mathematics

MDPI

Article

Non-Asymptotic Bounds of AIPW Estimators for Means with Missingness at Random

Fei Wang [1] and Yuhao Deng [2,*]

[1] College of Science, Minzu University of China, Beijing 100081, China
[2] School of Mathematical Sciences, Peking University, Beijing 100871, China
* Correspondence: dengyuhao@pku.edu.cn

Abstract: The augmented inverse probability weighting is well known for its double robustness in missing data and causal inference. If either the propensity score model or the outcome regression model is correctly specified, the estimator is guaranteed to be consistent. Another important property of the augmented inverse probability weighting is that it can achieve first-order equivalence to the oracle estimator in which all nuisance parameters are known, even if the fitted models do not converge at the parametric root-n rate. We explore the non-asymptotic properties of the augmented inverse probability weighting estimator to infer the population mean with missingness at random. We also consider inferences of the mean outcomes on the observed group and on the unobserved group.

Keywords: augmented inverse probability weighting; causal inference; double robustness; missing data; non-asymptotic

MSC: 62G32

Citation: Wang, F.; Deng, Y. Non-Asymptotic Bounds of AIPW Estimators for Means with Missingness at Random. *Mathematics* **2023**, *11*, 818. https://doi.org/10.3390/math11040818

Academic Editor: Christophe Chesneau

Received: 31 December 2022
Revised: 1 February 2023
Accepted: 3 February 2023
Published: 6 February 2023

1. Introduction

Missingness is an important issue in statistics. Suppose we are interested in the population mean of an outcome. To estimate the population mean, we randomly draw n independent units to form a sample. It is well known that the empirical average of the outcome in this sample is an unbiased estimator of the population mean. However, if there is missingness, we would only have measures of outcomes on $r \leq n$ units. Missingness may rely on the background characters of units, so the observed units and unobserved units may be different in covariates. As a result, estimating the population mean by the observed sample average is biased. Rubin [1] formalized the concept of "missing-at-random", saying that missingness depends on observed values but not on unobserved values. More specifically, missingness of an outcome depends on the observed covariates rather than outcomes [2]. Under the missing-at-random assumption, the population mean is identifiable.

With missingness at random, the population mean can be estimated by inverse probability weighting [3]. This approach is based on correct model specification on propensity scores and involves the estimation of propensity scores. Hirano et al. [4] discussed the consequence of using estimated propensity scores in inverse probability weighting. Using estimated propensity scores could lead to a more efficient estimator than using the true propensity score, but the exact properties of the former are complex. To improve the inverse probability weighting, Robins et al. [5] proposed augmented inverse probability weighting (AIPW) that combines the inverse probability weighting with outcome regression. If both the propensity score model and the outcome regression model are known, the AIPW estimator is the most efficient estimator in a set of regular and asymptotic linear estimators [6,7]. The AIPW estimator has double robustness, in that the estimator is consistent if either the

propensity score model or outcome regression model is correctly specified [8]. Farrell [9] discussed AIPW with high-dimensional covariates.

The asymptotic properties of the AIPW estimator have been well-studied in recent years. Chernozhukov et al. [10] found that if both the fitted propensity score model and the outcome regression model are sup-norm consistent and converge not too slowly—for example, if all models converge at rate $o_p(n^{-1/4})$—, then, the AIPW estimator with estimated model parameters can achieve first-order equivalence with the oracle estimator in which both models are known. The (empirical) AIPW estimator and oracle estimator differ by $o_p(n^{-1/2})$. Therefore, the large sample asymptotic law of the AIPW estimator is the same as that of the oracle one, converging in distribution to a normal distribution. Newey and Robins [11] and Kennedy [12] discussed more general conditions to achieve root-n convergence with estimated nuisance models by cross-fitting (also called sample-splitting).

Existing works mainly focus on the asymptotic properties of the AIPW estimator. In this paper, we study the AIPW estimator from a non-asymptotic view. The main focus is on the non-asymptotic error bounds of the estimated mean outcomes by AIPW. Zhang and Chen [13] and Zhang and Wei [14] reviewed a series of concentration inequalities that can be used to bound the tail probability of an estimator. We apply the concentration inequalities to study the behavior of doubly robust estimators of the population mean, the population mean in the observed group and the population mean in the unobserved group by AIPW. We will first discuss bounded outcomes, and then extend the results to subGaussian outcomes. Furthermore, we conduct a simulation study to compare the non-asymptotic error bounds for bounded outcomes and sub-Gaussian outcomes.

2. Notations

2.1. Missingness at Random

Consider a sample $\mathcal{I} = \{1, \ldots, n\}$ randomly drawn from a super-population. For each unit $i \in \mathcal{I}$, let Z_i be the missing indicator, where $Z_i = 1$ if the outcome Y_i is observed and $Z_i = 0$ if the outcome Y_i is not observed. We collect a vector of covariates X_i for each unit, which is predictable of the missing propensity or outcomes. The observed variables of the ith unit is $O_i = (Z_i, Z_iY_i, X_i)$, for $i = 1, \ldots, n$. The observed data $\{O_i\}_{i=1}^n$ are n independent and identically distributed (iid) copies of $O = (Z, ZY, X)$. Define the propensity score

$$e(x) = P(Z = 1 \mid X = x)$$

and the outcome regression

$$m(x) = E(Y \mid X = x).$$

In many scenarios in biostatistics, economics and social sciences, we are interested in the mean outcome in the overall population $\tau = E(Y)$. We assume missing at random (MAR), that is, whether a unit is missing is independent of its outcome.

Assumption 1 (Missingness at random)**.** $Z \perp\!\!\!\perp Y \mid X$.

Under Assumption 1, $m(x) = E(Y \mid Z = 1, X = x)$. Moreover, we also assume the propensity score is bounded far from 0. Each unit has a positive probability of being observed.

Assumption 2 (Positivity)**.** $e(X) > \eta > 0$, *where* η *is a known constant.*

2.2. Augmented Inverse Probability Weighting

The information of X can be summarized into a one-dimensional scalar, the propensity score $e(X)$. Conditioning on the propensity score rather than all the covariates, we have

$Z \perp\!\!\!\perp Y \mid e(X)$ [3]. The inverse probability weighting (Horvitz-Thompson [15]) utilizes this property and identifies the target estimand by

$$\tau = E\left\{\frac{ZY}{e(X)}\right\}. \tag{1}$$

Another approach to identifying τ is outcome regression. Since we can estimate the mean outcome in observed units, this information can be generalized to the whole population at each level of covariates X. Therefore, we can simply express τ as

$$\tau = E[E\{Y \mid Z = 1, X\}] = E\{m(X)\}, \tag{2}$$

which is also identifiable from observed data. The inner expectation in the second formula is taken over Y given X in the observable population, and the outer expectation is taken over X in the whole population.

