Journal of Machine Learning Research 25 (2024) 1-59 Submitted 5/23; Revised 11/23; Published 1/24

Localized Debiased Machine Learning: Efficient Inference on Quantile Treatment Effects and Beyond

Nathan Kallus KALLUS@CORNELL.EDU
Cornell Tech
Cornell University
2 West Loop Rd, NY 10044, USA

Xiaojie Mao MAOXJ@SEM.TSINGHUA.EDU.CN
School of Economics and Management
Tsinghua University
Beijing, 100084, China

Masatoshi Uehara MU223@CORNELL.EDU
Cornell Tech
Cornell University
2 West Loop Rd, NY 10044, USA

Editor: David Sontag

Abstract

We consider estimating a low-dimensional parameter in an estimating equation involving high-dimensional nuisance functions that depend on the target parameter as an input. A central example is the efficient estimating equation for the (local) quantile treatment effect ((L)QTE) in causal inference, which involves the covariate-conditional cumulative distribution function evaluated at the quantile to be estimated. Existing approaches based on flexibly estimating the nuisances and plugging in the estimates, such as debiased machine learning (DML), require we learn the nuisance at all possible inputs. For (L)QTE, DML requires we learn the *whole* covariate-conditional cumulative distribution function. We instead propose *localized* debiased machine learning (LDML), which avoids this burdensome step and needs only estimate nuisances at a *single* initial rough guess for the target parameter. For (L)QTE, LDML involves learning just two regression functions, a standard task for machine learning methods. We prove that under lax rate conditions our estimator has the same favorable asymptotic behavior as the infeasible estimator that uses the unknown true nuisances. Thus, LDML notably enables practically-feasible and theoretically-grounded efficient estimation of important quantities in causal inference such as (L)QTEs when we must control for many covariates and/or flexible relationships, as we demonstrate in empirical studies.

Keywords: Causal Inference, Neyman Orthogonality, Cross-fitting, Instrumental Variables, Conditional Value at Risk, Expectiles

1. Introduction

In this paper, we consider estimating parameters $\theta^* = (\theta_1^*,\, \theta_2^*) \in \Theta = \Theta_1 \times \Theta_2 \subseteq \mathbb{R}^d$ defined as the (unique) solution to the following d-dimensional estimating equation:

$$\mathbb{P}\psi(Z; \theta, \eta_1^*(Z; \theta_1), \eta_2^*(Z)) = \mathbf{0}, \quad \theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2, \tag{1}$$

where $Z \in \mathcal{Z}$ are observed random variables, $\eta_1^*(Z; \theta_1)$ and $\eta_2^*(Z)$ are two unknown nuisance functions, and $\mathbf{0}$ is the zero vector in $\mathbb{R}^d$. We hope to estimate θ^* based on $(Z_1, \ldots, Z_N)$, N independent and identically distributed (iid) draws from the distribution $\mathbb{P}$. As we will show, estimating equations of the form above are prevalent in efficient estimation in causal inference and missing data problems, with quantile treatment effect (QTE) estimation (Section 1.1) as a prominent example, among many others (Section 1.2).

One important feature of Eq. (1) is that the nuisance $\eta_1^*(Z; \theta_1)$ involves the parameters to be estimated as an input. This *parameter-dependent* nuisance raises several challenges and causes existing methods to be unstable and computationally burdensome. Specifically, we could potentially use the observed data to estimate the nuisances $\eta_1^*(Z; \theta_1)$ and $\eta_2^*(Z)$ and then solve a sample analogue of Eq. (1) based on the estimated nuisances in order to estimate θ^*, possibly using cross-fitting (Robins et al., 2008; Zheng and van der Laan, 2011; Chernozhukov et al., 2018a). However, this requires estimating the nuisance $\eta_1^*(Z; \theta_1)$ for *all* possible $\theta_1 \in \Theta_1$, *i.e.*, learning *infinitely* many functions of Z, and then solving for the root of an estimated function. For example, when estimating QTE (see Section 1.1), this involves estimating a *whole* conditional cumulative distribution function, or equivalently, *infinitely* many binary probability regressions. This can be very unstable, especially in causal inference with observational data where typically a large number of covariates need to be conditioned on to remove confounding. Although one may discretize the space of Θ_1 and estimate $\eta_1^*(Z; \theta_1)$ only for finitely many θ_1, this can still be computationally burdensome when the discrete grid is large, and the resulting estimator can be sensitive to the discretization scheme.

In this paper, we propose a localized debiased machine learning (LDML) approach that only requires estimating $\eta_1^*(Z; \theta_1)$ at a *single* θ_1 value, without estimating it for all possible values or ad-hoc discretized values of θ_1. Importantly, our estimator is asymptotically equivalent to an oracle estimator that knows the whole continuum of nuisance function $\eta_1^*(Z; \theta_1)$ for all $\theta_1 \in \Theta_1$. In other words, asymptotically, our method does not incur any loss even though it only estimates the nuisance function at a single θ_1 value. In the case of QTE (and other parameters in Section 1.2), the resulting asymptotic variance coincides with the corresponding semiparametric efficiency bound. Moreover, estimating this far simpler nuisance reduces to standard classification and regression tasks, *i.e.*, fitting conditional expectations (regression) and conditional binary probabilities (classification), for which many machine learning methods exist. In particular, our approach will be shown to be largely insensitive to *how* these conditional expectation functions are estimated, so we may directly use off-the-shelf machine learning methods and treat them as black-box regression or classification algorithms (*e.g.*, random forests, gradient boosting, neural networks). Therefore, our proposed method notably enables practical and efficient estimation using time-tested machine learning methods to solve Eq. (1).

In comparison, existing approaches for debiased and efficient estimation with black-box nuisance estimators either focus on settings where nuisances do not depend on the target parameters (*i.e.*, $\eta_1^*(Z;\theta_1)$ does not appear) or treat nuisances as abstract objects so that one must estimate a continuum of nuisances (*i.e.*, estimate $\eta_1^*(Z;\theta_1)$ for all $\theta_1 \in \Theta_1$) when applying to Eq. (1) thus precluding the use of standard machine-learning algorithms (e.g., Tsiatis, 2006; Robins et al., 2008; Zheng and van der Laan, 2011; Robins et al., 2013; Chernozhukov et al., 2018a; Bravo et al., 2020). Similarly, existing works specifically on the efficient estimation of QTEs either apply similar debiased approaches using a continuum of nuisances (Belloni et al., 2017; Díaz, 2017) or use specific non-black-box nuisance estimators like polynomial sieves and local polynomial kernel regression and make explicit smoothness restrictions (Firpo, 2007; Frölich and Melly, 2013). See an extensive literature review in Section 7. Compared to these works, our proposal is fully generic, flexible, and machine-learning driven in that it handles many important examples that fit into Eq. (1), as we review in the next two subsections. Specifically, our method only requires estimating $\eta_1^*(Z;\theta_1)$ for a single θ_1 and doing so only at slow, nonparametric rates. This involves estimating a conditional expectation function in all of our examples, which can be implemented by standard regression and classification machine-learning methods. Our method can be seen as a variant of DML applied to an alternative *localized* estimating equation that we prove asymptotically equivalent to the original Eq. (1). In particular, we develop special *localized* estimators for the new *localized* nuisance that appears in the *localized* equation. Compared to the analysis of DML which we build upon (Chernozhukov et al., 2018a), our asymptotic analysis requires a more careful handling of nonlinear estimating equations and relaxing nuisance rate conditions. Our proposed method applied to estimating (L)QTEs has been implemented in the DoubleML python package (Bach et al., 2022).

1.1 Motivating Example: Quantile Treatment Effects

A primary motivation of considering Eq. (1) is the estimation of QTE. In this case, we consider a population of units, each associated with some baseline covariates $X \in \mathcal{X}$, two potential outcomes $Y(0), Y(1) \in \mathbb{R}$ for each of two possible treatments, and a treatment indicator $T \in \{0,1\}$. We are interested in the γ-quantile of $Y(1)$: the θ_1^* such that $\mathbb{P}(Y(1) \leq \theta_1^*) = \gamma$ (assuming existence and uniqueness) for $\gamma \in (0,1)$. And, similarly, we are interested in the quantile of $Y(0)$ and in the difference of the quantiles, known as the quantile treatment effect (QTE), but these estimation questions are analogous so for brevity we focus just on θ_1^*, the γ-quantile of $Y(1)$. Compared to the average outcome and the average treatment effect (ATE), the quantile of outcomes and the QTE provide a more robust assessment of the effects of treatment and are very important quantities in program evaluation.

We do not observe the potential outcomes but instead only the realized factual outcome corresponding to the assigned treatment, $Y = Y(T)$. Hence, we only observe $Z = (X, T, Y)$. Ignorable treatment assignment with respect to X assumes that $Y(1) \perp\!\!\!\perp T \mid X$ (*i.e.*, no unobserved confounders) and overlap assumes that $\mathbb{P}(T = 1 \mid X) > 0$, and these together ensure that θ_1^* is identifiable from observations of Z. Specifically, a straightforward identification is given by the so-called inverse propensity weighting (IPW) equation:

$$\begin{aligned} &\mathbb{P}\psi^{\mathrm{IPW}}(Z;\theta_1^*,\eta_2^*(Z))=0, \qquad (2)\\ &\text{where} \quad \psi^{\mathrm{IPW}}(Z;\theta_1,\eta_2(Z))=\mathbb{I}\left[T=1\right]\mathbb{I}\left[Y\le\theta_1\right]/\eta_2(Z)-\gamma, \quad \eta_2^*(Z)=\mathbb{P}\left(T=1\mid X\right). \end{aligned}$$

Here estimating the propensity score function η_2^* amounts to learning a conditional probability function from a binary response, for which many standard machine learning methods exist. Once we construct an estimator $\hat\eta_2$, we can obtain the standard IPW estimator $\hat\theta_1^{\mathrm{IPW}}$ by solving $\frac{1}{N}\sum_{i=1}^N \psi^{\mathrm{IPW}}(Z_i;\theta_1,\hat\eta_2(Z_i))=0$. Generally, the error of the IPW estimator can heavily depend on the particular method used to construct $\hat\eta_2$ and its convergence rate can be slowed down by that of $\hat\eta_2$, prohibiting the use of general nonparametric machine learning methods and potentially leading to unstable estimates.

Instead, one can alternatively obtain the following estimating equation from the efficient influence function for θ^* (Tsiatis, 2006):

$$\begin{aligned} &\mathbb{P}\psi(Z;\theta_1^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))=0, \qquad (3)\\ &\text{where} \quad \psi(Z;\theta_1,\eta_1(Z;\theta_1),\eta_2(Z))=\mathbb{I}\left[T=1\right]\left(\mathbb{I}\left[Y\le\theta_1\right]-\eta_1(Z;\theta_1)\right)/\eta_2(Z)+\eta_1(Z;\theta_1)-\gamma,\\ &\quad \eta_1^*(Z;\theta_1)=\mathbb{P}\left(Y\le\theta_1\mid X,T=1\right). \end{aligned}$$

An important feature of the above is that it satisfies a property known as *Neyman orthogonality*: the moment $\mathbb{P}\psi(Z;\theta_1,\eta_1(Z;\theta_1),\eta_2(Z))$ has *zero* derivatives with respect to the nuisances at $\theta_1^*,\eta_1^*,\eta_2^*$. This means that the estimating equation is robust to small perturbations in the nuisances so that estimation errors therein contribute only to higher-order error terms in the final estimate of θ_1^*. Neyman orthogonality is leveraged in many works on semiparametric inference, such as Robins et al. (1994b); Robins and Rotnitzky (1995); van der Laan and Robins (2003); van der Laan and Rose (2011) (see a more detailed review in Section 7). In particular, Chernozhukov et al. (2018a) recently proposed a debiased machine learning (DML) approach that also builds on Neyman orthogonality. Their approach is as follows: split the data randomly into K folds, $\mathcal{D}_1,\dots,\mathcal{D}_K$, and then for each $k=1,\dots,K$, use all but the k^{th} fold to construct nuisance estimates $\hat\eta_1^{(k)}$, $\hat\eta_2^{(k)}$, and finally solve the empirical estimating equation $\frac{1}{N}\sum_{k=1}^K\sum_{i\in\mathcal{D}_k}\psi(Z_i;\theta_1,\hat\eta_1^{(k)}(Z_i;\theta_1),\hat\eta_2^{(k)}(Z_i))=0$ to obtain the estimator $\hat\theta$. They prove that as long as the estimates $\hat\eta_1^{(k)}$, $\hat\eta_2^{(k)}$ converge to η_1^*, η_2^* faster than $N^{-1/4}$, the estimate $\hat\theta_1$ will have similar behavior to the *oracle* estimate that solves $\frac{1}{N}\sum_{i=1}^N\psi(Z_i;\theta_1,\eta_1^*(Z_i;\theta_1),\eta_2^*(Z_i))=0$, *i.e.*, the empirical estimating equation using the *true* nuisance functions. As a result, the estimate $\hat\theta_1$ is asymptotically normal and semiparametrically *efficient*. Since, apart from the mild rate requirement on $\hat\eta_1^{(k)}$, $\hat\eta_2^{(k)}$, no metric entropy conditions are assumed, this allows one to successfully use machine learning methods to learn nuisances and achieve asymptotically normal and efficient estimation.

The problem with this approach for estimating quantiles of outcomes (similarly, QTEs), however, is that it requires the estimation of a very complex nuisance function: $\eta_1^*(Z;\theta_1)$ is the *whole* conditional cumulative distribution function of a real-valued outcome, potentially conditioned on high-dimensional covariates. While certainly nonparametric methods for estimating conditional distributions exist such as kernel estimators, this learning problem is *much harder* to do in a flexible, blackbox, machine-learning manner, compared to just estimating a single regression function. This indeed stands in stark contrast to the estimation of average treatment effect (ATE), where applying DML requires a far simpler nuisance function given by the regression of outcome on covariates and treatment, $\mathbb{E}[Y \mid X, T]$, for which a long list of practice-proven machine learning methods can be directly and successfully applied. The key difference is that the nuisance function in ATE estimation does *not* depend on the target parameter and can therefore be estimated in an independent manner whereas the nuisance function in QTE estimation *does* depend on the target parameter. This issue makes DML, despite its theoretical benefits, untenable in practice for the important task of QTE estimation.

1.2 Estimating Equations with Incomplete Data

More generally, we can consider parameters $(\theta_1^*, \theta_2^*) \in \Theta_1 \times \Theta_2$ defined as the solution to the following estimating equation on the (unavailable) complete data:

$$\mathbb{P}[U(Y(1); \theta_1) + V(\theta_2)] = 0, \tag{4}$$

for some given functions $U(y; \theta_1)$ and $V(\theta_2)$. Below we provide some concrete examples.

Example 1 (Quantile of Potential Outcome)**.** In Section 1.1, we consider the quantile θ_1^* defined as $\mathbb{P}(Y(1) \le \theta_1^*) = \gamma$ for $\gamma \in (0, 1)$. This corresponds to Eq. (4) with

$$U(y; \theta_1) = \mathbb{I}[Y(1) \le \theta_1] - \gamma, \quad V(\theta_2) = 0.$$

Example 2 (CVaR of Potential Outcome)**.** Another example is conditional value at risk (CVaR) $\theta_2^* = \mathbb{P}[Y(1)\mathbb{I}[F_1(Y(1)) \ge \gamma]]/(1-\gamma)$, where F_1 is the cumulative distribution function of $Y(1)$. This gives the expectation of $Y(1)$ conditioned on being above the γ-quantile (again, assuming uniqueness). CVaR is also known as expected shortfall, a popular risk measure in risk management and optimization (Rockafellar and Uryasev, 2002). Letting

$$U(y; \theta_1) = \Big(\mathbb{I}[y \le \theta_1], \ \max\{\theta_1, (1-\gamma)^{-1}(y - \gamma\theta_1)\}\Big), \quad V(\theta_2) = (-\gamma, \ -\theta_2), \tag{5}$$

Eq. (4) defines (θ_1^*, θ_2^*) as the quantile and CVaR of $Y(1)$.

Example 3 (Expectile of Potential Outcome)**.** Yet another example is the expectile, a measure for asymmetric risk (Newey and Powell, 1987). The γ-expectile of $Y(1)$ is defined by the following asymmetric least squares problem:

$$\theta_1^* = \operatorname*{argmin}_{\theta_1 \in \mathbb{R}} \mathbb{P}\left[|\gamma - \mathbb{I}(Y(1) - \theta_1 \le 0)| (Y(1) - \theta_1)^2\right].$$

Its first-order condition corresponds to Eq. (4) with

$$U(y; \theta_1) = (1-\gamma)(Y(1) - \theta_1) - (1 - 2\gamma)\max(Y(1) - \theta_1, 0), \quad V(\theta_2) = 0. \tag{6}$$

We cannot directly use estimating equations above for estimation since they depend on the counterfactual outcome $Y(1)$ that is partially observed. Under ignorable treatment assignment and overlap, we could also derive an IPW estimating equation akin to Equation (2).

However, the IPW equation is not Neyman-orthogonal so it is amenable to general machine learning nuisance estimation. Instead, we can consider the following general-purpose efficient and Neyman orthogonal estimating equation (Tsiatis, 2006, Theorems 10.1 and 10.2) for the estimand (θ_1^*, θ_2^*) defined by Eq. (4):

$$\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z)) = \frac{\mathbb{I}\left[T=1\right]}{\eta_2^*(Z)}\Big(U(Y;\theta_1)-\eta_1^*(Z;\theta_1)\Big)+\eta_1^*(Z;\theta_1)+V(\theta_2), \tag{7}$$

$$\text{where}\quad \eta_1^*(Z;\theta_1)=\mathbb{E}\left[U(Y;\theta_1)\mid X, T=1\right],\quad \eta_2^*(Z)=\mathbb{P}\left(T=1\mid X\right).$$

This orthogonal estimating equation again involves a nuisance $\eta_1^*(Z;\theta_1)$ that involves the target parameter as input. This occurs for all examples above, whether estimating quantiles, CVaR, or expectiles, and more generally, whenever $U(y;\theta_1)$ is not linear in θ_1. And, in such cases, learning $\eta_1^*(Z;\theta_1)$ for *all* θ_1 is practically difficult, which may involve learning a whole conditional distribution function or a whole continuum of conditional expectation functions given potentially high-dimensional covariates.

In the above examples, we consider parameters in terms of the counterfactual outcome $Y(1)$. We can similarly consider parameters of $Y(0)$ and their differences as the treatment effects. In Appendix A, we further consider the setting where the the treatment assignment is not ignorable but a binary instrumental variable (IV) is available. In this case, we focus on parameters defined for the *complier* subpopulation, such as the local QTE (LQTE). We present the efficient estimating equations for these local parameters and show they also satisfy Neyman orthogonality and also involve some estimand-dependent nuisance functions $\eta_1^*(Z;\theta_1)$. Thus the same technical challenge also appears in the IV setting.

1.3 Main Idea: Localization

The primary goal of this paper can be understood as extending DML to effectively tackle the case where nuisances depend on the target parameters. In particular, we propose a technique called *localization* to alleviate this dependence. This will enable efficient estimation of important quantities such as QTEs in the presence of high-dimensional nuisances by using and debiasing black-box machine learning methods for the standard regression task.

Specifically, we will show that under a condition (Assumption 1) that holds for all of our examples, we can equivalently consider solving the orthogonal estimating equation

$$\mathbb{P}\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))=0, \tag{8}$$

where the function ψ is given in Equation (7). Notably, Equation (8) involves $\eta_1^*(Z;\theta_1)$ at the *single* value $\theta_1=\theta_1^*$, as opposed to the infinitely many possible values for θ_1. This formulation considerably reduces the need of nuisance estimation: now we only need to estimate $\eta_1^*(Z;\theta_1^*)$ and $\eta_2^*(Z)$, both functions only of Z but not of θ_1. This enables us to apply our localization technique.

The basic idea of localization as it applies to the estimation of the quantile of outcomes is as follows. While perhaps inefficient, the estimator $\hat{\theta}_1^{\text{IPW}}$ based on the IPW estimating equation in Equation (2) relies only on estimating a binary regression η_1^*. This is amenable to machine learning approaches but may have a slow convergence rate in general. Despite its slow rate, this rough initial guess can sufficiently localize our nuisance estimation. We consider estimating $\eta_1^*(Z; \hat{\theta}_1^{\mathrm{IPW}})$ as an proxy for $\eta_1^*(Z; \theta_1^*)$, since $\hat{\theta}_1^{\mathrm{IPW}}$ provides an initial estimate of θ_1^*. For estimating the quantiles, this means we only have to regress the binary response $\mathbb{I}[Y \le \hat{\theta}_1^{\mathrm{IPW}}]$ on X, treating $\hat{\theta}_1^{\mathrm{IPW}}$ as fixed. In particular, we propose a special three-way data splitting procedure that debiases such plug-in nuisance estimates in order to obtain an estimate for θ^* with near-oracle performance. In the rest of this paper, we will further generalize this technique to all parameters and also thoroughly analyze its asymptotic properties.

Our proposal can be viewed as an application of the DML framework by Chernozhukov et al. (2018a) for nonlinear orthogonal estimating equations to Eq. (8) with nuisances $\eta_1^*(Z; \theta_1^*)$ and $\eta_2^*(Z)$. Our paper contributes to explicitly characterizing the challenge of parameter-dependent nuisances when estimating important causal parameters such as (L)QTEs. This complements Chernozhukov et al. (2018a) as all of their examples are for linear estimating equations and their treatment to nonlinear estimating equations focuses on abstract nuisances. Our paper proposes a localization technique and a new data splitting procedure to practically estimate $\eta_1^*(Z; \theta_1^*)$ and implement DML for Eq. (8). Importantly, we rigorously establish when it suffices to focus on Eq. (8), provide thorough asymptotic analysis of the proposed approach, and provide an asymptotically valid general variance estimator and confidence interval. Notably, our asymptotic analysis is based on a different proof than Chernozhukov et al. (2018a) so that we permit more flexible rate conditions on the nuisance estimation (see Appendix F for a detailed discussion). Moreover, our theoretical guarantees for variance estimation and confidence interval also complement Chernozhukov et al. (2018a), since they only provide such guarantees for linear estimating equations.

Notation. We let d_1, d_2 be the dimensions of θ_1^*, θ_2^*, respectively, where $d_1 + d_2 = d$. For $f : \mathbb{R}^d \to \mathbb{R}^m$, $\partial_{\theta^\top} f(\theta)$ is the $m \times d$-matrix-valued function with entry $\frac{\partial f_i(\theta)}{\partial \theta_j}$ in position (i, j) and $\partial_{\theta^\top} f(\theta)|_{\theta=\theta_0}$ is its evaluation at θ_0. For $g : \mathbb{R}^d \to \mathbb{R}$, $\partial_\theta \partial_{\theta^\top} g(\theta)$ is the $d \times d$-matrix-valued function with entry $\frac{\partial g(\theta)}{\partial \theta_i \partial \theta_j}$ in position (i, j). We use $\sigma_{\max}(\partial_\theta \partial_{\theta^\top} g(\theta))$ to denote its largest singular value. We let $\mathbb{P}(Z \in A)$ and $\mathbb{E}[Z \mid Z \in A]$ for measurable sets A denote probabilities and expectations with respect to $\mathbb{P}$. We let $\mathbb{P}f(Z) = \int f \, \mathrm{d}\mathbb{P}$ for measurable functions f denote expectations with respect to Z alone, while we let $\mathbb{E}f(Z; Z_1, \ldots, Z_n)$ denote expectations with respect to Z *and* the data. Thus, if $\hat{\varphi}$ depends on the data, $\mathbb{P}f(Z; \hat{\varphi})$ remains a function of the data while $\mathbb{E}f(Z; \hat{\varphi})$ is a number. We let $\mathbb{P}_N$ denote the empirical expectation: $\mathbb{P}_N f(Z) = \frac{1}{N}\sum_{i=1}^N f(Z_i)$ for any measurable function f. Moreover, for vector-valued function $f(Z) = (f_1(Z), \ldots, f_d(Z))$, we let $\mathbb{P}f^2(Z) := (\mathbb{P}f_1^2(Z), \ldots, \mathbb{P}f_d^2(Z))$. For any $x \in \mathbb{R}^d$, we denote the open ball centered at x with radius δ as $\mathcal{B}(x; \delta)$. For $p > 0$ and a probability measure $\mathbb{Q}$, we denote $\|f\|_{\mathbb{Q},p} = \left(\int |f|^p \, \mathrm{d}\mathbb{Q}\right)^{1/p}$. For a set of functions $\mathcal{F}$, we define the covering number $N(\epsilon, \mathcal{F}, \|\cdot\|_{\mathbb{Q},2})$ as the minimal number N of functions $f_1, \ldots, f_N$ such that $\sup_{f \in \mathcal{F}} \inf_{i=1,\ldots,N} \|f - f_i\|_{\mathbb{Q},2} \le \epsilon$. For positive deterministic sequence a_n and random variable sequence X_n, $X_n = o_{\mathbb{P}}(a_n)$ means $\mathbb{P}(|X_n|/a_n > \epsilon) \to 0 \, \forall \epsilon > 0$ and $X_n = O_{\mathbb{P}}(a_n)$ means for any $\epsilon > 0$, there exists $M > 0$ such that $\limsup_{n \to \infty} \mathbb{P}(|X_n|/a_n \ge M) \le \epsilon$.

2. Method

We next present our methodology, first motivating the localization technique, and then explicitly stating our meta-algorithm.

2.1 Motivation

Ideally, if the nuisances η_1^* and η_2^* were both known, then Eq. (1) suggests that θ^* could be estimated by solving the following estimating equation:

$$\mathbb{P}_N\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right]=0. \tag{9}$$

Under standard regularity conditions for Z-estimation (van der Vaart, 1998), the resulting oracle estimator $\tilde{\theta}$ that solves Eq. (9) is asymptotically linear (and hence $\sqrt{N}$-consistent and asymptotically normal):

$$\sqrt{N}(\tilde{\theta}-\theta^*)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}-{J^*}^{-1}\psi(Z_i;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i))+o_{\mathbb{P}}(1), \tag{10}$$
$$\text{where}\quad J^*=\partial_{\theta^\top}\left\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right]\right\}|_{\theta=\theta^*}.$$

Furthermore, if ${J^*}^{-1}\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ is the semiparametrically efficient influence function for θ^*, then $\tilde{\theta}$ also achieves the efficiency lower bound, that is, has minimal asymptotic variance among all regular estimators (van der Vaart, 1998).

Since η_1^* and η_2^* are actually unknown, the oracle estimator $\tilde{\theta}$ is of course infeasible. Instead, we must estimate the nuisance functions. A direct application of DML would require us to learn the whole functions η_1^* and η_2^*. That is, in order to solve Eq. (9) we would need to estimate infinitely many nuisance functions, $H_1=\{\eta_1^*(\cdot,\theta_1):\theta_1\in\Theta_1\}$.

To avoid the daunting task of estimating infinitely many nuisances, we will instead attempt to target the following alternative oracle estimating equation

$$\mathbb{P}_N\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]=0. \tag{11}$$

Although Eq. (11) appears very similar to Eq. (9), it involves $\eta_1^*(Z;\theta_1)$ only at the *single* value $\theta_1=\theta_1^*$. In other words, among the whole family of nuisances H_1, *only* $\eta_1^*(Z;\theta_1^*)\in H_1$ is relevant for Eq. (11).

The (infeasible) estimators that solve each of Eqs. (9) and (11) have the same leading asymptotic behavior as long as the respective associated Jacobian matrices coincide, as posited by the following assumption.

Assumption 1 (Invariant Jacobian)**.** $\partial_{\theta^\top}\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}=J^*$.

Assumption 1 means that solving Eq. (9) or (11) will have the same asymptotic behavior. Both, however, are infeasible since they involve unknown nuisances. Nonetheless, Eq. (11) motivates our new algorithm, which eschews estimating $H_1=\{\eta_1^*(\cdot\,;\theta_1):\theta_1\in\Theta_1\}$ in full.

It it easy to verify that Assumption 1 holds for estimating equations with incomplete data (Section 1.2), which includes QTE estimation. In particular, the estimating equation ψ in

Eq. (7) satisfies that

$$\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right] = \mathbb{P}\left[\frac{\mathbb{I}\left[T=1\right]}{\eta_2^*(Z)}U(Y;\theta_1) - \frac{\mathbb{I}\left[T=1\right]-\eta_2^*(Z)}{\eta_2^*(Z)}\eta_1^*(Z;\theta_1)\right] + V(\theta_2)$$
$$= \mathbb{P}\left[U\left(Y(1);\theta_1\right)\right] + V(\theta_2), \tag{12}$$

which does not depend on $\eta_1^*(Z;\theta_1)$ at all. Thus whether fixing $\eta_1^*(Z;\theta_1)$ at $\theta_1=\theta_1^*$ or not, the Jacobian matrix of the estimating equation remains the same.

