Journal of Machine Learning Research 27 (2026) 1-50 Submitted 11/25; Revised 4/26; Published 7/26

Efficient Inference under Label Shift in Unsupervised Domain Adaptation

Seong-ho Lee SEONGHO@UOS.AC.KR
Department of Statistics
University of Seoul
Seoul, 02504, South Korea

Yanyuan Ma YZM63@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Jiwei Zhao* JIWEI.ZHAO@WISC.EDU
Departments of Statistics and of Biostatistics & Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706, USA

Editor: Kun Zhang

Abstract

In many real-world applications, researchers aim to deploy models trained in a source domain to a target domain, where obtaining labeled data is often expensive, time-consuming, or even infeasible. While most existing literature assumes that the source and target data follow the same joint distribution, distribution shifts are common in practice. This paper considers a particular type of distribution shift, label shift, and develops an efficient inference procedure for general parameters characterizing the unlabeled target population. A central idea is to model the outcome density ratio between the labeled source data and unlabeled target data. To this end, we propose a progressive estimation strategy that unfolds in three stages: an initial heuristic guess, a consistent estimation, and ultimately, an efficient estimation. This self-evolving process is novel in the statistical literature and of independent interest. We also highlight the connection between our approach and prediction-powered inference (PPI), which uses machine learning models to improve statistical inference in related settings. We rigorously establish the asymptotic properties of the proposed estimators and demonstrate their superior performance compared to existing methods. Through simulation studies and multiple real-world applications, we illustrate both the theoretical contributions and practical benefits of our approach.

Keywords: distribution shift, efficient estimation, efficient influence function, label shift, semiparametric statistics

1. Introduction

In many real-world applications, researchers are interested in deploying models and algorithms trained on a source domain to a target domain where obtaining labeled data is often expensive, time-consuming, or even infeasible. For example, in healthcare, disease pre-

*. Corresponding author

diction models trained on data from one hospital may be applied to another population of patients at risk but not yet diagnosed. In social science, models are often evaluated for their generalizability across survey populations that might differ demographically or culturally, for which the outcome of interest is not fully collected and time-consuming to collect. In autonomous driving, models trained on simulated data or limited environment must adapt to new, unseen locations. In the literature, this type of research problems was termed unsupervised domain adaptation (UDA, also known as transductive transfer learning); see an excellent review paper by Kouw and Loog (2019).

In such scenarios where acquiring gold-standard labels in the target domain is costly and labor-intensive, machine learning (ML)-based algorithms and models offer an efficient solution, substantially reducing the cost of outcome collection and accelerating scientific research across disciplines (Wang et al., 2023). For instance, in electronic health records (EHR), true disease statuses are accessible through expert-led manual chart review, and this manual review requires substantial costs due to the large scale nature of EHR. As a cost-effective solution, ML models can provide approximation of the disease statuses with accuracy comparable to manual review, thereby streamlining downstream clinical analyses (Yang et al., 2022; Zhou et al., 2022). Similarly, although population biobanks provide extensive data for high-throughput genetic research, many phenotypes and disease outcomes remain unmeasured due to prohibitive costs. Resolving this practical issue, ML approaches have been proven effective in facilitating novel genetic discoveries and advancing scientific inquiry (An et al., 2023; McCaw et al., 2024; Miao et al., 2024).

However, the inherent complexity and the use of ML predictions pose significant challenges in statistical inferences. Using such predictions without recognizing their differences from observed gold-standard data can result in biased outcomes and misleading scientific conclusions. To address this, researchers have introduced methods that fuse extensive ML predictions with limited gold-standard data to ensure the validity of ML-assisted statistical inference. Over the past five years, this topic has gained significant importance and popularity in the literature. For easier exposition, consider the availability of both labeled data

$$(Y_i, \mathbf{X}_i), i = 1, \ldots, n, \text{ drawn as a random sample from source population } \mathcal{P}, \tag{1}$$

and unlabeled data

$$\mathbf{X}_i, i = n+1, \ldots, N, \text{ drawn as a random sample from target population } \mathcal{Q}, \tag{2}$$

where the assumption $\mathcal{P} = \mathcal{Q}$ is tentatively adopted for simplicity. Wang et al. (2020) proposed a *post-prediction inference* procedure, and Motwani and Witten (2023) conducted a more rigorous investigation on this *inference after prediction* procedure, and showed that, the method proposed in Wang et al. (2020) could only provide a valid inference in some restrictive scenarios in the context of least squares estimation. More recently, Angelopoulos et al. (2023a) proposed the approach termed *prediction-powered inference* (PPI), which yields valid inference even when the ML predictions were of a low quality. To illustrate,

PPI estimates the outcome mean $\theta = \mathrm{E}(Y)$ as

$$\widehat{\theta} \;=\; \underbrace{\frac{1}{N-n}\sum_{i=n+1}^{N}\widehat{y}_i}_{\substack{\text{Direct estimation using}\\ \text{predictions on unlabeled data}}} \;+\; \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)}_{\substack{\text{Rectifier: measure the bias of using}\\ \text{predictions on labeled data}}}, \tag{3}$$

where $\widehat{y}_i$'s, representing the conditional mean $\mathrm{E}(Y \mid \mathbf{x}_i)$, are the predicted outcomes produced by the ML algorithm/model. A few extensions of PPI were also proposed in the past two years, including but not limited to, Zrnic and Candès (2024) which considered the availability of the ML predictions, and Angelopoulos et al. (2023b); Miao et al. (2023); Gronsbell et al. (2024) which investigated the statistical efficiency gains from different perspectives.

It is worthwhile to note, in the statistical literature, terms as *semi-supervised inference*, researchers have also proposed different methods of utilizing the unlabeled data for different purposes (Chakrabortty and Cai, 2018; Yuval and Rosset, 2022; Azriel et al., 2022; Livne et al., 2022; Hou et al., 2023; Song et al., 2024), including settings under high dimensionality (Zhang et al., 2019; Cai and Guo, 2020; Zhang and Bradic, 2022; Deng et al., 2024; Ying et al., 2026).

Despite significant progress in conducting inference under such a scenario, several key aspects remain unclear in the literature. This paper is motivated by the challenge of distribution shift (Quiñonero-Candela et al., 2009) between source domain labeled data (1) and target domain unlabeled data (2), in the sense that the unlabeled data were collected under conditions different from those of the labeled data. The scenario $\mathcal{P} \neq \mathcal{Q}$ is more realistic and more applicable in many practical settings. Indeed, various forms of distribution shift have been studied extensively in the literature, including covariate shift (the conditional distribution of Y given $\mathbf{X}$ remains the same while the marginal distribution of $\mathbf{X}$ shifts), and label shift (the conditional distribution of $\mathbf{X}$ given Y remains the same while the marginal distribution of Y shifts). In particular, label shift has been further extended to more complex scenarios, such as open-set label shift (Garg et al., 2022), latent label shift (Roberts et al., 2022), and relaxed label shift (Garg et al., 2023).

The goal of this paper is to develop an efficient procedure for estimation and inference under *label shift*, where *efficiency* is defined in the context of statistical estimation (Bickel et al., 1993; Tsiatis, 2006). To relate this statistical efficiency to the machine learning focus on prediction accuracy, we note that both share a common pursuit of precision improvement through bias and variance reduction: while prediction accuracy usually seeks to minimize the mean squared error, which includes the bias and variance of individual predictions, statistical efficiency in our work aims to minimizing the mean squared error of estimators. Because our procedure ensures that the bias is dominated by the variance, hence the problem further reduces to minimize exclusively the variance of the parameter estimator. In this shared context, our efficiency-driven approach provides an estimation-level counterpart to the prediction-level precision typically sought in ML tasks. It is worth noting that in the original PPI paper, Angelopoulos et al. (2023a) addressed scenarios involving distribution shift. However, under label shift, their method is limited to cases with discrete Y, similar to the approaches in Lipton et al. (2018) and Tian et al. (2023). Moreover, estimation efficiency was not considered in their work. To the best of our knowledge, efficient estimation with

label shift under this setting has not been explored in the literature, despite the availability of various methods proposed in both statistical and machine learning contexts (Storkey, 2009; Zhang et al., 2013; Du Plessis and Sugiyama, 2014; Iyer et al., 2014; Nguyen et al., 2016; Tasche, 2017; Kim et al., 2024; Lee et al., 2025). We also emphasize that the setting considered in this paper falls under the broad umbrella of UDA, where the outcome Y *is not* observed in the unlabeled dataset $\mathcal{Q}$. This is fundamentally different from settings where Y *is* observed in $\mathcal{Q}$, as in Li and Luedtke (2023) and Qiu et al. (2024), for reasons discussed below.

Under label shift, the most challenging step is identifying the distribution discrepancy, characterized by the density ratio $\rho(y)$ defined in (8) in Section 2. In the setting of UDA, this step is particularly challenging due to the absence of outcome data Y from the population $\mathcal{Q}$. This absence is the key distinction from other settings pointed out in the previous paragraph. Under this setting, Lee et al. (2025) proposed estimation procedures for the general parameter $\boldsymbol{\theta}$ in the unlabeled data population $\mathcal{Q}$ defined in (6), strategically avoiding the challenge of identifying $\rho(y)$. These procedures guarantee consistency but not efficiency, as the efficiency is achieved only if a heuristic guess on $\rho(y)$ is correct, which is almost impossible due to the absence of Y in $\mathcal{Q}$. In contrast, our work offers a way to guarantee efficiency in estimating $\boldsymbol{\theta}$ through a novel way of estimating $\rho(y)$. A critical insight enabling the estimation of $\rho(y)$ is that, although $\rho(y)$ is a *local* feature of $\mathcal{Q}$, its estimation can be reframed as a problem of estimating a *global* feature of $\mathcal{Q}$. Leveraging this insight, we propose estimation procedures for $\rho(y)$ which, in turn, enable efficient estimation of $\boldsymbol{\theta}$ in $\mathcal{Q}$. For estimating $\rho(y)$ itself, we also introduce a progressive estimation process, evolving through three stages: an initial heuristic guess, a consistent estimation, and ultimately an efficient estimation. Different from the traditional one-step estimation (Bickel, 1975) commonly used in semiparametric statistics (Bickel et al., 1993) and targeted learning (van der Laan et al., 2011), our estimation is multi-step, where in each step, we adopt the same procedure to obtain an improved estimate based on the current estimate. This self-driven evolutionary process is not common in the statistical literature and might be of independent interest. The detailed methodology underlying our proposal is presented in Section 3.

The structure of the paper is as follows. In Section 2, we first review the label shift literature and some key results such as the efficient influence function. Section 3 presents the methodology and Section 4 presents the corresponding theory, respectively. In Section 5, we specifically discuss the discrete label case which is of primary interest in the ML community. From Sections 3 to 5, the data available to analysts are both labeled and unlabeled data, and the ML prediction model available to analysts is $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, where E_p denotes the expectation with respect to $\mathcal{P}$. Similar to the setting in Angelopoulos et al. (2023a), this ML prediction model can be fitted from the labeled data $\mathcal{P}$ or from some external data that has the same distribution as the labeled data. Section 6 contains the simulation results and Section 7 contains two data applications, respectively. The paper is concluded with some discussions in Section 8. All the technical details are deferred in the Supplementary Material.

2. Review: Label Shift and Efficient Influence Function (EIF)

We start by briefly reviewing the label shift assumption and some existing results under label shift. Throughout the paper, we observe both source domain labeled data (1) and

target domain unlabeled data

$$\mathbf{X}_i, i = n+1, \ldots, N, \text{ drawn as a random sample from population } \mathcal{Q}, \text{ where } \mathcal{P} \neq \mathcal{Q}. \quad (4)$$

We denote p_Y and q_Y the density of Y in $\mathcal{P}$ and $\mathcal{Q}$, $p_{\mathbf{X}|Y}$ and $q_{\mathbf{X}|Y}$ the conditional density of $\mathbf{X}$ given Y in $\mathcal{P}$ and $\mathcal{Q}$, respectively.

We emphasize that we allow the distributions in $\mathcal{P}$ and $\mathcal{Q}$ to be different. In the literature, various discussions have explored different formats of distribution shifts (Quiñonero-Candela et al., 2009), such as covariate shift, label shift, etc. In this paper, we consider the label shift assumption,

$$p_Y(y) \neq q_Y(y), \quad \text{and} \quad p_{\mathbf{X}|Y}(\mathbf{x}, y) = q_{\mathbf{X}|Y}(\mathbf{x}, y) \equiv f_{\mathbf{X}|Y}(\mathbf{x}, y) \text{ for some } f_{\mathbf{X}|Y}. \quad (5)$$

This means that in both population $\mathcal{P}$ and population $\mathcal{Q}$, as long as the label Y is given, the covariate $\mathbf{X}$ follows the same distribution. We use the notation $f_{\mathbf{X}|Y}$ to emphasize that the $\mathbf{X} \mid Y$ distribution is identical in the two distinct populations. The distributions of $\mathcal{P}$ and $\mathcal{Q}$ differ because the marginal distributions of Y are different. Label shift has been shown to be reasonable and practical in the anticausal learning setting (Schölkopf et al., 2012), where Y causes $\mathbf{X}$—examples include diseases causing symptoms or objects causing sensory observations (Storkey, 2009; Zhang et al., 2013; Du Plessis and Sugiyama, 2014; Iyer et al., 2014; Nguyen et al., 2016; Tasche, 2017; Kim et al., 2024; Lee et al., 2025). It is also suitable in computer vision applications, such as predicting object locations, directions, and human poses (Martinez et al., 2017; Yang et al., 2018; Guo et al., 2020).

In this paper, we consider the parameter of interest $\boldsymbol{\theta}$ that satisfies the global characteristic

$$\mathrm{E}_q\{\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})\} = \mathbf{0}, \quad (6)$$

where $\mathbf{U}(\cdot)$ is a pre-specified known function, and E_q denotes the expectation with respect to $\mathcal{Q}$. The overarching goal of this paper is to investigate the efficient estimation and inference of $\boldsymbol{\theta}$, under the assumption in (5).

The definition of $\boldsymbol{\theta}$ in (6) is quite general. It includes the minimizer of a given loss function $\ell(\cdot)$ in the target population $\mathcal{Q}$, such as $\mathrm{argmin}_{\boldsymbol{\theta}} \mathrm{E}_q\{\ell(Y, \mathbf{X}, \boldsymbol{\theta})\}$. For illustration purposes, we sometimes also present the results for a simpler scenario defined as

$$\theta = \mathrm{E}_q\{s(Y, \mathbf{X})\}, \quad (7)$$

with $s(\cdot)$ a pre-specified known function. This corresponds to $U(Y, \mathbf{X}, \theta) = s(Y, \mathbf{X}) - \theta$.

2.1 The Robust Estimation Procedure

In both ML and statistics, there has been significant progress in developing methods for prediction and estimation under label shift. Here, we briefly review the robust estimation procedure from Lee et al. (2025), as it forms the basis of the new method proposed in this paper.

When pooling the data from both populations together, to facilitate the presentation, the population indicator R is introduced in that $R = 1$ if the sample is from $\mathcal{P}$ and $R = 0$ if from $\mathcal{Q}$, with r_i being its sample realization. Further, the proportion of the sample from $\mathcal{P}$

is defined as $\pi = n/N$, and the density ratio of Y, given that the support of q_Y is contained in that of p_Y, is defined as

$$\rho(y) = q_Y(y)/p_Y(y), \tag{8}$$

which is not straightforward to estimate because of the absence of Y observations from $\mathcal{Q}$. The key idea of Lee et al. (2025) is to construct a locally efficient influence function $\boldsymbol{\phi}^*_{\text{eff}}$ by replacing $\rho(\cdot)$ in the efficient influence function (EIF) $\boldsymbol{\phi}_{\text{eff}}$ with an arbitrary working model $\rho^*(\cdot)$. The EIF is a central tool in semiparametric inference that characterizes the efficiency bound, representing the optimal data utilization to reach the theoretical precision limit. In this framework, unknown components such as the density ratio $\rho(\cdot)$ are treated as nuisance functions—auxiliary parts of the model that must be accounted for to estimate the target parameter $\boldsymbol{\theta}$. A primary advantage of the EIF-based approach is its first-order insensitivity, or Neyman orthogonality, to estimation errors in these nuisance functions. This property ensures that the final inference may remain valid even if the ML models used to estimate the nuisance functions are not perfectly specified. This approach maintains the validity of the estimating function by ensuring it has mean zero at the true parameter, a property that holds even without requiring a correct identification of $\rho(\cdot)$. Depending on how other nuisance components are handled, the estimators proposed in Lee et al. (2025) enjoys some nice robustness properties such as single flexibility and double flexibility.

For the simpler estimand in (7), the proposed singly flexible estimator in Lee et al. (2025) is

$$\begin{aligned}\widetilde{\theta} &= \frac{1}{N-n}\sum_{i=n+1}^{N}\widetilde{w}_i\widehat{\mathrm{E}}_p\{\widehat{a}^*(Y)\rho^*(Y)\mid \mathbf{x}_i\} \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}\rho^*(y_i)\left[s(y_i,\mathbf{x}_i)-\widetilde{w}_i\widehat{\mathrm{E}}_p\{\widehat{a}^*(Y)\rho^*(Y)\mid \mathbf{x}_i\}\right],\end{aligned} \tag{9}$$

where $\widetilde{w}_i = [\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{x}_i\}]^{-1}$'s are the estimates of the sample weights, $\widehat{a}^*(y)$ can be obtained by solving the equation

$$s(y,\mathbf{x}_i) = \sum_{i=1}^{N}\widetilde{w}_i\widehat{\mathrm{E}}_p\{a(Y)\rho^*(Y)\mid \mathbf{x}_i\}\frac{r_i\widetilde{K}_l(y-y_i)}{\sum_{j=1}^{N}r_j\widetilde{K}_l(y-y_j)}, \tag{10}$$

where $\sum_{i=1}^{N}\ell(\mathbf{x}_i)\frac{r_i\widetilde{K}_l(y-y_i)}{\sum_{j=1}^{N}r_j\widetilde{K}_l(y-y_j)}$ is the approximation of the conditional expectation $\mathrm{E}\{\ell(\mathbf{X})\mid y\}$, $\widetilde{K}(\cdot)$ is the kernel function and l is the bandwidth. This equation is known as a Fredholm equation of the first kind, for which a detailed discussion and methods to solve it are in Lee et al. (2025). Note that, the EIF for estimating θ is

$$\phi_{\text{eff}} = \frac{r}{\pi}\rho(y)\left[s(y,\mathbf{x})-w(\mathbf{x})\mathrm{E}_p\{a(Y)\rho(Y)\mid \mathbf{x}\}\right]+\frac{1-r}{1-\pi}\left[w(\mathbf{x})\mathrm{E}_p\{a(Y)\rho(Y)\mid \mathbf{x}\}-\theta\right],$$

where $w(\mathbf{x}) = \left[\mathrm{E}_p\{\rho^2(Y)+\pi/(1-\pi)\rho(Y)\mid \mathbf{x}\}\right]^{-1}$, and $a(y)$ satisfies

$$\mathrm{E}\left[w(\mathbf{X})\mathrm{E}_p\{a(Y)\rho(Y)\mid \mathbf{X}\}\mid y\right] = s(y,\mathbf{x}).$$

Algorithm 1 Algorithm $[\theta = \mathrm{E}_q\{s(Y, \mathbf{X})\},\ \rho^*(y)]$

Data Input: data from $\mathcal{P}$: $(r_i = 1, y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, data from $\mathcal{Q}$: $(r_j = 0, \mathbf{x}_j)$, $j = n+1, \ldots, N$, and $\pi = n/N$.
Algorithm/Model Input: the AI prediction model $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, and the density ratio model $\rho^*(y)$.
Output: $\widetilde{\theta}$, as defined in (9).
Do:
(a) compute $\widetilde{w}_i = [\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}]^{-1}$ for $i = 1, \ldots, N$;
(b) obtain $\widehat{a}^*(y)$ by solving (10);
(c) construct $\widetilde{\theta}$ as in (9).

For convenience of later use, we summarize this robust estimation procedure in Algorithm 1, and call it

$$\textsc{Algorithm}\ [\theta, \rho^*(y)],$$

where the first argument θ indicates the estimand and the second argument $\rho^*(y)$ indicates how the density ratio $\rho(y)$ is implemented in the algorithm. Meanwhile, we also summarize the key notations used throughout the paper in Table 1 for a better reference.

Table 1: Summary of key notations.

Notation	Description
$\mathcal{P}, \mathcal{Q}$	Source (labeled) and Target (unlabeled) populations
R	Population indicator ($R = 1$ for $\mathcal{P}$, $R = 0$ for $\mathcal{Q}$)
$Y, \mathbf{X}$	Outcome and Covariates
π	Proportion of samples from $\mathcal{P}$
$\rho(y)$	Outcome density ratio $q_Y(y)/p_Y(y)$
E_p, E_q	Expectations with respect to $\mathcal{P}$, $\mathcal{Q}$
δ_0	Kernel approximation of $q_Y(y_0)$
$\boldsymbol{\theta}$	Target parameter in $\mathcal{Q}$
$\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})$	Estimating function for $\boldsymbol{\theta}$
ϕ_{eff}	Efficient Influence Function (EIF)

2.2 The Limitation and Our Novel Contributions

An obvious limitation of Lee et al. (2025) is that the proposed estimators are not efficient. The performance of their estimators heavily depends on how lucky they are in choosing $\rho^*(\cdot)$: if the working model $\rho^*(\cdot)$ is unfortunately very badly chosen, their estimators may have substantial standard error and poor inference performance.

In contrast, we propose an efficient estimator with the smallest variability among all regular asymptotically linear estimators (Bickel et al., 1993; Tsiatis, 2006). A key step in achieving efficiency is estimating $\rho(\cdot)$, a step omitted in Lee et al. (2025). Notably, although $\rho(y)$ is a function of y hence describes *local* features of $\mathcal{Q}$, estimating $\rho(y)$ can almost be reframed as a problem of estimating a *global* feature in $\mathcal{Q}$, using a specially

defined $\mathbf{U}(\cdot)$ function in (6). While technical challenges must be addressed, this insight is central to developing our efficient estimator for the general parameter $\boldsymbol{\theta}$. The details will be presented in Section 3.

3. Proposed Methodology

Note that $\rho(y)$ is defined as the density ratio $\rho(y) = q_Y(y)/p_Y(y)$. Between $p_Y(y)$ and $q_Y(y)$, estimating $p_Y(y)$ is relatively straightforward since Y-data are available in the population $\mathcal{P}$. However, estimating $q_Y(y)$ is challenging due to the absence of observations. In Section 3.1, we first introduce an initial consistent estimator for $\rho(y)$, which is already sufficient for the efficient estimation of $\boldsymbol{\theta}$ proposed in Section 3.3. Additionally, an efficient estimation of $\rho(y)$ is also feasible and is presented in Section 3.2. The progression in estimating $\rho(y)$—from the heuristic guess in Lee et al. (2025), to the consistent estimation in Section 3.1, and finally to the efficient estimation in Section 3.2, is notably achieved by repeating the same procedure, hence can be viewed as a self-driven evolutionary process.

