Transfer Learning with General Estimating Equations

Han Yan[1] , Song Xi Chen[1,2]
[1] Guanghua School of Management, Peking University,
[2] School of Mathematical Science, Peking University

Abstract

We consider statistical inference for parameters defined by general estimating equations under the covariate shift transfer learning. Different from the commonly used density ratio weighting approach, we undertake a set of formulations to make the statistical inference semiparametric efficient with simple inference. It starts with re-constructing the estimation equations to make them Neyman orthogonal, which facilitates more robustness against errors in the estimation of two key nuisance functions, the density ratio and the conditional mean of the moment function. We present a divergence-based method to estimate the density ratio function, which is amenable to machine learning algorithms including the deep learning. To address the challenge that the conditional mean is parametric-dependent, we adopt a nonparametric multiple-imputation strategy that avoids regression at all possible parameter values. With the estimated nuisance functions and the orthogonal estimation equation, the inference for the target parameter is formulated via the empirical likelihood without sample splittings. We show that the proposed estimator attains the semiparametric efficiency bound, and the inference can be conducted with the Wilks' theorem. The proposed method is further evaluated by simulations and an empirical study on a transfer learning inference for ground-level ozone pollution.

1 Introduction

The past decades have witnessed rapid development of statistical learning techniques in many fields of applications. Most of the techniques rely on a commonly adopted model where the training and testing data are sampled from the same distribution. However, this homogeneity between the training and testing data is frequently violated in practice since diverse datasets are increasingly available. An enduring challenge in statistical inference is to generalize an inference procedure from one data domain to another to achieve the goal of generalization and fully use of data information.

Transfer learning (TL) has become an active and promising area in dealing with distribution mismatch problems and has achieved considerable success in a wide range of applications, such as computer vision (Kulis et al., 2011), bioinformatics (Hanson et al., 2020) and precision medicine (Mo et al., 2021). See Zhuang et al. (2020) for reviews.

Suppose that the distributions for the source and target samples are $P_{\boldsymbol{X},Y}$ and $Q_{\boldsymbol{X},Y}$, respectively, where $\boldsymbol{X}$ is the vector of covariates and Y is the response or outcome. Two popular settings of TL that have been considered in the literature are posterior drift and covariate shift. The posterior drift TL assumes the marginal distributions of covariates are invariant, while the conditional distributions $P_{Y|\boldsymbol{X}}$ and $Q_{Y|\boldsymbol{X}}$ may differ. On the other hand, covariate shift TL is for the situations where the marginal covariate distributions may differ, while a common conditional distribution is shared across the two domains. Recently, there has been a growing literature on statistical inference on the posterior drift TL, including the classification problems (Reeve et al., 2021), the linear and generalized linear models (Li et al., 2022b), and the Gaussian graphical models (Li et al., 2023). Unlike the posterior drift, the covariate shift TL has responses inaccessible to the target sample. Such a setting is well motivated by various real-world scenarios, where the same study is conducted with different covariate populations while the law that governs the input-output determination is kept across the domains. For instance, in medical data analysis (Guan and Liu, 2021), covariate shift is reasonable for health records across different patients, and clinical outcomes are usually scarce because of ethical concerns. Despite its importance in applications, there has been a limited amount of literature on the statistical theory on the covariate shift TL relative to those for the posterior shift TL.

In this paper, we consider statistical inference on parameters defined via general estimating equations (GEE) in the context of the covariate shift TL. The GEE is a general framework for semi-parametric inference, and is appealing for requiring less stringent distributional assumptions on the data, and yet can encompass a wide range of model structures and parameters. The goal is to efficiently estimate and make inference for a p-dimensional parameter $\boldsymbol{\theta}_0$ defined through $\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{X}, Y, \boldsymbol{\theta}_0)\} = \mathbf{0}$. Inference for $\boldsymbol{\theta}_0$ under this situation is more challenging than the conventional GEE problems because on one hand, Y is available in the target sample. On the other hand, directly using the sample from the source population P leads to biased estimates since $P \neq Q$ in general.

1.1 Related works

We review related works so as to situate our study within a broader context and discuss the gaps between existing results and the goal of this paper.

Covariate shift The covariate shift, as an important scenario of TL, is also called the domain adaptation (Pan and Yang, 2010) and has been investigated in the machine learning literature, such as Gretton et al. (2009) and Ryan and Culp (2015), with a focus to correct for estimating bias in the empirical risk minimization or model selection due to the covariate shift. The standard strategy adopted in the existing literature is the so-called importance reweighting with the density ratio between $Q_{\boldsymbol{X}}$ and $P_{\boldsymbol{X}}$; see Kouw and Loog (2019) and the references therein. The covariate shift problem has also been studied from the perspectives of statistical methodologies and theories. Lei and Candès (2021) studied conformal prediction under covariate shift. Non-parametric classification under covariate shift is explored in Kpotufe and Martinet (2021) and non-parametric regression is investigated in Ma et al. (2023). Cai et al. (2023) considered contextual multi-armed bandits under the covariate shift. In comparison, the semi-parametric inference under the covariate shift is less-explored.

Missing data and causal inference The covariate shift TL is closely related to missing data problems and causal inference, since the assumption $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$ is equivalent to the missing at random (MAR) condition. The sample estimating equation employed in this work shares a similar form as the augmented inverse-probability weighted (AIPW) estimator (Robins et al., 1994) and its variants (e.g., Rotnitzky et al., 2012 and Chernozhukov et al., 2018). Both the AIPW method and our proposed method require the estimation of the conditional mean function $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})|\boldsymbol{X}\}$. A key distinction of the GEE considered in this paper from the aforementioned literature is that their estimand is commonly linear in $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})$, for example, the average treatment effect (ATE) problem corresponds to the estimating function $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}) = Y - \theta$, while we are interested in more general cases where $\boldsymbol{\theta}$ may depend nonlinearly on $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})$, such as the quantile and quantile regression. For the nonlinear cases, to estimate the conditional mean function $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta})$, one has to regress $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})$ on $\boldsymbol{X}$ repeatedly for each $\boldsymbol{\theta}$ during its optimization, which can be too computationally intensive to be practical, especially for some time consuming optimizations requiring thousands of iterations before convergence.

Recently, Chen et al. (2024) considered the GEE problem with missing data and proposed a neural network based inverse probability weighting estimator. Compared with Chen et al. (2024), our methods have two appealing advantages. One is being doubly robust in that our estimator is consistent if either the density ratio or the conditional mean function is consistently estimated. The other important feature is that we can conveniently employ Wilks' theorem for the inference of $\boldsymbol{\theta}_0$, while Chen et al. (2024) has to resort to a Bootstrap method to facilitate the inference. More detailed comparisons are presented in Section 6.1.

Empirical likelihood The empirical likelihood (EL) approach introduced in Owen (1988) has been demonstrated to be powerful for statistical inference of GEEs, for having appealing properties such as Wilks' theorem (Qin and Lawless, 1994) and Bartlett correction (Chen and Cui, 2007). When an unknown nuisance function is present in the estimation equations, Hjort et al. (2009) and Wang and Chen (2009) showed that asymptotically the empirical likelihood ratio statistic with a plugged-in estimate of the nuisance function can be weighted-sum-of-χ^2 distributed, which is non-pivotal, and a bootstrap procedure has to be used to approximate the distribution of the EL ratio. Bravo et al. (2020) proposed a two-step procedure for empirical likelihood inference of semi-parametric models, employing a modified sample estimating function, which leads to an asymptotically χ^2 distributed EL ratio statistic. They also considered the GEE with missing data to illustrate their methodology. However, nuisance functions in Bravo et al. (2020) are estimated with conventional kernel smoothing whose performance may deteriorate with increase of dimensionality. In our proposal, the nuisance functions are estimated in a more flexible way that accommodates modern ML algorithms.

1.2 Our contributions

The investigation in this work contributes to several aspects.

First, we construct a modified moment function for the GEE inference in the presence of covariate shift, which has the advantage of being Neyman orthogonal (Neyman, 1959) that permits elimination of the first-order effect of the nuisance function estimation, including a density ratio function $r(\boldsymbol{x})$ and a conditional moment function $\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})$.

The second contribution is in proposing a novel estimation methods for the two nuisance functions, which both enable the use of flexible nonparametric tools, including the linear sieves and generic black-box machine learning algorithms. The density ratio function is estimated by a divergence minimization approach. The estimation of the conditional moment function $\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})$ is more challenging since it requires estimating the nuisance for infinitely many $\boldsymbol{\theta}$. To tackle such a problem, we employ a multiple imputation approach which bypasses the involvement of the parameter $\boldsymbol{\theta}$ and just needs to estimate the conditional density $p(y|\boldsymbol{X})$. Instead of the conventional kernel smoothing estimator, novel estimation methods for the density ratio and the conditional density are presented, which can utilize a broad array of nonparametric methods.

Thirdly, by employing the EL method, the proposed estimation is shown to be both doubly robust and semi-parametric efficient. Different from the double machine learning approach, the construction of the sample estimating function used for the EL does not require sample splitting. Furthermore, the log EL ratio statistics admits the Wilks' theorem which greatly facilitates the inference. For comparison, we also investigate the theoretical properties of density ratio weighting (DRW) estimation, which is commonly adopted in covariate shift problems. It is found that the DRW not only requires more stringent conditions than the proposed method to attain the same asymptotic variance, but also make the EL ratio statistics asymptotically weighted χ^2-distributed, which makes the subsequent inference tedious. We also discuss a growing dimension scenario, where the nuisance functions are estimated with the deep neural networks to circumvent the curse of dimensionality.

1.3 Orginization

The paper is organized as follows. Section 2 describes the setup of the GEE problem with the covariate shift TL. In Section 3 the orthogonal moment equations are constructed. Section 4 discusses the estimation of the two nuisance functions and establishes their theoretical properties. The inference with the EL is presented in Section 5, where the scenario with a growing dimension is also investigated. In we discuss comparisons between our study and some related works. Section 7 and 8 report numerical experiment results and a case study on ground-level ozone pollution, respectively. Finally, concluding discussions are given in Section 9.

2 Notation and problem setup

We first introduce notations used in this study. We use $\mathbb{1}(\mathcal{A})$ as the indicator function of an event $\mathcal{A}$. For any vector $\boldsymbol{v} = (v_1, \cdots, v_d)^{\mathrm{T}}$, let $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$ and $\|\boldsymbol{v}\|_p$ denote its L^p norm. For a function $f : \mathcal{X} \to \mathbb{R}$, its supreme is denoted by $\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$, and its L_p-norm under a distribution F is denoted by $\|f\|_{L_p(F)} = (\mathbb{E}_F|f(X)|^p)^{1/p}$ for any $p \geqslant 1$. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leqslant Cb_n$.

Suppose a source sample $\mathcal{D}_S$ has n independently and identically distributed (i.i.d.) observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ from a source population $\boldsymbol{Z} \sim P = P_{\boldsymbol{X}} \times P_{Y|\boldsymbol{X}}$ where $\boldsymbol{Z}_i = (\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}$ consists of a d-dimensional covariate $\boldsymbol{X}_i$ and a response/label Y_i. In this study, we take the response variable as a scalar, as the case of multivariate responses can be readily extended. The target population is $\boldsymbol{Z} \sim Q = Q_{\boldsymbol{X}} \times Q_{Y|\boldsymbol{X}}$. Observations of the target sample $\mathcal{D}_T$ are $\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_N$ with $N = n + m$, while the responses Y_i in $\mathcal{D}_T$ are *not* accessible. We introduce a binary variable δ to indicate whether the data is drawn from the source ($\delta = 0$) or the target ($\delta = 1$) population. Let $\tau = \mathbb{P}(\delta = 1)$ denote the proportion of target observations in the entire N observations, which is approximated by m/N.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ be a p-dimensional parameter taking values in $\boldsymbol{\Theta} \subset \mathbb{R}^p$. For a set of estimating equations $\{g_i(\boldsymbol{Z}, \boldsymbol{\theta})\}_{i=1}^r$, the true parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ of the target population is identified by the moment condition

$$\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}_0)\} = \mathbf{0}, \tag{2.1}$$

where $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}) = (g_1(\boldsymbol{Z}, \boldsymbol{\theta}), \ldots, g_r(\boldsymbol{Z}, \boldsymbol{\theta}))^{\mathrm{T}}$ with $r \geqslant p$, which is necessary for identifying $\boldsymbol{\theta}_0$. If $\boldsymbol{\theta}_0$ depends on Q_Y, then without further distributional conditions it is impossible to identify $\boldsymbol{\theta}_0$ using the observed data due to the missingness of Y in $\mathcal{D}_T$.

Following the standard setting in the covariate shift literature, we assume that the conditional distributions $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$ so that the information of Y can be transferred from the source sample $\mathcal{D}_S$ to the target sample $\mathcal{D}_T$, while the covariate distributions $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ can differ. Our interest is the inference on $\boldsymbol{\theta}_0$ with the combined sample $\mathcal{D} = \mathcal{D}_S \cup \mathcal{D}_T$ under the covariate shift. The following conditions are required for the sample and target populations.

Condition 1. (i) The covariate distributions $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ are absolutely continuous with densities $p_0(\boldsymbol{x})$ and $q_0(\boldsymbol{x})$ supported on $\mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact. (ii) The conditional distributions $P_{Y|\boldsymbol{X}=\boldsymbol{x}} = Q_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for every $\boldsymbol{x} \in \mathcal{X}$.

Condition 2. (i) The parameter $\boldsymbol{\theta}_0 \in \mathrm{int}(\Theta)$ is the unique solution to the moment condition $\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})\} = 0$. (ii) $\mathbb{E}_Q\{\sup_{\boldsymbol{\theta}\in\Theta} \|\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})\|_2^\alpha\} < \infty$ for some $\alpha > 2$. (iii) The eigenvalues of $\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})^{\otimes 2}\}$ are bounded away from zero and infinity. (iv) $\mathbf{g}(\boldsymbol{z}, \boldsymbol{\theta})$ is continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$ with $\mathbb{E}_Q\{\sup_{\boldsymbol{\theta}\in\mathcal{N}} \|\partial\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\|_2\} < \infty$, and $\mathbb{E}_Q\{\partial\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\}$ is of full rank.

Condition 1 summarizes assumptions about the sample and target populations under the setting of covariate shift, where Condition 1 (ii) is necessary for inferring information regarding the target domain from the source domain with accessible responses. Such a condition is also required in semi-supervised learning problems (e.g., Ryan and Culp, 2015). However, more challenging than the semi-supervised learning setup, we do not assume that $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ are the same. Condition 2 for the estimating functions are standard regularity conditions in the literature of general estimation equations (e.g., Newey and Smith, 2004).

3 Orthogonal moment functions

To address the problems caused by the covariate shift, the most common existing method is via a density ratio weighting (DRW) approach, see Sugiyama et al. (2007) for an empirical risk minimization, Lei and Candès (2021) for conformal predictions, and Ma et al. (2023) for the kernel ridge regression under the covariate shift. However, as will be revealed shortly, the DRW method may not be suitable for the inference of the GEEs under the covariate shift. We will propose to modify the DRW moment functions into an orthogonal moment function that is more robust against the estimation error of the density ratio function.

Let $r_0(\boldsymbol{x}) = q_0(\boldsymbol{x})/p_0(\boldsymbol{x})$ be the density ratio of $Q_{\boldsymbol{X}}$ and $P_{\boldsymbol{X}}$. Using $r_0(\boldsymbol{X})$ to weigh for $\boldsymbol{Z}$ from the source population, it holds that

$$\mathbb{E}_P\{r_0(\boldsymbol{X})\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\} = \mathbb{E}_Q\left\{\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\},$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The above relations reveal the central role of the density ratio function in the identification of $\boldsymbol{\theta}_0$ using the fully observed source sample. With a consistent $\widehat{r}(\boldsymbol{x})$, we can obtain an estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ from the following density ratio weighting (DRW) moment function

$$\tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta},\widehat{r}) = \widehat{r}(\boldsymbol{X}_i)\mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta}) \quad \text{for } i = 1,\ldots,n, \tag{3.1}$$

with either the empirical likelihood or the generalized method of moments approach.

While being the most popular and natural strategy for tackling the covariate shift, the DRW method for the GEE problem has several drawbacks. First, the accuracy of $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ crucially depends on that of $\widehat{r}$, which may not be high quality when r_0 has a complex structure or the model of r_0 is misspecified. A more important problem arises in the inference as the estimation error of $\widehat{r}$ may have a first-order effect on $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$. This is because the asymptotic distribution of $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ depends on that of the partial sum $n^{-\frac{1}{2}}\sum_{i=1}^n \tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta}_0,\widehat{r})$, which can be decomposed as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta}_0,\widehat{r}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta}_0,r_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta}_0)\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\} := T_n + R_n,$$

where T_n is usually asymptotically Gaussian, However, the second term R_n, which gathers effects of the plugged-in estimator $\widehat{r}$, may not have a weak limit, especially when $\widehat{r}$ is obtained from some black-box machine learning methods. As shown in Section 6.1, $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ requires quite strong conditions to be asymptotically normal. Even under such a case, Theorem 6.1 shows that the EL ratio statistic using (3.1) as moment functions has a weighed-χ^2 limiting distribution, whose quantiles require a Bootstrap procedure to approximate. See also Hjort et al. (2009) on the weighted-χ^2 phenomenon associated with the EL with plugged-in nuisance function estimators.

To alleviate the effect of estimation error in the nuisance function $r_0(\boldsymbol{x})$, we opt for adjusting the sample moment function $\tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta},\widehat{r})$ by employing the first-order correction advocated in the semiparametric literature such as Newey (1994). To illustrate the idea, let $\boldsymbol{\mu}(\boldsymbol{\theta}, \widehat{r}) = \mathbb{E}_P\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}, \widehat{r})\}$ and consider the first-order von Mises expansion

$$\boldsymbol{\mu}(\boldsymbol{\theta}, r_0) = \boldsymbol{\mu}(\boldsymbol{\theta}, \widehat{r}) + \int \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta}, \widehat{r}) dP(\boldsymbol{z}) + R_2(\widehat{r}, r_0), \tag{3.2}$$

where $\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta}, \widehat{r})$ is the pathwise derivative of $\boldsymbol{\mu}(\boldsymbol{\theta}, r)$ at $\widehat{r}$ and $R_2(\widehat{r}, r_0)$ is the reminder term. The expansion suggests that $\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta}, \widehat{r})$ represents the plugged-in effect of $\widehat{r}$, and the bias of the weighted moment function $\tilde{\mathbf{g}}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \widehat{r})$, namely $\boldsymbol{\mu}(\boldsymbol{\theta}, \widehat{r}) - \boldsymbol{\mu}(\boldsymbol{\theta}, r_0)$, can be corrected by adding back $\int \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta}, \widehat{r}) dP(\boldsymbol{z})$. As will be shown in Theorem 3.1, the pathwise derivative satisfies

$$\int \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta}, \widehat{r}) dP(\boldsymbol{z}) = \mathbb{E}_Q\{\mathbf{m}_0(\boldsymbol{X}, \boldsymbol{\theta})\} - \mathbb{E}_P\{\mathbf{m}_0(\boldsymbol{X}, \boldsymbol{\theta})\widehat{r}(\boldsymbol{X})\}.$$

With an estimated conditional mean function $\widehat{\mathbf{m}}(\boldsymbol{x}, \boldsymbol{\theta})$ using the source sample $\mathcal{D}_S$, the above quantity can be approximated by $\sum_{i=1}^N \tilde{\boldsymbol{\psi}}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$, where $\boldsymbol{W}_i = (\boldsymbol{Z}_i, \delta_i)$, $\widehat{\boldsymbol{\eta}} = (\widehat{r}, \widehat{\mathbf{m}})$ and

$$\tilde{\boldsymbol{\psi}}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = \frac{\delta_i}{\tau}\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta}) - \frac{1-\delta_i}{1-\tau}\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\widehat{r}(\boldsymbol{X}_i)$$

for $i = 1, \cdots, N$. Adding the adjustment $\tilde{\boldsymbol{\psi}}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$ to the weighted moment function $\tilde{\mathbf{g}}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \widehat{r})$ leads to the following moment function

$$\boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = \frac{1-\delta_i}{1-\tau}\widehat{r}(\boldsymbol{X}_i)\{\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta}) - \widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} + \frac{\delta_i}{\tau}\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta}). \tag{3.3}$$

A direct calculation verifies that $\mathbb{E}\{\boldsymbol{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\} = \mathbf{0}$ at $\boldsymbol{\eta}_0 = (r_0, \mathbf{m}_0)$, which implies that $\boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$ is a valid moment function for identifying $\boldsymbol{\theta}_0$. We now establish the key aspects regarding the proposed moment function. Let F be the mixture of the source and target distributions. To derive the pathwise derivative of the functional $r \mapsto \mathbb{E}_F\{\widetilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r)\}$, let $\{F_\tau, \tau \in [0, 1)\}$ be a collection of regular parametric submodels satisfying $F_0 = F$ and the mean-square differentiability condition (see, e.g., Van der Vaart, 2000). The true nuisance function under the submodel F_τ is denoted as $\boldsymbol{\eta}(F_\tau)$ such that $r(F_\tau)$ is the true covariate density ratio under F_τ

Theorem 3.1. *Under Conditions 1 and 2, the following results hold. (i) For any* $\boldsymbol{\theta} \in \Theta$,

$$\frac{\partial}{\partial \tau}\mathbb{E}_F\{\widetilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\}\Big|_{\tau=0} = \mathbb{E}_F\left\{\boldsymbol{\varphi}(\boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{\eta}_0)S_0(\boldsymbol{W})\right\}, \tag{3.4}$$

where $\boldsymbol{\eta}_0(\boldsymbol{x}, \boldsymbol{\theta}) = (r_0(\boldsymbol{x}), \mathbf{m}_0(\boldsymbol{x}, \boldsymbol{\theta}))$ *and* $\boldsymbol{\varphi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{\delta}{p}\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}) - \frac{1-\delta}{1-p}r(\boldsymbol{x})\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})$.

(ii) Let $\boldsymbol{\Psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \widetilde{\mathbf{g}}(\boldsymbol{w}, \boldsymbol{\theta}, r) + \boldsymbol{\varphi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\eta})$ *or equivalently,*

$$\boldsymbol{\Psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1-\delta}{1-p}r(\boldsymbol{x})\{\mathbf{g}(\boldsymbol{z}, \boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})\} + \frac{\delta}{p}\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}), \tag{3.5}$$

then $\frac{\partial}{\partial\tau}\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0} = \mathbf{0}$.

(iii) For any candidate $\boldsymbol{\eta}(\boldsymbol{x},\boldsymbol{\theta}) = (r(\boldsymbol{x}),\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}))$,

$$\|\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\}\|_1 \leqslant \|r(\boldsymbol{X}) - r_0(\boldsymbol{X})\|_{L_2(P_{\boldsymbol{X}})}\Big(\sum_{j=1}^{r}\|m_j(\boldsymbol{X},\boldsymbol{\theta}) - m_{0j}(\boldsymbol{X},\boldsymbol{\theta})\|_{L_2(P_{\boldsymbol{X}})}\Big).$$

Theorem 3.1 (i) shows the pathwise derivative function of $\mathbb{E}_F\{\widetilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}$ is $\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta})$, reflecting the local effect the density ratio $r(F_\tau)$ on $\mathbb{E}_F\{\widetilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}$. The property $\frac{\partial}{\partial\tau}\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0} = \mathbf{0}$ in Theorem 3.1 (ii) is the so-called Neyman orthogonality (Neyman, 1959, Chernozhukov et al., 2018), which means the proposed sample moment function $\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta})$ is orthogonal to the nuisance functions. Based on such a property, perturbing the nuisance function $\boldsymbol{\eta}$ locally around $\boldsymbol{\eta}_0$ does not have the first-order effect on $\mathbb{E}\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\}$. The Neyman orthogonality is an important notion in semi-parametric inference as it enables the estimating function to be locally insensitive to the nuisance function. Compared with the debiased machine learning (DML) proposed by Chernozhukov et al. (2018) which also utilized Neyman orthogonal moments, the problem considered here is more challenging, as $\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta})$ is parameter-dependent. Theorem 3.1 (iii) reveals that the bias of the moment functions $\mathbb{E}\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\}$ is bounded by the *product* of the L_2 norms of the estimation errors of the two nuisance functions, which is related to the double robustness property introduced by Robins et al. (1994) for the augmented inverse probability weighting (AIPW) estimator.

4 Estimation of nuisance functions

Given the important roles played by the two nuisance functions $r(\boldsymbol{x})$ and $\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta})$, this section proposes estimators of the two nuisance functions and discuss their theoretical properties.

4.1 Density ratio estimation

We first present estimators to the density ratio r. Conventional approaches, such as the kernel smoothing or the classification-based methods, typically estimate the density functions of the target domain (numerator) and the source domain (denominator), respectively, to form the ratio estimator. However, such density ratio estimators can be quite unstable when the dimension is large or the denominator density is close to zero. We take an approach that *directly* estimate the density ratio based on the dual characteristic of the ϕ-divergence, which can be solved via an empirical risk minimization problem and can accommodates a variety of machine learning algorithms.

For any two distributions P and Q with densities p_0 and q_0 and suppose that P is absolutely continuous with respect to Q, their ϕ-divergence is

$$D_\phi(Q\|P) = \int_{\mathcal{X}} \phi\left(\frac{q_0(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right) p_0(\boldsymbol{x}) d\boldsymbol{x}, \tag{4.1}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a convex and lower semicontinuous function. Different choices of ϕ lead to different divergences, such as the KL divergence for $\phi(u) = u \log u$, the squared Hellinger distance for $\phi(u) = (\sqrt{u}-1)^2$, and the Pearson's χ^2-divergence for $\phi(u) = (u-1)^2$. See Sason and Verdú (2016) for more examples. Moreover, the class of the Cressie-Read power divergence can be represented as ϕ-divergences as shown in Maji et al. (2019).

Let $\phi_*(v) = \sup_{u\in\mathbb{R}}\{uv - \phi(u)\}$ be the Fenchel dual function of ϕ. The dual representation theorem (Rockafellar, 1997) implies that

$$\begin{aligned} D_\phi(Q\|P) &= \sup_{f:\mathcal{X}\to\mathrm{Dom}(\phi_*)} \{\mathbb{E}_Q(f) - \mathbb{E}_P(\phi_*(f))\} \\ &= \mathbb{E}_Q(f_0) - \mathbb{E}_P(\phi_*(f_0)), \end{aligned} \tag{4.2}$$

where $\mathrm{Dom}(\phi_*)$ denotes the domain of ϕ_*, and the supreme is attained at $f_0(\boldsymbol{x}) = \phi'\left(\frac{q_0(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right) = \phi'(r_0(\boldsymbol{x}))$. For each ϕ-function, we define

$$\ell_{1,\phi}(r) = \phi_*\{\phi'(r)\} \quad \text{and} \quad \ell_{2,\phi}(r) = \phi'(r). \tag{4.3}$$

Then the relationship (4.2) induces an identification condition for the density ratio r_0 as presented in the following lemma.