More generally, a sufficient condition for Assumption 1 is the below Fréchet-derivative orthogonality condition.

Proposition 1 (Sufficient Conditions for Invariant Jacobian)**.** *Assume that the map* $(\theta,\eta_1(\cdot;\theta_1'))\mapsto \mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2^*(Z))\right]$ *is Fréchet differentiable at* $(\theta^*,\eta_1^*(\cdot,\theta_1^*))$*. Namely, assume that there exists a bounded linear operator* $\mathcal{D}_{\eta_1^*}$*, such that for any* $(\theta,\eta_1(\cdot,\theta_1'))$ *within a small open neighborhood* $\mathcal{N}$ *around* $(\theta^*,\eta_1^*(\cdot,\theta_1^*))$*,*

$$\begin{aligned}&\|\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z,\theta_1'),\eta_2^*(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\\ &\qquad - \partial_{\theta^\top}\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}(\theta-\theta^*) - \mathcal{D}_{\eta_1^*}[\eta_1(\cdot,\theta_1')-\eta_1^*(\cdot,\theta_1^*)]\|\\ &= o(\|\theta-\theta^*\|) + o(\{\mathbb{P}\left[\eta_1(Z,\theta_1')-\eta_1^*(Z;\theta_1^*)\right]^2\}^{1/2}).\end{aligned}$$

Assume further that there exists $C>0$ *such that for any* $(\theta,\eta_1(\cdot,\theta_1'))\in\mathcal{N}$

$$\mathcal{D}_{\eta_1^*}[\eta_1(\cdot,\theta_1')-\eta_1^*(\cdot,\theta_1^*)] = 0, \tag{13}$$
$$\mathbb{P}\left[\left\|\eta_1^*(Z,\theta_1')-\eta_1^*(Z;\theta_1^*)\right\|^2\right]^{1/2} \le C\|\theta_1'-\theta_1^*\|.$$

Then Assumption 1 is satisfied.

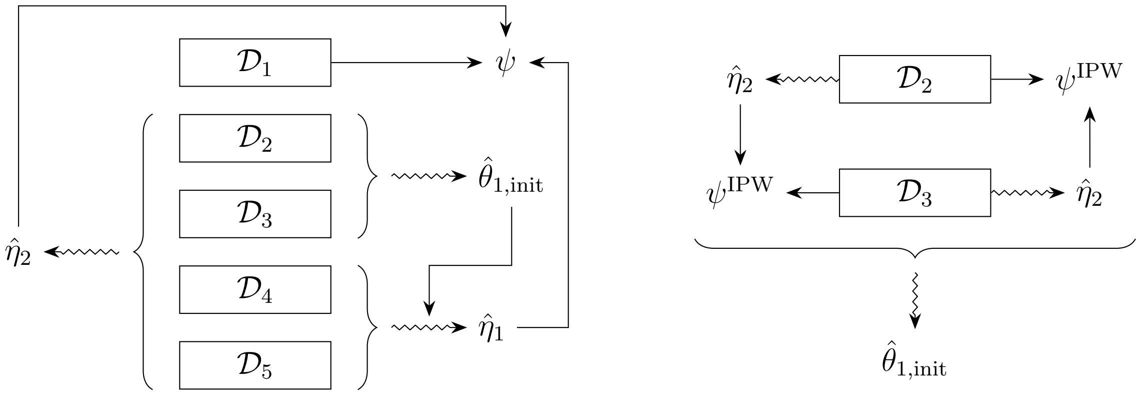
Equation (13) is an orthogonality condition using the Fréchet derivative, which is stronger than the Gâteaux derivative in Neyman orthogonality (Assumption 2 condition vii.). This condition is automatically satisfied for the efficient estimating equation with incomplete data because, following Equation (12), we have that $\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2^*(Z))\right]$ does not depend on η_1 at all, so that its Fréchet derivative with respect to η_1 trivially exists and is always 0. Therefore, all of our examples in Section 1.2 satisfy our Assumption 1 and are therefore amenable to our localization approach. This can be understood as a consequence of double robustness and parallels how double robustness yields the usual Gâteaux-derivative Neyman orthogonality of Chernozhukov et al. (2018a).

2.2 The LDML Meta-Algorithm

Motivated by the new (infeasible) estimating equation in Eq. (11), we propose to estimate θ^* by the following (feasible) three-way sample splitting method, which we term localized debiased machine learning (LDML). The algorithm has two parts: three-way-cross-fold nuisance estimation and solving the estimating equation.

We start by discussing how we estimate the nuisances that we will then plug into Eq. (11).

Definition 1 (3-way-cross-fold nuisance estimation)**.** Fix integers $K\ge 3$, $K'\in[1,K-2]$.



(a) Iteration $k = 1$ of step 2 of Definition 1 in the LDML meta-algorithm with $K = 5, K' = 2$.

(b) In some cases, cross-fitted IPW (Definition 4) can be used for $\hat\theta_{1,\text{init}}$.

Figure 1: Sketch of the LDML estimation procedure and a possible initial guess estimator. Squiggly arrows "$\rightsquigarrow$" denote estimation. Plain arrows "$\rightarrow$" denote plugging in.

1. Randomly permute the data indices and let $\mathcal{D}_k = \{\lceil (k-1)N/K\rceil+1, \dots, \lceil kN/K\rceil\}$, $k = 1, \dots, K$ be a random even K-fold split of the data.
2. For $k = 1, \dots, K$:
   (a) Set $\mathcal{H}_{k,1} = \{1, \dots, K'+\mathbb{I}[k \le K']\}\backslash\{k\}$, $\mathcal{H}_{k,2} = \{K'+\mathbb{I}[k \le K']+1, \dots, K\}\backslash\{k\}$.
   (b) Use only $\mathcal{D}_k^{C,1} = \left\{Z_i : i \in \bigcup_{k'\in\mathcal{H}_{k,1}} \mathcal{D}_{k'}\right\}$ to construct an initial estimator $\hat\theta_{1,\text{init}}^{(k)}$ of θ_1^*. Use only $\mathcal{D}_k^{C,2} = \left\{Z_i : i \in \bigcup_{k'\in\mathcal{H}_{k,2}} \mathcal{D}_{k'}\right\}$ to construct estimator $\hat\eta_1^{(k)}(\cdot\,; \hat\theta_{1,\text{init}}^{(k)})$ of $\eta_1^*(\cdot\,; \hat\theta_{1,\text{init}}^{(k)})$. Use only $\mathcal{D}_k^{C,1} \cup \mathcal{D}_k^{C,2}$ to construct estimator $\hat\eta_2^{(k)}$ of η_2^*.

For illustration the first iteration of step 2 above is sketched in Fig. 1(a) along with the plugging of estimated nuisances into the estimating equation (see Definitions 2 and 5).

Notice that since $\mathcal{D}_k^{C,1}$ and $\mathcal{D}_k^{C,2}$ are disjoint, $\eta_1^*(\cdot\,; \hat\theta_{1,\text{init}}^{(k)})$ is a fixed, nonrandom function with respect to the data $\mathcal{D}_k^{C,2}$. That is, the η_1^* nuisance estimation task in step 2b appears as estimating a *single* $\eta_1^*(\cdot\,; \theta_1') \in H_1$ for $\theta_1' = \hat\theta_{1,\text{init}}^{(k)}$ rather than the estimation of *all* of H_1.

A natural question is, what might be a reasonable initial estimator. In the examples in Sections 1.1 and 1.2, we can use an IPW estimate for $\hat\theta_{1,\text{init}}^{(k)}$ (see Fig. 1(b) and Definition 4).

Given these nuisance estimates, we can obtain the LDML estimator for θ^* by approximately solving the average of the estimate of Eq. (11) in each fold.

Definition 2 (LDML)**.** We let the estimator $\hat\theta$ be given by (approximately) solving

$$\overline{\Psi}(\theta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k} \psi(Z_i; \theta, \hat\eta_1^{(k)}(Z_i; \hat\theta_{1,\text{init}}^{(k)}), \hat\eta_2^{(k)}(Z_i)) = 0. \tag{14}$$

In fact, we allow for an approximate least-squares solution, which is useful if the empirical estimating equation has no exact solution. Namely, we let $\hat{\theta}$ be any satisfying

$$\|\overline{\Psi}(\hat{\theta})\| \leq \inf_{\theta\in\Theta} \|\overline{\Psi}(\theta)\| + \varepsilon_N, \quad \text{for an approximation error } \varepsilon_N. \tag{15}$$

In Appendix D Definition 5, we give an alternative LDML estimator obtained by averaging solutions to Eq. (11) estimated in each fold separately. These two LDML estimators are asymptotically equivalent and all results in this paper apply to both, thus we focus on Definition 2 in the main text. Moreover, both estimators depend on the random splitting in Definition 1. To reduce the variance from this, we may aggregate estimates from multiple different sample splitting realizations. See Appendix E for a detailed discussion.

3. Theoretical Analysis

In this section, we provide the sufficient conditions that guarantee the proposed estimator $\hat{\theta}$ in Definition 2 to be consistent and asymptotically normal. In particular, although the proposed estimator relies on plug-in nuisance estimators, it is asymptotically equivalent to the *infeasible* estimator based on Eq. (9) with *true* nuisances, that is, it satisfies Eq. (10). While some of our conditions are analogous to those in Chernozhukov et al. (2018a), some are not and our proof takes a different approach that enables weaker conditions for convergence rates of the nuisance estimators.

Our asymptotic normality results may be stated uniformly over a sequence of models $\mathcal{P}_N$ for any data generating distribution $\mathbb{P} \in \mathcal{P}_N$. Our first set of assumptions ensure that θ^* is reasonably identified by the given estimating equation for all $\mathbb{P} \in \mathcal{P}_N$. We also assume that our estimating equation satisfies the Neyman orthogonality condition with respect to a nuisance realization set $\mathcal{T}_N \subset [\mathcal{Z} \to \mathbb{R}]^2$ that contains the nuisance estimates $\hat{\eta}_1(\cdot\,;\hat{\theta}_{1,\text{init}})$ and $\hat{\eta}_2(\cdot)$ with high probability (Assumption 3 condition i.). Note the set $\mathcal{T}_N$ consists of pairs of functions of the data Z alone and *not* θ_1. Therefore, we denote members of the set as $(\eta_1(\cdot\,;\theta_1'), \eta_2(\cdot)) \in \mathcal{T}_N$, where $\eta_1(\cdot\,;\theta_1')$ is simply understood as a symbol representing of some fixed function of Z alone.

Assumption 2 (Regularity of Estimating Equations)**.** Assume there exist positive constants c_1 to c_7 such that the following conditions hold for all $\mathbb{P} \in \mathcal{P}_N$:

i. Θ is a compact set and it contains a ball of radius $c_1 N^{-1/2}\log N$ centered at θ^*.

ii. The map $(\theta, a, b) \mapsto \mathbb{P}\left[\psi(Z;\theta,a,b)\right]$ is twice continuously Gâteaux-differentiable.

iii. For any $\theta \in \Theta$, $2\|\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\| \geq \|J^*(\theta-\theta^*)\| \wedge c_2$.

iv. J^* is non-singular with singular values bounded between positive constants c_3 and c_4.

v. Singular values of the covariance matrix Σ are bounded between constants c_5 and c_6:

$$\Sigma := \mathbb{E}\left[J^{*-1}\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))^\top J^{*-\top}\right]. \tag{16}$$

vi. The nuisance realization set $\mathcal{T}_N$ contains the true nuisance parameters $(\eta_1^*(\cdot\,;\theta_1^*),\eta_2^*(\cdot))$. Moreover, the parameter space Θ is bounded and for each $(\eta_1(\cdot\,;\theta_1'),\eta_2(\cdot))\in\mathcal{T}_N$, the function class $\mathcal{F}_{\eta,\theta_1'}=\{Z\mapsto\psi_j(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z)) : j=1,\dots,d,\theta\in\Theta\}$ is suitably measurable and its uniform covering entropy satisfies the following condition: for positive constants a, v, and $q>2$, $\sup_{\mathbb{Q}}\log N(\epsilon\|F_{\eta,\theta_1'}\|_{\mathbb{Q},2},\mathcal{F}_{\eta,\theta_1'},\|\cdot\|_{\mathbb{Q},2})\le v\log(a\epsilon)$ $\forall\epsilon\in(0,1]$, where $F_{\eta,\theta_1'}$ is a measurable envelope for $\mathcal{F}_{\eta,\theta_1'}$ that satisfies $\|F_{\eta,\theta_1'}\|_{\mathbb{P},q}\le c_7$.

vii. $\partial_r\left\{\mathbb{P}\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*)+r(\eta_1(Z;\theta_1')-\eta_1^*(Z;\theta_1^*)),\eta_2^*(Z)+r(\eta_2(Z)-\eta_2^*(Z))\right\}\big|_{r=0}=0$ for all $(\eta_1(\cdot\,;\theta_1'),\eta_2(\cdot))\in\mathcal{T}_N$.

Assumption 2 conditions i.–v. constitute standard identification and regularity conditions for Z-estimation (with uniform guarantees; see also Remark 1 below). Assumption 2 condition vi. requires that ψ is a well-estimable function of θ for any *fixed* set of nuisances. Importantly, while it imposes a metric entropy condition on ψ, this condition does *not* impose metric entropy conditions on our nuisance estimators, so flexible machine learning nuisance estimators are allowed. This assumption is very mild as Θ is finite-dimensional, so it can be ensured by some continuity and compactness condition via standard empirical process theory (Vaart and Wellner, 2023; Kosorok, 2008). Finally, Assumption 2 condition vii. is the Neyman orthogonality condition (Chernozhukov et al., 2018a). We will show how these conditions are ensured in the incomplete data setting in Section 1.2. In particular, the Neyman orthogonality condition holds for the incomplete-data efficient estimating equations in Section 1.2 with respect to the class of all square integrable functions, let alone any reasonable nuisance realization set $\mathcal{T}_N$.

Our second set of assumptions involve conditions on our nuisance estimators.

Assumption 3 (Nuisance Estimation Conditions)**.** For any $\mathbb{P}\in\mathcal{P}_N$:

i. For some sequence of constants $\Delta_N\to0$, the nuisance estimates $(\hat\eta_1^{(k)}(\cdot\,;\hat\theta_{1,\text{init}}^{(k)}),\hat\eta_2^{(k)}(\cdot))$ belong to the realization set $\mathcal{T}_N$ for all $k=1,\dots,K$ with probability[1] at least $1-\Delta_N$.

ii. For some sequence of constants $\delta_N,\tau_N\to0$, the statistical rates r_N, r_N', $\lambda_N'(\theta)$ satisfy:

$$r_N\coloneqq\sup_{(\eta_1(\cdot;\theta_1'),\eta_2)\in\mathcal{T}_N,\theta\in\Theta}\left\|\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z))\right]-\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\right\|\le\delta_N\tau_N,$$

$$r_N'\coloneqq\sup_{\substack{\theta\in\mathcal{B}(\theta^*;\tau_N),\\(\eta_1(\cdot;\theta_1'),\eta_2)\in\mathcal{T}_N}}\left\|\left(\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z))-\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]^2\right)^{1/2}\right\|\le\tfrac{\delta_N}{\log N},$$

$$\lambda_N'(\theta)\coloneqq\sup_{\substack{r\in(0,1),\\(\eta_1(\cdot;\theta_1'),\eta_2)\in\mathcal{T}_N}}\|\partial_r^2 f(r;\theta,\eta_1(\cdot;\theta_1'),\eta_2)\|\le\left(\|\theta-\theta^*\|+N^{-1/2}\right)\delta_N,\quad\forall\theta\in\mathcal{B}(\theta^*;\tau_N),$$

where $f(r;\theta,\eta_1(\cdot\,;\theta_1'),\eta_2)\coloneqq\mathbb{P}[\psi(Z;\theta^*+r(\theta-\theta^*),\eta_1\left(Z;\theta_1',r\right),\eta_2\left(Z;r\right))]$,

$$\eta_1\left(Z;\theta_1',r\right)\coloneqq\eta_1^*(Z;\theta_1^*)+r(\eta_1(Z;\theta_1')-\eta_1^*(Z;\theta_1^*)),\ \eta_2\left(Z;r\right)\coloneqq\eta_2^*(Z)+r(\eta_2(Z)-\eta_2^*(Z)).$$

iii. The solution approximation error in (15) satisfies $\varepsilon_N\le\delta_N N^{-1/2}$.

Here our condition on $\lambda_N'(\theta)$ differs from the counterpart condition in Chernozhukov et al. (2018a), which also leads to a different proof strategy. Our condition and proof generally

[1] The probability is with respect to both the randomness in data sampling and sample splitting in the cross-fitting, as both impact the realizations of the nuisance estimates.

require weaker conditions for convergence rates of nuisance estimators. See the discussions in Appendix F for more details. Moreover, the constants $\Delta_N, \delta_N, \tau_N$ are all prespecified and do not depend on any particular instance $\mathbb{P}$. In Section 5, we will verify that Assumption 3 holds for our examples in the incomplete data setting under mild conditions on the smoothness of the estimating equations and slow rate conditions on the nuisance estimators (for example, see the conditions in Theorem 3 and its proof).

Our key result in this paper is the following theorem, which shows that the asymptotic distribution of our estimator is identical to the (infeasible) oracle estimator solving the estimating equation in Eq. (9) with known nuisances.

Theorem 1 (Asymptotic Behavior of LDML)**.** *Assume Assumptions 1 to 3 hold with*

$$\begin{aligned} &\max\{\log^2 N(1+N^{-1/2+1/q}),\, \delta_N \log N\}/\sqrt{N} \le \tau_N \le \delta_N, \\ &\max\{r'_N \log^{1/2}(1/r'_N),\, N^{-1/2+1/q}\log(1/r'_N)\} \le \delta_N. \end{aligned} \tag{17}$$

Then the estimator $\hat{\theta}$ given in Definition 2 is asymptotically linear and converges to a Gaussian distribution uniformly over $\mathbb{P} \in \mathcal{P}_N$:

$$\sqrt{N}\Sigma^{-1/2}(\hat{\theta}-\theta^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} -\Sigma^{-1/2}J^{*-1}\psi(Z_i;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i)) + O_{\mathbb{P}}(\rho_N) \rightsquigarrow \mathcal{N}(0, I_d),$$

where Σ is given in Eq. (16), the remainder term satisfies $\rho_N = (N^{-1/2+1/q} + r'_N)\log N + r'_N \log^{1/2}(1/r'_N) + N^{-1/2+1/q}\log(1/r'_N) + \delta_N \lesssim \delta_N$, and the $O_{\mathbb{P}}$ term depends only on constants pre-specified in Assumptions 1 to 3 and no instance-specific constants.

The conditions in Eq. (17) and $\rho_N \lesssim \delta_N$ are fairly mild because Assumption 2 condition vi. requires $q > 2$ (so $N^{-1/2+1/q} \to 0$) and Assumption 3 condition ii. requires $r'_N \le \frac{\delta_N}{\log N}$.

Remark 1 (Uniform vs non-uniform convergence)**.** To obtain a non-uniform convergence result, we need only need set $\mathcal{P}_N = \{\mathbb{P}\}$ as a constant singleton in Theorem 1. In this case, much of Assumption 2 simplifies: the existence of the constants c_4, c_6 is trivial, the non-singularity of J^* is enough for c_3 to exist, and θ^* being in the interior of Θ is enough for c_1 to exist. Further, we can relax condition iv. by allowing c_5 to be zero (in which case we rephrase the asymptotic normality in Theorem 1 by putting Σ on the right-hand side of the limit rather than inverting it). Uniformity, however, is important in practice. Without uniformity, for any given sample size N there may always exist some bad instances such that the normal approximation suggested by the convergence is inaccurate (Kasy, 2019).

4. Variance Estimation and Inference

In the previous section we established the asymptotic normality of the LDML estimator under lax conditions. This suggests that if we can estimate its asymptotic variance, then we can easily construct confidence intervals on θ. In this section we provide a variance estimator and prove its consistency, resulting in asymptotically calibrated confidence intervals. For DML, Chernozhukov et al. (2018a) provides variance estimates only for estimating functions ψ that are linear in θ, which already excludes estimand-dependent nuisances. Our results are therefore notable both for handling nonlinear and non-differentiable estimating equations and for handling estimand-dependent nuisances.

Definition 3 (LDML variance estimator)**.** Given $\hat{\theta}$ from Definition 2 and $\hat{J}$, set

$$\widehat{\Sigma} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\hat{J}^{-1}\psi(Z_i;\hat{\theta},\hat{\eta}_1^{(k)}(Z_i;\hat{\theta}_{1,\mathrm{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z_i))\psi(Z_i;\hat{\theta},\hat{\eta}_1^{(k)}(Z_i;\hat{\theta}_{1,\mathrm{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z_i))^\top\hat{J}^{-\top}.$$

We next establish the consistency of $\widehat{\Sigma}$, which relies on the following assumption.

Assumption 4. Assume that $\|\hat{J}-J^*\| = \rho_{J,N} \lesssim \delta_N$ and that for some $C, \beta > 0$,

$$\begin{aligned} &m_N := \sup\nolimits_{(\eta_1(\cdot;\theta_1'),\eta_2)\in\mathcal{T}_N}\mathbb{P}[\|\psi(Z;\theta^*,\eta_1(Z;\theta_1'),\eta_2(Z))\|^4]^{1/4} \le C \quad \forall\theta\in\mathcal{B}(\theta^*;\tau_N),\\ &\mathbb{P}[\|\psi(Z;\theta,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i)) - \psi(Z;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i))\|^2] \le C\|\theta-\theta^*\|^\beta. \end{aligned} \tag{18}$$

Here, Eq. (18) implies continuity $\theta\mapsto\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ in terms of L_2 norm in the range space. Note that this condition can be satisfied even if $\theta\mapsto\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ is non-differentiable. For example, in the estimation of QTEs, the efficient estimating equation in Eq. (3) involves the indicator function $\mathbb{I}\left[Y\le\theta_1\right]$, so the map $\theta\mapsto\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ is obviously not differentiable. However, the condition in Eq. (18) amounts to

$$\mathbb{P}[\left(\mathbb{P}\left(T=1\mid X\right)\right)^{-1}\left(\mathbb{I}[Y\le\theta_1]-\mathbb{I}[Y\le\theta_1^*]\right)^2]\le C|\theta_1-\theta_1^*|^\beta.$$

In Assumption 5, we will assume that $\mathbb{P}\left(T=1\mid X\right)\ge\epsilon_\pi$ for a positive constant ϵ_π. Then the condition above follows if the cumulative distribution function of $Y\left(1\right)$ is smooth enough, so that $|\mathbb{P}(Y(1)\le\theta_1)-\mathbb{P}(Y(1)\le\theta_1^*)|\le C\epsilon_\pi\left|\theta_1-\theta_1^*\right|^\beta$ for any $\theta_1\in\mathcal{B}(\theta_1^*;\tau_N)$.

Under Assumption 4, we now show that the variance estimator in Definition 3 is consistent and it leads to asymptotically valid confidence intervals.

Theorem 2. *Assume the assumptions in Theorem 1 and Assumption 4. Then,*

$$\begin{aligned} &\hat{\Sigma} = \Sigma + O_{\mathbb{P}}(\rho_N'') \to \Sigma, \quad \textit{uniformly over } \mathbb{P}\in\mathcal{P}_N,\\ &\textit{where } \rho_N'' = N^{-1/2+1/q}(\log N)^{1/2} + N^{-1/4}(\log N)^{1/2} + r_N' + \rho_{J,N} + N^{-\beta/4} \lesssim \delta_N. \end{aligned}$$

Given some $\zeta\in\mathbb{R}^d$*, the confidence interval* $\mathrm{CI} := [\zeta^\top\hat{\theta}\pm\Phi^{-1}(1-\alpha/2)\sqrt{\zeta^\top\hat{\Sigma}^2\zeta/N}]$ *obeys*

$$\sup\nolimits_{\mathbb{P}\in\mathcal{P}_N}\left|\mathbb{P}(\zeta^\top\theta^*\in\mathrm{CI})-(1-\alpha)\right|\to 0,\ \textit{as } N\to\infty.$$

In Assumption 4, we assumed that we have a consistent estimator $\hat{J}$ for J^*. How to construct such an estimator depends on the problem. When $\theta\mapsto\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ is differentiable, an estimator may easily be constructed as follows:

$$\hat{J} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\partial_{\theta^\top}\psi(Z;\theta,\hat{\eta}_1^{(k)}(Z;\hat{\theta}_{1,\mathrm{init}}),\hat{\eta}_2^{(k)}(Z))|_{\theta=\hat{\theta}}.$$

However, the estimating equation for QTE is not differentiable. Thus we rely on deriving the form of J^* and estimate it directly, which we discuss in detail in Remark 4.

With finite sample, the variance of the LDML estimator also depends on the uncertain sample splitting in Definition 1. This uncertainty can be additionally accounted for when multiple sample splitting realizations are used, which we discuss in Appendix E.

Remark 2 (Estimating and Conducting Inference on Treatment Effects)**.** Suppose we have two sets of parameters, $\theta^{*(0)}$, $\theta^{*(1)}$, each identified by its own estimating equation, $\psi^{(0)}$, $\psi^{(1)}$, and we are interested in estimating the difference, $\tau^* = \theta^{*(1)} - \theta^{*(0)}$. For example, $\theta^{*(0)}$, $\theta^{*(1)}$ can be the quantile and/or CVaR of $Y(0)$, $Y(1)$, respectively, and we are interested in the QTE and/or CVaR treatment effect. To do this, we can concatenate the two estimating equations and augment them with the additional equation $\theta^{*(1)} - \theta^{*(0)} - \tau^* = 0$. Estimating this set of estimating equations with LDML is equivalent to applying LDML to each of $\psi^{(0)}$, $\psi^{(1)}$ and letting $\hat\tau$ be the difference of the estimates $\hat\theta^{(0)}, \hat\theta^{(1)}$, where we may use the same data and the same folds for the two LDML procedures. For QTE and for other estimating equations with incomplete data, we can even share the nuisance estimates of the propensity score (*i.e.*, $\hat\eta_2^{(0),(k)} = 1 - \hat\eta_2^{(1),(k)}$ in the below equation). The variance estimate one would derive for $\hat\tau$ from the augmented estimating equations is equivalent to

$$\widehat{\Sigma}_\tau = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\omega_{i,k}\omega_{i,k}^\top, \quad \text{where} \quad \omega_{i,k} = (\hat J^{(1)})^{-1}\psi^{(1)}(Z_i;\hat\theta^{(1)},\hat\eta_1^{(1),(k)}(Z_i;\hat\theta_{1,\mathrm{init}}^{(1),(k)}),\hat\eta_2^{(1),(k)}(Z_i))$$
$$- (\hat J^{(0)})^{-1}\psi^{(0)}(Z_i;\hat\theta^{(0)},\hat\eta_1^{(0),(k)}(Z_i;\hat\theta_{1,\mathrm{init}}^{(0),(k)}),\hat\eta_2^{(0),(k)}(Z_i)).$$

5. Estimating Equations with Incomplete Data

In this section, we apply our method and theory to general estimating equations with incomplete data presented in Eq. (4), which subsumes the estimation of QTEs, quantile of potential outcomes, CVaR treatment effect, CVaR of potential outcomes, expectile treatment effect, and expectile of potential outcomes. We will proceed to further specialize these results to quantile and CVaR estimation, deferring the case of expectiles to the appendix (Appendix B). We also defer the case of using IVs to estimate the solution to *local* estimating equations, such as those that describe the LQTE, to appendix (Appendix A).

As motivated in Section 1.1, under unconfoundedness, there is a very natural initial estimator: the IPW estimator. As we will show, the LDML estimate for this problem using the IPW initial estimator can be computed using just blackbox algorithms for (possibly binary) regression, which is the standard supervised learning task in machine learning. And, under lax conditions, the estimate is efficient, asymptotically normal, and amenable to inference.

Recall that θ is defined by the complete-data estimating equations in Eq. (4), namely, $\mathbb{P}[U(Y(1);\theta_1) + V(\theta_2)] = 0$. Assuming ignorability and overlap, θ is identified from the incomplete-data observations $Z = (X, T, Y)$ where $Y = Y(T)$. In particular, Eq. (7) provides a Neyman-orthogonal estimating equation identifying θ. For better interpretability, we give our nuisances names: we denote $\pi^*(t \mid x) = \mathbb{P}(T = t \mid X = x)$, $\mu_j^*(x,t;\theta_1) = \mathbb{E}\left[U_j(Y;\theta_1) \mid X = x, T = t\right]$, and $\mu^*(x,t;\theta_1) = [\mu_1^*(x,t;\theta_1), \ldots, \mu_d^*(x,t;\theta_1)]^\top$. For estimating parameters corresponding to $Y(1)$, our estimand-independent nuisance is the propensity score $\eta_2^*(Z) = \pi^*(1 \mid X)$, and our estimand-dependent nuisance is $\eta_1^*(Z;\theta_1) = \mu^*(X,1;\theta_1)$. The case for $Y(0)$ is symmetric; and it also need the symmetric ignorability and overlap assumptions for identifiability: $Y(0) \perp\!\!\!\perp T \mid X$ and $\mathbb{P}(T = 1 \mid X) < 1$. Treatment effects (*e.g.*, QTEs) can be estimated by differences of estimates, where we can use the same data, the same fold splits, and the same estimates of π^* for both treatments (see Remark 2).