3.1 An Initial Step for Consistently Estimating $\rho(y)$

For estimating $q_Y(y)$, without direct observations, we must infer information about $q_Y(y)$ indirectly through the available $\mathbf{X}$ observations from $\mathcal{Q}$ and the link between $\mathcal{P}$ and $\mathcal{Q}$. Therefore, the core difficulty in estimating $\rho(y)$ lies in constructing a consistent estimator of $q_Y(y)$ that can overcome the absence of the outcome in $\mathcal{Q}$ by leveraging the information in the covariates from $\mathcal{Q}$, the data from $\mathcal{P}$, and the distributional link between the two populations.

Consider the estimation of $q_Y(y)$ at an arbitrary point $y = y_0$ in the support of Y. The quantity $q_Y(y_0)$ is essentially a local feature of $\mathcal{Q}$, whereas the method reviewed in Algorithm $[\theta, \rho^*(y)]$ applies only to a quantity θ that represents a global feature of $\mathcal{Q}$. To bridge this gap, we approximate $q_Y(y_0)$ by a global feature and thereby facilitate its estimation. Specifically, we can approximate $q_Y(y_0)$ by

$$\delta_0 \equiv \mathrm{E}_q\{K_h(Y - y_0)\},$$

where $K_h(y) = K(y/h)/h$, with $K(\cdot)$ being a kernel function and h a bandwidth. Now, we can view δ_0 as an unknown parameter in $\mathcal{Q}$ defined in (7) corresponding to $s(Y, \mathbf{X}) = K_h(Y - y_0)$. This formulation allows for utilization of the EIF for estimating δ_0, with its detailed derivation presented in Supplementary Material A.1. Thus, one can apply the Algorithm $[\delta_0, \rho^*(y)]$ to obtain the estimator

$$\begin{aligned}\widetilde{\delta}_0 \;=\; & \frac{1}{N-n}\sum_{i=n+1}^{N} \widetilde{w}_i \widehat{\mathrm{E}}_p\{\hat{a}^*(Y)\rho^*(Y) \mid \mathbf{x}_i\} \\ & + \frac{1}{n}\sum_{i=1}^{n} \rho^*(y_i)\Big[K_h(y_i - y_0) - \widetilde{w}_i \widehat{\mathrm{E}}_p\{\hat{a}^*(Y)\rho^*(Y) \mid \mathbf{x}_i\}\Big],\end{aligned}$$

where $\widetilde{w}_i$ is the same as that in (9), $\widehat{a}^*(y)$ is also the same as the solver of (10) except for replacing $s(y, \mathbf{x})$ by $K_h(y - y_0)$. Consequently, the estimator of $\rho(y_0)$ is obtained as

$$\widetilde{\rho}(y_0) = \frac{\widetilde{\delta}_0}{\widehat{p}_Y(y_0)} \quad \text{where} \quad \widehat{p}_Y(y_0) = n^{-1} \sum_{i=1}^{n} K_h(y_i - y_0). \tag{11}$$

Note that we can estimate $\rho(y)$ at any $y = y_0$ value above, hence we can estimate the entire function $\rho(\cdot)$, i.e., we can obtain $\widetilde{\rho}(\cdot)$. In practice, we may preselect a set of points $\{t_1, \ldots, t_{M_n}\}$ on the support and implement the estimator $\widetilde{\rho}(y_0)$ at each $y_0 = t_k$ for $k = 1, \ldots, M_n$, then construct a smoothing curve that interpolates $\{\widetilde{\rho}(t_1), \ldots, \widetilde{\rho}(t_{M_n})\}$ to establish $\widetilde{\rho}(\cdot)$. Here M_n can be set to $[n^{1/4}]$ to ensure sufficient precision.

Remark 1 *To ensure consistency in presentation throughout the paper, even for the purpose of estimating $\rho(y)$, we assume here that an ML prediction model $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ is available. However, it is important to highlight that a correctly specified model of $\mathrm{E}_p(\cdot \mid \mathbf{x})$ is NOT required for the purpose of estimating $\rho(y)$ in the above procedure. This is a key attribute of the locally efficient influence function, as discovered in Lee et al. (2025) and referred to as "double flexibility". This characteristic distinguishes our proposed method here from others in the literature, such as Kim et al. (2024) which relies on a correctly specified model of $\mathrm{E}_p(\cdot \mid \mathbf{x})$ for estimating $\rho(y)$.*

3.2 Efficiently Estimating $\rho(y)$

Equipped with $\widetilde{\rho}(y)$, one can further refine the estimation for $\rho(y)$ by first applying the Algorithm $[\delta_0, \widetilde{\rho}(y)]$ to obtain $\widehat{\delta}_0$, where the input model for the density ratio is replaced with the estimated $\widetilde{\rho}(y)$ in Section 3.1 in lieu of $\rho^*(y)$, then formulating the refined estimator as

$$\widehat{\rho}(y_0) = \frac{\widehat{\delta}_0}{\widehat{p}_Y(y_0)} \quad \text{where} \quad \widehat{p}_Y(y_0) = n^{-1} \sum_{i=1}^{n} K_h(y_i - y_0), \tag{12}$$

and

$$\begin{aligned} \widehat{\delta}_0 \;=\; & \frac{1}{N-n} \sum_{i=n+1}^{N} \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\} \\ & + \frac{1}{n} \sum_{i=1}^{n} \widetilde{\rho}(y_i) \left[K_h(y_i - y_0) - \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\} \right], \end{aligned}$$

where $\widehat{w}_i = [\widehat{\mathrm{E}}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}]^{-1}$, and $\widehat{a}(y)$ represents the solution by solving the equation (10) in the Algorithm $[\delta_0, \widetilde{\rho}(y)]$.

This refinement procedure further reduces the variability of the $\rho(y)$ estimator. In fact, as we will demonstrate in Section 4, the resulting estimator $\widehat{\rho}(y)$ achieves the minimum variance at a given kernel function and bandwidth. Thus, we complete a self-improvement process by iteratively applying our method, starting from a likely inconsistent initial guess $\rho^*(y)$, progressing to a consistent estimate $\widetilde{\rho}(y)$, and ultimately achieving an efficient estimator $\widehat{\rho}(y)$.

Remark 2 *The self-driven evolutionary process of iteratively applying the same algorithm to achieve the efficient estimation of $\rho(y)$ given fixed kernel function and bandwidth, as described in Sections 3.1 and 3.2, is fundamentally different from the classical one-step estimation methods in semiparametric statistics and targeted learning (Bickel, 1975; Bickel et al., 1993; van der Laan et al., 2011). Though we estimate $\rho(y)$ via estimating $p_Y(y_0)$ and $q_Y(y_0)$ at each support point y_0, our goal here is to estimate the entire function $\rho(y)$ instead of a parameter with fixed dimensionality. More importantly, a key innovation in our approach is the approximation of $q_Y(y_0)$ by δ_0, which can then be estimated via implementing the locally efficient influence function. This evolutionary strategy is essential because, unlike standard semiparametric settings where a consistent initial estimator is readily available, obtaining a consistent density ratio $\rho^*(y)$ is remarkably difficult in our setting due to the lack of labels in the target population. While classical one-step estimators rely on the initial consistency to reach the efficiency bound, such an approach would fail if the initial guess is inconsistent. Our three-stage strategy addresses this by utilizing Stage 2 as a crucial stepping stone to transform a possibly inconsistent heuristic guess $\rho^*(y)$ into a consistent estimator $\widetilde{\rho}(y)$, which then facilitates the attainment of the efficiency bound in Stage 3. Subsequently, this progressive evolution ensures both the validity and efficiency of $\widehat{\boldsymbol{\theta}}$ even when starting from an unreliable initial model.*

Remark 3 *It is worth noting that our approach to estimating $\rho(y)$ differs from existing methods, such as Lipton et al. (2018) and Tian et al. (2023), in two key aspects. First, these existing methods were designed for discrete Y and hence require discretization to handle continuous Y, which inevitably leads to information loss. In contrast, our method directly handles continuous Y. Second, the existing methods predominantly treat density ratio estimation as their final goal. In contrast, the estimator of $\rho(y)$ constructed here is built as a stepping stone. By incorporating this density ratio estimate into the efficient influence function framework as detailed in the following section, we complete the semiparametric efficient inference procedure for the general parameter $\boldsymbol{\theta}$ in $\mathcal{Q}$. Notably, $\boldsymbol{\theta}$ also includes $\rho(y)$ as a special case.*

3.3 Efficiently Estimating θ

We now return to the primary objective of our work: to devise an efficient estimator of the general parameter $\boldsymbol{\theta}$ defined in (6). Given the consistent estimate of $\rho(y)$, $\widetilde{\rho}(y)$, presented in Section 3.1, the procedure is overall similar to the construction of $\widehat{\delta}_0$ in Section 3.2. For estimating $\boldsymbol{\theta}$, the EIF (Lee et al., 2025) is

$$\begin{aligned}\boldsymbol{\phi}_{\text{eff}}(r, ry, \mathbf{x}, \boldsymbol{\theta}) \;=\; & \mathbf{A}(\boldsymbol{\theta}) \Big(\frac{r}{\pi} \rho(y) \big[\mathbf{U}(y, \mathbf{x}, \boldsymbol{\theta}) - w(\mathbf{x}) \mathrm{E}_p\{ \mathbf{U}(Y, \mathbf{x}, \boldsymbol{\theta}) \rho^2(Y) + \mathbf{a}(Y) \rho(Y) \mid \mathbf{x} \} \big] \\ & + \frac{1-r}{1-\pi} w(\mathbf{x}) \mathrm{E}_p\{ \mathbf{U}(Y, \mathbf{x}, \boldsymbol{\theta}) \rho^2(Y) + \mathbf{a}(Y) \rho(Y) \mid \mathbf{x} \} \Big), \end{aligned} \tag{13}$$

where $\mathbf{A}(\boldsymbol{\theta}) \equiv [\mathrm{E}_q\{ \partial \mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^{\mathrm{T}} \}]^{-1}$, $w(\mathbf{x}) = \big[\mathrm{E}_p\{ \rho^2(Y) + \pi/(1-\pi) \rho(Y) \mid \mathbf{x} \} \big]^{-1}$, and $\mathbf{a}(y)$ needs to satisfy

$$\mathrm{E}\big[w(\mathbf{X}) \mathrm{E}_p\{ \mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta}) \rho^2(Y) + \mathbf{a}(Y) \rho(Y) \mid \mathbf{X} \} \mid y \big] = \mathrm{E}\{ \mathbf{U}(y, \mathbf{X}, \boldsymbol{\theta}) \mid y \}.$$

Algorithm 2 Algorithm for efficiently estimating $\boldsymbol{\theta}$ defined in (6)

Data Input: data from $\mathcal{P}$: $(r_i = 1, y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, data from $\mathcal{Q}$: $(r_j = 0, \mathbf{x}_j)$, $j = n+1, \ldots, N$, and $\pi = n/N$.
Algorithm/Model Input: the AI prediction model $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, and the density ratio model $\rho^*(y)$.
Output: $\widehat{\boldsymbol{\theta}}$, via solving (15).
Do:
(a) follow Algorithm $[\delta_0, \rho^*(y)]$ to compute $\widetilde{\delta}_0$ and then compute $\widetilde{\rho}(y)$ in (11);
(b) follow Algorithm $[\boldsymbol{\theta}, \widetilde{\rho}(y)]$ to compute $\widehat{\boldsymbol{\theta}}$, via solving (15).

To be more specific, given the consistent estimate $\widetilde{\rho}(y)$ and the ML prediction model $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, we first compute the estimate for the weight function $\widehat{w}_i = [\widehat{\mathrm{E}}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}]^{-1}$ at each observation $i = 1, \ldots, N$. Then, we obtain the function $\widehat{\mathbf{a}}(y)$ by solving the integral equation

$$\begin{aligned} & \sum_{i=1}^{N} \widehat{w}_i \widehat{\mathrm{E}}_p\{\mathbf{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\} r_i \widetilde{K}_l(y - y_i) \\ = & \sum_{i=1}^{N} \left[\mathbf{U}(y, \mathbf{x}_i, \boldsymbol{\theta}) - \widehat{w}_i \widehat{\mathrm{E}}_p\{\mathbf{U}(Y, \mathbf{x}_i, \boldsymbol{\theta})\widetilde{\rho}^2(Y) \mid \mathbf{x}_i\}\right] r_i \widetilde{K}_l(y - y_i) \end{aligned}$$

with respect to $\mathbf{a}(y)$, where $\widetilde{K}_l(y - y_i) = \widetilde{K}\{(y - y_i)/l\}/l$, $\widetilde{K}(\cdot)$ is a kernel function, and l is a bandwidth, chosen to approximate $\mathrm{E}(\cdot \mid y)$ by smoothing over the neighborhood of y.

Afterwards, we calculate

$$\widehat{\mathbf{b}}(\mathbf{x}_i) = \widehat{w}_i \widehat{\mathrm{E}}_p\{\mathbf{U}(Y, \mathbf{x}_i, \boldsymbol{\theta})\widetilde{\rho}^2(Y) + \widehat{\mathbf{a}}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}, \tag{14}$$

for each sample $i = 1, \ldots, N$. Finally, we obtain $\widehat{\boldsymbol{\theta}}$ by solving the following estimating equation

$$\underbrace{\frac{1}{N-n}\sum_{i=n+1}^{N} \widehat{\mathbf{b}}(\mathbf{x}_i)}_{\substack{\text{Direct estimating equation using} \\ \text{predictions on unlabeled data}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{\rho}(y_i)\{\mathbf{U}(y_i, \mathbf{x}_i, \boldsymbol{\theta}) - \widehat{\mathbf{b}}(\mathbf{x}_i)\}\right]}_{\substack{\text{Rectifier: measure the bias of using} \\ \text{predictions on labeled data}}} = \mathbf{0}, \tag{15}$$

with respect to $\boldsymbol{\theta}$. The summary of the implementation can be found in the Algorithm 2. It can be seen that for solving $\widehat{\boldsymbol{\theta}}$, the term $\widehat{\mathbf{b}}(\mathbf{x}_i)$ serves as the predicted value of the estimating equation $\mathbf{U}$, say, $\widehat{\mathbf{U}}(y_i, \mathbf{x}_i, \boldsymbol{\theta})$.

To conclude this section, we elaborate a little more on the efficient estimator for the simplest case $\theta = \mathrm{E}_q(Y)$, where the efficient estimator can be simplified as

$$\widehat{\theta} = \underbrace{\frac{1}{N-n}\sum_{i=n+1}^{N} \widehat{y}_i}_{\substack{\text{Direct estimation} \\ \text{using predictions}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \widetilde{\rho}(y_i)(y_i - \widehat{y}_i)}_{\substack{\text{Rectifier: measure the bias of using} \\ \text{predictions on labeled data}}}, \tag{16}$$

with $\widehat{y}_i = \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}$. Clearly, our proposed efficient estimator has the similar appearance as the existing PPI-type estimator in the absence of distribution shifts, e.g., in Angelopoulos et al. (2023a). However, due to the presence of label shift and the goal of achieving efficient estimation, our estimator incorporates several key differences compared to the original PPI estimator. First, to ensure efficient estimation, the density ratio model $\rho(y)$ plays a crucial role in the "rectifier" component. Accurate estimation of $\rho(y)$ is essential in the presence of label shift, whereas in the original PPI framework, $\rho(y) = 1$ by assumption. Second, the construction of the predictive values $\widehat{y}_i$'s is more sophisticated. Unlike the simple conditional mean estimates obtained from an ML prediction model, our approach defines the predictive values $\widehat{y} = \widehat{\mathbf{b}}(\mathbf{x})$ as in (14). This structure ensures that $\widehat{\mathrm{E}}(\widehat{Y} \mid y) = \widehat{\mathrm{E}}\{\widehat{\mathbf{b}}(\mathbf{X}) \mid y\} = y$ where $\widehat{\mathrm{E}}$ is the approximation to the expectation of $\mathbf{X} \mid Y$, guarantees the consistency of the direct estimation term in (16) for $\mathrm{E}_q(Y)$. Furthermore, $\widehat{y} = \widehat{\mathbf{b}}(\mathbf{x})$ is constructed to make the estimator in (16) efficient. In contrast, in the original PPI, the direct estimation term in (3) can be biased. Even when it is unbiased, the estimator is not efficient in the presence of label shift.

Remark 4 *We emphasize that unlike traditional UDA methods that heavily rely on the quality of the estimated individual pseudo-labels in the target domain (Liu et al., 2022), our proposed framework does not require to generate any pseudo-labels hence completely breaks free from such constraint. Instead, we only use a ML prediction model in the source domain to calculate an estimating equation. Further, by construction, our procedure automatically corrects for any potential biases that may be inherent in this procedure. Thus, the method has the advantage of not relying on the correctness of any pseudo-labels, and it simultaneously relies minimally on the quality of the ML model in that the performance of the final estimator only depends on the ML model in terms of efficiency, not consistency.*

Remark 5 *If the ultimate goal is to obtain high-quality pseudo-labels for the target population $\mathcal{Q}$, our general framework can be adapted to train a prediction model optimized for the target population. Suppose we postulate a prediction model $m(\mathbf{x}, \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. To find the prediction model that minimizes the mean squared error in the target population, we can specify the estimating function as $\mathbf{U}(y, \mathbf{x}, \boldsymbol{\theta}) = \frac{\partial m(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\{y - m(\mathbf{x}, \boldsymbol{\theta})\}$. By solving the proposed estimating equation with this $\mathbf{U}$, we will obtain an efficient estimator $\widehat{\boldsymbol{\theta}}$, and the resulting model $m(\mathbf{x}, \widehat{\boldsymbol{\theta}})$ can subsequently be deployed to assign pseudo-labels to the unlabeled target data.*

4. Asymptotic Theory

4.1 Asymptotic Properties of the Estimator in Section 3.1

As we outlined in Remark 1, a key feature of the proposed estimator $\widetilde{\rho}(y)$ in Section 3.1 is that its implementation does NOT require a correctly specified model of $\mathrm{E}_p(\cdot \mid \mathbf{x})$ to consistently estimate $\rho(y)$. To present our theory reflecting this flexibility in the choice of $\mathrm{E}_p(\cdot \mid \mathbf{x})$, we use superscript † throughout this Section 4.1. Specifically, we use $p^\dagger_{Y|\mathbf{X}}(y, \mathbf{x})$ and $\mathrm{E}^\dagger_p(\cdot \mid \mathbf{x})$ to denote the chosen conditional distribution function and conditional expectation with respect to the model of Y given $\mathbf{X}$ in the labeled population $\mathcal{P}$ that may or may not be correct.

To facilitate the presentation of the theoretical properties, we introduce some more notations. Let us denote

$$\begin{aligned} u^{*\dagger}(t,y) &\equiv p_Y(y)\int \frac{\rho^*(t)p^{\dagger}_{Y|\mathbf{X}}(t,\mathbf{x})}{\mathrm{E}^{\dagger}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{x}\}} f_{\mathbf{X}|Y}(\mathbf{x},y)d\mathbf{x}, \\ \mathcal{I}^{*\dagger}(a)(y) &\equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}^{\dagger}_p\{a(Y)\rho^*(Y)\mid \mathbf{X}\}}{\mathrm{E}^{\dagger}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{X}\}}\mid y\right] = \int a(t)u^{*\dagger}(t,y)dt, \\ v_h(y) &\equiv p_Y(y)K_h(y-y_0), \end{aligned}$$

and let $a_h^{*\dagger}(y)$ satisfy $\mathcal{I}^{*\dagger}(a_h^{*\dagger})(y) = v_h(y)$. For a function $a(y,\mathbf{x}) \in \mathbb{R}$, let $\|a(y,\mathbf{x})\|_p = [\int\{a(y,\mathbf{x})\}^p dyd\mathbf{x}]^{1/p}$ $(1 \leqslant p < \infty)$ be the L_p-norm of $a(y,\mathbf{x})$, and let $\|a(y,\mathbf{x})\|_\infty = \sup_{y,\mathbf{x}}|a(y,\mathbf{x})|$ be the sup-norm of $a(y,\mathbf{x})$. In addition, for a vector-valued function $\mathbf{a}(y,\mathbf{x}) = \{a_1(y,\mathbf{x}),\ldots,a_d(y,\mathbf{x})\}^{\mathrm{T}} \in \mathbb{R}^d$, let $\|\mathbf{a}(y,\mathbf{x})\|_\infty = \max_{k=1,\ldots,d}\|a_k(y,\mathbf{x})\|_\infty$.

We require the following regularity conditions to develop our theory.

(A1) $\rho^*(y)$ is bounded away from zero on the support of $p_Y(y)$, and has a bounded kth derivative for $k = 1,\ldots,m$.

(A2) $p^{\dagger}_{Y|\mathbf{X}}(y,\mathbf{x})$ is complete, i.e., $\mathrm{E}^{\dagger}_p\{g(Y)\mid \mathbf{x}\} = \int g(y)p^{\dagger}_{Y|\mathbf{X}}(y,\mathbf{x})dy = 0$ for all $\mathbf{x}$ implies $g(y) = 0$ almost surely. The function $\rho^*(\cdot)p^{\dagger}_{Y|\mathbf{X}}(\cdot,\mathbf{x})$ is square integrable.

(A3) $u^{*\dagger}(t,y)$ is square integrable, and has bounded kth derivatives with respect to t and y for $k = 1,\ldots,m$.

(A4) The supports of $f_{\mathbf{X}|Y}(\mathbf{x},y), p_Y(y), \rho(y)$ are compact. $\rho(y)$ is bounded, and $p(y)$ and $q(y)$ have bounded kth derivatives for $k = 1,\ldots,m$.

(A5) $K(\cdot)$ and $\widetilde{K}(\cdot)$ have compact support $[-1,1]$, are of order m, bounded, and have bounded kth derivatives for $k = 1,\ldots,m$.

(A6) $h \to 0$, $\{n\min(n,N-n)\}^{1/2}h \to \infty$, $nl^{2m} \to 0$, $nl^2 \to \infty$, and $nh^{-2m-1}l^{2m} \to 0$ as $n \to \infty$.