Lemma 4.1. *For any convex and lower semicontinuous function $\phi : \mathbb{R}_+ \to \mathbb{R}$, the true density ratio satisfies*

$$r_0 = \arg\min_{r\in\mathcal{F}} L_\phi(r) \quad \textit{with} \quad L_\phi(r) = \mathbb{E}_P\{\ell_{1,\phi}(r)\} - \mathbb{E}_Q\{\ell_{2,\phi}(r)\}, \tag{4.4}$$

where the candidate class $\mathcal{F}$ is any class of nonnegative functions that contains r_0.

The proof of the above lemma is presented in the supplementary material (SM). For each given ϕ function, r_0 can be uniquely determined from the population objective function (4.4). Table 1 lists some examples of commonly used divergence, along with the corresponding Fenchel conjugate function ϕ_* and the objective functions $\ell_{1,\phi}$ and $\ell_{2,\phi}$.

With the two samples from P and Q, the density ratio r_0 can be estimated with the sample objective function obtained by replacing the expectations in (4.4) with the corresponding empirical averages. The function class $\mathcal{F}$ in (4.4) is required to contain the true density ratio r_0, whose functional form is generally unknown in practice. As a practical

Table 1. Examples of ϕ-divergence, the associated Fenchel conjugate and the objective functions.

Divergence	$\phi(u)$	$\phi_*(v)$	$\ell_{1,\phi}(r)$	$\ell_{2,\phi}(r)$
Kullback-Leibler	$u\log(u)$	$\exp(v-1)$	r	$\log(r)+1$
Reverse KL	$-\log(u)$	$-1-\log(-v)$	$\log(r)+1$	$-r^{-1}$
Pearson χ^2	$(u-1)^2$	$v^2/4+v$	r^2-1	$2(r-1)$
Squared Hellinger	$(\sqrt{u}-1)^2$	$v/(v-1)$	$r^{\frac{1}{2}}-1$	$1-r^{-\frac{1}{2}}$

surrogate of $\mathcal{F}$, we use a candidate class $\mathcal{F}_N$ which may not exactly contain r_0 but has the universal approximation ability as described in Condition 4. Such a requirement can be satisfied by the linear sieves and numerous machine learning methods. The density ratio estimator $\widehat{r}$ is given by

$$\widehat{r} = \arg\min_{r\in\mathcal{F}_N}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell_{1,\phi}\{r(\boldsymbol{X}_i)\} - \frac{1}{m}\sum_{i=n+1}^{N}\ell_{2,\phi}\{r(\boldsymbol{X}_i)\}\right\}. \tag{4.5}$$

It is noted that we not only obtain the estimator $\widehat{r}$, but also an estimate of the divergence $D_\phi(Q\|P)$ by the sample objective function with $\widehat{r}$, which reveals the discrepancy of the source and the target populations. The procedure applies to any ϕ-divergence introduced by different choices of ϕ. For example, choosing $\phi(u) = u\log(u)$ corresponds to the KL-divergence in Table 1 such that

$$\widehat{r} = \arg\min_{r\in\mathcal{F}_N}\left\{\frac{1}{n}\sum_{i=1}^{n}r(\boldsymbol{X}_i) - \frac{1}{m}\sum_{i=n+1}^{N}\log\{r(\boldsymbol{X}_i)\}\right\}. \tag{4.6}$$

Since we have formulated the estimation for r_0 into an empirical risk minimization problem, a variety of computational methods for the optimization can be applied. If the candidate function space $\mathcal{F}_N$ is convex, then problem (4.6) is a convex programming problem, as was demonstrated in Nguyen et al. (2010) for $\mathcal{F}_N$ being the reproducing kernel Hilbert space (RKHS). For more general nonparametric function classes such as the deep neural networks (DNN), the optimization can be conducted via efficient computational algorithms such as the stochastic gradient descent. We advocate the use of the DNNs in the scenario of large-scale data, since the DNNs are more amenable to parallel computations (Goodfellow et al., 2016). Furthermore, compared with the RKHS employed in Nguyen et al. (2010), the DNNs also enjoyed the advantage of adaptivity to unknown low-dimensional structures of the underlying function, thus mitigating the curse of dimensionality, as will be discussed in Section 5.2.

Now we study the L_2-estimation error of the proposed density ratio estimator by first presenting the results for the Hölder function class.

Definition 1. *Let* $\beta = \lfloor \beta \rfloor + r > 0, r \in (0, 1]$ *where* $\lfloor \beta \rfloor$ *denotes the largest integer strictly smaller than* β*. For a finite constant* $B > 0$ *and a compact region* $\mathcal{X} \subset \mathbb{R}^d$*, the Hölder function class*

$$\mathcal{H}^{\beta}(\mathcal{X}, B) = \left\{ f : \mathcal{X} \to \mathbb{R}, \max_{\|\boldsymbol{\alpha}\|_1 \leqslant \lfloor \beta \rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_\infty \leqslant B, \max_{\|\boldsymbol{\alpha}\|_1 = \lfloor \beta \rfloor} \sup_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in \mathcal{X}} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_1) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_2)}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^r} \leqslant B \right\},$$

where $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$ *with* $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)^{\mathrm{T}} \in \mathbb{N}^d$ *and* $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^d \alpha_i$*.*

Condition 3. There exist constants $B_1 > 0$ and $\beta_1 \geqslant 1$ such that the target function $r_0 \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$.

Condition 4. Let the pseudo-dimension (see Pollard, 1990) of $\mathcal{F}_N$ be $\mathrm{Pdim}(\mathcal{F}_N)$, then (i) $\mathrm{Pdim}(\mathcal{F}_N) \log(N) = o(N)$; and (ii) there exists a constant $c_2 > 0$ such that for large enough n, $\inf_{r \in \mathcal{F}_N} \|r - r_0\|_\infty \leqslant c_2 \mathrm{Pdim}(\mathcal{F}_N)^{-\frac{\beta_1}{d}}$. (iii) There exists a positive constant M_1 such that $\|r\|_\infty \leqslant M_1$ and $\|\ell''_{i,\phi}(r)\|_\infty \leqslant M_1$ for $i = 1, 2$ and for every $r \in \mathcal{F}_N$.

The above two conditions are imposed on the true density ratio r_0 and the candidate function class $\mathcal{F}_N$, respectively. Condition 3 characterizes the smoothness of r_0, as commonly imposed in nonparametric function estimation. Condition 4 restricts the complexity of $\mathcal{F}_N$ and assumes the approximation error $\inf_{r \in \mathcal{F}_N} \|r - r_0\|_\infty$ converges to 0 with the increase of the pseudo-dimension $\mathrm{Pdim}(\mathcal{F}_N)$. Such a condition can be satisfied by various nonparametric function classes, including the linear sieves (Chen, 2007), such as the splines and the wavelets, and also the many machine learning methods, for example, the deep neural networks (Jiao et al., 2023). Condition 4 (iii) ensures that every function r in $\mathcal{F}_N$ as well as the second derivative of $\ell_{i,\phi}(r)$ are bounded by M_1, which can be practically achieved by a truncation operation. With the above conditions, we have the following result for the estimation error of the proposed estimator $\hat{r}$. To quantity the estimation performance, we define empirical L_2 error of $\widehat{r}$ as

$$\mathcal{E}_N(\widehat{r}) = [N^{-1} \sum_{i=1}^N \{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2]^{1/2}. \tag{4.7}$$

Theorem 4.1. *Under Conditions 1, 3, and 4, there exists a positive constant* C_1 *such that with probability at least* $1 - 2e^{-t}$*, for* N *large enough and any* $t > 0$*,*

$$\mathcal{E}_N(\widehat{r}) \leqslant C_1 \left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N) \log(N)}{N}} + \inf_{r \in \mathcal{F}_N} \|r - r_0\|_\infty + \sqrt{\frac{t}{N}} \right). \tag{4.8}$$

The theorem provides the non-asymptotic estimation error bound for $\widehat{r}$. The proof of the theorem is presented in Section B.2 of the SM, which is built on a scale-sensitive localization theory (Koltchinskii, 2011) to derive the tight bounds on the estimation errors.

In the proof of Theorem 4.1, we also show that the population L_2 error of $\widehat{r}$ can be bounded by half of the right-hand side of (4.8) with high probability, as a consequence of Talagrand's concentration The first two terms of the bounds in (4.8) correspond to the stochastic error determined by the complexity $\mathrm{Pdim}(\mathcal{F}_N)$ and the approximation error $\inf_{r\in\mathcal{F}_N}\|r-r_0\|_\infty$, respectively. Under Condition 4 (ii), the second term can be bounded by $\mathrm{Pdim}(\mathcal{F}_N)^{-\frac{\beta_1}{d}}$. Therefore, there is a trade-off between the first two terms on the right-hand side of (4.8) with respect to the increase of the complexity of the candidate class $\mathcal{F}_N$. Balancing the first two terms, it can be seen that $\mathrm{Pdim}(\mathcal{F}_N) = O(N^{-\frac{d}{2\beta_1+d}})$ is the optimal choice of pseudo-dimension up to some $\log(N)$ factor. In practice, while the underlying smoothness β_1 is generally unknown, we can specify the optimal $\mathrm{Pdim}(\mathcal{F}_N)$ with the cross-validation method. With such the optimal specification of $\mathrm{Pdim}(\mathcal{F}_N)$, the following convergence rate of $\widehat{r}$ can be obtained.

Corollary 4.1. *Under Conditions 1, 3, and 4, and taking* $\mathrm{Pdim}(\mathcal{F}_N) = O(N^{-\frac{d}{2\beta_1+d}})$, *we have*

$$\mathcal{E}_N(\widehat{r}) = O_p\left(N^{-\frac{\beta_1}{2\beta_1+d}}\log^{\frac{1}{2}}(N)\right).$$

Theorem 4.1 establishes the convergence rate of the proposed density ratio estimator. We note that the $N^{-\frac{\beta_1}{2\beta_1+d}}$ is the minimax lower bound for the density estimation problem as shown in Stone (1982) and Yang and Barron (1999). However, the density ratio estimation is a harder problem than the density estimation as the former is a two-sample problem. In the next theorem, the minimax lower bound for the density ratio estimation is derived.

Theorem 4.2. *Let* $\mathcal{M}^d(\beta_1, B_1) = \left\{(\mathbb{P},\mathbb{Q}) : d\mathbb{Q}/d\mathbb{P} = r_0 \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)\right\}$, *then there exists a positive constant* c_1 *such that*

$$\inf_{\tilde{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}^d(\beta,B)} \mathbb{E}\{\|r-r_0\|_{L_2(P)}\} \geqslant c_1 N^{-\frac{\beta_1}{2\beta_1+d}},$$

for large enough N, *where the infimum is taken over all estimators.*

The above theorem indicates that the minimax lower bound for the density ratio estimation is the same as that for the density estimation. It is worth noting that the convergence rate provided in Theorem 4.1 matches the lower bound up to a $\log(N)$ factor, meaning that the proposed estimator nearly attains the minimax bound. Moreover, as will be discussed in Section 5.2, if the true function r_0 has a low-dimensional support, then the estimation error of $\widehat{r}$ estimated with the DNNs can adaptively achieve a faster convergence rate depending on the low-dimensional structure instead of the nominal dimension d, thus alleviating the curse of dimensionality.

4.2 Conditional density estimation and multiple imputation

The goal is to estimate the conditional moment $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})|\boldsymbol{X}\}$. For a given $\boldsymbol{\theta}$, it can be simply estimated by regressing $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})$ on $\boldsymbol{X}$. However, it has to be conducted repeatedly in search of the optimal solution. In this section, we propose a multiple imputation method to bypass the issue.

To present the multiple imputation procedure, we first need to estimate the conditional density function $p_{Y|\boldsymbol{X}}$, then using it to conduct imputations for the responses of both the source and the target samples. The conditional density estimation is a conventional topic in statistics, whose development includes the kernel density estimation (Hall et al., 2004), the nearest neighbor (Li et al., 2022a), and the regression methods (Izbicki and Lee, 2017). However, the existing methods are mostly restricted to certain nonparametric forms, such as the kernel or the orthogonal basis functions. We propose a new scheme for conditional density estimation which is flexible enough to accommodate a wide range of nonparametric methods.

We note that the conditional density function is essentially a density ratio between the joint density $p_0(y, \boldsymbol{x})$ over the marginal density $p_0(\boldsymbol{x})$. However, the ϕ-divergence based density ratio estimation method described in Section 4.1 requires the support of the denominator density covers that of the numerator density to ensure the ϕ-divergence is well defined. For this reason, we express the conditional density as

$$p_0(y|\boldsymbol{x}) = \frac{p_0(y, \boldsymbol{x})}{p_0(\boldsymbol{x})} = \frac{p_0(y, \boldsymbol{x})}{p_0(\boldsymbol{x})\tilde{p}_0(y)}\tilde{p}_0(y) =: \tilde{r}_0(y, \boldsymbol{x})\tilde{p}_0(y), \tag{4.9}$$

where $\tilde{r}_0(y, \boldsymbol{x})$ is an auxiliary density ratio function between the source population $P_{\boldsymbol{X},Y}$ and an auxiliary population $\tilde{P}_{\boldsymbol{X},\tilde{Y}} = P_{\boldsymbol{X}} \times \tilde{P}_Y$, where $\tilde{P}_Y$ is supported on $\mathbb{R}$ with a known density $\tilde{p}_0(y)$. Such a transformation ensures that $\tilde{P}_{\boldsymbol{X},\tilde{Y}}$ is absolutely continuous with respect to the the source distribution $P_{\boldsymbol{X},Y}$. Hence, the ϕ-divergence based approach for estimating the auxillary density ratio $\tilde{r}_0(y, \boldsymbol{x})$ can be applied. Specifically, let $\mathcal{G}_N$ be a $(d+1)$-dimensional candidate function class that satisfies Condition 6 below, then the density ratio $\tilde{r}_0(y, \boldsymbol{x})$ can be estimated via the following sample criterion

$$\widehat{\tilde{r}}(y, \boldsymbol{x}) = \underset{p \in \mathcal{G}_N}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell_{1,\phi}\{p(\tilde{Y}_i, \boldsymbol{X}_i)\} - \frac{1}{n}\sum_{i=1}^{n} \ell_{2,\phi}\{p(Y_i, \boldsymbol{X}_i)\} \right\}, \tag{4.10}$$

where $\{\tilde{Y}_i\}_{i=1}^n$ are independently sampled from $\tilde{P}_Y$ and are independent of $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$. With $\widehat{\tilde{r}}(y, \boldsymbol{x})$, the conditional density is estimated by

$$\widehat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \widehat{\tilde{r}}(y, \boldsymbol{x})\tilde{p}_0(y). \tag{4.11}$$

This facilitates the multiple imputation of Wang and Chen (2009) for the estimation of $\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})$.

Using the conditional density estimator $\widehat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$, for any $\boldsymbol{X}_i \in \{\boldsymbol{X}_l\}_{l=1}^N$, we generate a sample $\{\tilde{Y}_i^\nu\}_{\nu=1}^\kappa$ independently from $\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)$ as advocated by the multiple nonparametric imputation of Wang and Chen (2009). Then, the imputed moment function is

$$\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}_i, \boldsymbol{\theta}) = \frac{1}{\kappa}\sum_{\nu=1}^{\kappa} \mathbf{g}(\boldsymbol{X}_i, \tilde{Y}_i^\nu, \boldsymbol{\theta}).$$

The most prominent advantage of such an imputation-based estimator is that it does not depend on any particular $\boldsymbol{\theta}$, and is in sharp contrast to the regression approach which shall regress $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})$ on each $(\boldsymbol{X}, \boldsymbol{\theta})$. As in Wang and Chen (2009), it requires that $\kappa \to \infty$ as $N \to \infty$ to attain the best efficiency. To establish the convergence rate of $\widehat{p}(y|\boldsymbol{x})$, the following conditions are required.

Condition 5. (i) The support of $\tilde{P}_Y$ covers that of P_Y, and (ii) the density function of $\tilde{P}_Y$ is uniformly bounded. (iii) There exist constants $B_2 > 0$ and $\beta_2 \geqslant 1$ such that the true conditional density function $p_{Y|\boldsymbol{X}} \in \mathcal{H}^{\beta_2}(\mathcal{Y} \times \mathcal{X}, B_2)$. (iv) $\inf_{y\in\mathcal{Y},\boldsymbol{x}\in\mathcal{X}} p_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) > 0$.

Condition 6. The pseudo-dimension of $\mathcal{G}_N$ satisfies (i) $\mathrm{Pdim}(\mathcal{G}_N)\log(N) = o(N)$, and (ii) there exists a constant $c_3 > 0$ such that for large enough n, $\inf_{p\in\mathcal{G}_N} \|p - p_{Y|\boldsymbol{X}}\|_\infty \leqslant c_3\mathrm{Pdim}(\mathcal{G}_N)^{-\frac{\beta_2}{d+1}}$. (iii) There exists a positive constant M_2 such that $\|p\|_\infty \leqslant M_2$ and $\|\ell''_{i,\phi}(p)\|_\infty \leqslant M_2$ for $i = 1, 2$ for every $p \in \mathcal{G}_N$.

Condition 7. There exists a positive constant $\sigma_g > 0$ such that $\mathbb{E}\{\exp(\lambda\|\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}\|^2)|\boldsymbol{X} = \boldsymbol{x}\} < \exp(\lambda\sigma_g^2)$ for all $0 \leqslant \lambda \leqslant \sigma_g^{-2}$ for each $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{x} \in \mathcal{X}$.

In the above conditions, Conditions 5 (i)-(ii) are regularity conditions for the auxiliary distribution $\tilde{P}_Y$. Conditions 5 (iii)-(iv) and Condition 6 are in analog to Conditions 3 and 4 for the density ratio estimation, respectively, requiring that the true conditional density function has β_2-smoothness and the candidate function class has a sufficient approximation ability. Condition 7 assumes that $\|\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})\|$ is sub-Gaussian conditional on $\boldsymbol{X}$. Though such a condition is not required in classic GEE literature, it is required here since we need to conduct nonparametric estimation for its conditional mean function.

To present the result on the convergence of multiple imputation estimator $\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta})$, similar to (4.7) of $\widehat{r}$, we define the empirical L_2 error of $\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta})$ as

$$\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = \sum_{j=1}^{r}[N^{-1}\sum_{i=1}^{N}\{\widehat{m}_{\kappa j}(\boldsymbol{X}_i, \boldsymbol{\theta}) - m_{0j}(\boldsymbol{X}_i, \boldsymbol{\theta})\}^2]^{1/2}, \tag{4.12}$$

where $\widehat{m}_{\kappa j}$ and m_{0j} are the j-th component of $\widehat{\mathbf{m}}_\kappa$ and $\mathbf{m}_0$, respectively.

Theorem 4.3. *Under Conditions 1, 5–7 and taking* $\mathrm{Pdim}(\mathcal{G}_N) = O(N^{-\frac{d+1}{2\beta_2+d+1}})$ *and* $\kappa \gtrsim N$*, for any* $\boldsymbol{\theta} \in \Theta$*,*

$$\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}})\} = O_p\left(N^{-\frac{\beta_2}{2\beta_2+d+1}}\log^{\frac{3}{2}}(N)\right).$$

The proof of the above theorem is similar to that of Theorem 4.1 and is given in Section B.2 of the SM. It is known from Yang and Barron (1999) that the $N^{-\frac{\beta_2}{2\beta_2+d+1}}$ rate matches the minimax lower bound for the $(d+1)$-dimensional conditional mean estimation problem. Hence, Theorem 4.3 shows that multiple imputation estimator $\widehat{\mathbf{m}}_\kappa(\boldsymbol{X},\boldsymbol{\theta})$ have the merit of being rate optimal up to the $\log(N)$ factor, while conveniently avoiding conducting infinitely many regressions at all possible $\boldsymbol{\theta}$. The effect of d in the above rate reveals the curse of dimensionality. However, the accommodation of flexible uses of modern machine learning algorithms in the proposed method provides the opportunity to improve the convergence rate since the low-dimensional structure, if the underlying distribution indeed posits, can be adaptively learned by the DNNs, as will be shown in 5.3.

5 Empirical likelihood inference

Using the orthogonal moment function $\boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})$ with $\widehat{\boldsymbol{\eta}}(\boldsymbol{X}_i,\boldsymbol{\theta}) = (\widehat{r}(\boldsymbol{X}_i),\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}_i,\boldsymbol{\theta}))$ estimated by the methods in Section 4.1 and 4.2, respectively, the EL estimator of $\boldsymbol{\theta}_0$ is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} L_N(\boldsymbol{\theta}) \tag{5.1}$$

where $L_N(\boldsymbol{\theta})$ is the profile EL

$$L_N(\boldsymbol{\theta}) = \sup\left\{\prod_{i=1}^N p_i \;\middle|\; p_i \geqslant 0,\ \sum_{i=1}^N p_i = 1,\ \sum_{i=1}^N p_i \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}(\boldsymbol{X}_i,\boldsymbol{\theta})) = \mathbf{0}\right\}. \tag{5.2}$$

In the following Section 5.1, we will investigate the asymptotic distribution of the EL estimator $\widehat{\boldsymbol{\theta}}$ and the inference for $\boldsymbol{\theta}_0$. In Section 5.2, we will discuss the scenario where the covariate dimension d and the parameter dimension p are allowed to grow the the increase of the sample size.

5.1 Asymptotic results for the EL inference

In this part, we discuss the large sample properties of the EL-based estimator $\widehat{\boldsymbol{\theta}}$ and then propose confidence regions for $\boldsymbol{\theta}_0$ based on the EL ratio. To present the result, we define $\boldsymbol{\Gamma} = \mathbb{E}\{\partial\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)/\partial\boldsymbol{\theta}\}$, $\boldsymbol{\Omega} = \mathbb{E}\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)^{\otimes 2}\}$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}$. The empirical estimation errors $\mathcal{E}_N(\widehat{r})$ and $\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}})$ are defined in (4.7) and (4.12), respectively.

Theorem 5.1. *Under Conditions 1 and 2, if the estimation errors satisfy*

$$\mathcal{E}_N(\widehat{r}) + \mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(1) \quad \textit{and} \quad \mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(N^{-\frac{1}{2}}), \tag{5.3}$$

for every $\boldsymbol{\theta} \in \Theta$*, then we have*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{5.4}$$

In this theorem, the requirement for the nuisance function estimation is only via their estimation errors (5.3). Specifically, under Conditions 3–7 where r_0 and $p_{Y|\boldsymbol{X}}$ have the smoothness of β_1 and β_2, respectively, then (5.3) is attainable provided that

$$\frac{\beta_1}{2\beta_1 + d} + \frac{\beta_2}{2\beta_2 + d + 1} > \frac{1}{2}, \tag{5.5}$$

using the proposed divergence-based density ratio estimator $\widehat{r}$ and the multiple-imputation estimator $\widehat{\mathbf{m}}_\kappa$ whose convergence rates are established in Theorem 4.1 and 4.3. It is remarkable that the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ reaches the semiparametric efficiency bound established in Chen et al. (2008) for problem (2.1), meaning that it has the optimal variance among the family of unbiased estimators for $\boldsymbol{\theta}_0$. Compared with the estimators in Chen et al. (2008), the proposed method accommodates more flexible uses of the ML methods for the nuisance function estimation and requires milder conditions to achieve the asymptotic normality in (5.4), as will be further discussed in Section 6.1.

Remark 1. Different from the cross-fitting adopted by Chernozhukov et al. (2018) and Kallus et al. (2024) among many others in the recent literature of semiparametric inference with machine learning methods, our theoretical results do not necessarily require the sample splitting procedure. It is noted that the sample splitting may alleviate potential overfitting problems and under some conditions may lead to faster convergence of the reminder terms as shown in Newey and Robins (2018). However, due to the heavy computational cost of the cross-fitting, we do not consider such a procedure in this study. Moreover, the reduced sample size caused by the sample splitting may deteriorate the empirical performance of the ML-based estimation, especially when the original sample size is not sufficiently large. Detailed comparisons for the proposed whole sample and the cross-fitting methods are of future interest.

We next consider the inference for $\boldsymbol{\theta}_0$. Let the log EL ratio be $\ell_N(\boldsymbol{\theta}) = -\log\{L_N(\boldsymbol{\theta})/N^{-N}\}$ for every $\boldsymbol{\theta} \in \Theta$, and let $R_N(\boldsymbol{\theta}_0) = 2\ell_N(\boldsymbol{\theta}_0) - 2\ell_N(\widehat{\boldsymbol{\theta}})$. The next theorem shows that the $R_N(\boldsymbol{\theta}_0)$ converges to a standard χ^2 distribution.

Theorem 5.2. *Under the same conditions as in Theorem 5.1, as* $N \to \infty$*,*

$$R_N(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_r.$$

The central χ^2 distribution in Theorem 5.2 brings convenience for the inference of $\boldsymbol{\theta}$. Different from other methods such as the Wald-type inference and the GMM, we do not require the estimation of the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ due to the self-normalization of the EL. Theorem 5.2 is often referred to as Wilks's theorem, as one of the most prominent benefits of the EL-based inference for the GEEs (Qin and Lawless, 1994). However, with the presence of nuisance functions, the log EL ratio no longer necessarily converges weakly to a central χ^2 distribution but may be a weighted sum of χ^2 distributions, whose critical values require Bootstrap to approximate, as demonstrated in Wang and Chen (2009) and Hjort et al. (2009). Due to the orthogonal estimating function, our method overcomes such a situation and restores Wilks's theorem of the log EL ratio, despite the involvement of two nuisance functions.

5.2 Circumventing the curse of dimensionality

In various modern scientific tasks, the dimension d of the covariate can be very large. In this part, we consider the inference for $\boldsymbol{\theta}$ with the presence of a high dimensional covariate. It is known that the increase in dimensionality deteriorates the convergence rates of estimators (Stone, 1982). Recently, it has been investigated that the DNNs can adaptively approximate high-dimensional functions with low-dimensional structures (Jiao et al., 2023). There have been increasing studies indicating that high-dimensional data tend to be supported on some low-dimensional manifolds in many applications, such as image analysis and natural language processing (Goodfellow et al., 2016). Therefore, we consider the following approximate manifold support condition.

Condition 8 (Approximate manifold support)**.** The covariate distributions $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ are concentrated on $\mathcal{M}_\rho$, a ρ-neighborhood of $\mathcal{M} \subset \mathcal{X}$, where $\mathcal{M}$ is a compact $d_{\mathcal{M}}$-dimensional Riemannian manifold (Lee, 2006) and $\mathcal{M}_\rho = \{\boldsymbol{x} \in \mathcal{X} : \inf\{\|\boldsymbol{x}-\boldsymbol{y}\|_2 : \boldsymbol{y} \in \mathcal{M}\} \leqslant \rho\},\ \rho \in (0,1)$.