This problem also admits a simpler but unstable (*i.e.*, non-orthogonal) estimating equation using IPW, which suggests a possible initial estimator, using $K' \geq 2$ in Definition 1:

Definition 4 (IPW Initial Estimator)**.** For each $k = 1, \ldots, K$ and $l \in \mathcal{H}_{k,1}$ as in Definition 1, use only the data in $\mathcal{D}_k^{C,1,l} = \left\{Z_i : i \in \bigcup_{k' \in \mathcal{H}_{k,1} \setminus \{l\}} \mathcal{D}_{k'}\right\}$ to construct a propensity score estimator $\hat{\pi}^{(k,l)}(1 \mid \cdot)$ for $\pi^*(1 \mid \cdot)$. Then let $\hat{\theta}_{1,\text{init}}^{(k)}$ be given by solving the following estimating equation (or, its least squares solution up to approximation error of ε_N):

$$\frac{1}{|\mathcal{D}_k^{C,1}|} \sum_{l \in \mathcal{H}_{k,1}} \sum_{i \in \mathcal{D}_l} \psi^{\text{IPW}}(Z_i;\theta,\hat{\pi}^{(k,l)}) = 0, \text{ where } \psi^{\text{IPW}}(Z;\theta,\pi) = \frac{\mathbb{I}(T=1)}{\pi(1 \mid X)} U(Y;\theta_1) + V(\theta_2).$$

This procedure is illustrated in Fig. 1(b). Note that, given a fixed θ_1', both $\pi^*(1 \mid \cdot)$ and $\mu^*(\cdot, 1; \theta_1')$ are conditional expectations of observable variables given X. Thus, in this setting, the whole LDML estimate using the IPW initial estimate can be computed given just blackbox algorithms for (possibly binary) regression.

5.1 Theoretical Analysis

We first study the LDML estimate for estimating equations with incomplete data by leveraging our general theory in Theorem 1. To this end, we assume a strong form of the overlap condition and specify the convergence rates of the initial estimator and nuisance estimators used. We consider a generic treatment level $t \in \{0, 1\}$ in these two assumptions.

Assumption 5 (Strong Overlap)**.** Assume that there exists a positive constant $\varepsilon_\pi > 0$ such that for any $\mathbb{P} \in \mathcal{P}_N$, $\pi(t \mid X) \geq \varepsilon_\pi$ almost surely.

Assumption 6 (Nuisance Estimation Rates)**.** Assume that for any $\mathbb{P} \in \mathcal{P}_N$: condition i. of Assumption 3 holds for a sequence of constants $\Delta_N \to 0$; with probability at least $1 - \Delta_N$, $\hat{\pi}^{(k)}(t \mid X) \geq \varepsilon_\pi$ for almost all realizations of X, and

$$\|(\mathbb{P}(\hat{\mu}^{(k)}(X,t;\hat{\theta}_{1,\text{init}}^{(k)}) - \mu^*(X,t;\hat{\theta}_{1,\text{init}}^{(k)}))^2)^{1/2}\| \leq \rho_{\mu,N},$$
$$(\mathbb{P}(\hat{\pi}^{(k)}(t \mid X) - \pi^*(t \mid X))^2)^{1/2} \leq \rho_{\pi,N}, \quad \|\hat{\theta}_{1,\text{init}}^{(k)} - \theta^*\| \leq \rho_{\theta,N}.$$

The following theorem establishes that the asymptotic distribution of our proposed estimator is similar to the (infeasible) one that solves the semiparametric efficient estimating equation in Eq. (7) with known nuisances. This theorem is proved by verifying conditions in Theorem 1, namely Assumptions 1 to 3.

Theorem 3 (LDML for Estimating Equations with Incomplete Data)**.** *Fix $t = 1$ and let the estimator $\hat{\theta}$ be given by applying Definition 2 to the estimating equation in Eq.* (7)*.*

Suppose Assumptions 5 and 6 hold and that there exist positive constants c', C, and c_1 to c_7 such that for any $\mathbb{P} \in \mathcal{P}_N$ the following conditions hold:

i. Conditions i. (with c_1), ii., v. (with c_5, c_6), and vi. (with c_7) of Assumption 2 and condition iii. of Assumption 3 for the estimating equation in Eq. (7).

ii. For $j = 1, \dots, d$, $\theta \mapsto \mathbb{P}\left[U_j(Y(t);\theta_1) + V(\theta_2)\right]$ is differentiable at any θ in a compact set Θ, and each component of its gradient is c'-Lipschitz continuous at θ^*. Moreover, for any $\theta \in \Theta$ with $\|\theta - \theta^*\| \geq \frac{c_3}{2\sqrt{d}c'}$, we have $2\|\mathbb{P}\left[U(Y(t);\theta_1) + V(\theta_2)\right]\| \geq c_2$.

iii. The singular values of $\partial_{\theta^\top}\mathbb{P}\left[U(Y(t);\theta_1) + V(\theta_2)\right]|_{\theta=\theta^*}$ are bounded between c_3 and c_4.

iv. For any $\theta \in \mathcal{B}(\theta^*; \frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N \varepsilon_\pi}) \cap \Theta$, $r \in (0,1)$, and $j = 1, \dots, d$, there exist $h_1(x,t;\theta_1), h_2(x,t;\theta_1)$ such that $\mathbb{P}\left[h_1(X,t;\theta_1)\right] < \infty$, $\mathbb{P}\left[h_2(X,t;\theta_1)\right] < \infty$ and almost surely

$$\left|\partial_r \mu_j^*(X,t;\theta_1^* + r(\theta_1 - \theta_1^*))\right| \leq h_1(X,t;\theta_1), \quad \left|\partial_r^2 \mu_j^*(X,t;\theta_1^* + r(\theta_1 - \theta_1^*))\right| \leq h_2(X,t;\theta_1).$$

v. For $j = 1, \dots, d$ and any $\theta \in \Theta$, we have $(\mathbb{P}(\mu_j^*(X,t;\theta_1))^2)^{1/2} \leq C$.

vi. For $j = 1, \dots, d$ and any $\theta \in \mathcal{B}(\theta^*; \max\{\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N \varepsilon_\pi}, \rho_{\theta,N}\}) \cap \Theta$.

$$\left\{\mathbb{P}\left[\mu_j^*(X,t;\theta_1) - \mu_j^*(X,t;\theta_1^*)\right]^2\right\}^{1/2} \leq C\|\theta_1 - \theta_1^*\|, \quad \left\|\left\{\mathbb{P}\left[\partial_{\theta_1}\mu_j^*(X,t;\theta_1)\right]^2\right\}^{1/2}\right\| \leq C,$$
$$\sigma_{\max}\left(\mathbb{P}\left[\partial_{\theta_1}\partial_{\theta_1^\top}\mu_j^*(X,t;\theta_1)\right]\right) \leq C, \ \sigma_{\max}\left(\partial_{\theta_2}\partial_{\theta_2^\top}V_j(\theta_2)\right) \leq C.$$

vii. $\rho_{\pi,N}(\rho_{\mu,N} + C\rho_{\theta,N}) \leq \frac{\varepsilon_\pi^3}{3}\delta_N N^{-1/2}$, $\rho_{\pi,N} \leq \frac{\delta_N^3}{\log N}$, $\rho_{\mu,N} + C\rho_{\theta,N} \leq \frac{\delta_N^2}{\log N}$, $\delta_N \leq \frac{4C^2\sqrt{d}+2\varepsilon_\pi}{\varepsilon_\pi^2}$, and $\delta_N \leq \min\{\frac{\varepsilon_\pi^2}{8C^2 d}\log N, \sqrt{\frac{\varepsilon_\pi^3}{2C\sqrt{d}}}\log^{1/2} N\}$.

Then $\hat{\theta}$ satisfies the conclusion of Theorem 1 for $\psi(Z;\theta^*, \eta_1^*(Z;\theta_1^*), \eta_2^*(Z))$ given in Eq. (7), and its asymptotic variance Σ attains the corresponding semiparametric efficiency bound.

An analogous result for the estimating equations involving $Y(0)$ holds when we change $t = 1$ to $t = 0$ everywhere in Theorem 3. See Remark 2 regarding estimation of the difference of the parameters (*i.e.*, the treatment effects) and inference thereon.

In Theorem 3, conditions ii. and iii. guarantee the identification conditions iii. and iv. of Assumption 2. Condition iv. enables exchange of integration, which together with conditions v., vi., and vii. imply the rate condition ii. of Assumption 3. Note condition vii. permits nonparametric rates for nuisance estimators. Focusing on the order in the sample size and up to polylog factors, the condition allows for $\rho_{\pi,N}\rho_{\mu,N} = o(N^{-1/2})$, $\rho_{\pi,N}\rho_{\theta,N} = o(N^{-1/2})$, $\rho_{\pi,N} = o(1)$, $\rho_{\mu,N} = o(1)$, $\rho_{\theta,N} = o(1)$. Note the first two restrictions are on *products*, permitting a trade-off between rates for different nuisances (see also Appendix F).

Remark 3 (Rate Conditions with IPW Initial Estimator)**.** In Appendix C, we prove that if the propensity nuisance estimators used to construct the IPW initial estimators (Definition 4) also have convergence rate $\rho_{\pi,N}$, then the initial estimators' convergence rates satisfy that $\rho_{\theta,N} = O\left(\rho_{\pi,N}\right)$. In this case, we are essentially imposing $\rho_{\pi,N} = o(N^{-1/4})$: condition

vii. of Theorem 3 requires $\rho_{\pi,N}\rho_{\theta,N} = o(N^{-1/2})$, so unless $\rho_{\theta,N}$ is somehow even faster than $\rho_{\pi,N}$, we must need both $\rho_{\theta,N}$ and $\rho_{\pi,N}$ to be $o(N^{-1/4})$.

5.2 Quantile and CVaR

Now we consider estimating quantile and (possibly) CVaR based on the semiparametrically efficient estimating equation in Eq. (3). Instantiating Eq. (7) for the simultaneous estimation of quantile and CVaR and rearranging, we obtain the following estimating equation:

$$\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z)) = \frac{\mathbb{I}\left[T=1\right]}{\eta_2^*(Z)}\begin{bmatrix}\mathbb{I}[Y\le\theta_1]-\eta_{1,1}^*(Z;\theta_1)\\ \frac{1}{1-\gamma}\big(\max(Y-\theta_1,0)-\eta_{1,2}^*(Z;\theta_1)\big)\end{bmatrix}+\begin{bmatrix}\eta_{1,1}^*(Z;\theta_1)-\gamma\\ \theta_1+\frac{1}{1-\gamma}\eta_{1,2}^*(Z;\theta_1)-\theta_2\end{bmatrix},$$

$$\text{where}\quad \eta_1^*(Z;\theta_1)=\begin{bmatrix}\mathbb{P}\left(Y\le\theta_1\mid X,T=1\right)\\ \mathbb{E}\left[\max(Y-\theta_1,0)\mid X,T=1\right]\end{bmatrix},\quad \eta_2^*(Z)=\mathbb{P}\left(T=1\mid X\right). \tag{19}$$

We use $F_t(\cdot\mid x)$ and $F_t(\cdot)$ to denote the conditional and unconditional cumulative distribution function of $Y(t)$, respectively: for any y, $F_t(y\mid x)=\mathbb{P}(Y(t)\le y\mid X=x)$ and $F_t(y)=\mathbb{P}(Y(t)\le y)$. The following proposition gives the asymptotic behavior of our proposed estimators for the quantile and CVaR of $Y(1)$. This conclusion is proved by verifying all conditions in Theorem 3. Analogous conclusions also hold for $Y(0)$ when all assumptions hold for $t=0$ instead of $t=1$.

Proposition 2 (LDML for Quantile and CVaR)**.** *Fix $t=1$ and Let the estimator $\hat\theta$ be given by applying Definition 2 to the estimating function in Eq.* (19)*. Suppose Assumptions 5 and 6 hold and there exist positive constants $c_1'\sim c_5'$ and $C\ge 1$, such that for any $\mathbb{P}\in\mathcal{P}_N$, the following conditions hold:*

i. Conditions i. (with c_1), ii., v. (with c_5, c_6) of Assumption 2, condition iii. of Assumption 3, and condition vii. of Theorem 3 for the estimating function in Eq. (19) *and the corresponding nuisance estimators.*

ii. $F_t(\theta_1)$ is twice differentiable with derivatives $f_t(\theta_1),\dot f_t(\theta_1)$ satisfying $0<c_1'\le f_t(\theta_1^*)$, $f_t(\theta_1)\le c_2'$, $|\dot f_t(\theta_1)|\le c_3'$ $\forall\theta_1\in\Theta_1$. Moreover, $|F_t(\theta_1^*)-F_t(\theta_1)|\ge c_4'$ for $|\theta_1-\theta_1^*|\ge\frac{c_1'}{2c_3'}$.

iii. At any $\theta\in\mathcal{B}(\theta^*;\max\{\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi},\rho_{\theta,N}\})\cap\Theta$, $F_t(\theta_1\mid X)$ is twice differentiable almost surely with first two order derivatives $f_t(\theta_1\mid X)$ and $\dot f_t(\theta_1\mid X)$ that satisfy $f_t(\theta_1\mid X)\le C$ and $|\dot f_t(\theta_1\mid X)|\le C$ almost surely.

iv. $2\|\mathbb{P}[U(Y(t);\theta_1)+V(\theta_2)]\|\ge c_5'$ for $\|\theta-\theta^*\|\ge\frac{\min\{\gamma,(1-\gamma)c_1',\gamma c_1'\}}{4\sqrt{2}\gamma\max\{c_2',c_3'\}}$ and $U(Y(t);\theta_1)+V(\theta_2)$ as given in Eq. (5)*.*

v. $\big(\mathbb{P}\big(\mathbb{E}[\max(Y-\theta_1,0)\mid X,T=t]^2\big)\big)^{1/2}\le C$ for any $\theta\in\Theta$.

Then $\hat\theta$ satisfies the conclusion of Theorem 1 for $\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ given in Eq. (19) *and for $J^*=\operatorname{diag}(f_t(\theta_1^*),-1)$. Moreover, under all conditions above except conditions iv. and v., the quantile estimator $\hat\theta_1$ alone still satisfies the analogous asymptotic linear expansion for $\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ given in Eq.* (3) *and for $J^*=f_t(\theta_1^*)$.*

Remark 4 (Estimating $f_t(\theta_1^*)$ for Variance Estimation)**.** If we want to conduct inference on the quantile or QTE using our method from Section 4, we need to estimate $f_t(\theta_1^*)$. Theoretically, for nominal asymptotic coverage, we need only do this consistently, regardless of rate. One simple approach is to use cross-fitted IPW kernel density estimation at $\hat{\theta}_1$:

$$\hat{J} = \frac{1}{Nh}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} \frac{\mathbb{I}\left[T_i = 1\right]}{\hat{\pi}^{(k)}(1 \mid X_i)} \kappa((Y_i - \hat{\theta}_1)/h),$$

where $\kappa(u)$ is a kernel function such as $\kappa(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and $h \to 0$ is a bandwidth. Under Assumption 5, $h \asymp N^{-1/5}$ would be the optimal bandwidth. While this together with any consistent estimate $\hat{\pi}^{(k)}$ suffices for asymptotic coverage, the estimate may be unstable. Generally, estimating f_t at any one point, known as counterfactual density estimation, is a challenging problem. The above simple estimator may be improved by introducing weight normalization or clipping. There also exist more sophisticated counterfactual density estimators (*e.g.*, Kennedy et al., 2021) that may lead to better variance estimation for counterfactual quantile estimators and better finite-sample coverage.

6. Empirical Results

We first study the behavior of LDML in a simulation study. We then demonstrate its use in estimating the QTE of 401(k) eligibility on net financial assets, and the LQTE of 401(k) participation using eligibility as IV. Replication code is available at `https://github.com/CausalML/LocalizedDebiasedMachineLearning`.

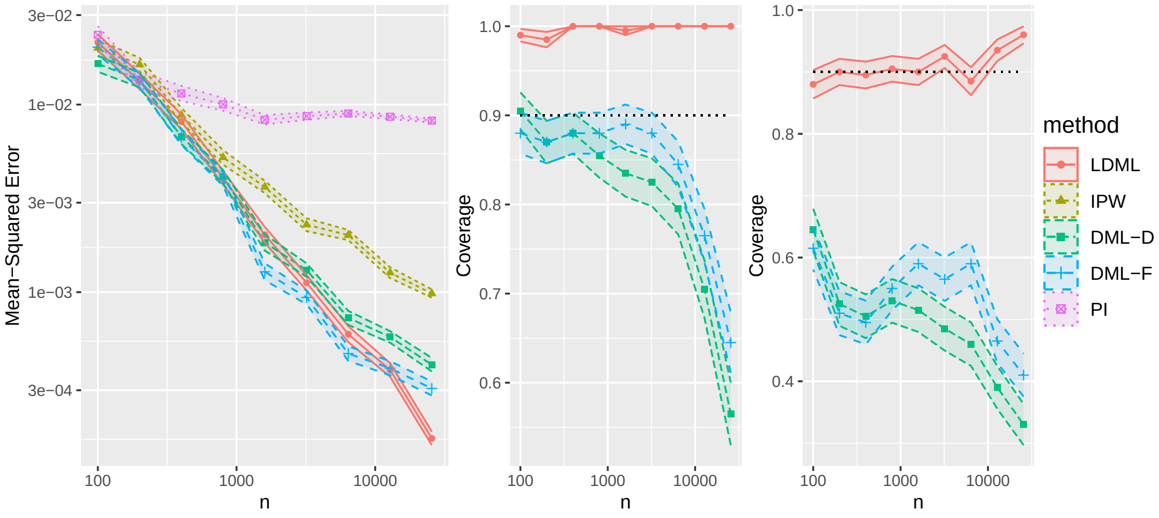
6.1 Simulation Study

First, we consider a simulation study to compare the performance of LDML estimates to benchmarks. We consider estimating θ_1^* as the second tertile of $Y(1)$ from incomplete data. The data is randomly generated according to the following process:

$$X \sim \text{Uniform}\left([0,1]^{20}\right),\ T \sim \text{Bernoulli}(\Phi(3(1 - X_1 - X_3))),\ Y(1) \sim \mathcal{N}(\mathbb{I}[X_1 + X_2 \leq 1],\, 2X_3),$$

and finally we have access to observations for the variables (X, T, Y) where $Y = Y(1)$ when $T = 1$ and Y is missing when $T = 0$.

We consider estimating θ_1^* using five different methods. First, we consider LDML applied to the efficient estimating equation (Eq. (3)) with $K = 5, K' = 2$, $\hat{\theta}_{1,\text{init}}^{(k)}$ estimated using 2-fold cross-fitted IPW with random-forest-estimated propensities, and $\hat{\pi}^{(k)}(1 \mid X)$, $\hat{\mu}^{(k)}(X, 1; \hat{\theta}_{1,\text{init}}^{(k)})$ similarly estimated by random forests. Second, we consider $K = 5$-fold cross-fitted IPW with random-forest-estimated propensities. Third, we consider DML with $K = 5$ and the estimand-dependent nuisance estimated using a discretization approach similar to the suggestion of Belloni et al. (2018): for $j = 1, \dots, 99$, fix $\theta_{1,j}$ to be the $j/100$ marginal quantile of Y and fit $\hat{\mu}^{(k)}(X, 1; \theta_{1,j})$ using random forests; then apply DML with the restricted discretized estimand range $\{\theta_{1,j} : j = 1, \dots, 99\}$. We refer to this method as DML-D for *discretized.* Fourth, we consider taking the empirical cross-fold average of the same counterfactual CDF estimator, $\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k}\hat{\mu}^{(k)}(X, 1; \cdot)$,



(a) Mean-squared error. (b) Coverage, est density. (c) Coverage, true density.

Figure 2: Results for the simulation study of estimating the second tertile of $Y(1)$ using different methods. The curves show the estimated mean squared errors. Shaded regions denote plus/minus one standard error for estimated mean-squared error or mean coverage. All results are computed from 250 replications of simulations.

and directly inverting this to estimate the marginal quantile. We refer to this method as PI for *plug-in*. Fifth, we consider DML with $K = 5$ and where the estimand-dependent nuisance is estimated using an approach similar to Meinshausen (2006); Bertsimas and Kallus (2014): namely, fit a random forest regression to the out-of-fold data $\{(X_i, Y_i) : i \notin \mathcal{D}_k, T_i = 1\}$ to obtain B regression trees $\tau_j : \text{support}(X) \to \{1, \ldots, \ell_j\}$, then set $\hat{\mu}^{(k)}(X, 1; \theta_1) = \sum_{i \notin \mathcal{D}_k : T_i = 1} \frac{\mathbb{I}[Y_i \le \theta_1]}{B} \sum_{j=1}^{B} \frac{\mathbb{I}[\tau_j(X_i) = \tau_j(X)]}{\sum_{i' \notin \mathcal{D}_k : T_{i'} = 1} \mathbb{I}[\tau_j(X_i) = \tau_j(X_{i'})]}$ for all θ_1. We refer to this method as DML-F for *forest*. For each method, we run it three times with new random fold splits (with the same data) and take the median of the three results to be the estimate.

For each of $n = 100, 200, \ldots, 25600$, we consider 250 replications of drawing a dataset of size n and constructing each of the above four estimates. We plot the mean-squared error of each method and n over the 250 replications in Fig. 2(a). The shaded regions show plus/minus one standard error of this as the sample mean of 250 squared errors. LDML offers competitive performance with DML, while avoiding fitting a continuum of nuisances (or approximating them with discretization), and even offers a marked improvement for large n. Methods without debiasing, IPW and PI, perform less well.

In Fig. 2(b) and (c), we additionally report the coverage of the true parameter by confidence intervals given by the estimand plus/minus 1.645 of an estimated standard error. The standard error for LDML is estimated as in Definition 3 and similarly for DML but using $\hat{\eta}_1^{(k)}(Z_i; \hat{\theta})$ in place of $\hat{\eta}_1^{(k)}(Z_i; \hat{\theta}_{1,\text{init}}^{(k)})$ using the DML estimate $\hat{\theta}$. We consider two choices for the density estimator $\hat{J}$: a kernel density estimator on a million iid draws from $Y(1)$ (true density) and the density estimator in Remark 4 (est density). The latter appears to underestimate the density by roughly a half on average, leading LDML to cover conser-

Table 1: The QTE of 401(k) eligibility in thousand dollars (and standard error) estimated by LDML using different regression methods (LASSO, neural network, boosting, and random forests), and raw differences of margianl quantiles that do not adjust for covariates. Here $\gamma \in \{25\%, 50\%, 75\%\}$ denotes the quantile level of QTE, and $K \in \{5, 15, 25\}$ denotes the number of folds used in the cross-fitting of LDML.

γ	K	LASSO	Neural Net	Boosting	Forest	Raw
25%	5	0.95 (0.24)	1.05 (0.19)	1.00 (0.20)	0.93 (0.29)	1.50 (0.25)
	15	0.95 (0.24)	1.06 (0.20)	1.00 (0.20)	0.93 (0.28)	
	25	0.95 (0.24)	1.03 (0.20)	1.00 (0.20)	0.93 (0.29)	
50%	5	4.74 (0.68)	5.56 (0.69)	4.47 (0.85)	3.64 (1.87)	8.98 (0.41)
	15	4.68 (0.68)	5.59 (0.68)	4.47 (0.85)	3.46 (1.85)	
	25	4.68 (0.68)	5.55 (0.67)	4.47 (0.85)	3.45 (1.85)	
75%	5	14.00 (4.14)	17.12 (4.10)	13.28 (5.11)	13.88 (11.32)	29.67 (1.35)
	15	13.94 (4.12)	16.86 (4.01)	13.29 (5.20)	14.30 (12.11)	
	25	13.93 (4.13)	16.87 (4.00)	13.29 (5.16)	14.29 (12.23)	

Table 2: The LQTE of 401(k) participation in thousand dollars (and standard error) estimated by LDML using different regression methods (LASSO, neural network, boosting, and random forests), and raw differences of marginal quantiles by eligibility that do not adjust for covariates. Here $\gamma \in \{25\%, 50\%, 75\%\}$ denotes the quantile level of LQTE, and $K \in \{5, 15, 25\}$ denotes the number of folds used in the cross-fitting of LDML.

γ	K	LASSO	Neural Net	Boosting	Forest	Raw
25%	5	1.75 (0.23)	2.06 (0.25)	1.57 (0.26)	1.91 (0.44)	4.18 (0.37)
	15	1.74 (0.23)	2.04 (0.25)	1.57 (0.26)	1.88 (0.44)	
	25	1.75 (0.23)	2.07 (0.25)	1.58 (0.26)	1.87 (0.44)	
50%	5	8.64 (0.60)	10.38 (0.66)	7.54 (0.60)	6.32 (1.12)	15.05 (0.67)
	15	8.55 (0.59)	10.64 (0.68)	7.53 (0.60)	6.12 (1.11)	
	25	8.52 (0.59)	10.45 (0.67)	7.51 (0.60)	6.08 (1.11)	
75%	5	22.02 (1.87)	31.86 (1.77)	20.54 (2.05)	19.28 (4.81)	38.59 (1.71)
	15	21.78 (1.86)	32.73 (1.73)	20.48 (2.05)	19.91 (5.07)	
	25	21.72 (1.89)	33.01 (1.76)	20.45 (2.04)	19.96 (5.24)	

vatively. Using the true density provides roughly the nominal 90% coverage predicted by the asymptotics, validating the theory. Counterfactual density estimation is indeed a challenging task and the variance estimator may benefit from plugging in more sophisticated counterfactual density estimators such as that of Kennedy et al. (2021). DML provides bad coverage regardless of the density estimator used, underestimating the variance.

6.2 Effect of 401(k) Eligibility on Net Financial Assets

Next we consider an empirical case study to demonstrate the estimation of QTE using LDML in practice and with a variety of machine learning nuisance estimators. We use data from Chernozhukov and Hansen (2004) to estimate the QTEs of 401(k) retirement plan eligibility on net financial assets ($N = 9915$). Eligibility for 401(k) (here considered the treatment, T ; 37% are eligible in the data) is not randomly assigned, but is argued in Chernozhukov and Hansen (2004) to be ignorable conditioned on certain covariates: age, income, family size, years of education, marital status, two-earner household status, availability of defined benefit pension plan to household, IRA participation, and home ownership status. Net financial assets (the outcome, Y) are defined as the sum of IRA and 401(k) balances, bank accounts, and other interest-earning accounts and assets minus non-mortgage debt. While Chernozhukov and Hansen (2004) considered controlling for these in a low-dimensional linear specification, it is not clear whether such is sufficient to account for all confounding. Consequently, Belloni et al. (2017) considered including higher-order terms and interactions, but needed to theoretically construct a continuum of LASSO estimates and may not be able to use generic black-box regression methods. Finally, Chernozhukov et al. (2018a) considered using generic machine learning methods, but only tackled ATE estimation.

In contrast, we will use LDML to estimate and conduct inference on the QTEs of 401(k) eligibility on net assets using a variety of flexible black-box regression methods. First, to understand the effect of different choices in the application of LDML to the problem, we consider estimating the 25%, 50%, and 75% QTE while varying K in $\{5, 15, 25\}$ and varying the nuisance estimators. We consider estimating both propensity score η_2^* and conditional cumulative distribution η_1^* with each of: boosting (using *R* package gbm), LASSO (using *R* package hdm), and a one-hidden-layer neural network (using *R* package nnet). For LASSO, we use a 275-dimensional expansion of the covariates by considering higher-order terms and interactions. In each instantiation of LDML, we construct folds so to ensure a balanced distribution of treated and untreated units, we let $K' = (K-1)/2$, we use the IPW initial estimator for $\hat{\theta}_{1,\mathrm{init}}$, we normalize propensity weights to have mean 1 within each treatment group, we use estimates given by solving the grand-average estimating equation as in Definition 2, and for variance estimation we estimate J^* using IPW kernel density estimation as in Remark 4. The solution to the LDML-estimated empirical estimating equation must occur at an observed outcome Y_i and that we can find the solution using binary search after sorting the data along outcomes. We re-randomize the fold construction and repeat each instantiation 100 times. We then remove the outlying 2.5% from each end and report $\hat{\theta}^{\mathrm{mean}}$, $\widehat{\Sigma}^{\mathrm{mean}}$ as in Appendix E. The resulting estimates and standard errors are shown in Table 1. The estimates appear overall roughly stable across methods and K.