The requirement on the bandwidth l in Condition (A6) regulates the Nadaraya-Watson estimator of $\mathrm{E}(\cdot \mid y)$ based on $\widetilde{K}_l(\cdot)$ to converge at rate $o_p(n^{-1/4})$. To implement our estimator in practice, the bandwidths can be chosen as follows: suppose we adopt a kernel function $K(\cdot)$ that is of order $m = 2$, Gaussian for instance, then the bandwidth l needs to be of order between $n^{-1/2}$ and $n^{-1/4}$ according to Condition (A6). If we choose $l = C_1 n^{-1/3}$ for some constant C_1, then subsequently the other bandwidth h needs to be of order greater than $n^{-1/15}$ and still converge to zero. In our numerical studies, we chose $h = C_2 n^{-1/16}$ for some constant C_2.

Below, we present three results, corresponding to the behavior of the operator $\mathcal{I}^{*\dagger}$, the asymptotic normal distribution of the estimator $\widetilde{\delta}_0$, and the convergence rate of the estimator $\widetilde{\rho}(y)$, respectively. Their proofs are contained in Supplementary Material A.2, A.3, A.4, respectively.

Lemma 6 *Under Conditions (A1)-(A4),*

(i) the linear operator $\mathcal{I}^{*\dagger} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ *is invertible,*

(ii) there exist constants C_1, C_2 *such that* $0 < C_1, C_2 < \infty$, $\|\mathcal{I}^{*\dagger}(a)\|_2 \leqslant C_1\|a\|_2$, *and* $\|\mathcal{I}^{*\dagger-1}(a)\|_2 \leqslant C_2\|a\|_2$ *for all* $a(y) \in L^2(\mathbb{R})$.

Theorem 7 *Assume Conditions (A1)-(A6). For any choice of* $\rho^*(y)$ *and* $\mathrm{E}_p^\dagger(\cdot \mid \mathbf{x})$, *at any* y_0, $\mathrm{var}(\widetilde{\delta}_0)^{-1/2}\{\widetilde{\delta}_0 - \mathrm{bias}(\widetilde{\delta}_0) - q_Y(y_0)\}$ *converges in distribution to the standard normal distribution, where*

$$\begin{aligned}
\mathrm{bias}(\widetilde{\delta}_0) &= q_Y^{(m)}(y_0)\frac{\int t^m K(t)dt}{m!}h^m + O(h^{m+1}) + o(n^{-1/2}h^{-1/2}),\\
\mathrm{var}(\widetilde{\delta}_0) &= n^{-1}\mathrm{E}_p\left\{\left(\rho(Y)\left[K_h(Y-y_0) - \frac{\mathrm{E}_p^\dagger\{a_h^{*\dagger}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\dagger\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{X}\}}\right]\right)^2\right\}\\
&\quad + (N-n)^{-1}\mathrm{E}_q\left(\left[\frac{\mathrm{E}_p^\dagger\{a_h^{*\dagger}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\dagger\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{X}\}} - \delta_0\right]^2\right)\\
&\quad + o[\{\min(n, N-n)\}^{-1/2}n^{-1/2}h^{-1}].
\end{aligned}$$

Theorem 7 guarantees that our proposed estimator $\widetilde{\delta}_0$ is consistent for $q_Y(y_0)$ regardless the chosen models for $\rho(y)$ and $\mathrm{E}_p(\cdot \mid \mathbf{x})$ are correct or not. According to the bias and standard error of $\widetilde{\delta}_0$ presented in Theorem 7, $\widetilde{\delta}_0$ converges to $q_Y(y_0)$ as long as the bandwidth h is chosen to satisfy the regularity conditions in Condition (A6). Since this result holds under arbitrary models $\rho^*(y)$ and $\mathrm{E}_p^\dagger(\cdot \mid \mathbf{x})$, we are allowed to freely choose the initial guess for $\rho(y)$ and the prediction model for $\mathrm{E}_p(\cdot \mid \mathbf{x})$.

Based on the above point-wise convergence behavior of the estimator of $q_Y(y_0)$, we further elaborate the uniform convergence rate of $\widetilde{\rho}(\cdot)$. For simplicity, we consider the function $\widetilde{\rho}(\cdot)$ constructed as a piecewise linear interpolation of $\{\widetilde{\rho}(t_1), \ldots, \widetilde{\rho}(t_{M_n})\}$. In short, the estimator converges uniformly to the true $\rho(y)$ under a few additional regularity conditions. Such convergence allows us to successfully implement the EIF for estimating a general parameter $\boldsymbol{\theta}$.

Theorem 8 *Assume Conditions (A1)-(A6). In addition, assume* $\rho(y)$ *is Lipschitz continuous. Then for any choice of* $\rho^*(y)$ *and* $\mathrm{E}_p^\dagger(\cdot \mid \mathbf{x})$,

$$\|\widetilde{\rho}(y) - \rho(y)\|_\infty = O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2}\log n] = o_p(1).$$

4.2 Asymptotic Properties of the Estimator in Section 3.2

To state the theoretical properties of $\widehat{\delta}_0$ and $\widehat{\rho}(y)$ studied in Section 3.2, we introduce some more notations. Define

$$\begin{aligned}
\widetilde{u}(t,y) &\equiv p_Y(y)\int \frac{\widetilde{\rho}(t)p_{Y|\mathbf{X}}(t,\mathbf{x})}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} f_{\mathbf{X}|Y}(\mathbf{x}, y)d\mathbf{x},\\
\widetilde{\mathcal{I}}(a)(y) &\equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{X}\}} \mid y\right] = \int a(t)\widetilde{u}(t,y)dt,
\end{aligned}$$

and let $\widetilde{a}_h(y)$ satisfy $\widetilde{\mathcal{I}}(\widetilde{a}_h)(y) = v_h(y)$. We also need the following regularity conditions to develop the theory.

(A1') $\widetilde{\rho}(y)$ is bounded away from zero on the support of $p_Y(y)$, and has a bounded kth derivative for $k = 1, \ldots, m$.

(A2') $p_{Y|\mathbf{X}}(y, \mathbf{x})$ is complete, i.e., $\mathrm{E}_p\{g(Y) \mid \mathbf{x}\} = \int g(y) p_{Y|\mathbf{X}}(y, \mathbf{x}) dy = 0$ for all $\mathbf{x}$ implies $g(y) = 0$ almost surely. The function $\widetilde{\rho}(\cdot) p_{Y|\mathbf{X}}(\cdot, \mathbf{x})$ is square integrable.

(A3') $\widetilde{u}(t, y)$ is square integrable, and has bounded kth derivatives with respect to t and y for $k = 1, \ldots, m$.

The following result concerns the asymptotic normality as well as the estimation efficiency of $\widehat{\delta}_0$, with the proof contained in Supplementary Material A.5.

Theorem 9 *Assume Conditions (A1')-(A3'), (A4)-(A6). If $\widetilde{\rho}(\cdot)$ satisfies $\sup_y |\widetilde{\rho}(y) - \rho(y)| = o_p(1)$ and $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ satisfies $|\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{a(Y) \mid \mathbf{x}\}| = o_p(n^{-1/4}\|a\|_2)$ for every square integrable $a(y)$ and $\mathbf{x}$, then for any y_0, $\mathrm{var}(\widehat{\delta}_0)^{-1/2}\{\widehat{\delta}_0 - \mathrm{bias}(\widehat{\delta}_0) - q_Y(y_0)\}$ converges in distribution to the standard normal distribution, where*

$$
\begin{aligned}
\mathrm{bias}(\widehat{\delta}_0) &= q_Y^{(m)}(y_0) \frac{\int t^m K(t) dt}{m!} h^m + O(h^{m+1}) + o(n^{-1/2} h^{-1/2}), \\
\mathrm{var}(\widehat{\delta}_0) &= n^{-1} \mathrm{E}_p \left\{ \left(\rho(Y) \left[K_h(Y - y_0) - \frac{\mathrm{E}_p\{a_h(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{X}\}} \right] \right)^2 \right\} \\
&\quad + (N-n)^{-1} \mathrm{E}_q \left(\left[\frac{\mathrm{E}_p\{a_h(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{X}\}} - \delta_0 \right]^2 \right) \\
&\quad + o[\{\min(n, N-n)\}^{-1/2} n^{-1/2} h^{-1}].
\end{aligned}
$$

Theorem 9 indicates that using suitable nuisance estimators for ρ and E_p and bandwidth h chosen as Condition (A6), $\mathrm{bias}(\widehat{\delta}_0)$ vanishes and $\mathrm{var}(\widehat{\delta}_0)$ achieves the semiparametric efficiency bound for estimating $\mathrm{E}_q\{K_h(Y - y_0)\}$, where the corresponding EIF is derived in Supplementary Material A.1. Hence, at fixed choices of $K(\cdot)$ and h, $\widehat{\delta}_0$ is the semiparametrically efficient estimator of $q_Y(y_0)$. An interesting observation is that for the nuisance estimator $\widetilde{\rho}$, uniform consistency is already sufficient to guarantee the efficiency of our proposed estimator $\widehat{\delta}_0$, and no specific convergence order is imposed on $\widetilde{\rho}$.

In terms of estimating $\rho(y_0) = q_Y(y_0)/p_Y(y_0)$ at an arbitrary y_0, we can construct $\widehat{\rho}(y_0) = \widehat{q}_Y(y_0)/\widehat{p}_Y(y_0)$ where $\widehat{p}_Y(y_0)$ is the standard Nadaraya-Watson kernel estimator. Because $\widehat{p}_Y(y_0)$ and $\widehat{q}_Y(y_0)$ are both efficient at the given kernel function $K(\cdot)$ and bandwidth h, by the delta-method, we can see that $\widehat{\rho}(y_0)$ is also efficient. We summarize this result in Corollary 10 below and omit its proof.

Corollary 10 *At any y_0, the estimator $\widehat{\rho}(y_0) = \widehat{q}_Y(y_0)/\widehat{p}_Y(y_0)$, where $\widehat{p}_Y(y_0)$ is the Nadaraya-Watson estimator and $\widehat{q}_Y(y_0)$ is given in Theorem 9, is efficient given the kernel function $K(\cdot)$ and the bandwidth h.*

We further present the uniform convergence result of $\widehat{\rho}(y)$. Since the bias and the variance of $\widehat{\rho}(y_0)$ are of the same order as those of $\widetilde{\rho}(y_0)$ as presented in Theorems 7 and 9, $\|\widehat{\rho}(y) - \rho(y)\|_\infty$ is also of the same order as $\|\widetilde{\rho}(y) - \rho(y)\|_\infty$ presented in Theorem 8. We state this result in Corollary 11 below and omit its proof.

Corollary 11 *Assume the same conditions as in Theorem 9. In addition, assume $\rho(y)$ is Lipschitz continuous. Then*

$$\|\widehat{\rho}(\cdot) - \rho(\cdot)\|_\infty = O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2} \log n] = o_p(1).$$

4.3 Asymptotic Properties of the Estimator in Section 3.3

For the estimator $\widehat{\boldsymbol{\theta}}$, its asymptotic normality and semiparametric efficiency can be found below, with the proof contained in Supplementary Material A.6. We first list some additional regularity conditions.

(B1) $\mathrm{E}_q\{\partial \mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$ and $\mathrm{E}\{\partial \boldsymbol{\phi}_{\mathrm{eff}}(R, RY, \mathbf{X}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$ are invertible, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is compact, and $\mathrm{E}\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{\phi}_{\mathrm{eff}}(R, RY, \mathbf{X}, \boldsymbol{\theta})\|_2\} < \infty$. $\mathbf{U}(y, \mathbf{x}, \boldsymbol{\theta})$ is twice differentiable with respect to y and its derivative is bounded.

(B2) The function $u(t, y)$ is bounded and has bounded derivatives with respect to t and y on its support. $\mathbf{a}(y)$ in (13) is bounded.

(B3) The bandwidth l satisfies $n(\log n)^{-4} l^2 \to \infty$ and $N^2 n^{-1} l^4 \to 0$.

Theorem 12 *Assume Conditions (A1'), (A2'), (A4), (A5), (B1)-(B3). Suppose $\widetilde{\rho}(\cdot)$ satisfies $\sup_y |\widetilde{\rho}(y) - \rho(y)| = o_p(1)$ and $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ satisfies $\|\widehat{\mathrm{E}}_p\{\mathbf{a}(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{\mathbf{a}(Y) \mid \mathbf{x}\}\|_\infty = o_p(n^{-1/4})$ for every bounded $\mathbf{a}(y)$ and $\mathbf{x}$, then*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N\left[\mathbf{0}, \mathrm{var}\{\sqrt{\pi}\boldsymbol{\phi}_{\mathrm{eff}}(R, RY, \mathbf{X}, \boldsymbol{\theta})\}\right]$$

in distribution as $n \to \infty$. Clearly, $\widehat{\boldsymbol{\theta}}$ achieves the semiparametric efficiency.

Theorem 12 supports that $\mathrm{var}(\widehat{\boldsymbol{\theta}})$ achieves the semiparametric efficiency bound for estimating $\boldsymbol{\theta}$, as long as the estimator $\widehat{\boldsymbol{\theta}}$ is implemented with a consistent $\widetilde{\rho}$ and an $o_p(n^{-1/4})$-consistent $\widehat{\mathrm{E}}_p$. This theoretical result differs from the results by Lee et al. (2025) on their singly flexible estimator $\widetilde{\boldsymbol{\theta}}$. Lee et al. (2025) provided (i) the consistency of $\widetilde{\boldsymbol{\theta}}$ given an arbitrary ρ^* and an $o_p(n^{-1/4})$-consistent $\widehat{\mathrm{E}}_p$, and (ii) the semiparametric efficiency of $\widetilde{\boldsymbol{\theta}}$ given further that ρ^* equals to the true ρ. We point out that (ii) is almost impossible in practice since there is almost no chance of an arbitrary chosen function ρ^* being equal to the true function ρ, an object challenging to identify under our problem setting as discussed earlier. On the other hand, our proposed estimator $\widehat{\boldsymbol{\theta}}$ can achieve the semiparametric efficiency given an $o_p(n^{-1/4})$-consistent $\widehat{\mathrm{E}}_p$ and a consistent $\widetilde{\rho}$, and we can easily provide $\widetilde{\rho}$ that is consistent for the true ρ by implementing our proposed estimators $\widetilde{\delta}_0$ or $\widehat{\delta}_0$.

In Theorems 9 and 12, we require the conditional expectation estimators $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ to attain the $o_p(n^{-1/4})$ convergence rate with respect to different norms in each theorem. Nonetheless, the requirements not significantly differ from each other. To see this, firstly, note that we regulate the gap between $\widehat{\mathrm{E}}_p$ and E_p by the absolute value in Theorem 9, while in Theorem 12, by the maximum of the sup-norms. We can treat these two norms interchangeably, as long as the parameter of interest has finite dimension which matches the dimension of the estimating equation, and the functions in the estimating equation are bounded uniformly on the support of variables. Secondly, we require the convergence rate

is proportional to $\|a(y)\|_2$ in Theorem 9, while it is not in Theorem 12. This gap is not significant as well. Specifically, consider a vector function $\mathbf{b}(y)$ not necessarily bounded and the convergence rate requirement in Theorem 12. As long as we can identify the expression of $\|\mathbf{b}(y)\|_\infty$, letting $\mathbf{a}(y) = \mathbf{b}(y)/\|\mathbf{b}(y)\|_\infty$ transforms $\mathbf{b}(y)$ into the bounded function $\mathbf{a}(y)$. Now, suppose $\widehat{\mathrm{E}}_p$ satisfies the convergence rate requirement in Theorem 12, then plugging $\mathbf{a}(y)$ in the requirement leads us to $\|\widehat{\mathrm{E}}_p\{\mathbf{b}(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{\mathbf{b}(Y) \mid \mathbf{x}\}\|_\infty = o_p(n^{-1/4}\|\mathbf{b}(y)\|_\infty)$, which now is essentially similar to the one in Theorem 9.

5. Discussion on Discrete Y

In Sections 2 to 4, our proposed methodology and theoretical analysis have focused on scenarios where the outcome Y is continuous. We point out here that the established results apply directly to the discrete setting.

Without loss of generality, assume Y takes values $1, \ldots, K$. To use our method for this case, all we need to do is to adopt bandwidths h and l that are less than 1, then no other changes are needed. In fact, expressions will be simplified. For example, to construct the efficient estimator of $\Pr(Y = y_0)$ in $\mathcal{Q}$, first calculate weights $\widehat{w}_i = [\widehat{\mathrm{E}}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}]^{-1}$, and then we solve

$$\sum_{i=1}^{N} \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\} \frac{r_i \mathbb{I}(y = y_i)}{\sum_{j=1}^{N} r_j \mathbb{I}(y = y_j)} = \mathbb{I}(y = y_0)$$

for $\widehat{a}(y)$, then estimate $\Pr(Y = y_0)$ in $\mathcal{Q}$ by

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\rho}(y_i)\left[\mathbb{I}(y_i = y_0) - \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}\right] + \frac{1}{N-n}\sum_{i=n+1}^{N} \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid \mathbf{x}_i\}.$$

In fact, estimating $\rho(y)$ is equivalent to estimating $\rho(k)$, $k = 1, \ldots, K$. Thus, the proposed estimators $\widetilde{\rho}$ and $\widehat{\rho}$ will achieve the parametric convergence rate $n^{-1/2}$. Consequently, $\widehat{\rho}(k)$'s, will also retain the parametric convergence rate. The estimator $\widehat{\boldsymbol{\theta}}$, together with $\widehat{\rho}(k)$'s, will preserve the established semiparametric efficiency, since the efficiency result of $\widehat{\boldsymbol{\theta}}$ provided in Theorem 12 and of $\widehat{\rho}(k)$'s in Corollary 10 is guaranteed as long as $\widetilde{\rho}(y)$ is consistent.

6. Simulation Studies

We present a simulation study to evaluate the performance of the proposed estimators. The simulation is designed to first demonstrate the density ratio estimators proposed in Section 3.1, then compare various estimators of the mean and variance of Y in $\mathcal{Q}$. We first introduce the estimators being compared, followed by the process of generating the simulated data, and how the estimators are implemented. The results are then illustrated using plots and tables and discussed in detail.

We compare the following estimators of $\mathrm{E}_q\{V(Y)\}$, where $V(Y) = Y$ and $V(Y) = Y^2 - \{\mathrm{E}_q(Y)\}^2$.

- `Shift-dependent`*: The shift dependent estimator $N^{-1}\sum_{i=1}^{N} r_i \rho^*(y_i) V(y_i)/\pi$ naively incorporates a working model $\rho^*(y)$.

- `BBSE` and `BBSE+PPI`: The estimators based on the Black-Box Shift Estimation (Lipton et al., 2018). The latter incorporates prediction-powered inference (PPI) proposed by Angelopoulos et al. (2023a).

- `ELSA` and `ELSA+PPI`: The estimators based on the Efficient Label Shift Adaptation (Tian et al., 2023), with the latter version utilizing the PPI framework (Angelopoulos et al., 2023a).

- `Doubly-flexible**` and `Singly-flexible*`: The estimators proposed by Lee et al. (2025) that incorporate a working model $\rho^*(y)$. `Doubly-flexible**` further adopts a working regression model $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta})$, while `Singly-flexible*` adopts a nonparametric regression estimator $\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\}$.

- `Efficient`: Variations of the efficient estimators proposed in Section 3.3 are implemented and denoted as `Efficient~` and `Efficient^`, each corresponding to setting $\rho(y)$ to be $\widetilde{\rho}(y)$ or $\widehat{\rho}(y)$. Note that these two estimators are asymptotically exactly the same.

- `Oracle`: The same efficient estimator is implemented with setting $\rho(y)$ to be the truth. Since the true $\rho(y)$ is unknown in practice, `Oracle` serves only as a benchmark.

The simulated data are generated with $N = 500$ based on the following mechanism:

$$R \sim \mathrm{Ber}(1/2),\ Y \mid R = 1 \sim N(0, 2),\ Y \mid R = 0 \sim N(\theta_1, \theta_2),\ \mathbf{X} \mid Y \sim N_3(\boldsymbol{\alpha} Y, \mathbf{I}_3),$$

where the mean of Y in $\mathcal{Q}$, $\theta_1 = 1$, the variance of Y in $\mathcal{Q}$, $\theta_2 = 1$, and $\boldsymbol{\alpha} = (-0.5, 0.5, 1)$.

In our simulation study, the working model for $\rho(y)$ is specified as

$$\rho^*(y) = c^* \rho(y) \exp(0.2y + 0.1y^2),$$

where $c^* = \pi / [N^{-1} \sum_{i=1}^{N} \{r_i \rho(y_i) \exp(0.2y_i + 0.1y_i^2))\}]$. This formulation incorrectly specifies $\rho(y)$. For the estimators $\widetilde{\rho}(y)$ and $\widehat{\rho}(y)$ described in Section 3.1, we select the bandwidth $h = 0.5n^{-1/16}$, where the order is chosen to satisfy the regularity condition listed in Condition (A6). For the estimation of $\mathrm{E}\{b(\mathbf{X}) \mid y\}$ in the implementation of `Efficient` and the estimators proposed by Lee et al. (2025), we adopt the Nadaraya-Watson estimator with bandwidth $l = 1.5n^{-1/3}$. In addition, for the implementation of `Efficient`, `Oracle`, and `Singly-flexible*`, the Nadaraya-Watson estimator of $\mathrm{E}_p\{a(Y) \mid \mathbf{x}\}$ is used as our machine-learning estimator, with bandwidth $3n^{-1/7}$. For `BBSE` and `ELSA`, as these methods are applicable to discrete Y only, we discretize the continuous Y into 9 bins using the quantiles of Y in $\mathcal{P}$. When implementing `PPI`, predictions for Y are generated using a linear regression model trained on the sample from $\mathcal{P}$. The standard errors of the estimates from these methods are estimated using the sandwich formula. For `Doubly-flexible**`, the working model $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta})$ is set to be a normal regression model with the wrongly transformed covariate

$$\left(X_1, \exp\left(\frac{X_2}{2}\right), \frac{X_3}{1 + \exp(X_2)} + 10 \right).$$

The parameter $\boldsymbol{\beta}$ is then estimated via maximum likelihood estimation.

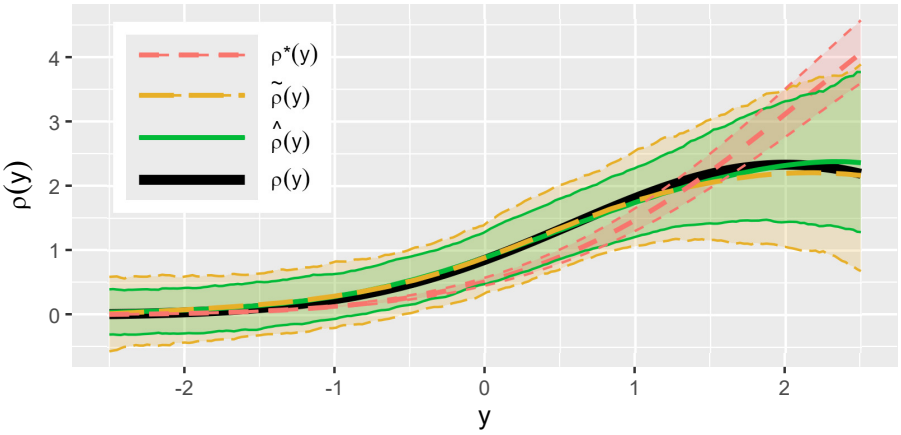

Figure 1: Plot of $\rho(y)$ estimates from 1000 simulation replicates. Lines: the average of 1000 estimates, Shades: the pointwise 90% empirical confidence band of 1000 estimates.