In the above condition, the dimension $d_{\mathcal{M}}$ of the manifold $\mathcal{M}$ can be regarded as an intrinsic dimension of the covariate. Throughout this section, we allow the nominal dimension d to diverge with the sample size, while taking the intrinsic dimension $d_{\mathcal{M}}$ as a fixed constant. Jiao et al. (2023) established that the fully connected DNNs can adaptively estimate a smooth function with the manifold assumption, hence alleviating the curse of dimensionality. Motivated by the development, we choose the function classes $\mathcal{F}_N$ in the density ratio and $\mathcal{G}_N$ in conditional density estimation as the DNNs with the ReLU activation function. The widths for $\mathcal{F}_N$ and $\mathcal{G}_N$ are specified as W_1 and W_2, and the depths are specified as D_1 and D_2, respectively. Let $\tilde{d}_{\mathcal{M}} = O(d_{\mathcal{M}} \log(d/\delta)/\delta^2)$ be an integer such that $d_{\mathcal{M}} \leqslant \tilde{d}_{\mathcal{M}} < d$, where $\delta \in (0,1)$ is a given constant. The following theorem gives the convergence rate of the DNN-based estimation of the nuisance functions under Condition 8.

Theorem 5.3. *Under Conditions 3–8, let the widths and depths of $\mathcal{F}_N$ and $\mathcal{G}_N$ be*

$W_i = 114(\lfloor\beta_i\rfloor + 1)^2 \tilde{d}_{\mathcal{M}}^{\lfloor\beta_i\rfloor+1}$ *and* $D_i = 21(\lfloor\beta_i\rfloor + 1)^2 N^{\tilde{d}_{\mathcal{M}}/2(\tilde{d}_{\mathcal{M}}+2\beta_i)}\lceil\log_2(8N^{\tilde{d}_{\mathcal{M}}/2(\tilde{d}_{\mathcal{M}}+2\beta_i)})\rceil$,

for $i = 1$ and 2. Then, the estimation errors of $\hat{r}$ and $\widehat{\mathbf{m}}_{\boldsymbol{\theta}}$ satisfy

$$\begin{aligned} &\mathcal{E}_N(\hat{r}) = O_p\left(d^{\frac{1}{2}} N^{-\frac{\beta_1}{\tilde{d}_{\mathcal{M}}+2\beta_1}} \log^{\frac{1}{2}}(N)\right) \quad \textit{and} \\ &\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = O_p\left((d+1)^{\frac{1}{2}} N^{-\frac{\beta_2}{(\tilde{d}_{\mathcal{M}}+1+2\beta_2)}} \log^{\frac{3}{2}}(N)\right), \quad \textit{respectively.} \end{aligned} \tag{5.6}$$

The above theorem shows that the DNN-based estimation for the nuisance functions is adaptive to the low-dimensional manifold structure, with the convergence rates depending on the intrinsic dimension $\tilde{d}_{\mathcal{M}}$ and a prefactor of the rate $\sqrt{d}$. In comparison, the convergence rates for $\hat{r}$ and $\widehat{\mathbf{m}}_{\boldsymbol{\theta}}$ established in Corollary 4.1 and Theorem 4.3 without the manifold condition are $O_p(N^{-\frac{\beta_1}{d+2\beta_1}})$ and $O_p(N^{-\frac{\beta_1}{d+1+2\beta_1}})$, respectively, up to some $\log(N)$ factors. Therefore, the effect of the dimensionality is substantially mitigated with the adaptivity of the DNN function classes to the underlying low-dimensional manifolds. Compared with classic structural methods that pre-assume some low-dimensional structures, such as the additive models, the DNNs can obtain considerably improved convergence rates and circumvent the curse of dimensionality without the knowledge of the specific low-dimensional function structure.

Remark 2. Aside from the manifold assumption considered above, there are several other low-dimensional structure conditions that the DNNs can be adaptive to. For example, Bauer and Kohler (2019) showed if the underlying function follows the β-smooth generalized hierarchical interaction model of the order $\tilde{d}$, then the estimation error of the sigmoid-activated DNN achieves the order of $O_p(\alpha(d)N^{-\frac{\beta}{\tilde{d}_{\mathcal{M}}+2\beta}})$ for some $\alpha(d)$ depending on d. See also Schmidt-Hieber (2020) for similar results. However, the $\alpha(d)$ may depend exponentially on d, while in (5.6) only the factors of the order $\sqrt{d}$ are involved. Hence, we mainly consider the manifold condition in this study.

We first discuss the inference for the fixed dimensional $\boldsymbol{\theta}$ with the presence of the covariate with a growing dimension d, namely p is a constant while d can increase with the sample size N. Such a scenario corresponds to the parameter depending on Y but not on $\boldsymbol{X}$. The following theorem specifies the regime for $(d, \tilde{d}_{\mathcal{M}}, \beta_1, \beta_2, N)$, where the estimator $\widehat{\boldsymbol{\theta}}$ and the log EL ratio $R_N(\boldsymbol{\theta}_0)$ have the same asymptotic distributions as those in Section 5.1.

Theorem 5.4. *Under Conditions 1–8 and suppose that $d = O(N^k)$ for some $k \geqslant 0$ and*

$$\frac{\beta_1}{2\beta_1 + \tilde{d}_{\mathcal{M}}} + \frac{\beta_2}{2\beta_2 + \tilde{d}_{\mathcal{M}} + 1} > \frac{2+k}{4}, \tag{5.7}$$

then $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ *and* $R_N(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_r$ *as* $N \to \infty$.

Compared with Condition (5.5) under the fixed d and without the manifold structure, the requirement in (5.7) replaces the d factors appeared on the denominators to the intrinsic dimension $\widetilde{d}_{\mathcal{M}}$, which provides the opportunity to allow the nominal dimension d grows with the polynomial rate of N. Chen et al. (2024) also considered a growing dimension scenario and applied the shallow neural network for nuisance function estimation, where the dimension was allowed to increase at the rate $d = o(\sqrt{\log(N)})$.

Next, we consider the case where both the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{X}$ diverge, namely $p, d \to \infty$ as $N \to \infty$, which implies the number of moment restrictions $r \to \infty$ since it is no less than the number of parameters p for the identification. The high dimensional EL without the nuisance functions has been investigated by Chen et al. (2009), Hjort et al. (2009), and Chang et al. (2015). The following extends their results to the covariate shift setting in the presence of high dimensional nuisance functions.

Theorem 5.5. *Under Conditions 1–8 and regime (5.7), if* $r^3p^2N^{-1} = o(1)$ *and* $r^3N^{2/\alpha-1} = o(1)$, *where* $\alpha > 2$ *is the order of moment defined in Condition 2, then as* $r, p, N \to \infty$, *(i) for any* $\boldsymbol{u}_n \in \mathbb{R}^p$ *with unit* L_2*-norm,*

$$\sqrt{N}\boldsymbol{u}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, 1); \tag{5.8}$$

(ii) the EL ratio statistic $R_N(\boldsymbol{\theta}_0)$ *satisfies*

$$(2r)^{-\frac{1}{2}}\{R_N(\boldsymbol{\theta}_0) - r\} \xrightarrow{d} \mathcal{N}(0, 1). \tag{5.9}$$

Although the estimating function involves nuisance functions, the above asymptotic distributions of $\widehat{\boldsymbol{\theta}}$ and $R_N(\boldsymbol{\theta}_0)$ recover those in Chang et al. (2015) in the absence of the nuisance functions, due to the Neyman-orthogonality of the construction for $\boldsymbol{\Psi}$. As established in Theorem 5.5 (i), the normalized EL estimator $\widehat{\boldsymbol{\theta}}$ remains asymptotic normal under $r^3p^2N^{-1} = o(1)$ and $r^3N^{2/\alpha-1} = o(1)$. The asymptotic normality (5.9) for $R_N(\boldsymbol{\theta}_0)$ is a natural substitute for the Wilks' theorem with diverging r, which conveniently facilitates the inference for $\boldsymbol{\theta}_0$.

Remark 3. The above analyses are under the regime where p and r diverge at rates slower than the sample size N. For the ultra high dimensional cases where $p, r \gg N$, one can utilize the penalized EL approach introduced by Chang et al. (2018) and Chang et al. (2021), while imposing sparsity structures on the model parameters.

6 Related methods

In this section, we will discuss the distinctions of our proposed approach to some popular methods in the missing data and causal inference literature.

6.1 Density ratio weighting estimation

In this part, we formally establish the theoretical properties of the density ratio weighting (DRW) estimation briefly discussed in Section 3, which is employed in Chen et al. (2008) and Chen et al. (2024) for the inference of GEEs in missing data problems.

Since the covariate shift setting $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$ is equivalent to the missing at random condition, the GEE problem (2.1) is closely related to that considered in Chen et al. (2008). By the Bayes rule, the density ratio function r_0 can be expressed as

$$r_0(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\delta=0)}{f(\boldsymbol{x}|\delta=1)} = \frac{\mathbb{P}(\delta=1)}{\mathbb{P}(\delta=0)} \frac{\mathbb{P}(\delta=0|\boldsymbol{X}=\boldsymbol{x})}{1-\mathbb{P}(\delta=0|\boldsymbol{X}=\boldsymbol{x})}. \tag{6.1}$$

Let $\pi_0(\boldsymbol{x}) = \mathbb{P}(\delta = 0|\boldsymbol{X} = \boldsymbol{x})$, which is the propensity score function in the missing data literature. Hence, (6.1) reveals that r_0 has a one-to-one correspondence with π_0. Therefore, the DRW estimator is essentially an IPW estimator. The most important advantage of the DRW estimator is that it does not need to estimate the conditional mean function $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta})$. However, unlike most classic IPW estimators, where the propensity score π_0 is estimated using the logistic or the least squares regression, in the density ratio weighting for the covariate shift, we often directly estimate r_0 instead of π_0. Therefore, the results shown for the IPW estimators may not directly hold for the DRW estimator. In the following, we explore whether and when the DRW estimator is as efficient as the proposed method.

Suppose the density ratio estimator $\widehat{r}$ satisfies

$$\widehat{L}_N(\widehat{r}) \leqslant \widehat{L}_N(r) + O_p(\epsilon_N^2), \quad \text{for all } r \in \mathcal{F}_N, \tag{6.2}$$

where $\mathcal{F}_N$ is the function class where the density ratio estimator is chosen from, ϵ_N is a positive sequence satisfying $\epsilon_N = o(N^{-\frac{1}{2}})$, and the objective function $\widehat{L}_N(r)$ is defined as

$$\widehat{L}_N(r) = \frac{1}{n}\sum_{i=1}^{n} \ell_1(\boldsymbol{X}_i, r(\boldsymbol{X}_i)) - \frac{1}{m}\sum_{i=n+1}^{n+m} \ell_2(\boldsymbol{X}_i, r(\boldsymbol{X}_i)),$$

where $\ell_i (i = 1, 2)$ are not necessarily the $\ell_{i,\phi}(i = 1, 2)$ introduced in (4.5).

With an estimated $\widehat{r}(\boldsymbol{x})$, the DRW moment function for the source sample is

$$\mathbf{g}^{\mathrm{drw}}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \widehat{r}) = \widehat{r}(\boldsymbol{X}_i)\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta}) \quad \text{for } i = 1, \ldots, n,$$

from which we can obtain an estimator of $\boldsymbol{\theta}_0$ defined as

$$\widehat{\boldsymbol{\theta}}^{\mathrm{drw}} = \arg\max_{\boldsymbol{\theta}\in\Theta} L_n^{\mathrm{drw}}(\boldsymbol{\theta}, \widehat{r}), \tag{6.3}$$

where $L_n^{\mathrm{drw}}(\boldsymbol{\theta}, \widehat{r})$ is the profile EL ratio

$$L_n^{\mathrm{drw}}(\boldsymbol{\theta}, \widehat{r}) = \sup\left\{\prod_{i=1}^{n} p_i \;\middle|\; p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \mathbf{g}^{\mathrm{drw}}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \widehat{r}) = \mathbf{0}\right\}.$$

For a given $\ell_2(\boldsymbol{x}, r)$, we let $\mathbf{m}_\ell(\boldsymbol{x}) = \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}_0)\{\partial\ell_2(\boldsymbol{x}, r_0)/\partial r\}^{-1}$. The following conditions are required to establish asymptotic properties of $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$.

Condition 9. (i) The objectives $\ell_1(\boldsymbol{x}, r)$ and $\ell_2(\boldsymbol{x}, r)$ are three-times continuously differentiable with respect to r and $\inf_{\boldsymbol{x}\in\mathcal{X}} \partial\ell_i(\boldsymbol{x}, r_0)/\partial r > 0$ for $i = 1, 2$. (ii) The partial derivatives satisfy $\partial\ell_1(\boldsymbol{x}, r)/\partial r = r(\boldsymbol{x})\partial\ell_2(\boldsymbol{x}, r)/\partial r$.

Condition 10. (i) The estimation error of $\widehat{r}$ satisfies $\|\widehat{r} - r_0\|_{L_2(P)} = O_p(\delta_N)$ for some $\delta_n = o(N^{-\frac{1}{4}})$. (ii) The bracketing integral (see Van der Vaart, 2000) of $\mathcal{F}_N$ satisfies $J_{[\,]}(\delta_N, \mathcal{F}_N, L_2(P)) = o(1)$. (iii) For every $i = 1, \cdots p$, there exists some $\tilde{m}_{\ell,j} \in \mathcal{F}_N$ such that $\|m_{\ell,j}(\boldsymbol{X}) - \tilde{m}_{\ell,j}(\boldsymbol{X})\|_{L_2(P)} = o(N^{-\frac{1}{4}})$, where $m_{\ell,j}(\boldsymbol{x})$ is the j-th element of $\mathbf{m}_\ell(\boldsymbol{x})$.

Condition 9 is regarding the requirements of the objective function and ensures r_0 is the solution to the population objective function. Condition 10 collects the assumptions for $\widehat{r}$ and the function class $\mathcal{F}_N$ to which $\widehat{r}$ belongs. Specifically, Condition 10 (i) requires the L_2-estimation error of $\widehat{r}$ to be $o_p(N^{-\frac{1}{4}})$. In comparison, our result only requires the *product* of the estimation errors of $\widehat{r}$ and $\widehat{\mathbf{m}}_{\boldsymbol{\theta}}$ to be $o_p(N^{-\frac{1}{2}})$, achieving more robustness against the nuisance function estimation errors. Condition 10 (ii) is a restriction for the complexity of $\mathcal{F}_N$, which is needed for stochastic equicontinuity. Condition 10 (iii) assumes that each element of $\mathbf{m}_\ell(\boldsymbol{x})$ can be approximated sufficiently well by the function class $\mathcal{F}_N$. It is worth noting that such a condition implicitly brings more smoothness for the conditional mean function $\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}_0)$. With the above conditions, the asymptotic distributions of the DRW estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ and the associated log EL ratio statistics $R_n^{\mathrm{drw}}(\boldsymbol{\theta}_0)$ can be derived.

Theorem 6.1. *Under Conditions 1–3, 4. (iii), 9 , and 10, the DRW estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ satisfies*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}^{\mathrm{drw}} - \boldsymbol{\theta}_0) \xrightarrow{d} \boldsymbol{U} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{6.4}$$

and the log EL ratio statistics $R_n^{\mathrm{drw}}(\boldsymbol{\theta}_0) \xrightarrow{d} \boldsymbol{U}^{\mathrm{T}}[\mathbb{E}\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}_0)^{\otimes 2}\}]^{-1}\boldsymbol{U}$.

The theorem reveals that the DRW estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ attains the semiparametric efficiency bound as the proposed estimator. However, it requires more stringent conditions on both the density ratio estimation error and the approximation ability of $\mathcal{F}_N$, which are not necessarily needed by our method. More importantly, the limiting distribution of the log EL ratio $R_n^{\mathrm{drw}}(\boldsymbol{\theta}_0)$ has a weighted χ^2 limiting distribution, since the covariance of $\boldsymbol{U}$ does not match with $\mathbb{E}\{\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta}_0)^{\otimes 2}\}$. Consequently, for the inference of $\boldsymbol{\theta}_0$, the above density ratio weighting method requires a Bootstrap procedure (Chen et al., 2024), which brings considerable computation burden especially when $\widehat{r}$ is estimated with complex ML algorithms such as the DNNs, while our proposed method can conveniently employ the Wilk's theorem of the EL ratio statistics.

6.2 Double machine learning methods

Our work is also closely related to the classical semiparametric estimation literature on constructing asymptotic normal estimators for low dimensional parameters with the presence of infinitely dimensional nuisance functions.

Building upon the Neyman orthogonality condition, our modified moment function shares similar spirits as the class of doubly robust estimators (Robins et al., 1994 and Rotnitzky et al., 2012) and recently proposed double machine learning methods (Chernozhukov et al., 2018). However, this study has the following important distinctions. First, both the doubly robust and the double machine learning literature commonly deal with linear functional estimation, such as the average treatment effect. Under such cases, the nuisances are typically the propensity score function $\pi_0(\boldsymbol{X}) = \mathbb{P}(\delta = 0|\boldsymbol{X})$ and the conditional mean function $m(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$, which both can be easily estimated by solving a regression problem. However, our interested GEE problem is more challenging, due to the presence of the nuisance function $\mathbf{m}(\boldsymbol{X},\boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})|\boldsymbol{X}\}$, which requires estimating for all possible $\boldsymbol{\theta}$. Therefore, our work complements the line of research of the doubly robust and double machine learning methods, by providing an effective approach to handle such parameter-dependent nuisance function. Utilizing the idea of the multiple imputation, we circumvent directly estimate $\mathbf{m}(\boldsymbol{X},\boldsymbol{\theta})$ at infinitely many $\boldsymbol{\theta}$ but only requires the estimation of the conditional density function $p(y|\boldsymbol{X})$. Instead of the conventional kernel smoothings, our novel method for the estimation $p(y|\boldsymbol{X})$ can employ a broad array of machine-learning algorithms. Moreover, the sample splitting procedure required in the DML can be bypassed in this study as discussed in Remark 1.

7 Simulation Study

This section reports the simulation results for the proposed methods, including the density ratio estimation, the conditional density estimation, and the inference for the GEEs.

7.1 Numerical results of density ratio estimation

In this part, we carried out simulations to evaluate the performances of the proposed density ratio estimation and compared it with other popular density ratio estimation methods.

The covariate of source sample $\{\boldsymbol{X}_i\}_{i=1}^n$ and the target sample $\{\boldsymbol{X}_{n+i}\}_{i=1}^{n+m}$ were generated as independent copies of $\boldsymbol{X}^0 = (X_1^0, \ldots, X_d^0)^{\mathrm{T}}$ and $\boldsymbol{X}^1 = (X_1^1, \ldots, X_d^1)^{\mathrm{T}}$, respectively. The sample sizes were chosen within the range $n \in \{1000, 2000, 5000\}$ for the source sample and $m = n/2$ for the target sample, to accommodate the common case where the source sample usually has more observations than the target sample. We considered two settings for the distributions of $\boldsymbol{X}^0$ and $\boldsymbol{X}^1$, corresponding to the compact and uncompact supports, respectively. In Setting S1, $\{X_i^0\}_{i=1}^d$ were independent distributed as Uniform$(0, 1)$, and $\{X_i^1\}_{i=1}^d$ were independent distributed as Beta$(6/5, 6/5)$. In Setting S2 $\boldsymbol{X}^0$ was distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ is the d-dimensional identity matrix, and $\boldsymbol{X}^1$ was distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_d)$, where $\boldsymbol{\Sigma}_d = (\sigma_{i,j})_{d\times d}$ for $\sigma_{i,j} = 0.5^{|i-j|}$. The dimensions were chosen as $d = 5$ and 20, respectively, to evaluate the estimation performances as the increase of the dimension.

For the proposed divergence-based density ratio (DDR) estimation, we chose the ϕ-divergence in (4.1) as the KL-divergence and estimated r_0 by

$$\widehat{r} = \arg\max_{r\in\mathcal{F}_N} \left\{ \frac{1}{m} \sum_{i=n+1}^{n+m} \log\left(r(\boldsymbol{X}_i)\right) - \frac{1}{n} \sum_{i=1}^{n} r(\boldsymbol{X}_i) \right\}.$$

The function class $\mathcal{F}_N$ was chosen as the deep neural network, whose tuning parameters were selected by the three-fold cross-validation. We chose the ReLU function as the activation function and adopted the Adam as the optimization algorithm. For comparison, we also estimated the density ratio function with three commonly used methods: (1) the kernel mean matching (KMM) proposed by Gretton et al. (2009); (2) the kernel smoothing (KS) method that first obtains the kernel smoothing density estimates $\hat{p}(\boldsymbol{x})$ and $\hat{q}(\boldsymbol{x})$ for the source and target distributions, where the bandwidths were selected from the leave-one-out cross validation, then taking their ratio; (3) the probabilistic classification (PC) approach adopted by Lei and Candès (2021), which used the ratio of posterior classification probabilities to estimate the density ratio function. Simulation results were based on 300 repetitions. The performances of the estimators were evaluated by the mean squared error

(MSE) calculated as $n^{-1}\sum_{i=1}^{n}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}^2$.

Table 2. Empirical average of mean squared errors (MSE) of $\widehat{r}$ based on 300 repetitions of the proposed divergence-based density ratio (DDR) estimation, the kernel mean matching (KMM), the kernel smoothing (KS), and the probabilistic classification (PC) methods. The empirical standard deviations of the MSEs are in parentheses.

Setting	n	$d=5$				$d=20$			
		DDR	KKM	KS	PC	DDR	KKM	KS	PC
S1	1000	0.64	0.75	0.86	0.81	0.93	1.24	1.85	1.56
		(0.35)	(0.29)	(0.31)	(0.28)	(0.43)	(0.52)	(0.68)	(0.59)
	2000	0.34	0.44	0.52	0.48	0.58	0.71	1.16	0.92
		(0.19)	(0.15)	(0.20)	(0.16)	(0.30)	(0.25)	(0.36)	(0.24)
	5000	0.22	0.31	0.38	0.34	0.36	0.45	0.81	0.67
		(0.09)	(0.11)	(0.09)	(0.07)	(0.13)	(0.15)	(0.25)	(0.19)
S2	1000	0.74	0.82	0.93	0.85	1.15	1.38	1.94	1.73
		(0.36)	(0.25)	(0.30)	(0.36)	(0.50)	(0.62)	(0.71)	(0.66)
	2000	0.37	0.51	0.65	0.49	0.64	0.78	1.25	1.04
		(0.16)	(0.13)	(0.22)	(0.15)	(0.34)	(0.30)	(0.41)	(0.28)
	5000	0.24	0.37	0.44	0.40	0.43	0.52	0.93	0.79
		(0.08)	(0.09)	(0.13)	(0.12)	(0.20)	(0.19)	(0.28)	(0.17)

As indicated in Table 2, the proposed DDR for the density ratio estimation achieved the best finite sample performances among the two settings. The KMM using the reproducing kernel Hilbert space had the second smallest MSE in most cases, followed by the PC method that used the posterior classification probability to indirectly estimate the density ratio. The conventional kernel smoothing (KS) method had the worst performances among the four candidates. The DDR had significantly improved estimation accuracy over the other methods especially for $d=20$, suggesting the advantage of the proposed method for the large dimensional scenarios. On the other hand, it was noted that the standard deviations of the DDR estimates were relatively large in some cases with small and moderate sample sizes, which became smaller as the sample size were larger.

7.2 Numerical results of conditional density estimation

We also conducted simulations to evaluate the finite sample performance of the proposed conditional density estimation methods in Section 4.2. The estimation was conducted using the source sample $\mathcal{D}_S$, where the covariates $\{\boldsymbol{X}_i\}_{i=1}^n$ generated independently from the Uniform$(0,1)^d$ distribution, and the responses $\{Y_i\}_{i=1}^n$ were generated according to the following three models:

$$\text{(M1):}\quad Y_i = 0.5\sum_{k=1}^{\lfloor d/2\rfloor}\sum_{j=2k}X_{j,i} - 0.5\sum_{k=1}^{\lfloor d/2\rfloor}\sum_{j=2k-1}X_{j,i}+\epsilon_i, \tag{7.1}$$

$$\text{(M2):}\quad Y_i = \sin\left(\pi\sum_{k=1}^{\lfloor d/2\rfloor}\sum_{j=2k}X_{j,i}\right)+\epsilon_i,$$

$$\text{(M3):} \qquad Y_i = \mathbb{1}\left(\sum_{k=1}^{\lfloor d/2 \rfloor}\sum_{j=2k} X_{j,i} < \sum_{k=1}^{\lfloor d/2 \rfloor}\sum_{j=2k-1} X_{j,i}\right) + \epsilon_i, \tag{7.2}$$

where the regression functions are linear, trigonometric, and piecewise constant, respectively, and the noises $\{\epsilon_i\}_{i=1}^n$ were independently distributed as $\mathcal{N}(0, \sigma_X^2)$, where $\sigma_X^2 = \max(0.5, |X_{i,1}|)$. The dimensions were chosen as $d = 5$ and 20, respectively. For the proposed ratio transformed conditional density estimation (RTCDE), we chose the auxiliary distribution $\tilde{P}_Y$ as the standard normal distribution to generate $\{\tilde{Y}_i\}_{i=1}^n$ in (4.10). The candidate function class $\mathcal{G}_N$ was chosen as the neural network, whose width and depth were selected from five-fold cross-validations. For comparison, we also conducted the conditional kernel density estimation (KCDE), the least squares conditional density estimation (LSCDE, Sugiyama et al., 2010) that uses the RKHS as its function class, and the FlexCode (Izbicki and Lee, 2017) method, which reformulates the conditional density estimation as a non-parametric orthogonal series problem. The sample sizes were chosen as $n \in \{1000, 2000, 5000\}$. To measure the accuracy of the estimates, we computed the empirical MSE $n^{-1}\sum_{i=1}^n \{\hat{p}(Y_i|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(Y_i, \boldsymbol{X}_i)\}^2$.

Table 3. Empirical average of mean squared errors (MSE) of the estimated conditional density based on 300 repetitions of the proposed ratio transformed conditional density estimation (RTCDE), the KCDE, the LSCDE, and the FlexCode. The empirical standard deviations of the MSEs of the 300 repetitions for each method are reported in parentheses.

Models	n	$d = 5$				$d = 20$			
		RTCDE	KCDE	LSCDE	FlexCode	RTCDE	KCDE	LSCDE	FlexCode
M1	1000	0.34 (0.13)	0.41 (0.17)	0.30 (0.15)	0.37 (0.11)	0.58 (0.30)	0.98 (0.26)	0.64 (0.29)	0.72 (0.33)
	2000	0.13 (0.08)	0.23 (0.09)	0.14 (0.09)	0.17 (0.06)	0.32 (0.12)	0.53 (0.16)	0.36 (0.14)	0.41 (0.19)
	5000	0.09 (0.05)	0.17 (0.09)	0.09 (0.04)	0.10 (0.04)	0.22 (0.13)	0.37 (0.15)	0.28 (0.11)	0.30 (0.13)
M2	1000	0.53 (0.23)	0.89 (0.30)	0.71 (0.26)	0.49 (0.21)	0.92 (0.34)	1.51 (0.50)	1.07 (0.29)	0.86 (0.28)
	2000	0.25 (0.12)	0.46 (0.18)	0.30 (0.11)	0.24 (0.09)	0.43 (0.20)	0.83 (0.24)	0.52 (0.18)	0.45 (0.20)
	5000	0.16 (0.09)	0.25 (0.11)	0.18 (0.06)	0.16 (0.05)	0.28 (0.16)	0.53 (0.19)	0.34 (0.11)	0.31 (0.13)
M3	1000	0.67 (0.18)	0.91 (0.36)	0.86 (0.31)	0.77 (0.25)	1.18 (0.53)	1.55 (0.65)	1.34 (0.42)	1.27 (0.50)
	2000	0.31 (0.12)	0.59 (0.18)	0.44 (0.13)	0.40 (0.15)	0.52 (0.23)	0.79 (0.27)	0.66 (0.19)	0.59 (0.26)
	5000	0.18 (0.07)	0.31 (0.11)	0.25 (0.09)	0.21 (0.07)	0.27 (0.11)	0.43 (0.18)	0.37 (0.13)	0.34 (0.12)

Table 3 suggests that the proposed RTCDE outperformed the other three methods in most experiments. For the small sample ($n = 1000$) and low dimensional ($d = 5$) scenarios, the LSCDE and the FlexCode had slightly smaller MSE than the RTCDE for the linear model (M1) and trigonometric model (M2), while the RTCDE showed faster convergence rates and had the best performances in large samples. The MSEs of the four estimators under the M3 model were larger than those under M1 and M2 because the underlying regression function was discontinuous, where the RTCDE still had superior performances compared to the others under such challenging cases.

