $\widehat{\boldsymbol{\beta}}_{\text{iw}}$ and $\widehat{\boldsymbol{\beta}}_{\text{adaiw}}$.

Algorithm S.2 Imputation-only method with ℓ_1 penalty (IM) and with weighted ℓ_1 penalty ($\text{IM}_{\text{aLasso}}$)

Input: $\mathscr{D}^{[0]} = \{\mathbf{D}_i = (S_iY_i, \mathbf{X}_i^\intercal, \mathbf{W}_i^\intercal, S_i, R_i)^\intercal : R_i = 0,\ i \in [n]\}$ and the preliminary estimator $\widetilde{\boldsymbol{\gamma}}^{[0]}$.

IM: Obtain $\widehat{\boldsymbol{\beta}}_{\text{im}}$ as

$$\widehat{\boldsymbol{\beta}}_{\text{im}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\, \widehat{\mathbb{E}}_{\mathcal{T}}^{[0]}\{G(\mathbf{X}^\intercal\boldsymbol{\beta}) - b(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[0]})\mathbf{X}^\intercal\boldsymbol{\beta}\} + \lambda\|\boldsymbol{\beta}\|_1\,.$$

$\text{IM}_{\text{aLasso}}$: Obtain $\widehat{\boldsymbol{\beta}}_{\text{adaim}}$ as

$$\widehat{\boldsymbol{\beta}}_{\text{adaim}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\, \widehat{\mathbb{E}}_{\mathcal{T}}^{[0]}\{G(\mathbf{X}^\intercal\boldsymbol{\beta}) - b(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[0]})\mathbf{X}^\intercal\boldsymbol{\beta}\} + \lambda\sum_{j=1}^{q} b_j|\beta_j|\,.$$

where $b_j = 1/|\hat{\beta}_{\text{r},j}|$ with

$$(\hat{\beta}_{\text{r},1}, \ldots, \hat{\beta}_{\text{r},q})^\intercal = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\, \widehat{\mathbb{E}}_{\mathcal{T}}^{[0]}\{G(\mathbf{X}^\intercal\boldsymbol{\beta}) - b(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\gamma}}^{[0]})\mathbf{X}^\intercal\boldsymbol{\beta}\} + n^{-2/3}\|\boldsymbol{\beta}\|_2^2\,.$$

Output: $\widehat{\boldsymbol{\beta}}_{\text{im}}$ and $\widehat{\boldsymbol{\beta}}_{\text{adaim}}$.

Algorithm S.3 TransGLM and $\text{TransGLM}_{\text{iw}}$

Input: $\mathscr{D}^{[1,1]} = \{\mathbf{D}_i = (S_iY_i, \mathbf{X}_i^\intercal, \mathbf{W}_i^\intercal, S_i, R_i)^\intercal : S_i = 1, R_i = 1,\ i \in [n]\}$, $\mathscr{D}^{[1,0]} = \{\mathbf{D}_i = (S_iY_i, \mathbf{X}_i^\intercal, \mathbf{W}_i^\intercal, S_i, R_i)^\intercal : S_i = 1, R_i = 0,\ i \in [n]\}$ and the preliminary estimators $\widetilde{\boldsymbol{\alpha}}^{[r]}, r = 0, 1$.

TransGLM: Obtain $\widehat{\boldsymbol{\beta}}_{\text{trans}}$ by applying the TransGLM algorithm (Tian and Feng, 2023) using $\mathscr{D}^{[1,0]}$ as the target sample and $\mathscr{D}^{[1,1]}$ as the source sample.

$\text{TransGLM}_{\text{iw}}$: Obtain $\widehat{\boldsymbol{\beta}}_{\text{trans,iw}}$ by applying the TransGLM algorithm (Tian and Feng, 2023) with the following modifications:

- Use $\mathscr{D}^{[1,0]}$ as the target sample, with each observation weighted by $\exp(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\alpha}}^{[0]})$.
- Use $\mathscr{D}^{[1,1]}$ as the source sample, with each observation weighted by $\exp(\boldsymbol{\Phi}^\intercal\widetilde{\boldsymbol{\alpha}}^{[1]})$.

Output: $\widehat{\boldsymbol{\beta}}_{\text{trans}}$ and $\widehat{\boldsymbol{\beta}}_{\text{trans,iw}}$.

S.3.2 Exact Data-Generating Mechanisms for the Simulation Studies

The exact data-generating mechanisms used in Section 5 are as follows. Let $\mathbf{Z} = (Z_1, \ldots, Z_{p+q})^\intercal$, where $Z_1 = 1$ and Z_j, $j \geq 2$, are independently generated from a standard normal distribution truncated to $(-1.5, 1.5)$. We set $\mathbf{X} = (Z_1, \ldots, Z_q)^\intercal$, $\mathbf{W} = (Z_{q+1}, \ldots, Z_{q+p})^\intercal$, $\phi(\mathbf{Z}) = \mathbf{Z}$, and use the logistic link $g(a) = b(a) = \text{expit}(a)$. For the minority group, define

$$\boldsymbol{\gamma}^{[0]} = (0,\ 0.8,\ -0.6,\ 0.5,\ 0.4,\ \mathbf{0}_{q-5}^\intercal,\ 0.70,\ -0.70,\ 0.50,\ \mathbf{0}_{p-3}^\intercal)^\intercal,$$

and

$$\boldsymbol{\alpha}^{[0]} = (0,\ 0.55,\ -0.55,\ 0.4125, \mathbf{0}_{q-4}^{\mathsf{T}}, 1.03125,\ -0.825,\ 0.61875, \mathbf{0}_{p-3}^{\mathsf{T}})^{\mathsf{T}}.$$

The majority outcome coefficient is fixed as a local perturbation of the minority coefficient:

$$\boldsymbol{\gamma}^{[1]} = \boldsymbol{\gamma}^{[0]} + (0,\ 0.10\mathbf{1}_4^{\mathsf{T}},\ \mathbf{0}_{p+q-5}^{\mathsf{T}})^{\mathsf{T}}.$$

The majority source–target sampling coefficient is also fixed:

$$\boldsymbol{\alpha}^{[1]} = \boldsymbol{\alpha}^{[0]} + (0,\ 0.01,\ -0.01,\ 0.0075,\ \mathbf{0}_{q-4}^{\mathsf{T}},\ 0.045,\ -0.0375,\ 0.03,\ \mathbf{0}_{p-3}^{\mathsf{T}})^{\mathsf{T}}.$$