Figure 1 presents the $\rho(y)$ estimates from 1000 simulation replicates. The working model $\rho^*(y)$ significantly deviates from the true density ratio $\rho(y)$ as suggested by its definition. In contrast, our proposed estimators $\widetilde{\rho}(y)$ and $\widehat{\rho}(y)$ proposed in Section 3.1 provide a much more accurate estimate of $\rho(y)$. We also observe from the empirical confidence bands that the variability of $\widehat{\rho}(y)$ is much less compared to $\widetilde{\rho}(y)$, i.e., $\widehat{\rho}(y)$ provides closer alignment with $\rho(y)$ than $\widetilde{\rho}(y)$. This improvement illustrates the effectiveness of our proposed self-improvement procedure.

We now present the results for the estimation of the mean and variance of Y in $\mathcal{Q}$. The estimators are compared in terms of mean squared error multiplied by 100 (MSE), bias multiplied by 10 (Bias), standard error multiplied by 10 (SE), asymptotic relative efficiency calculated as the estimator's MSE divided by the `Oracle`'s MSE (ARE), and a coverage rate of asymptotic confidence interval with a confidence level $1 - \alpha = 0.95$ (CI) based on 1000 simulation replicates.

Figure 2 and Table 2 illustrate the performance of estimates based on 1000 simulation replicates. `Shift-dependent`* exhibits significant bias, and consequently, its CI falls far short of the nominal 95% level. The `BBSE`, `ELSA`, and their PPI variants show larger MSEs compared to our proposed estimators. Specifically, `BBSE` and its PPI variant exhibit significant instability, although the incorporation of PPI noticably improves its performance. `ELSA` and `ELSA+PPI` performed identically in this setting because in implementing `ELSA+PPI`, an `ELSA`-based $\widehat{\rho}$ was first constructed to match the 1st and 2nd moments of $\mathbf{X}$, then $\widehat{Y}$ was constructed from the linear regression of Y on $\mathbf{X}$ in $\mathcal{P}$, leading to a zero augmentation term. In contrast, both `Doubly-flexible`** and `Singly-flexible`* by Lee et al. (2025) provide relatively consistent estimates with improved Bias and CI. `Efficient`$^\sim$ and `Efficient`$^\wedge$, implemented using our proposed method, show an improvement in MSE compared to the estimators by Lee et al. (2025). Remarkably, the ARE's of `Efficient`$^\sim$ and `Efficient`$^\wedge$ are close to 1, indicating that these estimators exhibit efficiency similar to `Oracle` which

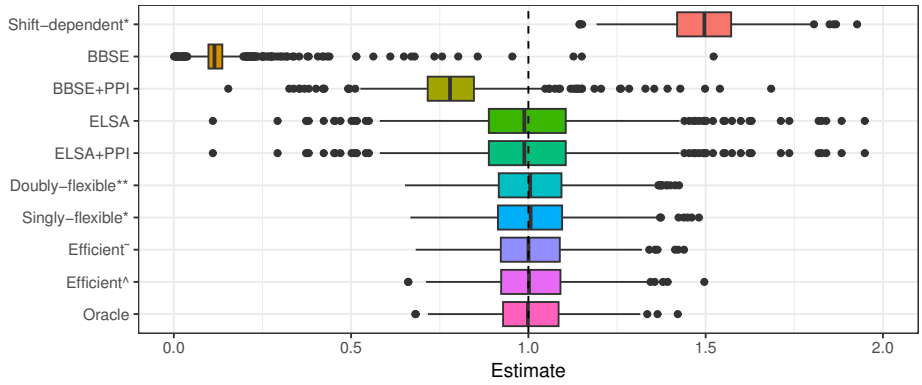

Figure 2: Boxplot of $\mathrm{E}_q(Y)$ estimates from 1000 simulation replicates.

Estimator	MSE	Bias	SE	ARE	CI
`Shift-dependent`*	25.8428	4.9437	1.1844	19.605	0.220
`BBSE`	237.9293	-8.2755	13.0236	180.496	0.205
`BBSE+PPI`	129.1417	-1.7068	11.2408	97.968	0.619
`ELSA`	4.7037	0.0637	2.1690	3.568	0.996
`ELSA+PPI`	4.7037	0.0637	2.1690	3.568	0.996
`Doubly-flexible`**	1.8189	0.0838	1.3461	1.380	0.951
`Singly-flexible`*	1.8039	0.0933	1.3399	1.368	0.952
`Efficient`~	1.4171	0.0620	1.1888	1.075	0.961
`Efficient`^	1.4096	0.0649	1.1855	1.069	0.966
`Oracle`	1.3182	0.0403	1.1474	1.000	0.961

Table 2: Summary of $\mathrm{E}_q(Y)$ estimates from 1000 simulation replicates.

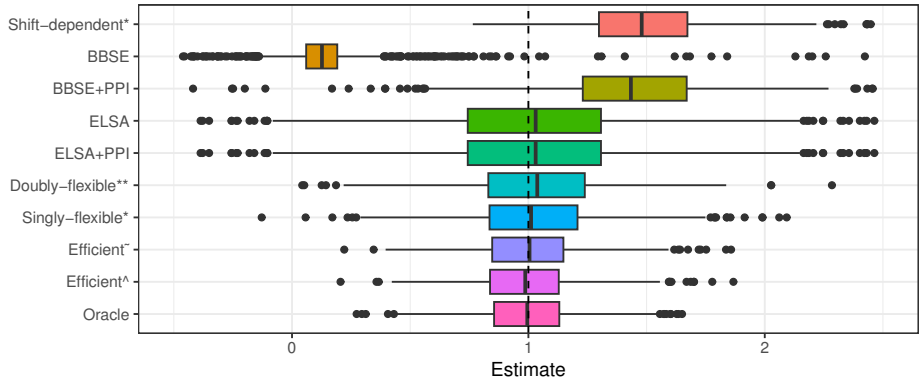

Figure 3: Boxplot of $\mathrm{var}_q(Y)$ estimates from 1000 simulation replicates.

Estimator	MSE	Bias	SE	ARE	CI
`Shift-dependent*`	31.9295	4.9199	2.7806	6.730	0.819
`BBSE`	413.3995	-9.7303	17.8617	87.140	0.464
`BBSE+PPI`	359.7540	3.8240	18.5870	75.832	0.863
`ELSA`	39.7373	0.3805	6.2954	8.376	1.000
`ELSA+PPI`	39.7373	0.3805	6.2954	8.376	1.000
`Doubly-flexible**`	9.8121	0.3287	3.1167	2.068	0.959
`Singly-flexible*`	8.7310	0.2380	2.9467	1.840	0.957
`Efficient~`	5.3925	0.0281	2.3232	1.137	0.965
`Efficient^`	5.1272	-0.1510	2.2604	1.081	0.968
`Oracle`	4.7441	-0.0969	2.1770	1.000	0.961

Table 3: Summary of $\mathrm{var}_q(Y)$ estimates from 1000 simulation replicates.

uses the true $\rho(y)$ which is unknown in practice. These results provide empirical support for Theorem 12. Furthermore, the confidence intervals computed using the asymptotic distribution provided in Theorem 12 closely align with the nominal 95% rate, further validating the accuracy of our asymptotic results.

Figure 3 and Table 3 present the boxplot and the summary table of the variance estimates from 1000 simulation replicates. Similar to the mean estimation results, `Shift-dependent*` estimator exhibits significant bias, with its CI again failing to reach the nominal 95% level. The discretization-based baseline methods `BBSE`, `ELSA`, and their PPI variants also struggle to achieve high efficiency. In addition, both `Doubly-flexible**` and `Singly-flexible*` provide relatively consistent estimates with lower Bias and improved CI. Notably, `Efficient~` and `Efficient^` demonstrate a more pronounced improvement in MSE compared to the estimators by Lee et al. (2025). In fact, the ARE's of both `Efficient~` and `Efficient^` are close to 1, indicating that the MSE's of these estimators are close that of `Oracle`. Moreover, the confidence intervals based on the asymptotic distribution derived in Theorem 12 once again demonstrate alignment with the nominal 95% level, reinforcing the validity of our theoretical results.

7. Real Data Applications

7.1 Plankton counting with computer vision

Now we analyze a real-world dataset and demonstrate an improvement in inference by our proposal from the existing methods. Our goal is to count the number of planktons from the images taken by the Imaging FlowCytobot, an automated submersible flow cytometry system at Woods Hole Oceanographic Institution (Olson et al., 2003; Orenstein et al., 2015). The dataset consists of the images of matters captured by the FlowCytobot from 2006 to 2014, the number of planktons in 2014 was predicted based on a ResNet trained on the dataset from 2006 to 2013. The distribution of the numbers of planktons shifted from 2013 to 2014, mainly due to the change in the base frequencies of plankton observations (Angelopoulos et al., 2023a). Accordingly, we consider the 2013 dataset with 421,238 labeled observations as taken from $\mathcal{P}$, and the 2014 dataset with 329,832 unlabeled observations as taken from $\mathcal{Q}$. To validate the label shift assumption, we conducted a conditional independence test using the `CondIndTests` package in `R` to check whether the ResNet predicted probabilities are conditionally independent of the population indicator given the label. The test yielded a p-value of approximately 0.5, indicating no significant evidence against the label shift assumption. This validation aligns with the observation by Angelopoulos et al. (2023a) that the distribution shift in this dataset was driven by changes in the class base frequencies.

To estimate the proportion of the planktons and compare inference results, we implemented the following methods: the prediction powered inference by Angelopoulos et al. (2023a) (`PPI*`), the shift dependent estimator $N^{-1}\sum_{i=1}^{N} r_i\rho^*(y_i)y_i/\pi$ where y_i is coded 1 if the image is a plankton or 0 otherwise (`Shift-dependent*`), the singly flexible estimator by Lee et al. (2025) (`Singly-flexible*`), and our proposed efficient estimators (`Efficient~` and `Efficient^`). To implement the existing methods, we constructed a naive density ratio estimator using the true labels Y_i and the predicted labels $\widehat{Y}_i$ in $\mathcal{P}$, specifically, $[\rho^*]_k = [q_Y^*]_k/[\widehat{p}_Y]_k$ where $q_Y^* = \widehat{C}^{-1}\widehat{q}_X$, $[\widehat{C}]_{k+1,l+1} = \sum_{i=1}^{n}\mathbb{I}(\widehat{Y}_i = k, Y_i = l)/\sum_{i=1}^{n}\mathbb{I}(Y_i = l)$, $[\widehat{q}_X]_{k+1} = (N-n)^{-1}\sum_{i=n+1}^{N}\mathbb{I}(\widehat{Y}_i = k)$, and $[\widehat{p}_Y]_{k+1} = n^{-1}\sum_{i=1}^{n}\mathbb{I}(Y_i = k)$. We constructed the conditional expectation estimators $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ using the last softmax layer of the ResNet, then used to implement the singly-flexible and our efficient estimators. We note that although the input raw image $\mathbf{x}$ is high-dimensional (dimension 224×224), our proposed method effectively handles this input by integrating kernel methods and ML methods. Specifically, the kernel smoothing operation in our procedure is exclusively applied to approximate functions given the one-dimensional Y. On the other hand, the high-dimensional feature extraction and prediction from $\mathbf{x}$ are all addressed entirely by the ML model (the ResNet in this analysis), which is well-suited for such high-dimensional data.

The inference results are provided in Table 4. We report the point estimate $\widehat{\theta}$, the estimated standard error $\widehat{\mathrm{sd}}(\widehat{\theta})$, and the width of the 95% confidence interval, and its lower and upper bound. Firstly, the `PPI*` provided the widest confidence interval and its width was drastically wider than the rest of the implemented methods. Other than the `PPI*`, all methods provided almost the same point estimates. In terms of the width of CI, our proposed methods `Efficient~` and `Efficient^` provided narrower confidence intervals compared to the `Shift-dependent*` and `Singly-flexible*`. In this analysis, the `Efficient~` and `Efficient^` resulted in the same interval estimate.

Estimator	$\widehat{\theta}$	$\widehat{\mathrm{sd}}(\widehat{\theta})$	CI Width	CI
`PPI*`	39176.00	2159.734	8466.000	[34943.00, 43409.00]
`Shift-dependent*`	39152.00	147.521	578.271	[38862.87, 39441.14]
`Singly-flexible*`	39152.01	168.780	661.606	[38821.21, 39482.81]
`Efficient~`	39152.01	110.477	433.062	[38935.48, 39368.54]
`Efficient^`	39152.01	110.477	433.062	[38935.48, 39368.54]

Table 4: Summary of inference on the number of planktons.

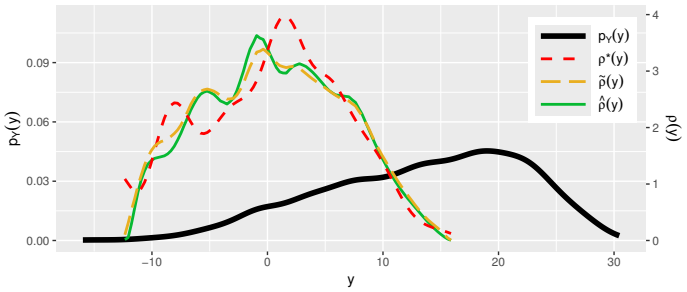

Figure 4: Temperature distribution in $\mathcal{P}$ and the density ratio estimates.

7.2 Temperature forecast with humidity

We now analyze the weather dataset of Szeged, Hungary from 2006 to 2016, and compare the performance of our proposed methods against the existing methods. The dataset is publicly available at the following link: `https://www.kaggle.com/datasets/budincsevity/szeged-weather`. Several weather characteristics of Szeged were recorded including temperature, humidity, precipitation, and visibility. Our analysis goal is to forecast the average temperature of Szeged in November and December using humidity (Kim et al., 2024). In this dataset, we consider the daily average humidity as a covariate and the daily average temperature as a response, and take 3,348 observations between January and October as taken from $\mathcal{P}$ and 671 observations between November and December as taken from $\mathcal{Q}$. The empirical distribution of the average temperature in $\mathcal{P}$ is visualized as the black curve in Figure 4 based on its kernel density estimate.

To forecast the average temperature of Szeged in November and December, we implemented the following methods: the shift dependent estimator $N^{-1}\sum_{i=1}^{N} r_i\rho^*(y_i)y_i/\pi$, the doubly-flexible (`Doubly-flexible**`) and singly-flexible (`Singly-flexible*`) estimators by Lee et al. (2025), and our proposed efficient estimators (`Efficient~` and `Efficient^`). To implement the existing methods, we created an arbitrary function $\rho^*(y)$ which is visualized in Figure 4. We adopted a normal regression model as a working model for $p_{Y|\mathbf{X}}$, and used the Nadaraya-Watson estimator based on the observations in $\mathcal{P}$ as $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$. Further, the bandwidth h is chosen as $1.5n^{-1/16}$ and l as $3n^{-1/3}$.

Figure 4 visualizes the density ratio estimates $\widetilde{\rho}$ and $\widehat{\rho}$, and Table 5 summarizes the inference results. We again report the point estimate $\widehat{\theta}$, the estimated standard error $\widehat{\mathrm{sd}}(\widehat{\theta})$, and the width of the 95% confidence interval, and its lower and upper bound. The point estimate from `Shift-dependent*`, 5.538, was much bigger than the rest of the methods,

Estimator	$\widehat{\theta}$	$\widehat{\mathrm{sd}}(\widehat{\theta})$	CI Width	CI
`Shift-dependent`*	5.538	0.153	0.602	[5.238, 5.839]
`Doubly-flexible`**	4.267	0.504	1.976	[3.280, 5.255]
`Singly-flexible`*	3.998	0.160	0.629	[3.683, 4.312]
`Efficient`~	4.086	0.123	0.482	[3.845, 4.327]
`Efficient`^	4.131	0.126	0.493	[3.884, 4.377]

Table 5: Summary of inference on the average temperature.

which were around 4.1. Also, we do not see much overlap of the CI of `Shift-dependent`* with the rest. On the other hand, we notice that the CI width of `Doubly-flexible`** was much wider than `Singly-flexible`* and our efficient estimators, possibly due to the departure of the chosen working model from the underlying true distribution. `Singly-flexible`* narrows the CI width from `Doubly-flexible`** by adopting an estimator of $\mathrm{E}_p(\cdot \mid \mathbf{x})$ that is more flexible hence more likely to cover the truth, but our efficient estimators `Efficient`~ and `Efficient`^ improve even more and provide the narrowest confidence intervals.

8. Conclusion

In this paper, we consider distribution shift, specifically label shift, and we propose an efficient inference procedure. The presence of distribution shift introduces additional complexity compared to cases without such shifts. Specifically, rather than simply predicting the conditional mean function as $\widehat{Y}$, one must predict $\widehat{Y}$ as $\widehat{\mathbf{b}}(\mathbf{X})$, still a function of the feature $\mathbf{X}$ but with a significantly more intricate structure. A crucial element for achieving efficient inference is modeling the outcome density ratio between labeled and unlabeled data $\rho(y)$. To this end, we introduce a progressive estimation process comprising three stages: an initial heuristic guess, a consistent estimation, and finally, an efficient estimation. This self-driven, iterative approach is unconventional in the statistical literature and may hold independent interest beyond its application in this context.

Finally, throughout the paper, we have assumed that the target label support is contained within the source label support. While this assumption is necessary to achieve identifiability, it restricts the direct application of the proposed method to other more complex domain adaptation scenarios, such as partial (Gu et al., 2021), few-shot (Lu et al., 2025), source-free (Lu et al., 2026a) and blended-target (Lu et al., 2026b) domain adaptations. Extending our semiparametric methodology to accommodating these cases requires additional assumptions to achieve identifiability. Under various assumptions, how to best handle these novel scenarios will be important directions for future research.

Acknowledgments

The research was supported in part by the National Research Foundation of Korea (NRF) grant (RS-2026-25468474) for Seong-ho Lee, by NIH (R01LM014401, R01NS131225) for Yanyuan Ma, and by NSF (DMS 1953526, 2122074, 2310942), NIH (R01DC021431) and the American Family Funding Initiative of UW-Madison for Jiwei Zhao. The authors would

like to thank the Action Editor and the three reviewers for their insightful comments which have helped improve the manuscript substantially.

Appendix A.1. Derivation of the EIF for δ_0

The efficient influence function for $\boldsymbol{\theta}$ that satisfies $\mathrm{E}_q\{\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})\} = 0$ is, by Proposition S.3 in the supplement of Lee et al. (2025),

$$\begin{aligned}\phi_{\text{eff}}(r, ry, \mathbf{x}) &= \frac{r}{\pi}\rho(y)\mathbf{A}\left[\mathbf{U}(y, \mathbf{x}, \boldsymbol{\theta}) - \frac{\mathrm{E}_p\{\mathbf{U}(Y, \mathbf{x}, \boldsymbol{\theta})\rho^2(Y) + \mathbf{a}(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}}\right] \\ &\quad + \frac{1-r}{1-\pi}\frac{\mathbf{A}\mathrm{E}_p\{\mathbf{U}(Y, \mathbf{x}, \boldsymbol{\theta})\rho^2(Y) + \mathbf{a}(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}},\end{aligned}$$

where $\mathbf{A} = [\mathrm{E}_q\{\partial\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\}]^{-1}$ and $\mathbf{a}(y)$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})\rho^2(Y) + \mathbf{a}(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{X}\}} \mid y\right] = \mathrm{E}\{\mathbf{U}(y, \mathbf{X}, \boldsymbol{\theta}) \mid y\}.$$

Note that $\delta_0 = \mathrm{E}_q\{K_h(Y - y_0)\}$ satisfies $\mathrm{E}_q\{U(Y, \mathbf{X}, \delta_0)\} = 0$ with $U(y, \mathbf{x}, \delta) = K_h(y - y_0) - \delta$. Then ϕ_{eff} reduces to

$$\begin{aligned}\phi_{\text{eff}}(r, ry, \mathbf{x}) &= -\frac{r}{\pi}\rho(y)\left[K_h(y - y_0) - \delta_0 - \frac{\mathrm{E}_p\{K_h(Y - y_0)\rho^2(Y) - \delta_0\rho^2(Y) + a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}}\right] \\ &\quad - \frac{1-r}{1-\pi}\frac{\mathrm{E}_p\{K_h(Y - y_0)\rho^2(Y) - \delta_0\rho^2(Y) + a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}},\end{aligned}$$

where $a(y)$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{K_h(Y - y_0)\rho^2(Y) - \delta_0\rho^2(Y) + a(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{X}\}} \mid y\right] = K_h(y - y_0) - \delta_0.$$

Now, let $\widetilde{a}(y) = a(y) + K_h(y - y_0)\rho(y) + \delta_0\pi/(1-\pi)\rho(y)$, then ϕ_{eff} further reduces to

$$\begin{aligned}\phi_{\text{eff}}(r, ry, \mathbf{x}) &= -\frac{r}{\pi}\rho(y)\left[K_h(y - y_0) - \frac{\mathrm{E}_p\{\widetilde{a}(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}}\right] \\ &\quad - \frac{1-r}{1-\pi}\left[\frac{\mathrm{E}_p\{\widetilde{a}(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}} - \delta_0\right],\end{aligned}$$

where $\widetilde{a}(y)$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{\widetilde{a}(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{X}\}} \mid y\right] = K_h(y - y_0).$$

■

Appendix A.2. Proof of Lemma 6

We first show that $\mathcal{I}^{*\dagger}$ is invertible at v_h by contradiction. Suppose there are a_1 and a_2 such that $a_1 \neq a_2$ and $\mathcal{I}^{*\dagger}(a_1) = \mathcal{I}^{*\dagger}(a_2) = v_h$. Then by Conditions (A1) and (A2), we have

$$\mathrm{E}_p^\dagger\{a_1(Y)\rho^*(Y) \mid \mathbf{x}\} \neq \mathrm{E}_p^\dagger\{a_2(Y)\rho^*(Y) \mid \mathbf{x}\}.$$

On the other hand, the efficient score under the posited models is

$$\begin{aligned}\phi_{\mathrm{eff}}^{*\dagger}(r, ry, \mathbf{x}) &= -\frac{r}{\pi}\rho^*(y)\left[K_h(y - y_0) - \frac{\mathrm{E}_p^\dagger\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p^\dagger\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}\right] \\ &\quad -\frac{1-r}{1-\pi}\left[\frac{\mathrm{E}_p^\dagger\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p^\dagger\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}} - \delta_0\right],\end{aligned}$$

where $a(y)$ satisfies $\mathcal{I}^{*\dagger}(a) = v_h$. Then letting $a = a_1$ and $a = a_2$ gives two distinct efficient scores, which contradicts the uniqueness of the efficient score. Therefore, $a_h^{*\dagger}$ is a unique solution to $\mathcal{I}^{*\dagger}(a) = v_h$, hence $\mathcal{I}^{*\dagger}$ is invertible at v_h.