7.3 Numerical results of estimation and inference of the GEE

We now present simulation results that examine the estimation accuracy of the estimator $\widehat{\boldsymbol{\theta}}$ based on the orthogonal estimating functions and the empirical coverage of the proposed inference procedure.

For the experiment results presented below, the covariates for the source and the target samples were generated in the same way as Setting S1 in Section 7.1, where the dimension was $d = 5$, and the responses were generated according to Model M2 in Section 7.2. The target parameter was $\theta_0 = Q_Y^{-1}(1/2)$, namely the median of Y for the target distribution, since it corresponds to a nonlinear estimating equation where the conventional AIPW cannot apply. In Section F of the SM, we report additional simulation results for the setting with $d = 20$, and results for the inference of the mean of Y for the target distribution.

The methods for comparison included the density ratio weighting (DRW), which is equivalent to the IPW of Chen et al. (2024), the multiple imputations (MI) proposed by Wang and Chen (2009), the proposed method having both the density ratio weighting and the multiple imputations with estimated nuisance functions (DRW-MI-E), the localized debiased machine learning (LDML) introduced by Kallus et al. (2024), and the covariance balancing (Imai and Ratkovic, 2014). The nuisance functions in the first four methods were all estimated with the deep neural networks for comparison fairness, whose widths and depths were chosen by the three-fold cross-validation. When using the multiple imputations, $\kappa = N/2$ imputations were made for each observation point. In addition, to evaluate the effects of nuisance function estimation errors, we also considered an oracle version of the DRW-MI, where the density ratio and the conditional density functions were used as the true ones (DRW-MI-T). To obtain the 95% confidence intervals, we employed the Wilks theorem for the DRW-MI-E and DRW-MI-T methods and the bootstrap approximations for the others, since they do not admit the Wilk theorem. Alternatively, one can use the asymptotic normality of the estimators of these four methods, where the asymptotic variances should be estimated.

Table 4 reports the performances of the six methods based on 300 simulation replications, where the estimation performances were measured with the empirical bias, standard deviation, and the MSE, and the inference performances were reflected from the empirical coverage probability and length of the confidence intervals (CI). The proposed DRW-MI-E method improved both the estimation and coverage performances than the DRW and MI alone, with the MSE converging to 0 and the coverage probability approaching the nominal level of 0.95. It is worth noting that the simulation results of the DRW-MI-E were comparable to that of the oracle method DRW-MI-T, where the nuisance functions used the true values, confirming the theoretical analysis that estimation for the nuisances does not have the first-order effect on the proposed orthogonal estimating function. The LDML also employed the same form of estimation function, while resorting to a two-step method for the estimation of the conditional mean function, which depended on an initial estimator and required three-fold sample splitting. As a result, its empirical performances were not as competitive as the DRW-MI-E that used the MI for the conditional mean function estimation. The CB method had the worst estimation accuracy and under-coveraged confidence intervals, since its theories require correctly specified balancing functions, which could not be satisfied under the nonlinear response models of the simulations.

Table 4. Empirical estimation and inference results for the median of the target population, based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	-0.0282	0.1759	0.0304	0.8791	0.7216
	MI	-0.0256	0.1801	0.0331	0.8900	0.7404
	DRW-MI-E	-0.0221	0.1648	0.0275	0.9326	0.7719
	DRW-MI-T	-0.0194	0.1610	0.0266	0.9437	0.7805
	LDML	0.0239	0.1810	0.0333	0.9048	0.7914
	CB	-0.0508	0.1964	0.0395	0.6612	0.8382
$n = 2000$	DRW	0.0204	0.1236	0.0157	0.8920	0.5083
	MI	-0.0216	0.1217	0.0153	0.9138	0.4885
	DRW-MI-E	-0.0180	0.1139	0.0130	0.9422	0.4729
	DRW-MI-T	0.0153	0.1062	0.0112	0.9540	0.4693
	LDML	0.0211	0.1302	0.0173	0.8987	0.4910
	CB	-0.0452	0.1516	0.0236	0.6865	0.5283
$n = 5000$	DRW	0.0159	0.0839	0.0073	0.9104	0.3665
	MI	-0.0164	0.0801	0.0067	0.9312	0.3572
	DRW-MI-E	-0.0135	0.0745	0.0058	0.9562	0.3691
	DRW-MI-T	-0.0129	0.0716	0.0052	0.9524	0.3634
	LDML	0.0160	0.0848	0.0074	0.9215	0.3820
	CB	0.0312	0.1209	0.0151	0.6928	0.4343

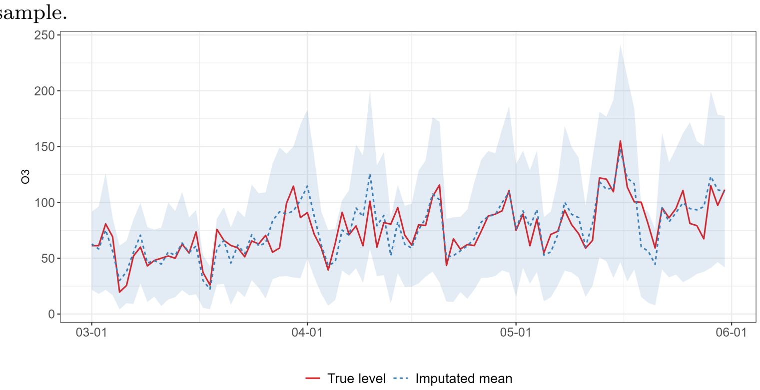
8 Case Study

Ground level ozone (O_3), as an air pollutant, has been at an elevated level across North China (Li et al., 2019) in the last decade. In this section, we demonstrate that the proposed method is well-suited for the transfer learning of the inference for the O_3 levels.

We focus on four major cities in North China, including Beijing, Xian, Jinan, and Taiyuan, where we used the first three cities as the source domain and Taiyuan as the target domain. The study period was the spring (March 1 to May 31) of 2018, a season when the ozone level is generally high. The response variable was the hourly O_3 obtained from China Meteorological Administration (CMA) monitoring sites, and the covariates included hourly $PM_{2.5}$, PM_{10}, nitrogen dioxide (NO_2), surface total solar radiation (TSR), surface air temperature (TEMP), relative humidity (HUMI), boundary layer height (BLH), the low (LCC), medium (MCC) and high (HCC) cloud cover percentages, as well as the one to three hour lagged terms of the above variables, where the first three variables were obtained from China Environmental Monitor Center sites in each city, the TSR data was collected from CMA, and the others were from the European Center for Medium-Range Weather Forecasts (ECMWF). We also included a DAY variable that counts for the number of days since March 1st to reflect the increasing radiation in the spring, which is highly statistically significant in modeling the O_3 as shown in Li et al. (2021). Our goal was to utilize the O_3 observations and the covariates of the source sample to assist the inferences at the target population of the O_3 in Taiyuan. To investigate the performances of the transfer learning methods, we assumed only the covariate variables of the target domain Taiyuan were observable during their implementations, while the true O_3 levels of the target sample were used to evaluate the quality of the transfer learning.

Distinctions between the distributions of some key variables of the target and the source samples are illustrated in Figure 1 of the SM, which reveals that directly using the source samples to make inferences about the O_3 of the target population would introduce biases. To apply the proposed method, we first estimated the covariate density ratio of the two samples and used the source sample to estimate the conditional density function. The density ratio and the conditional density functions were estimated with the neural networks, whose widths and heights were chosen from the five-fold cross-validation. From the estimated conditional density function, we conducted the multiple imputations with the number of imputations $\kappa = 200$. Figure 1 shows the 2.5% and 97.5% empirical quantiles as well as the empirical mean of the imputed values for O_3 of the target sample. The figure shows that most of the true O_3 values of the target were located within the 95% prediction region obtained from the multiple imputations, and the mean of the imputed values were well approximated to the true ones. Such a result not only verifies that the conditional density of the target sample was similar to that of the source, but also shows that our multiple imputation method produced high-quality surrogates for the O_3 on the target domain.

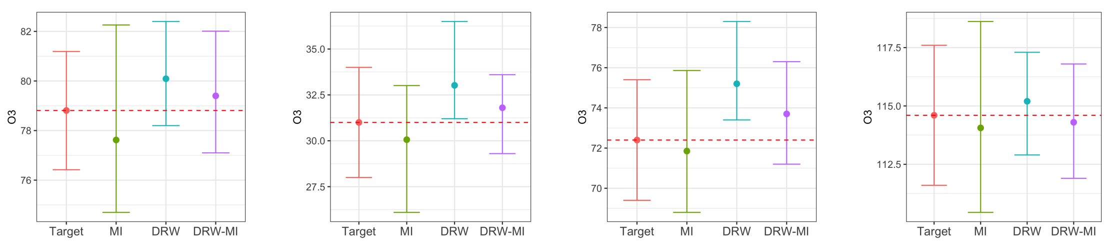
Figure 1. Illustration for the results of the multiple imputations for O_3 the target sample. The upper and lower boundaries of the blue region are the 2.5% and 97.5% empirical quantiles of the 200 imputations. The blue dotted line is the empirical mean of the imputed values. The red line indicates the true O_3 levels of the target sample.



We considered the estimation and inference for the mean and the α-quantiles ($\alpha = 25\%, 50\%$, and 75%) of the O_3 of the target domain in Taiyuan. The methods include the multiple imputation (MI), the density ratio weighting (DRW), and the proposed method (DRW-MI). Since the first two methods do not have Wilks' theorem, their confidence intervals (CIs) were derived by Bootstraps, where the CIs were derived based on the empirical 2.5% and 97.5% quantiles of the estimates from 200 resamplings.The CIs for the DRW-MI were via Wilks' theorem. As a baseline, we also considered an oracle method that used the O_3 of the target sample to conduct the inference for the four estimands, including the mean and the three quantiles of the O_3 of the target domain, where their CIs were obtained based on the asymptotic normalities. As reported in Figure 2, the CIs of the MI method were the largest among the four methods, indicating its being less favorable in terms of statistical efficiency. The estimation based on the DRW had a high proportion of overestimates compared with the estimation using the target sample. The proposed DRW-MI achieved the lowest bias among the three transfer learning methods, which verifies that the estimation based on the orthogonal estimating functions that used both the DRW and the MI can be regarded as a one-step bias correction of the DRW estimation. Moreover, the CIs of the DRW-MI had shorter lengths than those obtained with the target O_3 observations, indicating the benefit of the TL since it utilized both the information of the source and the target sample.

Figure 2. Estimation and 95% confidence intervals for the mean and three quantiles of the O_3 of the target population obtained from the target sample, the multiple imputations (MI), the density ratio weighting (DRW), and the density ratio weighting with multiple imputations (DRW-MI), respectively. As a comparison baseline, the red dotted line indicates the estimated value of the O_3 with the target sample.

(a) Mean (b) 25%-quantile (c) 50%-quantile (d) 75%-quantile



9 Discussion

This study investigates the statistical inference for general estimating equations with the covariate shift transfer learning. Instead of the common strategy of density ratio weighting, we construct an orthogonal estimating equation that is more robust against nuisance function estimation errors. To address the challenge that the conditional mean estimating function is parameter-dependent, we adopt a multiple-imputation approach that avoids conducting the regression at infinitely many parameters. Our estimation for the nuisance functions accommodates flexible uses of ML algorithms. The theoretical results reveal that the EL estimator based on the orthogonal estimating equation is semiparametric efficient. Compared with the related literature such as Chen et al. (2024), the inference does not require a Bootstrap procedure, as it is shown that the log El ratio restores the Wilks theorem, despite the presence of nuisance functions. We also discuss the DNN-based nuisance function estimation to alleviate the curse of dimensionality.

There are some intersting extensions that may be considered in subsequent research. First, in this work we focus on the transfer learning under the covariate shiftInvestigation on how to handle general distribution shift settings, such as the label shift and domain generalization, remains an important avenue for future work. Second, the density ratio function is required to be uniformly bounded, as a common assumption for nonparametric estimation. However, such a condition can be violated in scenarios where the target and the source domain have non-overlap regions, or the proportion of the target sample in the full sample converges to 0, namely $n/N \to 0$ as $N \to \infty$. Such scenarios have been considered in studies of semi-supervised learning, such as Zhang and Bradic (2022) and Chakrabortty et al. (2022). However, the current TL setting is more challenging due to the covariate shift, where the density ratio function should be estimated. In addition, while we have investigated the scenario where the dimension grows with the sample size in Section 5.2, it does not accommodate the ultra-high dimension regimes where the dimension can be larger than the sample size. Extensions to the ultra-high dimensional GEEs under transfer learning can possibly be achieved with the penalized EL established by Chang et al. (2018), while the effect of nuisance function estimation should be carefully examined.

References

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261.

Bravo, F., Escanciano, J. C., and Van Keilegom, I. (2020). Two-step semiparametric empirical likelihood inference. *The Annals of Statistics*, 48(1):1 – 26.

Cai, C., Cai, T. T., and Li, H. (2023). Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*.

Chakrabortty, A., Dai, G., and Tchetgen, E. T. (2022). A general framework for treatment effect estimation in semi-supervised and high dimensional settings. *arXiv preprint arXiv:2201.00468*.

Chang, J., Chen, S. X., and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1):283–304.

Chang, J., Chen, S. X., Tang, C. Y., and Wu, T. T. (2021). High-dimensional empirical likelihood inference. *Biometrika*, 108(1):127–147.

Chang, J., Tang, C. Y., and Wu, T. T. (2018). A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics*, 46(6B):3185–3216.

Chen, S. X. and Cui, H. (2007). On the second-order properties of empirical likelihood with moment restrictions. *Journal of Econometrics*, 141(2):492–516.

Chen, S. X., Peng, L., and Qin, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, 96(3):711–722.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2).

Chen, X., Liu, Y., Ma, S., and Zhang, Z. (2024). Causal inference of general treatment effects using neural networks with a diverging number of confounders. *Journal of Econometrics*, 238(1):105555.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press.

Guan, H. and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026.

Hanson, J., Litfin, T., Paliwal, K., and Zhou, Y. (2020). Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. *Bioinformatics*, 36(4):1107–1113.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, 37(3):1079–1111.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11:2800–2831.

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716.

Kallus, N., Mao, X., and Uehara, M. (2024). Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25(16):1–59.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media.

Kouw, W. M. and Loog, M. (2019). A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785.

Kpotufe, S. and Martinet, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323.

Kulis, B., Saenko, K., and Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR 2011*, pages 1785–1792.

Lee, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):911–938.

Li, K., Jacob, D. J., Liao, H., Shen, L., Zhang, Q., and Bates, K. H. (2019). Anthropogenic drivers of 2013–2017 trends in summer surface ozone in china. *Proceedings of the National Academy of Sciences*, 116(2):422–427.

Li, M., Neykov, M., and Balakrishnan, S. (2022a). Minimax optimal conditional density estimation under total variation smoothness. *Electronic Journal of Statistics*, 16(2):3937–3972.

Li, S., Cai, T. T., and Li, H. (2022b). 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.

Li, S., Cai, T. T., and Li, H. (2023). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, 118(543):2171–2183.

Li, S., Liu, R., Wang, S., and Chen, S. X. (2021). Radiative effects of particular matters on ozone pollution in six north china cities. *Journal of Geophysical Research: Atmospheres*, 126(24):e2021JD035963.

Ma, C., Pathak, R., and Wainwright, M. J. (2023). Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, (to appear).

Maji, A., Ghosh, A., Basu, A., and Pardo, L. (2019). Robust statistical inference based on the c-divergence family. *Annals of the Institute of Statistical Mathematics*, 71:1289–1322.

Mo, W., Qi, Z., and Liu, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and Statsitics*, pages 213–234.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, volume 2. Institute of Mathematical Statistics.

Qin, J. and Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, 22(1):300 – 325.

Reeve, H. W., Cannings, T. I., and Samworth, R. J. (2021). Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rockafellar, R. T. (1997). *Convex Analysis*, volume 18. Princeton University Press.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

Ryan, K. J. and Culp, M. V. (2015). On semi-supervised linear regression in covariate shift problems. *The Journal of Machine Learning Research*, 16(1):3183–3217.

Sason, I. and Verdú, S. (2016). f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *The Annals of Statistics*, 48(4):1875 – 1921.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of statistics*, pages 1040–1053.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

Sugiyama, M., Takeuchi, I., Suzuki, T., et al. (2010). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, 93(3):583–594.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490 – 517.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599.

Zhang, Y. and Bradic, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403.

Zhuang, F., Qi, Z., Duan, K., et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Supplemental Material for "Transfer Learning with General Estimating Equations"

Notations Throughout the supplementary material, we use c and C with different subscripts to denote generic finite positive constants and may be different in different uses. The empirical measure is denoted as $\mathbb{E}_n(\cdot)$. We use $\mathbb{1}(\mathcal{A})$ as the indicator function of an event $\mathcal{A}$. For any vector $\boldsymbol{v} = (v_1, \cdots, v_d)^{\mathrm{T}}$, let $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$ and $\|\boldsymbol{v}\|_p$ denote its L^p norm. For a function $f : \mathcal{X} \to \mathbb{R}$, its supreme is denoted by $\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$, and its L_p-norm under a distribution F that generates a random variable X is denoted by $\|f\|_{L_p(F)} = (\mathbb{E}_F|f(X)|^p)^{1/p}$ for any $p \geqslant 1$. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leqslant Cb_n$. Let $\mathrm{Pdim}(F)$ be the the Pseudo dimension (Pollard, 1990) of the function class $\mathcal{F}$. The ε-covering number of the function class F with respect to the metric d is denoted as $\mathcal{N}_d(\varepsilon, \mathcal{F})$.

A Proofs for Section 3

A.1 Proof of Theorem 3.1

In the sequel, we use $\mathbb{E}_0$ and $\mathbb{E}_\tau$ to denote the expectation under the true distribution F and the regular parametric submodel F_τ, respectively. The density function for F_τ is

$$f_\tau(\boldsymbol{w}) = p^\delta(1-p)^{1-\delta} f_\tau(\boldsymbol{y}|\boldsymbol{x})^{1-\delta} q_\tau(\boldsymbol{x})^\delta p_\tau(\boldsymbol{x})^{1-\delta},$$

and the score function is given by

$$S_\tau(\boldsymbol{w}) = (1-\delta)S_\tau(\boldsymbol{y}|\boldsymbol{x}) + \delta S_\tau^1(\boldsymbol{x}) + (1-\delta)S_\tau^0(\boldsymbol{x}),$$

where $S_\tau(\boldsymbol{y}|\boldsymbol{x}) = \partial \log f_\tau(\boldsymbol{y}|\boldsymbol{x})/\partial\tau$, $S_\tau^0(\boldsymbol{x}) = \partial \log p_\tau(\boldsymbol{x})/\partial\tau$ and $S_\tau^1(\boldsymbol{x}) = \partial \log q_\tau(\boldsymbol{x})/\partial\tau$, satisfying

$$\mathbb{E}_\tau\{S_\tau(\mathbf{Y}|\boldsymbol{X})|\boldsymbol{X}\} = \mathbf{0},\ \mathbb{E}_\tau\{\delta S_\tau^1(\boldsymbol{X})\} = \mathbf{0} \text{ and } \mathbb{E}_\tau\{(1-\delta)S_\tau^0(\boldsymbol{X})\} = \mathbf{0}. \tag{A.1}$$

(i) Since $\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\} = 0$, differentiating with respect to τ gives

$$\frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\}\Big|_{\tau=0} = \frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r_0)\}\Big|_{\tau=0} + \frac{\partial}{\partial\tau}\mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\}\Big|_{\tau=0}. \tag{A.2}$$

Under Condition 2 and the mean-squared differentiability of the submodel F_τ, for any $\boldsymbol{\theta} \in \Theta_0$, the differentiation and integration operators are exchangeable (see, e.g., Ibragimov and Has' Minskii, 1981) and it holds that

$$\frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r_0)\}\Big|_{\tau=0} = \mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r_0)S_0(\boldsymbol{W})\}. \tag{A.3}$$

We now calculate the right-hand side of (A.2).

$$\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\} = \mathbb{E}_\tau\left\{\frac{1-\delta}{1-p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})r(F_\tau)\right\} = \mathbb{E}_\tau\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\}.$$

Differentiating with respect to τ gives

$$\begin{aligned}
\frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}\Big|_{\tau=0} &= \frac{\partial}{\partial\tau}\mathbb{E}_\tau\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\}\Big|_{\tau=0}\\
&= \mathbb{E}_0\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})S_0(\boldsymbol{W})\right\}\\
&= \mathbb{E}_0\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})S_0(\boldsymbol{X}) + \frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})S_0(\mathbf{Y}|\boldsymbol{X})\right\}\\
&= \mathbb{E}_0\left\{\frac{\delta}{p}\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})S_0(\boldsymbol{W})\right\} + \mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\mathbf{Y}|\boldsymbol{X})\right\},
\end{aligned} \tag{A.4}$$

where the first term of (A.4) is from (A.1) and iterated expectation. We proceed to find a function $\mathbf{h}(\boldsymbol{W})$ such that the second term is equivalent to $\mathbb{E}_0\{\mathbf{h}(\boldsymbol{W})S_0(\boldsymbol{W})\}$. Note that

$$\begin{aligned}
\mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\mathbf{Y}|\boldsymbol{X})\right\} &= \mathbb{E}_0\left[\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)\{S_0(\boldsymbol{W}) - (1-\delta)S_0(\boldsymbol{X})\}\right]\\
&= \mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\boldsymbol{W})\right\} - \mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\boldsymbol{X})\right\},
\end{aligned}$$

and the second term is equivalent to

$$\mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\boldsymbol{X})\right\} = \mathbb{E}\left\{\frac{1-\delta}{1-p}r_0(\boldsymbol{X})\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})S_0(\boldsymbol{X})\right\} = \mathbb{E}\left\{\frac{1-\delta}{1-p}r_0(\boldsymbol{X})\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})S_0(\boldsymbol{W})\right\}, \tag{A.5}$$

where the first equality is by the iterated expectation, and the second equality is because of (A.1). Combining (A.2)-(A.5) gives

$$\frac{\partial}{\partial\tau}\mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}\Big|_{\tau=0} = \mathbb{E}_0\left\{\boldsymbol{\varphi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}_0)S_0(\boldsymbol{W})\right\},$$

where $\boldsymbol{\eta}_0(\boldsymbol{x}) = (r_0(\boldsymbol{x}),\mathbf{m}_0(\boldsymbol{x}))$ and

$$\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{\delta}{p}\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}) - \frac{1-\delta}{1-p}r(\boldsymbol{x})\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}),$$

It is straightforward to see that $\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}_0)\} = \mathbf{0}$ for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_0$. In addition, because the set of score functions is dense in $L_2(F)$, the influence function $\boldsymbol{\varphi}$ is uniquely determined.

(ii) Let $\boldsymbol{\Psi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}) = \tilde{\mathbf{g}}(\boldsymbol{w},\boldsymbol{\theta},r) + \boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta})$. Since $\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}_0)\} = \mathbf{0}$, replacing F by F_τ gives $\mathbb{E}_\tau\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\} = \mathbf{0}$. Differentiating this identity with respect to $\tau = 0$ gives

$$\mathbf{0} = \frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \tau}\mathbb{E}_\tau\left\{\boldsymbol{\varphi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}_0)\right\} + \frac{\partial}{\partial \tau}\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\} \qquad \text{(A.6)}\\
&= \mathbb{E}_0\left\{\boldsymbol{\varphi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}_0)S_0(\boldsymbol{W})\right\} + \frac{\partial}{\partial \tau}\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\\
&= \frac{\partial}{\partial \tau}\mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}\Big|_{\tau=0} + \frac{\partial}{\partial \tau}\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0} \qquad \text{(A.7)}\\
&= \frac{\partial}{\partial \tau}\mathbb{E}_0\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0},
\end{aligned}
$$

where (A.6) is from differentiation by parts and (A.7) is from the result in (i).

(iii) First, $\boldsymbol{\Psi}$ can be rewritten as

$$
\boldsymbol{\Psi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{\delta}{p}\mathbf{g}(\boldsymbol{z},\boldsymbol{\theta}) + \left\{\frac{1-\delta}{1-p}r(\boldsymbol{x}) - \frac{\delta}{p}\right\}\{\mathbf{g}(\boldsymbol{z},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{x},\boldsymbol{\theta})\}.
$$

Because $\mathbb{E}_F\{\delta\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta}_0)\} = \mathbf{0}$, we have

$$
\begin{aligned}
\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\} &= \mathbb{E}_F\left[\left\{\frac{1-\delta}{1-p}r(\boldsymbol{X}) - \frac{\delta}{p}\right\}\{\mathbf{g}(\boldsymbol{z},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{X},\boldsymbol{\theta})\}\right]\\
&= \mathbb{E}_F\left[\left\{\frac{1-\delta}{1-p}r(\boldsymbol{X}) - \frac{\delta}{p}\right\}\{\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{X},\boldsymbol{\theta})\}\right],
\end{aligned}
$$

implying that $\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\} = \mathbf{0}$ if either $r(\boldsymbol{x}) \overset{a.e.}{=} r_0(\boldsymbol{x})$ or $\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}_0) \overset{a.e.}{=} \mathbf{m}_0(\boldsymbol{x},\boldsymbol{\theta}_0)$.