The inverse probability weighting and outcome regression can be combined into the following expressions by appending an augmented term, referred to as the augmented inverse probability weighting (AIPW),

$$\tau = E\left\{\frac{ZY}{e(X)} + \left(1 - \frac{Z}{e(X)}\right)m(X)\right\} \tag{3}$$

$$= E\left\{\frac{Z(Y - m(X))}{e(X)} + m(X)\right\}. \tag{4}$$

Equation (3) can be understood as the ordinary inverse probability weighting, plus an augmented term, to correct for the potential bias of the propensity score. With a little transformation, Equation (4) can be understood as the ordinary outcome regression plus a correction for the potential bias of the outcome regression model. The augmented inverse probability weighting has a well-known property of double robustness, in that the equations above hold if either the propensity score model $e(x)$ or the outcome regression model $m(x)$ is correctly specified. That is,

$$\tau = E\left\{\frac{ZY}{e(X)} + \left(1 - \frac{Z}{e(X)}\right)m^*(X)\right\} \tag{5}$$

$$= E\left\{\frac{Z(Y - m(X))}{e^*(X)} + m(X)\right\} \tag{6}$$

for any functions $e^*(x)$ and $m^*(x)$. Another important property of AIPW is that the term in the expectation is the efficient influence function of τ (differing by a constant τ), so that estimation based on AIPW is the most efficient.

Define the oracle AIPW estimator as

$$\widehat{\tau}^* = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i Y_i}{e(X_i)} + \left(1 - \frac{Z_i}{e(X_i)}\right)m(X_i)\right\}. \tag{7}$$

It is an average of n independent random variables. We can easily prove that $\widehat{\tau}^*$ is an unbiased and consistent estimator for τ, with $E(\widehat{\tau}^*) = \tau$. In practice, the propensity score and the outcome regression models are unknown. Suppose we estimate them by $\widehat{e}(x)$ and $\widehat{m}(x)$. The (empirical) AIPW estimator becomes

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i Y_i}{\widehat{e}(X_i)} + \left(1 - \frac{Z_i}{\widehat{e}(X_i)}\right)\widehat{m}(X_i)\right\}. \tag{8}$$

Since Equation (8) involves fitted models, it is not an average of n independent random variables anymore.

3. Construction of Error Bounds for Bounded Outcomes

Lemma 1 (McDiarmid's inequality [16]). *Suppose $O_1, \ldots, O_n$ are independent random variables taking values in the set $\mathcal{O}$, and assume $f : \mathcal{O}^n \to \mathbb{R}$ satisfies the bounded difference condition (BDC)*

$$\sup_{o_1,\ldots,o_n,o_k' \in \mathcal{O}} |f(o_1,\ldots,o_n) - f(o_1,\ldots,o_{k-1},o_k',o_{k+1},\ldots,o_n)| \le c_k.$$

Then,

$$P(|f(O_1,\ldots,O_n) - E\{f(O_1,\ldots,O_n)\}| \ge t) \le 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right\}, \quad \forall\, t > 0.$$

We assume that $|Y| \le M$, where M is a positive constant. In fact, if Y is not bounded, we can transform Y to be a bounded random variable. In the following sections, we consider the non-asymptotic inferences of $\tau = E(Y)$, $\tau_1 = E(Y \mid Z = 1)$ and $\tau_0 = E(Y \mid Z = 0)$, respectively.

3.1. Mean Outcome

We first consider the oracle estimator $\widehat{\tau}^*$. To verify the bounded difference condition, suppose we replace the observation of the kth unit $O_k = (Z_k, Z_kY_k, X_k)$ with $O_k' = (Z_k', Z_k'Y_k', X_k')$. Then,

$$\begin{aligned}
|D^*| :&= |\widehat{\tau}^*(O_1,\ldots,O_k,\ldots,O_n) - \widehat{\tau}^*(O_1,\ldots,O_k',\ldots,O_n)| \\
&= \left|\frac{1}{n}\left\{\frac{Z_kY_k}{e(X_k)} + \left(1 - \frac{Z_k}{e(X_k)}\right)m(X_k)\right\} - \frac{1}{n}\left\{\frac{Z_k'Y_k'}{e(X_k')} + \left(1 - \frac{Z_k'}{e(X_k')}\right)m(X_k')\right\}\right| \\
&\le \frac{2}{n}\cdot\left(\frac{2}{\eta} - 1\right)M =: c^*.
\end{aligned}$$

By McDiarmid's inequality, since $E(\widehat{\tau}^*) = \tau$, we have

$$P\left(\sqrt{n}|\widehat{\tau}^* - \tau| \ge t\right) \le 2\exp\left\{-\frac{2t^2}{n^2c^{*2}}\right\} = 2\exp\left\{-\frac{t^2}{2(2/\eta - 1)^2M^2}\right\}. \tag{9}$$

To study the properties of the AIPW estimator (8) with estimated nuisance models, Chernozhukov et al. [10] proposed an approach by cross-fitting. Suppose the full sample is randomly divided into two halves, $\mathcal{I}_1$ and $\mathcal{I}_2$, with similar sample sizes $|\mathcal{I}_1|$ and $|\mathcal{I}_2|$. The models $\widehat{e}(x)$ and $\widehat{m}(x)$ are estimated using a half sample and then fitted on the other half. To be more specific, for each unit with covariates x_i in the $(3-l)$th half, $e(x_i)$ and $m(x_i)$ are estimated as $\widehat{e}^{(l)}(x_i)$ and $\widehat{m}^{(l)}(x_i)$, which are fitted using the lth half of sample ($l = 1, 2$). The AIPW estimator $\widehat{\tau}$ is given by averaging the estimated mean outcomes in these two halves of the full sample,

$$\widehat{\tau} = \frac{|\mathcal{I}_1|}{n}\widehat{\tau}^{(\mathcal{I}_1)} + \frac{|\mathcal{I}_2|}{n}\widehat{\tau}^{(\mathcal{I}_2)}, \tag{10}$$

where

$$\widehat{\tau}^{(\mathcal{I}_l)} = \frac{1}{|\mathcal{I}_l|}\sum_{i\in\mathcal{I}_l}\left\{\frac{Z_iY_i}{\widehat{e}^{(3-l)}(X_i)} + \left(1 - \frac{Z_i}{\widehat{e}^{(3-l)}(X_i)}\right)\widehat{m}^{(3-l)}(X_i)\right\}. \tag{11}$$