Now, we show $\mathcal{I}^{*\dagger}$ is invertible at any v in the range of $\mathcal{I}^{*\dagger}$ by contradiction. Suppose there are a_1 and a_2 such that $a_1 \neq a_2$ and $\mathcal{I}^{*\dagger}(a_1) = \mathcal{I}^{*\dagger}(a_2) = v$. This yields $\mathcal{I}^{*\dagger}(a_h^{*\dagger} + a_1 - a_2) = \mathcal{I}^{*\dagger}(a_h^{*\dagger}) + \mathcal{I}^{*\dagger}(a_1) - \mathcal{I}^{*\dagger}(a_2) = v_h$. Then letting $a = a_h^{*\dagger}$ and $a = a_h^{*\dagger} + a_1 - a_2$ gives two distinct efficient scores, and this again contradicts the uniqueness of the efficient score. Hence, there is a unique solution to $\mathcal{I}^{*\dagger}(a) = v$. Therefore, $\mathcal{I}^{*\dagger}$ is invertible.

Next, we show the existence of the constants C_1, C_2. Since

$$\|\mathcal{I}^{*\dagger}(a)\|_2^2 = \int\left\{\int a(t)u^{*\dagger}(t, y)dt\right\}^2 dy \leqslant \|a\|_2^2 \int \|u^{*\dagger}(\cdot, y)\|_2^2 dy,$$

there exists a constant C_1 such that $0 < C_1 < \infty$ and $\|\mathcal{I}^{*\dagger}(a)\|_2 \leqslant C_1\|a\|_2$ by Conditions (A3) and (A4). In addition, we have $\|\mathcal{I}^{*\dagger-1}(v)\|_2 \leqslant C_2\|v\|_2$ for some constant C_2 such that $0 < C_2 < \infty$ by the bounded inverse theorem, since $\mathcal{I}^{*\dagger}$ is bounded and invertible. ■

Appendix A.3. Proof of Theorem 7

Let us define

$$\begin{aligned}b^*(\mathbf{x}, a, \mathrm{E}_p) &\equiv \frac{\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}, \\ \mathcal{I}_{l,i}^{*\dagger}(a)(y) &\equiv \frac{r_i}{\pi}\widetilde{K}_l(y - y_i)b^*(\mathbf{x}_i, a, \mathrm{E}_p^\dagger), \\ \widehat{\mathcal{I}}^{*\dagger}(a)(y) &\equiv N^{-1}\sum_{i=1}^N \frac{r_i}{\pi}\widetilde{K}_l(y - y_i)b^*(\mathbf{x}_i, a, \mathrm{E}_p^\dagger), \\ v_{h,l,i}(y) &\equiv \frac{r_i}{\pi}\widetilde{K}_l(y - y_i)K_h(y - y_0), \\ \widehat{v}(y) &\equiv N^{-1}\sum_{i=1}^N v_{h,l,i}(y), \\ d_{h,l,i}(y) &\equiv \mathcal{I}^{*\dagger-1}\{v_{h,l,i} - \mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger})\}(y).\end{aligned}$$

Note that

$$\mathrm{E}\{b^*(\mathbf{X}, a_h^{*\dagger}, \mathrm{E}_p^\dagger) \mid y\} = \frac{\mathcal{I}^{*\dagger}(a_h^{*\dagger})(y)}{p_Y(y)} = \frac{v_h(y)}{p_Y(y)} = K_h(y - y_0). \tag{A.1}$$

First, we derive the norms of some functions. Since $\mathcal{I}^{*\dagger-1}$ is linear and bounded by Lemma 6 and $\|v_h\|_2 = O(h^{-1/2})$,

$$\|a_h^{*\dagger}\|_2 = \|\mathcal{I}^{*\dagger-1}(v_h)\|_2 = O(h^{-1/2}) \tag{A.2}$$

by Lemma 6. In addition, with $V_{h,l,i}(y) \equiv \frac{R_i}{\pi}\widetilde{K}_l(y - Y_i)K_h(y - y_0)$,

$$\begin{aligned}
\|\mathrm{E}\{\mathcal{I}^{*\dagger-1}(V_{h,l,i})\} - \mathcal{I}^{*\dagger-1}(v_h)\|_2 &= \|\mathcal{I}^{*\dagger-1}\{\mathrm{E}(V_{h,l,i})\} - \mathcal{I}^{*\dagger-1}(v_h)\|_2 \\
&= \|\mathcal{I}^{*\dagger-1}\{\mathrm{E}(V_{h,l,i}) - v_h\}\|_2 \\
&= O(h^{-1/2}l^m),
\end{aligned}$$

where the last equality holds by Lemma 6 and $\|\mathrm{E}(V_{h,l,i}) - v_h\|_2 = O(h^{-1/2}l^m)$ under Conditions (A4) and (A5). Also,

$$\begin{aligned}
& \|\mathrm{E}\{(\mathcal{I}^{*\dagger-1} \circ \mathcal{I}_{l,i}^{*\dagger})(a_h^{*\dagger})\} - a_h^{*\dagger}\|_2 \\
= \; & \|\mathrm{E}\{(\mathcal{I}^{*\dagger-1} \circ \mathcal{I}_{l,i}^{*\dagger})(a_h^{*\dagger})\} - \mathcal{I}^{*\dagger-1}(v_h)\|_2 \\
= \; & \|\mathcal{I}^{*\dagger-1}[\mathrm{E}\{\mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger})\} - v_h]\|_2 \\
= \; & O\left[\left\|\mathrm{E}\left\{\frac{R}{\pi}\widetilde{K}_l(y - Y)b^*(\mathbf{X}, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\right\} - v_h\right\|_2\right] \\
= \; & O\left[\|p_Y(y)\mathrm{E}\{b^*(\mathbf{X}, a_h^{*\dagger}, \mathrm{E}_p^\dagger) \mid y\} + O(h^{-1/2}l^m) - v_h\|_2\right] \\
= \; & O(h^{-1/2}l^m),
\end{aligned}$$

where the third equality holds by Lemma 6, the fourth equality holds by (A.2), Conditions (A4), and (A5), and the last equality holds by (A.1). Then,

$$\begin{aligned}
\|\mathrm{E}(D_{h,l,i})\|_2 &= \|\mathrm{E}\{\mathcal{I}^{*\dagger-1}(V_{h,l,i})\} - \mathrm{E}\{(\mathcal{I}^{*\dagger-1} \circ \mathcal{I}_{l,i}^{*\dagger})(a_h^{*\dagger})\} - \{\mathcal{I}^{*\dagger-1}(v_h) - a_h^{*\dagger}\}\|_2 \\
&\leqslant \|\mathrm{E}\{\mathcal{I}^{*\dagger-1}(V_{h,l,i})\} - \mathcal{I}^{*\dagger-1}(v_h)\|_2 + \|\mathrm{E}\{(\mathcal{I}^{*\dagger-1} \circ \mathcal{I}_{l,i})(a_h^{*\dagger})\} - a_h^{*\dagger}\|_2 \\
&= O(h^{-1/2}l^m).
\end{aligned} \tag{A.3}$$

In addition,

$$\begin{aligned}
& \left\|N^{-1}\sum_{i=1}^N d_{h,l,i} - \mathrm{E}(D_{h,l,i})\right\|_2 \\
= \; & \left\|N^{-1}\sum_{i=1}^N \mathcal{I}^{*\dagger-1}\{v_{h,l,i} - \mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger})\} - \mathrm{E}\left[\mathcal{I}^{*\dagger-1}\{V_{h,l,i} - \mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger})\}\right]\right\|_2 \\
= \; & \left\|\mathcal{I}^{*\dagger-1}\left[N^{-1}\sum_{i=1}^N v_{h,l,i} - \mathrm{E}(V_{h,l,i}) - N^{-1}\sum_{i=1}^N \mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger}) + \mathrm{E}\{\mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger})\}\right]\right\|_2 \\
= \; & O\left[\left\|N^{-1}\sum_{i=1}^N v_{h,l,i} - \mathrm{E}(V_{h,l,i})\right\|_2 + \left\|N^{-1}\sum_{i=1}^N \mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger}) - \mathrm{E}\{\mathcal{I}_{l,i}^{*\dagger}(a_h^{*\dagger})\}\right\|_2\right] \\
= \; & O_p\{h^{-1/2}(nl)^{-1/2}\},
\end{aligned} \tag{A.4}$$

where the second equality holds since $\mathcal{I}^{*\dagger-1}$ is linear by Lemma 6, the third equality holds since $\mathcal{I}^{*\dagger-1}$ is bounded by Lemma 6, and the last equality holds by (A.2) and Condition (A5). On the other hand, for any function $a(y)$,

$$\begin{aligned}\|(\widehat{\mathcal{I}}^{*\dagger}-\mathcal{I}^{*\dagger})(a)\|_2 &= \left\| N^{-1}\sum_{i=1}^{N}\frac{r_i}{\pi}\widetilde{K}_l(y-y_i)b^*(\mathbf{x}_i,a,\mathrm{E}_p^\dagger) - p_Y(y)\mathrm{E}\{b^*(\mathbf{x},a,\mathrm{E}_p^\dagger)\mid y\}\right\|_2 \\ &= O_p\left[\{(nl)^{-1/2}+l^m\}\|a\|_2\right] = o_p(n^{-1/4}\|a\|_2) \end{aligned} \tag{A.5}$$

under Conditions (A3)-(A6). Similarly,

$$\|\widehat{v}-v_h\|_2 = O_p\left\{(nl)^{-1/2}+l^m\right\}O(h^{-1/2}) = o_p(n^{-1/4}h^{-1/2}). \tag{A.6}$$

Now, we derive the asymptotic form of $b^*(\mathbf{x},\widehat{a}^{*\dagger},\mathrm{E}_p^\dagger)$. $\widehat{a}^{*\dagger}$ can be written as, since $\mathcal{I}^{*\dagger-1}$ is linear and bounded by Lemma 6,

$$\begin{aligned}\widehat{a}^{*\dagger}(y) &= \widehat{\mathcal{I}}^{*\dagger-1}(\widehat{v})(y) \\ &= \{\mathcal{I}^{*\dagger}+(\widehat{\mathcal{I}}^{*\dagger}-\mathcal{I}^{*\dagger})\}^{-1}(\widehat{v})(y) \\ &= \{\mathcal{I}^{*\dagger-1}-\mathcal{I}^{*\dagger-1}\circ(\widehat{\mathcal{I}}^{*\dagger}-\mathcal{I}^{*\dagger})\circ\mathcal{I}^{*\dagger-1}\}\{v_h+(\widehat{v}-v_h)\}(y)+o_p(n^{-1/2}h^{-1/2}) \\ &= a_h^{*\dagger}(y)+\mathcal{I}^{*\dagger-1}(\widehat{v}-v_h)(y)-\{\mathcal{I}^{*\dagger-1}\circ(\widehat{\mathcal{I}}^{*\dagger}-\mathcal{I}^{*\dagger})\}(a_h^{*\dagger})(y)+o_p(n^{-1/2}h^{-1/2}) \\ &= a_h^{*\dagger}(y)+\mathcal{I}^{*\dagger-1}\{\widehat{v}-\widehat{\mathcal{I}}^{*\dagger}(a_h^{*\dagger})\}(y)+o_p(n^{-1/2}h^{-1/2}) \\ &= a_h^{*\dagger}(y)+N^{-1}\sum_{i=1}^{N}d_{h,l,i}(y)+o_p(n^{-1/2}h^{-1/2}) \end{aligned} \tag{A.7}$$

uniformly in y by Condition (A4). Here, the third equality holds by (A.5) and

$$\|\widehat{v}\|_2 \leqslant \|v_h\|_2+\|\widehat{v}-v_h\|_2 = O(h^{-1/2})+o_p(n^{-1/4}h^{-1/2}) = O_p(h^{-1/2})$$

by (A.6), and the fourth equality holds by $a_h^{*\dagger}(y) = \mathcal{I}^{*\dagger-1}(v_h)(y)$, (A.5), and (A.6). Then,

$$\|\widehat{a}^{*\dagger}-a_h^{*\dagger}\|_2 = O(h^{-1/2}l^m)+O_p\{h^{-1/2}(nl)^{-1/2}\}+o_p(n^{-1/2}h^{-1/2}) = o_p(n^{-1/4}h^{-1/2}) \tag{A.8}$$

by (A.3), (A.4), (A.7), and Condition (A6), and

$$\|\widehat{a}^{*\dagger}\|_2 \leqslant \|a_h^{*\dagger}\|_2+\|\widehat{a}^{*\dagger}-a_h^{*\dagger}\|_2 = O_p(h^{-1/2}) \tag{A.9}$$

by (A.2). Hence,

$$\begin{aligned} & b^*(\mathbf{x},\widehat{a}^{*\dagger},\mathrm{E}_p^\dagger) \\ = \; & b^*(\mathbf{x},a_h^{*\dagger},\mathrm{E}_p^\dagger)+b^*(\mathbf{x},\widehat{a}^{*\dagger}-a_h^{*\dagger},\mathrm{E}_p^\dagger) \\ = \; & b^*(\mathbf{x},a_h^{*\dagger},\mathrm{E}_p^\dagger)+N^{-1}\sum_{i=1}^{N}b^*(\mathbf{x},d_{h,l,i},\mathrm{E}_p^\dagger)+o_p(n^{-1/2}h^{-1/2}) \end{aligned} \tag{A.10}$$

uniformly in $\mathbf{x}$ by Condition (A4), where the last equality holds by (A.7).

Now, we analyze $\widetilde{\delta}_0$. From the definition of $\widetilde{\delta}_0$,

$$\begin{aligned}
\widetilde{\delta}_0 - \delta_0 &= N^{-1}\sum_{i=1}^{N}\left[\frac{r_i}{\pi}\rho^*(y_i)\{K_h(y_i-y_0) - b^*(\mathbf{x}_i,\widehat{a}^{*\dagger},\mathrm{E}_p^\dagger)\} + \frac{1-r_i}{1-\pi}\{b^*(\mathbf{x}_i,\widehat{a}^{*\dagger},\mathrm{E}_p^\dagger)-\delta_0\}\right] \\
&= -N^{-1}\sum_{i=1}^{N}\phi_{\mathrm{eff}}^{*\dagger}(r_i, r_i y_i, \mathbf{x}_i) \\
&\quad -N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\{b^*(\mathbf{x}_i,\widehat{a}^{*\dagger},\mathrm{E}_p^\dagger) - b^*(\mathbf{x}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)\} \\
&= -N^{-1}\sum_{i=1}^{N}\phi_{\mathrm{eff}}^{*\dagger}(r_i, r_i y_i, \mathbf{x}_i) - T + o_p(n^{-1/2}h^{-1/2}), \qquad \text{(A.11)}
\end{aligned}$$

where

$$T \equiv N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\} b^*(\mathbf{x}_i, d_{h,l,j}, \mathrm{E}_p^\dagger).$$

By the property of the U-statistic, T can be written as

$$\begin{aligned}
& T \qquad \text{(A.12)} \\
&= N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\mathrm{E}\{b^*(\mathbf{x}_i, D_{h,l,j}, \mathrm{E}_p^\dagger) \mid r_i, r_i y_i, \mathbf{x}_i\} \\
&\quad +N^{-1}\sum_{j=1}^{N}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i) - \frac{1-R_i}{1-\pi}\right\} b^*(\mathbf{X}_i, d_{h,l,j}, \mathrm{E}_p^\dagger) \mid r_j, r_j y_j, \mathbf{x}_j\right] \\
&\quad -N^{-1/2}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i) - \frac{1-R_i}{1-\pi}\right\} b^*(\mathbf{X}_i, D_{h,l,j}, \mathrm{E}_p^\dagger)\right] + O_p(N^{-1}h^{-1/2}) \\
&= N^{-1}\sum_{j=1}^{N}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i) - \frac{1-R_i}{1-\pi}\right\} b^*(\mathbf{X}_i, d_{h,l,j}, \mathrm{E}_p^\dagger) \mid r_j, r_j y_j, \mathbf{x}_j\right] \\
&\quad +O_p(h^{-1/2}l^m + N^{-1}h^{-1/2}) \\
&= N^{-1}\sum_{j=1}^{N}\int\left[\rho^*(y)\mathrm{E}\{b^*(\mathbf{X}, d_{h,l,j}, \mathrm{E}_p^\dagger) \mid y\}p_Y(y) - \mathrm{E}\{b^*(\mathbf{X}, d_{h,l,j}, \mathrm{E}_p^\dagger) \mid y\}q_Y(y)\right]dy \\
&\quad +o_p(n^{-1/2}h^{-1/2}) \\
&= N^{-1}\sum_{j=1}^{N}\int\{\rho^*(y) - \rho(y)\}\,\mathcal{I}^{*\dagger}(d_{h,l,j})(y)dy + o_p(n^{-1/2}h^{-1/2}) \\
&= N^{-1}\sum_{j=1}^{N}\int\{\rho^*(y) - \rho(y)\}\,\{v_{h,l,j}(y) - \mathcal{I}_{l,j}^{*\dagger}(a_h^{*\dagger})(y)\}dy + o_p(n^{-1/2}h^{-1/2}) \\
&= N^{-1}\sum_{j=1}^{N}\frac{r_j}{\pi}\int\widetilde{K}_l(y-y_j)\,\{\rho^*(y) - \rho(y)\}\,\{K_h(y-y_0) - b^*(\mathbf{x}_j, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\}dy + o_p(n^{-1/2}h^{-1/2}),
\end{aligned}$$

where the second equality holds because (A.3) leads to

$$\mathrm{E}\{b^*(\mathbf{x}_i, D_{h,l,j}, \mathrm{E}_p^\dagger) \mid r_i, r_i y_i, \mathbf{x}_i\} = O(h^{-1/2} l^m),$$

and the third equality holds by Condition (A6). Further,

$$\begin{aligned}
&\int \widetilde{K}_l(y - y_j) \{\rho^*(y) - \rho(y)\} \{K_h(y - y_0) - b^*(\mathbf{x}_j, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\} dy \\
= &\int \widetilde{K}(t) \{\rho^*(y_j + lt) - \rho(y_j + lt)\} \{K_h(y_j + lt - y_0) - b^*(\mathbf{x}_j, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\} dt \\
= &\{\rho^*(y_j) - \rho(y_j)\} \{K_h(y_j - y_0) - b^*(\mathbf{x}_j, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\} + O(h^{-m-1} l^m) \\
= &\{\rho^*(y_j) - \rho(y_j)\} \{K_h(y_j - y_0) - b^*(\mathbf{x}_j, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\} + o(n^{-1/2} h^{-1/2}) \qquad \text{(A.13)}
\end{aligned}$$

under Conditions (A1), (A5), and (A6). Therefore, (A.11), (A.12), and (A.13) lead to

$$\begin{aligned}
&\widetilde{\delta}_0 - \delta_0 \\
= &-N^{-1} \sum_{i=1}^N \phi_{\mathrm{eff}}^{*\dagger}(r_i, r_i y_i, \mathbf{x}_i) \\
&-N^{-1} \sum_{i=1}^N \frac{r_i}{\pi} \{\rho^*(y_i) - \rho(y_i)\} \{K_h(y_i - y_0) - b^*(\mathbf{x}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\} + o_p(n^{-1/2} h^{-1/2}) \\
= &N^{-1} \sum_{i=1}^N \frac{r_i}{\pi} \rho(y_i) \left\{K_h(y_i - y_0) - b^*(\mathbf{x}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\right\} \\
&+N^{-1} \sum_{i=1}^N \frac{1 - r_i}{1 - \pi} \left\{b^*(\mathbf{x}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger) - \delta_0\right\} + o_p(n^{-1/2} h^{-1/2}). \qquad \text{(A.14)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathrm{var}(\widetilde{\delta}_0 - \delta_0) = &N^{-1} \mathrm{E}\left(\left[\frac{R_i}{\pi} \rho(Y_i) \left\{K_h(Y_i - y_0) - b^*(\mathbf{X}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\right\}\right]^2\right) \\
&+N^{-1} \mathrm{E}\left(\left[\frac{1 - R_i}{1 - \pi} \left\{b^*(\mathbf{X}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger) - \delta_0\right\}\right]^2\right) \\
&+o[\{\min(n, N - n)\}^{-1/2} n^{-1/2} h^{-1}] \\
= &n^{-1} \mathrm{E}\left(\frac{R_i}{\pi} \left[\rho(Y_i) \left\{K_h(Y_i - y_0) - b^*(\mathbf{X}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\right\}\right]^2\right) \\
&+(N - n)^{-1} \mathrm{E}\left[\frac{1 - R_i}{1 - \pi} \left\{b^*(\mathbf{X}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger) - \delta_0\right\}^2\right] \\
&+o[\{\min(n, N - n)\}^{-1/2} n^{-1/2} h^{-1}] \\
= &n^{-1} \mathrm{E}_p\left(\left[\rho(Y_i) \left\{K_h(Y_i - y_0) - b^*(\mathbf{X}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger)\right\}\right]^2\right) \\
&+(N - n)^{-1} \mathrm{E}_q\left[\left\{b^*(\mathbf{X}_i, a_h^{*\dagger}, \mathrm{E}_p^\dagger) - \delta_0\right\}^2\right] \\
&+o[\{\min(n, N - n)\}^{-1/2} n^{-1/2} h^{-1}].
\end{aligned}$$

In addition,

$$\begin{aligned}\delta_0 &= \int K_h(y-y_0)q_Y(y)dy \\ &= \int K(t)q_Y(y_0+ht)dt \\ &= q_Y(y_0)+q_Y^{(m)}(y_0)\frac{\int t^m K(t)dt}{m!}h^m+O(h^{m+1})\end{aligned}$$

under Conditions (A4), (A5), and (A6). ■

Appendix A.4. Proof of Theorem 8

Let

$$z_i \equiv \frac{r_i}{\pi}\rho(y_i)\left\{K_h(y_i-y_0)-b^*(\mathbf{x}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)\right\}+\frac{1-r_i}{1-\pi}\left\{b^*(\mathbf{x}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)-\delta_0\right\}.$$

We have

$$\begin{aligned}|z_i| &= O\left\{\frac{N}{n}(h^{-1}+h^{-1/2})+\frac{N}{N-n}(h^{-1/2}+1)\right\} \\ &\leqslant C_1\left\{\frac{N}{n}h^{-1}+\frac{N}{\min(n,N-n)}h^{-1/2}\right\} \qquad \text{(A.15)}\end{aligned}$$

for some constant $C_1 > 0$ uniformly for i, because $K_h(y_i-y_0) = O(h^{-1})$ by Condition (A5), $b^*(\mathbf{x}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger) = O(h^{-1/2})$ since ρ^* is bounded above and bounded away from zero by Condition (A1) and $\|a_h^{*\dagger}\|_2 = O(h^{-1/2})$ by (A.2), $\delta_0 = \mathrm{E}_q\{K_h(Y-y_0)\} = O(1)$, and $h = o(1)$ by Condition (A6). In addition,

$$\begin{aligned}&\mathrm{var}(Z_i) \\ \leqslant\ & \frac{1}{\pi}\mathrm{E}_p\left(\left[\rho(Y_i)\left\{K_h(Y_i-y_0)-b^*(\mathbf{X}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)\right\}\right]^2\right)+\frac{1}{1-\pi}\mathrm{E}_q\left[\left\{b^*(\mathbf{X}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)-\delta_0\right\}^2\right] \\ \leqslant\ & \frac{2}{\pi}\mathrm{E}_p\left\{\rho^2(Y_i)K_h^2(Y_i-y_0)\right\}+\frac{2}{\pi}\mathrm{E}_p\left\{\rho^2(Y_i)b^{*2}(\mathbf{X}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)\right\} \\ &+\frac{2}{1-\pi}\mathrm{E}_q\left\{b^{*2}(\mathbf{X}_i,a_h^{*\dagger},\mathrm{E}_p^\dagger)\right\}+\frac{2\delta_0^2}{1-\pi} \\ =\ & O\left\{\frac{N}{n}(h^{-1}+h^{-1})+\frac{N}{N-n}(h^{-1}+1)\right\} \\ \leqslant\ & C_2\frac{N}{\min(n,N-n)}h^{-1} \qquad \text{(A.16)}\end{aligned}$$

for some constant $C_2 > 0$ uniformly for i, because

$$\begin{aligned}\mathrm{E}_p\left\{\rho^2(Y_i)K_h^2(Y_i-y_0)\right\} &= \int \rho^2(y)K_h^2(y-y_0)p_Y(y)dy \\ &= h^{-1}\int \rho^2(y_0+ht)K^2(t)p_Y(y_0+ht)dt \\ &= h^{-1}\rho^2(y_0)p_Y(y_0)\int K^2(t)dt+O(1) \\ &= O(h^{-1})\end{aligned}$$

by Conditions (A4) and (A5), and $\sup_{\mathbf{x}} |b^*(\mathbf{x}, a_h^{*\dagger}, \mathrm{E}_p^\dagger)| = O(h^{-1/2})$ by Conditions (A1), (A4), and (A.2).