To introduce controlled nonlinear misspecification, let

$$u_Y(\mathbf{Z}) = 0.50 Z_2 Z_{q+1} + 0.25(Z_3^2 - \mu_2),$$

and

$$u_S(\mathbf{Z}) = 0.35 Z_2 Z_{q+1} - 0.175 Z_3 Z_{q+2} + 0.175(Z_4^2 - \mu_2),$$

where μ_2 is the second moment of the truncated normal covariate, approximated by the empirical second moment in each generated sample. For $r \in \{0, 1\}$, we generate the outcome Y in the following three settings:

$$\begin{aligned}
\text{Setting I:}\quad & \mathbb{P}(Y = 1 \mid \mathbf{Z}, R = r) = \operatorname{expit}\{\mathbf{Z}^{\mathsf{T}}\boldsymbol{\gamma}^{[r]}\},\\
& \mathbb{P}(S = 1 \mid \mathbf{Z}, R = r) = \operatorname{expit}\{\mathbf{Z}^{\mathsf{T}}\boldsymbol{\alpha}^{[r]}\};\\
\text{Setting II:}\quad & \mathbb{P}(Y = 1 \mid \mathbf{Z}, R = r) = \operatorname{expit}\{\mathbf{Z}^{\mathsf{T}}\boldsymbol{\gamma}^{[r]} + u_Y(\mathbf{Z})\},\\
& \mathbb{P}(S = 1 \mid \mathbf{Z}, R = r) = \operatorname{expit}\{\mathbf{Z}^{\mathsf{T}}\boldsymbol{\alpha}^{[r]}\};\\
\text{Setting III:}\quad & \mathbb{P}(Y = 1 \mid \mathbf{Z}, R = r) = \operatorname{expit}\{\mathbf{Z}^{\mathsf{T}}\boldsymbol{\gamma}^{[r]}\},\\
& \mathbb{P}(S = 1 \mid \mathbf{Z}, R = r) = \operatorname{expit}\{\mathbf{Z}^{\mathsf{T}}\boldsymbol{\alpha}^{[r]} + u_S(\mathbf{Z})\}.
\end{aligned}$$

S.3.3 Complete Binary-Outcome Simulation Results

This subsection provides additional binary-outcome simulation results that complement the main figures in Section 5. The first two tables give the complete benchmark comparisons for the sample-size and dimensionality grids, including methods omitted from the main figures for readability. The remaining tables vary the amount of minority target-unlabeled covariate data, examine AIMS through ablation and tuning-sensitivity analyses, and report the majority-side estimator used as auxiliary information. Unless otherwise stated, the tables use the simulation design in Section 5, with $q = 100$ and default AIMS tuning constants $C_\tau = 2$, $q_\tau = 0.8$, and $a = 5$, except in the tables where $n_{\mathcal{T},0}$, q_τ, or C_τ is explicitly varied. The main-grid and supplementary ablation and sensitivity summaries were generated in separate Monte Carlo batches under the same nominal reference design; hence, repeated reference settings may have slightly different Monte Carlo summaries.

Table S.2 expands Figure 2 to the full benchmark set along the sample-size axis. The ranking is consistent with the figure: AIMS has the smallest error in every reported cell, and $\text{TransGLM}_{\text{iw}}$ is typically the strongest external competitor. The importance-weighting-only and imputation-only estimators improve as $n_{\mathcal{S},0}$ increases, but they remain less accurate than AIMS in most cells, underscoring the value of combining weighting, imputation, and majority-guided transfer information.

Method	$n_{\mathcal{S},0}$ = 300	400	500	600
	Setting I			
AIMS	**0.315 (0.177)**	**0.259 (0.173)**	**0.218 (0.152)**	**0.255 (0.157)**
IW	0.785 (0.235)	0.665 (0.180)	0.618 (0.196)	0.574 (0.205)
$\text{IW}_{\text{aLasso}}$	1.740 (1.428)	1.426 (1.205)	1.101 (0.830)	0.815 (0.411)
IM	0.977 (0.224)	0.819 (0.155)	0.741 (0.188)	0.697 (0.180)
$\text{IM}_{\text{aLasso}}$	0.726 (0.249)	0.521 (0.157)	0.421 (0.164)	0.382 (0.148)
IM_{RF}	1.577 (0.165)	1.561 (0.143)	1.553 (0.137)	1.539 (0.119)
IM_{XGB}	1.802 (0.510)	1.449 (0.332)	1.295 (0.359)	1.183 (0.315)
CORAL	1.678 (0.271)	1.622 (0.269)	1.572 (0.221)	1.540 (0.210)
TransFusion	0.867 (0.210)	0.849 (0.188)	0.835 (0.174)	0.826 (0.158)
TransGLM	0.986 (0.219)	0.956 (0.197)	0.938 (0.177)	0.922 (0.165)
$\text{TransGLM}_{\text{iw}}$	0.428 (0.235)	0.366 (0.173)	0.359 (0.159)	0.334 (0.145)
	Setting II			
Method	$n_{\mathcal{S},0}$ = 300	400	500	600
AIMS	**0.303 (0.177)**	**0.267 (0.166)**	**0.228 (0.154)**	**0.228 (0.137)**
IW	0.731 (0.236)	0.653 (0.212)	0.583 (0.177)	0.533 (0.177)
$\text{IW}_{\text{aLasso}}$	1.557 (1.264)	1.215 (0.746)	1.010 (0.582)	0.775 (0.515)
IM	0.832 (0.221)	0.680 (0.164)	0.599 (0.142)	0.565 (0.162)
$\text{IM}_{\text{aLasso}}$	0.644 (0.268)	0.489 (0.174)	0.356 (0.144)	0.321 (0.124)
IM_{RF}	1.406 (0.176)	1.396 (0.148)	1.382 (0.147)	1.366 (0.126)
IM_{XGB}	1.477 (0.498)	1.191 (0.343)	1.047 (0.293)	0.933 (0.262)
CORAL	1.585 (0.320)	1.561 (0.287)	1.505 (0.279)	1.479 (0.235)
TransFusion	0.831 (0.222)	0.817 (0.189)	0.801 (0.191)	0.789 (0.158)
TransGLM	0.931 (0.242)	0.906 (0.214)	0.869 (0.196)	0.863 (0.171)
$\text{TransGLM}_{\text{iw}}$	0.412 (0.225)	0.376 (0.185)	0.337 (0.165)	0.314 (0.154)
	Setting III			
Method	$n_{\mathcal{S},0}$ = 300	400	500	600
AIMS	**0.314 (0.190)**	**0.288 (0.158)**	**0.269 (0.158)**	**0.258 (0.143)**
IW	0.706 (0.266)	0.607 (0.207)	0.581 (0.213)	0.522 (0.186)
$\text{IW}_{\text{aLasso}}$	1.663 (1.176)	1.204 (0.800)	0.948 (0.565)	0.839 (0.490)
IM	0.923 (0.248)	0.793 (0.193)	0.714 (0.177)	0.603 (0.173)
$\text{IM}_{\text{aLasso}}$	0.711 (0.295)	0.484 (0.194)	0.403 (0.136)	0.323 (0.140)
IM_{RF}	1.522 (0.172)	1.498 (0.163)	1.510 (0.140)	1.472 (0.112)
IM_{XGB}	1.674 (0.565)	1.353 (0.418)	1.234 (0.324)	1.033 (0.264)
CORAL	1.643 (0.305)	1.546 (0.291)	1.528 (0.249)	1.463 (0.221)
TransFusion	0.847 (0.217)	0.825 (0.208)	0.835 (0.180)	0.797 (0.147)
TransGLM	0.977 (0.258)	0.928 (0.210)	0.930 (0.190)	0.884 (0.159)
$\text{TransGLM}_{\text{iw}}$	0.414 (0.228)	0.382 (0.241)	0.330 (0.123)	0.347 (0.131)

Table S.2: Binary-outcome sample-size grid. The columns vary the labeled minority-source sample size $n_{\mathcal{S},0} \in \{300, 400, 500, 600\}$, while $p = 400$, $q = 100$, $n_{\mathcal{S},1} = 2000$, $n_{\mathcal{T},0} = 2000$, and $n_{\mathcal{T},1} = 3000$ are fixed. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications; boldface indicates the smallest Monte Carlo mean within each setting and column.

