Nonparametric Estimation of the Average Causal Effect using the Front-door Criterion

Ana P. L. Cavalcante[1] **Rafael Izbicki**[2] **Rafael B. Stern**[1]

[1]Department of Statistics, University of São Paulo, São Paulo, BRA
[2]Department of Statistics, Federal University of São Carlos, São Paulo, BRA

Abstract

Existing estimators for the average causal effect that are based on the front-door criterion are largely parametric or semiparametric, relying on restrictive assumptions that may not hold in practice. To address this gap, we propose three nonparametric methods, `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S`, the latter of which is applicable to continuous treatments. Specifically, when unobservable confounders prevent the identification of the causal effect of X on Y through the backdoor criterion, the front-door criterion allows identification through mediators, $\mathbb{W}$. Our approach estimates either conditional densities, conditional odds ratios, or conditional density ratios, using these quantities to re-weight instances and identify the causal effect. We establish that under mild nonparametric assumptions, all proposed estimators converge to the true causal effect. In the binary treatment case, simulation studies reveal that `Frontdoor-CDE` and `Frontdoor-Odds` outperform parametric methods when model assumptions are violated, while remaining competitive when assumptions hold. For continuous treatments, `Frontdoor-S` demonstrates consistency and the ability to capture complex data structures.

1 INTRODUCTION

1.1 RELATED WORK

Causal inference has seen significant advancements in both theory and applications over the past few decades. The increasing need to quantify and understand causal relationships has emerged in various disciplines, including statistics, machine learning, epidemiology, and economics.

Building on the seminal works of scholars such as Judea Pearl [Pearl, 2009], Donald Rubin [Imbens and Rubin, 2015], and others, causal inference provides solid methodologies for estimating various causal quantities. Among these, the Average Causal Effect (ACE), also referred to as the Average Treatment Effect (ATE), is of central interest.

In the context of structured causal models (SCMs) [Glymour et al., 2016], when X is a binary, the ACE of X on Y is

$$\text{ACE}_{X,Y} = \mathbb{E}[Y \mid \text{do}(X = 1)] - \mathbb{E}[Y \mid \text{do}(X = 0)], \quad (1)$$

where $\mathbb{E}[Y \mid \text{do}(X = x)]$ is the average value of Y when X is set to x through an intervention.

In order to estimate $\text{ACE}_{X,Y}$, a common strategy is to control for the effect of confounders, that is, variables that are common causes of both X and Y. If $\mathbf{Z}$ satisfies the backdoor criterion [Pearl, 2009], then

$$\mathbb{E}[Y \mid \text{do}(X = x)] = \mathbb{E}\left[\frac{Y \cdot \mathbb{I}(X = x)}{f(x|\mathbf{Z})}\right]. \quad (2)$$

Equation 2 motivates inverse probability weighting estimators (IPW) [Rosenbaum and Rubin, 1983]. Specifically, consider that $\hat{f}(x|\mathbf{Z})$ is an estimator for $f(x|\mathbf{Z})$. IPW estimates $\mathbb{E}[Y \mid \text{do}(X = x)]$ by starting from Equation 2, plugging in $\hat{f}$ in the place of f and, finally, substituting the expectation by an empirical average, that is,

$$\hat{\mathbb{E}}_{IPW}[Y \mid \text{do}(X = x)] = \sum_{i=1}^{n} \frac{Y_i \cdot \mathbb{I}(X_i = x)}{n\hat{f}(x|\mathbf{Z}_i)}. \quad (3)$$

However, due to unobserved variables, it might not be possible to control for the effect of confounders. In such a case, it is not possible to use common estimators such as IPW, the adjustment formula [Greenland and Robins, 1986], and matching [Cochran and Rubin, 1973]. An alternative approach consists of controlling the effect of mediator variables, that are affected by X and affect Y.

The front-door criterion [Pearl, 1995] ensures causal identification based on mediator variables, even in the presence of unobserved confounders:

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

Definition 1.1. A set of variables W satisfies the front-door criterion for estimating the causal effect of X on Y if

1. W intercepts all directed paths from X to Y,
2. There is no back-door path from X to W, and
3. All back-door paths from W to Y are blocked by X.

Theorem 1.1. [**Front-door identification**] If W satisfies the front-door criterion, Definition 1.1, for estimating the causal effect of X on Y, then

$$\begin{aligned}\mathbb{E}[Y \mid \text{do}(X = x)] &= \mathbb{E}\left[Y \cdot \frac{f(W \mid x)}{f(W \mid X)}\right] \\ &= \mathbb{E}\left[Y \cdot \frac{f(x \mid W)f(X)}{f(X \mid W)f(x)}\right].\end{aligned}$$

Variations of Theorem 1.1 have been used for causal estimation. For instance, Bellemare et al. [2024] proposes an estimator that assumes a linear model. Also, Fulcher et al. [2020] proposes both parametric and semiparametric estimators, including a doubly robust semiparametric locally efficient estimator. A modified version of the latter is proposed by Gupta et al. [2021]. Another extension, incorporating nonparametric methods, is introduced by Guo et al. [2023]. Also, within the nonparametric framework, Chernozhukov et al. [2022] proposes to circumvent the need for explicit estimation of conditional densities, and Singh et al. [2024] addresses continuous treatments.