Let $\boldsymbol{\Delta}(\boldsymbol{x},\boldsymbol{\theta}) = \mathbf{m}_0(\boldsymbol{x},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}) = (\Delta_1,\ldots,\Delta_r)^{\mathrm{T}}$. Since $\mathbb{E}_F[\{(1-p)^{-1}(1-\delta)r_0(\boldsymbol{X}) - p^{-1}\delta\}\boldsymbol{\Delta}(\boldsymbol{X},\boldsymbol{\theta})] = \mathbf{0}$, we have

$$
\begin{aligned}
|\mathbb{E}_F\{\Psi_j(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\}| &= \left|\mathbb{E}_F\left[\left\{\frac{1-\delta}{1-p}r_0(\boldsymbol{X}) - \frac{1-\delta}{1-p}r(\boldsymbol{X})\right\}\Delta_j(\boldsymbol{X},\boldsymbol{\theta}_0)\right]\right|\\
&= |\mathbb{E}_P[\{r_0(\boldsymbol{X}) - r(\boldsymbol{X})\}\Delta_j(\boldsymbol{X},\boldsymbol{\theta}_0)]|\\
&\leqslant \mathbb{E}_P\{|r_0(\boldsymbol{X}) - r(\boldsymbol{X})||\Delta_j(\boldsymbol{X},\boldsymbol{\theta}_0)|\}\\
&\leqslant \|r - r_0\|_{L_2(P_X)}\|m_j(\cdot,\boldsymbol{\theta}_0) - m_{0j}(\cdot,\boldsymbol{\theta}_0)\|_{L_2(P_X)}, \qquad \text{(A.8)}
\end{aligned}
$$

which completes the proof. □

B Proofs for Section 4

B.1 Proof of Lemma 4.1

According to Fenchel dual representation (Rockafellar, 1997), each convex ϕ can be expressed by:

$$
\phi(u) = \sup_{v\in\mathbb{R}}\{uv - \phi_*(v)\}.
$$

By the definition of $D_\phi(Q\|P)$, we have

$$
D_\phi(Q\|P) = \int \phi\left(\frac{q_0(x)}{p_0(x)}\right)p_0(\boldsymbol{x})d\boldsymbol{x}
$$

$$
\begin{aligned}
&= \int \sup_{v(\boldsymbol{x})} \left(v(\boldsymbol{x}) \frac{q_0(x)}{p_0(x)} - \phi_*(v(\boldsymbol{x})) \right) p_0(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sup_v \int \{ v(\boldsymbol{x}) q_0(\boldsymbol{x}) - \phi_*(v(\boldsymbol{x})) p_0(\boldsymbol{x}) \} d\boldsymbol{x} \\
&\geqslant \sup_v \mathbb{E}_Q\{v(\boldsymbol{X})\} - \mathbb{E}_P\{\phi_*(v(\boldsymbol{X}))\},
\end{aligned}
$$

where the supremum in the last two equality is taken over all measurable functions from $\mathcal{X} \to \mathrm{dom}(\phi_*)$. Since for each fixed $\boldsymbol{x}$ in the third equality, $v(\boldsymbol{x})q_0(\boldsymbol{x}) - \phi_*(v(\boldsymbol{x}))p_0(\boldsymbol{x})$ is maximized at $v_*(\boldsymbol{x}) = \phi_*^{-1}(q_0(\boldsymbol{x})/p_0(\boldsymbol{x})) = \phi_*^{-1}(r_0(\boldsymbol{x}))$. By the convex duality theorem, we have $v_*(\boldsymbol{x}) = \phi'(r_0(\boldsymbol{x}))$. Therefore,

$$
\phi'(r_0) = \arg\max_v [\mathbb{E}_Q\{v(\boldsymbol{X})\} - \mathbb{E}_P\{\phi_*(v(\boldsymbol{X}))\}],
$$

which implies that

$$
r_0 = \arg\min_r [\mathbb{E}_P\{\ell_{1,\phi}(r) - \mathbb{E}_Q\{\ell_{2,\phi}(r)\}\}],
$$

where the arg min is taken over all nonnegative functions with the domain $\mathcal{X}$. □

B.2 Proof of Theorem 4.1

Our proof proceeds in several steps. In Step 1, we present an error decomposition for $\|\widehat{r} - r_0\|^2_{L_2(P)}$. In Steps 2 - 4, we investigate the deviations between the sample and population excess risks via empirical process theories. Finally, we bound the empirical estimation error by the L_2 error in Step 5. Throughout the proof, we assume $M_1 \geqslant B_1$ without loss of generality. For any $r \in \mathcal{F}_N$, we define its empirical error as $\|r - r_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n (r(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i))^2$.

Step 1: Error decomposition. Denote $\ell_1(r, \boldsymbol{x}) = \phi_*\{\phi'(r(\boldsymbol{x}))\}, \ell_2(r, \boldsymbol{x}) = -\phi'(r(\boldsymbol{x}))$. Let $\mathcal{L}_1(r) = \mathbb{E}_P\{\ell_1(r, \boldsymbol{X})\}$, $\mathcal{L}_2(r) = \mathbb{E}_Q\{\ell_2(r, \boldsymbol{X})\}$, and $\widehat{\mathcal{L}}_1(r) = n^{-1}\sum_{i=1}^n \ell_1(r, \boldsymbol{X}_i)$, $\widehat{\mathcal{L}}_2(r) = m^{-1}\sum_{i=n+1}^{n+m} \ell_2(r, \boldsymbol{X}_i)$. The population and the sample criterion function are:

$$
\mathcal{L}(r) := \mathcal{L}_1(r) + \mathcal{L}_1(r) \quad \text{and} \quad \widehat{\mathcal{L}}(r) := \widehat{\mathcal{L}}_1(r) + \widehat{\mathcal{L}}_2(r).
$$

For any $r_1, r_2 : \mathcal{X} \to [0, \infty)$, let

$$
d_\phi(r_1, r_2) := \mathcal{L}_\phi(r_1) - \mathcal{L}_\phi(r_2) \quad \text{and} \quad \widehat{d}_\phi(r_1, r_2) = \widehat{\mathcal{L}}_\phi(r_1) - \widehat{\mathcal{L}}_\phi(r_2).
$$

Given a function class $\mathcal{F}_N$, we define the best approximation for r_0 realized by $\mathcal{F}_N$ and the corresponding approximation error as:

$$
r_N := \arg\min_{r \in \mathcal{F}_N} \|r - r_0\|_\infty \quad \text{and} \quad \varepsilon_N := \|r_N - r_0\|_\infty.
$$

Note that r_N and ε_N are both deterministic and depend only on the architecture of $\mathcal{F}_N$ and the target function r_0.

By the compactness of $\boldsymbol{X}$ and Condition 4, it can be shown that there exists a positive constant L, such that for every $r, r' \in \mathcal{F}_N$,

$$|\ell_i(r, \boldsymbol{x}) - \ell_i(r', \boldsymbol{x})| \leqslant L|r(\boldsymbol{x}) - r'(\boldsymbol{x})|, \quad (i = 1, 2),$$

for all $\boldsymbol{x} \in \mathcal{X}$, and there exists positive constants c_1 and c_2 such that

$$c_1\|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant \mathcal{L}_i(r) - \mathcal{L}_i(r_0) \leqslant c_2\|\widehat{r} - r_0\|_{L_2(P)}^2, \quad (i = 1, 2), .$$

Therefore, we have the following error decomposition:

$$c_1\|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant d_\phi(\widehat{r}, r_0) = d_\phi(\widehat{r}, r_N) + d_\phi(r_N, r_0) \leqslant d_\phi(\widehat{r}, r_N) + c_2\varepsilon_N^2. \tag{B.1}$$

We next bound $d_\phi(\widehat{r}, r_N)$ by analyzing the process $\sup_{r \in \mathcal{F}_N} |d_\phi(r, r_N) - \widehat{d}_\phi(r, r_N)|$, mainly based on techniques of the local Rademacher complexity analysis of empirical risk minimization (Bartlett et al., 2005 and Koltchinskii, 2011). First, we introduce some quantities that are necessary in this approach. Let $\{\varepsilon_i\}_{i=1}^{n+m}$ be i.i.d symmetric, $\{-1, 1\}$-valued random variables that are independent of $\{\boldsymbol{X}_i\}_{i=1}^{n+m}$. For any function class $\mathcal{F}$, we define

$$\mathcal{R}_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(\boldsymbol{X}_i), \quad \mathcal{R}_m(\mathcal{F}) := \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i f(\boldsymbol{X}_i).$$

The Rademacher complexities are defined as $\bar{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}\{\mathcal{R}_n(\mathcal{F})\}$ and $\bar{\mathcal{R}}_m(\mathcal{F}) = \mathbb{E}\{\mathcal{R}_m(\mathcal{F})\}$, where the expectations are taken over both the $\boldsymbol{X}_i$s and the ε_is. The empirical Rademacher complexities, which are conditioned on the data, are denoted by $\widehat{R}_n(\mathcal{F}) = \mathbb{E}_\varepsilon\{R_n(\mathcal{F})\}$ and $\widehat{R}_m(\mathcal{F}) = \mathbb{E}_\varepsilon\{R_m(\mathcal{F})\}$. For the candidate function class $\mathcal{F}_N$, let the shifted (centered) function class be

$$\mathcal{F}_N^* := \{r - r_N : r \in \mathcal{F}_N\}.$$

The population version of the localized Rademacher complexities are defined as:

$$\bar{\mathcal{R}}_n(\delta, \mathcal{F}_N^*) := \bar{\mathcal{R}}_n\left\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_{L_2(P)} \leqslant \delta\right\} \quad \text{and} \quad \bar{\mathcal{R}}_m(\delta, \mathcal{F}_N^*) := \bar{\mathcal{R}}_m\left\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_{L_2(Q)} \leqslant \delta\right\},$$

where $\delta > 0$ is a localization scale. Similarly, the empirical localized Rademacher complexities are defined as:

$$\widehat{\mathcal{R}}_n(\delta, \mathcal{F}_N^*) := \widehat{\mathcal{R}}_n\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_n \leqslant \delta\} \quad \text{and} \quad \widehat{\mathcal{R}}_m(\delta, \mathcal{F}_N^*) := \widehat{\mathcal{R}}_m\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_m \leqslant \delta\}.$$

A crucial parameter in the localized Rademacher complexity approach is the critical radius, which is defined as δ_n and δ_m that satisfy the following inequalities:

$$\delta_n^2 \geqslant \bar{\mathcal{R}}_n(\delta_n, \mathcal{F}_N^*), \quad \delta_m^2 \geqslant \bar{\mathcal{R}}_m(\delta_m, \mathcal{F}_N^*). \tag{B.2}$$

For $j = 1$ and 2, denote the supreme deviations between $\widehat{\mathcal{L}}_j(r) - \widehat{\mathcal{L}}_j(r_0)$ and $\mathcal{L}_j(r) - \mathcal{L}_j(r_0)$ restricted in the localized ball centered at r_0 with the radius s as

$$\lambda_N^j(s) = \sup_{\|r - r_N\|_{L_2(P)} \leqslant s} \left| \left(\widehat{\mathcal{L}}_j(r) - \widehat{\mathcal{L}}_j(r_N)\right) - \left(\mathcal{L}_j(r) - \mathcal{L}_j(r_N)\right), \right| \tag{B.3}$$

and denote the supreme deviations between $d_\phi(r, r_N)$ and $\widehat{d}_\phi(r, r_N)$ restricted in $d_\phi(r, r_N)$ as

$$\lambda_N(s) = \sup_{\|r - r_N\|_{L_2(P)} \leqslant s} \left| \widehat{d}_\phi(r, r_N) - d_\phi(r, r_N) \right|, \tag{B.4}$$

where $s > 0$ is a radius to be varied.

Step 2. Tail bound of $\lambda_N(s)$. We first estimate an upper bound of the expectation of $\lambda_N(s)$ for the s in the range $[\delta_n \vee \delta_m, \infty)$. Let

$$\mathcal{G}_N^j(s) = \left\{ g : g = \ell_j(r) - \ell_j(r_0) \ \text{ for } r \in \mathcal{F}_N \text{ and } d_\phi(r, r_0) \leqslant s^2 \right\}$$

for $j = 1$ and 2. Then by standard symmetrization arguments, we have

$$\mathbb{E}\left\{\lambda_N^1(s)\right\} \leqslant 2\bar{\mathcal{R}}_n\left\{\mathcal{G}_N^1(s)\right\} \quad \text{and} \quad \mathbb{E}\left\{\lambda_N^2(s)\right\} \leqslant 2\bar{\mathcal{R}}_m\left\{\mathcal{G}_N^2(s)\right\}. \tag{B.5}$$

Since both $\phi_* \circ \phi'$ and ϕ' are L-Lipschitz continuous, by the Ledoux-Talagrand contraction inequality due to Ledoux and Talagrand (1991), it holds that $\bar{\mathcal{R}}_n\left\{\mathcal{G}_N^1(s)\right\} \leqslant 2L\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*)$ and $\bar{\mathcal{R}}_m\left\{\mathcal{G}_N^2(s)\right\} \leqslant 2L\bar{\mathcal{R}}_m(s, \mathcal{F}_N^*)$. Therefore,

$$\mathbb{E}\left\{\lambda_N^1(s)\right\} \leqslant 4L\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*) \quad \text{and} \quad \mathbb{E}\left\{\lambda_N^2(s)\right\} \leqslant 4L\bar{\mathcal{R}}_m(s, \mathcal{F}_N^*).$$

Since $\mathcal{F}_N^*$ is star-shaped around r_N (if $r \in \mathcal{F}_N^*$, then for any $\alpha \in (0,1)$, $\alpha r \in \mathcal{F}_N^*$), the function $\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*)/s$ is non-increasing with resepct to s according to Lemma 13.6 of Wainwright (2019). As $s > \delta_n$ and $\delta_n^2 > \bar{\mathcal{R}}_n\left\{\delta_n, \mathcal{F}_N^*\right\}$, it holds that $\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*) \leqslant s\delta_n$. Similarly, we also have $\bar{\mathcal{R}}_m(s, \mathcal{F}_N^*) \leqslant s\delta_m$ for $s \geqslant \delta_m$, which delivers the upper bounds

$$\mathbb{E}\left\{\lambda_N^1(s)\right\} \leqslant 4Ls\delta_n \quad \text{and} \quad \mathbb{E}\left\{\lambda_N^2(s)\right\} \leqslant 4Ls\delta_m \quad (\forall s \geqslant \delta_n \vee \delta_m). \tag{B.6}$$

We next bound the deviation between $\lambda_N^j(s)$ and $\mathbb{E}\left\{\lambda_N^j(s)\right\}$ for $j = 1$ and 2. Note that for any $r \in \mathcal{F}_N$, we have $\|\ell_j(r) - \ell_j(r_N)\|_\infty \leqslant L\|r - r_N\|_\infty \leqslant 2M_1L$, by the Lipschitz condition of $\phi_* \circ \phi'$ and ϕ' and the boundness of $r \in \mathcal{F}_N$. In addition, the variance of $\ell_j(r) - \ell_j(r_N)$ can be upper bounded by

$$\begin{aligned}
\operatorname{Var}(\ell_j(r) - \ell_j(r_N)) \leqslant & \mathbb{E}\left\{(\ell_j(r) - \ell_j(r_N))^2\right\} \\
\leqslant & L^2\left(\|r - r_N\|_{L_2(P)}^2 + \|r - r_N\|_{L_2(Q)}^2\right) \\
\leqslant & 2(M_1L)^2\|r - r_N\|_{L_2(P)}^2 \leqslant 2(M_1Ls)^2,
\end{aligned} \tag{B.7}$$

where the second inequality is implied by the Lipschitz condition, the third inequality is due to $\|f\|_{L_2(Q)}^2 = \|f \cdot r_0\|_{L_2(P)}^2 \leqslant B^2\|f\|_{L_2(P)}^2$ for any $f : \mathcal{X} \to \mathbb{R}$, and the last inequality is because of the localization condition $\|r - r_N\|_{L_2(P)} \lesssim d_\phi(r, r_N) \leqslant s$. Consequently, for any $u > 0$ it holds that

$$\mathbb{P}\left\{\lambda_N^j(s) \geqslant \mathbb{E}\{\lambda_N^j(s)\} + u\right\} \leqslant 2\exp\left(\frac{-(n \wedge m)u^2}{8e\operatorname{Var}(\ell_\phi(r) - \ell_\phi(r_0)) + 8M_1Lu}\right)$$

$$\leqslant 2\exp\left(-\frac{C_t(n\wedge m)u^2}{(M_1Ls)^2+M_1Lu}\right),$$

for some universal constant $C_t > 0$, by applying Talagrand's concerntration equality (Talagrand, 1994) and (B.7). Therefore, we have

$$\left(\mathbb{P}\left\{\lambda_N^1(s)\geqslant 4Ls\delta_n+u\right\}\vee\mathbb{P}\left\{\lambda_N^2(s)\geqslant 4Ls\delta_m+u\right\}\right)\leqslant 2\exp\left(-\frac{C_t(n\wedge m)u^2}{(M_1Ls)^2+M_1Lu}\right),$$

for any $s \geqslant (\delta_n \vee \delta_m)$ and $u > 0$. Since

$$\lambda_N(s)\leqslant\lambda_N^1(s)+\lambda_N^2(s)$$

for any $s \geqslant 0$, we have

$$\mathbb{P}\left\{\lambda_N(s)\geqslant 4Ls(\delta_n+\delta_m)+u\right\}\leqslant 4\exp\left(-\frac{C_t(n\wedge m)u^2}{(2M_1Ls)^2+2M_1Lu}\right), \tag{B.8}$$

for any $s \geqslant (\delta_n \vee \delta_m)$ and $u > 0$. Denoting $\delta_N := \delta_n + \delta_m$ and setting $s = \delta_N, u = M_1L\delta_N^2$, then we have

$$\mathbb{P}\left\{\lambda_N(\delta_N)\geqslant C_1\delta_N^2\right\}\leqslant 4\exp\left(-C_2(n\wedge m)\delta_N^2\right), \tag{B.9}$$

where $C_1 = (4 + M_1)L$ and $C_2 = C_t/6$. In addition, setting $u = M_1Ls\delta_N$ yields

$$\mathbb{P}\left\{\lambda_N(s)\geqslant C_1s\delta_N\right\}\leqslant 2\exp\left(-\frac{C_tns^2\delta_N^2}{s^2+s\delta_N}\right)\leqslant 4\exp\left(-C_2(n\wedge m)\delta_N^2\right), \tag{B.10}$$

for any $s \geqslant \delta_N$.

Let

$$\mathcal{A}_1=\left\{\exists\, r\in\mathcal{F}_N:\|r-r_N\|_{L_2(P)}\leqslant\delta_N \quad\text{and}\quad \left|\widehat{d}_\phi(r,r_N)-d_\phi(r,r_N)\right|\geqslant C_1\delta_N^2\right\}. \tag{B.11}$$

Combining (B.4) with (B.9) yields that

$$\mathbb{P}(\mathcal{A}_1)\leqslant 4\exp\left(-C_2(n\wedge m)\delta_N^2\right). \tag{B.12}$$

The above tail bound (B.10) controls the largest deviation $\left|\widehat{d}_\phi(r,r_N)-d_\phi(r,r_N)\right|$ for r within the local ball $\|r-r_N\|_{L_2(P)}\leqslant\delta_N$. It remains to estimate an tail bound of the deviation $\left|\widehat{d}_\phi(r,r_N)-d_\phi(r,r_N)\right|$ outside this local region. We define the following event

$$\mathcal{A}_2=\left\{\exists\, r\in\mathcal{F}_N:\|r-r_N\|_{L_2(P)}>\delta_N \quad\text{and}\quad \left|\widehat{d}_\phi(r,r_N)-d_\phi(r,r_N)\right|\geqslant 2C_1\delta_N\|r-r_N\|_{L_2(P)}\right\}$$

However, bounding $\mathbb{P}(\mathcal{A}_2)$ is more delicate, since the function r that satisfies the requirement in $\mathcal{A}_2$ is random. In the following step, we will use a "peeling" argument to address the problem.

Step 3: Bound the event $\mathcal{A}_2$ with the peeling argument. For $m \in \mathbb{N}_+$, we define the events

$$\mathcal{S}_m := \left\{r \in \mathcal{F}_N : 2^{m-1}\delta_N < \|r - r_N\|_{L_2(P)} \leqslant 2^m \delta_N\right\}.$$

By the boundness of $r \in \mathcal{F}_N$, we have $\|r - r_N\|_{L_2(P)} \leqslant 2M_1$. Hence, any $r \in \mathcal{F}_N \cap \left\{\|r - r_N\|_{L_2(P)} > \delta_N\right\}$ must locate in some $\mathcal{S}_m$ for $m \in [\![K]\!]$, where $K \leqslant 2\log(M_1/\delta_N) + 1$. Since $\mathcal{A}_2$ is a subset of $\cup_{m=1}^{K} S_m$, by the union bound we have $\mathbb{P}(\mathcal{A}_2) \leqslant \sum_{m=1}^{M} \mathbb{P}(\mathcal{A}_2 \cap S_m)$.

Note that if $r_m \in \mathcal{A}_2 \cap S_m$, then we can take $s_m = 2^m \delta_N$, and r_m satisfies

$$\|r_m - r_N\|_{L_2(P)} \leqslant s_m \quad \text{and} \quad \left|\widehat{d}_\phi(r_m, r_0) - d_\phi(r_m, r_0)\right| \geqslant 2C_1\delta_N \|r - r_N\|_{L_2(P)} > C_1 \delta_N s_m,$$

where the last inequality is due to $2\|r - r_N\|_{L_2(P)} > 2^{m+1}\delta_N > s_m = 2^m \delta_N$. As a result, $\mathcal{A}_2 \cap \mathcal{S}_m \subset \{\lambda_N(s_m) \geqslant C_1 s_m \delta_N\}$. Then according to (B.10), we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{A}_2) &\leqslant \sum_{m=1}^{K} \mathbb{P}(\mathcal{A} \cap \mathcal{S}_m) \leqslant 2 \sum_{m=1}^{K} \exp\left(-C_2(n \wedge m)\delta_N^2\right) \\ &\leqslant 4\exp(-C_2(n \wedge m)\delta_N^2 + \log K) \leqslant 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right),\end{aligned} \tag{B.13}$$

where the last inequality holds provided that

$$\frac{C_2(n \wedge m)\delta_N^2}{2} \geqslant \log\left(2\log(M_1/\delta_N) + 1\right). \tag{B.14}$$

The complement of $\mathcal{A}_2$ is composed by $\mathcal{A}_2^c = \mathcal{B}_1 \cup \mathcal{B}_2$, where

$$\mathcal{B}_1 = \left\{r \in \mathcal{F}_N : \ \|r - r_N\|_{L_2(P)} \leqslant \delta_N\right\} \ \text{and} \ \mathcal{B}_2 = \left\{r \in \mathcal{F}_N : \ \left|\widehat{d}_\phi(r, r_N) - d_\phi(r, r_N)\right| < 2C_1\delta_N\|r - r_N\|_{L_2(P)}\right\}.$$

Therefore, (B.13) implies that

$$\mathbb{P}(\mathcal{B}_1 \cup \mathcal{B}_2) \geqslant 1 - 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right).$$

If $\widehat{r} \in \mathcal{B}_1$, then we have $d(\widehat{r}, r_N) \leqslant c_2\delta_N^2$ since $d(\widehat{r}, r_N) \leqslant c_2\|\widehat{r} - r_N\|_{L_2(P)}^2$. Moreover, if $\widehat{r} \in \mathcal{B}_2$, since $c_1\|\widehat{r} - r_N\|_{L_2(P)}^2 \leqslant d_\phi(\widehat{r}, r_N)$, and $\widehat{d}(\widehat{r}, r_N) \leqslant 0$ by the definition of $\widehat{r}$, we have $d_\phi(\widehat{r}, r_N) < 4c_1^{-2}C_1^2\delta_N^2$. This together with (B.12) leads to

$$\mathbb{P}\left\{d_\phi(\widehat{r}, r_N) < (c_2 \vee 4c_1^{-2}C_1^2)\delta_N^2\right\} \geqslant 1 - 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right). \tag{B.15}$$

Let $C_3 = c_2 \vee 4c_1^{-2}C_1^2$ and $C_4 = C_2/2$, combining (B.1) and (B.15), we obtain

$$\mathbb{P}\left\{c_1\|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant C_3\delta_N^2 + c_2\varepsilon_N^2\right\} \geqslant 1 - 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right). \tag{B.16}$$

Therefore, the estimation error $\|\widehat{r}-r_0\|_{L_2(P)}$ relies on the critical radius δ_N and the approximation error ε_N. In the next step, we provide an upper bound of the critical radius δ_N.

Step 4: Estimation of the critical radius δ_N. In this step, we first estimate the empirical critical radiuses $\widehat{\delta}_n$ and $\widehat{\delta}_m$ satisfying

$$\widehat{\delta}_n^2 \geqslant k\widehat{\mathcal{R}}_n(\widehat{\delta}_n, \mathcal{F}_N^*), \quad \widehat{\delta}_m^2 \geqslant k\widehat{\mathcal{R}}_m(\widehat{\delta}_m, \mathcal{F}_N^*), \tag{B.17}$$

where k is a fixed positive constant, $\widehat{\mathcal{R}}_n(\delta_n, \mathcal{F}_N^*)$ and $\widehat{\mathcal{R}}_n(\delta_m, \mathcal{F}_N^*)$ are localized empirical Rademacher complexities, respectively, then use Proposition 14.25 of Wainwright (2019) to obtain that

$$\mathbb{P}(C_4\delta_n \leqslant \widehat{\delta}_n \leqslant C_5\delta_n) \geqslant 1 - C_6\exp(-C_7 n\delta_n^2) \tag{B.18}$$

for some generic constants $C_4, \cdots, C_7 > 0$.

By the Dudley's chaining, we have

$$\widehat{\mathcal{R}}_n(s, \mathcal{F}_N^*) \leqslant \inf_{0<\alpha<s}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^s \sqrt{\log\left(\mathcal{N}_2(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)d\varepsilon\right)}\right\}, \tag{B.19}$$

where $\boldsymbol{X}_1^n = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$. Since for any $\|f\|_n \leqslant \max_{1\leqslant i\leqslant n}|f(\boldsymbol{X}_i)|$, we have $\mathcal{N}_2(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n) \leqslant \mathcal{N}_\infty(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)$. Since $\|f\|_\infty \leqslant 2M$ for $f \in \mathcal{F}_N^*$, according to Theorem 12.2 of Anthony and Bartlett (1999), we have

$$\log\left(\mathcal{N}_\infty(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)\right) \leqslant \mathrm{Pdim}(\mathcal{F}_N^*)\left(\frac{4eMn}{\varepsilon\mathrm{Pdim}(\mathcal{F}_N^*)}\right).$$

When $n > \mathrm{Pdim}(\mathcal{F}_N^*)$, let $\alpha = s\sqrt{\mathrm{Pdim}(\mathcal{F}_N^*)/n}$ in (B.19), we have

$$\inf_{0<\alpha<s}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^s \sqrt{\log\left(\mathcal{N}_2(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)d\varepsilon\right)}\right\} \leqslant 16s\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N^*)}{n}\left(\log\frac{4eM}{s} + \frac{3}{2}\log n\right)}.$$

Therefore, if $s \geqslant 1/n$ and $n \geqslant (4eM)^2$, the localized empirical Rademacher complexity can be upper bounded by

$$\widehat{\mathcal{R}}_n(s, \mathcal{F}_N^*) \leqslant 32s\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N^*)}{n}\log(n)}.$$

With such result, we find that the $\widehat{\delta}_n$ satisfying $\widehat{\delta}_n^2 \geqslant \widehat{\mathcal{F}}_n(\widehat{\delta}_n, \mathcal{F}_N^*)$ can be taken as

$$\widehat{\delta}_n = 32k\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N^*)}{n}\log(n)} + u = 32k\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N)}{n}\log(n)} + u, \tag{B.20}$$

for any $u \geqslant 0$. The empirical critical value $\hat{\delta}_m$ can be taken similarly. Using (B.16), (B.18), and (B.20), we obtain that for any $u \geqslant 0$,

$$\mathbb{P}\left\{\|\widehat{r}-r_0\|_{L_2(P)}^2 \leqslant C_8\left(\xi_N + \epsilon_N^2 + u\right)\right\} \geqslant 1 - C_9\exp(-N\xi_N - Nu), \tag{B.21}$$

for some universal constants C_8 and $C_9 > 0$, where ξ_N represents the stochastic error in the estimation and is defined as

$$\xi_N = \mathrm{Pdim}(\mathcal{F}_N)\left(\frac{\log(n)}{n} + \frac{\log(m)}{m}\right).$$

Since $N\xi_N \asymp \mathrm{Pdim}(\mathcal{F}_N)\log(N)$, we have $\exp(-N\xi_N) < C_9^{-1}$ for large enough N. Therefore, (B.21) implies that for large enough N and any $t \geqslant 0$, it holds that

$$\mathbb{P}\left\{\|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant C_8\left(\xi_N + \epsilon_N^2 + \frac{t}{N}\right)\right\} \geqslant 1 - \exp(-t).$$

Step 5: Bound the empirical error by the L_2 error.