Table S.3 gives the corresponding full benchmark comparison along the dimensionality axis in Figure 3. Several benchmarks are less competitive at larger p: the random-forest and boosting imputation variants have noticeably larger errors, and TransGLM without importance weighting is dominated by its weighted version. AIMS remains stable across dimensions and gives the smallest error throughout the grid.

Table S.4 varies the minority target-unlabeled sample size while holding $n_{\mathcal{S},0} = 400$ and $p = 400$ fixed. This comparison separates the role of unlabeled target covariates from the labeled-minority sample-size and dimensionality effects considered above. The ordering is stable over $n_{\mathcal{T},0} \in \{1000, 2000, 3000\}$: AIMS is the leading method in all three settings, $\text{TransGLM}_{\text{iw}}$ is the closest external benchmark, and TransFusion is less accurate under this design.

Table S.5 compares the intermediate AIMS estimators and aggregation variants at the reference design point. The rows $\widetilde{\boldsymbol{\beta}}^{[0]}$ and $\widehat{\boldsymbol{\beta}}^{[0]}_{\text{Deb}}$ are the preliminary minority estimator and the dense debiased estimator from the covariate shift correction step. $\text{AIMS}_{\text{min-o}}$ is the minority-only thresholded estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\text{Thr}}$ from Algorithm 1, $\text{AIMS}_{\text{maj-g}}$ is the majority-guided transfer estimator $\widehat{\boldsymbol{\beta}}^{[0]}_{\text{KTr}}$ from Algorithm 2, and the AIMS rows are the final aggregated estimators $\widehat{\boldsymbol{\beta}}^{[0]}_{\text{AIMS}}$ from Algorithm 3, with the row labeled AIMS using the default $a = 5$. The first three rows have larger errors than the majority-guided and aggregated estimators, showing the benefit of incorporating majority-group information in this local-transfer design. Across the finite temperatures $a \in \{1, 2, 5, 10\}$, aggregation is stable; the hard-selection limit $a = \infty$ is slightly less stable.

Table S.6 reports the corresponding majority-side estimators along the dimensionality axis, with errors measured relative to $\bar{\boldsymbol{\beta}}^{[1]}$. The rows $\widetilde{\boldsymbol{\beta}}^{[1]}$, $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Deb}}$, and $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Thr}}$ are the preliminary, dense debiased, and thresholded majority estimators from the covariate shift correction step. The dense debiased vector has larger squared ℓ_2 error because it is not sparsified, while the preliminary and thresholded sparse estimators are of the same order. The transfer step uses $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Thr}}$ as the offset in Algorithm 2; its error remains stable as p varies.

Tables S.7 and S.8 summarize tuning sensitivity. Varying the bootstrap quantile q_τ has only a modest effect over the range 0.6–0.9, indicating that the calibration is not driven by a single quantile choice. The threshold constant C_τ has a more visible bias-variance trade-off: very large thresholds are conservative and inflate the error, while moderate thresholds perform best in this reference design.

Method	p = 50	100	200	400	800
Setting I					
AIMS	**0.286 (0.161)**	**0.283 (0.185)**	**0.285 (0.160)**	**0.259 (0.173)**	**0.257 (0.186)**
IW	0.697 (0.270)	0.672 (0.240)	0.629 (0.217)	0.665 (0.180)	0.691 (0.243)
$\text{IW}_{\text{aLasso}}$	1.457 (1.142)	1.188 (0.715)	1.138 (0.862)	1.426 (1.205)	1.241 (0.983)
IM	0.724 (0.202)	0.737 (0.211)	0.752 (0.183)	0.819 (0.155)	0.923 (0.192)
$\text{IM}_{\text{aLasso}}$	0.486 (0.187)	0.471 (0.204)	0.473 (0.152)	0.521 (0.157)	0.639 (0.209)
IM_{RF}	1.510 (0.148)	1.472 (0.148)	1.490 (0.126)	1.561 (0.143)	1.598 (0.146)
IM_{XGB}	1.094 (0.324)	1.154 (0.368)	1.243 (0.287)	1.449 (0.332)	1.742 (0.417)
CORAL	1.646 (0.280)	1.574 (0.271)	1.569 (0.257)	1.622 (0.269)	1.595 (0.256)
TransFusion	0.860 (0.185)	0.826 (0.189)	0.813 (0.165)	0.849 (0.188)	0.816 (0.192)
TransGLM	0.987 (0.212)	0.941 (0.213)	0.922 (0.185)	0.956 (0.197)	0.934 (0.204)
$\text{TransGLM}_{\text{iw}}$	0.387 (0.147)	0.357 (0.184)	0.355 (0.161)	0.366 (0.173)	0.376 (0.181)
Setting II					
AIMS	**0.303 (0.144)**	**0.299 (0.166)**	**0.263 (0.145)**	**0.267 (0.166)**	**0.228 (0.153)**
IW	0.563 (0.172)	0.645 (0.470)	0.569 (0.188)	0.653 (0.212)	0.602 (0.211)
$\text{IW}_{\text{aLasso}}$	1.050 (0.693)	1.177 (0.726)	1.100 (0.849)	1.215 (0.746)	1.172 (0.711)
IM	0.559 (0.152)	0.609 (0.174)	0.599 (0.161)	0.680 (0.164)	0.728 (0.174)
$\text{IM}_{\text{aLasso}}$	0.395 (0.145)	0.427 (0.167)	0.410 (0.155)	0.489 (0.174)	0.556 (0.203)
IM_{RF}	1.261 (0.128)	1.310 (0.145)	1.317 (0.132)	1.396 (0.148)	1.412 (0.167)
IM_{XGB}	0.826 (0.235)	0.941 (0.288)	1.007 (0.302)	1.191 (0.343)	1.435 (0.430)
CORAL	1.531 (0.262)	1.531 (0.260)	1.475 (0.256)	1.561 (0.287)	1.540 (0.311)
TransFusion	0.800 (0.160)	0.832 (0.197)	0.775 (0.166)	0.817 (0.189)	0.815 (0.221)
TransGLM	0.871 (0.166)	0.902 (0.210)	0.844 (0.198)	0.906 (0.214)	0.887 (0.228)
$\text{TransGLM}_{\text{iw}}$	0.353 (0.186)	0.369 (0.176)	0.331 (0.151)	0.376 (0.185)	0.347 (0.173)
Setting III					
AIMS	**0.249 (0.173)**	**0.265 (0.158)**	**0.243 (0.151)**	**0.288 (0.158)**	**0.254 (0.159)**
IW	0.627 (0.227)	0.607 (0.216)	0.651 (0.218)	0.607 (0.207)	0.665 (0.258)
$\text{IW}_{\text{aLasso}}$	1.287 (0.797)	1.340 (0.926)	1.266 (0.936)	1.204 (0.800)	1.312 (1.119)
IM	0.625 (0.202)	0.681 (0.181)	0.731 (0.197)	0.793 (0.193)	0.863 (0.198)
$\text{IM}_{\text{aLasso}}$	0.440 (0.182)	0.424 (0.180)	0.471 (0.166)	0.484 (0.194)	0.585 (0.226)
IM_{RF}	1.395 (0.158)	1.411 (0.146)	1.471 (0.127)	1.498 (0.163)	1.543 (0.144)
IM_{XGB}	0.970 (0.306)	1.064 (0.309)	1.241 (0.342)	1.353 (0.418)	1.642 (0.459)
CORAL	1.590 (0.307)	1.527 (0.279)	1.580 (0.246)	1.546 (0.291)	1.568 (0.246)
TransFusion	0.850 (0.212)	0.798 (0.191)	0.851 (0.183)	0.825 (0.208)	0.828 (0.182)
TransGLM	0.939 (0.231)	0.893 (0.203)	0.954 (0.210)	0.928 (0.210)	0.932 (0.203)
$\text{TransGLM}_{\text{iw}}$	0.403 (0.228)	0.394 (0.211)	0.375 (0.168)	0.382 (0.241)	0.397 (0.254)