Despite these recent advances, to the best of our knowledge, the number of nonparametric estimators based on the front-door criterion remains limited, especially when compared to the broad array of methods developed under the more widely studied back-door criterion.

1.2 NOVELTY

We propose three new nonparametric estimators for the average causal effect based on the front-door criterion. The first two, `Frontdoor-CDE` and `Frontdoor-Odds` estimate, respectively, the conditional density and the conditional odds ratio of X given W, using these quantities to re-weight instances according to Theorem 1.1 and estimate the causal effect. The third method, `Frontdoor-S`, estimates the interventional expectation $\mathbb{E}[Y \mid \text{do}(X = x)]$ in settings with a continuous treatment, X.

We establish that under mild nonparametric assumptions, `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S` converge to the true causal effect. For the binary treatment setting, simulation studies demonstrate that `Frontdoor-CDE` and `Frontdoor-Odds` outperform parametric estimators when their assumptions are violated, while remaining competitive when the assumptions hold. In the continuous treatment setting, simulation results demonstrate that `Frontdoor-S` is consistent and effective in capturing complex relationships in the data.

The remainder of this paper is organized as follows. Section 2 describes the implementation and intuition behind `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S`. Section 3 proves that all methods are consistent under mild nonparametric assumptions. Section 4 presents simulation studies comparing `Frontdoor-CDE` and `Frontdoor-Odds` to other parametric and semiparametric approaches, both when parametric assumptions are satisfied and when they are not, focusing on binary treatments. Additionally, Section 4 presents simulation results for `Frontdoor-S` in the continuous treatment setting.

2 METHODOLOGY

In the following, consider that W is an observable set of variables that satisfies the front-door criterion for the causal effect of X on Y. In particular, Theorem 1.1 holds and provides causal identification for $\mathbb{E}[Y \mid \text{do}(X = x)]$. Next, we use this identification and introduce three new estimators for $\mathbb{E}[Y \mid \text{do}(X = x)]$.

2.1 CONDITIONAL DENSITY APPROACH

The IPW estimator in Equation 3 can be obtained from Equation 2 following two steps. First, let $\hat{f}(x|\mathbf{Z})$ be a conditional density estimator (CDE) for $f(x|\mathbf{Z})$ and approximate $\mathbb{E}\left[\frac{Y \cdot \mathbb{I}(X=x)}{f(x|\mathbf{Z})}\right]$ by $\mathbb{E}\left[\frac{Y \cdot \mathbb{I}(X=x)}{\hat{f}(x|\mathbf{Z})}\right]$. Second, approximate the latter expectation by an empirical average, thus obtaining the IPW estimator, $\sum_{i=1}^{n} \frac{Y_i \mathbb{I}(X_i=x)}{n\hat{f}(x|\mathbf{Z}_i)}$.

Using Theorem 1.1, one obtains the estimator

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \cdot \frac{\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)}. \tag{4}$$

However, since Equation 4 involves a ratio of random variables, it can have a high variance. In order to reduce this variance, it is useful to note that

$$\mathbb{E}\left[\frac{f(W|x)}{f(W|X)}\right] = 1. \tag{5}$$

Nonetheless, $\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}(W_i|x)}{\hat{f}(W_i|X_i)}$ can often be distant from 1. Hence, a way to control the variance in Equation 4 is to normalize the weights $\frac{\hat{f}(W_i|x)}{\hat{f}(W_i|X_i)}$ so that they sum to 1. Drawing inspiration from the Hájek estimator [Dorfman

and Valliant, 1997], we propose the `Frontdoor-CDE`:

$$\widehat{\mathbb{E}}_f[Y \mid \text{do}(X = x)] := \left(\sum_{i=1}^{n} \frac{\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} \right)^{-1} \times \sum_{i=1}^{n} Y_i \cdot \frac{\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)}. \tag{6}$$

Any conditional density estimator can be used within Equations 4 and 6. Several methods are available in the literature, including kernel-based approaches [Hyndman et al., 1996, Ichimura and Fukuda, 2010], mixture density networks [Bishop, 1994], normalizing flows [Papamakarios et al., 2021], and quantile-based methods [Takeuchi et al., 2009, Dey et al., 2024]. Each estimator has strengths suited to different applications [Dalmasso et al., 2020]. We use the Flexible Conditional Density Estimator (FlexCoDE) [Izbicki and Lee, 2017], a nonparametric method that converts regression models into conditional density estimators. Its adaptability allows for choosing regression techniques that best fit the data structure.