Next, we consider estimating a range of QTEs. We focus on nuisance estimation using LASSO and fix $K = 15$. We then estimate the $10\%, 11\%, \ldots, 89\%,$ and 90% quantiles and QTEs. We plot the resulting LDML estimates with 90% confidence intervals in Fig. 3 and compare these to the raw unadjusted marginal quantiles within each treatment group.

6.3 Effect of 401(k) Participation on Net Financial Assets

Next, we estimate the effect of 401(k) participation on net assets. Participation in a 401(k) plan (here considered the treatment, T ; 26% participate) is not randomly assigned: individuals with a preference for saving may save more in non-retirement accounts than others whether they were to participate in retirement savings or not. There may be many other confounding factors, such as the possibility of higher financial acumen of savers leading to higher net worth otherwise. It is unlikely that we can control for all these factors using

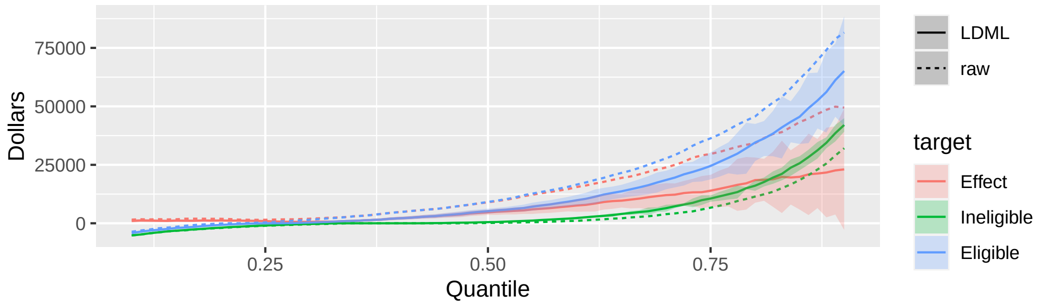


Figure 3: LDML estimates of a range of quantiles and QTEs with confidence 90% intervals and comparison to raw unadjusted marginal quantiles by treatment group.

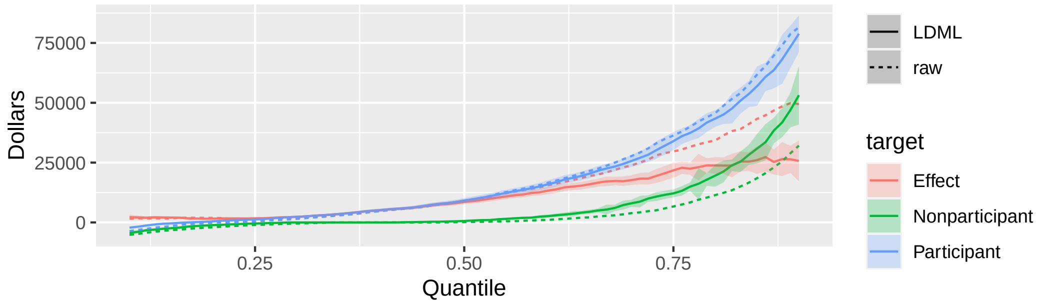


Figure 4: LDML estimates of a range of local quantiles and LQTEs with confidence 90% intervals and comparison to raw unadjusted marginal quantiles by treatment group.

observable covariates. Instead, we rely on instrumenting on eligibility since, as argued in Section 6.2, eligibility is ignorable given covariates. Additionally, one cannot participate if one is ineligible, ensuring monotonicity, and some eligible individuals do participate, ensuring relevance. Assuming that eligibility cannot affect net assets except through its effect on participation, we have that eligibility for a 401(k) (here considered as W) is valid IV. We can therefore use it to estimate local quantiles by and LQTEs of 401(k) Participation on the population of individuals that would participate if eligible.

We use LDML for the LQTE as developed in Appendix A. Again, we consider the impact of different choices in the application of LDML. We repeat the same specification as above, using each possible nuisance estimator to fit the conditional probabilities Eqs. (22) and (25). We display the results for the 25%, 50%, and 75% quantiles while varying K and the nuisance estimators in Table 2. The qualitative results regarding the stability of LDML across methods and K remain the same. Then, focusing as before on nuisance estimation using LASSO and on $K = 15$, we also estimate a range of local quantiles and QTEs, which we plot along with 90% confidence intervals in Fig. 3. Again, we compare to the raw unadjusted marginal quantiles within each treatment group.

7. Related Literature

Semiparametric Estimation, Neyman orthogonality, and Debiased Machine Learning. Our work is closely related to the classical semiparametric estimation literature on constructing $\sqrt{N}$-consistent and asymptotically normal estimators for low dimensional target parameters in the presence of infinitely dimensional nuisances, typically estimated by conventional nonparametric estimators such as kernel or series estimators (e.g., Newey, 1990, 1994; Newey et al., 1998; Ibragimov and Hasminskii, 1981; Levit, 1976; Bickel et al., 1998; Bickel, 1982; Robinson, 1988; var der Vaart, 1991; Andrews, 1994; Linton, 1996; Chen et al., 2003; Ai and Chen, 2003, 2012). Our work builds on the Neyman orthogonality condition introduced by Neyman (1959)). This condition plays a critical role in many works that go beyond the aforementioned literature, such as targeted learning (e.g., van der Laan and Rose, 2011; van der Laan and Rose, 2018), missing or censored data (e.g., Robins et al., 1994a; Robins and Rotnitzky, 1995; van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2006), inference for coefficients in high dimensional linear models (e.g., Belloni et al., 2016, 2014c; Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014; Chernozhukov et al., 2015; Ning et al., 2017), and semiparametric estimation with nuisances that involve high dimensional covariates (e.g., Belloni et al., 2017; Smucler et al., 2019; Chernozhukov et al., 2018b; Farrell, 2015; Belloni et al., 2014a,b; Bradic et al., 2019; Bravo et al., 2020).

Chernozhukov et al. (2018a) highlight the debiased machine learning (DML) approach that combines orthogonal estimating equations with cross-fitting, so that the traditional Donsker assumption on nuisance estimators can be relaxed, and a broad array of black-box machine learning algorithms can be used instead. Their work follows from a body of earlier literature that also leverage Neyman orthogonality and sample splitting (or cross-fitting) for flexible semiparametric inference (Klaassen, 1987; Zheng and van der Laan, 2011; Fan et al., 2012; Bickel, 1982; Robins et al., 2013; Schick, 1986; Robins et al., 2008; van der Laan and Rose, 2011; van der Laan and Robins, 2003). The DML aproach has been applied in numerous works on many different problems, such as heterogeneous treatment effect estimation (Kennedy, 2020; Nie and Wager, 2017; Curth et al., 2020; Semenova and Chernozhukov, 2020; Oprescu et al., 2019; Fan et al., 2020), causal effects of continuous treatments (Colangelo and Lee, 2020; Oprescu et al., 2019), instrumental variable estimation (Singh and Sun, 2019; Syrgkanis et al., 2019), partial identification (Bonvini and Kennedy, 2019; Kallus et al., 2019; Semenova, 2017; Yadlowsky et al., 2018), difference-in-difference models (Lu et al., 2019; Chang, 2020; Zimmert, 2018), off-policy evaluation (Kallus and Uehara, 2020; Demirer et al., 2019; Zhou et al., 2018; Athey and Wager, 2017), generalized method of moments (Chernozhukov et al., 2016; Belloni et al., 2018), improved machine learning nuisance estimation (Farrell et al., 2018; Cui and Tchetgen, 2019), statistical learning with nuisances (Foster and Syrgkanis, 2019), causal inference with surrogate observations (Kallus and Mao, 2020), linear functional estimation (Chernozhukov et al., 2018d,c; Bradic et al., 2019), etc. Our work complements this line of research by proposing a simple but effective way to handle estimand-dependent nuisances. This type of nuisances frequently appears in efficient estimation of complex causal effects such as QTEs, and applying DML directly would require estimating a continuum of nuisances, which is challenging in practice.

Efficient estimation of (L)QTE. Firpo (2007) first considered efficient estimation of QTE and proposed an IPW estimator based on propensity scores estimated by a logistic sieve estimator. Under strong smoothness conditions, this IPW estimator is $\sqrt{N}$-consistent and achieves the semiparametric efficiency bound. Frölich and Melly (2013) consider a weighted estimator for LQTE with weights estimated by local linear regressions using high-order kernels and show that their estimator is also semiparametrically efficient. Although these purely weighted methods bypass the estimation of nuisances that depend on the estimand, their favorable behavior is restricted to certain nonparametric weight estimators and strong smoothness requirements. Díaz (2017) proposed a Targeted Minimum Loss Estimator (TMLE) estimator for efficient QTE estimation. Built on the efficient influence function with nuisances that depends on the quantile itself, this estimator requires estimating a whole conditional cumulative distribution function, which as discussed may be very challenging in practice using flexible machine learning methods. Belloni et al. (2017) similarly consider efficient estimation of LQTE with high-dimensional covariates by using a Neyman-orthogonal estimating equation and discretizing a continuum of LASSO estimators for the estimand-dependent nuisance. In contrast, our proposed estimator can leverage a wide variety of flexible machine learning methods for the standard regression task to estimate nuisances, since we require estimating conditional cumulative distribution function only at a *single* point, which amounts to a binary regression problem.

Estimand-dependent nuisances. Besides (local) quantiles and CVaR, many efficient estimation problems involve nuisances that depends on the estimand (e.g., Tsiatis, 2006; Chen et al., 2005). Previous approaches estimate the whole continuum of the estimand-dependent nuisances either by positing simple parametric model for conditional distributions (Tsiatis, 2006, Chap 10), using sieve estimators (Chen et al., 2005), or discretizing a hypothetical continuum of regression estimators (Belloni et al., 2017). In contrast, our proposed method obviates the need to estimate infinitely many nuisances by fitting nuisances only at a preliminary estimate of the parameter of interest. This idea was briefly mentioned by Robins et al. (1994b), focusing on parametric models for nuisance estimation. Our paper rigorously develops this approach and admits flexible machine learning methods for estimating nuisances that depend on the estimand.

8. Conclusion

In many causal inference and missing data settings, the efficient influence function involves nuisances that depend on the estimand of interest. A key example provided was that of QTE under ignorable treatment assignment and LQTE estimation using an IV, where in both cases the efficient influence function depends on the conditional cumulative distribution function evaluated at the quantile of interest. This structure, common to many other important problems, makes the application of existing debiased machine learning methods difficult in practice. In quantile estimation, it requires we learn the whole conditional cumulative distribution function. To avoid this difficulty, we proposed the LDML approach, which localized the nuisance estimation step to an initial rough guess of the estimand. This was motivated by the fact that in many applications, the oracle estimating equation is asymptotically equivalent to one where the nuisance is evaluated at the true parameter value, which our localization approach targets. Assuming only standard identification conditions, Neyman orthogonality, and lax rate conditions on our nuisance estimates, we proved the LDML enjoys the same favorable asymptotics as the oracle estimator that solves the estimating equation with the *true* nuisance functions. This newly enables the practical efficient estimation of important quantities such as QTEs using machine learning.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1846210. Xiaojie Mao acknowledges support from National Natural Science Foundation of China (Grant Nos. 72322001, 72201150, and 72293561) and National Key R&D Program of China (2022ZD0116700).

References

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Chunrong Ai and Xiaohong Chen. Semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170:442–457, 2012.

Donald W. K. Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72, 1994.

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Philipp Bach, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. Doubleml-an object-oriented implementation of double machine learning in python. *J. Mach. Learn. Res.*, 23:53–1, 2022.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973, 2005.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28 (2):29–50, 2014a.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014b.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014c.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34 (4):606–619, 2016.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1): 233–298, 2017.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. High-dimensional econometrics and regularized gmm. *arXiv preprint arXiv: 1806.01888*, 2018.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*, 2014.

P. J. Bickel. On adaptive estimation. *Annals of Statistics*, 10(3):647–671, 1982.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Matteo Bonvini and Edward H Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *arXiv preprint arXiv:1912.02793*, 2019.

Jelena Bradic, Victor Chernozhukov, Whitney K. Newey, and Yinchu Zhu. Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

Francesco Bravo, Juan Carlos Escanciano, and Ingrid Van Keilegom. Two-step semiparametric empirical likelihood inference. *The Annals of Statistics*, 48(1):1–26, 2020.

Neng-Chieh Chang. Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191, 2020.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *LSE Research Online Documents on Economics*, 2003.

Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

Victor Chernozhukov and Christian Hansen. The effects of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and statistics*, 86(3):735–751, 2004.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. 2015.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

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

Victor Chernozhukov, Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*, 2018b.

Victor Chernozhukov, Whitney Newey, James Robins, and Rahul Singh. Double/de-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018c.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Learning l2 continuous regression functionals via regularized riesz representers. *arXiv preprint arXiv:1809.05224*, 8, 2018d.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

Yifan Cui and Eric Tchetgen Tchetgen. Bias-aware model selection for machine learning of doubly robust functionals. *arXiv preprint arXiv:1911.02029*, 2019.

Alicia Curth, Ahmed M Alaa, and Mihaela van der Schaar. Semiparametric estimation and inference on structural target functions using machine learning and influence functions. *arXiv preprint arXiv:2008.06461*, 2020.

Mert Demirer, Vasilis Syrgkanis, Greg Lewis, and Victor Chernozhukov. Semi-parametric efficient policy learning with continuous actions. *arXiv preprint arXiv:1905.10116*, 2019.

Iván Díaz. Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference*, 190:39–51, 2017.

Jianqing Fan, Shaojun Guo, and Ning Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, (just-accepted):1–39, 2020.

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

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *arXiv preprint arXiv:1809.09953*, 2018.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75:259–276, 2007.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Markus Frölich and Blaise Stéphane Melly. Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357, 2013.

I. A. Ibragimov and R. Z. Hasminskii. Statistical estimation : asymptotic theory. 1981.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, pages 467–475, 1994.

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

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

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.

Maximilian Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 8 (1):1–19, 2019.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Semiparametric counterfactual density estimation. *arXiv preprint arXiv:2102.12034*, 2021.

Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562, 1987.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, New York, NY, 2008.

B. Ya. Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability and Its Applications*, 20(4):723–740, 1976.

Oliver B. Linton. Edgeworth approximation for minpin estimators in semiparametric regression models. *Econometric Theory*, 12(1):30–60, 1996.

Chen Lu, Xinkun Nie, and Stefan Wager. Robust nonparametric difference-in-differences estimation. *arXiv preprint arXiv:1905.11622*, 2019.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5 (2):99–135, 1990.

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

Whitney K Newey and James L Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

Whitney K. Newey, Fushing Hsieh, and James Robins. Undersmoothing and bias corrected functional estimation. 1998.

Jerzy Neyman. Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, pages 416–44, 1959.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

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

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR, 2019.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. 89:846–866, 1994a.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994b.

James M Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, Aad van der Vaart, HEI Health Review Committee, et al. New statistical approaches to semiparametric regression with application to air pollution research. *Research report (Health Effects Institute)*, (175):3, 2013.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

Vira Semenova. Machine learning for set-identified linear models. *arXiv preprint arXiv:1712.10024*, 2017.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 2020.

Rahul Singh and Liyang Sun. De-biased machine learning for compliers. *arXiv preprint arXiv:1909.05244*, 2019.

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

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. In *Advances in Neural Information Processing Systems*, pages 15193–15202, 2019.

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

AW van der Vaart and Jon A Wellner. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer, 2023.

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

Mark J. van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality.* Springer Series in Statistics,. Springer New York, New York, NY, 2003.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data.* 2011.

Mark J van der Laan and Sherri Rose. *Targeted learning in data science.* Springer, 2018.

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

Aad var der Vaart. On differentiable functionals. *Annals of Statistics*, 19(1):178–204, 1991.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.

Michael Zimmert. Efficient difference-in-differences estimation with high-dimensional common trend confounding. *arXiv preprint arXiv:1809.01643*, 2018.

Appendix A. LDML Estimates for Local Estimating Equations using Instrumental Variable

Instead of assuming ignorable treatment assignment, we may have access to an instrumental variable (IV). We considier a binary IV denoted as $W \in \{0,1\}$ and assume that it satisfies identification conditions in Imbens and Angrist (1994) (namely, for potential treatments $T(w)$ and potential outcomes $Y(t,w)$, we have exclusion $Y(t) := Y(t,w) = Y(t,1-w)$, exogeneity $(Y(t),T(w)) \perp W \mid X$, overlap $\mathbb{P}(W=1 \mid X) \in (0,1)$, relevance $\mathbb{P}(T(1)=1) > \mathbb{P}(T(0)=1)$, and monotonicity $T(1) \geq T(0)$). We seek to use observations of $Z=(X,W,T,Y)$ to estimate *local* parameters defined by the following estimating equation conditionally on the subpopulation of compliers (*i.e.*, $T(1) > T(0)$):

$$\mathbb{P}[U(Y(1);\theta_1) + V(\theta_2) \mid T(1) > T(0)] = 0. \tag{20}$$

For example, specializing Eq. (20) to the functions $U(y;\theta_1), V(\theta_2)$ in Eq. (5) gives the *local* quantile and CVaR, which in turn gives the *local* QTE (LQTE).

Following Belloni et al. (2017), a Neyman orthogonal estimating equation for θ^* is given by

$$\psi(Z;\theta,\theta_2^{\text{aux}*},\eta_1^*(Z;\theta_1),\eta_2^*(Z)) = \begin{bmatrix}\psi_1(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z)) \\ \psi_2(Z;\theta_2^{\text{aux}*},\eta_2^*(Z))\end{bmatrix}, \tag{21}$$

where

$$\begin{aligned}\psi_1(Z;\theta,\eta_1(Z;\theta_1),\eta_2(Z)) = \Bigg(&\eta_{1,1}(Z;\theta_1) - \eta_{1,2}(Z;\theta_1) + \frac{W}{\eta_{2,1}(Z)}\left(TU(Y;\theta_1) - \eta_{1,1}(Z;\theta_1)\right) \\ &- \frac{1-W}{1-\eta_{2,1}(Z)}\left(TU(Y;\theta_1) - \eta_{1,2}(Z;\theta_1)\right)\Bigg) \times \frac{1}{\theta_2^{\text{aux}}} + V(\theta_2), \\ \psi_2(Z;\theta_2^{\text{aux}},\eta_2(Z)) = \eta_{2,2}(Z) - \eta_{2,3}(Z) &+ \frac{W}{\eta_{2,1}(Z)}\left(T - \eta_{2,2}(Z)\right) - \frac{1-W}{1-\eta_{2,1}(Z)}\left(T - \eta_{2,3}(Z)\right) - \theta_2^{\text{aux}}.\end{aligned}$$

with nuisance functions

$$\eta_1^*(Z;\theta_1) = \begin{bmatrix}\mathbb{E}\left[TU(Y;\theta_1) \mid X, W=1\right] \\ \mathbb{E}\left[TU(Y;\theta_1) \mid X, W=0\right]\end{bmatrix}, \quad \eta_2^*(Z) = \begin{bmatrix}\mathbb{P}(W=1 \mid X) \\ \mathbb{P}(T=1 \mid X, W=1) \\ \mathbb{P}(T=1 \mid X, W=0)\end{bmatrix}. \tag{22}$$

Here the second estimating equation $\mathbb{E}\left[\psi_2(Z;\theta_2^{\text{aux}*},\eta_2^*(Z))\right] = 0$ identifies the compliance probability, denoted by the following auxiliary parameter $\theta_2^{\text{aux}*}$:

$$\theta_2^{\text{aux}*} = \mathbb{E}\left[\mathbb{P}(T=1 \mid X, W=1) - \mathbb{P}(T=1 \mid X, W=0)\right] = \mathbb{P}(T(1) > T(0)).$$

By redefining $\tilde\theta_1 = \theta_1$, $\tilde\theta_2 = (\theta_2, \theta_2^{\text{aux}})$, and $\tilde\theta = \left(\tilde\theta_1, \tilde\theta_2\right)$, the estimating equation becomes

$$\mathbb{P}\left[\psi(Z;\tilde\theta,\eta_1^*(Z;\tilde\theta_1),\eta_2^*(Z))\right] = \mathbf{0}, \tag{23}$$

which apparently fits into our general framework in Eq. (1). Therefore, we can directly apply our LDML algorithm in Section 2.2 to estimate the local parameters $\theta^* = (\theta_1^*, \theta_2^*)$. We can also use the theory in Sections 3 and 4 to analyze the asymptotic distribution of the resulting estimators and estimate their asymptotic variances.

A.1 Estimating Local Quantiles

In particular, we take the local quantile estimation as an example, namely, the solution θ_1^* to the local estimating equation in Eq. (20) with

$$U\left(Y;\theta_1\right)=\mathbb{I}\left[Y\leq\theta_1\right],\quad V\left(\theta_2\right)=-\gamma. \tag{24}$$

Its orthogonal estimating equation involves the following nuisance functions:

$$\eta_1^*(Z;\theta_1)=\begin{bmatrix}\mathbb{P}\left(T=1,Y\leq\theta_1\mid X,W=1\right)\\ \mathbb{P}\left(T=1,Y\leq\theta_1\mid X,W=0\right)\end{bmatrix}. \tag{25}$$

For better readability, we denote the event of being a complier, *i.e.*, $T(1)>T(0)$, as $\mathcal{C}$, the nuisance functions as $\tilde{\pi}^*(X)=\mathbb{P}(W=1\mid X)$, $\nu_w^*(X)=\mathbb{P}\left(T=1\mid X,W=w\right)$, and $\tilde{\mu}_w^*(X;\theta_1)=\mathbb{P}\left(T=1,Y\leq\theta_1\mid X,W=w\right)$ for $w\in\{0,1\}$. We fit estimators for the nuisance functions based on the sample-splitting scheme given in Definition 1, which we denote as $\hat{\tilde{\pi}}^{(k)}(X)$, $\hat{\nu}_w^{(k)}(X)$ and $\hat{\tilde{\mu}}^{(k)}(X;\hat{\theta}_{1,\text{init}})=(\hat{\tilde{\mu}}_1^{(k)}(X;\hat{\theta}_{1,\text{init}}),\hat{\tilde{\mu}}_0^{(k)}(X;\hat{\theta}_{1,\text{init}}))$ respectively for $k=1,\ldots,K$. Finally, we obtain the estimator $\hat{\theta}=\left(\hat{\theta}_1,\hat{\theta}_2^{\text{aux}}\right)$ by searching approximate solutions over $\Theta=\Theta_1\times\Theta_2\subseteq\mathbb{R}\times\mathbb{R}$ to the empirical estimating equations in Definition 2 or Definition 5, specialized to Eqs. (21) and (24).

We next assume a strong form of the overlap and relevance assumptions and specify the convergence rates of the initial estimator and nuisance estimators. We again consider a generic treatment level $t\in\{0,1\}$ in these two assumptions.

Assumption 7 (Strong Overlap and Relevance Assumptions)**.** Assume that there exists a positive constant $\epsilon>0$ such that for any $\mathbb{P}\in\mathcal{P}_N$, $\epsilon\leq\tilde{\pi}^*(X)\leq1-\epsilon$ holds almost surely, and $\theta_2^{\text{aux}\,*}\geq\epsilon$.

Assumption 8 (Nuisance Estimation Rates)**.** Assume that for any $\mathbb{P}\in\mathcal{P}_N$: with probability at least $1-\Delta_N$, for $w=0,1$,

$$\left\|\left\{\mathbb{P}\left[\hat{\tilde{\mu}}_w^{(k)}\left(X;\hat{\theta}_{1,\text{init}}^{(k)}\right)\right)-\tilde{\mu}_w^*\left(X;\hat{\theta}_{1,\text{init}}^{(k)}\right)\right)\right]^2\right\}^{1/2}\right\|\leq\tilde{\rho}_{\mu,N},\ \left\{\mathbb{P}\left[\hat{\nu}_w^{(k)}(X)-\nu_w^*(X)\right]^2\right\}^{1/2}\leq\tilde{\rho}_{\nu,N},$$

$$\left\{\mathbb{P}\left[\hat{\tilde{\pi}}^{(k)}(X)-\tilde{\pi}^*(X)\right]^2\right\}^{1/2}\leq\tilde{\rho}_{\pi,N},\quad|\hat{\theta}_{1,\text{init}}^{(k)}-\theta_1^*|\leq\tilde{\rho}_{\theta,N},$$

and $\epsilon\leq\hat{\tilde{\pi}}^{(k)}(X)\leq1-\epsilon,0\leq\hat{\tilde{\mu}}_w^{(k)}\left(X;\hat{\theta}_{1,\text{init}}^{(k)}\right)\right)\leq1$, $0\leq\hat{\nu}_w^{(k)}(X)\leq1$ almost surely.

In the following theorem, we derive the asymptotic distribution of the local quantile estimator, which is proved by verifying all assumptions in Theorem 1.

Proposition 3 (LDML for Local Quantile)**.** *Fix $t=1$ and let $\Theta=(\Theta_1,\Theta_2)\subseteq\mathbb{R}^2$ be a compact set where $\theta_2^{\text{aux}}\geq\epsilon$ for any $\theta_2^{\text{aux}}\in\Theta_2$ and ϵ given in Assumption 7. Let $(\hat{\theta}_1,\hat{\theta}_2^{\text{aux}})$ be the LDML estimator given in either Definition 2 or Definition 5, specialized to Eqs.* (21) *and* (24)*. Suppose that there exist constants c',C such that the following conditions hold for any instance $\mathbb{P}\in\mathcal{P}_N$:*

i. *Conditions i. (with c_1), ii., v. (with c_5, c_6) and condition iii. of Assumption 3 for the estimating equation in Eqs. (21) and (24).*

ii. *For any $\theta_1 \in \Theta_1$, the distribution function of $Y(t)$ for compliers, denoted as $F_t(\theta_1 \mid \mathcal{C})$, is twice continuously differentible. Its first two order derivatives $f_t(\theta_1 \mid \mathcal{C})$ and $\dot{f}_t(\theta_1 \mid \mathcal{C})$ satisfy that $f_t(\theta_1 \mid \mathcal{C}) \le c_1'$, $\left|\dot{f}_t(\theta_1 \mid \mathcal{C})\right| \le c_2'$ for any $\theta_1 \in \Theta_1$, and $f_t(\theta_1^* \mid \mathcal{C}) \ge c_3' > 0$.*

iii. $2\|\mathbb{P}\left[\psi(Z;\theta,\theta_2^{\mathrm{aux}},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\| \ge c_2$ *for all* $\theta = (\theta_1, \theta_2^{\mathrm{aux}}) \in \Theta$ *such that* $\|\theta - \theta^*\| \ge \frac{c_3}{2\sqrt{d}c_{\mathrm{Lip}}}$ *where* $c_{\mathrm{Lip}} := \max\left\{\sqrt{\left(\frac{c_1'}{\epsilon^2}\right)^2 + \left(\frac{c_2'}{\epsilon}\right)^2}, \sqrt{\left(\frac{2}{\epsilon^3}\right)^2 + \left(\frac{c_1'}{\epsilon^2}\right)^2}\right\}$.

iv. *For any $\theta_1 \in \mathcal{B}(\theta_1^*; \max\{\frac{4\tilde{\rho}_{\pi,N}}{\epsilon^2(1-\epsilon)\delta_N}, \rho_{\theta,N}\}) \cap \Theta$ and $w \in \{0,1\}$, the conditional distribution of $Y(t)$ given $X, T(w) = 1$, denoted as $F_{t,w}(\theta_1 \mid X)$, is twice differentiable almost surely with first two order derivatives $f_{t,w}(\theta_1 \mid X)$ and $\dot{f}_{t,w}(\theta_1 \mid X)$ that satisfy $f_{t,w}(\theta_1 \mid X) \le C$ and $\left|\dot{f}_{t,w}(\theta_1 \mid X)\right| \le C$ almost surely.*

v. *The nuisance estimator convergence rates satisfy that $\tilde{\rho}_{\pi,N} \le \frac{\delta_N^3}{\log N}$, $\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N} \le \frac{\delta_N^2}{\log N}$, $\tilde{\rho}_{\pi,N}\left(\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N}\right) \le \frac{\epsilon^4(1-\epsilon)^3}{4\left(\epsilon^3+(1-\epsilon)^3\right)}\delta_N N^{-1/2}$, $\tilde{\rho}_{\pi,N}\tilde{\rho}_{\nu,N} \le \frac{\epsilon^3(1-\epsilon)^3}{8\left(\epsilon^3+(1-\epsilon)^3\right)}\delta_N N^{-1/2}$ with δ_N satisfying that $\delta_N \le \frac{\epsilon^3(1-\epsilon)^2}{4C+3\epsilon^2(1-\epsilon)}$, $\frac{\delta_N}{\log N} \le \frac{1}{C_\epsilon}$ for a positive constant C_ϵ given in Eq. (44) .*

Then $(\hat{\theta}_1, \hat{\theta}_2^{\mathrm{aux}})$ satisfies the conclusion of Theorem 1 for $\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ given in Eq. (21) and

$$J^{*-1} = \begin{bmatrix} \frac{1}{f_1\left(\theta_1^* \mid \mathcal{C}\right)} & -\frac{\gamma}{\theta_2^{\mathrm{aux}\,*} f_1\left(\theta_1^* \mid \mathcal{C}\right)} \\ 0 & -1 \end{bmatrix}.$$

In particular, the local quantile estimator $\hat{\theta}_1$ is asymptotically linear with the following influence function:

$$\frac{1}{f_1\left(\theta_1^* \mid \mathcal{C}\right)}\psi_1(Z_i;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i)) - \frac{\gamma}{\theta_2^{\mathrm{aux}\,*} f_1\left(\theta_1^* \mid \mathcal{C}\right)}\psi_2(Z_i;\theta_2^{\mathrm{aux}\,*},\eta_2^*(Z_i)),$$

where $\psi_1(Z_i;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i))$ and $\psi_2(Z_i;\theta_2^{\mathrm{aux}\,*},\eta_2^*(Z_i))$ are given in Eq. (21). Analogous conclusion for local quantiles of $Y(0)$ holds when all assumptions above hold for $t = 0$.