Table S.3: Binary-outcome dimensionality grid. The columns vary the dimension $p \in \{50, 100, 200, 400, 800\}$, while $n_{\mathcal{S},0} = 400$, $q = 100$, $n_{\mathcal{S},1} = 2000$, $n_{\mathcal{T},0} = 2000$, and $n_{\mathcal{T},1} = 3000$ are fixed. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications; boldface indicates the smallest Monte Carlo mean within each setting and column.

Setting I			
Method	$n_{\mathcal{T},0}$		
	1000	2000	3000
AIMS	**0.273 (0.175)**	**0.259 (0.173)**	**0.259 (0.154)**
$\text{IM}_{\text{aLasso}}$	0.528 (0.228)	0.521 (0.157)	0.519 (0.191)
$\text{TransGLM}_{\text{iw}}$	0.411 (0.243)	0.366 (0.173)	0.377 (0.171)
TransFusion	0.833 (0.208)	0.849 (0.188)	0.810 (0.180)
Setting II			
Method	$n_{\mathcal{T},0}$		
	1000	2000	3000
AIMS	**0.280 (0.165)**	**0.267 (0.166)**	**0.260 (0.168)**
$\text{IM}_{\text{aLasso}}$	0.474 (0.220)	0.489 (0.174)	0.468 (0.179)
$\text{TransGLM}_{\text{iw}}$	0.390 (0.211)	0.376 (0.185)	0.386 (0.207)
TransFusion	0.810 (0.196)	0.817 (0.189)	0.799 (0.177)
Setting III			
Method	$n_{\mathcal{T},0}$		
	1000	2000	3000
AIMS	**0.294 (0.179)**	**0.288 (0.158)**	**0.276 (0.161)**
$\text{IM}_{\text{aLasso}}$	0.509 (0.206)	0.484 (0.194)	0.502 (0.185)
$\text{TransGLM}_{\text{iw}}$	0.395 (0.174)	0.382 (0.241)	0.378 (0.159)
TransFusion	0.847 (0.191)	0.825 (0.208)	0.848 (0.191)

Table S.4: Sensitivity to the minority target-unlabeled sample size. The columns vary $n_{\mathcal{T},0} \in \{1000, 2000, 3000\}$, while $n_{\mathcal{S},0} = 400$, $p = 400$, $q = 100$, $n_{\mathcal{S},1} = 2000$, and $n_{\mathcal{T},1} = 3000$ are fixed. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications; boldface indicates the smallest Monte Carlo mean within each setting and column.

Method	Setting I	Setting II	Setting III
$\widetilde{\boldsymbol{\beta}}^{[0]}$	0.554 (0.193)	0.522 (0.149)	0.553 (0.172)
$\widehat{\boldsymbol{\beta}}^{[0]}_{\text{Deb}}$	1.879 (0.411)	1.923 (0.344)	2.014 (0.686)
$\text{AIMS}_{\text{min-o}}$	0.708 (0.214)	0.632 (0.114)	0.711 (0.168)
$\text{AIMS}_{\text{maj-g}}$	**0.242 (0.150)**	0.274 (0.161)	0.263 (0.173)
AIMS	0.247 (0.147)	**0.260 (0.150)**	**0.262 (0.157)**
AIMS ($a = 1$)	0.278 (0.123)	0.275 (0.128)	0.292 (0.135)
AIMS ($a = 2$)	0.256 (0.134)	0.263 (0.138)	0.270 (0.147)
AIMS ($a = 10$)	0.255 (0.158)	0.263 (0.157)	0.266 (0.166)
AIMS ($a = \infty$)	0.278 (0.186)	0.282 (0.175)	0.281 (0.183)

Table S.5: AIMS ablation at the reference binary-outcome design ($n_{\mathcal{S},0} = 400$, $p = 400$, $q = 100$, $n_{\mathcal{S},1} = 2000$, $n_{\mathcal{T},0} = 2000$, $n_{\mathcal{T},1} = 3000$, $C_\tau = 2$, and $q_\tau = 0.8$). Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications; boldface indicates the smallest Monte Carlo mean within each column.