2.2 ODDS APPROACH

Another possible source for the variance of the estimator in Equation 6 is that it involves a ratio of random variables, $\frac{\hat{f}(W_i|x)}{\hat{f}(W_i|X_i)}$, which is sensible when the denominator is close to 0. `Frontdoor-Odds` avoids this issue by focusing on the second line of Theorem 1.1 when X is a binary variable. In this case, $\frac{f(x|w)}{f(x'|w)}$ can be written as:

$$O_1(w, x', x) := \frac{f(x \mid w)}{f(x' \mid w)} = \begin{cases} 1 & \text{, if } x' = x, \\ \frac{\mathbb{P}(X=x|w)}{1-\mathbb{P}(X=x|w)} & \text{, otherwise.} \end{cases}$$

That is, whenever $O_1(w, x', x)$ is not 1, it is a conditional odds ratio. A similar result applies to $\frac{f(x')}{f(x)}$:

$$O_2(x', x) := \frac{f(x')}{f(x)} = \begin{cases} 1 & \text{, if } x' = x, \\ \frac{1-\mathbb{P}(X=x)}{\mathbb{P}(X=x)} & \text{, otherwise.} \end{cases}.$$

Using the above notation, it follows from Theorem 1.1 that

$$\mathbb{E}[Y \mid \text{do}(X = x)] = \mathbb{E}\left[Y \cdot O_1(W, X, x) \cdot O_2(X, x)\right] \tag{7}$$

We propose a similar plugin estimator as the one in Equation 6. Letting $\widehat{O}_1$ and $\widehat{O}_2$ be estimators for O_1 and O_2, the `Frontdoor-Odds` estimator is defined as:

$$\widehat{\mathbb{E}}_o[Y \mid \text{do}(X = x)] := \left(\sum_{i=1}^{n} \hat{O}(W_i, X_i, x) \right)^{-1} \times \sum_{i=1}^{n} Y_i \cdot \hat{O}(W_i, X_i, x). \tag{8}$$

Any conditional odds ratio estimator can be used. In this work, we use the method described in Dalmasso et al. [2021]. For convenience, let $O_1(w) := O_1(w, 0, 1)$. Since

$$\mathbb{P}(X = 1 \mid W) = \frac{O_1(W)}{1 + O_1(W)},$$

the cross-entropy loss is

$$L(O_1) = \sum_{i=1}^{n} \left(-X_i \cdot O_1(W_i) + \log(1 + O_1(W_i)) \right). \tag{9}$$

That is, $O_1(w)$ can be estimated using an arbitrary neural network that has (X, W) as input, outputs $O_1(W)$, and minimizes $L(O_1)$. More details about the neural network that was tested can be found in Section 4.

2.3 CONTINUOUS TREATMENT

When X is a continuous random variable, using the previous approaches is challenging. To overcome this barrier, we propose `Frontdoor-S`. Define $\lambda(w, x', x) := \frac{f(w|x)}{f(w|x')}$. Letting $\widehat{\lambda}$ be an estimator of λ, the `Frontdoor-S` is:

$$\widehat{\mathbb{E}}_s[Y \mid \text{do}(X = x)] := \left(\sum_{i=1}^{n} \widehat{\lambda}(W_i, X_i, x) \right)^{-1} \times \sum_{i=1}^{n} Y_i \cdot \widehat{\lambda}(W_i, X_i, x). \tag{10}$$

We propose an estimator for λ based on data augmentation. Let $\{(W_i, X_i)\}_{i=1}^n$ be the original dataset. Consider the augmented datasets,

$$\mathcal{D}_1 = \{(W_i, X_i, X_i', S_i = 1)\}_{i=1}^n,$$
$$\mathcal{D}_2 = \{(W_i, X_i', X_i, S_i = 0)\}_{i=1}^n,$$

where $X' = (X_1', \ldots, X_n')$ is a random permutation of X. The full augmented dataset is $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, and $(\tilde{W}, \tilde{X}, \tilde{X}', \tilde{S})$ is a randomly sampled observation from $\mathcal{D}$.

$$(\tilde{X}, \tilde{X}') = \begin{cases} (X_i, X_i'), & \text{if } \tilde{S} = 1, \\ (X_i', X_i), & \text{if } \tilde{S} = 0, \end{cases}$$

for some $i \in \{1, \ldots, n\}$. Hence,

$$\frac{\mathbb{P}(\tilde{S} = 1|\tilde{w}, \tilde{x}, \tilde{x}')}{\mathbb{P}(\tilde{S} = 0|\tilde{w}, \tilde{x}, \tilde{x}')} = \frac{f(\tilde{w}|\tilde{x}, \tilde{x}', \tilde{S} = 1)\mathbb{P}(\tilde{S} = 1|\tilde{x}, \tilde{x}')}{f(\tilde{w}|\tilde{x}, \tilde{x}', \tilde{S} = 0)\mathbb{P}(\tilde{S} = 0|\tilde{x}, \tilde{x}')} = \lambda(\tilde{w}, \tilde{x}', \tilde{x}).$$

We propose estimating λ through an estimator for the conditional odds of $\tilde{S}$ in the full augmented dataset. Here, we use the method described in Dalmasso et al. [2021] and in the end of Subsection 2.2.