As a comparison, the oracle estimator can be decomposed by

$$\widehat{\tau}^* = \frac{|\mathcal{I}_1|}{n}\widehat{\tau}^{*(\mathcal{I}_1)} + \frac{|\mathcal{I}_2|}{n}\widehat{\tau}^{*(\mathcal{I}_2)}, \tag{12}$$

where

$$\widehat{\tau}^{*(\mathcal{I}_l)} = \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} \left\{ \frac{Z_i Y_i}{e(X_i)} + \left(1 - \frac{Z_i}{e(X_i)}\right) m(X_i) \right\}. \tag{13}$$

We follow the assumptions in Chernozhukov et al. [10] as follows.

Assumption 3 (Sup-norm Consistency). $\sup_x |\widehat{m}(x) - m(x)| \to_p 0, \sup_x |\widehat{e}(x) - e(x)| \to_p 0$ *for any* x *on its support.*

Assumption 4 (Risk Decay). $E\{\widehat{m}(X) - m(X)\}^2 \cdot E\{\widehat{e}(X) - e(X)\}^2 = o(n^{-1})$.

Assumption 3 states that the fitted values of the propensity score and outcome regression models are consistent to the true value at any data point. Assumption 4 states that the convergence rates of models are not too slow. A wide range of nonparametric estimators can satisfy Assumptions 3 and 4. The deviation of $\widehat{\tau}$ is bounded as follows.

Theorem 1. *Let* $\widehat{\tau}$ *be the AIPW estimator of* $\tau = E(Y)$ *by cross-fitting. Under Assumptions 1–4, for any* $0 < \epsilon < 1$,

$$|\widehat{\tau} - \tau| \leq \sqrt{\frac{2(2/\eta - 1)^2 M^2 \log(2/\epsilon)}{n} + o(n^{-1})} \tag{14}$$

with probability larger than $1 - \epsilon$.

Proof. For $l = 1, 2$,

$$\begin{aligned}
\widehat{\tau}^{(\mathcal{I}_l)} - \widehat{\tau}^{*(\mathcal{I}_l)} = {} & \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} Z_i \left\{ \frac{1}{\widehat{e}^{(3-l)}(X_i)} - \frac{1}{e(X_i)} \right\} \{Y_i - m(X_i)\} \\
& - \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} \{\widehat{m}^{(3-l)}(X_i) - m(X_i)\} \left\{ \frac{Z_i}{e(X_i)} - 1 \right\} \\
& - \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} Z_i \left\{ \frac{1}{\widehat{e}^{(3-l)}(X_i)} - \frac{1}{e(X_i)} \right\} \{\widehat{m}^{(3-l)}(X_i) - m(X_i)\}.
\end{aligned}$$

Note that the first two terms are sums of cross products of a mean zero random variable multiplied by an independent $o_p(1)$ random variable (since $\mathcal{I}_l$ can be considered as fixed by conditioning on). The first two terms are $o_p(n^{-1/2})$ random variables. Assumption 4 implies that the third term is also $o_p(n^{-1/2})$ by Cauchy–Schwarz. Therefore, $|\widehat{\tau}^{(\mathcal{I}_l)} - \widehat{\tau}^{*(\mathcal{I}_l)}| = o_p(n^{-1/2})$ and thus $\sqrt{n}|\widehat{\tau} - \widehat{\tau}^*| = o_p(1)$. From the concentration inequality (9), we have

$$P\left(\sqrt{n}|\widehat{\tau} - \tau| \geq t + o(1)\right) \leq 2 \exp\left\{ -\frac{t^2}{2(2/\eta - 1)^2 M^2} \right\}. \tag{15}$$

Let the right hand side be ϵ, so

$$t = \sqrt{2(2/\eta - 1)^2 M^2 \log(2/\epsilon)}. \tag{16}$$

□

The first-order equivalence of $\widehat{\tau}$ and $\widehat{\tau}^*$ is important. If there are high-dimensional covariates, the convergence rates of fitted models usually cannot achieve the rate of $O_p(n^{-1/2})$. In addition, the risk decay assumption allows nonparametric estimation of models, for example, by spline or kernel regression. Provided that the estimated models do not converge

too slowly, the AIPW estimator $\widehat{\tau}$ can enjoy good asymptotic and non-asymptotic properties similar to those of the oracle estimator $\widehat{\tau}^*$.

3.2. Mean Outcome in the Observed Group

We can also study the non-asymptotic bound for $\tau_1 = E(Y \mid Z = 1)$. This estimate can be estimated by

$$\widehat{\tau}_1 = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i}, \tag{17}$$

where no nuisance models are involved. This empirical average estimator $\widehat{\tau}_1$ is unbiased and consistent for τ_1, because

$$\begin{aligned} E(\widehat{\tau}_1) &= E\left\{ \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i} \right\} \\ &= E\left[E\left\{ \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i} \;\middle|\; Z_1, \ldots, Z_n \right\} \right] \\ &= E\left\{ \frac{\sum_{i=1}^n Z_i \cdot E(Y_i \mid Z_i)}{\sum_{i=1}^n Z_i} \right\} \\ &= E\left\{ \frac{\sum_{i=1}^n Z_i \cdot E(Y_i \mid Z_i = 1)}{\sum_{i=1}^n Z_i} \right\} \\ &= \tau_1. \end{aligned}$$

The double robustness of $\widehat{\tau}_1$ is trivial because the propensity score model and the outcome regression model are not involved.