Now, let $t_1, \ldots, t_{M_n}$ be arbitrary points on the support of $p_Y(y)$, and M_n is either finite or diverging to infinity as $n \to \infty$, and $\widetilde{q}_Y(t_k) = \widetilde{\delta}_0$ at $y_0 = t_k$. Then for $\epsilon = C\{\min(n, N-n)h\}^{-1/2} \log M_n$ with some constant $C > 0$, (A.14) leads to

$$\begin{aligned}
& \Pr\left(\max_{k=1,\ldots,M_n} |\widetilde{q}_Y(t_k) - \mathrm{E}_q\{K_h(Y - t_k)\}| \geqslant \epsilon \right) \\
\leqslant\ & M_n \Pr(|\widetilde{\delta}_0 - \delta_0| \geqslant \epsilon) \\
=\ & M_n \Pr\left\{ \left| N^{-1} \sum_{i=1}^N Z_i + o_p(n^{-1/2}h^{-1/2}) \right| \geqslant \epsilon \right\} \\
\leqslant\ & M_n \Pr\left\{ \left| N^{-1} \sum_{i=1}^N Z_i \right| \geqslant \epsilon + \left| o_p(n^{-1/2}h^{-1/2}) \right| \right\} \\
\leqslant\ & M_n \Pr\left(\left| N^{-1} \sum_{i=1}^N Z_i \right| \geqslant 2\epsilon \right) \\
\leqslant\ & 2M_n \exp\left(-\frac{2\epsilon^2}{2B\epsilon/3 + V} \right)
\end{aligned}$$

by the Bernstein inequality where

$$\begin{aligned}
B &= C_1[n^{-1}h^{-1} + \{\min(n, N-n)\}^{-1}h^{-1/2}], \\
V &= C_2\{\min(n, N-n)^{-1}h^{-1}\},
\end{aligned}$$

by (A.15) and (A.16). Also, $\mathrm{E}_q\{K_h(Y - t_k)\} = q_Y(t_k) + O(h^m)$ by Conditions (A4) and (A5). These imply

$$\max_{k=1,\ldots,M_n} |\widetilde{q}_Y(t_k) - q_Y(t_k)| = O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2} \log M_n].$$

On the other hand,

$$\max_{k=1,\ldots,M_n} |\widehat{p}_Y(t_k) - p_Y(t_k)| = O(h^m) + O_p\{(nh)^{-1/2} \log M_n\}$$

under Conditions (A4)-(A6). Therefore,

$$\begin{aligned}
\max_{k=1,\ldots,M_n} |\widetilde{\rho}(t_k) - \rho(t_k)| &= \max_{k=1,\ldots,M_n} \left| \frac{\widetilde{q}_Y(t_k) - q_Y(t_k)}{\widehat{p}_Y(t_k)} + q_Y(t_k) \left\{ \frac{1}{\widehat{p}_Y(t_k)} - \frac{1}{p_Y(t_k)} \right\} \right| \\
&= O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2} \log M_n + (nh)^{-1/2} \log M_n] \\
&= O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2} \log M_n],
\end{aligned}$$

since p_Y and q_Y are bounded and bounded away from zero by Condition (A4).

Now, let $t_1 < \cdots < t_{M(\delta_n)}$ be equally-spaced points on the support of $\rho(y)$, say $[a, b]$ under Condition (A4), with grid spacing distance $\delta_n = C_\delta n^{m(1-m)/(2m+1)}$ for some constant $C_\delta > 0$. Let $y \in [a, b]$ and t_k be the point closest to y among $\{t_1, \ldots, t_{M(\delta_n)}\}$, then

considering that $\widetilde{\rho}(\cdot)$ is a piecewise linear interpolation of $\widetilde{\rho}(t_1), \ldots, \widetilde{\rho}(t_{M(\delta_n)})$, we get

$$\begin{aligned}
& |\widetilde{\rho}(y) - \rho(y)| \\
\leqslant\ & |\widetilde{\rho}(y) - \widetilde{\rho}(t_k)| + |\widetilde{\rho}(t_k) - \rho(t_k)| + |\rho(t_k) - \rho(y)| \\
\leqslant\ & |\widetilde{\rho}(t_{k+1}) - \widetilde{\rho}(t_k)| + |\widetilde{\rho}(t_k) - \rho(t_k)| + |\rho(t_k) - \rho(y)| \\
\leqslant\ & |\widetilde{\rho}(t_{k+1}) - \rho(t_{k+1})| + |\rho(t_{k+1}) - \rho(t_k)| + |\rho(t_k) - \widetilde{\rho}(t_k)| + |\widetilde{\rho}(t_k) - \rho(t_k)| + |\rho(t_k) - \rho(y)| \\
\leqslant\ & 3 \max_{k=1,\ldots,M(\delta_n)} |\widetilde{\rho}(t_k) - \rho(t_k)| + 2L\delta_n,
\end{aligned}$$

where $L > 0$ is the Lipschitz constant of $\rho(\cdot)$. Note that

$$n^{m(1-m)/(2m+1)} = o\{(nl^{2m})^{m/(2m+1)}\} = o(h^m)$$

under Condition (A6). Also note that $M(\delta_n) \asymp n^{m(m-1)/(2m+1)}$. Therefore,

$$\begin{aligned}
\|\widetilde{\rho}(y) - \rho(y)\|_\infty &= \sup_{y \in [a,b]} |\widetilde{\rho}(y) - \rho(y)| \\
&\leqslant 3 \max_{k=1,\ldots,M(\delta_n)} |\widetilde{\rho}(t_k) - \rho(t_k)| + 2L\delta_n \\
&= O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2} \log\{M(\delta_n)\}] + o(h^m) \\
&= O(h^m) + O_p[\{\min(n, N-n)h\}^{-1/2} \log n].
\end{aligned}$$

■

Appendix A.5. Proof of Theorem 9

Let us define

$$\begin{aligned}
\widetilde{b}(\mathbf{x}, a, \mathrm{E}_p) &\equiv \frac{\mathrm{E}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}, \\
\widetilde{\mathcal{I}}_{l,i}(a)(y) &\equiv \frac{r_i}{\pi} \widetilde{K}_l(y - y_i) \widetilde{b}(\mathbf{x}_i, a, \mathrm{E}_p), \\
\widehat{\mathcal{I}}(a)(y) &\equiv N^{-1} \sum_{i=1}^{N} \frac{r_i}{\pi} \widetilde{K}_l(y - y_i) \widetilde{b}(\mathbf{x}_i, a, \widehat{\mathrm{E}}_p), \\
d_{h,l,i}(y) &\equiv \widetilde{\mathcal{I}}^{-1}\{v_{h,l,i} - \widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)\}(y).
\end{aligned}$$

Note that

$$\mathrm{E}\{\widetilde{b}(\mathbf{X}, \widetilde{a}_h, \mathrm{E}_p) \mid y\} = \frac{\widetilde{\mathcal{I}}(\widetilde{a}_h)(y)}{p_Y(y)} = \frac{v_h(y)}{p_Y(y)} = K_h(y - y_0). \tag{A.17}$$

We also define the kth Gateaux derivative of $g(\cdot, \mu)$ with respect to μ at μ_1 in the direction μ_2 as

$$\frac{\partial^k g(\cdot, \mu_1)}{\partial \mu^k}(\mu_2) \equiv \frac{\partial^k g(\cdot, \mu)}{\partial \mu^k}(\mu_2)\Big|_{\mu=\mu_1} \equiv \frac{\partial^k g(\cdot, \mu_1 + h\mu_2)}{\partial h^k}\Big|_{h=0}.$$

Then,

$$\begin{aligned}
&\frac{\partial \widetilde{b}(\mathbf{x}, a, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \qquad (A.18)\\
= \;& \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}\\
& - \frac{\mathrm{E}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}]^2}\\
= \;& \widetilde{b}(\mathbf{x}, a, \mathrm{E}_p)\\
& \times \left[\frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}} - \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} \right],\\
&\frac{\partial^2 \widetilde{b}(\mathbf{x}, a, \widetilde{\mathrm{E}}_p)}{\partial {\mathrm{E}_p}^2}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \qquad (A.19)\\
= \;& \frac{-2(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{[\widetilde{\mathrm{E}}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}]^3}\\
& \times \Big[(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}\widetilde{\mathrm{E}}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}\\
& - \widetilde{\mathrm{E}}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\} \Big].
\end{aligned}$$

First, we derive the norms of some functions. We have

$$\|\widetilde{a}_h\|_2 = \|\widetilde{\mathcal{I}}^{-1}(v_h)\|_2 = O(h^{-1/2}), \qquad (A.20)$$

because $\widetilde{\mathcal{I}}^{-1}$ is bounded by Lemma 6 and $\|v_h\|_2 = O(h^{-1/2})$. In addition,

$$\|\mathrm{E}\{\widetilde{\mathcal{I}}^{-1}(V_{h,l,i})\} - \widetilde{\mathcal{I}}^{-1}(v_h)\|_2 = \|\widetilde{\mathcal{I}}^{-1}\{\mathrm{E}(V_{h,l,i}) - v_h\}\|_2 = O(h^{-1/2}l^m),$$

where the last equality holds by Lemma 6 and $\|\mathrm{E}(V_{h,l,i}) - v_h\|_2 = O(h^{-1/2}l^m)$ under Conditions (A4) and (A5). Also,

$$\begin{aligned}
\|\mathrm{E}\{(\widetilde{\mathcal{I}}^{-1} \circ \widetilde{\mathcal{I}}_{l,i})(\widetilde{a}_h)\} - \widetilde{a}_h\|_2 &= \|\mathrm{E}\{(\widetilde{\mathcal{I}}^{-1} \circ \widetilde{\mathcal{I}}_{l,i})(\widetilde{a}_h)\} - \widetilde{\mathcal{I}}^{-1}(v_h)\|_2\\
&= \|\widetilde{\mathcal{I}}^{-1}[\mathrm{E}\{\widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)\} - v_h]\|_2\\
&= O\left[\left\| \mathrm{E}\left\{ \frac{R}{\pi} \widetilde{K}_l(y - Y)\widetilde{b}(\mathbf{X}, \widetilde{a}_h, \mathrm{E}_p) \right\} - v_h \right\|_2 \right]\\
&= O\Big[\|p_Y(y)\mathrm{E}\{\widetilde{b}(\mathbf{X}, \widetilde{a}_h, \mathrm{E}_p) \mid y\} + O(h^{-1/2}l^m) - v_h\|_2 \Big]\\
&= O(h^{-1/2}l^m),
\end{aligned}$$

where the third equality holds by Lemma 6, the fourth equality holds by (A.20), Conditions (A4), and (A5), and the last equality holds by (A.17). Then,

$$\begin{aligned}
\|\mathrm{E}(D_{h,l,i})\|_2 &= \|\mathrm{E}\{\widetilde{\mathcal{I}}^{-1}(V_{h,l,i})\} - \mathrm{E}\{(\widetilde{\mathcal{I}}^{-1} \circ \widetilde{\mathcal{I}}_{l,i})(\widetilde{a}_h)\} - \{\widetilde{\mathcal{I}}^{-1}(v_h) - \widetilde{a}_h\}\|_2\\
&\leqslant \|\mathrm{E}\{\widetilde{\mathcal{I}}^{-1}(V_{h,l,i})\} - \widetilde{\mathcal{I}}^{-1}(v_h)\|_2 + \|\mathrm{E}\{(\widetilde{\mathcal{I}}^{-1} \circ \widetilde{\mathcal{I}}_{l,i})(\widetilde{a}_h)\} - \widetilde{a}_h\|_2\\
&= O(h^{-1/2}l^m). \qquad (A.21)
\end{aligned}$$

In addition,

$$\begin{aligned}
& \left\| N^{-1} \sum_{i=1}^{N} d_{h,l,i} - \mathrm{E}(D_{h,l,i}) \right\|_2 \\
= & \left\| N^{-1} \sum_{i=1}^{N} \widetilde{\mathcal{I}}^{-1}\{v_{h,l,i} - \widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)\} - \mathrm{E}\left[\widetilde{\mathcal{I}}^{-1}\{V_{h,l,i} - \widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)\}\right] \right\|_2 \\
= & \left\| \widetilde{\mathcal{I}}^{-1} \left[N^{-1} \sum_{i=1}^{N} v_{h,l,i} - \mathrm{E}(V_{h,l,i}) - N^{-1} \sum_{i=1}^{N} \widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h) + \mathrm{E}\{\widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)\} \right] \right\|_2 \\
= & O\left[\left\| N^{-1} \sum_{i=1}^{N} v_{h,l,i} - \mathrm{E}(V_{h,l,i}) \right\|_2 + \left\| N^{-1} \sum_{i=1}^{N} \widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h) - \mathrm{E}\{\widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)\} \right\|_2 \right] \\
= & O_p\{h^{-1/2}(nl)^{-1/2}\}, \qquad \text{(A.22)}
\end{aligned}$$

where the second equality holds since $\widetilde{\mathcal{I}}^{-1}$ is linear by Lemma 6, the third equality holds since $\mathcal{I}^{-1}$ is bounded by Lemma 6, and the last equality holds by (A.20) and Condition (A5). On the other hand, for $a(y)$ such that $\|a\|_2 < \infty$, (A.18) leads to

$$\begin{aligned}
& \left| \frac{\partial \widetilde{b}(\mathbf{x}, a, \mathrm{E}_p)}{\partial \mathrm{E}_p} (\widehat{\mathrm{E}}_p - \mathrm{E}_p) \right| \\
= & \left| \widetilde{b}(\mathbf{x}, a, \mathrm{E}_p) \right| \\
& \times \left| \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}} - \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} \right| \\
\leqslant & \frac{\|a\|_2 \|\widetilde{\rho}(\cdot) p_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_2}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} \\
& \times \left| \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{a(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}} - \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} \right| \\
= & o_p(n^{-1/4}\|a\|_2) \qquad \text{(A.23)}
\end{aligned}$$

uniformly in $\mathbf{x}$ by Condition (A4), because $\widetilde{\rho}(y)$ is bounded away from zero by Condition (A1), $\widetilde{\rho}(y)$ is bounded above since its derivative is bounded on a compact support by Conditions (A1) and (A4), and $\widetilde{\rho}(\cdot) p^{\dagger}_{Y|\mathbf{X}}(\cdot, \mathbf{x}, \boldsymbol{\beta})$ is square integrable by Condition (A2), and $|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y) \mid \mathbf{x}\}| = o_p(n^{-1/4}\|a\|_2)$. Similarly, for $\widetilde{\mathrm{E}}_p = \mathrm{E}_p + \alpha(\widehat{\mathrm{E}}_p - \mathrm{E}_p)$ with $\alpha \in (0, 1)$, (A.19) leads to

$$\left| \frac{\partial^2 \widetilde{b}(\mathbf{x}, a, \widetilde{\mathrm{E}}_p)}{\partial {\mathrm{E}_p}^2} (\widehat{\mathrm{E}}_p - \mathrm{E}_p) \right| = o_p(n^{-1/2}\|a\|_2)$$

uniformly in $\mathbf{x}$. Thus,

$$\begin{aligned}\widetilde{b}(\mathbf{x},a,\widehat{\mathrm{E}}_p)-\widetilde{b}(\mathbf{x},a,\mathrm{E}_p) &= \frac{\partial\widetilde{b}(\mathbf{x},a,\mathrm{E}_p)}{\partial\mathrm{E}_p}(\widehat{\mathrm{E}}_p-\mathrm{E}_p)+\frac{1}{2}\frac{\partial^2\widetilde{b}(\mathbf{x},a,\widetilde{\mathrm{E}}_p)}{\partial{\mathrm{E}_p}^2}(\widehat{\mathrm{E}}_p-\mathrm{E}_p)\\ &= \frac{\partial\widetilde{b}(\mathbf{x},a,\mathrm{E}_p)}{\partial\mathrm{E}_p}(\widehat{\mathrm{E}}_p-\mathrm{E}_p)+o_p(n^{-1/2}\|a\|_2) \qquad (A.24)\\ &= o_p(n^{-1/4}\|a\|_2)\end{aligned}$$

uniformly in $\mathbf{x}$. Hence,

$$\begin{aligned}\|(\widehat{\mathcal{I}}-\widetilde{\mathcal{I}})(a)\|_2 &= \left\|N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\widetilde{K}_l(y-y_i)\widetilde{b}(\mathbf{x}_i,a,\widehat{\mathrm{E}}_p)-p_Y(y)\mathrm{E}\{\widetilde{b}(\mathbf{x},a,\mathrm{E}_p)\mid y\}\right\|_2\\ &\leqslant \left\|N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\widetilde{K}_l(y-y_i)\left\{\widetilde{b}(\mathbf{x}_i,a,\widehat{\mathrm{E}}_p)-\widetilde{b}(\mathbf{x}_i,a,\mathrm{E}_p)\right\}\right\|_2\\ &\quad+\left\|N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\widetilde{K}_l(y-y_i)\widetilde{b}(\mathbf{x}_i,a,\mathrm{E}_p)-p_Y(y)\mathrm{E}\{\widetilde{b}(\mathbf{x},a,\mathrm{E}_p)\mid y\}\right\|_2\\ &= o_p(n^{-1/4}\|a\|_2)+O_p\left[\{(nl)^{-1/2}+l^m\}\|a\|_2\right]=o_p(n^{-1/4}\|a\|_2) \qquad (A.25)\end{aligned}$$

under Conditions (A3)-(A6). Similarly,

$$\|\widehat{v}-v_h\|_2=O_p\left\{(nl)^{-1/2}+l^m\right\}O(h^{-1/2})=o_p(n^{-1/4}h^{-1/2}). \qquad (A.26)$$

Furthermore,

$$\begin{aligned}&\widehat{\mathcal{I}}(\widetilde{a}_h)(y)\\ =\;& N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\widetilde{K}_l(y-y_i)\left\{\widetilde{b}(\mathbf{x}_i,\widetilde{a}_h,\mathrm{E}_p)+\frac{\partial\widetilde{b}(\mathbf{x}_i,\widetilde{a}_h,\mathrm{E}_p)}{\partial\mathrm{E}_p}(\widehat{\mathrm{E}}_p-\mathrm{E}_p)+o_p(n^{-1/2}h^{-1/2})\right\}\\ =\;& N^{-1}\sum_{i=1}^N\widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)(y)+p_Y(y)\frac{\partial\mathrm{E}\{\widetilde{b}(\mathbf{X},\widetilde{a}_h,\mathrm{E}_p)\mid y\}}{\partial\mathrm{E}_p}(\widehat{\mathrm{E}}_p-\mathrm{E}_p)\\ &+o_p[\{(nl)^{-1/2}+l^m\}n^{-1/4}h^{-1/2}]+o_p(n^{-1/2}h^{-1/2})\\ =\;& N^{-1}\sum_{i=1}^N\widetilde{\mathcal{I}}_{l,i}(\widetilde{a}_h)(y)+o_p(n^{-1/2}h^{-1/2}), \qquad (A.27)\end{aligned}$$

where the first equality holds by (A.20) and (A.24), the second equality holds by (A.20), (A.23), and Conditions (A3)-(A5), and the last equality holds by (A.17) and Condition (A6).