Method	Setting I				
	p				
	50	100	200	400	800
$\widetilde{\boldsymbol{\beta}}^{[1]}$	0.202 (0.101)	0.187 (0.078)	0.194 (0.115)	0.199 (0.095)	0.189 (0.086)
$\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Deb}}$	0.680 (0.179)	0.688 (0.208)	0.673 (0.200)	0.671 (0.184)	0.650 (0.177)
$\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Thr}}$	0.250 (0.181)	0.237 (0.180)	0.240 (0.174)	0.224 (0.165)	0.206 (0.168)
Method	Setting II				
	p				
	50	100	200	400	800
$\widetilde{\boldsymbol{\beta}}^{[1]}$	0.216 (0.111)	0.202 (0.112)	0.193 (0.099)	0.184 (0.092)	0.162 (0.089)
$\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Deb}}$	0.735 (0.226)	0.734 (0.275)	0.697 (0.196)	0.684 (0.177)	0.642 (0.150)
$\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Thr}}$	0.301 (0.199)	0.277 (0.207)	0.251 (0.187)	0.232 (0.183)	0.198 (0.167)
Method	Setting III				
	p				
	50	100	200	400	800
$\widetilde{\boldsymbol{\beta}}^{[1]}$	0.209 (0.114)	0.193 (0.094)	0.177 (0.086)	0.192 (0.078)	0.180 (0.076)
$\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Deb}}$	0.712 (0.327)	0.678 (0.183)	0.656 (0.179)	0.648 (0.179)	0.623 (0.171)
$\widehat{\boldsymbol{\beta}}^{[1]}_{\mathrm{Thr}}$	0.270 (0.215)	0.247 (0.164)	0.213 (0.171)	0.257 (0.187)	0.210 (0.158)

Table S.6: Majority-side diagnostic performance along the dimensionality axis ($q = 100$, $n_{\mathcal{S},0} = 400$, $n_{\mathcal{T},0} = 2000$, $n_{\mathcal{S},1} = 2000$, and $n_{\mathcal{T},1} = 3000$). Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[1]}\|_2^2$ over 100 replications.

Setting	q_τ			
	0.6	0.7	0.8	0.9
Setting I	0.273 (0.159)	0.267 (0.160)	0.247 (0.147)	**0.243 (0.143)**
Setting II	0.276 (0.147)	0.262 (0.150)	0.260 (0.150)	**0.255 (0.147)**
Setting III	0.279 (0.175)	0.270 (0.170)	**0.262 (0.157)**	0.264 (0.149)

Table S.7: Sensitivity to the bootstrap quantile q_τ at the reference binary-outcome design with $n_{\mathcal{S},0} = 400$, $p = 400$, $q = 100$, $n_{\mathcal{T},0} = 2000$, and $C_\tau = 2$. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications; boldface indicates the smallest Monte Carlo mean within each row.

Setting	C_τ				
	1.0	1.5	2.0	2.5	3.0
Setting I	0.212 (0.123)	**0.163 (0.134)**	0.247 (0.147)	0.349 (0.157)	0.464 (0.163)
Setting II	0.211 (0.107)	**0.170 (0.130)**	0.260 (0.150)	0.393 (0.162)	0.466 (0.147)
Setting III	0.229 (0.119)	**0.186 (0.141)**	0.262 (0.157)	0.393 (0.177)	0.476 (0.161)

Table S.8: Sensitivity to the threshold constant C_τ at the reference binary-outcome design with $n_{\mathcal{S},0} = 400$, $p = 400$, $q = 100$, $n_{\mathcal{T},0} = 2000$, and $q_\tau = 0.8$. The column $C_\tau = 2$ reproduces the default AIMS results reported in Tables S.5 and S.7. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications; boldface indicates the smallest Monte Carlo mean within each row.

S.3.4 Continuous-Outcome Simulation Results

To assess the robustness of the empirical findings to the outcome type, we repeat the sample-size and dimensionality grids in Section 5 with a continuous outcome; the results are reported in Tables S.9 and S.10, respectively. All covariate distributions, coefficient vectors $\boldsymbol{\alpha}^{[r]}$ and $\boldsymbol{\gamma}^{[r]}$, nonlinear functions $u_Y(\mathbf{Z})$ and $u_S(\mathbf{Z})$, and source-sampling mechanisms are kept the same as in the binary-outcome simulation. The reported continuous-outcome tables use $q = 100$, $n_{\mathcal{S},1} = 2000$, $n_{\mathcal{T},1} = 3000$, and $n_{\mathcal{T},0} = 2000$. Along the sample-size axis, $p = 400$ and $n_{\mathcal{S},0} \in \{300, 400, 500, 600\}$; along the dimensionality axis, $n_{\mathcal{S},0} = 400$ and $p \in \{50, 100, 200, 400, 800\}$. We use the same implementation and tuning constants as in the main simulation, namely $C_\tau = 2$, $q_\tau = 0.8$, $B = 500$, and aggregation temperature $a = 5$, and report Monte Carlo averages over 100 replications.

Specifically, we set $g(a) = b(a) = a$, so both the target and imputation working models use the identity link. Conditional on $\mathbf{Z}$ and $R = r$, we generate $Y = \mathbf{Z}^\top \boldsymbol{\gamma}^{[r]} + \epsilon$ in Settings I and III and $Y = \mathbf{Z}^\top \boldsymbol{\gamma}^{[r]} + u_Y(\mathbf{Z}) + \epsilon$ in Setting II, where $\epsilon \sim N(0, 1)$. Thus, Setting II keeps the density ratio model correctly specified while misspecifying the imputation model, and Setting III keeps the imputation model correctly specified while misspecifying the density ratio model. Tables S.9 and S.10 show that the continuous-outcome results are consistent with the binary-outcome findings: AIMS has the smallest error across the reported sample-size and dimensionality grids, while $\text{TransGLM}_{\text{iw}}$ is the closest external competitor.