To verify the bounded difference condition, suppose we replace the observation of the kth unit $O_k = (Z_k, Z_k Y_k, X_k)$ with $O_k' = (Z_k', Z_k' Y_k', X_k')$. If $Z_k = Z_k' = 1$,

$$\begin{aligned} |D_1| :&= |\widehat{\tau}_1(O_1, \ldots, O_k, \ldots, O_n) - \widehat{\tau}_1(O_1, \ldots, O_k', \ldots, O_n)| \\ &= \left| \frac{\sum_{i \neq k} Z_i Y_i + Y_k}{\sum_{i \neq k} Z_i + 1} - \frac{\sum_{i \neq k} Z_i Y_i + Y_k'}{\sum_{i \neq k} Z_i + 1} \right| \\ &= \frac{1}{\sum_{i \neq k} Z_i + 1} \cdot |Y_k - Y_k'| \\ &\leq \frac{1}{n\eta} \cdot 2M =: c_1 \end{aligned}$$

with probability larger than

$$\delta_n = P\left(\sum_{i=1}^n Z_i \geq n\eta \right) = \sum_{m=[n\eta]+1}^n C_n^m \eta^m (1-\eta)^{1-m}. \tag{18}$$

By the weak law of large numbers, we know that $\delta_n \to 1$ as $n \to \infty$ because $E(Z_i) > \eta$. If $Z_k = Z_k' = 0$,

$$\begin{aligned} |D_1| :&= |\widehat{\tau}_1(O_1, \ldots, O_k, \ldots, O_n) - \widehat{\tau}_1(O_1, \ldots, O_k', \ldots, O_n)| \\ &= \left| \frac{\sum_{i \neq k} Z_i Y_i}{\sum_{i \neq k} Z_i} - \frac{\sum_{i \neq k} Z_i Y_i}{\sum_{i \neq k} Z_i} \right| = 0. \end{aligned}$$

If $Z_k = 1$ and $Z_k' = 0$, using the equality

$$\frac{c}{d} - \frac{a}{b} = \frac{1}{d}\left\{ (c-a) - \frac{a}{b}(d-b) \right\} \tag{19}$$

for any nonzero $a, b, c, d \in \mathbb{R}$,

$$\begin{aligned}|D_1| :&= |\widehat{\tau}_1(O_1, \ldots, O_k, \ldots, O_n) - \widehat{\tau}_1(O_1, \ldots, O_k', \ldots, O_n)| \\ &= \left| \frac{\sum_{i \neq k} Z_i Y_i + Y_k}{\sum_{i \neq k} Z_i + 1} - \frac{\sum_{i \neq k} Z_i Y_i}{\sum_{i \neq k} Z_i} \right| \\ &= \frac{1}{\sum_{i \neq k} Z_i + 1} \cdot \left| Y_k - \frac{\sum_{i \neq k} Z_i Y_i}{\sum_{i \neq k} Z_i} \right| \\ &\leq \frac{1}{n\eta} \cdot 2M = c_1\end{aligned}$$

with probability larger than δ_n. The result is similar for the case with $Z_k = 0$ and $Z_k' = 1$. In summary, $|D_1| \leq c_1$. According to the McDiarmid's inequality,

$$P\left(\sqrt{n}|\widehat{\tau}_1 - \tau_1| \geq t\right) \leq 2\exp\left\{-\frac{\eta^2 t^2}{2M^2}\right\} + 1 - \delta_n, \quad \forall\, t > 0. \tag{20}$$

Theorem 2. *Let $\widehat{\tau}_1$ be the empirical average estimator of $\tau_1 = E(Y \mid Z = 1)$ in (17). Under Assumption 2, for any $1 - \delta_n < \epsilon < 1$,*

$$|\widehat{\tau}_1 - \tau_1| \leq \sqrt{\frac{2M^2}{n\eta^2} \log\left(\frac{2}{\epsilon + \delta_n - 1}\right)} \tag{21}$$

with probability larger than $1 - \epsilon$.

Proof. Let the right hand side of (20) be ϵ, so

$$t = \sqrt{\frac{2M^2}{\eta^2} \log\left(\frac{2}{\epsilon + \delta_n - 1}\right)}. \tag{22}$$

□

3.3. Mean Outcome in the Unobserved Group

We further assume $e(X) < 1 - \eta$, so that there would be missing units conditioning on X. Otherwise, the estimand $\tau_0 = E(Y \mid Z = 0)$ would be meaningless. To estimate τ_0, we must use the information of the observed units because there are no observations of Y in the $Z = 0$ group. The estimand τ_0 should address a covariate shift from the overall population to the $Z = 0$ group, which imposes a weight $\{1 - e(x)\}/P(Z = 0)$. The inverse probability weighting formula of τ_0 is

$$\tau_0 = E\left\{\frac{ZY}{e(X)} \cdot \frac{1 - e(X)}{P(Z = 0)}\right\}. \tag{23}$$

By appending an augmented term and estimating $P(Z = 0)$ by empirical average, the oracle doubly robust estimator for τ_0 is given by [17]

$$\widehat{\tau}_0^* = \frac{1}{\sum_{i=1}^n (1 - Z_i)} \cdot \sum_{i=1}^n \frac{Z_i Y_i (1 - e(X_i)) - (Z_i - e(X_i)) m(X_i)}{e(X_i)}. \tag{24}$$

It can be shown that $E(\widehat{\tau}_0^*) = \tau_0$ because

$$\begin{aligned}E(\widehat{\tau}_0^*) &= E\left\{\frac{1}{\sum_{i=1}^n (1 - Z_i)} \cdot \sum_{i=1}^n \frac{Z_i Y_i (1 - e(X_i)) - (Z_i - e(X_i)) m(X_i)}{e(X_i)}\right\} \\ &= E\left[E\left\{\frac{1}{\sum_{i=1}^n (1 - Z_i)} \cdot\right.\right.\end{aligned}$$

$$
\begin{aligned}
&\quad \sum_{i=1}^{n} \frac{Z_i Y_i (1-e(X_i)) - (Z_i - e(X_i))m(X_i)}{e(X_i)} \;\Big|\; Z_1, \ldots, Z_n, X_1, \ldots, X_n \Bigg\} \Bigg] \\
&= E\left\{ \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \sum_{i=1}^{n} \frac{Z_i m(X_i)(1-e(X_i)) - (Z_i - e(X_i))m(X_i)}{e(X_i)} \right\} \\
&= E\left\{ \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \sum_{i=1}^{n} (1-Z_i)m(X_i) \right\} \\
&= E\left[E\left\{ \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \sum_{i=1}^{n} (1-Z_i)m(X_i) \;\Big|\; Z_1, \ldots, Z_n \right\} \right] \\
&= E\left[\frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \sum_{i=1}^{n} (1-Z_i) E\{m(X_i) \mid Z_i = 0\} \right] \\
&= E\left\{ \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \sum_{i=1}^{n} (1-Z_i) E(Y_i \mid Z_i = 0) \right\} \\
&= \tau_0.
\end{aligned}
$$

In fact, $E(\widehat{\tau}_0^*) = \tau_0$ if either $e(x)$ or $m(x)$ is correctly specified.