Appendix B. LDML Estimates for Expectiles

We can also apply our method and analysis to estimating the γ-expectile θ_1 of $Y(1)$, as defined in Eq. (6). Instantiating Eq. (7) for expectiles and rearranging, we get the following efficient estimating function from incomplete data:

$$\begin{aligned}
&\psi(Z;\theta_1,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\\
=&\frac{\mathbb{I}(T=1)}{\eta_{2,2}^*(Z)}\left[(1-\gamma)\left(Y-\eta_{2,1}^*(Z)\right)-(1-2\gamma)\left(\max\left(Y-\theta_1,0\right)-\eta_1^*(Z;\theta_1)\right)\right]\\
&+\left[(1-\gamma)\eta_{2,1}^*(Z)-(1-2\gamma)\eta_1^*(Z;\theta_1)\right], \qquad (26)\\
\text{where}\qquad &\eta_1^*(Z;\theta_1)=\mathbb{E}\left[\max(Y-\theta_1,0)\mid X,T=1\right],\\
&\eta_2^*(Z;\theta_1)=\begin{bmatrix}\mathbb{E}\left[Y\mid X,T=1\right]\\ \mathbb{P}\left(T=1\mid X\right)\end{bmatrix}.
\end{aligned}$$

The next result gives the asymptotic behavior of LDML applied to these equations.

Proposition 4. *Fix $t=1$ and let the estimator $\hat\theta_1$ be given by applying either Definition 2 or Definition 5 to the estimating function in Eq.* (26). *Suppose Assumptions 5 and 6 hold and there exist positive constants C, c_1', c_2', such that for any $\mathbb{P}\in\mathcal{P}_N$, the following conditions hold:*

i. Conditions i. (with c_1), ii., v. (with c_5, c_6) of Assumption 2, condition iii. of Assumption 3, and condition vii. of Theorem 3 for the estimating function in Eq. (26) *and the corresponding nuisance estimators.*

ii. $F_t(\theta_1)$ is continuous at θ_1^*, and $|-(1-2\gamma)F_t(\theta_1^*)-\gamma|\ge c_1'>0$. Moreover, for any $\theta\in\Theta$ such that $\|\theta-\theta^*\|\ge\frac{c_1'}{2}$, $2\left|\mathbb{P}\left[U(Y(t);\theta_1)\right]\right|\ge c_2'$ for $U(Y(t);\theta_1)$ given in Eq. (6).

iii. At any $\theta_1\in\mathcal{B}(\theta^*;\max\{\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi},\rho_{\theta,N}\})\cap\Theta_1$, $F_t(\theta_1\mid X)$ is almost surely differentiable with first-order derivative $f_t(\theta_1\mid X)$, and second-order derivative $\dot f_t(\theta_1\mid X)$ that satisfies $f_t(\theta_1\mid X)\le C$ and $\left|\dot f_t(\theta_1\mid X)\right|\le C$ almost surely;

iv. For any $\theta_1\in\Theta_1$,

$$\left\{\mathbb{P}\left[\mathbb{E}[\max\{Y(t)-\theta_1,0\}\mid X]\right]^2\right\}^{1/2}\le C,\qquad\left\{\mathbb{P}\left[\mathbb{E}\left[Y(t)\mid X\right]\right]^2\right\}^{1/2}\le C.$$

Then $\hat\theta_1$ satisfies the conclusion of Theorem 1 for $\psi(Z;\theta_1^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))$ given in Eq. (26) *and $J^*=-\gamma-(1-2\gamma)F_t(\theta_1^*)$. Analogous conclusion for expectile of $Y(0)$ holds when all assumptions above hold for $t=0$.*

When constructing confidence intervals, we only need to estimate $F_t(\theta_1^*)$ to estimate J^*. This can be easily estimated by the inverse propensity reweighted estimator

$$\frac{1}{N}\sum_k\sum_{i\in\mathcal{D}_k}\frac{\mathbb{I}\left[T_i=t\right]}{\hat\pi^{(k)}(t\mid X_i)}\mathbb{I}\left[Y\le\hat\theta_1\right].$$

Alternatively, it can be estimated by an imputation estimator based on $\hat\mu^{(k)}$ or a LDML estimator that uses both $\hat\pi^{(k)}$ and $\hat\mu^{(k)}$ (see Remark 4).

Appendix C. Theoretical Analysis of IPW Initial Estimator

In this part, we show that the IPW initial estimator given in Definition 4 can satisfy the conditions on $\hat{\theta}_{1,\text{init}}$ in Assumption 6.

Proposition 5 (IPW Initial Estimator Rate)**.** *Fix $t = 1$ and let the initial estimator $\hat{\theta}^{(k)}_{1,init}$ be constructed according to Definition 4 for $k = 1, \ldots, K$. Assume the following (for $t = 1$):*

i. For each $k \in \{1, \ldots, K\}$ and $l \in \mathcal{H}_{k,1}$, $\hat{\pi}^{(k,l)}$ satisfies the same conditions as for $\hat{\pi}^{(k)}$ in Assumption 6.

ii. Conditions ii., iii., and v. in Theorem 3 (with constants c_2 to c_4 and C) hold.

iii. There exists a nuisance realization set Π_N that contains the true propensity score π^* and also the propensity score estimators $\hat{\pi}^{(k,l)}$ for $k = 1, \ldots, K$ and $l \in \mathcal{H}_{k,1}$ with at least probability $1 - \Delta_N$. Moreover, any $\pi \in \Pi_N$ satisfies that $\pi(t \mid X) \geq \epsilon_\pi$.

iv. For each $\pi \in \Pi_N$, the function class $\mathcal{G}_\pi = \{(X, T, Y) \mapsto \frac{\mathbb{I}[T=t]}{\pi(t|X)} U_j(Y; \theta_1) + V_j(\theta_2) : j = 1, \ldots, d, \theta \in \Theta\}$ is suitably measurable and its uniform covering entropy satisfies the following condition: for positive constants a', v' and $q' > 2$, $\sup_{\mathbb{Q}} \log N(\epsilon \|G_\pi\|_{\mathbb{Q},2}, \mathcal{G}_\pi, \|\cdot\|_{\mathbb{Q},2}) \leq v' \log(a'\epsilon)\ \forall \epsilon \in (0, 1]$, where G_π is a measurable envelope for $\mathcal{G}_\pi$. There exists a positive constant c_8 such that for any $\mathbb{P} \in \mathcal{P}_N$, $\|G_\pi\|_{\mathbb{P},q'} \leq c_8$.

v. $\left(\frac{K'}{N}\right)^{1/2} \log\left(\frac{K'}{N}\right) + \left(\frac{K'}{N}\right)^{1-\frac{1}{q'}} \log\left(\frac{K'}{N}\right) \leq \delta_N \rho_{\pi,N}$;

Then there exists a constant c that only depends on pre-specified constants in the conditions above such that with probability $1 - c(\log N)^{-1}$, $\rho_{\theta,N} \leq 2c_3^{-1}\left(C\sqrt{d}\epsilon_\pi^{-1} + 1\right)\rho_{\pi,N}$.

In Remark 3, we discuss the corresponding rate conditions on other nuisance estimators when using the IPW initial estimator, based on the conclusion in Proposition 5.

Appendix D. An Alternative LDML Estimator

In Definition 2, we construct an LDML estimator by first averaging estimates of the equation in Eq. (11) over all folds and then solving the grand-average equation approximately. Below we provide an alternative LDML estimator that first solves the estimate of Eq. (11) from each fold separately and then averages these solutions.

Definition 5 (LDML2)**.** For $k = 1, \ldots, K$, construct $\hat{\theta}^{(k)}$ by (approximately) solving

$$\overline{\Psi}^{(k)}(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} \psi(Z_i; \theta, \hat{\eta}_1^{(k)}(Z_i; \hat{\theta}^{(k)}_{1,\text{init}}), \hat{\eta}_2^{(k)}(Z_i)) = 0. \tag{27}$$

In fact, we allow for an approximate least-squares solution, which is useful if the empirical estimating equation has no exact solution. Namely, we let $\hat{\theta}^k$ be any satisfying

$$\|\overline{\Psi}^{(k)}(\hat{\theta}^{(k)})\| \leq \inf_{\theta \in \Theta} \|\overline{\Psi}^{(k)}(\theta)\| + \varepsilon_N. \tag{28}$$

Then, we let the final estimator be

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)}. \tag{29}$$

We can easily follow previous proofs for the LDML estimator in Definition 2 to show that Theorems 1 to 3 and Proposition 2 also apply to $\hat{\theta}$ in Definition 5, provided that ϵ_N in Eq. (28) is $o\left(N^{-1/2}\right)$ (*i.e.*, condition iii. in Assumption 3). For example, we demonstrate this at the end of the proof for Theorem 1. Thus the two LDML estimators in Definition 2 and Definition 5 are asymptotically equivalent.

Appendix E. Practical Considerations

The proposed LDML estimator $\hat{\theta}$ in Definition 2 or Definition 5 relies on nuisance estimates based on random sample splitting (Definition 1). Although the uncertainty due to sample splitting does not affect the asymptotic theory, it may influence the finite-sample performance of the LDML estimator.

To make the results more robust to sample splitting, we may consider aggregating the estimates over different random splitting realizations. In particular, it is possible to use many other different ways of splitting data. For example, in both Definitions 2 and 5 we may average more than just K solutions or equations. For each k, we can permute over all $\binom{K-1}{K'}$ splits of $\{1, \dots, K\} \setminus \{k\}$ into K' and $K - 1 - K'$ folds used for fitting $\hat{\theta}^{(k)}_{1,\mathrm{init}}$ and $\hat{\eta}^{(k)}_1(\cdot; \hat{\theta}^{(k)}_{1,\mathrm{init}}), \hat{\eta}^{(k)}_2$. Or, we could even permute over all $\sum_{K'=1}^{K-2} \binom{K-1}{K'}$ ways to split $\{1, \dots, K\} \setminus \{k\}$ into two. Or, we can even repeat the initial random splitting into K folds many times over and average the resulting estimates from either Definition 5 or 2, or take their median to avoid outliers, or solve the grand-mean of estimating equations. All of these procedures can provide improved finite-sample performance in practice as they can only reduce variance without affecting bias, and we do recommend these, but they have no effect on the leading asymptotic behavior, which remains the same whether you use one or more splits of the data into folds and/or one or more splits of $\{1, \dots, K\} \setminus \{k\}$ into two.

With estimates from multiple random splitting realizations, we may also improve variance estimation and to account for the variance due to random splitting. In particular, letting $\hat{\theta}_s$, $\widehat{\Sigma}_s$ be the parameter and variance estimates for each run of LDML for $s = 1, \dots, S$, we can let $\hat{\theta}^{\mathrm{mean}} = \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}_s$ and $\widehat{\Sigma}^{\mathrm{mean}} = \frac{1}{S} \sum_{s=1}^{S} (\widehat{\Sigma}_s + \frac{1}{S} (\hat{\theta}_s - \hat{\theta}^{\mathrm{mean}})(\hat{\theta}_s - \hat{\theta}^{\mathrm{mean}})^\top)$ be the final parameter and variance estimates. Like $\hat{\theta}^{\mathrm{mean}}$, the first term in $\widehat{\Sigma}^{\mathrm{mean}}$ reduces the variance in the estimate $\widehat{\Sigma}_s$ itself. The second term in $\widehat{\Sigma}^{\mathrm{mean}}$ accounts for the variance of $\hat{\theta}^{\mathrm{mean}}$ due to random splitting. Notice that the second term vanishes as $S \to \infty$; indeed then $\hat{\theta}^{\mathrm{mean}}$ has no variance due to random splitting as it is fully averaged over. Because $\hat{\theta}_s$ are each consistent, the second term also vanishes as $N \to \infty$. Removing the $\frac{1}{S}$ factor in the second term we can instead get an estimate of the variance of each single $\hat{\theta}_s$, rather than of $\hat{\theta}^{\mathrm{mean}}$, accounting for random splitting. This procedure extends a similar proposal by Chernozhukov et al. (2018a) for inference in linear estimating equations.

Appendix F. Comparison with Chernozhukov et al. (2018a)

Our proof of Theorem 1 and the proof of Theorem 3.3 in Chernozhukov et al. (2018a) are overall similar, but critically differ in Step II. In Step II, both proofs are based on the following decomposition:

$$\|J^{*-1}\sqrt{N}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]+\sqrt{N}(\hat{\theta}-\theta^*)\|\leq\varepsilon_N N^{1/2}+2\mathcal{I}_4+2\mathcal{I}_5, \tag{30}$$

where

$$\mathcal{I}_4\coloneqq\sqrt{N}\sup_{r\in(0,1),(\eta_1(\cdot;\theta_1'),\eta_2)\in\mathcal{T}_N}\|\partial_r^2 f(r;\hat{\theta},\eta_1(\cdot;\theta_1'),\eta_2)\|,$$
$$\mathcal{I}_5\coloneqq\mathbb{G}_N\left[\psi(Z;\hat{\theta},\hat{\eta}_1(Z,\hat{\theta}_{1,\text{init}}),\hat{\eta}_2(Z))-\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\|,$$

$\mathcal{I}_5=O_{\mathbb{P}}(\delta_N)$ is proved analogously in both proofs, and $\varepsilon_N N^{1/2}=O(\delta_N)$ is assumed in both proofs.

However, our proof and the proof in Chernozhukov et al. (2018a) assume different rate on λ_N' in Assumption 3 and thus $\mathcal{I}_4=\sqrt{N}\lambda_N'(\hat{\theta})$:

$$\text{Our condition}\quad \lambda_N'(\theta)\leq\left(\|\hat{\theta}-\theta^*\|+N^{-1/2}\right)\delta_N, \tag{31}$$
$$\text{Condition in Chernozhukov et al. (2018a)}\quad \lambda_N'(\theta)\leq N^{-1/2}\delta_N. \tag{32}$$

Under our condition, $\mathcal{I}_4\leq\left(\sqrt{N}\|\hat{\theta}-\theta^*\|+1\right)\delta_N$, then jointly considering the left hand side and right hand side in Eq. (30) gives $\|\hat{\theta}-\theta^*\|=O_p(N^{-1/2})$, which in turn implies that $\mathcal{I}_4=O(\delta_N)$, and thus the asserted conclusion in Theorem 1. In contrast, the counterpart condition in Chernozhukov et al. (2018a) guarantees that $\mathcal{I}_4=O(\delta_N)$ directly without needing to consider both sides of Eq. (30) jointly.

Now we use the example of estimating equation for incomplete data to show that the condition Eq. (32) in Chernozhukov et al. (2018a) generally requires stronger conditions for the convergence rates of nuisance estimators than our condition Eq. (31).

According to Eq. (43), under suitable regularity conditions,

$$\|\partial_r^2 f(r;\hat{\theta},\mu(X,T;\theta_1'),\pi)\|=O(\rho_{\pi,N}\rho_{\mu,N})+O_p(\rho_{\pi,N}\rho_{\theta,N})+O(\|\hat{\theta}-\theta^*\|^2)+O(\rho_{\pi,N}\|\hat{\theta}_1-\theta_1^*\|)$$

Since Step I in the proof of Theorem 1 already proves that $\|\hat{\theta}-\theta^*\|\leq\frac{\rho_{\pi,N}}{\delta_N}$, we need $\rho_{\pi,N}\rho_{\mu,N}\leq\delta_N N^{-1/2}$, $\rho_{\pi,N}\rho_{\theta,N}\leq\delta_N N^{-1/2}$, and $\rho_{\pi,N}\leq\delta_N^2$ to guarantee our condtion. Thus our condition in Eq. (31) only requires that the product error rates to vanish faster than $O(N^{-1/2})$, which is common in debiased machine learning for linear estimating equation (Chernozhukov et al., 2018a).

In contrast, to guarantee the condition in Chernozhukov et al. (2018a) given in Eq. (32), we need to assume that $\rho_{\pi,N}\leq\delta_N^{3/2}N^{-1/4}$, besides the conditions on product error rates. Therefore, following the proof in Chernozhukov et al. (2018a) directly will require the propensity score to converge faster than $O(N^{-1/4})$, no matter how fast the initial estimator $\hat{\theta}_{1,\text{init}}$ and the regression estimator $\hat{\mu}(\cdot,\hat{\theta}_{1,\text{init}})$ converge.

Appendix G. Proofs

G.1 Proofs for Section 2

Proof for Proposition 1. For any $\theta = (\theta_1, \theta_2)$ such that $(\theta, \eta_1^*(\cdot, \theta)) \in \mathcal{N}$, the asserted Fréchet differentiability and orthogonality condition imply that

$$\begin{aligned}&\|\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z,\theta_1),\eta_2^*(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] \\ &- \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}(\theta-\theta^*)\| = o(\|\theta-\theta^*\|).\end{aligned}$$

This means that $J^* = \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}$. □

G.2 Proofs for Section 3

Proof for Theorem 1. Fix any sequence $\{P_N\}_{N\geq 1}$ that generates the observed data $(Z_1, \ldots, Z_N)$ and satisfies that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, to prove that the asserted conclusion holds uniformly over $P \in \mathcal{P}_N$, we only need to prove

$$\sqrt{N}\Sigma^{-1/2}(\hat{\theta}-\theta^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Sigma^{-1/2}\left[J^{*-1}\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] + O_{P_N}(\rho_N) \xrightarrow{d} \mathcal{N}(0, I_d).$$

For $k = 1, \ldots, K$, we use $\mathbb{P}_{N,k}$ to represent the empirical average operator based on $\mathcal{D}_k$. For example, $\mathbb{P}_{N,k}\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] = \frac{1}{|\mathcal{D}_k|}\sum_{i\in\mathcal{D}_k}\psi(Z_i;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i))$. Analogously, $\mathbb{P}_N$ is the empirical average operator for the whole dataset, i.e., $\mathbb{P}_N f(Z) = \frac{1}{N}\sum_{i=1}^{N} f(Z_i)$. $\mathbb{G}_{N,k}$ is the empirical process operator $\sqrt{N}\left(\mathbb{P}_{N,k} - \mathbb{P}\right)$. Moreover, for a given N, $\mathbb{P}_{N,k}$, $\mathbb{P}_N$ and the population average operator $\mathbb{P}$ are all derived from the underlying true distribution P_N, but we supress such dependence for ease of notation. Throughout the proof, we condition on the event $(\hat{\eta}_1(\cdot, \hat{\theta}_{1,\text{init}}), \hat{\eta}_2(\cdot)) \in \mathcal{T}_N$, which happens with at least P_N-probability $1 - \Delta_N$ according to Assumption 3 condition i.. All statements involving $o(\cdot)$, $O_{P_N}(\cdot)$ or $\lesssim$ notations in this proof depend on only constants pre-specified in Assumptions 2 and 3, and do not depend on constants specific to the instance P_N. This should be clear from the proof, and the fact that the maximal inequality in Lemma 6.2 of Chernozhukov et al. (2018a) only depend on pre-specified parameters. Here we prove the asymptotic distribution of $\hat{\theta}$ given in Definition 2 first.

Step I: Prove a preliminary convergence rate for $\hat{\theta}$: $\|\hat{\theta} - \theta^*\| \leq \tau_N$ with P_N-probability $1 - o(1)$. Here we prove this by showing that with P_N-probability $1 - o(1)$,

$$\left\|\mathbb{P}\left[\psi(Z;\hat{\theta},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\right\| = o(\tau_N) \tag{33}$$

so that Assumption 2 implies that $\|J^*(\hat{\theta}-\theta^*)\| \wedge c_2 = o(\tau_N)$. Since the singular values of J^* are lower bounded by $c_3 > 0$, we can conclude that with P_N-probability $1 - o(1)$, $\|\hat{\theta} - \theta^*\| \leq \tau_N$ for N exceeding an instance-independent threshold.

In order to prove Eq. (33), we use the following decomposition:

$$\mathbb{P}\left[\psi(Z;\hat{\theta},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] = (a) + (b) + (c) + (d) + (e),$$

where

$$(a) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left[\psi(Z;\hat{\theta},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] - \mathbb{P}\left[\psi(Z;\hat{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right],$$

$$(b) = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbb{P}\left[\psi(Z;\hat{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}_{N,k}\left[\psi(Z;\hat{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\right\},$$

$$(c) = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbb{P}_{N,k}\left[\psi(Z;\hat{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}_{N,k}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\right\},$$

$$(d) = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbb{P}_{N,k}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\right\},$$

$$(e) = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbb{P}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\right\}.$$

Denote $\mathcal{I}_{1,k} = \sup_{\theta\in\Theta}\|\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] - \mathbb{P}\left[\psi(Z;\theta,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\|$ and $\mathcal{I}_{2,k} = \sup_{\theta\in\Theta}\|\mathbb{P}_{N,k}\left[\psi(Z;\theta,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\|$. Then obviously,

$$(a) + (e) \le \frac{2}{K}\sum_{k=1}^{K}\mathcal{I}_{1,k}, (b) + (d) \le \frac{2}{K}\sum_{k=1}^{K}\mathcal{I}_{2,k}.$$

Moreover, according to Eq. (15),

$$\begin{aligned}
(c) &\le \frac{1}{K}\sum_{k=1}^{K}\|\mathbb{P}_{N,k}\left[\psi(Z;\hat{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\| + \frac{1}{K}\sum_{k=1}^{K}\|\mathbb{P}_{N,k}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\| \\
&\le \frac{2}{K}\sum_{k=1}^{K}\|\mathbb{P}_{N,k}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\| + \varepsilon_N \\
&\le \frac{2}{K}\sum_{k=1}^{K}\left\|\mathbb{P}_{N,k}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right]\right\| \\
&+ \frac{2}{K}\sum_{k=1}^{K}\left\|\mathbb{P}\left[\psi(Z;\theta^*,\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\right\| + \varepsilon_N \\
&\le \frac{2}{K}\sum_{k=1}^{K}\mathcal{I}_{1,k} + \frac{2}{K}\sum_{k=1}^{K}\mathcal{I}_{2,k} + \varepsilon_N.
\end{aligned}$$

Therefore,

$$\mathbb{P}\left[\psi(Z;\hat{\theta},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] \le \frac{4}{K}\sum_{k=1}^{K}\mathcal{I}_{1,k} + \frac{4}{K}\sum_{k=1}^{K}\mathcal{I}_{2,k} + \varepsilon_N.$$

Note that Assumption 3 condition ii. implies that $\mathcal{I}_{1,k} \le \delta_N\tau_N$ and the Assumption 3 condition iii. implies that $\varepsilon_N \le \delta_N N^{-1/2} = o(\tau_N)$.

To bound $\mathcal{I}_{2,k}$, note that conditionally on $\hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)$, the function class $\mathcal{F}_{\hat{\eta}^{(k)}, \hat{\theta}_{1,\text{init}}^{(k)}} = \{\psi_j(\cdot; \theta, \hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(\cdot)) : j = 1, \dots, d, \theta \in \Theta\}$ satisfies the asserted entropy condition in Assumption 2, and has envelope $F_{1,\hat{\eta}^{(k)}, \hat{\theta}_{1,\text{init}}^{(k)}}$ that satisfies

$$\sup_{\theta\in\Theta} \mathbb{P}\left[\psi(Z; \theta, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))\right]^2 \leq \mathbb{P}\left[F^2_{1,\hat{\eta}^{(k)}, \hat{\theta}_{1,\text{init}}^{(k)}}\right] < C_{q,c_7}$$

for a positive constant C_{q,c_7} that only depends on q and c_7 specified in Assumption 2.

Then conditionally on $\hat{\theta}_{1,\text{init}}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)$, we can use Lemma 6.2 eq. (A.1) in Chernozhukov et al. (2018a) to prove that with P_N-probability $1 - o(1)$,

$$\sup_{\theta\in\Theta} \mathbb{G}_{N,k}\left[\psi(Z; \theta, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))\right] \lesssim \log N(1 + N^{-1/2+1/q}). \tag{34}$$

This also holds unconditionally according to Lemma 6.1 of in Chernozhukov et al. (2018a). This further implies that $\mathcal{I}_{2,k} \lesssim N^{-1/2} \log N(1+N^{-1/2+1/q}) = o(\tau_N)$. Thus $\mathbb{P}\left[\psi(Z; \hat{\theta}, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z))\right] \leq 4\delta_N \tau_N + 4N^{-1/2} \log N(1 + N^{-1/2+1/q}) + \delta_N N^{-1/2} = o(\tau_N)$.

Step II: Linearization and $\sqrt{N}$−Consistency. In Step I, we proved that $\|\hat{\theta} - \theta^*\| \leq \tau_N$ with P_N-probability $1 - o(1)$. Conditioned on this event, we will show that

$$\begin{aligned} &\|\sqrt{N}\mathbb{P}_N[\psi(Z; \theta^*, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z))] + \sqrt{N} J^*(\hat{\theta} - \theta^*)\| \\ \leq &\varepsilon_N N^{1/2} + \mathcal{I}_3 + \mathcal{I}_4 + \frac{1}{K}\sum_{k=1}^K \mathcal{I}_{5,k}, \end{aligned} \tag{35}$$

where

$$\begin{aligned} \mathcal{I}_3 &:= \inf_{\theta\in\Theta} \sqrt{N} \left\| \frac{1}{K}\sum_{k=1}^K \mathbb{P}_N[\psi(Z; \theta, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))] \right\|, \\ \mathcal{I}_4 &:= \sqrt{N} \sup_{r\in(0,1), (\eta_1(\cdot;\theta_1'), \eta_2)\in\mathcal{T}_N} \|\partial_r^2 f(r; \hat{\theta}, \eta_1(\cdot; \theta_1'), \eta_2)\|, \\ \mathcal{I}_{5,k} &:= \sup_{\|\theta-\theta^*\|\leq\tau_N} \|\mathbb{G}_{N,k}\left[\psi(Z; \theta, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)) - \psi(Z; \theta^*, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z))\right]\|. \end{aligned}$$

Here Assumption 3 condition ii. guarantees that $\mathcal{I}_4 \leq \delta_N \left(1 + \sqrt{N}\|\hat{\theta} - \theta^*\|\right)$ and the assumption that $\varepsilon_N = \delta_N N^{-1/2}$ guarantees that $\varepsilon_N N^{1/2} \leq \delta_N$. In step III and IV, we will further bound $\mathcal{I}_{5,k} \lesssim \rho_N' := (N^{-1/2+1/q} + r_N')\log N + r_N' \log^{1/2}(1/r_N') + N^{-1/2+1/q} \log(1/r_N') \lesssim \delta_N$ and $\mathcal{I}_3 \leq \mathcal{I}_4 + \frac{1}{K}\sum_{k=1}^K \mathcal{I}_{5,k}$ respectively.