Next, we show that under mild nonparametric conditions `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S` converge to $\mathbb{E}[Y \mid \text{do}(X = x)]$.

3 THEORETICAL RESULTS

Let $\widehat{f}$, $\widehat{O}_1$, $\widehat{O}_2$ and $\widehat{\lambda}$ be estimators for f, O_1, O_2 and λ respectively. Subsection 3.1 discusses the assumptions that are required for obtaining consistency of the proposed estimators. Subsection 3.2 presents the convergence results.

3.1 ASSUMPTIONS

As a starting assumption, we consider that the data are i.i.d.

Assumption 1 (i.i.d. data)**.** $(X_i, W_i, Y_i)_{i=1}^n$ are i.i.d.

Also, all estimators use an empirical weighted average of Y_i values as an estimate for a weighted expectation of Y. The convergence of such an estimate relies on the law of large numbers. Such a uniform convergence is obtained by assuming that the conditional expectation of Y is uniformly bounded:

Assumption 2 (Bounded conditional expectation for Y)**.** There exists $M > 0$ such that

$$\sup_{x,w} \mathbb{E}[|Y| \mid X = x, W = w] < M.$$

Besides Assumption 2, `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S` require additional properties of the conditional density and odds ratio estimators. For simplicity, similar assumptions are presented side-by-side for each type of estimator.

Since instances are i.i.d. (Assumption 1), the order in which they are obtained brings no information about the ACE. Hence, this order should not be used by the estimator. That is, no matter the order in which the instances are inserted, the estimator should be the same. This condition is formalized in Assumption 3.

Assumption 3 (Invariance to permutation of instances)**.**

(a) $\widehat{f}$ is invariant to permutations of instances.

(b) $\widehat{O}_1$ and $\widehat{O}_2$ are invariant to permutations of instances.

(c) $\widehat{\lambda}$ is invariant to permutations of instances.

We also require for $\widehat{f}$, $\widehat{O}_1$, $\widehat{O}_2$ and $\widehat{\lambda}$ to converge their respective target functions. The type of convergence that is required varies slightly according to the method:

Assumption 4 (Convergence of estimator)**.**

(a) $\mathbb{E}[(\widehat{f}(W|x) - f(W|x))^2] = o(1)$ and also $\mathbb{E}[(\widehat{f}(W|X) - f(W|X))^2] = o(1)$.

(b) $\mathbb{E}[(\widehat{O}_1(W, X, x) - O_1(W, X, x))^2] = o(1)$ and also $\mathbb{E}[(\widehat{O}_2(X, x) - O_2(X, x))^2] = o(1)$.

(c) $\mathbb{E}[|\widehat{\lambda}(W, X, x) - \lambda(W, X, x)|] = o(1)$.

That is, we assume that $\widehat{f}$, $\widehat{O}_1$ and $\widehat{O}_2$ converge to their respective target functions in quadratic mean, whereas $\widehat{\lambda}$ converges to its target functions in absolute mean.

In addition to convergence, we also require some of the estimators and target functions to be smooth functions. Specifically, we assume that they have finite second moment:

Assumption 5 (Smoothness of estimators)**.**

(a) There exists $K > 0$ such that

$$\max\left(\mathbb{E}\left[\hat{f}^2(W|x)\right], \mathbb{E}\left[\hat{f}^2(W|X)\right]\right) < K,$$

(b) There exists $K > 0$ such that

$$\max\left(\mathbb{E}\left[\hat{O}_1^2(W, X, x)\right], \mathbb{E}\left[O_2^2(X, x)\right]\right) < K.$$

Note that `Frontdoor-S` does not require an assumption analogous to Assumption 5.

Finally, since the weights in `Frontdoor-CDE` are a ratio of random variables, it might be unstable if the denominator is close to 0. In order to avoid this behavior, we require the denominator to be uniformly far from 0:

Assumption 6 (Denominator uniformly far from 0)**.** There exists $\delta > 0$ such that

$$\inf_{x,w} \min\{\widehat{f}(w \mid x), f(w \mid x)\} > \delta.$$

Since both `Frontdoor-Odds` and `Frontdoor-S` do not use a ratio estimator, they do not require an assumption analogous to Assumption 6.