To verify the bounded difference condition, suppose we replace the observation of the kth unit $O_k = (Z_k, Z_k Y_k, X_k)$ with $O_k' = (Z_k', Z_k' Y_k', X_k')$. If $Z_k = Z_k'$,

$$
\begin{aligned}
|D_0^*| :&= |\widehat{\tau}_0(O_1, \ldots, O_k, \ldots, O_n) - \widehat{\tau}_0(O_1, \ldots, O_k', \ldots, O_n)| \\
&= \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \left| \frac{Z_k Y_k (1-e(X_k)) - (Z_k - e(X_k))m(X_k)}{e(X_k)} \right. \\
&\qquad \left. - \frac{Z_k' Y_k' (1-e(X_k')) - (Z_k' - e(X_k'))m(X_k')}{e(X_k')} \right| \\
&\le \frac{2}{n\eta} \cdot \frac{2M}{\eta} =: c_0^*
\end{aligned}
$$

with probability larger than

$$
\delta_n' = P\left(\sum_{i=1}^{n} (1-Z_i) \ge n\eta \right) = \sum_{m=0}^{[n(1-\eta)]} C_n^m \eta^m (1-\eta)^{n-m}. \tag{25}
$$

By the weak law of large numbers, $\delta_n' \to 1$ as $n \to \infty$ because $E(1-Z_i) > \eta$. If $Z_k \neq Z_k'$, without loss of generality, suppose $Z_k = 0$ and $Z_k' = 1$. Use Equation (19),

$$
\begin{aligned}
|D_0^*| :&= |\widehat{\tau}_0^*(O_1, \ldots, O_k, \ldots, O_n) - \widehat{\tau}_0^*(O_1, \ldots, O_k', \ldots, O_n)| \\
&= \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \cdot \left| \frac{Z_k Y_k (1-e(X_k)) - (Z_k - e(X_k))m(X_k)}{e(X_k)} \right. \\
&\qquad \left. - \frac{Z_k' Y_k' (1-e(X_k')) - (Z_k' - e(X_k'))m(X_k')}{e(X_k')} - \widehat{\tau}_0^{*\prime} \right| \\
&\le \frac{1}{n\eta} \cdot \left(\frac{2}{\eta} - 1 + 1 \right) M \le c_0^*
\end{aligned}
$$

with probability larger than δ_n'. By McDiarmid's inequality,

$$
P\left(\sqrt{n} |\widehat{\tau}_0^* - \tau_0| \ge t \right) \le 2 \exp\left\{ -\frac{\eta^4 t^2}{8M^2} \right\} + 1 - \delta_n'. \tag{26}
$$

If the models $e(x)$ and $m(x)$ are unknown, we use cross-fitting to obtain the AIPW estimator

$$\widehat{\tau}_0 = \frac{1}{\sum_{i=1}^n (1-Z_i)} \cdot \sum_{i=1}^n \frac{Z_i Y_i (1-\widehat{e}(X_i)) - (Z_i - \widehat{e}(X_i))\widehat{m}(X_i)}{\widehat{e}(X_i)}. \tag{27}$$

This estimator by cross-fitting can achieve first-order equivalence with the oracle estimator $\widehat{\tau}_0^*$ (which we will prove later). In fact, the estimator (27) can be further expressed as

$$\widehat{\tau}_0 = \frac{|\mathcal{I}_1|}{\sum_{i=1}^n (1-Z_i)} \widehat{\tau}_0^{(\mathcal{I}_1)} + \frac{|\mathcal{I}_2|}{\sum_{i=1}^n (1-Z_i)} \widehat{\tau}_0^{(\mathcal{I}_2)}, \tag{28}$$

with

$$\widehat{\tau}_0^{(\mathcal{I}_l)} = \frac{1}{|\mathcal{I}_l|} \cdot \sum_{i \in \mathcal{I}_l} \frac{Z_i Y_i (1-\widehat{e}^{(3-l)}(X_i)) - (Z_i - \widehat{e}^{(3-l)}(X_i))\widehat{m}^{(3-l)}(X_i)}{\widehat{e}^{(3-l)}(X_i)}, \tag{29}$$

where $\widehat{m}^{(3-l)}(x)$ and $\widehat{e}^{(3-l)}$ are models fitted by the $(3-l)$th half of the sample ($l = 1, 2$). As a comparison, the oracle estimator can be decomposed by

$$\widehat{\tau}_0^* = \frac{|\mathcal{I}_1|}{\sum_{i=1}^n (1-Z_i)} \widehat{\tau}_0^{*(\mathcal{I}_1)} + \frac{|\mathcal{I}_2|}{\sum_{i=1}^n (1-Z_i)} \widehat{\tau}_0^{*(\mathcal{I}_2)}, \tag{30}$$

with

$$\widehat{\tau}_0^{*(\mathcal{I}_l)} = \frac{1}{|\mathcal{I}_l|} \cdot \sum_{i \in \mathcal{I}_l} \frac{Z_i Y_i (1-e(X_i)) - (Z_i - e(X_i))m(X_i)}{e(X_i)}. \tag{31}$$

Theorem 3. *Let $\widehat{\tau}_0$ be the AIPW estimator of $\tau_0 = E(Y \mid Z = 0)$ by cross-fitting given in (27). Under Assumptions 1–4 and $e(x) < 1 - \eta$, for any $1 - \delta_n' < \epsilon < 1$,*

$$|\widehat{\tau}_0 - \tau_0| \leq \sqrt{\frac{8M^2}{n\eta^4} \log\left(\frac{2}{\epsilon + \delta_n' - 1}\right) + o(n^{-1})} \tag{32}$$

with probability larger than $1 - \epsilon$.