In this step, we show that with high probability, the empirical error $\|\widehat{r} - r\|_n$ is at most twice the L_2 error if r is in a given neighboring ball around r_0.

Let $g(r) = (r - r_0)^2$ for every $r \in \mathcal{F}_N$. Then since $g(r) = (r + r_0)(r - r_0)$, we have $|g(r)| \leqslant 3M_1|r - r_0| \geqslant 9M_1^2$, implying that $g(r)$ has a Lipschitz constant of $3M_1$, and $g(r)$ is a bounded function. Furthermore, if r is restricted to a radius with $\|r - r_0\|_{L_2(P)} \leqslant \xi$ for some fixed constant $\xi > 0$, then

$$\mathrm{Var}\{g(r)\} \leqslant \mathbb{E}\{g^2(r)\} \leqslant \mathbb{E}\{(r - r_0)^4\} \leqslant 9M_1^2\xi^2.$$

By applying Theorem 2.1 of Bartlett et al. (2005), which is based on Talagrand's concentration, for every r with $\|r - r_0\|_{L_2(P)} \leqslant \xi$, it holds that

$$\begin{aligned}\|r - r_0\|_n^2 - \|r - r_0\|_{L_2(P)} &\leqslant 3\widehat{\mathcal{R}}_n(g(r) : r \in \mathcal{F}_N, \|r - r_0\|_{L_2(P)} \leqslant \xi) + 3M_1\xi\sqrt{\frac{2t}{n}} + \frac{12M_1^2 t}{n} \\ &\leqslant 18M_1\widehat{\mathcal{R}}_n(\xi, \mathcal{F}_N^*) + 3M_1\xi\sqrt{\frac{2t}{n}} + \frac{12M_1^2 t}{n}, \qquad \text{(B.22)}\end{aligned}$$

with probability at least $1 - e^{-t}$ where the second inequality is due to $(r - r_0) \in \mathcal{F}_N^*$, the Lipschitz continuity of $g(r)$, and iterated expectations.

Now, suppose that the radius ξ satisfies

$$\xi^2 \geqslant 36M_1\widehat{\mathcal{R}}_n(\xi, \mathcal{F}_N^*), \quad \text{and} \quad \xi^2 \geqslant \frac{72M_1^2 t}{n}, \qquad \text{(B.23)}$$

then (B.22) implies that with probability at least $1 - e^{-t}$,

$$\|r - r_0\|_n^2 \leqslant \xi^2/2 + \xi^2/2 + \xi^2/6 < 2\xi^2 \quad \text{for all } r \text{ satisfies (B.23) and } \|r - r_0\|_{L_2(P)} \leqslant \xi.$$

As shown in the calculation of the previous step, for large enough n,

$$\xi = C_8(\xi_N + \epsilon_N^2 + \frac{t}{N})$$

satisfies the requirement in (B.23) for any given $t > 0$. This together with $\mathbb{P}(\|r - r_0\|_{L_2(P)} \leqslant \xi) > 1 - e^{-t}$ implies that $\mathbb{P}(\|r - r_0\|_n^2 \leqslant \xi) > 1 - 2e^{-t}$, which completes the proof of Theorem 4.1. □

B.3 Proof of Theorem 4.2

We will apply Yang-Barron's version of Fano's method (Yang and Barron, 1999) to derive the lower bound for the density ratio estimation.

Part 1. Let us first consider a sub-class of $\mathcal{M}^d(\beta_1, B_1)$ defined by

$$\mathcal{M}_1 = \left\{(\mathbb{P}_0, \mathbb{Q}) : \mathbb{P}_0 \text{ is the uniform distribution }, \ d\mathbb{Q}/d\mathbb{P} \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1), \ \inf_{x\in\mathcal{X}} d\mathbb{Q}(x) > c_0 > 0\right\}.$$

Then for any two distinct elements $(\mathbb{P}_0, \mathbb{Q}_1)$ and $(\mathbb{P}_0, \mathbb{Q}_2)$ in $\mathcal{M}_1$, their KL-divergence $D\left((\mathbb{P}_0, \mathbb{Q}_1)\|(\mathbb{P}_0, \mathbb{Q}_2)\right)$ can be bounded by

$$\begin{aligned}
D\left((\mathbb{P}_0, \mathbb{Q}_1)\|(\mathbb{P}_0, \mathbb{Q}_2)\right) =& D(\mathbb{Q}_1\|\mathbb{Q}_2) = \int_{x\in\mathcal{X}} \log\left(\frac{d\mathbb{Q}_1(x)}{d\mathbb{Q}_2(x)}\right) d\mathbb{Q}_1(x) \\
\leqslant& \int_{x\in\mathcal{X}} \left(\frac{d\mathbb{Q}_1(x)}{d\mathbb{Q}_2(x)} - 1\right) d\mathbb{Q}_1(x) = \int_{x\in\mathcal{X}} \left(\frac{d\mathbb{Q}_1(x)}{d\mathbb{Q}_2(x)}\right)^2 d\mathbb{Q}_2(x) - 1 \\
=& \int_{x\in\mathcal{X}} \left(\frac{(d\mathbb{Q}_1(x) - d\mathbb{Q}_2(x))}{d\mathbb{Q}_2(x)}\right)^2 d\mathbb{Q}_2 \\
\leqslant& c_0^{-1} \int_{x\in\mathcal{X}} \left(d\mathbb{Q}_1(x) - d\mathbb{Q}_2(x)\right)^2 dx. \qquad \text{(B.24)}
\end{aligned}$$

The above bound together with $D\left((\mathbb{P}_0^{\otimes n}, \mathbb{Q}_1^{\otimes m})\|(\mathbb{P}_0^{\otimes n}, \mathbb{Q}_2^{\otimes m})\right) = mD(\mathbb{Q}_1\|\mathbb{Q}_2)$ implies that for any $\varepsilon > 0$, the ε-covering number of $\mathcal{M}_1$ in the square-root KL divergence has an upper bound:

$$\mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_1) \leqslant \mathcal{N}_{L_2(\mu)}\left(\sqrt{\frac{c_0}{m}}\varepsilon, \mathcal{Q}_1\right),$$

where $\mathcal{Q}_1$ is the function class of $\mathbb{Q}$ that is the second element of $(\mathbb{P}, \mathbb{Q}) \in \mathcal{M}_1$. By definition, we know that $\mathcal{Q}_1$ is a sub-class of $\mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$, whose covering number is known from classical theory (see e.g., Giné and Nickl, 2021). Therefore we obtain

$$\log \mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_1) \leqslant \log \mathcal{N}_{L_2(\mu)}\left(\sqrt{\frac{c_0}{m}}\varepsilon, \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)\right) \asymp \left(\frac{B\sqrt{m}}{\varepsilon}\right)^{\frac{d}{\beta}}. \qquad \text{(B.25)}$$

Applying Yang-Barron's version of Fano's method, we choose $(\varepsilon_n, \delta_n)$ that satisfies

$$\varepsilon_m^2 \geqslant \mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_1) \quad \text{and} \quad \log M(2\delta_m; d, \Theta) \geqslant 4\varepsilon_m^2 + \log 2. \qquad \text{(B.26)}$$

Since the estimand is the density ratio function that belongs to $\mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$, we have

$$\log M(2\delta_m; d, \Theta) \asymp (\frac{1}{\delta_m})^{\frac{d}{\beta}}. \qquad \text{(B.27)}$$

With (B.25) and (B.27), $(\varepsilon_n, \delta_n)$ that ensures (B.26) can be specified as $\varepsilon_m^2 \asymp m^{\frac{d}{2\beta+d}}$ and $\delta_m^2 \asymp m^{-\frac{2\beta_1}{2\beta_1+d}}$.

According to Yang and Barron (1999), a minimax lower bound for the sub-class $\mathcal{M}_1$ is given by

$$\inf_{\widehat{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}_1} \mathbb{E}\|\widehat{r} - d\mathbb{Q}/d\mathbb{P}\|^2 \geqslant \frac{\delta_m^2}{2} \asymp m^{-\frac{2\beta_1}{2\beta_1+d}}. \tag{B.28}$$

Part 2. Let us first consider another sub-class of $\mathcal{M}^d(\beta_1, B_1)$ defined by

$$\mathcal{M}_2 = \left\{(\mathbb{P}, \mathbb{Q}_0) : \mathbb{Q}_0 \text{ is the uniform distribution },\ d\mathbb{Q}_0/d\mathbb{P} \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1),\ 0 < c_1 < d\mathbb{P}(x) < c_2 < \infty\right\}.$$

For any two distinct elements $(\mathbb{P}_1, \mathbb{Q}_0)$ and $(\mathbb{P}_2, \mathbb{Q}_0)$ in $\mathcal{M}_2$, with the same argument as in (B.24), we can obtain

$$D\left((\mathbb{P}_0, \mathbb{Q}_1)\|(\mathbb{P}_0, \mathbb{Q}_2)\right) \leqslant c_1^{-1} \int_{x\in\mathcal{X}} \left(d\mathbb{P}_1(x) - d\mathbb{P}_2(x)\right)^2 dx.$$

Since $d\mathbb{Q}_0(x) = 1$, we write $dP_i(x) = r_i^{-1}(x)$ with $r_i(x) \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$ for $i = 1, 2$. Then the above quantity can be upper bounded by

$$\begin{aligned} c_1^{-1} \int_{x\in\mathcal{X}} \left(d\mathbb{P}_1(x) - d\mathbb{P}_2(x)\right)^2 dx =& c_1^{-1} \int_{x\in\mathcal{X}} \left(\frac{1}{r_1(x)} - \frac{1}{r_2(x)}\right)^2 dx \\ \leqslant& c_2^4 c_1^{-1} \int_{x\in\mathcal{X}} (r_1(x) - r_2(x))^2 dx. \end{aligned} \tag{B.29}$$

Therefore, the square-root covering number of $\mathcal{M}_2$ in KL-divergence can be upper bounded by the covering number of $\mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$ in the $L_2(\mu)$-norm, leading to

$$\log \mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_2) \leqslant \log \mathcal{N}_{L_2(\mu)}\left(\sqrt{\frac{c_1}{c_2^4 n}}\varepsilon, \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)\right) \asymp \left(\frac{B\sqrt{n}}{\varepsilon}\right)^{\frac{d}{\beta}}. \tag{B.30}$$

for any $\varepsilon > 0$. The rest procedure is similar to Part I and we omit here for simplicity. The conclusion is for the sub-class $\mathcal{M}_2$, a minimiax lower bound is given by

$$\inf_{\widehat{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}_2} \mathbb{E}\|\widehat{r} - d\mathbb{Q}/d\mathbb{P}\|^2 \geqslant \frac{\delta_n^2}{2} \asymp n^{-\frac{2\beta_1}{2\beta_1+d}}. \tag{B.31}$$

Since $\mathcal{M}_1$ and $\mathcal{M}_2$ are both sub-class of $\mathcal{M}^d(\beta_1, B_1)$, their minimax lower bounds are also lower bounds of $\mathcal{M}^d(\beta_1, B_1)$. Combining the results in Part I and II, we obtain:

$$\inf_{\widehat{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}^d(\beta_1,B_1)} \mathbb{E}\|\widehat{r} - d\mathbb{Q}/d\mathbb{P}\|^2 \gtrsim n^{-\frac{2\beta_1}{2\beta_1+d}} + m^{-\frac{2\beta_1}{2\beta_1+d}} \asymp N^{-\frac{2\beta_1}{2\beta_1+d}}, \tag{B.32}$$

which completes the proof of Theorem 4.2. □

B.4 Proof of Theorem 4.3

For any given distribution $\tilde{P}_Y$ supported on $\mathbb{R}$ with a known density $\tilde{p}_0(y)$, we let $\tilde{P} = \tilde{P}_Y \times P_{\boldsymbol{X}}$ be the distribution of $(\tilde{Y}, \boldsymbol{X})$ for $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$ and $Y \sim \tilde{P}_Y$, which is independent of $\boldsymbol{X}$, and let

$$\tilde{r}_0(y, \boldsymbol{x}) = \frac{p_0(y, \boldsymbol{x})}{p_0(\boldsymbol{x})\tilde{p}_0(y)},$$

be the true density ratio function between P and $\tilde{P}$. Then, under Conditions 5 and 6, applying Theorem 4.2 leads to

$$\mathbb{E}_N\{(\hat{\tilde{r}} - \tilde{r}_0)^2\} = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}}\log(N)\right),$$

for the estimator $\hat{\tilde{r}}$. Since $\hat{p}_{Y|\boldsymbol{X}} = \hat{\tilde{r}}\tilde{p}_Y$ and $p_{Y|\boldsymbol{X}}(y, \boldsymbol{x}) = \tilde{r}_0\tilde{p}_Y$, where $\tilde{p}_Y$ is a bounded function, we have

$$\mathbb{E}_N\{(\hat{p}_{Y|\boldsymbol{X}} - p_{Y|\boldsymbol{X}})^2\} = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}}\log(N)\right). \tag{B.33}$$

For any $\boldsymbol{\theta}$, let $\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \int \mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)dy$ be the conditional mean function with the estimated conditional density function $\hat{p}_{Y|\boldsymbol{X}}$, then

$$\mathbb{E}_N\{\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta}) - \mathbf{m}_0(\boldsymbol{X}, \boldsymbol{\theta})\}^2 = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{X}}\left[\int g(y, \boldsymbol{X}_i, \boldsymbol{\theta})\{\hat{p}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)\}dy\right]^2.$$

Since there exists a constant $c > 0$ such that $p_0(y|\boldsymbol{X}) > c$, we have

$$\begin{aligned}
&\left\{\int |\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})||\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|dy\right\}^2 \\
&\leqslant c^{-1}\left\{\int |\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})||\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|\sqrt{p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)}dy\right\}^2 \\
&\leqslant c^{-1}\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}_i}(y|\boldsymbol{X}_i)|^2dy\int p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)dy \\
&\leqslant c^{-1}\log^2(N)\int |\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}_i}(y|\boldsymbol{X})|^2dy+ \\
&\quad + c^{-1}\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 I(\|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 > \log(N))|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|^2dy \\
&=: I_{1i} + I_{2i}, \quad \text{say}.
\end{aligned} \tag{B.34}$$

Note that as $\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)$ and $p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)$ are uniformaly bounded by a constant $M > 0$, we have

$$|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X})|^2 \leqslant 4M^2 + Mp_0(y|\boldsymbol{X}_i) \leqslant (4M^2m^{-1} + M)p_0(y|\boldsymbol{X}_i). \tag{B.35}$$

Hence, I_{2i} can be bounded by

$$I_{2i} = c^{-1}\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 I(\|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 > \log(N))|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|^2dy$$

$$
\begin{aligned}
&\lesssim \int \|\mathbf{g}(y,\boldsymbol{X}_i,\boldsymbol{\theta})\|^2 I(\|\mathbf{g}(y,\boldsymbol{X}_i,\boldsymbol{\theta})\|^2 > \log(N)) p_0(y|\boldsymbol{X}_i)dy \\
&\lesssim \left\{ \left(\int \|\mathbf{g}(y,\boldsymbol{X}_i,\boldsymbol{\theta})\|^4 p_0(y|\boldsymbol{X}_i)dy \right) \left(\int I(\|\mathbf{g}(y,\boldsymbol{X}_i,\boldsymbol{\theta})\|^2 > \log(N)) p_0(y|\boldsymbol{X}_i)dy \right) \right\}^{1/2} \\
&\lesssim N^{-1},
\end{aligned}
$$

which implies $\mathbb{E}_N(I_{2i}) \lesssim N^{-1}$. For the I_{1i} term, it can be seen that

$$
\mathbb{E}_N(I_{1i}) \lesssim \log^2(N)\mathbb{E}_N\{(\hat{p}_{Y|\boldsymbol{X}} - p_{Y|\boldsymbol{X}})^2\} = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}} \log^3(N) \right).
$$

Hence,

$$
\mathbb{E}_N\{\widehat{\mathbf{m}}(\boldsymbol{X},\boldsymbol{\theta}) - \mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})\}^2 = \mathbb{E}_N(I_{1i}) + \mathbb{E}_N(I_{2i}) = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}} \log^3(N) \right),
$$

which together with $\|\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}_i,\boldsymbol{\theta}) - \widehat{\mathbf{m}}(\boldsymbol{X}_i,\boldsymbol{\theta})\| = O_p(1/\sqrt{\kappa})$ complete the proof of Theorem 4.3. □

C Proofs for Section 5

C.1 Proof for the consistency of $\widehat{\boldsymbol{\theta}}$

Given the estimated $\widehat{\boldsymbol{\eta}}$, for any $\boldsymbol{\theta}$, we let $\boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}) = \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\boldsymbol{\eta})$, $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}) = N^{-1}\sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})$, and $\widehat{\Omega}(\boldsymbol{\theta},\boldsymbol{\eta}) = N^{-1}\sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})\boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})^{\mathrm{T}}$. With the EL estimator $\widehat{\boldsymbol{\theta}}$, we write $\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\eta}}) = \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})$, $\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})$, and $\widehat{\Omega}(\boldsymbol{\eta}) = \widehat{\Omega}(\widehat{\boldsymbol{\theta}},\boldsymbol{\eta})$.

Lemma C.1. *Under Conditions 1 and 2, if the estimation errors satisfy*

$$
\mathcal{E}_N(\widehat{r}) + \mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(1) \quad \text{and} \quad \mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(N^{-\frac{1}{2}}), \tag{C.1}
$$

then we have

$$
\frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\boldsymbol{\eta}_0) + o_p(1). \tag{C.2}
$$

Proof. Note that for each $i = 1,\cdots,N$,

$$
\boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}) - \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\boldsymbol{\eta}_0) = R_{1,i}(\widehat{\boldsymbol{\eta}}) + R_{2,i}(\widehat{\boldsymbol{\eta}}) + R_{3,i}(\widehat{\boldsymbol{\eta}}),
$$

where

$$
\begin{aligned}
R_{1,i}(\widehat{\boldsymbol{\eta}}) &= \left\{ \frac{\delta_i}{p} - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) \right\} \{\widehat{\mathbf{m}}(\boldsymbol{X}_i,\boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i,\boldsymbol{\theta})\}, \\
R_{2,i}(\widehat{\boldsymbol{\eta}}) &= \frac{1-\delta_i}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\{\widehat{\mathbf{m}}(\boldsymbol{X}_i,\boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i,\boldsymbol{\theta})\},
\end{aligned}
$$

$$R_{3,i}(\widehat{\boldsymbol{\eta}}) = \frac{1-\delta_i}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\{\mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta}) - \mathbf{m}_0(\boldsymbol{X}_i,\boldsymbol{\theta})\}.$$

Let $R_j(\widehat{\boldsymbol{\eta}}) = N^{-\frac{1}{2}}\sum_{i=1}^N R_{j,i}(\widehat{\boldsymbol{\eta}})$ for $j = 1, 2, 3$. Then (C.2) can be shown if $R_j(\widehat{\boldsymbol{\eta}}) = o_p(1)$ for $j = 1, 2, 3$. For the first term,

$$\begin{aligned}\mathbb{E}\{R_1^2(\widehat{\boldsymbol{\eta}})|\{\boldsymbol{X}_i\}_{i=1}^N\} &= \mathbb{E}_N\left[\left\{\frac{\delta_i}{p} - \frac{1-\delta_i}{1-p}r_0(\boldsymbol{X}_i)\right\}^2\{\widehat{\mathbf{m}}(\boldsymbol{X}_i,\boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i,\boldsymbol{\theta})\}^2\right] \\ &\lesssim \mathbb{E}_N\left[\{\widehat{\mathbf{m}}(\boldsymbol{X}_i,\boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i,\boldsymbol{\theta})\}^2\right] = \mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(1), \end{aligned} \tag{C.3}$$

where the first equality is due to

$$\mathbb{E}_N\{R_{1,i}(\widehat{\boldsymbol{\eta}})R_{1,i'}(\widehat{\boldsymbol{\eta}})|\{\boldsymbol{X}_i\}_{i=1}^N\} = 0,$$

for each $i \neq i'$, by the independence of $(\boldsymbol{X}_i, \delta_i)$ and $(\boldsymbol{X}_i', \delta_i')$, and $\mathbb{E}_N\left\{\frac{\delta_i}{p} - \frac{1-\delta_i}{1-p}r_0(\boldsymbol{X}_i)|\boldsymbol{X}_i\right\} = 0$ for each $1 \leqslant i \leqslant N$. Therefore, $R_1(\widehat{\boldsymbol{\eta}}) = o_p(1)$. For the second term, we have

$$\begin{aligned}R_2(\widehat{\boldsymbol{\eta}}) &= \sqrt{N}\mathbb{E}_N\left\{\frac{1-\delta_i}{1-p}\left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right|\left|\widehat{\mathbf{m}}(\boldsymbol{X}_i,\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{X}_i,\boldsymbol{\theta})\right|\right\} \\ &\lesssim \sqrt{N}\mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}_0}) = o_p(1),\end{aligned}$$

by the Cauchy-Schwarz inequality and (C.1). Finally, for the third term,

$$\begin{aligned}\mathbb{E}\{R_3^2(\widehat{\boldsymbol{\eta}})|\{\delta_i,\boldsymbol{X}_i\}_{i=1}^N\} &= \mathbb{E}_N\left[\frac{1-\delta_i}{(1-p)^2}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\{\mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta}_0) - \mathbf{m}_0(\boldsymbol{X}_i,\boldsymbol{\theta})\}^2|\{\delta_i,\boldsymbol{X}_i\}_{i=1}^N\right] \\ &\lesssim \mathbb{E}_n\left[\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\mathrm{Var}(\mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta})|\boldsymbol{X}_i)\}_{i=1}^N\right] \\ &\lesssim \mathcal{E}_n(\widehat{r}) = o_p(1).\end{aligned}$$

Therefore, we have $R_3(\widehat{\boldsymbol{\eta}}) = o_p(1)$. Since

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}) - \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta},\boldsymbol{\eta}) = R_1(\widehat{\boldsymbol{\eta}}) + R_2(\widehat{\boldsymbol{\eta}}) + R_3(\widehat{\boldsymbol{\eta}}),$$

the proof of Lemma C.1 is finished. □

Lemma C.2. *Under Conditions 1 and 2, if the estimation errors satisfy*

$$\mathcal{E}_N(\widehat{r}) + \mathcal{E}_N(\widehat{\mathbf{m}}) = o_p(1) \quad \textit{and} \quad \mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}) = o_p(N^{-\frac{1}{2}}), \tag{C.4}$$

then $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$.

Proof. The EL estimator $\widehat{\boldsymbol{\theta}}$ can be written as the solution to the saddle point problem (Newey and Smith, 2004):

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta}\sup_{\lambda\in\widehat{\Lambda}_N(\boldsymbol{\theta})}\frac{1}{N}\sum_{i=1}^N \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})), \tag{C.5}$$

where $\rho(v) = \log(1+v)$ and $\widehat{\Lambda}_N(\boldsymbol{\theta}) = \{\lambda : \lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) \in (-1, \infty)\}$. For any $\xi \in (1/\alpha, 1/2)$ where α is defined in Condition 2 (ii), let $\tilde{\lambda} = N^{-\xi}\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})/\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\|$. By Lemma A1 of Newey and Smith (2004), $\max_{i \leqslant N} |\tilde{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_i(\widehat{\boldsymbol{\eta}})| = o_p(1)$, and $\tilde{\lambda} \in \Lambda_N(\widehat{\boldsymbol{\theta}})$ with probability approaching 1. Thus, for any $\dot{\lambda} \in (\tilde{\lambda}, 0)$. Let ρ_k be the k-th derivative function of ρ. Then since $\rho_2(0) = -1$, with probability approaching 1 we have $\rho_2(\dot{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_i(\widehat{\boldsymbol{\eta}})) \geqslant -C(i = 1, \cdots, N)$ for some positive constant C_1. In addition, by the Cauchy-Schwarz inequality, Condition 2 (iii), and the uniform weak law of large numbers it can easily be derived that $N^{-1}\sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}))^{\otimes 2} \leqslant C_2\mathbf{I}_r$ for some positive constant C_2 with probability approaching 1, meaning that the largest eigenvalue of $N^{-1}\sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$ is bounded from above with probability approaching 1. Taking the Taylor expansion for $\rho(\tilde{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))$ at 0 gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \rho(\tilde{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) &= \tilde{\lambda}\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) + \frac{1}{2}\tilde{\lambda}^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^N \rho_2(\dot{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))^{\otimes 2}\right\}\tilde{\lambda} \\ &\geqslant N^{-\xi}\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - \frac{C_1C_2}{2}\|\tilde{\lambda}\|^2 \geqslant N^{-\xi}\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - C_3N^{-2\xi}, \quad \text{(C.6)}\end{aligned}$$

with probability approaching 1, where $C_3 = C_1C_2/2$.