Method	$n_{\mathcal{S},0}$ = 300	400	500	600
Setting I				
AIMS	**0.093 (0.053)**	**0.067 (0.046)**	**0.062 (0.037)**	**0.064 (0.045)**
IW	0.554 (0.243)	0.436 (0.159)	0.385 (0.159)	0.321 (0.137)
$\text{IW}_{\text{aLasso}}$	0.801 (0.374)	0.625 (0.266)	0.578 (0.233)	0.471 (0.195)
IM	0.323 (0.200)	0.232 (0.132)	0.193 (0.118)	0.171 (0.087)
$\text{IM}_{\text{aLasso}}$	0.197 (0.079)	0.158 (0.054)	0.124 (0.042)	0.107 (0.037)
IM_{RF}	1.900 (0.214)	1.748 (0.181)	1.643 (0.159)	1.599 (0.138)
IM_{XGB}	1.172 (0.223)	0.941 (0.170)	0.793 (0.160)	0.707 (0.151)
CORAL	2.052 (0.262)	2.016 (0.243)	2.026 (0.229)	2.045 (0.208)
TransFusion	0.969 (0.152)	0.957 (0.123)	0.949 (0.108)	0.949 (0.105)
TransGLM	1.079 (0.181)	1.045 (0.140)	1.035 (0.116)	1.028 (0.125)
$\text{TransGLM}_{\text{iw}}$	0.295 (0.148)	0.238 (0.133)	0.234 (0.135)	0.201 (0.108)
Setting II				
AIMS	**0.118 (0.066)**	**0.098 (0.045)**	**0.095 (0.043)**	**0.094 (0.053)**
IW	0.506 (0.211)	0.425 (0.195)	0.350 (0.148)	0.284 (0.117)
$\text{IW}_{\text{aLasso}}$	0.851 (0.411)	0.665 (0.281)	0.565 (0.246)	0.447 (0.204)
IM	0.249 (0.161)	0.197 (0.135)	0.145 (0.069)	0.128 (0.061)
$\text{IM}_{\text{aLasso}}$	0.254 (0.089)	0.226 (0.066)	0.199 (0.058)	0.184 (0.048)
IM_{RF}	1.664 (0.190)	1.511 (0.163)	1.400 (0.145)	1.353 (0.114)
IM_{XGB}	0.920 (0.169)	0.739 (0.152)	0.595 (0.123)	0.527 (0.118)
CORAL	2.193 (0.315)	2.172 (0.298)	2.192 (0.274)	2.214 (0.225)
TransFusion	1.146 (0.165)	1.134 (0.140)	1.118 (0.123)	1.118 (0.109)
TransGLM	1.144 (0.184)	1.111 (0.143)	1.096 (0.129)	1.088 (0.124)
$\text{TransGLM}_{\text{iw}}$	0.272 (0.130)	0.230 (0.150)	0.209 (0.112)	0.173 (0.093)
Setting III				
AIMS	**0.085 (0.054)**	**0.072 (0.040)**	**0.070 (0.042)**	**0.062 (0.038)**
IW	0.519 (0.162)	0.464 (0.159)	0.348 (0.121)	0.325 (0.116)
$\text{IW}_{\text{aLasso}}$	0.783 (0.288)	0.703 (0.297)	0.546 (0.234)	0.473 (0.187)
IM	0.327 (0.179)	0.258 (0.147)	0.189 (0.112)	0.173 (0.118)
$\text{IM}_{\text{aLasso}}$	0.191 (0.069)	0.155 (0.055)	0.121 (0.040)	0.106 (0.032)
IM_{RF}	1.811 (0.209)	1.725 (0.158)	1.609 (0.163)	1.531 (0.156)
IM_{XGB}	1.123 (0.213)	0.958 (0.171)	0.773 (0.167)	0.684 (0.161)
CORAL	2.024 (0.257)	2.040 (0.221)	2.027 (0.234)	1.980 (0.206)
TransFusion	0.961 (0.154)	0.997 (0.136)	0.976 (0.117)	0.957 (0.108)
TransGLM	1.040 (0.176)	1.074 (0.161)	1.038 (0.139)	1.012 (0.118)
$\text{TransGLM}_{\text{iw}}$	0.311 (0.136)	0.293 (0.124)	0.242 (0.106)	0.231 (0.115)

Table S.9: Continuous-outcome sample-size grid. The columns vary the labeled minority-source sample size $n_{\mathcal{S},0} \in \{300, 400, 500, 600\}$, while $p = 400$, $q = 100$, $n_{\mathcal{S},1} = 2000$, $n_{\mathcal{T},0} = 2000$, and $n_{\mathcal{T},1} = 3000$ are fixed. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications. Boldface indicates the smallest Monte Carlo mean within each setting and column.

Method	p = 50	100	200	400	800
Setting I					
AIMS	**0.093 (0.048)**	**0.100 (0.103)**	**0.086 (0.047)**	**0.067 (0.046)**	**0.073 (0.049)**
IW	0.446 (0.160)	0.445 (0.166)	0.444 (0.192)	0.436 (0.159)	0.436 (0.160)
$\text{IW}_{\text{aLasso}}$	0.640 (0.269)	0.663 (0.309)	0.643 (0.262)	0.625 (0.266)	0.613 (0.263)
IM	0.196 (0.128)	0.219 (0.137)	0.217 (0.152)	0.232 (0.132)	0.313 (0.176)
$\text{IM}_{\text{aLasso}}$	0.161 (0.051)	0.158 (0.048)	0.158 (0.047)	0.158 (0.054)	0.157 (0.061)
IM_{RF}	1.591 (0.180)	1.630 (0.169)	1.684 (0.177)	1.748 (0.181)	1.859 (0.189)
IM_{XGB}	0.743 (0.185)	0.828 (0.186)	0.870 (0.190)	0.941 (0.170)	1.095 (0.204)
CORAL	2.029 (0.220)	2.021 (0.225)	2.047 (0.251)	2.016 (0.243)	2.047 (0.218)
TransFusion	0.958 (0.121)	0.965 (0.140)	0.954 (0.138)	0.957 (0.123)	0.957 (0.125)
TransGLM	1.052 (0.142)	1.066 (0.148)	1.044 (0.149)	1.045 (0.140)	1.053 (0.136)
$\text{TransGLM}_{\text{iw}}$	0.243 (0.118)	0.238 (0.114)	0.250 (0.134)	0.238 (0.133)	0.245 (0.127)
Setting II					
AIMS	**0.119 (0.060)**	**0.117 (0.088)**	**0.116 (0.075)**	**0.098 (0.045)**	**0.105 (0.047)**
IW	0.408 (0.163)	0.409 (0.165)	0.405 (0.153)	0.425 (0.195)	0.399 (0.149)
$\text{IW}_{\text{aLasso}}$	0.637 (0.299)	0.729 (0.329)	0.649 (0.269)	0.665 (0.281)	0.658 (0.283)
IM	0.163 (0.115)	0.171 (0.131)	0.171 (0.104)	0.197 (0.135)	0.223 (0.112)
$\text{IM}_{\text{aLasso}}$	0.258 (0.066)	0.258 (0.067)	0.244 (0.069)	0.226 (0.066)	0.188 (0.056)
IM_{RF}	1.353 (0.163)	1.388 (0.159)	1.451 (0.161)	1.511 (0.163)	1.607 (0.153)
IM_{XGB}	0.580 (0.158)	0.616 (0.155)	0.668 (0.157)	0.739 (0.152)	0.848 (0.148)
CORAL	2.156 (0.256)	2.187 (0.289)	2.207 (0.294)	2.172 (0.298)	2.209 (0.270)
TransFusion	1.126 (0.129)	1.141 (0.160)	1.126 (0.149)	1.134 (0.140)	1.128 (0.142)
TransGLM	1.109 (0.143)	1.128 (0.167)	1.106 (0.158)	1.111 (0.143)	1.116 (0.148)
$\text{TransGLM}_{\text{iw}}$	0.237 (0.127)	0.208 (0.103)	0.226 (0.116)	0.230 (0.150)	0.225 (0.107)
Setting III					
AIMS	**0.088 (0.043)**	**0.082 (0.044)**	**0.079 (0.044)**	**0.072 (0.040)**	**0.070 (0.049)**
IW	0.430 (0.181)	0.431 (0.172)	0.438 (0.187)	0.464 (0.159)	0.430 (0.138)
$\text{IW}_{\text{aLasso}}$	0.631 (0.285)	0.660 (0.310)	0.661 (0.308)	0.703 (0.297)	0.656 (0.285)
IM	0.201 (0.125)	0.226 (0.158)	0.216 (0.137)	0.258 (0.147)	0.292 (0.157)
$\text{IM}_{\text{aLasso}}$	0.156 (0.050)	0.162 (0.056)	0.157 (0.047)	0.155 (0.055)	0.151 (0.055)
IM_{RF}	1.527 (0.152)	1.565 (0.167)	1.651 (0.179)	1.725 (0.158)	1.762 (0.189)
IM_{XGB}	0.727 (0.163)	0.779 (0.175)	0.861 (0.192)	0.958 (0.171)	1.044 (0.193)
CORAL	2.020 (0.237)	2.013 (0.266)	2.002 (0.244)	2.040 (0.221)	1.988 (0.237)
TransFusion	0.997 (0.129)	0.975 (0.136)	0.974 (0.132)	0.997 (0.136)	0.963 (0.124)
TransGLM	1.069 (0.148)	1.053 (0.154)	1.048 (0.148)	1.074 (0.161)	1.043 (0.149)
$\text{TransGLM}_{\text{iw}}$	0.259 (0.122)	0.265 (0.122)	0.276 (0.155)	0.293 (0.124)	0.280 (0.109)