3.2 MAIN RESULT

Using the above assumptions, we establish the consistency of `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S`:

Theorem 3.1. Let W satisfy the frontdoor criterion for the causal effect of X on Y. Also, let Assumption 1 and Assumption 2 be valid:

(a) If the part (a) of Assumption 3, Assumption 4, and Assumption 5 hold and also Assumption 6 holds, then

$$\widehat{\mathbb{E}}_f[Y \mid \text{do}(X = x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)].$$

(b) If the part (b) of Assumption 3, Assumption 4, and Assumption 5 hold, then

$$\widehat{\mathbb{E}}_o[Y \mid \text{do}(X = x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)]$$

(c) If the part (c) of Assumption 3 and Assumption 4 hold, then

$$\widehat{\mathbb{E}}_s[Y \mid \text{do}(X = x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)]$$

The proof of Theorem 3.1 can be found in Appendix B.

Next, we complement the theoretical properties of the proposed methods with their empirical performance in simulated data.

4 SIMULATION STUDY

4.1 BINARY TREATMENT

We use simulations to compare the performance of `Frontdoor-CDE` and `Frontdoor-Odds` to previously proposed estimators. In all of the datasets, we generate variables, W, that satisfy the front-door criterion. For each dataset, we estimate the ACE with the following estimators:

1. `Frontdoor-CDE`(Equation 6): FlexCoDE [Izbicki and Lee, 2017] together with Random Forest regression [Breiman, 2001] was used for estimating $\widehat{f}(w|x)$. The Fourier series truncation size was set to 50.
2. `Frontdoor-Odds`(Equation 8): O_2 was estimated using the empirical odds ratio. O_1 was estimated using a neural network with input (X, W) and the loss in Equation 9. We employed a fully connected neural network with three layers. The first layer consisted of 64 units, followed by a second layer with 32 units, and a final output layer with a single unit. A ReLU activation function was applied after each of the hidden layers, and a dropout layer with a probability of 0.5 was used after the first hidden layer to prevent overfitting. For optimization, the model was trained for a maximum of 20 epochs using the Adam optimizer, with an initial learning rate controlled by a one-cycle learning rate scheduler, reaching a maximum of 0.1. Early stopping was applied with a patience of 3 epochs.
3. `Logistic`: A parametric version of the estimator in Equation 8. The conditional odds ratio is estimated plugging in $\hat{P}(X = 1|W)$, obtained from the maximum likelihood estimate in a logistic regression.
4. `Linear`: The ordinary least squares estimate obtained assuming a linear model. Let $\mathbb{E}[W_i|X] = \alpha_i + \beta_i X$ and $\mathbb{E}[Y|W_i, X] = \mu_i + \delta_i X + \gamma_i W_i$. Let $\widehat{\beta}_i$ and $\widehat{\gamma}_i$ be the ordinary least square estimators obtained from linear regression. The estimator for the ACE is $\sum_{i=1}^{d} \beta_i \cdot \gamma_i$, where d is dimension of W.

Next, we describe the data generating procedures that were employed for comparing the above estimators.

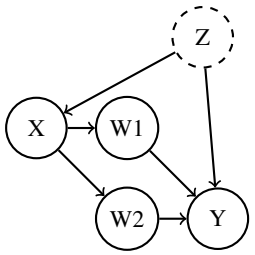


Figure 1: Causal graph that was used in all simulations with binary treatment. We aim to establish the causal effect of X on Y. Z is an unobserved confounder. W_1 and W_2 are mediators which, together, satisfy the front-door criterion.

4.1.1 Data-generating procedure

We compare the previous methods in various simulation scenarios. In all of the scenarios, the relevant random variables were generated according to the directed acyclic graph (DAG) in Figure 1. All variables were treated as observable, except for the sole confounder, Z. As a result, it is not possible to estimate the causal effect by controlling for confounders [Pearl, 1995]. As an alternative, (W_1, W_2) satisfy the front-door criterion. Hence, it is possible to estimate the causal effect using `Frontdoor-CDE`, `Frontdoor-Odds`, `Logistic`, or `Linear`.

While the validity of `Logistic` and `Linear` relies on parametric assumptions, the validity of `Frontdoor-CDE` and `Frontdoor-Odds` is nonparametric. Hence, we compare these methods when the parametric assumptions are satisfied and also when they are not.