Now, we derive the asymptotic form of $\widetilde{b}(\mathbf{x}, \widehat{a}, \widehat{\mathrm{E}}_p)$. $\widehat{a}$ can be written as, since $\widetilde{\mathcal{I}}^{-1}$ is linear and bounded by Lemma 6,

$$\begin{aligned}
\widehat{a}(y) &= \widehat{\mathcal{I}}^{-1}(\widehat{v})(y) \\
&= \{\widetilde{\mathcal{I}} + (\widehat{\mathcal{I}} - \widetilde{\mathcal{I}})\}^{-1}(\widehat{v})(y) \\
&= \{\widetilde{\mathcal{I}}^{-1} - \widetilde{\mathcal{I}}^{-1} \circ (\widehat{\mathcal{I}} - \widetilde{\mathcal{I}}) \circ \widetilde{\mathcal{I}}^{-1}\}\{v_h + (\widehat{v} - v_h)\}(y) + o_p(n^{-1/2}h^{-1/2}) \\
&= \widetilde{a}_h(y) + \widetilde{\mathcal{I}}^{-1}(\widehat{v} - v_h)(y) - \{\widetilde{\mathcal{I}}^{-1} \circ (\widehat{\mathcal{I}} - \widetilde{\mathcal{I}})\}(\widetilde{a}_h)(y) + o_p(n^{-1/2}h^{-1/2}) \\
&= \widetilde{a}_h(y) + \widetilde{\mathcal{I}}^{-1}\{\widehat{v} - \widehat{\mathcal{I}}(\widetilde{a}_h)\}(y) + o_p(n^{-1/2}h^{-1/2}) \\
&= \widetilde{a}_h(y) + N^{-1}\sum_{i=1}^{N} d_{h,l,i}(y) + o_p(n^{-1/2}h^{-1/2}) \qquad \text{(A.28)}
\end{aligned}$$

uniformly in y by Condition (A4). Here, the third equality holds by (A.25) and

$$\|\widehat{v}\|_2 \leqslant \|v_h\|_2 + \|\widehat{v} - v_h\|_2 = O(h^{-1/2}) + o_p(n^{-1/4}h^{-1/2}) = O_p(h^{-1/2})$$

by (A.26), the fourth equality holds by $\widetilde{a}_h(y) = \widetilde{\mathcal{I}}^{-1}(v_h)(y)$, (A.25), and (A.26), and the last equality holds by (A.27). Then

$$\|\widehat{a} - \widetilde{a}_h\|_2 = O(h^{-1/2}l^m) + O_p\{h^{-1/2}(nl)^{-1/2}\} + o_p(n^{-1/2}h^{-1/2}) = o_p(n^{-1/4}h^{-1/2}) \quad \text{(A.29)}$$

by (A.21), (A.22), (A.28), and Condition (A6), and

$$\|\widehat{a}\|_2 \leqslant \|\widetilde{a}_h\|_2 + \|\widehat{a} - \widetilde{a}_h\|_2 = O_p(h^{-1/2}) \qquad \text{(A.30)}$$

by (A.20). Further, (A.18) leads to

$$\begin{aligned}
& \left| \frac{\partial \widetilde{b}(\mathbf{x}, \widehat{a}, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) - \frac{\partial \widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \right| \\
= & \left| \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{(\widehat{a} - \widetilde{a}_h)(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} \right. \\
& \left. - \frac{\mathrm{E}_p\{(\widehat{a} - \widetilde{a}_h)(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}]^2} \right| \\
\leqslant & \frac{|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{(\widehat{a} - \widetilde{a}_h)(Y)\widetilde{\rho}(Y) \mid \mathbf{x}\}|}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}} \\
& + \frac{\|\widehat{a} - \widetilde{a}_h\|_2 \|\widetilde{\rho}(\cdot) p_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_2 |(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}|}{[\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}]^2} \\
= & o_p(n^{-1/4}n^{-1/4}h^{-1/2}) + o_p(n^{-1/4}h^{-1/2}n^{-1/4}) \\
= & o_p(n^{-1/2}h^{-1/2}), \qquad \text{(A.31)}
\end{aligned}$$

because $\widetilde{\rho}(y)$ is bounded away from zero and above by Conditions (A1) and (A4), $|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y) \mid \mathbf{x}\}| = o_p(n^{-1/4}\|a\|_2)$, and $\|\widehat{a} - \widetilde{a}_h\|_2 = o_p(n^{-1/4}h^{-1/2})$ by (A.29). Hence,

$$\begin{aligned}
& \widetilde{b}(\mathbf{x}, \widehat{a}, \widehat{\mathrm{E}}_p) \qquad\qquad (A.32)\\
= \;& \widetilde{b}(\mathbf{x}, \widehat{a}, \mathrm{E}_p) + \frac{\partial \widetilde{b}(\mathbf{x}, \widehat{a}, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n^{-1/2}h^{-1/2})\\
= \;& \widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p) + \widetilde{b}(\mathbf{x}, \widehat{a} - \widetilde{a}_h, \mathrm{E}_p) + \frac{\partial \widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n^{-1/2}h^{-1/2})\\
= \;& \widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p) + N^{-1}\sum_{i=1}^{N} \widetilde{b}(\mathbf{x}, d_{h,l,i}, \mathrm{E}_p) + \frac{\partial \widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n^{-1/2}h^{-1/2})
\end{aligned}$$

uniformly in $\mathbf{x}$ by Condition (A4), where the first equality holds by (A.24) and (A.30), the second equality holds by (A.31), and the last equality holds by (A.28).

Now, we analyze $\widehat{\delta}_0$. From the definition of $\widehat{\delta}_0$,

$$\begin{aligned}
\widehat{\delta}_0 - \delta_0 &= N^{-1}\sum_{i=1}^{N}\left[\frac{r_i}{\pi}\widetilde{\rho}(y_i)\{K_h(y_i - y_0) - \widetilde{b}(\mathbf{x}_i, \widehat{a}, \widehat{\mathrm{E}}_p)\} + \frac{1 - r_i}{1 - \pi}\{\widetilde{b}(\mathbf{x}_i, \widehat{a}, \widehat{\mathrm{E}}_p) - \delta_0\}\right]\\
&= -N^{-1}\sum_{i=1}^{N}\widetilde{\phi}_{\mathrm{eff}}(r_i, r_i y_i, \mathbf{x}_i)\\
&\quad -N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\widetilde{\rho}(y_i) - \frac{1 - r_i}{1 - \pi}\right\}\{\widetilde{b}(\mathbf{x}_i, \widehat{a}, \widehat{\mathrm{E}}_p) - \widetilde{b}(\mathbf{x}_i, \widetilde{a}_h, \mathrm{E}_p)\}\\
&= -N^{-1}\sum_{i=1}^{N}\widetilde{\phi}_{\mathrm{eff}}(r_i, r_i y_i, \mathbf{x}_i) - T_1 - T_2 + o_p(n^{-1/2}h^{-1/2}), \qquad (A.33)
\end{aligned}$$

where $\widetilde{\phi}_{\mathrm{eff}}(r, ry, \mathbf{x})$ is $\phi_{\mathrm{eff}}(r, ry, \mathbf{x})$ with $\rho(y)$ replaced by $\widetilde{\rho}(y)$,

$$\begin{aligned}
T_1 &\equiv N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\frac{r_i}{\pi}\widetilde{\rho}(y_i) - \frac{1 - r_i}{1 - \pi}\right\}\widetilde{b}(\mathbf{x}_i, d_{h,l,j}, \mathrm{E}_p),\\
T_2 &\equiv N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\widetilde{\rho}(y_i) - \frac{1 - r_i}{1 - \pi}\right\}\frac{\partial \widetilde{b}(\mathbf{x}_i, \widetilde{a}_h, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p).
\end{aligned}$$

By the property of the U-statistic, T_1 can be written as

$$
\begin{aligned}
& T_1 \qquad\qquad (A.34)\\
= \;& N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\widetilde{\rho}(y_i)-\frac{1-r_i}{1-\pi}\right\}\mathrm{E}\{\widetilde{b}(\mathbf{x}_i, D_{h,l,j}, \mathrm{E}_p)\mid r_i, r_iy_i, \mathbf{x}_i\}\\
& +N^{-1}\sum_{j=1}^{N}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\widetilde{\rho}(Y_i)-\frac{1-R_i}{1-\pi}\right\}\widetilde{b}(\mathbf{X}_i, d_{h,l,j}, \mathrm{E}_p)\mid r_j, r_jy_j, \mathbf{x}_j\right]\\
& -N^{-1/2}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\widetilde{\rho}(Y_i)-\frac{1-R_i}{1-\pi}\right\}\widetilde{b}(\mathbf{X}_i, D_{h,l,j}, \mathrm{E}_p)\right]+O_p(N^{-1}h^{-1/2})\\
= \;& N^{-1}\sum_{j=1}^{N}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\widetilde{\rho}(Y_i)-\frac{1-R_i}{1-\pi}\right\}\widetilde{b}(\mathbf{X}_i, d_{h,l,j}, \mathrm{E}_p)\mid r_j, r_jy_j, \mathbf{x}_j\right]\\
& +O_p(h^{-1/2}l^m+N^{-1}h^{-1/2})\\
= \;& N^{-1}\sum_{j=1}^{N}\int\left[\widetilde{\rho}(y)\mathrm{E}\{\widetilde{b}(\mathbf{X}, d_{h,l,j}, \mathrm{E}_p)\mid y\}p_Y(y)-\mathrm{E}\{\widetilde{b}(\mathbf{X}, d_{h,l,j}, \mathrm{E}_p)\mid y\}q_Y(y)\right]dy\\
& +o_p(n^{-1/2}h^{-1/2})\\
= \;& N^{-1}\sum_{j=1}^{N}\int\{\widetilde{\rho}(y)-\rho(y)\}\widetilde{\mathcal{I}}(d_{h,l,j})(y)dy+o_p(n^{-1/2}h^{-1/2})\\
= \;& N^{-1}\sum_{j=1}^{N}\int\{\widetilde{\rho}(y)-\rho(y)\}\{v_{h,l,j}(y)-\widetilde{\mathcal{I}}_{l,j}(\widetilde{a}_h)(y)\}dy+o_p(n^{-1/2}h^{-1/2})\\
= \;& N^{-1}\sum_{j=1}^{N}\frac{r_j}{\pi}\int\widetilde{K}_l(y-y_j)\{\widetilde{\rho}(y)-\rho(y)\}\{K_h(y-y_0)-\widetilde{b}(\mathbf{x}_j, \widetilde{a}_h, \mathrm{E}_p)\}dy+o_p(n^{-1/2}h^{-1/2}),
\end{aligned}
$$

where the second equality holds because (A.21) leads to

$$
\mathrm{E}\{\widetilde{b}(\mathbf{x}_i, D_{h,l,j}, \mathrm{E}_p)\mid r_i, r_iy_i, \mathbf{x}_i\}=O(h^{-1/2}l^m),
$$

and the third equality holds by Condition (A6). Further,

$$
\begin{aligned}
& \int\widetilde{K}_l(y-y_j)\{\widetilde{\rho}(y)-\rho(y)\}\{K_h(y-y_0)-\widetilde{b}(\mathbf{x}_j, \widetilde{a}_h, \mathrm{E}_p)\}dy\\
= \;& \int\widetilde{K}(t)\{\widetilde{\rho}(y_j+lt)-\rho(y_j+lt)\}\{K_h(y_j+lt-y_0)-\widetilde{b}(\mathbf{x}_j, \widetilde{a}_h, \mathrm{E}_p)\}dt\\
= \;& \{\widetilde{\rho}(y_j)-\rho(y_j)\}\{K_h(y_j-y_0)-\widetilde{b}(\mathbf{x}_j, \widetilde{a}_h, \mathrm{E}_p)\}+O(h^{-m-1}l^m)\\
= \;& \{\widetilde{\rho}(y_j)-\rho(y_j)\}\{K_h(y_j-y_0)-\widetilde{b}(\mathbf{x}_j, \widetilde{a}_h, \mathrm{E}_p)\}+o(n^{-1/2}h^{-1/2}) \qquad (A.35)
\end{aligned}
$$

under Conditions (A1), (A5), and (A6). On the other hand, T_2 can be written as

$$\begin{aligned} T_2 &= \mathrm{E}\left[\left\{\frac{R}{\pi}\widetilde{\rho}(Y)-\frac{1-R}{1-\pi}\right\}\frac{\partial\widetilde{b}(\mathbf{X},\widetilde{a}_h,\mathrm{E}_p)}{\partial\mathrm{E}_p}(\widehat{\mathrm{E}}_p-\mathrm{E}_p)\right]+O_p(n^{-1/2})o_p(n^{-1/4}h^{-1/2}) \\ &= \frac{\partial}{\partial\mathrm{E}_p}\mathrm{E}\left[\left\{\frac{R}{\pi}\widetilde{\rho}(Y)-\frac{1-R}{1-\pi}\right\}K_h(Y-y_0)\right](\widehat{\mathrm{E}}_p-\mathrm{E}_p)+o_p(n^{-1/2}h^{-1/2}) \\ &= o_p(n^{-1/2}h^{-1/2}), \end{aligned} \tag{A.36}$$

where the first equality holds by (A.20) and (A.23), and the second equality holds by (A.17). Therefore, (A.33), (A.34), (A.35) and (A.36) lead to

$$\begin{aligned} &\widehat{\delta}_0-\delta_0 \\ =\;& -N^{-1}\sum_{i=1}^N\widetilde{\phi}_{\mathrm{eff}}(r_i,r_iy_i,\mathbf{x}_i) \\ & -N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\{\widetilde{\rho}(y_i)-\rho(y_i)\}\{K_h(y_i-y_0)-\widetilde{b}(\mathbf{x}_i,\widetilde{a}_h,\mathrm{E}_p)\}+o_p(n^{-1/2}h^{-1/2}) \\ =\;& N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\rho(y_i)\left\{K_h(y_i-y_0)-\widetilde{b}(\mathbf{x}_i,\widetilde{a}_h,\mathrm{E}_p)\right\} \\ & +N^{-1}\sum_{i=1}^N\frac{1-r_i}{1-\pi}\left\{\widetilde{b}(\mathbf{x}_i,\widetilde{a}_h,\mathrm{E}_p)-\delta_0\right\}+o_p(n^{-1/2}h^{-1/2}) \\ =\;& N^{-1}\sum_{i=1}^N\frac{r_i}{\pi}\rho(y_i)K_h(y_i-y_0)-N^{-1}\sum_{i=1}^N\left\{\frac{r_i}{\pi}\rho(y_i)-\frac{1-r_i}{1-\pi}\right\}\widetilde{b}(\mathbf{x}_i,\widetilde{a}_h,\mathrm{E}_p) \\ & -\delta_0+o_p(n^{-1/2}h^{-1/2}). \end{aligned} \tag{A.37}$$

Now, for $a(y)$ such that $\|a\|_2<\infty$,

$$\begin{aligned} \|(\widetilde{\mathcal{I}}-\mathcal{I})(a)\|_2 &= \left[\int\left\{\int a(t)(\widetilde{u}-u)(t,y)dt\right\}^2dy\right]^{1/2} \\ &\leqslant \left[\int\|a\|_2^2\|(\widetilde{u}-u)(\cdot,y)\|_2^2dy\right]^{1/2} \\ &= o_p(\|a\|_2) \end{aligned} \tag{A.38}$$

since $\sup_y|\widetilde{\rho}(y)-\rho(y)|=o_p(1)$. Recall that $a_h(y)$ in $\boldsymbol{\phi}_{\mathrm{eff}}(r,ry,\mathbf{x})$ satisfies $\mathcal{I}(a_h)(y)=v_h(y)$, and $\widetilde{a}_h(y)$ satisfies $\widetilde{\mathcal{I}}(\widetilde{a}_h)(y)=v_h(y)$. Also, recall Lemma 6 to have that both $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ are linear, invertible, bounded, and their inverses are also bounded. Then,

$$\begin{aligned} \widetilde{a}_h(y) &= \widetilde{\mathcal{I}}^{-1}(v_h)(y) \\ &= \{\mathcal{I}+(\widetilde{\mathcal{I}}-\mathcal{I})\}^{-1}(v_h)(y) \\ &= \{\mathcal{I}^{-1}-\mathcal{I}^{-1}\circ(\widetilde{\mathcal{I}}-\mathcal{I})\circ\mathcal{I}^{-1}\}(v_h)(y)+o_p(h^{-1/2}) \\ &= a_h(y)-\{\mathcal{I}^{-1}\circ(\widetilde{\mathcal{I}}-\mathcal{I})\}(a_h)(y)+o_p(h^{-1/2}) \\ &= a_h(y)+o_p(h^{-1/2}), \end{aligned}$$

where the third equality holds by (A.38) and $\|v_h\|_2 = O(h^{-1/2})$, and the last equality holds by (A.38) and $\|a_h\|_2 = \|\mathcal{I}^{-1}(v_h)\|_2 = O(h^{-1/2})$. This holds uniformly for y by Condition (A4). This and $\sup_y |\widetilde{\rho}(y) - \rho(y)| = o_p(1)$ further lead to

$$\widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p) = b(\mathbf{x}, a_h, \mathrm{E}_p) + o_p(h^{-1/2})$$

where

$$b(\mathbf{x}, a, \mathrm{E}_p) = \frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}},$$

and this holds uniformly for $\mathbf{x}$ by Condition (A4). Then,

$$\begin{aligned}
& N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\}\widetilde{b}(\mathbf{x}_i, \widetilde{a}_h, \mathrm{E}_p) \\
= \; & N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\} b(\mathbf{x}_i, a_h, \mathrm{E}_p) \\
& + \mathrm{E}\left[\left\{\frac{R}{\pi}\rho(Y) - \frac{1-R}{1-\pi}\right\}\mathrm{E}\left\{\widetilde{b}(\mathbf{X}, \widetilde{a}_h, \mathrm{E}_p) - b(\mathbf{X}, a_h, \mathrm{E}_p) \mid Y\right\}\right] + o_p(n^{-1/2}h^{-1/2}) \\
= \; & N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\} b(\mathbf{x}_i, a_h, \mathrm{E}_p) + o_p(n^{-1/2}h^{-1/2}),
\end{aligned}$$

where the second equality holds since $\rho(y)$ is bounded under Condition (A4) and $\widetilde{b}(\mathbf{x}, \widetilde{a}_h, \mathrm{E}_p) - b(\mathbf{x}, a_h, \mathrm{E}_p) = o_p(h^{-1/2})$, and the last equality holds since $\mathrm{E}\{\widetilde{b}(\mathbf{X}, a_h, \mathrm{E}_p) - b(\mathbf{X}, a_h, \mathrm{E}_p) \mid y\} = K_h(y - y_0) - K_h(y - y_0) = 0$. This fact and (A.37) lead to

$$\begin{aligned}
\widehat{\delta}_0 - \delta_0 &= N^{-1}\sum_{i=1}^{N}\frac{r_i}{\pi}\rho(y_i)K_h(y_i - y_0) - N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\} b(\mathbf{x}_i, a_h, \mathrm{E}_p) \\
&\quad -\delta_0 + o_p(n^{-1/2}h^{-1/2}) \\
&= N^{-1}\sum_{i=1}^{N}\frac{r_i}{\pi}\rho(y_i)\left\{K_h(y_i - y_0) - b(\mathbf{x}_i, a_h, \mathrm{E}_p)\right\} \\
&\quad + N^{-1}\sum_{i=1}^{N}\frac{1-r_i}{1-\pi}\left\{b(\mathbf{x}_i, a_h, \mathrm{E}_p) - \delta_0\right\} + o_p(n^{-1/2}h^{-1/2}).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\delta_0 &= \int K_h(y - y_0)q_Y(y)dy \\
&= \int K(t)q_Y(y_0 + ht)dt \\
&= q_Y(y_0) + q_Y^{(m)}(y_0)\frac{\int t^m K(t)dt}{m!}h^m + O(h^{m+1})
\end{aligned}$$

under Conditions (A4)-(A6), and

$$\begin{aligned}
\mathrm{var}(\widehat{\delta}_0 - \delta_0) &= N^{-1}\mathrm{E}\left(\left[\frac{R_i}{\pi}\rho(Y_i)\left\{K_h(Y_i - y_0) - b(\mathbf{X}_i, a_h, \mathrm{E}_p)\right\}\right]^2\right) \\
&\quad + N^{-1}\mathrm{E}\left(\left[\frac{1-R_i}{1-\pi}\left\{b(\mathbf{X}_i, a_h, \mathrm{E}_p) - \delta_0\right\}\right]^2\right) \\
&\quad + o[\{\min(n, N-n)\}^{-1/2} n^{-1/2} h^{-1}] \\
&= n^{-1}\mathrm{E}\left(\frac{R_i}{\pi}\left[\rho(Y_i)\left\{K_h(Y_i - y_0) - b(\mathbf{X}_i, a_h, \mathrm{E}_p)\right\}\right]^2\right) \\
&\quad + (N-n)^{-1}\mathrm{E}\left[\frac{1-R_i}{1-\pi}\left\{b(\mathbf{X}_i, a_h, \mathrm{E}_p) - \delta_0\right\}^2\right] \\
&\quad + o[\{\min(n, N-n)\}^{-1/2} n^{-1/2} h^{-1}] \\
&= n^{-1}\mathrm{E}_p\left(\left[\rho(Y_i)\left\{K_h(Y_i - y_0) - b(\mathbf{X}_i, a_h, \mathrm{E}_p)\right\}\right]^2\right) \\
&\quad + (N-n)^{-1}\mathrm{E}_q\left[\left\{b(\mathbf{X}_i, a_h, \mathrm{E}_p) - \delta_0\right\}^2\right] \\
&\quad + o[\{\min(n, N-n)\}^{-1/2} n^{-1/2} h^{-1}].
\end{aligned}$$

Because the variance of $\widehat{\delta}_0$ is the variance of the efficient influence function of estimating $E_q\{K_h(Y-y)\}$, hence $\widehat{\delta}_0$ is efficient. Further at any fixed $K(\cdot), h$, $\widehat{\delta}_0$ is an efficient estimator of $q_Y(y)$. ■

Appendix A.6. Proof of Theorem 12

We recall the last two displays in the proof of Theorem S.2 in the supplement of Lee et al. (2025). These equations with $\rho^* = \widetilde{\rho}$ lead to

$$\begin{aligned}
\mathbf{0} &= N^{-1/2}\sum_{i=1}^{N}\left[\sqrt{\pi}\widetilde{\boldsymbol{\phi}}_{\mathrm{eff}}(r_i, r_i y_i, \mathbf{x}_i, \boldsymbol{\theta}) + \frac{r_i}{\sqrt{\pi}}\widetilde{\mathbf{A}}(\boldsymbol{\theta})\{\widetilde{\mathbf{b}}(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{U}(y_i, \mathbf{x}_i, \boldsymbol{\theta})\}\{\widetilde{\rho}(y_i) - \rho(y_i)\}\right] \\
&\quad + \left\{\widetilde{\mathbf{A}}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^{-1} + o_p(1)\right\}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1),
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\boldsymbol{\phi}}_{\mathrm{eff}}(r, ry, \mathbf{x}, \boldsymbol{\theta}) &= \frac{r}{\pi}\widetilde{\rho}(y)\widetilde{\mathbf{A}}(\boldsymbol{\theta})\{\mathbf{U}(y, \mathbf{x}, \boldsymbol{\theta}) - \widetilde{\mathbf{b}}(\mathbf{x}, \boldsymbol{\theta})\} + \frac{1-r}{1-\pi}\widetilde{\mathbf{A}}(\boldsymbol{\theta})\widetilde{\mathbf{b}}(\mathbf{x}, \boldsymbol{\theta}), \\
\widetilde{\mathbf{A}}(\boldsymbol{\theta}) &= \left(\mathrm{E}_p\left[\widetilde{\rho}(Y)\mathrm{E}\{\partial\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}} \mid Y\}\right]\right)^{-1}, \\
\widetilde{\mathbf{b}}(\mathbf{x}, \boldsymbol{\theta}) &= \frac{\mathrm{E}_p\{\mathbf{U}(Y, \mathbf{x}, \boldsymbol{\theta})\widetilde{\rho}^2(Y) + \widetilde{\mathbf{a}}(Y, \boldsymbol{\theta})\widetilde{\rho}(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{x}\}},
\end{aligned}$$

and $\widetilde{\mathbf{a}}(y, \boldsymbol{\theta})$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})\widetilde{\rho}^2(Y) + \widetilde{\mathbf{a}}(Y, \boldsymbol{\theta})\widetilde{\rho}(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\widetilde{\rho}^2(Y) + \pi/(1-\pi)\widetilde{\rho}(Y) \mid \mathbf{X}\}} \mid y\right] = \mathrm{E}\{\mathbf{U}(y, \mathbf{X}, \boldsymbol{\theta}) \mid y\}.$$