By the similar arguments as Lemma A2 of Newey and Smith (2004), it can be shown that if for any $\bar{\boldsymbol{\theta}} \in \Theta$ such that $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$ and $\widehat{\boldsymbol{\Psi}}(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = O_p(N^{-\frac{1}{2}})$, then

$$\bar{\lambda} = \underset{\lambda \in \widehat{\Lambda}_N(\bar{\boldsymbol{\theta}})}{\arg\max}\, N^{-1}\frac{1}{N}\sum_{i=1}^N \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))$$

exists with probability approaching 1, also it holds that

$$\sup_{\lambda \in \widehat{\Lambda}_N(\boldsymbol{\theta}_0)} \frac{1}{N}\sum_{i=1}^N \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) = O_p(N^{-1}), \quad \text{and} \quad \bar{\lambda} = O_p(N^{-\frac{1}{2}}). \quad \text{(C.7)}$$

Setting $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Then, according to Lemma C.1,

$$\widehat{\boldsymbol{\Psi}}(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{\Psi}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) + o_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}}),$$

which shows that (C.7) holds with $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Using the definition of the saddle point $(\widehat{\boldsymbol{\theta}}, \bar{\lambda})$, the inequality (C.6), and the claim (C.7) with , we have

$$\begin{aligned}N^{-\xi}\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - C_3N^{-2\xi} &\leqslant \frac{1}{N}\sum_{i=1}^N \rho(\tilde{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) \\ &\leqslant \frac{1}{N}\sum_{i=1}^N \rho(\widehat{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) \quad \text{(C.8)} \\ &\leqslant \sup_{\lambda \in \widehat{\Lambda}_N(\boldsymbol{\theta}_0)} \frac{1}{N}\sum_{i=1}^N \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\eta}})) = O_p(N^{-1}),\end{aligned}$$

implying that $\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\| = O_p(N^{-1+\xi}) + O_p(N^{-\xi}) = O_p(N^{-\xi})$, since $\xi < 1/2$. Now, suppose ϵ_N is an arbitrary sequence that converges to 0 and let $\tilde{\lambda} = \epsilon_N \widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})$, which implies $\tilde{\lambda} = o_p(N^{-\xi})$. Then, similar to (C.8), we have

$$\tilde{\lambda}^{\mathrm{T}} \|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\| - C_3\|\tilde{\lambda}\|^2 = O_p(N^{-1}),$$

which implies $\epsilon_N(1 - C_3\epsilon_N)\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\|^2 = O_p(N^{-1})$. Since $1 - C_3\epsilon_N = O(1)$, we have $\epsilon_N\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\|^2 = O_p(N^{-1})$ for any sequence $\epsilon_N = o(1)$. Then it follows that $\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\| = O_p(N^{-\frac{1}{2}})$. Similar to Lemma C.1, it implies that $\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\boldsymbol{\eta}_0) = \widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}}) + o_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}})$.

According to the uniform weak law of large numbers,

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta},\boldsymbol{\eta}_0) - \boldsymbol{\Psi}(\boldsymbol{\theta},\boldsymbol{\eta}_0)\| = o_p(1),$$

which together with $\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}},\boldsymbol{\eta}_0) = o_p(1)$ implies $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}},\boldsymbol{\eta}_0) = o_p(1)$. Since $\boldsymbol{\Psi}(\boldsymbol{\theta},\boldsymbol{\eta}_0) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\Psi}(\boldsymbol{\theta},\boldsymbol{\eta}_0)$ is continuous with respect to $\boldsymbol{\theta}$, $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}},\boldsymbol{\eta}_0) = o_p(1)$ implies $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$, which establishes the consistency of $\widehat{\boldsymbol{\theta}}$. □

C.2 Proof of Theorem 5.1

The saddle point $(\widehat{\boldsymbol{\theta}},\widehat{\lambda})$ to (C.5) satisfies $Q_{1,N}(\widehat{\boldsymbol{\theta}},\widehat{\lambda}) = 0$ and $Q_{2,N}(\widehat{\boldsymbol{\theta}},\widehat{\lambda}) = 0$, where

$$Q_{1,N}(\widehat{\boldsymbol{\theta}},\widehat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\widehat{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}}), \quad \text{and}$$

$$Q_{2,N}(\widehat{\boldsymbol{\theta}},\widehat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\widehat{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})}\left(\frac{\partial\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\widehat{\lambda}.$$

By Taylor expansion of $Q_{1,N}(\widehat{\boldsymbol{\theta}},\widehat{\lambda}) = 0$ and $Q_{2,N}(\widehat{\boldsymbol{\theta}},\widehat{\lambda}) = 0$ around $(\boldsymbol{\theta}_0, 0)$, we have

$$0 = Q_{1,n}(\boldsymbol{\theta}_0,0) + \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda}\widehat{\lambda} + o_p(\delta_N), \quad \text{and}$$

$$0 = Q_{2,n}(\boldsymbol{\theta}_0,0) + \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda}\widehat{\lambda} + o_p(\delta_N),$$

where $\delta_N = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \|\widehat{\lambda}\|$, leading to

$$\begin{pmatrix}\hat{\lambda}\\ \widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix} = \mathbf{S}_N^{-1}\begin{pmatrix}-Q_{1,N}(\boldsymbol{\theta}_0,0)+o_p(\delta_N)\\ -Q_{2,N}(\boldsymbol{\theta}_0,0)o_p(\delta_N)\end{pmatrix} = \mathbf{S}_N^{-1}\begin{pmatrix}-Q_{1,N}(\boldsymbol{\theta}_0,0)+o_p(\delta_N)\\ o_p(\delta_N)\end{pmatrix}, \tag{C.9}$$

where

$$\mathbf{S}_N = \begin{pmatrix}\frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda} & \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}}\\ \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda} & \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}}\end{pmatrix},$$

and the partial derivatives are

$$\frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\boldsymbol{\Psi}_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\eta}})}{\partial\boldsymbol{\theta}}, \qquad \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Psi}_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\eta}})^{\otimes 2},$$
$$\frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}} = 0, \qquad \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\boldsymbol{\Psi}_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\eta}})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}.$$

Using the dominated convergence theorem, we can show that $\|\widehat{\mathbf{m}}-\mathbf{m}_0\| = o_p(1)$ implies $\|\partial\widehat{\mathbf{m}}/\partial\boldsymbol{\theta} - \partial\mathbf{m}_0/\partial\boldsymbol{\theta}\| = o_p(1)$. With the continuous mapping theorem and the law of large numbers, we have

$$\begin{aligned}\frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}} &= \boldsymbol{\Gamma} + o_p(1), & \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda} &= -\boldsymbol{\Omega} + o_p(1),\\ \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\boldsymbol{\theta}} &= 0, & \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial\lambda} &= \boldsymbol{\Gamma}^{\mathrm{T}} + o_p(1),\end{aligned} \tag{C.10}$$

where

$$\boldsymbol{\Gamma} = \mathbb{E}\left\{\frac{\partial\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)}{\partial\boldsymbol{\theta}}\right\} \quad\text{and}\quad \boldsymbol{\Omega} = \mathbb{E}\left\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)^{\otimes 2}\right\}.$$

From Lemma C.1, we have

$$Q_{1,N}(\boldsymbol{\theta}_0,0) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + o_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}}), \tag{C.11}$$

where the last equality is due to the CLT. Combining (C.9), (C.10), and (C.11), and using the continuous mapping theorem, we have

$$\begin{pmatrix}\hat{\lambda}\\ \widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix} = \left(\begin{pmatrix}-\boldsymbol{\Omega} & \boldsymbol{\Gamma}\\ \boldsymbol{\Gamma}^{\mathrm{T}} & 0\end{pmatrix}^{-1} + o_p(1)\right)\begin{pmatrix}Q_{1,N}(\boldsymbol{\theta}_0,0)+o_p(\delta_N)\\ o_p(\delta_N)\end{pmatrix}, \tag{C.12}$$

assuming that the block matrix on the right-hand side is invertible. Since $\delta_N = \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| + \|\widehat{\lambda}\|$, we know that $\delta_N = O_P(N^{-\frac{1}{2}})$, which further implies that

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = \left\{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}\right\}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\sqrt{N}Q_{1,N}(\boldsymbol{\theta}_0,0) + o_p(1) \xrightarrow{d} \mathcal{N}(0,\left\{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}\right\}^{-1}),$$

which completes the proof of Theorem 5.1. □

C.3 Proof of Theorem 5.2

Since for every $\boldsymbol{\theta} \in \Theta$, the optimal empirical weight p_i is given by

$$p_i = \frac{1}{N}\frac{1}{1+\lambda(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})},$$

where $\lambda(\boldsymbol{\theta})$ satisfies $Q_{1,N}(\boldsymbol{\theta}, \lambda(\boldsymbol{\theta})) = 0$, the log EL statistics with a given $\boldsymbol{\theta}$ can be written as

$$\ell_N(\boldsymbol{\theta}) = \log\{1 + \lambda(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})\}.$$

With $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, solving $Q_{1,N}(\boldsymbol{\theta}_0, \lambda) = 0$ gives

$$\lambda(\boldsymbol{\theta}_0) = \Omega^{-1} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(N^{-\frac{1}{2}}).$$

Taking the expansion of $\ell_N(\boldsymbol{\theta}_0)$ leads to

$$\ell_N(\boldsymbol{\theta}_0) = -\frac{N}{2} Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)\Omega^{-1} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1). \tag{C.13}$$

Using the characteristic of $\widehat{\lambda}$ given in (C.12), and expanding $\ell_N(\widehat{\boldsymbol{\theta}})$ gives

$$\ell_N(\boldsymbol{\theta}_0) = -\frac{N}{2} Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)\mathbf{A} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1), \tag{C.14}$$

where

$$\mathbf{A} = -\boldsymbol{\Omega}^{-1}\{\mathbf{I} + \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\}.$$

Therefore, $R_N(\boldsymbol{\theta}_0)$ is equivalent to

$$\begin{aligned} R_N(\boldsymbol{\theta}_0) &= N Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)(\mathbf{A} - \boldsymbol{\Omega}^{-1}) Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1) \\ &= N Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1). \end{aligned}$$

Note that $(-\boldsymbol{\Omega})^{-\frac{1}{2}}\sqrt{N} Q_{1,N}(\boldsymbol{\theta}_0, 0)$ weakly converges to a standard normal distribution, and

$$(-\boldsymbol{\Omega})^{-\frac{1}{2}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}(-\boldsymbol{\Omega})^{-\frac{1}{2}}$$

is symmetric and idempotent with the trace equal to r. Hence, $R_N(\boldsymbol{\theta}_0) \xrightarrow{d} \chi_r^2$, which completes the proof of Theorem 5.2. □

C.4 Proof of Theorem 5.3

We present the proof for the density ratio estimation, since the conditional density estimation can be proved similarly. Throughout this proof, we take the compact covariate domain $\mathcal{X} = [0, 1]^d$ without loss of generality. The main idea for the proof, which is similar to that of Theorem 6.1 of Jiao et al. (2023), is to project the data to a low-dimensional space, where the DNN can be used to approximate the low-dimensional function.

Let $d_\delta = O(d_{\mathcal{M}} \log(d/\delta)/\delta^2)$ be an integer such that $d_{\mathcal{M}} \leqslant d_\delta < d$ for any $\delta \in (0, 1)$. According to Theorem 3.1 of Baraniuk and Wakin (2009), there exists a matrix $\mathbf{A} \in \mathbb{R}^{d_\delta \times d}$, which maps a manifold in $\mathbb{R}^d$ into a low-dimensional space $\mathbb{R}^{d_\delta}$ and approximately preserves the distance. To be more specific, such the matrix A satisfies $\mathbf{A}\mathbf{A}^{\mathrm{T}} = (d/d_\delta)I_{d_\delta}$, and

$$(1 - \delta)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leqslant \|\mathbf{A}\boldsymbol{x}_1 - \mathbf{A}\boldsymbol{x}_2\|_2 \leqslant (1 + \delta)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$$

for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}_\rho$. Using $\mathbf{A}$ as a projection operator, we have

$$\mathbf{A}(\mathcal{M}_\rho) \subset \mathbf{A}([0,1]^d) \subset \left[-\sqrt{\frac{d}{d_\delta}}, \sqrt{\frac{d}{d_\delta}}\right]^{d_\delta}.$$

We now show that for every $\mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho)$, there exists a unique $\boldsymbol{x} \in \mathcal{M}_\rho$ such that $\mathbf{A}\boldsymbol{x} = \mathbf{a}$. Suppose that $\boldsymbol{x}' \in \mathcal{M}_\rho$ is another point with $\mathbf{A}\boldsymbol{x}' = \mathbf{a}$. Then $(1-\delta)\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leqslant \|\mathbf{A}\boldsymbol{x} - \mathbf{A}\boldsymbol{x}'\|_2 \leqslant (1+\delta)\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ implies that $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 = 0$. Therefore, for any $\mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho)$, we can define $\boldsymbol{x}(\mathbf{a}) = \mathcal{S}_{\mathbf{A}}(\{\boldsymbol{x} \in \mathcal{M}_\rho, \mathbf{A}\boldsymbol{x} = \mathbf{a}\})$, where $\mathcal{S}_{\mathbf{A}}(\cdot)$ is a set function that maps a set to a unique element of this set. It can be shown that $\mathcal{S}_{\mathbf{A}} : \mathbf{A}(\mathcal{M}_\rho) \to \mathcal{M}_\rho$ is a differentiable function, because for every $\mathbf{a}_1, \mathbf{a}_2 \in \mathbf{A}(\mathcal{M}_\rho)$,

$$\frac{1}{1+\delta}\|\mathbf{a}_1 - \mathbf{a}_2\| \leqslant \|\boldsymbol{x}(\mathbf{a}_1) - \boldsymbol{x}(\mathbf{a}_2)\| \leqslant \frac{1}{1-\delta}\|\mathbf{a}_1 - \mathbf{a}_2\|,$$

and the norm of the derivative of $\mathcal{S}_{\mathbf{A}}$ is in the range $[(1+\delta)^{-1}, (1-\delta)^{-1}]$.

Given a function $f_0 : [0,1]^d \to \mathbb{R}$, with the operator $\boldsymbol{x}(\cdot)$, we can define its low-dimensional representation $\tilde{f}_0 : \mathbf{A}(\mathcal{M}_\rho) \to \mathbb{R}$ by

$$\tilde{f}_0(\mathbf{a}) = f_0(\boldsymbol{x}(\mathbf{a})), \quad \text{for every } \mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho) \subset \mathbb{R}^{d_\delta}.$$

Since $r_0 \in \mathcal{H}^{\beta_1}([0,1]^d B_1)$, we have $\tilde{f}_0 \in \mathcal{H}^{\beta}(\mathbf{A}(\mathcal{M}_\rho), B_1/(1-\delta)^{\beta_1})$. Since $\mathcal{M}_\rho$ is a compact space and $\mathbf{A}$ is a linear operator, by Whitney extension theorem (Fefferman, 2006), there exists $\tilde{F}_0 \in \mathcal{H}^{\beta_1}(E_\delta, B_1/(1-\delta)^{\beta_1})$ with $E_\delta = [-\sqrt{d/d_\delta}, \sqrt{d/d_\delta}]^{d_\delta}$, such that $\tilde{F}_0(\mathbf{a}) = \tilde{f}_0(\mathbf{a})$ for every $\mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho)$. According to Theorem 3.3 of Jiao et al. (2023), for any $N, M \in \mathbb{N}_+$, there exists a function $\tilde{f} : E_\delta : \mathbb{R}$ belongs to the DNN function class with the ReLU activation function, whose width $W = 38(s+1)^2 d_\delta^{s+1} J\lceil\log_2(8J)\rceil$ and depth $D = 21(s+1)^2 M\lceil\log_2(8M)\rceil$, where $s = \lfloor\beta_1\rfloor$ such that

$$\sup_{\mathbf{a} \in E_\delta \backslash \Omega(E_\delta)} |\tilde{f}(\mathbf{a}) - \tilde{F}_0(\mathbf{a})| \leqslant 36\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}\, d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta}, \tag{C.15}$$

where $\Omega(E_\delta)$ is a subset of E_δ whose Lebesgue measure is arbitrarily small, as well as $\Omega := \{\boldsymbol{x} \in \mathcal{M}_\rho : \mathbf{A}\boldsymbol{x} \in \Omega(E_\delta)\}$ does.

Let $\tilde{f}_* = \tilde{f} \circ \mathbf{A}$, meaning that $\tilde{f}_*(\boldsymbol{x}) = \tilde{f}(\mathbf{A}\boldsymbol{x})$ for every $\boldsymbol{x} \in [0,1]^d$. Then, $\tilde{f}_*$ is also a DNN whose width and depth are the same as $\tilde{f}$. For every $\boldsymbol{x} \in \mathcal{M}_\rho \backslash \Omega$ and $\mathbf{a} = \mathbf{A}\boldsymbol{x}$, by the definition of $\mathcal{M}_\rho$, there exists a $\tilde{\boldsymbol{x}} \in \mathcal{M}_\rho$ such that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leqslant \rho$. Then,

$$\begin{aligned}
|\tilde{f}_*(\boldsymbol{x}) - r_0(\boldsymbol{x})| &\leqslant |\tilde{f}(\mathbf{A}\boldsymbol{x}) - \tilde{F}_0(\mathbf{A}\boldsymbol{x})| + |\tilde{F}_0(\mathbf{A}\boldsymbol{x}) - \tilde{F}_0(\mathbf{A}\tilde{\boldsymbol{x}})| + |\tilde{F}_0(\mathbf{A}\tilde{\boldsymbol{x}}) - r_0(\boldsymbol{x})| \\
&\leqslant 36\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}\, d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta} + \frac{B_1}{1-\delta}\|\mathbf{A}\boldsymbol{x} - \mathbf{A}\tilde{\boldsymbol{x}}\| + |r_0(\tilde{\boldsymbol{x}}) - r_0(\boldsymbol{x})| \\
&\leqslant 36\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}\, d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta} + \frac{\rho B_1}{1-\delta}\sqrt{d/d_\delta} + \rho B_1
\end{aligned}$$

$$\leqslant (36 + C_\rho)\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta},$$

where the second inequality is by (C.15), the smoothness of $\tilde{F}_0$, and the definition of $\tilde{F}_0$. The third inequality is because $\|\mathbf{A}\| = \sqrt{d/d_\delta}$ and the smoothness of r_0. The positive constant C_ρ is taken such that $\rho \leqslant C_\rho(1-\delta)^{1-\beta}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta}(\sqrt{d/d_\delta}+1-\delta)^{-1}$. Since $P_{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure, we have

$$\|\tilde{f}_* - r_0\|_{L_2(P)}^2 \leqslant (36+C_\rho)^2\frac{B_1^2}{(1-\delta)^{2\beta_1}}(s+1)^4 d d_\delta^{3s}(JM)^{-4\beta_1/d_\delta}. \tag{C.16}$$

As shown in the proof of Theorem 4.1,

$$\mathbb{E}\{\|\widehat{r} - r_0\|_n^2\} \leqslant C\left(\frac{\text{Pdim}(\mathcal{F}_N)\log(N)}{N} + \epsilon_N^2\right),$$

for some positive constant C, where $\epsilon_N^2 = \inf_{f\in\mathcal{F}_N}\|\tilde{f}_* - r_0\|_{L_2(P)}^2$. According to Bartlett et al. (2019), for the DNN class $\mathcal{F}_N$ with width W and depth D, its pseodu-dimension is bounded by

$$\text{Pdim}(\mathcal{F}_N) \leqslant C_1 W^2 D^2 \log(W^2 D),$$

where C_1 is a positive constant. The approximation error $\epsilon_N^2 \leqslant \|\tilde{f}_* - r_0\|_{L_2(P)}^2$ is bounded by the right-hand side of (C.16). Therefore,

$$\mathbb{E}\{\|\widehat{r} - r_0\|_n^2\} \leqslant C_2\left(\frac{W^2D^2\log(W^2D)\log(N)}{N} + \frac{B_1^2}{(1-\delta)^{2\beta_1}}(s+1)^4 d d_\delta^{3s}(JM)^{-4\beta_1/d_\delta}\right).$$

Choosing $J = 1$ and $M = N^{D_\delta}$ with $D_\delta = d_\delta/(2(d_\delta + 2\beta_1))$ leads to

$$\mathbb{E}\{\|\widehat{r} - r_0\|_n^2\} \leqslant C_3 d d_\delta^{3\lfloor\beta_1\rfloor} N^{-\frac{2\beta_1}{2\beta_1+d_\delta}},$$

where the positive constant C_3 does not depend on N or d, which completes the proof. □

C.5 Proof of Theorem 5.4

With our Lemma C.1 and Theorem 5.3, the proof is obtained by assigning $\alpha(k) = 0$ and $M = 1$ in Theorem 2 of Chang et al. (2015), and hence is omitted here.

D Proofs for Section 5

D.1 Proof of Theorem 6.1

Lemma D.1. *Under Conditions 1–3, 4 (iii), 9, and 10,*

$$\sqrt{N}\mathbb{E}\left\{\frac{1-\delta}{1-p}\widehat{r}(\boldsymbol{X})\mathbf{m}(\boldsymbol{X})\right\} = \frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{\frac{\delta_i}{p}\mathbf{m}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p}r_0(\boldsymbol{X}_i)\mathbf{m}(\boldsymbol{X}_i)\right\} + o_p(1), \tag{D.1}$$

where the expectation is taken with respect to $\boldsymbol{X}$*.*

Recall that the criterion function for the estimation of r is defined as

$$\widehat{L}_N(r) = \frac{1}{N}\sum_{i=1}^{N} \ell(\delta_i, \boldsymbol{X}_i; r),$$

where

$$\ell(\delta, \boldsymbol{X}; r) = \frac{1-\delta}{1-p}\ell_1(\boldsymbol{X}; r) - \frac{\delta}{p}\ell_2(\boldsymbol{X}; r).$$

The directional derivative of $\ell(\delta, \boldsymbol{X}; r)$ with respect to r in the direction $u \in L_2(P)$ is given by

$$\begin{aligned}
\frac{d}{du}\ell(\delta, \boldsymbol{X}; r)[u] &:= \lim_{t\to 0}\frac{\ell(\delta, \boldsymbol{X}; r+tu) - \ell(\delta, \boldsymbol{X}; r)}{t} \\
&= \left\{\frac{1-\delta}{1-p}\frac{\partial}{\partial r}\ell_1(\boldsymbol{X}; r) - \frac{\delta}{p}\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r)\right\} u(\boldsymbol{X}) \\
&=: \ell^{(1)}(\delta, \boldsymbol{X}; r)u(\boldsymbol{X}), \quad \text{say.}
\end{aligned} \tag{D.2}$$

According to Condition 9. (ii), we have

$$\ell^{(1)}(\delta, \boldsymbol{X}; r) = \frac{1-\delta}{1-p}\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r)r(\boldsymbol{X}) - \frac{\delta}{p}\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r).$$

The first-order approximation error for $\ell(\delta, \boldsymbol{X}; r_0)$ is denoted as

$$e(\delta, \boldsymbol{X}, r-r_0) = \ell(\delta, \boldsymbol{X}; r) - \ell(\delta, \boldsymbol{X}; r_0) - \frac{d}{du}\ell(\delta, \boldsymbol{X}; r_0)[r-r_0].$$

With the above notations, for any $r \in \mathcal{F}_N$, it holds that

$$\begin{aligned}
\widehat{L}_N(r) =& \widehat{L}_N(r_0) + \{\widehat{L}_N(r) - \widehat{L}_N(r_0)\} \\
=& \widehat{L}_N(r_0) + \frac{1}{N}\sum_{i=1}^{N}\{\ell(\delta_i, \boldsymbol{X}_i; r) - \ell(\delta_i, \boldsymbol{X}_i; r_0)\} \\
=& \widehat{L}_N(r_0) + \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[r-r_0] + e(\delta_i, \boldsymbol{X}_i; r-r_0)\right\} \\
=& \widehat{L}_N(r_0) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[r-r_0]\right) + \frac{1}{N}\sum_{i=1}^{N} e(\delta_i, \boldsymbol{X}_i; r-r_0),
\end{aligned} \tag{D.3}$$

where the last equality is because

$$\mathbb{E}\left\{\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[r-r_0]\right\} = 0. \tag{D.4}$$

We will employ the Cramer-Wald device to establish (D.1). For any $\boldsymbol{v} \in \mathbb{R}^p$ with $\|\boldsymbol{v}\| = 1$, we define $\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{x}) = \mathbf{m}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{v} \cdot (\partial \ell_2(\boldsymbol{x}, r)/\partial r)^{-1}$. For any $r \in \mathcal{F}_N$, let

$$\bar{r}(r, \epsilon_N) = (1-\epsilon_N)r + \epsilon_N(r_0 + \tilde{m}_{\boldsymbol{v},\ell_2})$$

be a local alternative value around r and

$$\Pi_{\mathcal{F}_n}\bar{r}(r,\epsilon_N) = (1-\epsilon_N)r + \epsilon_N(r_* + \tilde{m}_*),$$

where $r_* = \arg\min_{r\in\mathcal{F}_N}\|r-r_0\|_{L_2(F)}$ and $\tilde{m}_* = \arg\min_{m\in\mathcal{F}_N}\|m-\tilde{m}_{\boldsymbol{v},\ell_2}\|_{L_2(F)}$. In the light of Condition 10, we have $\Pi_{\mathcal{F}_n}\bar{r}(r,\epsilon_N)\in\mathcal{F}_N$ and

$$\sup_{r\in\mathcal{F}_N}\|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(r,\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(r,\epsilon_N)\|_{L_2(F)} = o(\epsilon_N\cdot N^{-\frac{1}{4}}). \tag{D.5}$$

By substituting r with $\widehat{r}$ and $\Pi_{\mathcal{F}_n}\bar{r}(\widehat{r},\epsilon_N)$, respectively, we obtain

$$\widehat{L}_N(\widehat{r}) = \widehat{L}_N(r_0) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\widehat{r}-r_0]\right) + \frac{1}{N}\sum_{i=1}^N e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0) \tag{D.6}$$

and

$$\begin{aligned}\widehat{L}_N(\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) =& \widehat{L}_N(r_0) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0]\right)\\ &+ \frac{1}{N}\sum_{i=1}^N e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0).\end{aligned} \tag{D.7}$$

Subtracting (D.6) from (D.7) gives

$$\begin{aligned}\widehat{L}_N(\widehat{r}) =& \widehat{L}_N(\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\widehat{r}-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)]\right)\\ &+\frac{1}{N}\sum_{i=1}^N\left\{e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0) - e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0)\right\}.\end{aligned} \tag{D.8}$$

We will prove later in Subsection D.2 that

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N\left\{e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0) - e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0)\right\}\\ &= \epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right).\end{aligned} \tag{D.9}$$

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

$$\widehat{L}_N(\widehat{r}) - \widehat{L}_N(\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) \leqslant O(\epsilon_N^2),$$

which together with (D.8) and (D.9) yield

$$\begin{aligned}&\frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\widehat{r}-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)]\right)\\ &-\epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right) \leqslant O(\epsilon_N^2).\end{aligned} \tag{D.10}$$

For the term $\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)]\right)$, we have

$$\begin{aligned}
&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)]\right)\\
=&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)]\right)+\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)]\right)\\
=&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)]\right)+o_p(\epsilon_N),
\end{aligned}$$

where the last equality is due to (D.5) and the Chebyshev inequality. By the definition of $\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)$, we have

$$\begin{aligned}
&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)]\right)\\
=&\epsilon_N\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-r_0]\right)-\epsilon_N\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\tilde{m}_{\boldsymbol{v},\ell}]\right).
\end{aligned}\tag{D.11}$$

We now show that $\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-r_0]\right)=o_p(1)$. By (D.2),

$$\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\hat{r}-r_0]=\ell^{(1)}(\delta_i,\boldsymbol{X}_i;r_0)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}.$$

Let

$$\tilde{\mathcal{F}}_N=\left\{\ell^{(1)}(\delta,\boldsymbol{x};r_0)\{r(\boldsymbol{x})-r_0(\boldsymbol{x})\}:\ r\in\mathcal{F}_N,\|r-r_0\|_{L_2(F)}\leqslant\delta_N\right\},$$

then it is evident that

$$\log N_{[\,]}(\epsilon,\tilde{\mathcal{F}}_N,L_2(F))\lesssim\log N_{[\,]}(\epsilon,\mathcal{F}_N,L_2(F))$$

for any $\epsilon>0$. Therefore, the bracketing number of $\tilde{F}_N$ satisfies

$$\begin{aligned}
J_{[\,]}(\delta_N,\tilde{\mathcal{F}}_N,L_2(F))&=\int_0^{\delta_N}\sqrt{1+\log N_{[\,]}(\epsilon,\tilde{\mathcal{F}}_N,L_2(F))}d\epsilon\\
&\lesssim\int_0^{\delta_N}\sqrt{1+\log N_{[\,]}(\epsilon,\mathcal{F}_N,L_2(F))}d\epsilon\\
&=J_{[\,]}(\delta_N,\mathcal{F}_N,L_2(F))=o(1)
\end{aligned}$$

by Condition 10 (iii). Also, for every $f\in\tilde{F}_N$, it holds that $\|f\|_\infty=O(1)$ and $\|f\|_{L_2(F)}=O(\delta_N)$. By applying Lemma 3.4.2 of van der Vaart and Wellner (1996), we have

$$\mathbb{E}\|\mathbb{G}_N\|_{\tilde{F}_N}\lesssim J_{[\,]}(\delta_N,\tilde{\mathcal{F}}_N,L_2(F))\left(1+\frac{J_{[\,]}(\delta_N,\tilde{\mathcal{F}}_N,L_2(F))}{\delta_N^2\sqrt{N}}O(1)\right)=o(1),$$

which, by the Markov inequality, implies that

$$\sup_{r\in\mathcal{F}_N}\mathbb{G}_N\left(\ell^{(1)}(\delta,\boldsymbol{x};r_0)\{r(\boldsymbol{x})-r_0(\boldsymbol{x})\}\right)=o_p(1),\tag{D.12}$$

meaning that

$$\epsilon_N \mathbb{G}_N \left(\frac{d}{dr} \ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - r_0] \right) = o_p(\epsilon_N).$$

In addition, plugging $\tilde{m}_{\boldsymbol{v},\ell}(\boldsymbol{X}_i) = m_{\boldsymbol{v}}(\boldsymbol{X}_i) \cdot \{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i, r_0)\}^{-1}$ into the directional derivative specified in (D.2) gives

$$-\mathbb{G}_N \left(\frac{d}{dr} \ell(\delta_i, \boldsymbol{X}_i; r_0)[\tilde{m}_{\boldsymbol{v},\ell}] \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{\delta_i}{p} m_{\boldsymbol{v}}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right\}.$$

Combining the above results gives

$$\begin{aligned}
&\mathbb{G}_N \left(\frac{d}{dr} \ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \Pi_{\mathcal{F}_N} \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)] \right) \\
=& \frac{\epsilon_N}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{\delta_i}{p} m_{\boldsymbol{v}}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right\} + o_p(\epsilon_N).
\end{aligned}$$

Therefore, multiplying the both sides of (D.10) by $\sqrt{N}/\epsilon_N$ leads to

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{\delta_i}{p} m_{\boldsymbol{v}}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right\} + o_p(\epsilon_N) \\
&- \sqrt{N}(1-\epsilon_N) \mathbb{E} \left(\frac{1-\delta}{1-p} \{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X})\} m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right) = o_p(1) + O_p\left(\frac{\epsilon_N}{\sqrt{N}} \right) = o_p(1),
\end{aligned}$$

which completes the proof of Lemma D.1.