Proof. For $l = 1, 2$,

$$\begin{aligned}
\widehat{\tau}_0^{(\mathcal{I}_l)} - \widehat{\tau}_0^{*(\mathcal{I}_l)} = {} & \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} Z_i \left\{ \frac{1}{\widehat{e}^{(3-l)}(X_i)} - \frac{1}{e(X_i)} \right\} \{Y_i - m(X_i)\} \\
& - \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} \{\widehat{m}^{(3-l)}(X_i) - m(X_i)\} \left\{ \frac{Z_i}{e(X_i)} - 1 \right\} \\
& - \frac{1}{|\mathcal{I}_l|} \sum_{i \in \mathcal{I}_l} Z_i \left\{ \frac{1}{\widehat{e}^{(3-l)}(X_i)} - \frac{1}{e(X_i)} \right\} \{\widehat{m}^{(3-l)}(X_i) - m(X_i)\}.
\end{aligned}$$

Note that the first two terms are sums of cross products of a mean zero random variable multiplied by an independent $o_p(1)$ random variable (since $\mathcal{I}_l$ can be considered as fixed by conditioning on), so the first two terms are $o_p(n^{-1/2})$ random variables. Assumption 4 implies that the third term is also $o_p(n^{-1/2})$ by Cauchy–Schwarz. Therefore, $|\widehat{\tau}_0^{(\mathcal{I}_l)} - \widehat{\tau}_0^{*(\mathcal{I}_l)}| = o_p(n^{-1/2})$; Thus, $\sqrt{n}|\widehat{\tau}_0 - \widehat{\tau}_0^*| = o_p(1)$. From (26), we have

$$P\left(\sqrt{n}|\widehat{\tau}_0 - \tau_0| \geq t + o(1)\right) \leq 2\exp\left\{-\frac{\eta^4 t^2}{8M^2}\right\} + 1 - \delta_n'. \tag{33}$$

Let the right-hand-side be ϵ, so

$$t = \sqrt{\frac{8M^2}{\eta^4} \log\left(\frac{2}{\epsilon + \delta_n' - 1}\right)}. \tag{34}$$

□

4. Construction of Error Bounds for SubGaussian Outcomes

Recently the McDiarmid's inequality has been extended to boundless situations [18]. In this section, we assume that Y is sub-Gaussian, i.e.,

$$Ee^{tY} \leq e^{t^2\sigma^2/2}, \quad \forall\, t \in \mathbb{R},$$

for some $\sigma^2 > 0$ [13]. Define the sub-Gaussian norm

$$\|Y\|_G := \sup_{k \geq 1} \left[\frac{E(Y^{2k})}{(2k-1)!!}\right]^{1/2k}$$

for a random variable Y. The random variable Y is sub-Gaussian if $\|Y\|_G < \infty$.

Lemma 2 (Zhang–Lei's inequality [18])**.** *Suppose $O_1, \ldots, O_n$ are independent random variables taking values in the set $\mathcal{O}$, and assume $f : \mathcal{O}^n \to \mathbb{R}$ is a function. Define*

$$D_{f,O_k}(o) = f(o_1, \ldots, o_{k-1}, O_k, o_{k+1}, \ldots, o_n) - E\{f(o_1, \ldots, o_{k-1}, O_k, o_{k+1}, \ldots, o_n)\}.$$

If $\{D_{f,O_k}(o)\}_{i=1}^n$ have finite $\|\cdot\|_G$-norm, then for $\forall\, t > 0$,

$$P(|f(O_1, \ldots, O_n) - E\{f(O_1, \ldots, O_n)\}| \geq t) \leq 2\exp\left\{\frac{-t^2}{16 \sup_{o \in \mathcal{O}^n} \sum_{i=1}^n \|D_{f,O_i}(o)\|_G^2}\right\}.$$

Now we assume that the conditional mean outcome $m(X)$ and residual $Y - m(X)$ are sub-Gaussian, so that Y is also subGaussian. For $\tau = E(Y)$, let $\widehat{\tau}^*$ be the oracle AIPW estimator. Then,

$$D_{\widehat{\tau}^*,O_k}(o) = \frac{1}{n}\left\{\frac{Z_k Y_k}{e(X_k)} + \left(1 - \frac{Z_k}{e(X_k)}\right) m(X_k) - \tau\right\}.$$

To verify that $D_{\widehat{\tau}^*,O_k}(o)$ is sub-Gaussian, it suffices to prove that the $2k$-th moment norm of $D_{\widehat{\tau}^*,O_k}(o)$ is finite for every $k \geq 1$. By the triangle inequality,

$$\begin{aligned}
\|D_{\widehat{\tau}^*,O_k}(o)\|_{2k} &= \left\|\frac{1}{n}\left\{\frac{Z_k Y_k}{e(X_k)} + \left(1 - \frac{Z_k}{e(X_k)}\right) m(X_k) - \tau\right\}\right\|_{2k} \\
&\leq \left\|\frac{Z_k\{Y_k - m(X_k)\}}{ne(X_k)}\right\|_{2k} + \left\|\frac{m(X_k) - \tau}{n}\right\|_{2k} \\
&\leq \left\|\frac{Y_k - m(X_k)}{n\eta}\right\|_{2k} + \left\|\frac{m(X_k) - \tau}{n}\right\|_{2k} \\
&< \infty,
\end{aligned}$$

where $\|W\|_{2k} := [\mathrm{E}|W|^{2k}]^{1/2k}$ for a random variable W.

Let $\nu = \|m(X)\|_G$ and $\sigma = \|Y - m(X)\|_G$, so $\|D_{\widehat{\tau}^*,O_k}(o)\|_G \leq (\sigma/\eta + \nu)/n$. Therefore,

$$P\left(\sqrt{n}|\widehat{\tau}^* - \tau| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{16(\sigma/\eta + \nu)^2}\right\}. \tag{35}$$

Considering that the AIPW estimator $\widehat{\tau}$ is first-order equivalent to the oracle version $\widehat{\tau}^*$, we have the following error bound for $\widehat{\tau}$.

Theorem 4. *Let $\widehat{\tau}$ be the AIPW estimator of $\tau = E(Y)$ by cross-fitting. Under Assumptions 1–4, for any $0 < \epsilon < 1$,*

$$|\widehat{\tau} - \tau| \leq \sqrt{\frac{16(\sigma/\eta + \nu)^2 \log(2/\epsilon)}{n} + o(n^{-1})} \tag{36}$$

with probability larger than $1 - \epsilon$.