Table S.10: Continuous-outcome dimensionality grid. The columns vary the dimension $p \in \{50, 100, 200, 400, 800\}$, while $n_{\mathcal{S},0} = 400$, $q = 100$, $n_{\mathcal{S},1} = 2000$, $n_{\mathcal{T},0} = 2000$, and $n_{\mathcal{T},1} = 3000$ are fixed. Each cell reports the Monte Carlo mean (empirical standard deviation) of $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^{[0]}\|_2^2$ over 100 replications. Boldface indicates the smallest Monte Carlo mean within each setting and column.

S.3.5 Additional T2D Benchmark Results

Table S.11 reports the complete T2D real-data comparison, including the adaptive-Lasso variants, RF-based imputation, TransGLM variants, and the MAP-based phenotyping baseline that are omitted from the main-text table for readability.

	AIMS	IW	$\text{IW}_{\text{aLasso}}$	IM	$\text{IM}_{\text{aLasso}}$	IM_{RF}	IM_{XGB}	CORAL	TransFusion	TransGLM	$\text{TransGLM}_{\text{iw}}$	$\text{IM}_{\text{PheNorm}}$	IM_{MAP}
BSS	**0.36**	-0.11	-0.22	0.23	0.24	-1.00	-1.07	-0.11	0.01	0.00	0.00	-0.20	-0.29
GOF	**0.16**	-0.85	-1.87	-0.02	-0.02	-1.36	-1.46	-0.88	-0.25	-0.78	-0.82	-0.45	-0.54
AUC	**0.89**	0.64	0.62	0.81	0.83	0.78	0.81	0.62	0.81	0.71	0.70	0.82	0.82

Table S.11: Full predictive performance comparison for the T2D risk models evaluated on the validation data. This table includes additional benchmark variants omitted from Table 1 for readability.

S.3.6 Additional $\Delta\Delta G$ Results on the Target Majority Subgroup

Table S.12 compares the preliminary estimator $\widetilde{\boldsymbol{\beta}}^{[1]}$ ($\text{AIMS}_{\text{Init}}$), the dense debiased estimator $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Dense\_Deb}}$ ($\text{AIMS}_{\text{Dense\_Deb}}$), and the thresholded estimator $\widehat{\boldsymbol{\beta}}^{[1]}_{\text{Thr}}$ (AIMS_{Thr}) for the majority subgroup. The thresholded estimator achieves the strongest Pearson and Spearman correlations with the held-out target-majority $\Delta\Delta G$ labels, while the preliminary estimator yields slightly smaller RMSE and MAE. The dense debiased estimator performs worst under the prediction-error metrics RMSE and MAE, which is consistent with its role as an intermediate dense debiased estimator rather than a final prediction rule.

	$\text{AIMS}_{\text{Init}}$	$\text{AIMS}_{\text{Dense\_Deb}}$	AIMS_{Thr}
RMSE	2.24	3.03	2.31
MAE	1.54	2.20	1.57
Pearson	0.06	0.14	0.20
Spearman	0.09	0.20	0.31

Table S.12: Predictive performance of the preliminary, dense debiased, and thresholded AIMS estimators on the target majority subgroup, evaluated using held-out ground-truth $\Delta\Delta G$ labels.

References

Nicole D Armstrong, Amit Patki, Vinodh Srinivasasainagendra, Tian Ge, Leslie A Lange, Leah Kottyan, Bahram Namjou, Amy S Shah, Laura J Rasmussen-Torvik, Gail P Jarvik, et al. Variant level heritability estimates of type 2 diabetes in African Americans. Scientific Reports, 14(1):14009, 2024.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. Management Science, 67(5):2964–2984, 2021.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. The Annals of Statistics, 46(3): 1352–1382, 2018. doi: 10.1214/17-AOS1587.

Adam Baybutt and Manu Navjeevan. Doubly-robust inference for conditional average treatment effects with high-dimensional controls. Journal of Econometrics, 253:106180, 2026.

Jeffrey Braithwaite. Changing how we think about healthcare improvement. BMJ, 361: k2014, 2018.

Tianxi Cai, Mengyan Li, and Molei Liu. Semi-supervised triply robust inductive transfer learning. Journal of the American Statistical Association, pages 1–11, 2024.

Victor M Castro, Vivian Gainer, Nich Wattanasin, Barbara Benoit, Andrew Cagan, Bhaswati Ghosh, Sergey Goryachev, Reeta Metta, Heekyong Park, David Wang, et al. The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. Journal of the American Medical Informatics Association, 29(4):643–651, 2022.

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

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

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. Statistical Science, 39(1):165–191, 2024.

Arnak S Dalalyan and Alexandre B Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In International Conference on Computational Learning Theory, pages 97–111. Springer, 2007.

Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. Journal of Business & Economic Statistics, 40(1):313–327, 2022.

Subharup Guha, Mengqi Xu, and Yi Li. Enhancing inference for small cohorts via transfer learning and weighted integration of multiple datasets. arXiv preprint arXiv:2505.07153, 2025.

Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Jose Posada, Scott Lanyon Fleming, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Applied Clinical Informatics, 12(04):808–815, 2021.

Zelin He, Ying Sun, and Runze Li. TransFusion: Covariate-shift robust transfer learning for high-dimensional regression. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, volume 238 of Proceedings of Machine Learning Research, pages 703–711. PMLR, 2024.

Iván Martín Hernández, Yves Dehouck, Ugo Bastolla, José Ramón López-Blanco, and Pablo Chacón. Predicting protein stability changes upon mutation using a simple orientational potential. Bioinformatics, 39(1):btad011, 2023.