Consequently, with P_N-probability $1 - o(1)$,

$$\begin{aligned} &\|\sqrt{N}\mathbb{P}_N[\psi(Z; \theta^*, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z))] + \sqrt{N} J^*(\hat{\theta} - \theta^*)\| \\ \lesssim &\left(\delta_N \left(1 + \sqrt{N}\|\hat{\theta} - \theta^*\|\right)\right) + \rho_N' + \delta_N. \end{aligned} \tag{36}$$

This implies that

$$\begin{aligned}
&\sqrt{N}\|\hat\theta-\theta^*\|-\|\sqrt{N}J^{*-1}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]\|\\
\le&\|\sqrt{N}(\hat\theta-\theta^*)+\sqrt{N}J^{*-1}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]\|\\
\lesssim&\|J^{*-1}\|\left[\left(\delta_N\left(1+\sqrt{N}\|\hat\theta-\theta^*\|\right)\right)+\rho_N'+\delta_N\right]
\end{aligned}$$

and

$$\sqrt{N}\|\hat\theta-\theta^*\|\lesssim\frac{1}{c_3}\left[\left(\delta_N\left(2+\sqrt{N}\|\hat\theta-\theta^*\|\right)\right)+\rho_N\right]+\|\sqrt{N}J^{*-1}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]\|.$$

By Assumption 2 condition v. and Markov inequality, $\|\sqrt{N}J^{*-1}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]\|=O_{P_N}(\sqrt{c_6})$. Thus, with P_N-probability $1-o(1)$,

$$\sqrt{N}\|\hat\theta-\theta^*\|\lesssim\delta_N+\rho_N'.$$

Plugging this back into Eq. (36) gives

$$\|\sqrt{N}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]+\sqrt{N}J^*(\hat\theta-\theta^*)\|=O_{P_N}(\delta_N+\rho_N').$$

Since $\|J^{*-1}\|\le 1/c_3$ and $\|\Sigma^{-1/2}\|\le 1/\sqrt{c_5}$, we further have

$$\begin{aligned}
&\|\Sigma^{-1/2}J^{*-1}\sqrt{N}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]+\Sigma^{-1/2}\sqrt{N}(\hat\theta-\theta^*)\|\\
&\le\|\Sigma^{-1/2}\|\|J^{*-1}\|\|\sqrt{N}\mathbb{P}_N[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]+\sqrt{N}J^*(\hat\theta-\theta^*)\|\\
&\lesssim\delta_N+\rho_N'=\rho_N.
\end{aligned}$$

Now we prove the decomposition Eq. (35). Note that for any $\theta\in\Theta$ and $(\eta_1(\cdot,\theta_1),\eta_2)\in\mathcal{T}_N$

$$\begin{aligned}
&\sqrt{N}\left\{\frac{1}{K}\sum_{k=1}^K\mathbb{P}_{N,k}\left[\psi(Z;\theta,\eta_1(Z,\theta_1),\eta_2(Z))\right]\right\}\\
&=\frac{1}{K}\sum_{k=1}^K\mathbb{G}_{N,k}\left[\psi(Z;\theta,\eta_1(Z,\theta_1),\eta_2(Z)))-\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]+\sqrt{N}\mathbb{P}_N\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\\
&+\frac{1}{K}\sum_{k=1}^K\sqrt{N}\Big\{\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z,\theta_1),\eta_2(Z))\right]-\mathbb{P}\left[\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\Big\}\qquad(37)
\end{aligned}$$

If we apply Eq. (37) with $\theta = \hat{\theta}$ and $(\eta_1(\cdot, \theta_1), \eta_2)$ equal $(\hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)})$ for the kth fold, and apply Eq. (15), then

$$\begin{aligned}
&\Bigg\| \frac{1}{K}\sum_{k=1}^{K} \mathbb{G}_{N,k}\left[\psi(Z;\hat{\theta}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)) - \psi(Z;\theta^*, \eta_1^*(Z;\theta_1^*), \eta_2^*(Z))\right] \\
&+\sqrt{N}\left\{\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left[\psi(Z;\hat{\theta}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*, \eta_1^*(Z;\theta_1^*), \eta_2^*(Z))\right]\right\} \\
&+\sqrt{N}\mathbb{P}_N\left[\psi(Z;\theta^*, \eta_1^*(Z;\theta_1^*), \eta_2^*(Z))\right]\Bigg\| \\
=&\sqrt{N}\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_{N,k}\left[\psi(Z;\hat{\theta}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))\right]\right\| \\
\le&\sqrt{N}\inf_{\theta\in\Theta}\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_{N,k}\left[\psi(Z;\theta, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))\right]\right\| + \varepsilon_N\sqrt{N}. \qquad (38)
\end{aligned}$$

Here

$$\|\mathbb{G}_{N,k}\left[\psi(Z;\hat{\theta}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)) - \psi(Z;\theta^*, \eta_1^*(Z;\theta_1^*), \eta_2^*(Z))\right]\| \le \mathcal{I}_{5,k} \qquad (39)$$

and the second order tayler expansion at $r = 0$ gives that for some data-dependent $\tilde{r} \in (0, 1)$,

$$\begin{aligned}
&\sqrt{N}\left\{\mathbb{P}\left[\psi(Z;\hat{\theta}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z))\right] - \mathbb{P}\left[\psi(Z;\theta^*, \eta_1^*(Z;\theta_1^*), \eta_2^*(Z))\right]\right\} \\
=&\sqrt{N}\left[f(1;\hat{\theta}, \hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}) - f(0;\hat{\theta}, \hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)})\right] \\
=&\sqrt{N}\left\{J^*(\hat{\theta} - \theta^*) + \partial_r^2 f(r;\hat{\theta}, \hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)})|_{r=\tilde{r}}\right\} \qquad (40)
\end{aligned}$$

where the third equality uses the Neyman orthogonality in Assumption 2 condition vii..

Combining Eq. (38), Eq. (39) and Eq. (40) gives decomposition Eq. (35).

Step III: bounding $\mathcal{I}_{5,k}$. To bound $\mathcal{I}_{5,k}$, we still condition on $\hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}$, and then apply Lemma 6.2 in Chernozhukov et al. (2018a) with function class

$$\mathcal{F}'_{\hat{\eta}^{(k)}, \hat{\theta}_{1,\text{init}}^{(k)}} = \{\psi_j(\cdot;\theta, \hat{\eta}_1^{(k)}(\cdot, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}) - \psi_j(\cdot;\theta^*, \eta_1^*(\cdot, \theta^*), \eta_2^*) : j = 1, \ldots, d, \theta \in \Theta, \|\theta - \theta^*\| \le \tau_N\}.$$

We can verify that $\mathcal{F}'_{\hat{\eta}^{(k)}, \hat{\theta}_{1,\text{init}}^{(k)}}$ satisfies similar entropy condition with envelope $F_{1,\hat{\eta}^{(k)}, \hat{\theta}_{1,\text{init}}^{(k)}} + F_{1,\eta^*,\theta_1^*}$. Moreover, Assumption 3 implies that

$$\sup_{\|\theta-\theta^*\|\le\tau_N} \|\psi(Z;\theta, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)) - \psi(Z;\theta^*, \eta_1^*(Z, \theta^*), \eta_2^*(Z))\|_{\mathbb{P},2} \le r'_N.$$

Thus conditionally on $\hat{\theta}_{1,\text{init}}, \hat{\eta}_1^{(k)}(Z, \hat{\theta}_{1,\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z)$, we can use Lemma 6.2 eq. (A.1) in Chernozhukov et al. (2018a) to show that with P_N-probability $1 - o(1)$,

$$\mathcal{I}_{5,k} \lesssim (N^{-1/2+1/q} + r'_N)\log N + r'_N \log^{1/2}(1/r'_N) + N^{-1/2+1/q}\log(1/r'_N),$$

which also holds unconditionally according to Lemma 6.1 in Chernozhukov et al. (2018a) .

Step IV: bounding $\mathcal{I}_3$. Let $\overline{\theta} = \theta^* - J^{*-1}\mathbb{P}_N\left[\psi(Z;\theta^*,\eta_1^*(Z,\theta^*),\eta_2^*(Z))\right]$.
Since $\mathbb{P}\left[\psi(Z;\theta^*,\eta_1^*(Z,\theta^*),\eta_2^*(Z))\right]=0$, J^* is nonsingular with singular values bounded away from 0 by c_3, and $\|\mathbb{P}_N\left[\psi(Z;\theta^*,\eta_1^*(Z,\theta^*),\eta_2^*(Z))\right]\| = O_{P_N}(N^{-1/2})$, $\|\overline{\theta}-\theta^*\| = O_{P_N}(N^{-1/2}) = o_{P_N}(\tau_N)$. According to Assumption 2 condition i., $\overline{\theta}\in\Theta$ with P_N probability $1-o(1)$. Therefore,

$$\mathcal{I}_3 \le \sqrt{N}\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_N[\psi(Z;\overline{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))]\right\|$$

Then apply the linearization Eq. (37) and taylor expansion similar to Eq. (40) with $\theta=\overline{\theta}$ and $(\eta_1(\cdot,\theta_1),\eta_2)$ equal $(\hat{\eta}_1^{(k)}(\cdot,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)})$ for the kth fold, we can get that

$$\begin{aligned}&\sqrt{N}\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_{N,k}[\psi(Z;\overline{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{1,\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))]\right\|\\ \le&\sqrt{N}\|\mathbb{P}_N[\psi(Z;\overline{\theta},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]+J^*(\overline{\theta}-\theta^*)\|+\mathcal{I}_4+\frac{1}{K}\sum_{k=1}^{K}\mathcal{I}_{5,k}=\mathcal{I}_4+\frac{1}{K}\sum_{k=1}^{K}\mathcal{I}_{5,k}.\end{aligned}$$

where the last equality here holds because $\mathbb{P}_N[\psi(Z;\overline{\theta},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))]+J^*(\overline{\theta}-\theta^*)=0$ as a consequence of the special construction of $\overline{\theta}$.

Extension: $\hat{\theta}$ defined in Definition 5. By applying step I to IV to sample estimating equation Eq. (28), we can get that for $k=1,\ldots,K$,

$$\sqrt{N/K}\Sigma^{-1/2}(\hat{\theta}^{(k)}-\theta^*) = \frac{1}{\sqrt{N/K}}\sum_{i\in\mathcal{D}_k}\Sigma^{-1/2}J^{*-1}\psi(Z_i;\theta^*,\eta_1^*(Z_i;\theta_1^*),\eta_2^*(Z_i))+O_P(\rho_{N/K}).$$

Since K is a fixed integer that does not grow with N, the equation above implies that the asserted conclusion in Theorem 1 also holds for $\hat{\theta}=\frac{1}{K}\sum_{k=1}^{K}\hat{\theta}^{(k)}$. □

G.3 Proofs for Section 4

Proof of Theorem 2. We still consider data generating processes $\{P_N\}_{N\ge 1}$ defined in the proof for Theorem 1, and define $\otimes a = aa^\top$. Now we prove that

$$\|\mathbb{P}_{N,k}[\otimes\psi(Z;\hat{\theta},\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{\text{init}}^{(k)}),\hat{\eta}_2^{(k)}(Z))]-\mathbb{P}[\otimes\psi(Z;\theta^*,\eta_1^*(Z,\theta_1^*),\eta_2^*(Z))]\| = O_{P_N}(\rho_N''). \quad (41)$$

for any $k\in[1,\cdots,K]$. Then, the statement in Theorem 2 is immediately concluded. For all $j,l\in[1,\cdots,d]$ $(d=d_1+d_2)$, Eq. (41) follows once we have $\mathcal{I}_{jl}=O_{P_N}(\rho_N'')$, where

$$\mathcal{I}_{jl} := |\mathbb{P}_{N,k}[\psi_j(Z;\hat{\theta},\hat{\eta}_1^{(k)},\hat{\eta}_2^{(k)})\psi_l(Z;\hat{\theta},\hat{\eta}_1^{(k)},\hat{\eta}_2^{(k)})]-\mathbb{P}[\psi_j(Z;\theta^*,\eta_1^*,\eta_2^*)\psi_l(Z;\theta^*,\eta_1^*,\eta_2^*)]|.$$

Here, to simplify the notation, we use $\hat{\eta}_1^{(k)}=\hat{\eta}_1^{(k)}(Z,\hat{\theta}_{\text{init}}^{(k)}),\eta_1^*=\eta_1^*(Z,\theta_1^*),\hat{\eta}_2^{(k)}=\hat{\eta}_2^{(k)}(Z,\hat{\theta}_{\text{init}}^{(k)}),\eta_2^*=\eta_2^*(Z,\theta_2^*)$. Obviously we have $\mathcal{I}_{jl}\le\mathcal{I}_{jl,1}+\mathcal{I}_{jl,2}$, where

$$\begin{aligned}\mathcal{I}_{jl,1}&=|\mathbb{P}_{N,k}[\psi_j(Z;\hat{\theta},\hat{\eta}_1^{(k)},\hat{\eta}_2^{(k)})\psi_l(Z;\hat{\theta},\hat{\eta}_1^{(k)},\hat{\eta}_2^{(k)})]-\mathbb{P}_{N,k}[\psi_j(Z;\theta^*,\eta_1^*,\eta_2^*)\psi_l(Z;\theta^*,\eta_1^*,\eta_2^*)]|,\\ \mathcal{I}_{jl,2}&=|\mathbb{P}_{N,k}[\psi_j(Z;\theta^*,\eta_1^*,\eta_2^*)\psi_l(Z;\theta^*,\eta_1^*,\eta_2^*)]-\mathbb{P}[\psi_j(Z;\theta^*,\eta_1^*,\eta_2^*)\psi_l(Z;\theta^*,\eta_1^*,\eta_2^*)]|,\end{aligned}$$

and we show that each term here is $O_p(\rho''_N)$.

We first bound $\mathcal{I}_{jl,2}$. This is upper bounded as

$$\begin{aligned}\mathbb{P}[\mathcal{I}^2_{jl,2}] &\leq N^{-1}\mathbb{P}[\psi^2_j(Z;\theta^*,\eta^*_1,\eta^*_2)\psi^2_l(Z;\theta^*,\eta^*_1,\eta^*_2)] \\ &\leq N^{-1}\{\mathbb{P}[\psi^4_j(Z;\theta^*,\eta^*_1,\eta^*_2)]\mathbb{P}[\psi^4_l(Z;\theta^*,\eta^*_1,\eta^*_2)]\}^{1/2} \\ &\leq N^{-1}\mathbb{P}[\|\psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^4] \leq N^{-1}C^4.\end{aligned}$$

Here, we use the fourth moment assumption in Assumption 4. From conditional Markov inequality, we have $\mathcal{I}_{jl,2} = O_{P_N}(1/N^{-1/2})$.

Next, we bound $\mathcal{I}_{jl,1}$. Following the proof of Theorem 3.2 (Chernozhukov et al., 2018a), we have

$$\begin{aligned}\mathcal{I}^2_{jl,1} &\leq R_N \times \left\{\mathbb{P}_{N,k}[\|\psi(Z;\theta^*,\eta^*_1,\eta^*_2\|^2] + R_N\right\}, \\ R_N &= \mathbb{P}_{N,k}[\|\psi(Z;\hat{\theta},\hat{\eta}_1,\hat{\eta}_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2].\end{aligned}$$

In addition, from the fourth moment assumption in Assumption 4

$$\mathbb{P}[\mathbb{P}_{N,k}[\|\psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2]] = \mathbb{P}[\|\psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] \leq C^2.$$

It follows from Markov inequality that

$$\mathbb{P}_{N,k}[\|\psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] = O_{P_N}(1).$$

It remains to bound R_N. We have

$$\begin{aligned}R_N &= \mathbb{P}_{N,k}[\|\psi(Z;\hat{\theta},\hat{\eta}_1,\hat{\eta}_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] \\ &\leq \mathbb{P}_{N,k}[\|\psi(Z;\hat{\theta},\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] + \mathbb{P}_{N,k}[\|\psi(Z;\hat{\theta},\hat{\eta}_1,\hat{\eta}_2) - \psi(Z;\hat{\theta},\eta^*_1,\eta^*_2)\|^2].\end{aligned} \tag{42}$$

Then, the first term of Eq. (42) is upper bounded with P_N-probability $1-o(1)$ as

$$\begin{aligned}&\mathbb{P}_{N,k}[\|\psi(Z;\hat{\theta},\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] \\ &= \frac{1}{\sqrt{N}}\mathbb{G}_{N,k}[\|\psi(Z;\hat{\theta},\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] + \mathbb{P}[\|\psi(Z;\hat{\theta},\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] \\ &\leq \sup_{\theta\in\Theta}\frac{1}{\sqrt{N}}\mathbb{G}_{N,k}[\|\psi(Z;\theta,\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] + \mathbb{P}[\|\psi(Z;\hat{\theta},\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2] \\ &\lesssim N^{-1/2}\log N\{1+N^{-1/2+2/q}\} + \|\hat{\theta}-\theta^*\|^{\beta}_2 = N^{-1/2}\log N\{1+N^{-1/2+2/q}\} + N^{-\beta/2}.\end{aligned}$$

In the last inequality, we use Lemma 6.2 (Chernozhukov et al., 2018a). Here, the envelops exists since $\|\psi(Z;\theta,\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2 \leq CF^2_{\eta^*,\theta^*}$ for some constant C depending on d_1, d_2. According to condition vi. in Assumption 2, it satisfies the moment condition $\|F^2_{\eta^*,\theta^*}\|_{\mathbb{P},q/2} \leq c_1$. In addition, the metricy entropy assumption is satisfied since

$$\begin{aligned}&\sup_{\mathbb{Q}}\log N(\epsilon\|C\mathcal{F}^2_{1,\eta^*,\theta^*}\|_{\mathbb{Q},2}, \left\{\|\psi(Z;\theta,\eta^*_1,\eta^*_2) - \psi(Z;\theta^*,\eta^*_1,\eta^*_2)\|^2 : \theta\in\Theta\right\}, \|\cdot\|_{\mathbb{Q},2}) \\ &\lesssim \sup_{\mathbb{Q}}\log\{N(\epsilon\|\mathcal{F}_{1,\eta^*,\theta^*}\|_{\mathbb{Q},2}, \mathcal{F}_{1,\eta,\theta'_1}, \|\cdot\|_{\mathbb{Q},2})\}^2 \lesssim v\log(a/\epsilon).\end{aligned}$$

Similarly, with P_N-probability probability $1-o(1)$, the second term of Eq. (42) can be upper bounded as follows:

$$\begin{aligned}
&\mathbb{P}_{N,k}[\|\psi(Z;\hat\theta,\hat\eta_1,\hat\eta_2)-\psi(Z;\hat\theta,\eta_1^*,\eta_2^*)\|^2] \\
&=\frac{1}{\sqrt{N}}\mathbb{G}_{N,k}[\|\psi(Z;\hat\theta,\hat\eta_1,\hat\eta_2)-\psi(Z;\hat\theta,\eta_1^*,\eta_2^*)\|^2]+\mathbb{P}[\|\psi(Z;\hat\theta,\hat\eta_1,\hat\eta_2)-\psi(Z;\hat\theta,\eta_1^*,\eta_2^*)\|^2] \\
&\le \sup_{\theta\in\Theta}\frac{1}{\sqrt{N}}\mathbb{G}_{N,k}[\|\psi(Z;\theta,\hat\eta_1,\hat\eta_2)-\psi(Z;\theta,\eta_1^*,\eta_2^*)\|^2]+\sup_{\theta\in\mathcal{B}(\theta^*;\tau_N)}\mathbb{P}[\|\psi(Z;\theta,\hat\eta_1,\hat\eta_2)-\psi(Z;\theta,\eta_1^*,\eta_2^*)\|^2] \\
&\lesssim N^{-1/2}\log N\{1+N^{-1/2+2/q}\}+\{r_N'\}^2.
\end{aligned}$$

In the last inequality, we use Lemma 6.2 (Chernozhukov et al., 2018a) and Assumption 3. In the end, we have

$$R_N=O_{P_N}\left(N^{-1/2+1/q}(\log N)^{1/2}+N^{-1/4}(\log N)^{1/2}+r_N'\right)+N^{-\beta/4}.$$

This concludes the proof. □

G.4 Proofs for Section 5

Proof for Theorem 3. In this part, we prove the asymptotic distribution of our estimators corresponding to the general estimating equation Eq. (7). We prove this by verifying all conditions in the assumptions for Theorem 1.

Verifying Assumption 1.

$$\begin{aligned}
J^* &= \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right]\}|_{\theta=\theta^*} \\
&=\partial_\theta\mathbb{P}\left\{\frac{\mathbb{I}(T=t)}{\pi^*(t\mid X)}U(Y;\theta_1)-\frac{\mathbb{I}(T=t)-\pi^*(t\mid X)}{\pi^*(t\mid X)}\mu^*(X,t;\theta_1)+V(\theta_2)\right\}|_{\theta=\theta^*} \\
&=\partial_\theta\mathbb{P}\left\{\frac{\mathbb{I}(T=t)}{\pi^*(t\mid X)}U(Y;\theta_1)+V(\theta_2)\right\}|_{\theta=\theta^*} \\
&=\partial_\theta\mathbb{P}\left\{\frac{\mathbb{I}(T=t)}{\pi^*(t\mid X)}U(Y;\theta_1)-\frac{\mathbb{I}(T=t)-\pi^*(t\mid X)}{\pi^*(t\mid X)}\mu^*(X,t;\theta_1^*)+V(\theta_2)\right\}|_{\theta=\theta^*} \\
&=\partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}.
\end{aligned}$$

Verifying Assumption 2. We first verify conditions iii. and iv. in Assumption 2. We denote that $J_{jk}(\theta)=\partial_{\theta^{(k)}}\mathbb{P}\left[U_j(Y(t);\theta_1)+V_j(\theta_2)\right]$ where $\theta^{(k)}$ is the k^{th} component of $\theta=(\theta_1,\theta_2)$. By condition ii., $J_{jk}(\theta)$ is Lipschitz continuous at θ^* with Lipschitz constant c'. So for any $\varepsilon>0$, if θ belongs to the open ball $\mathcal{B}(\theta^*;\epsilon/c')$, then

$$|J_{jk}(\theta)-J_{jk}(\theta^*)|=|\partial_{\theta^{(k)}}\mathbb{P}\left[U_j(Y(t);\theta_1)+V_j(\theta_2)\right]-\partial_{\theta^{(k)}}\mathbb{P}\left[U_j(Y(t);\theta_1^*)+V_j(\theta_2^*)\right]|\le\varepsilon.$$

By Taylor expansion, for any $\theta\in\mathcal{B}(\theta^*;\delta)$, there exists $\overline{\theta}\in\mathcal{B}(\theta^*;\|\theta-\theta^*\|)$ such that

$$\begin{aligned}
\|\mathbb{P}\left[U(Y(t);\theta_1)+V(\theta_2)\right]\| &= \|J(\overline{\theta})(\theta-\theta^*)\|\ge\|J(\theta^*)(\theta-\theta^*)\|-\|(J(\overline{\theta})-J(\theta^*))(\theta-\theta^*)\| \\
&\ge\|J(\theta^*)(\theta-\theta^*)\|-\varepsilon\sqrt{d}\|\theta-\theta^*\|\ge\|J(\theta^*)(\theta-\theta^*)\|-\frac{1}{2}\|J(\theta^*)(\theta-\theta^*)\| \\
&=\frac{1}{2}\|J(\theta^*)(\theta-\theta^*)\|,
\end{aligned}$$

where the second last inequality holds if we choose $\varepsilon \leq \frac{c_3}{2\sqrt{d}} \leq \frac{1}{2\sqrt{d}}\sigma_{\min}(J(\theta^*))$, where $\sigma_{\min}(J(\theta^*))$ is the smallest singular value of $J(\theta^*)$. Thus

$$\inf_{\theta \in \mathcal{B}(\theta^*;\varepsilon/c')} 2\|\mathbb{P}\left[U(Y(t);\theta_1) + V(\theta_2)\right]\| \geq \|J(\theta^*)(\theta - \theta^*)\|.$$

Moreover, for any $\theta \in \Theta \setminus \mathcal{B}(\theta^*; \frac{c_3}{2\sqrt{d}c'})$, $2\|\mathbb{P}\left[U(Y(t);\theta_1) + V(\theta_2)\right]\| \geq c_2$ according to condition ii.. Therefore,

$$2\|\mathbb{P}\left[U(Y(t);\theta_1) + V(\theta_2)\right]\| \geq J^*(\theta - \theta^*) \wedge c_2, \quad \text{where } J^* = J(\theta^*).$$

Moreover, the singular values J^* are bounded between c_3, c_4 according to condition iii..

We then verify condition vii. in Assumption 2: for any $(\eta_1(\cdot;\theta_1'), \eta_2) \in \mathcal{T}_N$,

$$\begin{aligned}&\partial_r\big\{\mathbb{P}\left[\psi(Z;\theta^*, \eta_1(Z;\theta_1^*) + r(\eta_1(\cdot;\theta_1') - \eta_1^*(Z;\theta_1^*)), \eta_2^*(Z))\right]\big\}|_{r=0}\\ =&\partial_r\mathbb{P}\left\{\frac{\mathbb{I}(T=t)}{\pi^*(t \mid X) + r(\pi(t \mid X) - \pi^*(t \mid X))}\big(U(Y;\theta_1^*) - \mathbb{E}[U(Y;\theta_1^*) \mid X, T]\big)\right\}|_{r=0} = 0.\end{aligned}$$

Verifying Assumption 3. We take $\mathcal{T}_N$ to be the set that contains all $(\mu(\cdot,\theta_1'), \pi(\cdot))$ that satisfies $\|\theta_1' - \theta_1^*\| \leq \rho_{\theta,N}$ and

$$\left\|\left\{\mathbb{P}\left[\mu\left(X,T;\theta_1'\right) - \mu^*\left(X,T;\theta_1'\right)\right]^2\right\}^{1/2}\right\| \leq \rho_{\mu,N}, \left\{\mathbb{P}\left[\pi(T \mid X) - \pi^*(T \mid X)\right]^2\right\}^{1/2} \leq \rho_{\pi,N},$$

with $\rho_{\pi,N}(\rho_{\mu,N} + C\rho_{\theta,n}) \leq \frac{\varepsilon_\pi^3}{3}\delta_N N^{-1/2}$, $\rho_{\pi,N} \leq \frac{\delta_N^3}{\log N}$, and $\rho_{\mu,N} + C\rho_{\theta,N} \leq \frac{\delta_N^2}{\log N}$.

Then Assumption 6 and condition vii. in Theorem 3 guarantee that the nuisance estimates $(\hat{\mu}(,\hat{\theta}_{1,\text{init}}), \hat{\pi}) \in \mathcal{T}_N$ with probability, namely, condition i. in Assumption 3 is satisfied.

Before verifying other conditions, first note that the condition vi. states that

$$\left\{\mathbb{P}\left[\mu^*(X,T;\theta_1) - \mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2} \leq C\|\theta_1 - \theta_1^*\|, \quad \forall \|\theta_1 - \theta_1^*\| \leq \rho_{\theta,N},$$

which implies that for any $(\mu(\cdot,\theta_1'), \pi(\cdot)) \in \mathcal{T}_N$,

$$\begin{aligned}&\left\|\left\{\mathbb{P}\left[\mu(X,T;\theta_1') - \mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\|\\ \leq &\left\|\left\{\mathbb{P}\left[\mu(X,T;\theta_1') - \mu^*(X,T;\theta_1')\right]^2\right\}^{1/2}\right\| + \left\|\left\{\mathbb{P}\left[\mu^*(X,T;\theta_1') - \mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\| = \rho_{\mu,N} + C\rho_{\theta,N}.\end{aligned}$$

Now we verify the condition on r_N: for any $(\eta_1(\cdot;\theta_1'), \eta_2(\cdot)) = (\mu(\cdot,\theta_1'), \pi(\cdot)) \in \mathcal{T}_N$, $\theta \in \Theta$,

$$\begin{aligned}&\|\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z))\right] - \mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\|\\ \leq&\|\mathbb{P}\big(\frac{\mathbb{I}(T=t)}{\pi(t \mid X)} - \frac{\mathbb{I}(T=t)}{\pi^*(t \mid X)}\big)\big(\mu^*(X,T;\theta_1^*) - \mu(X,T;\theta_1')\big)\| + \|\mathbb{P}\frac{\mathbb{I}(T=t) - \pi^*(t \mid X)}{\pi^*(t \mid X)}[\mu^*(X,T;\theta_1^*) - \mu(X,T;\theta_1')]\|\end{aligned}$$

The above can be further upper bounded by

$$\frac{1}{\varepsilon_\pi}\left\{\mathbb{P}\left[\pi(t \mid X)-\pi^*(t \mid X)\right]^2\right\}^{1/2}\left\|\left\{\mathbb{P}\left[\mu^*(X,T;\theta_1)-\mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\|$$

$$+\frac{1}{\varepsilon_\pi}\left\{\mathbb{P}\left[\pi(t \mid X)-\pi^*(t \mid X)\right]^2\right\}^{1/2}\left\|\left\{\mathbb{P}\left[\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\| \leq \frac{4C}{\varepsilon_\pi}\sqrt{d}\rho_{\pi,N}.$$

Thus, the condition on r_N is satisfied with τ_N such that $\tau_N=\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}$.