This leads to

$$
\begin{aligned}
& -\left\{\mathbf{A}(\boldsymbol{\theta})^{-1}+o_p(1)\right\}\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})+o_p(1) \qquad \text{(A.39)}\\
= \;& N^{-1/2}\sum_{i=1}^{N}\sqrt{\pi}\left[\widetilde{\mathbf{A}}(\boldsymbol{\theta})^{-1}\widetilde{\boldsymbol{\phi}}_{\text{eff}}(r_i, r_i y_i, \mathbf{x}_i, \boldsymbol{\theta})+\frac{r_i}{\pi}\{\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta})-\mathbf{U}(y_i,\mathbf{x}_i,\boldsymbol{\theta})\}\{\widetilde{\rho}(y_i)-\rho(y_i)\}\right]\\
= \;& N^{-1/2}\sum_{i=1}^{N}\sqrt{\pi}\left[\frac{r_i}{\pi}\widetilde{\rho}(y_i)\{\mathbf{U}(y_i,\mathbf{x}_i,\boldsymbol{\theta})-\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta})\}+\frac{1-r_i}{1-\pi}\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta})\right.\\
& \left.+\frac{r_i}{\pi}\{\rho(y_i)-\widetilde{\rho}(y_i)\}\{\mathbf{U}(y_i,\mathbf{x}_i,\boldsymbol{\theta})-\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta})\}\right]\\
= \;& N^{-1/2}\sum_{i=1}^{N}\sqrt{\pi}\left[\frac{r_i}{\pi}\rho(y_i)\{\mathbf{U}(y_i,\mathbf{x}_i,\boldsymbol{\theta})-\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta})\}+\frac{1-r_i}{1-\pi}\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta})\right]\\
= \;& \sqrt{n}N^{-1}\sum_{i=1}^{N}\frac{r_i}{\pi}\rho(y_i)\mathbf{U}(y_i,\mathbf{x}_i,\boldsymbol{\theta})-\sqrt{n}N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i)-\frac{1-r_i}{1-\pi}\right\}\widetilde{\mathbf{b}}(\mathbf{x}_i,\boldsymbol{\theta}).
\end{aligned}
$$

Let

$$
\begin{aligned}
\mathcal{I}(\mathbf{a})(y,\boldsymbol{\theta}) &\equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}_p\{\mathbf{a}(Y,\boldsymbol{\theta})\rho(Y)\mid\mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y)+\pi/(1-\pi)\rho(Y)\mid\mathbf{X}\}}\mid y\right],\\
\mathbf{v}(y,\boldsymbol{\theta}) &\equiv p_Y(y)\mathrm{E}\left[\mathbf{U}(y,\mathbf{X},\boldsymbol{\theta})-\frac{\mathrm{E}_p\{\mathbf{U}(Y,\mathbf{X},\boldsymbol{\theta})\rho^2(Y)\mid\mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y)+\pi/(1-\pi)\rho(Y)\mid\mathbf{X}\}}\mid y\right],
\end{aligned}
$$

and $\widetilde{\mathcal{I}}(\mathbf{a})(y,\boldsymbol{\theta}), \widetilde{\mathbf{v}}(y,\boldsymbol{\theta})$ be $\mathcal{I}(\mathbf{a})(y,\boldsymbol{\theta}), \mathbf{v}(y,\boldsymbol{\theta})$ with ρ replaced by $\widetilde{\rho}$. For $\mathbf{a}(y,\boldsymbol{\theta})=\{a_1(y,\boldsymbol{\theta}),\ldots,a_p(y,\boldsymbol{\theta})\}^{\mathrm{T}}$, let $\|\mathbf{a}\|_\infty=\max_{k=1,\ldots,p}\sup_{y,\boldsymbol{\theta}}|a_k(y,\boldsymbol{\theta})|$. Then for any $\mathbf{a}(y,\boldsymbol{\theta})$ with $\|\mathbf{a}\|_\infty<\infty$,

$$
\|(\widetilde{\mathcal{I}}-\mathcal{I})(\mathbf{a})\|_\infty = o_p(1) \qquad \text{(A.40)}
$$

since $\sup_y|\widetilde{\rho}(y)-\rho(y)|=o_p(1)$, and the support of Y and the parameter space of $\boldsymbol{\theta}$ are compact under the regularity conditions. Similarly,

$$
\|\widetilde{\mathbf{v}}-\mathbf{v}\|_\infty = o_p(1). \qquad \text{(A.41)}
$$

Note that $\mathbf{a}(y,\boldsymbol{\theta})$ in $\boldsymbol{\phi}_{\text{eff}}(r,ry,\mathbf{x},\boldsymbol{\theta})$ satisfies $\mathcal{I}(\mathbf{a})(y,\boldsymbol{\theta})=\mathbf{v}(y,\boldsymbol{\theta})$, and $\widetilde{\mathbf{a}}(y,\boldsymbol{\theta})$ satisfies $\widetilde{\mathcal{I}}(\widetilde{\mathbf{a}})(y,\boldsymbol{\theta})=\widetilde{\mathbf{v}}(y,\boldsymbol{\theta})$. Also, we recall Lemma S.1 in the supplement of Lee et al. (2025) to have that both $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ are linear, invertible, bounded, and their inverses are also bounded. Then,

$$
\begin{aligned}
\widetilde{\mathbf{a}}(y,\boldsymbol{\theta}) &= \widetilde{\mathcal{I}}^{-1}(\widetilde{\mathbf{v}})(y,\boldsymbol{\theta})\\
&= \{\mathcal{I}+(\widetilde{\mathcal{I}}-\mathcal{I})\}^{-1}\{\mathbf{v}+(\widetilde{\mathbf{v}}-\mathbf{v})\}(y,\boldsymbol{\theta})\\
&= \{\mathcal{I}^{-1}-\mathcal{I}^{-1}\circ(\widetilde{\mathcal{I}}-\mathcal{I})\circ\mathcal{I}^{-1}\}\{\mathbf{v}+(\widetilde{\mathbf{v}}-\mathbf{v})\}(y,\boldsymbol{\theta})+o_p(1)\\
&= \mathbf{a}(y,\boldsymbol{\theta})+\mathcal{I}^{-1}(\widetilde{\mathbf{v}}-\mathbf{v})(y,\boldsymbol{\theta})-\{\mathcal{I}^{-1}\circ(\widetilde{\mathcal{I}}-\mathcal{I})\}(\mathbf{a})(y,\boldsymbol{\theta})+o_p(1)\\
&= \mathbf{a}(y,\boldsymbol{\theta})+o_p(1),
\end{aligned}
$$

where the third equality holds by (A.40), and the fourth and last equality hold by (A.40) and (A.41). This holds uniformly for $(y,\boldsymbol{\theta})$ by the compactness of the support of Y and the

parameter space of $\boldsymbol{\theta}$ under the regularity conditions. This and $\sup_y |\widetilde{\rho}(y) - \rho(y)| = o_p(1)$ further lead to

$$\widetilde{\mathbf{b}}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{b}(\mathbf{x}, \boldsymbol{\theta}) + o_p(1)$$

where

$$\mathbf{b}(\mathbf{x}, \boldsymbol{\theta}) = \frac{\mathrm{E}_p\{\mathbf{U}(Y, \mathbf{x}, \boldsymbol{\theta})\rho^2(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}},$$

and this holds uniformly for $(\mathbf{x}, \boldsymbol{\theta})$ by the compactness of the support of $\mathbf{X}$ and the parameter space of $\boldsymbol{\theta}$ under the regularity conditions. Then,

$$\begin{aligned}
& N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\}\widetilde{\mathbf{b}}(\mathbf{x}_i, \boldsymbol{\theta}) \\
= \; & N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\}\mathbf{b}(\mathbf{x}_i, \boldsymbol{\theta}) \\
& + \mathrm{E}\left[\left\{\frac{R}{\pi}\rho(Y) - \frac{1-R}{1-\pi}\right\}\mathrm{E}\left\{\widetilde{\mathbf{b}}(\mathbf{X}, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{X}, \boldsymbol{\theta}) \mid Y\right\}\right] + o_p(n^{-1/2}) \\
= \; & N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\}\mathbf{b}(\mathbf{x}_i, \boldsymbol{\theta}) + o_p(n^{-1/2}),
\end{aligned}$$

where the second equality holds since $\rho(y)$ is bounded under the regularity condition and $\widetilde{\mathbf{b}}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{x}, \boldsymbol{\theta}) = o_p(1)$, and the last equality holds since $\mathrm{E}\{\widetilde{\mathbf{b}}(\mathbf{X}, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{X}, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(y, \mathbf{X}, \boldsymbol{\theta}) - \mathbf{U}(y, \mathbf{X}, \boldsymbol{\theta}) \mid y\} = \mathbf{0}$. This fact and (A.39) lead to

$$\begin{aligned}
& -\left\{\mathbf{A}(\boldsymbol{\theta})^{-1} + o_p(1)\right\}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1) \\
= \; & \sqrt{n}N^{-1}\sum_{i=1}^{N}\frac{r_i}{\pi}\rho(y_i)\mathbf{U}(y_i, \mathbf{x}_i, \boldsymbol{\theta}) - \sqrt{n}N^{-1}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\rho(y_i) - \frac{1-r_i}{1-\pi}\right\}\mathbf{b}(\mathbf{x}_i, \boldsymbol{\theta}) + o_p(1) \\
= \; & N^{-1/2}\sum_{i=1}^{N}\sqrt{\pi}\left[\frac{r_i}{\pi}\rho(y_i)\{\mathbf{U}(y_i, \mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{x}_i, \boldsymbol{\theta})\} + \frac{1-r_i}{1-\pi}\mathbf{b}(\mathbf{x}_i, \boldsymbol{\theta})\right] + o_p(1).
\end{aligned}$$

■

Appendix A.7. Sensitivity Analysis to the Magnitude of Label Shift

To evaluate our proposed method under varying magnitudes of label shift, we conducted additional simulations by increasing the distance between the source and target distributions. Specifically, we varied the target distribution $Y \mid R = 0 \sim N(\mu, \sigma^2)$ by taking $(\mu, \sigma^2) \in \{(1.25, 1.25), (1.5, 1.5)\}$. We measured the shift magnitude using the L_2-norm distance between the true density ratio and unity, $\|\rho - 1\|_{L_2(p_Y)} = [\int\{\rho(y) - 1\}^2 p_Y(y)dy]^{1/2}$, which was 0.7820 for the original setting in Section 6, and increased to 0.9508 and 1.2411 for the new settings here, respectively.

Tables A.1 and A.2 summarize the performance of the estimators under these increased shift magnitudes based on 1000 simulation replicates. While the MSE of all estimators

Estimator	MSE	Bias	SE	ARE	CI
`Shift-dependent*`	59.4159	7.2168	2.7095	37.799	0.910
`BBSE`	134.7256	-10.9149	3.9504	85.709	0.177
`BBSE+PPI`	19.7269	-2.5926	3.6081	12.550	0.514
`ELSA`	5.9311	0.1049	2.4343	3.773	0.997
`ELSA+PPI`	5.9311	0.1049	2.4343	3.773	0.997
`Doubly-flexible**`	4.7610	0.6805	2.0742	3.029	0.905
`Singly-flexible*`	2.5004	0.2160	1.5672	1.591	0.934
`Efficient~`	1.7122	-0.0707	1.3073	1.089	0.959
`Efficient^`	1.7639	-0.0793	1.3264	1.122	0.961
`Oracle`	1.5719	-0.0136	1.2543	1.000	0.962

Table A.1: Summary of $\mathrm{E}_q(Y)$ estimates under $Y \sim N(1.25, 1.25)$ in $\mathcal{Q}$.

Estimator	MSE	Bias	SE	ARE	CI
`Shift-dependent*`	111.4633	9.1794	5.2181	52.259	0.999
`BBSE`	288.0817	-13.3593	10.4748	135.066	0.163
`BBSE+PPI`	92.1207	-3.2685	9.0288	43.190	0.449
`ELSA`	7.8770	0.1342	2.8048	3.693	0.997
`ELSA+PPI`	7.8770	0.1342	2.8048	3.693	0.997
`Doubly-flexible**`	30.7471	2.4528	4.9755	14.416	0.740
`Singly-flexible*`	7.9418	0.6688	2.7390	3.724	0.871
`Efficient~`	2.4120	-0.3573	1.5122	1.131	0.929
`Efficient^`	2.5150	-0.4310	1.5270	1.179	0.931
`Oracle`	2.1329	-0.1767	1.4504	1.000	0.952

Table A.2: Summary of $\mathrm{E}_q(Y)$ estimates under $Y \sim N(1.5, 1.5)$ in $\mathcal{Q}$.

naturally increases as the shift becomes more severe, our proposed estimators `Efficient~` and `Efficient^` consistently maintain the lowest MSE and achieve valid confidence interval coverage. These results demonstrate the robustness of the proposed inference procedure across different label shift magnitudes.

Appendix A.8. Robustness to Violations of the Label Shift Assumption

To empirically investigate the robustness of our estimators when the label shift assumption is violated, we conducted additional sensitivity analyses. To perturb the label shift assumption, we introduced independent Gaussian noise to the covariates of $\mathcal{Q}$. Specifically, while the source covariates were generated as in the original setting as in Section 6, the target covariates were perturbed such that

$$\mathbf{X} \mid (Y, R = 0) \overset{d}{=} (\mathbf{X} \mid Y, R = 1) + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}).$$

We evaluated the performance under two noise levels by setting the standard deviation σ_ϵ to 0.25 and 0.5.

Tables A.3 and A.4 summarize the results based on 1000 simulation replicates for $\sigma_\epsilon = 0.25$ and $\sigma_\epsilon = 0.5$, respectively. Although the MSE of all estimators increases with higher noise levels, our proposed estimators `Efficient~` and `Efficient^` maintain their superiority over the competing methods. Remarkably, they remain robust and provide valid

Estimator	MSE	Bias	SE	ARE	CI
`Shift-dependent*`	25.8420	4.9437	1.1846	19.605	0.218
`BBSE`	1194.6947	-7.7334	33.7050	906.376	0.196
`BBSE+PPI`	846.2374	-1.1942	29.0802	642.013	0.618
`ELSA`	5.1135	0.0800	2.2610	3.879	0.996
`ELSA+PPI`	5.1135	0.0800	2.2610	3.879	0.996
`Doubly-flexible**`	1.8635	0.2177	1.3483	1.414	0.953
`Singly-flexible*`	1.8022	0.0914	1.3400	1.367	0.949
`Efficient~`	1.4399	0.0538	1.1994	1.092	0.963
`Efficient^`	1.4304	0.0585	1.1952	1.085	0.966
`Oracle`	1.3181	0.0215	1.1484	1.000	0.968

Table A.3: Summary of $\mathrm{E}_q(Y)$ estimates under $\sigma_\epsilon = 0.25$.

Estimator	MSE	Bias	SE	ARE	CI
`Shift-dependent*`	25.8454	4.9441	1.1844	19.109	0.219
`BBSE`	2275.5139	-10.1415	46.6352	1682.450	0.204
`BBSE+PPI`	1629.0685	-3.2963	40.2470	1204.487	0.611
`ELSA`	7.0544	0.1323	2.6540	5.216	0.998
`ELSA+PPI`	7.0544	0.1323	2.6540	5.216	0.998
`Doubly-flexible**`	2.2568	0.5851	1.3844	1.669	0.939
`Singly-flexible*`	1.8520	0.0473	1.3607	1.369	0.948
`Efficient~`	1.5198	0.0054	1.2334	1.124	0.963
`Efficient^`	1.4999	0.0067	1.2253	1.109	0.965
`Oracle`	1.3525	-0.0618	1.1619	1.000	0.965

Table A.4: Summary of $\mathrm{E}_q(Y)$ estimates under $\sigma_\epsilon = 0.5$.

confidence interval coverage even as the prediction quality degrades. These findings seem to suggest that our EIF-based inference procedure is resilient to practical violations or uncertainties in the underlying label shift mechanism. More rigorous study on how to quantify such robustness will be challenging and warrants further investigations.

References

Ulzee An, Ali Pazokitoroudi, Marcus Alvarez, Lianyun Huang, Silviu Bacanu, Andrew J Schork, Kenneth Kendler, Päivi Pajukanta, Jonathan Flint, Noah Zaitlen, et al. Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nature Genetics*, 55(12):2269–2276, 2023.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

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

Peter J Bickel, J Klaassen, YA'Acov Ritov, and Jon A Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore, 1993.

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

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal and safe estimation for high-dimensional semi-supervised learning. *Journal of the American Statistical Association*, 119(548):2748–2759, 2024.

Marthinus Christoffel Du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, 35:22531–22546, 2022.

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. Rlsbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, pages 10879–10928. PMLR, 2023.

Jessica Gronsbell, Jianhui Gao, Yaqi Shi, Zachary R McCaw, and David Cheng. Another look at inference after prediction. *arXiv preprint arXiv:2411.19908*, 2024.

Xiang Gu, Xi Yu, yan yang, Jian Sun, and Zongben Xu. Adversarial reweighting for partial domain adaptation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14860–14872. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/7ce3284b743aefde80ffd9aec500e085-Paper.pdf`.

Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, and Dacheng Tao. Ltf: A label transformation framework for correcting label shift. In *International Conference on Machine Learning*, pages 3843–3853. PMLR, 2020.

Jue Hou, Zijian Guo, and Tianxi Cai. Surrogate assisted semi-supervised inference for high dimensional risk prediction. *Journal of Machine Learning Research*, 24(265):1–58, 2023.

Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538. PMLR, 2014.

Hwanwoo Kim, Xin Zhang, Jiwei Zhao, and Qinglong Tian. RetaSA: A nonparametric functional estimation approach for addressing continuous target shift. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Seong-ho Lee, Yanyuan Ma, and Jiwei Zhao. Doubly flexible estimation under label shift. *Journal of the American Statistical Association*, 120(549):278–290, 2025.

Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4): 1041–1054, 2023.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. Deep unsupervised domain adaptation: A review of recent advances and perspectives, 2022. URL `https://arxiv.org/abs/2208.07422`.

Ilan Livne, David Azriel, and Yair Goldberg. Improved estimators for semi-supervised high-dimensional regression model. *Electronic Journal of Statistics*, 16(2):5437–5487, 2022.

Yuwu Lu, Haoyu Huang, Wai Keung Wong, Xue Hu, Zhihui Lai, and Xuelong Li. Adaptive dispersal and collaborative clustering for few-shot unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 2025.

Yuwu Lu, Yifan Lan, Huan Yang, Zhihui Lai, and Xuelong Li. Exploring generic knowledge and reactivating source model for source-free universal domain adaptation. *IEEE Transactions on Multimedia*, 2026a.

Yuwu Lu, Yihan Yang, Wai Keung Wong, Anne Toomey, Zhihui Lai, and Xuelong Li. Energy-driven explicit alignment network: A blended-target domain adaptation approach. *IEEE Transactions on Multimedia*, 2026b.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

Zachary R McCaw, Jianhui Gao, Xihong Lin, and Jessica Gronsbell. Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nature Genetics*, pages 1–10, 2024.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *arXiv preprint arXiv:2311.14220*, 2023.

Jiacheng Miao, Yixuan Wu, Zhongxuan Sun, Xinran Miao, Tianyuan Lu, Jiwei Zhao, and Qiongshi Lu. Valid inference for machine learning-assisted genome-wide association studies. *Nature Genetics*, pages 1–9, 2024.

Keshav Motwani and Daniela Witten. Revisiting inference after prediction. *Journal of Machine Learning Research*, 24(394):1–18, 2023.

Tuan Duong Nguyen, Marthinus Christoffel, and Masashi Sugiyama. Continuous target shift adaptation in supervised learning. In *Asian Conference on Machine Learning*, pages 285–300. PMLR, 2016.

Robert J Olson, Alexi Shalapyonok, and Heidi M Sosik. An automated submersible flow cytometer for analyzing pico-and nanophytoplankton: Flowcytobot. *Deep Sea Research Part I: Oceanographic Research Papers*, 50(2):301–315, 2003.

Eric C Orenstein, Oscar Beijbom, Emily E Peacock, and Heidi M Sosik. Whoi-plankton-a large scale fine grained visual recognition benchmark dataset for plankton classification. *arXiv preprint arXiv:1510.00745*, 2015.

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

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Manley Roberts, Pranav Mani, Saurabh Garg, and Zachary Lipton. Unsupervised learning under latent label shift. *Advances in Neural Information Processing Systems*, 35:18763–18778, 2022.

B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. Omnipress, 2012.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 119(546):1065–1075, 2024.

Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

Dirk Tasche. Fisher consistency for prior probability shift. *The Journal of Machine Learning Research*, 18(1):3338–3369, 2017.

Qinglong Tian, Xin Zhang, and Jiwei Zhao. Elsa: Efficient label shift adaptation through the lens of semiparametric models. In *International Conference on Machine Learning*, pages 34120–34142. PMLR, 2023.

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

Mark J van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

Xue Yang, Hao Sun, Xian Sun, Menglong Yan, Zhi Guo, and Kun Fu. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access*, 6:50839–50849, 2018.

Chao Ying, Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Dependable exploitation of high-dimensional unlabeled data in an assumption-lean framework. *arXiv preprint arXiv:2603.27869*, 2026.

Oren Yuval and Saharon Rosset. Semi-supervised empirical risk minimization: Using unlabeled data to improve prediction. *Electronic Journal of Statistics*, 16(1):1434–1460, 2022.

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

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

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

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216, 2022.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.