None of these assumptions depend on the form in which Z and X are generated. Hence, in all scenarios, they were generated in the same way:

$$Z \sim \mathcal{N}(0, 1), \text{ and}$$
$$X \mid Z \sim \begin{cases} \text{Ber}(0.8), & \text{if } Z \geq 0, \\ \text{Ber}(0.2), & \text{if } Z < 0. \end{cases}$$

Also, `Logistic` requires that X follows a logistic regression given $\mathbb{W} = (W_1, W_2)$. This condition is satisfied by generating $\mathbb{W}$ from a normal distribution with a mean that depends on X. This condition is not satisfied when, for instance, conditional on X, $\mathbb{W}$ follows a mixture of normals. Hence, we consider two methods for generating $\mathbb{W}$:

1. **Normal distribution**

$$W_1 \mid X \sim \begin{cases} \mathcal{N}(0, 1) & \text{, when } X = 0, \\ \mathcal{N}(0, 1) & \text{, when } X = 1. \end{cases}$$

$$W_2 \mid X \sim \begin{cases} \mathcal{N}(0, 1) & \text{, when } X = 0, \\ \mathcal{N}(10, 1) & \text{, when } X = 1, \text{ and} \end{cases}$$

2. **Normal-mixture distribution**

$$W_i \mid X \sim \begin{cases} \frac{1}{8}\mathcal{N}(0, 0.7^2) + \frac{7}{8}\mathcal{N}(4, 0.7^2) & \text{, when } X = 1, \\ \mathcal{N}(2, 0.7^2) & \text{, when } X = 0. \end{cases}$$

Similarly, `Linear` requires a linear link between $\mathbb{W}$ and Y, that is, $\mathbb{E}[Y|\mathbb{W}, Z] = \alpha + \beta \cdot \mathbb{W} + g(Z)$. Hence, we consider scenarios in which this relation is satisfied and also others in which there exists a nonlinear link between Y and $\mathbb{W}$:

(a) **Linear link**

$$Y \mid W_1, W_2, Z \sim \mathcal{N}(W_1 + W_2 + Z, 1), \text{ and}$$

(b) **Squared- difference link**

$$Y \mid W_1, W_2, Z \sim \mathcal{N}((W_1 - W_2)^2 + Z, 1).$$

Using the above considerations, we considered 4 scenarios: 1a, 1b, 2a, and 2b. In 1a all methods have valid assumption, in 1b all except for `Linear`, in 2a all except for `Logistic`, and in 2b all except for `Linear` and `Logistic`. Each scenario was generated 100 times, with sample sizes of 100, 500, and 1,000.

4.1.2 Results

For each scenario, each method was evaluated according to the mean squared error (MSE) between the ACE estimates and the true ACE value.

- **1a. Normal distribution with linear link**: This is the only scenario in which all methods have valid assumptions. The simulation results are summarized in Figure 2. All methods obtain estimates that are close to the true ACE value, as the sample size increases. However, both Figure 2 and Table 1 show that, given its parametric nature, `Linear` converges the fastest.
- **1b. Normal distribution with squared-difference link**: In this scenario all methods except for `Linear` have valid assumptions. The simulation results are summarized in Figure 3. As expected, while `Frontdoor-CDE`, `Frontdoor-Odds`, and `Logistic` approach the true ACE value as the sample size increases, `Linear` is still relatively far, even for a sample of size 1,000. Table 2 shows that, `Frontdoor-Odds` and `Logistic` obtain similar rates, slightly outperforming `Frontdoor-CDE`.
- **2a. Normal-mixture distribution with linear link**: In this scenario, all methods except for `Logistic` have valid assumptions. The simulation results are summarized in Figure 4. As expected, all methods approach the true ACE value except for `Logistic`, which remains distant, even for a sample of size 1,000. Table 3 shows that, due to the parametric nature of `Linear`, it converges faster than `Frontdoor-Odds`, which slightly outperforms `Frontdoor-CDE`.

Figure 2: Boxplots for the absolute error obtained for each method in scenario 1a, in which $\mathbb{W}$ follows a normal distribution and Y has a linear link with $\mathbb{W}$.

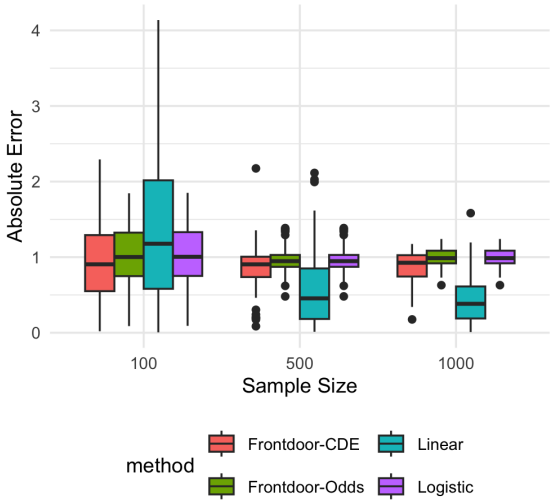


Table 1: Mean squared error (MSE) for each method across sample sizes in scenario 1a, in which $\mathbb{W}$ follows a normal distribution and Y has a linear link with $\mathbb{W}$.