Next, we verify the condition on r_N': for any θ such that $\|\theta-\theta^*\|\leq\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}$, and any $(\eta_1(\cdot;\theta_1'),\eta_2(\cdot))=(\mu(\cdot,\theta_1'),\pi(\cdot))\in\mathcal{T}_N$,

$$\begin{aligned}&\left\|\left\{\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z))-\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]^2\right\}^{1/2}\right\|\\ \leq&\|\mathbf{A}\|+\|\mathbf{B}\|+\|\mathbf{C}\|,\end{aligned}$$

where $\mathbf{A},\mathbf{B},\mathbf{C}$ are three d-dimensional vector whose i-th elements are given as follows:

$$\begin{aligned}\mathbf{A}_i&=\left\{\mathbb{P}\left[\Big(\frac{\mathbb{I}(T=t)}{\pi(t\mid X)}-\frac{\mathbb{I}(T=t)}{\pi^*(t\mid X)}\Big)\big(\mu_i^*(X,T;\theta_1)-\mu_i^*(X,T;\theta_1^*)\big)\right]^2\right\}^{1/2}\\ \mathbf{B}_i&=\left\{\mathbb{P}\left[\Big(\frac{\mathbb{I}(T=t)}{\pi(t\mid X)}-\frac{\mathbb{I}(T=t)}{\pi^*(t\mid X)}\Big)\big(\mu_i^*(X,T;\theta_1^*)-\mu_i(X,T;\theta_1')\big)\right]^2\right\}^{1/2}\\ \mathbf{C}_i&=\left\{\mathbb{P}\left[\frac{\mathbb{I}(T=t)-\pi^*(t\mid X)}{\pi^*(t\mid X)}\big(\mu_i^*(X,T;\theta_1^*)-\mu_i(X,T;\theta_1')\big)\right]^2\right\}^{1/2}.\end{aligned}$$

It follows that

$$\begin{aligned}&\left\|\left\{\mathbb{P}\left[\psi(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z))-\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]^2\right\}^{1/2}\right\|\\ \leq&\frac{4C^2\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi^2}+\frac{1}{\varepsilon_\pi}\left(\rho_{\mu,N}+C\rho_{\theta,N}\right)+\frac{1}{\varepsilon_\pi}\left(\rho_{\mu,N}+C\rho_{\theta,N}\right)\end{aligned}$$

So when $\rho_{\pi,N}\leq\frac{\delta_N^3}{\log N}$, and $\rho_{\mu,N}+C\rho_{\theta,N}\leq\frac{\delta_N^2}{\log N}$, $r_N'=\frac{\delta_N^2}{\varepsilon_\pi^2\log N}\left(4C^2\sqrt{d}+2\varepsilon_\pi\right)\leq\frac{\delta_N}{\log N}$ if $\delta_N\leq\frac{\varepsilon_\pi^2}{4C^2\sqrt{d}+2\varepsilon_\pi}$.

Finally, to verify the condition on λ_N', we note that for any θ such that $\|\theta-\theta^*\|\leq\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}$, and any $(\eta_1(\cdot;\theta_1'),\eta_2(\cdot))=(\mu(\cdot,\theta_1'),\pi(\cdot))\in\mathcal{T}_N$

$$\begin{aligned}&f(r;\theta,\eta_1(Z;\theta_1'),\eta_2)\\ =\mathbb{P}\Big\{&\frac{\mathbb{I}(T=t)}{\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))}\left[\mu^*(X,T;\theta_1^*+r(\theta_1-\theta_1^*))-\mu^*(X,T;\theta_1^*)\right.\\ &\left.-r\big(\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\big)\right]+\left[\mu^*(X,t;\theta_1^*)+r\big(\mu(X,t;\theta_1')-\mu^*(X,t;\theta_1^*)\big)\right]+V(\theta_2^*+r(\theta_2-\theta_2^*))\Big\}\end{aligned}$$

Thus the first-order derivative is

$$\begin{aligned}
&\partial_r f(r;\theta,\eta_1(Z;\theta_1'),\eta_2)\\
=&-\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\big(\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))\big)^2}\big(\pi(T\mid X)-\pi^*(T\mid X)\big)\big[\mu^*(X,T;\theta_1^*+r(\theta_1-\theta_1^*))\\
&-\mu^*(X,T;\theta_1^*)-r\big(\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\big)\big]\Big\}+\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))}\\
&\times\partial_{\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}(\theta_1-\theta_1^*)\Big\}-\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))}\\
&\times\big[\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\big]\Big\}+\mathbb{P}\Big\{\big[\mu(X,t;\theta_1')-\mu^*(X,t;\theta_1^*)\big]\Big\}+\partial_{\overline{\theta}_2^\top}V(\overline{\theta}_2)|_{\overline{\theta}_2=\theta_2^*+r(\theta_2-\theta_2^*)}(\theta_2-\theta_2^*).
\end{aligned}$$

The second order derivative is

$$\begin{aligned}
&\partial_r^2 f(r;\theta,\eta_1(Z;\theta_1'),\eta_2)\\
=&\mathbb{P}\Big\{\frac{2\mathbb{I}(T=t)}{\big(\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))\big)^3}\big(\pi(T\mid X)-\pi^*(T\mid X)\big)^2\big[\mu^*(X,T;\theta_1^*+r(\theta_1-\theta_1^*))\\
&-\mu^*(X,T;\theta_1^*)-r\big(\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\big)\big]\Big\}-\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\big(\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))\big)^2}\\
&\times\big(\pi(T\mid X)-\pi^*(T\mid X)\big)\partial_{\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}(\theta_1-\theta_1^*)\Big\}\\
&+\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\big(\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))\big)^2}\big(\pi(T\mid X)-\pi^*(T\mid X)\big)\big[\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\big]\Big\}\\
&+\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))}\operatorname{diag}\big[(\theta_1-\theta_1^*)^\top\big]\big[\partial^2_{\overline{\theta}_1,\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}\big](\theta_1-\theta_1^*)\Big\}\\
&-\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)\big(\pi(T\mid X)-\pi^*(T\mid X)\big)}{\big(\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))\big)^2}\partial_{\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}(\theta_1-\theta_1^*)\Big\}\\
&+\mathbb{P}\Big\{\frac{\mathbb{I}(T=t)}{\big(\pi^*(T\mid X)+r(\pi(T\mid X)-\pi^*(T\mid X))\big)^2}\big(\pi(T\mid X)-\pi^*(T\mid X)\big)\big[\mu(X,T;\theta_1')-\mu^*(X,T;\theta_1^*)\big]\\
&\quad+\operatorname{diag}(\theta_2-\theta_2^*)^\top\partial^2_{\overline{\theta}_2,\overline{\theta}_2^\top}V(\overline{\theta}_2)|_{\overline{\theta}_2=\theta_2+r(\theta_2-\theta_2^*)}(\theta_2-\theta_2^*)
\end{aligned}$$

Above, we use condition iv. in Theorem 3 to ensure exchange of integration and differentiation so we can get terms $\partial_{\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}$ and $\partial^2_{\overline{\theta}_1,\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}$.

We can verify that

$$\begin{aligned}
&\Big\|\mathbb{P}\Big[\big(\pi(T\mid X)-\pi^*(T\mid X)\big)\partial_{\overline{\theta}_1^\top}\mu^*(X,T;\overline{\theta}_1)|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}(\theta_1-\theta_1^*)\Big]\Big\|\\
&\le\Big\{\mathbb{P}\big[\big(\pi(T\mid X)-\pi^*(T\mid X)\big)\big]^2\Big\}^{1/2}\times\sqrt{d}\sup_{j,\|\theta_1-\theta_1^*\|\le\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}}\left\|\mathbb{P}\left\{\Big[\partial_{\overline{\theta}_1}\mu_j^*(X,t;\overline{\theta}_1)\Big]^2\right\}^{1/2}\right\|\times\|\theta_1-\theta_1^*\|\\
&\le C\sqrt{d}\rho_{\pi,N}\|\theta_1-\theta_1^*\|,
\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathbb{P}\left[\operatorname{diag}\left[(\theta_1-\theta_1^*)^\top\right]\left[\partial^2_{\overline{\theta},\overline{\theta}^\top}\mu^*(X,T;\overline{\theta})\right]|_{\overline{\theta}_1=\theta_1^*+r(\theta_1-\theta_1^*)}(\theta_1-\theta_1^*)\right]\right\|\\ \leq&\sqrt{d}\|\theta_1-\theta_1^*\|^2\sup_{j,\|\theta-\theta^*\|\leq\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}}\|\mathbb{P}\left[\partial_{\overline{\theta}_1}\partial_{\overline{\theta}_1^\top}\mu_j^*(X,T;\overline{\theta}_1)\right]\|\leq C\sqrt{d}\|\theta_1-\theta_1^*\|^2,\end{aligned}$$

and

$$\sup_{r\in(0,1)}\left\|\left\{\mathbb{P}\left[\mu^*(X,T;\theta_1^*+r(\theta_1-\theta_1^*))-\mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\|\leq C\sqrt{d}\|\theta_1-\theta_1^*\|.$$

Thus for any θ such that $\|\theta-\theta^*\|\leq\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}$,

$$\begin{aligned}&\|\partial_r^2 f(r;\theta,\mu(X,T;\theta_1'),\pi)\|\\ \leq&\frac{C\sqrt{d}}{\varepsilon_\pi^2}\rho_{\pi,N}\|\theta_1-\theta_1^*\|+\frac{1}{\varepsilon_\pi^2}\rho_{\pi,N}(\rho_{\mu,N}+C\rho_{\theta,N})+C\|\theta_2-\theta_2^*\|^2.\end{aligned}\tag{43}$$

Given $\rho_{\pi,N}\leq\frac{\delta_N^3}{\log N}$, when $\frac{\delta_N}{\log N}\leq\frac{\varepsilon_\pi^2}{8C^2d}$ and $\frac{\delta_N^2}{\log N}\leq\frac{\varepsilon_\pi^3}{2C\sqrt{d}}$, $\frac{4C^2d}{\varepsilon_\pi^2\delta_N}\rho_{\pi,N}\|\theta-\theta^*\|+\frac{C\sqrt{d}}{\varepsilon_\pi^3}\rho_{\pi,N}\|\theta_1-\theta_1^*\|\leq\delta_N\|\theta-\theta^*\|$. Moreover, when $\rho_{\pi,N}(\rho_{\mu,N}+C\rho_{\theta,n})\leq\frac{\varepsilon_\pi^3}{3}\delta_N N^{-1/2}$, $\frac{3}{\varepsilon_\pi^3}\rho_{\pi,N}(\rho_{\mu,N}+C\rho_{\theta,n})\leq\delta_N N^{-1/2}$. Consequently, $\|\partial_r^2 f(r;\theta,\mu(X,T;\theta_1'),\pi)\|\leq\delta_N(\|\theta-\theta^*\|+N^{-1/2})$.

Semiparametric efficiency. For this estimating equation with incomplete data, Theorems 10.1 and 10.2 in Tsiatis (2006) show that the corresponding semiparametric efficient influence is

$$\psi^{\text{eff}}(Z)=J^{*-1}\psi(Z;\theta^*,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z)),$$

where ψ is given in Equation (7). So the asymptotic variance Σ coincides with the semiparametric efficiency bound. □

Proof for Proposition 2. First note that

$$\mathbb{P}\left[U(Y(t);\theta_1)+V(\theta_2)\right]=\begin{bmatrix}F_t(\theta_1)-\gamma\\ \theta_1+\frac{1}{1-\gamma}\mathbb{E}[Y(t)-\theta_1]^+-\theta_2\end{bmatrix}.$$

When $F_t(\theta_1)$ is differentiable, $\mathbb{P}\left[U(Y(t);\theta_1)+V(\theta_2)\right]$ is also differentiable by Leibnitz integral rule, with derivative

$$J(\theta)=\begin{bmatrix}f_t(\theta_1)&0\\ \frac{F_t(\theta_1)-\gamma}{1-\gamma}&-1\end{bmatrix},\text{ and }J^*=J(\theta^*)\begin{bmatrix}f_t(\theta_1^*)&0\\0&-1\end{bmatrix}.$$

Now we prove Proposition 2 by verifying the assumptions in Theorem 3.

Verifying condition i in Theorem 3. We only need to verify that condition vi. of Assumption 2 hold. Since Θ is compact, $\{y\mapsto\mathbb{I}\left[y\leq\theta_1\right],\theta\in\Theta\}$, $\{y\mapsto\max\{\theta_1,\frac{1}{1-\gamma}(y-\theta_1)\}-\theta_2,\theta\in\Theta\}$ are obviously Donsker classes, so condition vi. of Assumption 2 is satisfied.

Verifying conditions ii. and iii. in Theorem 3. It is straightforward to show that $J(\theta)$ is invertible with the following matrix as its inverse:

$$J^{-1}(\theta) = \begin{bmatrix} \frac{1}{f_t(\theta_1)} & 0 \\ -\frac{F_t(\theta_1)-\gamma}{f_t(\theta_1)(1-\gamma)} & -1 \end{bmatrix}.$$

Note that $\sigma_{\max}(J(\theta^*)) \leq 2\max\{f_t(\theta_1^*), 1\} \leq 2\max\{c_2', 1\}$ and

$$\sigma_{\min}(J(\theta^*)) = 1/\sigma_{\max}(J^{-1}(\theta^*)) \geq \min\{\frac{f_t(\theta_1^*)}{2}, \frac{(1-\gamma)f_t(\theta_1^*)}{2\gamma}, \frac{1}{2}\} \geq \frac{1}{2}\min\{1, \frac{1-\gamma}{\gamma}c_1', c_1'\}.$$

Thus condition iii. in Theorem 3 is satisfied with $c_3 = \frac{1}{2}\min\{1, \frac{1-\gamma}{\gamma}c_1', c_1'\}$ and $c_4 = 2\max\{1, c_2'\}$. When we estimate quantile only, then only $f_t(\theta_1)$ in $J(\theta)$ matters. Then condition iii. in Theorem 3 is satisfied with $c_3 = c_1'$ and $c_4 = 2c_2'$.

Since $f_t(\theta_1) \leq c_2'$ and $\dot{f}_t(\theta_1) \leq c_3'$, it follows that each element in $J(\theta)$ is Lipschtiz continuous at θ^* with Lipschitz constant $c' = \max\{c_2', c_3'\}$. Moreover, for $\theta \in \Theta$ such that $\|\theta - \theta^*\| \geq \frac{c_3}{2\sqrt{2c'}}$, we have $2\|\mathbb{P}\left[U(Y(t); \overline{\theta}_1) + V(\overline{\theta}_2)\right]\| \geq c_5'$. This means that condition ii. in Theorem 3 is satisfied with $c' = \max\{c_2', c_3'\}$ and $c_2 = c_5'$. When we estimate quantile only, we only require $|F(\theta_1^*) - F(\theta_1)| \geq c_4'$ for $|\theta_1 - \theta_1^*| \geq \frac{c_1'}{2c_3'}$. Then condition ii. in Theorem 3 is satisfied with $c' = c_3'$ and $c_2 = 2c_4'$.

Verifying condition iv. in Theorem 3. This condition can be verified by the following facts: for any θ_1 such that $|\theta_1 - \theta_1^*| \leq \frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N \varepsilon_\pi}$,

$$\begin{aligned} &|\partial_r \mu_1^*(X, t; \theta_1^* + r(\theta_1 - \theta_1^*))| = |\partial_r\{F_t(\theta_1^* + r(\theta_1 - \theta_1^*) \mid X) - \gamma\}| \\ =& |f_t(\theta_1^* + r(\theta_1 - \theta_1^*) \mid X)|\,|\theta_1 - \theta_1^*| \leq C\,|\theta_1 - \theta_1^*| \\ &\left|\partial_r^2 \mu_1^*(X, t; \theta_1^* + r(\theta_1 - \theta_1^*))\right| = \left|\dot{f}_t(\theta_1^* + r(\theta_1 - \theta_1^*) \mid X)\right| |\theta_1 - \theta_1^*|^2 \leq C\,|\theta_1 - \theta_1^*|^2 \end{aligned}$$

and

$$\begin{aligned} |\partial_r \mu_2^*(X, t; \theta_1^* + r(\theta_1 - \theta_1^*))| &= |\theta_1 - \theta_1^*| \left|1 - \frac{1}{1-\gamma}\left(1 - F_t(\theta_1^* + r(\theta_1 - \theta_1^*) \mid X)\right)\right| \leq |\theta_1 - \theta_1^*| \\ \left|\partial_r^2 \mu_2^*(X, t; \theta_1^* + r(\theta_1 - \theta_1^*))\right| &= |\theta_1 - \theta_1^*|^2\,|f_t(\theta_1^* + r(\theta_1 - \theta_1^*) \mid X)| \leq C\,|\theta_1 - \theta_1^*|^2. \end{aligned}$$

Verifying conditions v. and vi. in Theorem 3. For any $(\theta_1, \theta_2) \in \Theta$,

$$\begin{aligned} \left\{\mathbb{P}\left[\mu_1^*(X, t; \theta_1)\right]^2\right\}^{1/2} &= |F_t(\theta_1 \mid X) - \gamma| \leq 1 \\ \left\{\mathbb{P}\left[\mu_2^*(X, t; \theta_1)\right]^2\right\}^{1/2} &= \left\{\mathbb{P}\left[\mathbb{E}[\max(Y(t) - \theta_1, 0) \mid X]\right]^2\right\}^{1/2} \leq C. \end{aligned}$$

By first-order Taylor expansion, for any θ_1 such that $|\theta_1 - \theta_1^*| \leq \max\{\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N \varepsilon_\pi}, \rho_{\theta,N}\}$, there exists $\tilde{\theta}_1$ between θ_1 and θ_1^* such that

$$\begin{aligned} \left\{\mathbb{P}\left[\mu_1^*(X, t; \theta_1) - \mu_1^*(X, t; \theta_1^*)\right]^2\right\}^{1/2} &= \left\{\mathbb{P}\left[(\theta_1 - \theta_1^*) f_t(\tilde{\theta}_1 \mid X)\right]^2\right\}^{1/2} \leq C\,|\theta_1 - \theta_1^*| \\ \left\{\mathbb{P}\left[\mu_2^*(X, t; \theta_1) - \mu_2^*(X, t; \theta_1^*)\right]^2\right\}^{1/2} &= \left\{\mathbb{P}\left[(\theta_1 - \theta_1^*)(F_t(\tilde{\theta}_1 \mid X) - 1)\right]^2\right\}^{1/2} \leq |\theta_1 - \theta_1^*|. \end{aligned}$$

Moreover, for any θ_1 such that $|\theta_1 - \theta_1^*| \leq \max\{\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N \varepsilon_\pi}, \rho_{\theta,N}\}$, $\left(\partial_{\theta_2}\partial_{\theta_2^\top} V_j(\theta_2)\right) = 0 \leq C$

$$
\begin{aligned}
&\left\{\mathbb{P}\left[\partial_{\theta_1}\mu_1^*(X,t;\theta_1)\right]^2\right\}^{1/2} = \left\{\mathbb{P}\left[f_t(\theta_1 \mid X)\right]^2\right\}^{1/2} \leq C, \\
&\left\{\mathbb{P}\left[\partial_{\theta_1}\mu_2^*(X,t;\theta_1)\right]^2\right\}^{1/2} = \left\{\mathbb{P}\left[F_t(\theta_1 \mid X) - 1\right]^2\right\}^{1/2} \leq 1, \\
&\left|\mathbb{P}\left[\frac{\partial^2}{\partial\theta_1^2}\mu_1^*(X,t;\theta_1)\right]\right| = \left|\mathbb{P}\left[\dot{f}_t(\theta_1 \mid X)\right]\right| \leq C, \left|\mathbb{P}\left[\frac{\partial^2}{\partial\theta_1^2}\mu_2^*(X,t;\theta_1)\right]\right| = \left|\mathbb{P}\left[f_t(\theta_1 \mid X)\right]\right| \leq C.
\end{aligned}
$$

□

Proof for Proposition 5. We follow the proof of Theorem 1 to consider any sequence of data generating process $P_N \in \mathcal{P}_N$ but we suppress it for ease of notation. We prove the conclusion for a generic $k \in \{1, \ldots, K\}$. For $l \in \mathcal{H}_{k,1}$, we denote $\mathbb{P}_{N,l}$ and $\mathbb{G}_{N,l}$ as the empirical average operator and empirical process operator for data in the $\mathcal{D}_l$. Throughout the proof, we condition on the event that the convergence rate of propensity score estimator $\hat{\pi}^{(k,l)}$ in mean squared error is $\rho_{\pi,N}$ and it is lower bounded by ϵ_π, which holds with at least probability $1 - \Delta_N$ according to Assumption 6. In this proof, all notations $\lesssim$ only involve pre-specified constants and not any instance-dependent constants.

We use the following decomposition analogous to that in Step I of proof for Theorem 1.

$$
\begin{aligned}
&\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\pi^*)\right] \\
=&\frac{1}{K'}\sum_{l\in\mathcal{H}_{k,1}}\left\{\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\pi^*)\right] - \mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\hat{\pi}^{(k,l)})\right]\right\} \\
+&\frac{1}{K'}\sum_{l\in\mathcal{H}_{k,1}}\left\{\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\hat{\pi}^{(k,l)})\right] - \mathbb{P}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\hat{\pi}^{(k,l)})\right]\right\} \\
+&\frac{1}{K'}\sum_{l\in\mathcal{H}_{k,1}}\left\{\mathbb{P}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\hat{\pi}^{(k,l)})\right] - \mathbb{P}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\theta^*,\hat{\pi}^{(k,l)})\right]\right\} \\
+&\frac{1}{K'}\sum_{l\in\mathcal{H}_{k,1}}\left\{\mathbb{P}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\theta^*,\hat{\pi}^{(k,l)})\right] - \mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta^*,\hat{\pi}^{(k,l)})\right]\right\} \\
+&\frac{1}{K'}\sum_{l\in\mathcal{H}_{k,1}}\left\{\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta^*,\hat{\pi}^{(k,l)})\right] - \mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta^*,\pi^*)\right]\right\}
\end{aligned}
$$

By following the Step I of proof for Theorem 1, we can also analogously show that

$$
\left\|\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}_{\mathrm{init}}^{(k)},\pi^*)\right]\right\| \leq \frac{4}{K'}\sum_{l\in\mathcal{H}_{k,1}}\mathcal{I}'_{1,l} + \frac{4}{K'}\sum_{l\in\mathcal{H}_{k,1}}\mathcal{I}'_{2,l} + \epsilon_N
$$

where

$$
\begin{aligned}
\mathcal{I}'_{1,l} &= \sup_{\theta\in\Theta}\left\|\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta,\pi^*)\right] - \mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)})\right]\right\| \\
\mathcal{I}'_{2,l} &= \sup_{\theta\in\Theta}\left\|\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)})\right] - \mathbb{P}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)})\right]\right\|.
\end{aligned}
$$

Bounding $\mathcal{I}'_{1,l}$. Note that by condition v. of Theorem 3,

$$\begin{aligned}\mathcal{I}'_{1,l} =& \left\|\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta,\pi^*) - \psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)})\right]\right\| \\ =& \sup_{\theta\in\Theta}\left\|\mathbb{P}\left[\frac{\mu^*(X,t;\theta_1)}{\hat{\pi}^{(k,l)}(X)}\left(\hat{\pi}^{(k,l)}(X) - \pi^*(X)\right)\right]\right\| = \frac{\sqrt{d}\rho_{\pi,N}}{\epsilon_\pi}\max_j\sup_{\theta\in\Theta}\left\{\mathbb{P}\left[\mu_j^*(X,t;\theta_1)\right]^2\right\}^{1/2} \leq \frac{C\sqrt{d}\rho_{\pi,N}}{\epsilon_\pi}.\end{aligned}$$

Bounding $\mathcal{I}'_{2,l}$. Note that

$$\sqrt{\frac{N}{K'}}\mathcal{I}'_{2,l} = \sqrt{\frac{N}{K'}}\sup_{\theta\in\Theta}\left\|\mathbb{P}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)}\right] - \mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)}\right]\right\| = \sup_{\theta\in\Theta}\left\|\mathbb{G}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)}\right]\right\|$$

Given that condition vi. in Assumption 2 is satisfied for the estimating equation ψ^{IPW}, we can follow the end of step I in the proof for Theorem 1 to prove that with P_N probability $1 - c(\log N)^{-1}$ for a constant c that depends on only constants in the assumptions,

$$\sup_{\theta\in\Theta}\left\|\mathbb{G}_{N,l}\left[\psi^{\mathrm{IPW}}(Z;\theta,\hat{\pi}^{(k,l)}\right]\right\| \lesssim \log\left(\frac{N}{K'}\right) + \left(\frac{N}{K'}\right)^{-1/2+1/q'}\log\left(\frac{N}{K'}\right),$$

so that $\mathcal{I}'_{2,l} \lesssim \left(\frac{K'}{N}\right)^{1/2}\log\left(\frac{K'}{N}\right) + \left(\frac{K'}{N}\right)^{1-\frac{1}{q'}}\log\left(\frac{K'}{N}\right) \leq \delta_N\rho_{\pi,N} < \rho_{\pi,N}$.

Therefore, with P_N-probability $1 - c(\log N)^{-1}$,

$$\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}^{(k)}_{\mathrm{init}},\pi^*)\right] \leq \left(\frac{C\sqrt{d}}{\epsilon_\pi} + 1\right)\rho_{\pi,N}.$$

The proof of Theorem 3 shows that conditions ii. and iii. there imply

$$\|J^*(\hat{\theta}^{(k)}_{\mathrm{init}} - \theta^*)\| \wedge c_0 \leq 2\left\|\mathbb{P}\left[\psi^{\mathrm{IPW}}(Z;\hat{\theta}^{(k)}_{\mathrm{init}},\pi^*\right]\right\| \leq 2\left(\frac{C\sqrt{d}}{\epsilon_\pi} + 1\right)\rho_{\pi,N}.$$

Therefore, with probability $1 - c(\log N)^{-1}$:

$$\rho_{\theta,N} = \left\|\hat{\theta}^{(k)}_{1,\mathrm{init}} - \theta^*\right\| \leq \left\|\hat{\theta}^{(k)}_{\mathrm{init}} - \theta^*\right\| \leq \frac{2}{c_3}\left(\frac{C\sqrt{d}}{\epsilon_\pi} + 1\right)\rho_{\pi,N}.$$

□

G.5 Proofs for Appendix

Proof of Proposition 3. In this part, we prove the asymptotic distribution of our estimator $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2^{\mathrm{aux}}\right) \in \Theta_1 \times \Theta_2 \subseteq \mathbb{R}_2$ corresponding to Eqs. (21) and (24). We denote $\theta = (\theta_1, \theta_2^{\mathrm{aux}})$. We prove this by verifying all conditions in the assumptions in Theorem 1.