The proof is similar to the proof of Theorem 1. Since $\widehat{\tau}$ and $\widehat{\tau}^*$ are first-order equivalent, we can replace the left-hand-side of Inequality (35) with $P(\sqrt{n}|\widehat{\tau} - \tau| \geq t + o(1))$, and then solve this inequality. Furthermore, we can apply Lemma 2 to study the non-asymptotic error bounds for $\widehat{\tau}_1$ and $\widehat{\tau}_0$, the (doubly robust) estimators for $\tau_1 = E(Y \mid Z = 1)$ and $\tau_0 = E(Y \mid Z = 0)$. Given that Y is sub-Gaussian, $\|D_{\widehat{\tau}_z, O_k}(o)\|_G$ is a weighted average of sub-Gaussian random variables where the weights are bounded ($z = 1, 0$). Whenever the event with high probability that $\sum_{i=1}^{n} Z_i$ is bounded away from 0 or 1 occurs, the error bounds of $\widehat{\tau}_z$ would then be obtained.

5. Simulation

Suppose there are two independent covariates, X_1 and X_2, both following a uniform distribution on $[-1, 1]$. Denote $X = (1, X_1, X_2)$. The propensity score is $e(x) = expit(1 - 0.5X_1 + 0.5X_2)$, where $expit(x) = 1/\{1 + \exp(-x)\}$. Since X is bounded, the propensity score is also bounded away from zero and one. Next, we consider two sorts of outcomes. The first is the binary case: $P(Y = 1 \mid X) = expit(X_1 + X_2)$. The second is the continuous case: Y follows a normal distribution with mean $X_1 + X_2$ and standard deviation of value one.

We set the sample size n from 100 to 1000 with step 100. Under each sample size, we generate data 1000 times and calculate the average width of error bounds. Figure 1 displays the widths of error bounds in Theorem 1 (based on McDiarmid's inequality) and Theorem 4 (based on Zhang–Lei's inequality) respectively. We set the parameters in the formulas of error bounds at their true values.

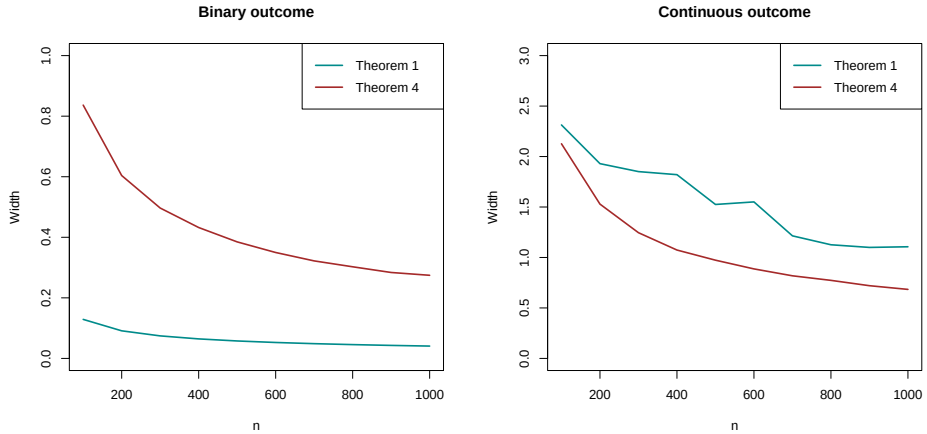

Figure 1. Width of the error bounds with the increasing of sample sizes (using true parameters).

In practice, the parameters η, σ and ν are unknown, so we need to estimate them. We fit a logistic regression of Z on X as the propensity score, denoted by $\widehat{e}(x)$. We can estimate η by $\widehat{\eta} = \min_i\{\widehat{e}(X_i)\}$. An outcome regression is fitted by a logistic regression for binary outcomes and linear regression for continuous outcomes, denoted by $\widehat{m}(x)$. If $Y \in \{0, 1\}$ is binary, we can transform Y to $Y - 0.5$, so that $M = 0.5$. If Y is continuous, the bound

of Y can be estimated by $\widehat{M} = \max_i\{Y_i - \sum_{j=1}^n Y_j/n\}$. The sub-Gaussian norms σ and ν are estimated based on the empirical $2k$-th moments of $Y_i - \widehat{m}(X_i)$ and $\widehat{m}(X_i)$, respectively, denoted by $\widehat{\sigma}$ and $\widehat{\nu}$. Then, we can obtain two error bounds according to Theorem 1 and Theorem 4, respectively.

Figure 2 displays the widths of error bounds in Theorem 1 and Theorem 4. When the sample size n is small, the empirical error bounds might be slightly shorter than the oracle versions. In the binary case, the error bound based on Theorem 1 (McDiarmid's inequality) performs better. In the continuous case, the error bound based on Theorem 4 (Zhang–Lei's inequality) performs better. In fact, if the outcome is binary, it is naturally bounded. We do not need information about higher-order moments. If the outcome is continuous, the estimation of the bound M would be unstably affected by extreme values of outcomes, so the error bound by McDiarmid's inequality could be unstable and too wide. On the contrary, Zhang–Lei's inequality just requires finite moments of the outcomes without putting restrictions on the maximum value, so Zhang–Lei's inequality is more appropriate for boundless outcomes.

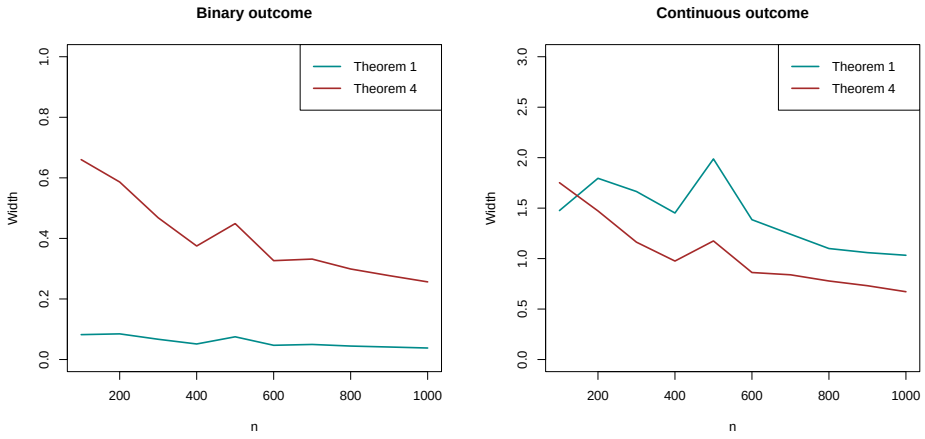

Figure 2. Width of the error bounds with the increasing of sample sizes (using estimated parameters).