D.2 Proof of (D.9)

First, for any candidate r we can decompose $e(\delta, \boldsymbol{X}, r - r_0)$ as

$$\begin{aligned}
&e(\delta, \boldsymbol{X}, r - r_0) \\
=& \ell(\delta, \boldsymbol{X}; r) - \ell(\delta, \boldsymbol{X}; r_0) - \frac{d}{du} \ell(\delta, \boldsymbol{X}; r_0)[r - r_0] \\
=& \frac{1}{2} \left\{ \frac{1-\delta}{1-p} \frac{\partial^2}{\partial r^2} \ell_1(\boldsymbol{X}; r_0) - \frac{\delta}{p} \frac{\partial^2}{\partial r^2} \ell_2(\boldsymbol{X}; r_0) \right\} \{r(\boldsymbol{X}) - r_0(\boldsymbol{X})\}^2 + R(\delta, \boldsymbol{X}, r),
\end{aligned} \tag{D.13}$$

where the remainder term $R(\delta, \boldsymbol{X}, r)$ is

$$R(\delta, \boldsymbol{X}, r) = \frac{1}{2} \int_{r_0(\boldsymbol{X})}^{r(\boldsymbol{X})} \left\{ \frac{1-\delta}{1-p} \frac{\partial^3}{\partial r^3} \ell_1(\boldsymbol{X}; t) - \frac{\delta}{p} \frac{\partial^3}{\partial r^3} \ell_2(\boldsymbol{X}; t) \right\} \{r(\boldsymbol{X}) - t\}^2 dt,$$

and the last equality of (D.13) is due to the following Taylor's theorem

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2}(b-a)^2 + \int_a^b \frac{f'''(t)}{2}(b-t)^2 dt.$$

Let

$$\ell^{(2)}(\delta, \boldsymbol{X}) := \frac{1-\delta}{1-p}\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X}; r_0) - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0).$$

Then, according to Condition 9.(i), we have

$$\begin{aligned}
&\frac{\partial}{\partial r}\ell_1(\boldsymbol{X}; r_0) = r_0(\boldsymbol{X})\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r_0),\\
&\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X}; r_0) = r_0(\boldsymbol{X})\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0) + \frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r_0),
\end{aligned}$$

which implies that

$$\ell^{(2)}(\delta, \boldsymbol{X}) = \frac{1-\delta}{1-p}\left\{r_0(\boldsymbol{X})\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0) + \frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r_0)\right\} - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0). \tag{D.14}$$

The last term in (D.8) can be written as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \{e(\delta_i, \boldsymbol{X}_i; \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - r_0) - e(\delta_i, \boldsymbol{X}_i; \widehat{r} - r_0)\}\\
=&\frac{1}{2N}\sum_{i=1}^N \left\{\frac{1-\delta}{1-p}\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X}; r_0) - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0)\right\}\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
&- \frac{1}{2N}\sum_{i=1}^N \left\{\frac{1-\delta}{1-p}\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X}; r_0) - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0)\right\}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
&+ \frac{1}{N}\sum_{i=1}^N \{R(\delta_i, \boldsymbol{X}_i, \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)) - R(\delta_i, \boldsymbol{X}_i, \widehat{r})\}\\
=:&E_{1,N} + E_{2,N} + E_{3,N}, \quad \text{say.}
\end{aligned}$$

For the term $\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2$, we have

$$\begin{aligned}
&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
=&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) + \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
=&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) + (1-\epsilon_N)(\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)) + \epsilon_N\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}^2\\
=&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i)\}^2 + (1-\epsilon_N)^2\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2 + \epsilon_N^2\tilde{m}_{\boldsymbol{v},\ell_2}^2(\boldsymbol{X}_i)\\
&+ 2(1-\epsilon_N)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i)\}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\\
&+ 2\epsilon_N\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
&+ 2(1-\epsilon_N)\epsilon_N\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i).
\end{aligned} \tag{D.15}$$

Using (D.15), we can decompose $E_{1,N} + E_{2,N}$ as

$$\begin{aligned}
&E_{1,N} + E_{2,N}\\
=&\frac{1}{2N}\sum_{i=1}^N \ell^{(2)}(\delta_i, \boldsymbol{X}_i)[\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2 - \{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{2N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)\}^2\\
&+\frac{(1-\epsilon_N)^2-1}{2N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}^2+\frac{\epsilon_N^2}{2N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\tilde{m}^2_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
&+\frac{1-\epsilon_N}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)\}\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\\
&+\frac{\epsilon_N}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
&+\frac{\epsilon_N(1-\epsilon_N)}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
=&\frac{1}{2}\mathbb{E}[\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)\}^2]\{1+o_p(1)\}\\
&+\frac{\epsilon_N^2-2\epsilon_N}{2}\mathbb{E}[\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}^2]\{1+o_p(1)\}+\frac{\epsilon_N^2}{2}\mathbb{E}\{\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\tilde{m}^2_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}\{1+o_p(1)\}\\
&+(1-\epsilon_N)\mathbb{E}[\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)\}\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}]\{1+o_p(1)\}\\
&+\epsilon_N\mathbb{E}\{\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}\\
&+\frac{\epsilon_N(1-\epsilon_N)}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
\leqslant&O_p(\epsilon_N^2\delta_N^2)+O_p(\epsilon_N\delta_N^2)+O_p(\epsilon_N^2)+O_p(\epsilon_N\delta_N^2)+O_p(\epsilon_N^2\delta_N)\\
&+\frac{\epsilon_N(1-\epsilon_N)}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i),
\end{aligned}
\tag{D.16}
$$

where the expectations are taken with respect to $(\delta_i,\boldsymbol{X}_i)$, and the last equality is by the uniform boundness of $\ell^{(2)}(\delta,\boldsymbol{X})$, the approximation error in (D.5), and the bounded moment of $\|\tilde{m}_{\boldsymbol{v},\ell_2}\|^2$. For the last term in (D.16), we note that

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
=&\frac{1}{\sqrt{N}}\mathbb{G}_N\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right)\\
&+\mathbb{E}\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right),
\end{aligned}
\tag{D.17}
$$

where the expectation is taken with respect to $(\delta_i,\boldsymbol{X}_i)$. By the stochastic equicontinuity which can be derived with the similar arguments as for (D.12), we can obtain

$$
\mathbb{G}_N\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right)=o_p(1). \tag{D.18}
$$

In the light of (D.14) and $\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)=m_{\boldsymbol{v}}(\boldsymbol{X}_i)\cdot\{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i,r_0)\}^{-1}$, the expectation term can be written as

$$
\begin{aligned}
&\mathbb{E}\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right)\\
=&\mathbb{E}\left(\frac{1-\delta}{1-p}\left\{r_0(\boldsymbol{X})\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0)+\frac{\partial}{\partial r}\ell_2(\boldsymbol{X};r_0)\right\}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\cdot\{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i,r_0)\}^{-1}\right)\\
&-\mathbb{E}\left\{\frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\cdot\{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i,r_0)\}^{-1}\right\}\\
=&\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right),
\end{aligned}
\tag{D.19}
$$

where the last equality is due to $\mathbb{E}\{(1-\delta)r_0(\boldsymbol{X})f(\boldsymbol{X})\}=\mathbb{E}\{\delta f(\boldsymbol{X})\}$ for any $f(\boldsymbol{X})$. Combining (D.16), (D.17), (D.18), and (D.19), and taking the convergence rate $\delta_N=o_p(N^{-\frac{1}{4}})$, we obtain

$$
E_{1,N}+E_{2,N}=\epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right)+o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right). \tag{D.20}
$$

For the term $E_{3,N}$, we let

$$
\ell^{(3)}(\delta,\boldsymbol{X};t)=\frac{1-\delta}{1-p}\frac{\partial^3}{\partial r^3}\ell_1(\boldsymbol{X};t)-\frac{\delta}{p}\frac{\partial^3}{\partial r^3}\ell_2(\boldsymbol{X};t).
$$

Due to $\frac{\partial}{\partial r}\ell_1(\boldsymbol{X},t)=t\cdot\frac{\partial}{\partial r}\ell_2(\boldsymbol{X},t)$ imposed in Condition 9, we have

$$
\ell^{(3)}(\delta,\boldsymbol{X};t)=\frac{1-\delta}{1-p}\left\{t\cdot\frac{\partial^3}{\partial r^3}\ell_2(\boldsymbol{X};t)+\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};t)+\frac{\partial}{\partial r}\ell_2(\boldsymbol{X};t)\right\}-\frac{\delta}{p}\frac{\partial^3}{\partial r^3}\ell_2(\boldsymbol{X};t), \tag{D.21}
$$

which is uniformly bounded by some positive constant c_ℓ according to Condition 9.(ii).

then $E_{3,N}$ can be decomposed as

$$
\begin{aligned}
E_{3,N}=&\frac{1}{N}\sum_{i=1}^{N}\{R(\delta_i,\boldsymbol{X}_i,\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N))-R(\delta_i,\boldsymbol{X}_i,\widehat{r})\}\\
=&\frac{1}{2N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)}\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-t\}^2dt\\
&-\frac{1}{2N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)}\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)\{\widehat{r}(\boldsymbol{X}_i)-t\}^2dt\\
=&\frac{1}{2N}\sum_{i=1}^{N}\int_{\widehat{r}(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)}\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-t\}^2dt\\
&-\frac{1}{2N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)}\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)[\{\widehat{r}(\boldsymbol{X}_i)-t\}^2-\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-t\}^2]dt\\
=:&D_{1,N}+D_{2,N},\quad\text{say.}
\end{aligned}
$$

For the term $D_{1,N}$, we have

$$
\begin{aligned}
2|D_{1,N}| &= \frac{1}{N}\left|\sum_{i=1}^{N}\int_{\hat{r}(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)} \ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-t\}^2 dt\right| \\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}\left|\int_{\hat{r}(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)} \{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-t\}^2 dt\right| \\
&= \frac{c_\ell}{N}\sum_{i=1}^{N}(1-s_i)\left|\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)\}^3\right| \quad (\text{for some } s_i\in(0,1)) \\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)|^3 \\
&\leqslant \frac{2c_\ell}{N}\sum_{i=1}^{N}\{|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)|^3+|\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)|^3\},
\end{aligned}
$$

where the first inequality is from the uniform boundness of $\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)$, the second equality is by applying the mean value theorem, and the last inequality is from the inequality $(a+b)^3 \leqslant 2(a^3+b^3)$ for any positive a and b. From (D.5) it can be easily derived that $\max_{1\leqslant i\leqslant N}|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)| = o_p(1)$. For the term $|\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)|$, we have

$$
\frac{1}{N}\sum_{i=1}^{N}|\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)| = \epsilon_N\frac{1}{N}\sum_{i=1}^{N}\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)-\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\} = O_p(\epsilon_N), \quad \text{(D.22)}
$$

$$
\frac{1}{N}\sum_{i=1}^{N}|\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)|^2 = \epsilon_N^2\frac{1}{N}\sum_{i=1}^{N}\{\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)-\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}^2 = O_p(\epsilon_N^2), \quad \text{(D.23)}
$$

Using Lemma 2 of Owen (1990), it holds that $\max_{1\leqslant i\leqslant N}|\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)| = o_p(\sqrt{N})$, which together with the uniform boundness of $\hat{r}$ and r_0 and $\epsilon_N = o_p(N^{-\frac{1}{2}})$ imply that

$$
\max_{1\leqslant i\leqslant N}|\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-\hat{r}(\boldsymbol{X}_i)| = \epsilon_N\max_{1\leqslant i\leqslant N}|\hat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)-\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)| = o_p(1). \quad \text{(D.24)}
$$

Therefore, $|D_{1,N}|$ can be bounded by

$$
|D_{1,N}| \leqslant o_p(\epsilon_N^2\delta_N^2)+o_p(\epsilon_N^2) = o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right), \quad \text{(D.25)}
$$

where the equality is due to $\epsilon_N = o(N^{-\frac{1}{2}})$ and $\delta_N = o(N^{-\frac{1}{4}})$.

For the term $D_{2,N}$, we have

$$
\begin{aligned}
2|D_{2,N}| &= \frac{1}{N}\left|\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\hat{r}(\boldsymbol{X}_i)} \ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)[\{\hat{r}(\boldsymbol{X}_i)-t\}^2-\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-t\}^2]dt\right| \\
&= \frac{1}{N}\left|\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\hat{r}(\boldsymbol{X}_i)} \ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)[\{\hat{r}(\boldsymbol{X}_i)-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)\}\{\hat{r}(\boldsymbol{X}_i)+\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\hat{r},\epsilon_N)-2t\}]dt\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^N \int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)} \left|\widehat{r}(\boldsymbol{X}_i) - \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \left|\widehat{r}(\boldsymbol{X}_i) + \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - 2t\right| dt \\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^N \{\left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\widehat{r}(\boldsymbol{X}_i) - \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \\
&\qquad \cdot \left|\widehat{r}(\boldsymbol{X}_i) + \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - 2\{s_i\widehat{r}(\boldsymbol{X}_i) + (1-s_i)r_0(\boldsymbol{X}_i)\}\right|\} \quad (\text{for some } s_i \in (0,1)) \\
&= \frac{c_\ell}{N}\sum_{i=1}^N \{\left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\widehat{r}(\boldsymbol{X}_i) - \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \\
&\qquad \cdot \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i) + 2(1-s_i)\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\right|\} \\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^N \{\left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left(\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| + \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right|\right) \\
&\qquad \cdot \left(\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| + \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right| + 2\left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right|\right)\} \\
&= \frac{c_\ell}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|^2 \\
&\quad + \frac{c_\ell}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right| \\
&\quad + \frac{2c_\ell}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right|^2 \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \\
&\quad + \frac{c_\ell}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right| \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \\
&\quad + \frac{c_\ell}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right|^2 \\
&\quad + \frac{2c_\ell}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right|^2 \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right|, \qquad \text{(D.26)}
\end{aligned}
$$

where the first inequality is from the uniform boundness of $\ell^{(3)}(\delta_i, \boldsymbol{X}_i; t)$ and the second inequality is by applying the mean value theorem. By the uniform boundness of $\widehat{r}$ and r_0, the approximation error in (D.5), (D.23), $\|\widehat{r} - r_0\|_{L_2(P)} = O_p(\delta_N)$, and the Cauchy-Schwarz inequality, we can obtain

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|^2 = O_p(\epsilon_N^2\delta_N^2), \\
&\frac{1}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right| \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| \left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \widehat{r}(\boldsymbol{X}_i)\right| = O_p(\epsilon_N^2\delta_N), \\
&\frac{1}{N}\sum_{i=1}^N \left|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\right|^2 \left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right| = O_p(\epsilon_N\delta_N^2).
\end{aligned}
$$

By the uniform boundness of $\widehat{r}$ and r_0, $\|\widehat{r} - r_0\|_{L_2(P)} = O_p(\delta_N)$, and (D.24), we have

$$\frac{1}{N}\sum_{i=1}^{N} |\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)| \, |\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - \widehat{r}(\boldsymbol{X}_i)|^2 = O_p(\epsilon_N^2)$$
$$\frac{1}{N}\sum_{i=1}^{N} |\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)|^2 \, |\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - \widehat{r}(\boldsymbol{X}_i)| = O_p(\epsilon_N \delta_N^2),$$

where the second result is obtained from the Cauchy-Schwarz inequality. Collecting the above results and plugging them into (D.26), we can bound $|D_{2,N}|$ by

$$\begin{aligned} |D_{2,N}| &\leqslant O_p(\epsilon_N^2 \delta_N^2) + O_p(\epsilon_N \delta_N^2) + O_p(\epsilon_N \delta_N^2) + O_p(\epsilon_N^2) \\ &= o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right), \end{aligned} \tag{D.27}$$

where the equality is due to $\epsilon_N = o(N^{-\frac{1}{2}})$ and $\delta_N = o_p(N^{-\frac{1}{4}})$.

To sum up, we have shown that

$$E_{3,N} = D_{1,N} + D_{2,N} = o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right),$$

which together with the result for $E_{1,N} + E_{2,N}$ in (D.20) yield

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} \{e(\delta_i, \boldsymbol{X}_i; \Pi_{\mathcal{F}_N} \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - r_0) - e(\delta_i, \boldsymbol{X}_i; \widehat{r} - r_0)\} \\ =&\epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X})\} m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right), \end{aligned}$$

which is the desired result. □

E Additional simulation results

In this part, we report additional results of the numerical simulations, including the inference for the mean of Y of the target population with the dimension of the covariate $d = 5$ in Table 1, and the inference for the mean and median Y of the target population with $d = 10$ in Table 2 and 3, respectively.

Table 1. Empirical estimation and inference results for $\theta = \mathbb{E}_Q(Y)$ of the target population with $d = 5$ based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	-0.0168	0.1322	0.0175	0.9048	0.4087
	MI	0.0203	0.1471	0.0217	0.8736	0.3716
	DRW-MI-E	-0.0135	0.1304	0.0171	0.9265	0.3824
	DRW-MI-T	-0.0125	0.1271	0.0163	0.9374	0.3791
	LDML	-0.0117	0.1426	0.0204	0.8592	0.3617
	CB	0.0370	0.1683	0.0297	0.7332	0.4204
$n = 2000$	DRW	-0.0149	0.1006	0.0103	0.9102	0.2817
	MI	-0.0182	0.1120	0.0129	0.8914	0.2546
	DRW-MI-E	-0.0118	0.0937	0.0089	0.9350	0.2972
	DRW-MI-T	-0.0121	0.0922	0.0086	0.9550	0.2935
	LDML	0.0130	0.1105	0.0124	0.9008	0.2780
	CB	0.0302	0.1319	0.0183	0.7298	0.3064
$n = 5000$	DRW	0.0105	0.0772	0.0061	0.9163	0.1708
	MI	-0.0127	0.0869	0.0078	0.9081	0.1665
	DRW-MI-E	0.0084	0.0673	0.0046	0.9437	0.1812
	DRW-MI-T	-0.0081	0.0660	0.0043	0.9481	0.1845
	LDML	-0.0119	0.0882	0.0078	0.9083	0.1713
	CB	0.0267	0.0941	0.0096	0.7510	0.1964

Table 2. Empirical estimation and inference results for $\theta = \mathbb{E}_Q(Y)$ of the target population with $d = 20$ based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	0.0815	0.3048	0.0995	0.7296	1.1592
	MI	-0.0902	0.3407	0.1242	0.7381	1.2157
	DRW-MI-E	0.0521	0.2485	0.0645	0.8168	0.9052
	DRW-MI-T	0.0347	0.2019	0.0419	0.8477	0.8895
	LDML	0.0609	0.3601	0.1334	0.7201	1.3162
	CB	-0.1308	0.2724	0.0864	0.5942	0.8125
$n = 2000$	DRW	0.0701	0.2382	0.0616	0.7640	0.8619
	MI	-0.0736	0.2619	0.0631	0.7774	0.9015
	DRW-MI-E	-0.0452	0.1829	0.0355	0.8851	0.7824
	DRW-MI-T	-0.0301	0.1681	0.0291	0.9174	0.7637
	LDML	0.0492	0.2128	0.0477	0.7831	0.8459
	CB	-0.0945	0.2209	0.0559	0.5781	0.7037
$n = 5000$	DRW	0.0539	0.1839	0.0367	0.8152	0.6729
	MI	0.0569	0.2007	0.0435	0.8347	0.7138
	DRW-MI-E	-0.0335	0.1362	0.0196	0.9214	0.6042
	DRW-MI-T	0.0304	0.1120	0.0135	0.9436	0.5814
	LDML	0.0369	0.1783	0.0331	0.8152	0.7221
	CB	-0.0901	0.1821	0.0395	0.6515	0.5981

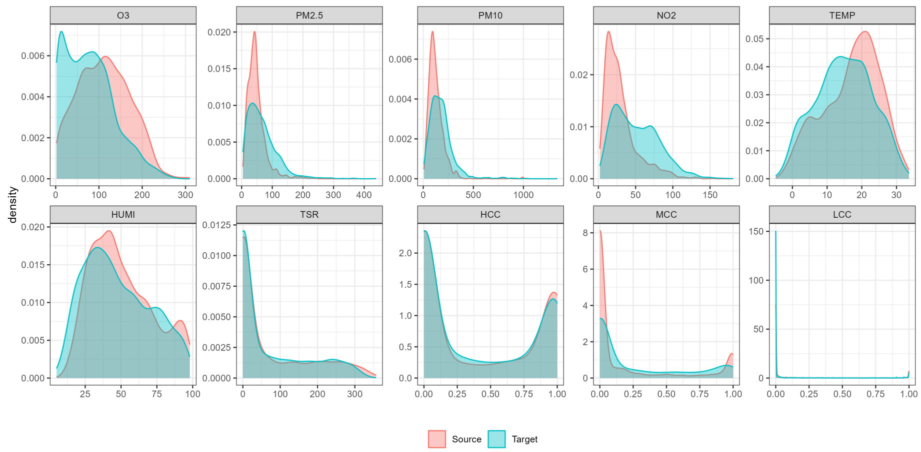
Table 3. Empirical estimation and inference results for $\theta = Q_Y^{-1}(1/2)$ of the target population with $d = 20$ based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	-0.0943	0.3420	0.1258	0.7169	1.2011
	MI	-0.0962	0.3541	0.1346	0.7215	1.2142
	DRW-MI-E	0.0731	0.2685	0.0774	0.8280	1.0204
	DRW-MI-T	0.0527	0.2301	0.0557	0.8505	0.9969
	LDML	-0.0693	0.3318	0.1148	0.7119	1.2650
	CB	-0.1436	0.2817	0.1001	0.5523	0.8856
$n = 2000$	DRW	0.0815	0.2740	0.0817	0.7593	0.8619
	MI	-0.0856	0.2802	0.0858	0.7324	0.8242
	DRW-MI-E	-0.0528	0.2129	0.0481	0.8613	0.7907
	DRW-MI-T	0.0493	0.1891	0.0381	0.9038	0.7741
	LDML	0.0566	0.2547	0.0681	0.7918	0.8109
	CB	-0.1231	0.2037	0.0566	0.5390	0.7074
$n = 5000$	DRW	0.0652	0.1971	0.0431	0.8098	0.6872
	MI	-0.0690	0.2085	0.0482	0.8209	0.7524
	DRW-MI-E	-0.0341	0.1381	0.0203	0.9209	0.6507
	DRW-MI-T	-0.0318	0.1152	0.0143	0.9367	0.5901
	LDML	0.0392	0.1801	0.0339	0.8247	0.7349
	CB	-0.1056	0.1618	0.0373	0.5607	0.5890

F Additional case study results

Figure 1 in the SM illustrates the distinctions between the distributions of some key variables of the target and the source samples, which reveals that directly using the source samples to make inferences about the O_3 of the target population would introduce biases.

Figure 1. Density plots for the O_3 and covariate variables of the source and the target samples.



References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*, volume 9. Cambridge University Press, 1999.

Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Jinyuan Chang, Song Xi Chen, and Xiaohong Chen. High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185 (1):283–304, 2015.

Charles Fefferman. Whitney's extension problem for. *Annals of mathematics*, pages 313–359, 2006.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical Estimation: Asymptotic Theory*. SpringerVerlag, New York, 1981.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Art Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18 (1):90 – 120, 1990.

David Pollard. *Empirical Processes: Theory and Applications*, volume 2. Institute of Mathematical Statistics, 1990.

R Tyrrell Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1997.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

Aad W van der Vaart and Jon A Wellner. Weak convergence and empirical processes. *Springer Series in Statistics*, 126:505, 1996.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.