Verifying Assumption 1. Similar to the proof of Theorem 3, we can easily show that

$$J^* = \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\theta_2^{\mathrm{aux}},\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}$$

does not depend on $\eta_1^*(Z;\theta_1)$ at all. Thus Assumption 1 holds trivially:

$$J^* = \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1),\eta_2^*(Z))\right]\}|_{\theta=\theta^*} = \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\}|_{\theta=\theta^*}.$$

Verifying Assumption 2. We first verify conditions iii. and iv. in Assumption 2. We can easily derive that

$$\mathbb{P}\left[\psi(Z;\theta,\theta_2^{\mathrm{aux}},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right] = \begin{bmatrix} \frac{\mathbb{P}(\mathcal{C})F_1(\theta_1|\mathcal{C})}{\theta_2^{\mathrm{aux}}} - \gamma \\ \theta_2^{\mathrm{aux}\,*} - \theta_2^{\mathrm{aux}} \end{bmatrix}$$

and its Jacobian matrix is given by

$$J(\theta) = \partial_\theta\{\mathbb{P}\left[\psi(Z;\theta,\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\} = \begin{bmatrix} \frac{\mathbb{P}(\mathcal{C})f_1(\theta_1|\mathcal{C})}{\theta_2^{\mathrm{aux}}} & -\frac{\mathbb{P}(\mathcal{C})F_1(\theta_1|\mathcal{C})}{\left(\theta_2^{\mathrm{aux}}\right)^2} \\ 0 & -1 \end{bmatrix}.$$

This means that

$$J(\theta^*) = \begin{bmatrix} f_1\left(\theta_1^* \mid \mathcal{C}\right) & -\frac{\gamma}{\theta_2^{\mathrm{aux}\,*}} \\ 0 & -1 \end{bmatrix}, \quad (J(\theta^*))^{-1} = \begin{bmatrix} \frac{1}{f_1\left(\theta_1^*|\mathcal{C}\right)} & -\frac{\gamma}{\theta_2^{\mathrm{aux}\,*}f_1\left(\theta_1^*|\mathcal{C}\right)} \\ 0 & -1 \end{bmatrix}.$$

Therefore,

$$\sigma_{\max}\left(J(\theta^*)\right) \le 2\max\left\{f_1\left(\theta_1^* \mid \mathcal{C}\right), \frac{\gamma}{\theta_2^{\mathrm{aux}\,*}}, 1\right\} \le 2\max\left\{c_1', \frac{\gamma}{\epsilon}, 1\right\},$$
$$\sigma_{\max}\left(J^{-1}\left(\theta^*\right)\right) \le 2\max\left\{\frac{1}{f_1\left(\theta_1^* \mid \mathcal{C}\right)}, \frac{\gamma}{\theta_2^{\mathrm{aux}\,*}f_1\left(\theta_1^* \mid \mathcal{C}\right)}, 1\right\} \le 2\max\left\{\frac{1}{c_3'}, \frac{\gamma}{c_3'\epsilon}, 1\right\}.$$

The latter implies that $\sigma_{\min}\left(J\left(\theta^*\right)\right) = 1/\sigma_{\max}\left(J^{-1}\left(\theta^*\right)\right) \ge \frac{1}{2}\min\left\{c_3', c_3'\epsilon/\gamma, 1\right\}$. Therefore, condition iv. in Assumption 2 is satisfied with $c_3 = \frac{1}{2}\min\left\{c_3', c_3'\epsilon/\gamma, 1\right\}$, $c_4 = 2\max\left\{c_1', \frac{\gamma}{\epsilon}, 1\right\}$. Moreover, for any $(\theta_1, \theta_2^{\mathrm{aux}}) \in \Theta_1 \times \Theta_2$ and $t = 1$, $f_t(\theta_1 \mid \mathcal{C}) \le c_1'$, $\left|\dot{f}_t(\theta_1 \mid \mathcal{C})\right| \le c_2'$, so we have that entries in $J(\theta)$ are all Lipschtiz with $c_{\mathrm{Lip}} := \max\left\{\sqrt{\left(\frac{c_2'}{\epsilon}\right)^2 + \left(\frac{c_1'}{\epsilon^2}\right)^2}, \sqrt{\left(\frac{2}{\epsilon^3}\right)^2 + \left(\frac{c_1'}{\epsilon^2}\right)^2}\right\}$ as a valid Lipschitz constant. Moreover, we have $2\|\mathbb{P}\left[\psi(Z;\theta,\theta_2^{\mathrm{aux}},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\| \ge c_2$ for all $\theta = (\theta_1, \theta_2^{\mathrm{aux}}) \in \Theta$ such that $\|\theta - \theta^*\| \ge \frac{c_3}{2\sqrt{d}c_{\mathrm{Lip}}}$. By following the proof of Theorem 3, we can easily verify condition iii. in Assumption 2.

Next, we verify condition vi. in Assumption 2. For any fixed $\eta_1\left(Z;\theta_1'\right)$ and η_2, the class $\mathcal{F}_{\eta,\theta_1'} = \{\psi_j(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z)) : j = 1,\ldots,d, \theta \in \Theta\}$ depend on θ only through $\{\mathbb{I}\left[Y \le \theta_1\right] : \theta_1 \in \Theta_1\}$ and $\{\theta_2^{\mathrm{aux}} : \theta_2^{\mathrm{aux}} \in \Theta_2\}$. Since the latter two classes are Donsker class, vi. in Assumption 2 for the function class $\mathcal{F}_{\eta,\theta_1'} = \{\psi_j(Z;\theta,\eta_1(Z;\theta_1'),\eta_2(Z)) : j = 1,\ldots,d, \theta \in \Theta\}$ has to be satisfied as well.

Verifying Assumption 3. We take $\mathcal{T}_N$ to be the set that contains all $(\eta_1\left(\cdot;\theta_1'\right) = \tilde{\mu}(\cdot,\theta_1'), \eta_2\left(\cdot\right) = \left(\nu_w\left(\cdot\right), \tilde{\pi}(\cdot)\right))$ that satisfies the following conditions: for $w = 0, 1$,

$$\left\|\left\{\mathbb{P}\left[\tilde{\mu}_w\left(X;\hat{\theta}_{1,\mathrm{init}}^{(k)}\right) - \tilde{\mu}_w^*\left(X;\hat{\theta}_{1,\mathrm{init}}^{(k)}\right)\right]^2\right\}^{1/2}\right\| \le \tilde{\rho}_{\mu,N}, \ \left\{\mathbb{P}\left[\nu_w(X) - \nu_w^*(X)\right]^2\right\}^{1/2} \le \tilde{\rho}_{\nu,N},$$

$$\left\{\mathbb{P}\left[\tilde{\pi}^{(k)}(X) - \tilde{\pi}^*(X)\right]^2\right\}^{1/2} \le \tilde{\rho}_{\pi,N}, \quad |\theta_1' - \theta_1^*| \le \tilde{\rho}_{\theta,N},$$

and $\epsilon \le \hat{\tilde{\pi}}^{(k)}(X) \le 1-\epsilon, 0 \le \hat{\tilde{\mu}}_w^{(k)}\left(X;\hat{\theta}^{(k)}_{1,\text{init}})\right) \le 1,\ 0 \le \hat{\nu}_w^{(k)}(X) \le 1$ almost surely. Moreover, $\tilde{\rho}_{\pi,N} \le \frac{\delta_N^3}{\log N}$, $\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N} \le \frac{\delta_N^2}{\log N}$, $\tilde{\rho}_{\pi,N}\left(\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N}\right) \le \frac{\epsilon^4(1-\epsilon)^3}{4\left(\epsilon^3+(1-\epsilon)^3\right)}\delta_N N^{-1/2}$, $\tilde{\rho}_{\pi,N}\tilde{\rho}_{\nu,N} \le \frac{\epsilon^3(1-\epsilon)^3}{8\left(\epsilon^3+(1-\epsilon)^3\right)}\delta_N N^{-1/2}$ with δ_N satisfying that $\delta_N \le \frac{\epsilon^3(1-\epsilon)^2}{4C+3\epsilon^2(1-\epsilon)}$, $\frac{\delta_N}{\log N} \le \frac{1}{C_\epsilon}$ for a positive constant C_ϵ given in Eq. (44).

Then Assumption 8 and Proposition 3 condition v. ensure that the nuisance estimates $(\hat{\mu}(,\hat{\theta}_{1,\text{init}}),\hat{\pi}) \in \mathcal{T}_N$ with probability, namely, condition i. in Assumption 3 is satisfied.

Before verifying other conditions, we first note that

$$\begin{aligned}\tilde{\mu}_w^*\left(X;\theta_1\right) &= \mathbb{P}\left(T=1, Y\le\theta_1 \mid X, W=w\right)\\ &= \mathbb{P}\left(T(w)=1, Y(1)\le\theta_1 \mid X\right) = F_{1,w}\left(\theta_1 \mid X\right) v_w\left(X\right).\end{aligned}$$

It follows from Item iv. that for any $\theta_1 \in \mathcal{B}(\theta_1^*; \max\{\frac{4\tilde{\rho}_{\pi,N}}{\epsilon^2(1-\epsilon)\delta_N}, \rho_{\theta,N}\}) \cap \Theta$,

$$\left[\mathbb{P}\left[\left(\tilde{\mu}_w^*\left(X;\theta_1\right) - \tilde{\mu}_w^*\left(X;\theta_1^*\right)\right)^2\right]\right]^{1/2} \le C\|\theta_1-\theta_1^*\|.$$

This means that for any $(\mu(\cdot,\theta_1'),\pi(\cdot)) \in \mathcal{T}_N$,

$$\begin{aligned}&\left\|\left\{\mathbb{P}\left[\mu(X,T;\theta_1') - \mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\|\\ \le &\left\|\left\{\mathbb{P}\left[\mu(X,T;\theta_1') - \mu^*(X,T;\theta_1')\right]^2\right\}^{1/2}\right\| + \left\|\left\{\mathbb{P}\left[\mu^*(X,T;\theta_1') - \mu^*(X,T;\theta_1^*)\right]^2\right\}^{1/2}\right\| = \tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N}.\end{aligned}$$

Next, we verify Assumption 3 condition ii.. We first verify the condition on r_N. By following the proof of Theorem 3, we can show that for any $(\eta_1(\cdot;\theta_1'),\eta_2(\cdot)) \in \mathcal{T}_N$,

$$\begin{aligned}&\|\mathbb{P}\left[\psi(Z;\theta,\theta_2^{\text{aux}},\eta_1(Z;\theta_1'),\eta_2(Z))\right] - \mathbb{P}\left[\psi(Z;\theta,\theta_2^{\text{aux}},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]\|\\ \le&\left\|\frac{1}{\theta_2^{\text{aux}}}\mathbb{P}\left(\frac{W-\tilde{\pi}\left(X\right)}{\tilde{\pi}\left(X\right)\left(1-\tilde{\pi}\left(X\right)\right)} - \frac{W-\tilde{\pi}^*\left(X\right)}{\tilde{\pi}^*\left(X\right)\left(1-\tilde{\pi}^*\left(X\right)\right)}\right)\left(\tilde{\mu}_W^*\left(X;\theta_1\right) - \tilde{\mu}_W^*\left(X;\theta_1^*\right)\right)\right\|\\ +&\left\|\frac{1}{\theta_2^{\text{aux}}}\mathbb{P}\left(\frac{W-\tilde{\pi}\left(X\right)}{\tilde{\pi}\left(X\right)\left(1-\tilde{\pi}\left(X\right)\right)} - \frac{W-\tilde{\pi}^*\left(X\right)}{\tilde{\pi}^*\left(X\right)\left(1-\tilde{\pi}^*\left(X\right)\right)}\right)\left(\tilde{\mu}_W^*\left(X;\theta_1^*\right) - \tilde{\mu}_W\left(X;\theta_1'\right)\right)\right\| \le \frac{4}{\epsilon^2\left(1-\epsilon\right)}\tilde{\rho}_{\pi,N}.\end{aligned}$$

The last inequality holds because

$$\tilde{\mu}_w^*\left(X;\theta_1\right) = \mathbb{P}\left(T\left(w\right)=1, Y\left(1\right)\le\theta_1 \mid X\right) \in [0,1],\ \text{almost surely},$$

and so is $\tilde{\mu}_w\left(X;\theta_1\right)$. This means that the condition on r_N is satisfied with $\tau_N = \frac{4\tilde{\rho}_{\pi,N}}{\epsilon^2(1-\epsilon)\delta_N}$.

Next, we verify the condition on r'_N. Again, by following the proof of Theorem 3, we have that for any $\|\theta-\theta^*\| \leq \frac{4\tilde{\rho}_{\pi,N}}{\epsilon^2(1-\epsilon)\delta_N}$ and any $(\eta_1(\cdot;\theta'_1), \eta_2(\cdot)) \in \mathcal{T}_N$,

$$
\begin{aligned}
&\left\|\left\{\mathbb{P}\left[\psi(Z;\theta,\theta_2^{\mathrm{aux}},\eta_1(Z;\theta'_1),\eta_2(Z)) - \psi(Z;\theta,\theta_2^{\mathrm{aux}},\eta_1^*(Z;\theta_1^*),\eta_2^*(Z))\right]^2\right\}^{1/2}\right\| \\
\leq&\left\|\left\{\mathbb{P}\left(\frac{W-\tilde{\pi}\left(X\right)}{\tilde{\pi}\left(X\right)\left(1-\tilde{\pi}\left(X\right)\right)} - \frac{W-\tilde{\pi}^*\left(X\right)}{\tilde{\pi}^*\left(X\right)\left(1-\tilde{\pi}^*\left(X\right)\right)}\right)^2\left(\tilde{\mu}_W^*\left(X;\theta_1\right)-\tilde{\mu}_W^*\left(X;\theta_1^*\right)\right)^2\right\}^{1/2}\right\| \\
+&\left\|\left\{\mathbb{P}\left(\frac{W-\tilde{\pi}\left(X\right)}{\tilde{\pi}\left(X\right)\left(1-\tilde{\pi}\left(X\right)\right)} - \frac{W-\tilde{\pi}^*\left(X\right)}{\tilde{\pi}^*\left(X\right)\left(1-\tilde{\pi}^*\left(X\right)\right)}\right)^2\left(\tilde{\mu}_W^*\left(X;\theta_1^*\right)-\tilde{\mu}_W\left(X;\theta'_1\right)\right)^2\right\}^{1/2}\right\| \\
+&\left\|\left\{\mathbb{P}\left(\frac{W-\tilde{\pi}^*\left(X\right)}{\tilde{\pi}^*\left(X\right)\left(1-\tilde{\pi}^*\left(X\right)\right)}\right)^2\left(\tilde{\mu}_W^*\left(X;\theta_1^*\right)-\tilde{\mu}_W\left(X;\theta'_1\right)\right)^2\right\}^{1/2}\right\|.
\end{aligned}
$$

The above can be further upper bounded by

$$
\begin{aligned}
&\frac{4C}{\epsilon^3\left(1-\epsilon\right)^2\delta_N}\tilde{\rho}_{\pi,N} + \frac{1}{\epsilon\left(1-\epsilon\right)}\left(\tilde{\rho}_{\mu,N}+C\tilde{\rho}_{\theta,N}\right) \\
&+\frac{1}{\epsilon\left(1-\epsilon\right)}\left\{\mathbb{P}\left[\left(\tilde{\mu}_1^*\left(X;\theta_1^*\right)-\tilde{\mu}_1\left(X;\theta'_1\right)\right)^2\right]\right\}^{1/2} + \frac{1}{\epsilon\left(1-\epsilon\right)}\left\{\mathbb{P}\left[\left(\tilde{\mu}_0^*\left(X;\theta_1^*\right)-\tilde{\mu}_0\left(X;\theta'_1\right)\right)^2\right]\right\}^{1/2} \\
\leq&\frac{4C}{\epsilon^3\left(1-\epsilon\right)^2\delta_N}\tilde{\rho}_{\pi,N} + \frac{3}{\epsilon\left(1-\epsilon\right)}\left(\tilde{\rho}_{\mu,N}+C\tilde{\rho}_{\theta,N}\right).
\end{aligned}
$$

Therefore, if $\tilde{\rho}_{\pi,N} \leq \frac{\delta_N^3}{\log N}$ and $\tilde{\rho}_{\mu,N}+C\tilde{\rho}_{\theta,N} \leq \frac{\delta_N^2}{\log N}$, then $r'_N = \frac{\delta_N^2}{\log N}\left(\frac{4C}{\epsilon^3(1-\epsilon)^2}+\frac{3}{\epsilon(1-\epsilon)}\right) \leq \frac{\delta_N}{\log N}$ given $\delta_N \leq \frac{\epsilon^3(1-\epsilon)^2}{4C+3\epsilon^2(1-\epsilon)}$.

Finally, we verify the condition on λ'_N. Note that in this case $V\left(\theta_2\right)=0$ and denote

$$
\tilde{\psi}_1\left(Z;\theta,\eta_1\left(Z;\theta_1\right),\eta_2\left(Z\right)\right) \coloneqq \theta_2^{\mathrm{aux}}\psi_1(Z;\theta,\eta_1(Z;\theta'_1),\eta_2(Z)).
$$

Then for any $(\eta_1(\cdot;\theta'_1),\eta_2(\cdot)) \in \mathcal{T}_N$ and $\theta \in \mathcal{B}\left(\theta^*;\frac{4\tilde{\rho}_{\pi,N}}{\epsilon^2(1-\epsilon)\delta_N}\right)$, letting $\theta_r = \theta^*+r\left(\theta-\theta^*\right)$, $\eta_{1,r}(Z) = \eta_1^*(Z;\theta_1)+r\left(\eta_1(Z;\theta'_1)-\eta_1^*(Z;\theta_1)\right), \eta_2^*(Z)$, $\eta_{2,r}(Z) = \eta_2^*(Z)+r\left(\eta_2(Z)-\eta_2^*(Z)\right)$, we have

$$
\begin{aligned}
\left|\partial_r^2\mathbb{P}\left[\psi_1(Z;\theta_r,\eta_{1,r}(Z),\eta_{2,r}(Z))\right]\right| = &\left|\partial_r^2\mathbb{P}\left[\psi_1(Z;\theta_r,\eta_{1,r}(Z),\eta_{2,r}(Z))\right]\right| \times \frac{1}{\theta_2^{\mathrm{aux}\,*}+r\left(\theta_2^{\mathrm{aux}}-\theta_2^{\mathrm{aux}\,*}\right)} \\
&+2\left|\partial_r\mathbb{P}\left[\psi_1(Z;\theta_r,\eta_{1,r}(Z),\eta_{2,r}(Z))\right]\right| \times \frac{\theta_2^{\mathrm{aux}}-\theta_2^{\mathrm{aux}\,*}}{\left(\theta_2^{\mathrm{aux}\,*}+r\left(\theta_2^{\mathrm{aux}}-\theta_2^{\mathrm{aux}\,*}\right)\right)^2} \\
&+\left|\mathbb{P}\left[\psi_1(Z;\theta_r,\eta_{1,r}(Z),\eta_{2,r}(Z))\right]\right| \times \frac{2\left(\theta_2^{\mathrm{aux}}-\theta_2^{\mathrm{aux}\,*}\right)^2}{\left(\theta_2^{\mathrm{aux}\,*}+r\left(\theta_2^{\mathrm{aux}}-\theta_2^{\mathrm{aux}\,*}\right)\right)^3}.
\end{aligned}
$$

By following the proof of Theorem 3, we can bound each term above and prove that

$$
\begin{aligned}
&\left|\partial_r^2\mathbb{P}\left[\psi_1(Z;\theta^*+r\left(\theta-\theta^*\right),\eta_1^*(Z;\theta_1)+r\left(\eta_1(Z;\theta'_1)-\eta_1^*(Z;\theta_1)\right),\eta_2^*(Z)+r\left(\eta_2(Z)-\eta_2^*(Z)\right))\right]\right| \\
\leq&4\left(\frac{1}{\epsilon^4}+\frac{1}{\left(1-\epsilon\right)^3\epsilon}\right)\tilde{\rho}_{\pi,N}\left(\tilde{\rho}_{\mu,N}+C\tilde{\rho}_{\theta,N}\right)+\left(C_{\epsilon,1}\frac{\tilde{\rho}_{\pi,N}}{\delta_N}+C_{\epsilon,2}\left(\tilde{\rho}_{\mu,N}+C\tilde{\rho}_{\theta,N}\right)\right)\|\theta-\theta^*\|,
\end{aligned}
$$

where

$$C_{\epsilon,1} = 4C\left(\frac{1}{\epsilon^3(1-\epsilon)} + \frac{1}{\epsilon^2(1-\epsilon)^2}\right) + \frac{1}{\epsilon^2}\left(\frac{2}{\epsilon^2} + \frac{2}{(1-\epsilon)^2} + \frac{4C}{\epsilon^3(1-\epsilon)} + \frac{4C}{\epsilon^2(1-\epsilon)^2}\right)$$
$$+ 16\left(1 + \frac{1}{\epsilon} + \frac{1}{1-\epsilon}\right)\frac{1}{\epsilon^5(1-\epsilon)}, \quad C_{\epsilon,2} = \frac{2}{\epsilon^2}\left(1 + \frac{1}{\epsilon} + \frac{1}{1-\epsilon}\right).$$

Also define

$$C_\epsilon = C_{\epsilon,1} + C_{\epsilon,2}. \tag{44}$$

Since $\tilde{\rho}_{\pi,N} \leq \frac{\delta_N^3}{\log N}$ and $\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N} \leq \frac{\delta_N^2}{\log N}$, if $\frac{\delta_N}{\log N} \leq \frac{1}{C_\epsilon}$, then

$$\left(C_{\epsilon,1}\frac{\tilde{\rho}_{\pi,N}}{\delta_N} + C_{\epsilon,2}\left(\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N}\right)\right) \leq C_\epsilon \frac{\delta_N^2}{\log N} \leq \delta_N.$$

Morevoer, when $\tilde{\rho}_{\pi,N}\left(\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N}\right) \leq \frac{\epsilon^4(1-\epsilon)^3}{8\left(\epsilon^3+(1-\epsilon)^3\right)}\delta_N N^{-1/2}$, we have

$$4\left(\frac{1}{\epsilon^4} + \frac{1}{(1-\epsilon)^3\epsilon}\right)\tilde{\rho}_{\pi,N}\left(\tilde{\rho}_{\mu,N} + C\tilde{\rho}_{\theta,N}\right) \leq \frac{1}{2}\delta_N N^{-1/2}.$$

Plus, we can similarly show that given $\tilde{\rho}_{\pi,N}\tilde{\rho}_{\nu,N} \leq \frac{\epsilon^3(1-\epsilon)^3}{8\left(\epsilon^3+(1-\epsilon)^3\right)}\delta_N N^{-1/2}$,

$$\left|\partial_r^2 \mathbb{P}\left[\psi_2(Z;\theta_2^{\mathrm{aux}\,*} + r\left(\theta_2^{\mathrm{aux}} - \theta_2^{\mathrm{aux}\,*}\right), \eta_2^*(Z) + r\left(\eta_2(Z) - \eta_2^*(Z)\right))\right]\right| \leq 4\left(\frac{1}{\epsilon^3} + \frac{1}{(1-\epsilon)^3}\right)\tilde{\rho}_{\pi,N}\tilde{\rho}_{\nu,N} \leq \frac{1}{2}\delta_N N^{-1/2}.$$

Then Assumption 3 condition ii. follows from

$$\begin{aligned}&\left\|\partial_r^2 \mathbb{P}\left[\psi(Z;\theta^* + r\left(\theta - \theta^*\right), \eta_1^*(Z;\theta_1) + r\left(\eta_1(Z;\theta_1') - \eta_1^*(Z;\theta_1)\right), \eta_2^*(Z) + r\left(\eta_2(Z) - \eta_2^*(Z)\right))\right]\right\| \\ \leq& \delta_N\|\theta - \theta^*\| + \delta_N N^{-1/2},\end{aligned}$$

Therefore, we have

$$\sqrt{N}\begin{bmatrix}\hat{\theta}_1 - \theta_1^* \\ \hat{\theta}_2^{\mathrm{aux}} - \theta_2^{\mathrm{aux}\,*}\end{bmatrix} = \frac{1}{\sqrt{N}}\sum_{i=1}^N J^{-1}\left(\theta^*\right)\begin{bmatrix}\psi_1(Z_i;\theta^*, \eta_1^*(Z_i;\theta_1^*), \eta_2^*(Z_i)) \\ \psi_2(Z_i;\theta_2^{\mathrm{aux}\,*}, \eta_2^*(Z_i))\end{bmatrix} + O_{\mathbb{P}}\left(\rho_N\right).$$

□

Proof for Proposition 4. We only need to verify the conditions in Theorem 3.

Verifying condition i. in Theorem 3. We only need to verify that condition vi. of Assumption 2 hold. Since Θ is compact, $\{y \mapsto (1-\gamma)y - \theta_1, \theta \in \Theta\}$, $\{y \mapsto (1-2\gamma)\max\{y - \theta_1, 0\}, \theta \in \Theta\}$ are obviously Donsker classes, condition vi. of Assumption 2 is satisfied.

Verifying condition ii. and iii. in Theorem 3. According to Eq. (6), the estimating function for complete data is given by $U(Y(1);\theta_1) = (1-\gamma)\,(Y(1)-\theta_1) - (1-2\gamma)\max\,(Y(1)-\theta_1, 0)$. It follows that

$$\begin{aligned}\frac{\partial}{\partial\theta_1}\mathbb{P}\left[U(Y(t);\theta_1)\right] &= -(1-\gamma) - (1-2\gamma)\frac{\partial}{\partial\theta_1}\mathbb{P}\left[\max\left(Y(t)-\theta_1, 0\right)\right] \\ &= -(1-\gamma) - (1-2\gamma)\frac{\partial}{\partial\theta_1}\int_{\theta_1}^{\infty}(y-\theta_1)f_t(y)dy = -\gamma-(1-2\gamma)F_t(\theta_1).\end{aligned}$$

Here the differentiability of $\frac{\partial}{\partial\theta_1}\mathbb{P}\left[U(Y(t);\theta_1)\right]$ is guaranteed by Leibniz integral rule, the continuity of its derivative at θ_1^* is guaranteed by the continuity of $F_t(\theta_1)$ at θ_1^* , and $J(\theta_1^*) = \frac{\partial}{\partial\theta_1}\mathbb{P}\left[U(Y(t);\theta_1)\right]|_{\theta_1=\theta_1^*} = -\gamma-(1-2\gamma)F_t(\theta_1^*)$, whose singular value $|-\gamma-(1-2\gamma)F_t(\theta_1^*)|$ is bounded between c_4' and $\max\{\gamma, 1-\gamma\}$. Moreover, $\frac{\partial}{\partial\theta_1}\mathbb{P}\left[U(Y(t);\theta_1)\right] \leq \max\{\gamma, 1-\gamma\}$, which implies that $\mathbb{P}\left[U(Y(t);\theta_1)\right]$ is Lipschitz continuous with Lipschitz constant $\max\{\gamma, 1-\gamma\} \leq 1$. Therefore, the constants c' in condition ii. and constant c_3 in iii. of Theorem 3 can be set as $c_3 = c_1', c' = 1$. The assumption that $\|\theta-\theta^*\| \geq \frac{c_3}{2c'} = \frac{c_1'}{2}$, $2\mathbb{P}\left[U(Y(t);\theta_1)\right] \geq c_2'$ for any $\theta \in \Theta$ ensures the condition ii. of Theorem 3 with constant $c_2 = c_2'$.

Verifying condition iv. in Theorem 3. Note that for any $\theta \in \mathcal{B}(\theta^*; \frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}) \cap \Theta$, we have $\mu^*(X, 1; \theta_1^* + r(\theta-\theta_1^*)) = (1-\gamma)\eta_{2,1}^*(Z) - (1-2\gamma)\eta_1^*(Z; \theta_1^* + r(\theta_1-\theta_1^*))$. Thus

$$\begin{aligned}&|\partial_r\mu^*(X, 1; \theta_1^* + r(\theta_1-\theta_1^*))| = |-\gamma(\theta_1-\theta_1^*) - (1-2\gamma)(\theta_1-\theta_1^*)F_t(\theta_1^* + r(\theta-\theta_1^*) \mid X)| \leq 2|\theta_1-\theta_1^*|, \\ &\left|\partial_r^2\mu^*(X, 1; \theta_1^* + r(\theta_1-\theta_1^*))\right| = |1-2\gamma||\theta_1-\theta_1^*|f_t(\theta_1^* + r(\theta_1-\theta_1^*) \mid X) \leq C|1-2\gamma||\theta_1-\theta_1^*|,\end{aligned}$$

which trivially imply condition iv. in Theorem 3.

Verifying condition iv in Theorem 3. Again $\mu^*(X, 1; \theta_1) = (1-\gamma)\eta_{2,1}^*(Z) - (1-2\gamma)\eta_1^*(Z; \theta_1)$. The the asserted assumpton iv means that $\left\{\mathbb{P}[\eta_{2,1}^*(Z)]^2\right\}^{1/2} \leq C$ and $\left\{\mathbb{P}[\eta_1^*(Z;\theta_1)]^2\right\}^{1/2} \leq C$ for any $\theta \in \Theta$, thus $\left\{\mathbb{P}\left[\mu^*(X, 1; \theta_1)\right]^2\right\}^{1/2}$ is upper bounded by $|1-\gamma| + |1-2\gamma|\, C \leq 2C$ for any $\theta \in \Theta$. Plus, for any $\theta_1 \in \mathcal{B}(\theta_1^*; \max\{\frac{4C\sqrt{d}\rho_{\pi,N}}{\delta_N\varepsilon_\pi}, \rho_{\pi,N}\}) \cap \Theta$, we have

$$\begin{aligned}\{\mathbb{P}\left[\frac{\partial}{\partial\theta_1}\mu^*(X, 1; \theta_1)\right]^2\}^{1/2} &\leq \sup_x |-\gamma-(1-2\gamma)F_t(\theta_1 \mid X = x)| \leq 2, \\ \mathbb{P}\left[\frac{\partial^2}{\partial\theta_1^2}\mu^*(X, 1; \theta_1)\right] &\leq |1-2\gamma|\mathbb{P}\left[f_t(\theta_1 \mid X)\right] \leq C|1-2\gamma|,\end{aligned}$$

and there exists $\tilde{\theta}_1$ between θ_1 and θ_1^* such that

$$\left\{\mathbb{P}\left[\mu^*(X, 1; \theta_1) - \mu^*(X, 1; \theta_1^*)\right]^2\right\}^{1/2} = |\theta_1-\theta_1^*|\left\{\mathbb{P}\left[\frac{\partial}{\partial\theta_1}\mu^*(X, 1; \tilde{\theta}_1)\right]^2\right\}^{1/2} \leq 2|\theta_1-\theta_1^*|.$$

□