	Sample Size		
Method	**100**	**500**	**1,000**
`Frontdoor-CDE`	1.08	0.820	0.808
`Frontdoor-Odds`	1.19	0.950	0.998
`Linear`	3.11	0.571	0.290
`Logistic`	1.20	0.950	0.998

- **2b. Normal-mixture with squared-difference link**: The results are summarized in Figure 5. The estimates for `Logistic` and `Linear` are far from the true ACE value, even for large sample sizes. Also, `Frontdoor-CDE` and `Frontdoor-Odds` obtain estimates that are much closer to the true ACE value. As the sample size increases, `Frontdoor-CDE` and `Frontdoor-Odds` perform similarly.

To sum up, the proposed nonparametric estimators, `Frontdoor-CDE` and `Frontdoor-Odds`, outperformed parametric methods when their assumptions were invalid and remained competitive otherwise. When all methods were valid (Scenario 1a), `Linear` converged faster, but `Frontdoor-CDE` and `Frontdoor-Odds` performed well. In scenarios with nonlinear relationships (1b, 2b) or misspecified density assumptions (2a), parametric estimators failed, while `Frontdoor-CDE` and `Frontdoor-Odds` provided accurate estimates. Notably, `Frontdoor-Odds` performed best in fully nonparametric settings, demonstrating its robustness.

Figure 3: Boxplots for the absolute error obtained for each method in scenario 1b, in which $\mathbb{W}$ follows a normal distribution and Y has a squared-difference link with $\mathbb{W}$.

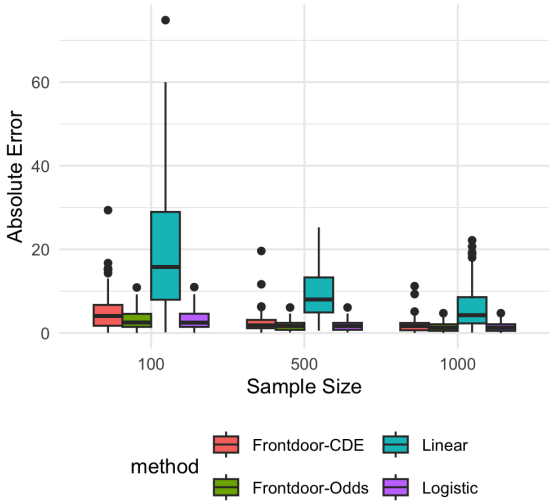


Table 2: Mean squared error (MSE) for each method across sample sizes in scenario 1b, in which $\mathbb{W}$ follows a normal distribution and Y has a squared-difference link with $\mathbb{W}$.

	Sample Size		
Method	**100**	**500**	**1,000**
`Frontdoor-CDE`	45.5	12.2	6.24
`Frontdoor-Odds`	16.4	4.41	2.97
`Linear`	674.	117.	61.1
`Logistic`	16.5	4.41	2.98

Figure 4: Boxplots for the absolute error obtained for each method in scenario 2a, in which $\mathbb{W}$ follows a normal-mixture distribution and Y has a linear link with $\mathbb{W}$.

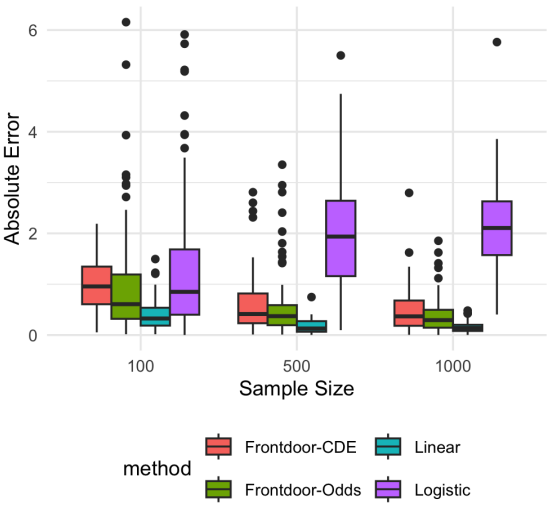


Table 3: MSE for each method across sample sizes in scenario 2a, in which $\mathbb{W}$ follows a normal-mixture distribution and Y has a linear link with $\mathbb{W}$.

	Sample Size		
Method	**100**	**500**	**1,000**
`Frontdoor-CDE`	1.17	0.660	0.414
`Frontdoor-Odds`	2.08	0.678	0.262
`Linear`	0.230	0.0452	0.0341
`Logistic`	3.50	5.35	5.30

Figure 5: Boxplots for the absolute error obtained for each method in scenario 2b, in which $\mathbb{W}$ follows a normal-mixture distribution and Y has a squared-difference link with $\mathbb{W}$.