6. Discussion: Relation to Causal Inference

Compared with asymptotic arguments, the non-asymptotic inference provides more accurate error bounds when the sample size is not large enough. However, the non-asymptotic error bounds may be conservative. It is of great interest to investigate the performance of different types of non-asymptotic error bounds. In the simulation study, we find that one type of error bound may outperform another in certain scenarios. Massive efforts have been devoted to shortening the error bounds with weaker conditions on the distribution (e.g., moments) of outcome variables [13,19–22]. In practice, we can construct several error bounds as long as the conditions to derive concentration inequalities are satisfied and choose the shortest one.

This work is closely related to causal inference. As a missing data problem, causal inference aims to make statistical inferences on the difference of the average potential outcomes under the treated and under the control [23]. Let $Z \in \{1, 0\}$ be the treatment indicator and $Y(z)$ be the potential outcome associated with the treatment $z \in \{1, 0\}$. For each unit, only one of these two potential outcomes, either $Y(1)$ or $Y(0)$, is observable, while the other is missing. To identify the average causal effect, we usually assume causal consistency $Y(Z) = Y$ and unconfoundedness $(Y(1), Y(0)) \perp\!\!\!\perp Z \mid X$. Unconfoundedess, similar to missingness at random, says that the treatment assignment Z can only rely on observed baseline covariates X, rather than potential outcomes $(Y(1), Y(0))$, which could be missing by design [2].

We take the inference on $Y(1)$ as an example. $E\{Y(1)\}$ is the population mean, $E\{Y(1) \mid Z = 1\}$ is the population mean in the observed group (where $Y(1)$ is observable) by treating $Z = 1$ as observed and $Z = 0$ as missing, and $E\{Y(1) \mid Z = 0\}$ is the population

mean in the unobserved group (where $Y(1)$ is unobservable) by treating $Z = 0$ as observed and $Z = 1$ as missing. It is also straightforward to infer the average causal effect $E\{Y(1) - Y(0)\}$, since the AIPW estimator just becomes a combination of two parts, corresponding to $E\{Y(1)\}$ and $E\{Y(0)\}$, respectively. Under causal consistency, uncounfoundedness, positivity, boundedness (sub-Gaussian), sup-norm consistency and risk decay, the non-asymptotic bounds of the tail probability of AIPW estimators can be similarly established by McDiarmid's inequality or Zhang–Lei's inequality.

Other estimands that may be of interest are the average causal effect on the treated (ATT) $E\{Y(1) - Y(0) \mid Z = 1\}$ and the average causal effect on the control (ATC) $E\{Y(1) - Y(0) \mid Z = 0\}$. Take the ATT as an example. It can be decomposed as $E\{Y(1) \mid Z = 1\}$ and $E\{Y(0) \mid Z = 1\}$. The former corresponds to the mean outcome in the observed group, as if $Y(1)$ is observed in the $Z = 1$ group, and the latter corresponds to the mean outcome in the unobserved group, as if $Y(0)$ is unobserved in the $Z = 1$ group. By examining the expression of the estimator of ATT given in [17], the bounded difference condition can hold with a high probability, so the non-asymptotic law can be established.

Author Contributions: Conceptualization, F.W.; Methodology, Y.D.; Investigation, F.W. and Y.D.; Writing—original draft, F.W. and Y.D.; Writing—review & editing, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (Grant No. 2021YFF0901400), National Natural Science Foundation of China (Grant No. 12026606) and Fundamental Research Funds for the Central Universities (Grant No. 2022QNPY73). This work is also partly supported by Novo Nordisk A/S.

Data Availability Statement: Data availability is not applicable to this article as no new data were created or analysed in this study.

Acknowledgments: We thank the editor of this special issue "New Advances in High-Dimensional and Non-asymptotic Statistics" and reviewers for their kind comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
2. Mealli, F.; Rubin, D.B. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **2015**, *102*, 995–1000. [CrossRef]
3. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
4. Hirano, K.; Imbens, G.W.; Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **2003**, *71*, 1161–1189. [CrossRef]
5. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* **1995**, *90*, 106–121. [CrossRef]
6. Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **1998**, *66*, 315–331. [CrossRef]
7. Tsiatis, A.A. *Semiparametric Theory and Missing Data*; Springer: Berlin/Heidelberg, Germany, 2006.
8. Bang, H.; Robins, J.M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **2005**, *61*, 962–973. [CrossRef]
9. Farrell, M.H. Robust inference on average treatment effects with possibly more covariates than observations. *J. Econom.* **2015**, *189*, 1–23. [CrossRef]
10. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *Econom. J.* **2018**, *21*, C1–C68. [CrossRef]
11. Newey, W.K.; Robins, J.R. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv* **2018**, arXiv:1801.09138.
12. Kennedy, E.H. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv* **2020**, arXiv:2004.14497.
13. Zhang, H.; Chen, S.X. Concentration inequalities for statistical inference. *Commun. Math. Res.* **2020**, *37*, 1–85.
14. Zhang, H.; Wei, H. Sharper sub-weibull concentrations. *Mathematics* **2022**, *10*, 2252. [CrossRef]
15. Horvitz, D.G.; Thompson, D.J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **1952**, *47*, 663–685. [CrossRef]
16. McDiarmid, C. On the method of bounded differences. *Surv. Comb.* **1989**, *141*, 148–188.

17. Mercatanti, A.; Li, F. Do debit cards increase household spending? evidence from a semiparametric causal analysis of a survey. *Ann. Appl. Stat.* **2014**, *8*, 2485–2508. [CrossRef]
18. Zhang, H.; Lei, X. Non-asymptotic optimal prediction error for growing-dimensional partially functional linear models. *arXiv* **2022**, arXiv:2009.04729.
19. Hsu, D.J.; Kakade, S.M.; Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **2012**, *17*, 1–6. [CrossRef]
20. Bennett, G. Probability inequalities for the sum of independent random variables. *J. Am. Stat. Assoc.* **1962**, *57*, 33–45. [CrossRef]
21. Bernstein, S. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukr. Sect. Math* **1924**, *1*, 38–49.
22. Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.
23. Ding, P.; Li, F. Causal inference. *Stat. Sci.* **2018**, *33*, 214–237. [CrossRef]