Chuan Hong, Everett Rush, Molei Liu, Doudou Zhou, Jiehuan Sun, Aaron Sonabend, Victor M Castro, Petra Schubert, Vidul A Panickan, Tianrun Cai, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. npj Digital Medicine, 4(1):151, 2021.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In Advances in Neural Information Processing Systems, pages 601–608, 2007.

Masahiro Kato. Triple/debiased lasso for statistical inference of conditional average treatment effects. arXiv preprint arXiv:2403.03240, 2024.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics, 17(2):3008–3049, 2023.

Seok-Jin Kim, Hongjie Liu, Molei Liu, and Kaizheng Wang. Transfer learning of CATE with kernel ridge regression. arXiv preprint arXiv:2502.11331, 2025.

Max Lam, Chia-Yen Chen, Zhiqiang Li, Alicia R Martin, Julien Bryois, Xixian Ma, Helena Gaspar, Masashi Ikeda, Beben Benyamin, Brielin C Brown, et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. Nature Genetics, 51(12):1670–1678, 2019.

Pascal Leuenberger, Stefan Ganscha, Abdullah Kahraman, Valentina Cappelletti, Paul J. Boersema, Christian von Mering, Manfred Claassen, and Paola Picotti. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science, 355(6327): eaai7825, 2017.

Sai Li and Linjun Zhang. Multi-dimensional domain generalization with low-rank structures. Journal of the American Statistical Association, pages 1–13, 2025.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1):149–173, 2022.

Sai Li, Tianxi Cai, and Rui Duan. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. The Annals of Applied Statistics, 17 (4):2970–2992, 2023.

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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637): 1123–1130, 2023.

Molei Liu, Yi Zhang, Katherine P Liao, and Tianxi Cai. Augmented transfer regression learning with semi-non-parametric nuisance models. Journal of Machine Learning Research, 24(293):1–50, 2023.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. The Journal of Machine Learning Research, 16(1):559–616, 2015.

Ethan B Ludmir, Walker Mainwaring, Timothy A Lin, Austin B Miller, Amit Jethanandani, Andres F Espinoza, Jacob J Mandel, Steven H Lin, Benjamin D Smith, Grace L Smith, et al. Factors associated with age disparities among cancer clinical trial participants. JAMA Oncology, 5(12):1769–1773, 2019.

Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R Robertson, Jason M Torres, N William Rayner, Anthony J Payne, Valgerdur Steinthorsdottir, Robert A Scott, Niels Grarup, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics, 50(11):1505–1513, 2018.

Josep M Mercader and Jose C Florez. The genetic basis of type 2 diabetes in Hispanics and Latin Americans: challenges and opportunities. Frontiers in Public Health, 5:329, 2017.

Karel GM Moons, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward. Risk prediction models: II. External validation, model updating, and impact assessment. Heart, 98(9):691–698, 2012.

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

Rahul Nikam, A Kulandaisamy, K Harini, Divya Sharma, and M Michael Gromiha. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Research, 49(D1):D420–D424, 2021.

Edward P. O'Brien, Bernard R. Brooks, and D. Thirumalai. Effects of pH on proteins: Predictions for ensemble and single-molecule pulling experiments. Journal of the American Chemical Society, 134(2):979–987, 2012.

Emily D Parker, Janice Lin, Troy Mahoney, Nwanneamaka Ume, Grace Yang, Robert A Gabbay, Nuha A ElSayed, and Raveendhara R Bannuru. Economic costs of diabetes in the US in 2022. Diabetes Care, 47(1):26–43, 2024.

Ross L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine, 8(4):431–440, 1989. doi: 10.1002/sim.4780080407.

Hongxiang Qiu, Eric Tchetgen Tchetgen, and Edgar Dobriban. Efficient and multiply robust risk estimation under general forms of dataset shift. The Annals of Statistics, 52 (4):1796–1824, 2024.

Sashank Jakkam Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pages 2949–2955. AAAI Press, 2015.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. The Econometrics Journal, 24(2):264–289, 2021.

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

Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In Domain Adaptation in Computer Vision Applications, pages 153–171. Springer, 2017.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. Annals of Statistics, 48(2):811–837, 2020.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544):2684–2697, 2023.

Ye Tian, Peng Wu, and Zhiqiang Tan. Semi-supervised regression analysis with model misspecification and high-dimensional data. arXiv preprint arXiv:2406.13906, 2024.

Peng-Sheng Ting, Likwang Chen, Wei-Chih Yang, Tien-Shang Huang, Chau-Chung Wu, and Yen-Yuan Chen. Gender and age disparity in the initiation of life-supporting treatments: a population-based cohort study. BMC Medical Ethics, 18:62, 2017.

Martin Tollinger, Karin A. Crowhurst, Lewis E. Kay, and Julie D. Forman-Kay. Site-specific contributions to the pH dependence of protein stability. Proceedings of the National Academy of Sciences, 100(8):4545–4550, 2003.

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

Anurag Verma, Jennifer E Huffman, Alex Rodriguez, Mitchell Conery, Molei Liu, Yuk-Lam Ho, Youngdae Kim, David A Heise, Lindsay Guare, Vidul Ayakulangara Panickan, et al. Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. Science, 385(6706):eadj1182, 2024.

Costantino Vetriani, Dennis L. Maeder, Nicola Tolliday, Kitty S.-P. Yip, Timothy J. Stillman, K. Linda Britton, David W. Rice, Horst H. Klump, and Frank T. Robb. Protein thermostability above 100°C: A key role for ionic interactions. Proceedings of the National Academy of Sciences, 95(21):12300–12305, 1998.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

Fan Wang, Zhongyi Han, Xingbo Liu, Xin Gao, and Yilong Yin. Unveiling the superior paradigm: a comparative study of source-free domain adaptation and unsupervised domain adaptation. arXiv preprint arXiv:2411.15844, 2024.

Kaizheng Wang. Pseudo-Labeling for kernel ridge regression under covariate shift. The Annals of Statistics, 54(1):252–276, 2026.

Joicymara S Xavier, Thanh-Binh Nguyen, Malancha Karmarkar, Stephanie Portelli, Pâmela M Rezende, João P L Velloso, David B Ascher, and Douglas E V Pires. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Research, 49(D1):D475–D479, 2021.

Yunxin Xu, Di Liu, and Haipeng Gong. Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy. Nature Computational Science, 4(11):840–850, 2024.

Sheng Yu, Yumeng Ma, Jessica Gronsbell, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, et al. Enabling phenotypic big data with PheNorm. Journal of the American Medical Informatics Association, 25(1):54–60, 2018.

Keyao Zhan, Xin Xiong, Zijian Guo, Tianxi Cai, and Molei Liu. Transfer learning targeting mixed population: A distributional robust perspective. arXiv preprint arXiv:2407.20073, 2024.

Doudou Zhou, Molei Liu, Mengyan Li, and Tianxi Cai. Doubly robust augmented model accuracy transfer inference with high dimensional features. Journal of the American Statistical Association, 120(549):524–534, 2025.