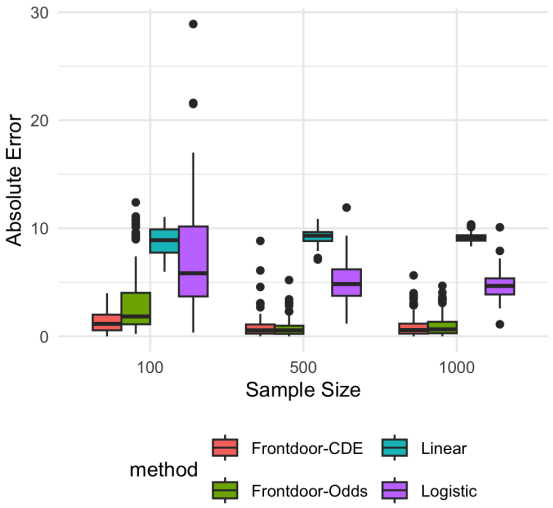


Table 4: MSE for each method across sample sizes in scenario 2c, in which $\mathbb{W}$ follows a normal-mixture distribution and Y has a squared-difference link with $\mathbb{W}$.

	Sample Size		
Method	**100**	**500**	**1,000**
`Frontdoor-CDE`	2.84	2.21	2.11
`Frontdoor-Odds`	18.1	1.34	2.00
`Linear`	78.5	85.3	83.7
`Logistic`	81.8	30.8	23.3

From the simulations presented in this section and the assumptions discussed in Subsection 3.1, the choice between the proposed estimators depends on the underlying characteristics of the data and the available sample size. Both estimators are highly flexible, however, their practical suitability often varies with the structure of the problem.

`Frontdoor-CDE` relies on conditional density estimation and is thus particularly well suited for situations where the distribution of $\mathbb{W}$ given X is sufficiently smooth to allow accurate nonparametric estimation, and also when $\mathbb{W}$ is low-dimensional. We emphasize that careful selection of the conditional density estimator is crucial in this context, and refer the reader to Dalmasso et al. [2020] for a discussion of estimators suited to different data scenarios.

In contrast, `Frontdoor-Odds` relies on estimating conditional odds of X given $\mathbb{W}$, which can be more stable and computationally tractable, especially when $\mathbb{W}$ is high-dimensional. In our study, we used the same neural network architecture for `Frontdoor-Odds` across all scenarios and sample sizes to maintain comparability. This sometimes led to a higher incidence of outliers in smaller samples, highlighting the importance of tuning architectures and hyperparameters to the specific data complexity and sample size. For further practical guidance on selecting classification models and their tuning procedures, we refer to James et al. [2021], Goodfellow et al. [2016].

From a computational perspective, `Frontdoor-CDE` generally requires substantially more resources due to the explicit density estimation step, whereas `Frontdoor-Odds` is typically faster and works well with modern neural network frameworks. Taken together, our findings suggest that practitioners might prefer `Frontdoor-CDE` when accurate conditional density estimation is feasible, the data dimensionality is not excessively high, and computational resources allow. In contrast, `Frontdoor-Odds` emerges as a compelling choice for higher-dimensional problems, or when the use of flexible classifiers is particularly advantageous.

4.2 CONTINUOUS TREATMENT

We conducted a simulation study to evaluate the performance of `Frontdoor-S` in estimating $\mathbb{E}[Y \mid \text{do}(X = x)]$ in a setting where X is a continuous treatment and W satisfies the front-door criterion.

The conditional density ratio was estimated using a neural network with input $(\tilde{W}, \tilde{X}, \tilde{X}', \tilde{S})$, output $\lambda(\tilde{W}, \tilde{X}', \tilde{X})$, and that minimizes the loss specified in Equation 9. The network architecture consisted of three fully connected layers. The first with 64 units, the second with 32 units, and a final output layer with a single unit. Each hidden layer was followed by a ReLU activation function and a dropout layer with a dropout rate of 0.2 to prevent overfitting. The model

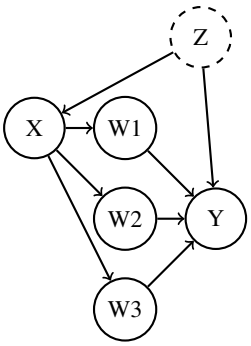


Figure 6: Causal graph that was used in the continuous treatment simulation. We aim to establish the causal effect of X on Y. Z is an unobserved confounder. W_1, W_2 and W_3 are mediators which, together, satisfy the front-door criterion.

was trained using the Adam optimizer for up to 30 epochs, with early stopping based on validation performance and a patience of 3 epochs.

The data was generated according to Figure 6, as follows:

$$
\begin{aligned}
Z &\sim \mathcal{N}(0,1) \\
X \mid Z &\sim \begin{cases} \mathcal{N}(2,1), & \text{if } Z \geq 0 \\ \mathcal{N}(0,1), & \text{if } Z < 0 \end{cases} \\
W_i \mid X &\sim \mathcal{N}(X^2, 0.5^2), \text{ with } i = 1,2,3 \\
Y \mid W, Z &\sim \text{Bernoulli}(p), \\
&\text{where } \text{logit}(p) = 2(W_1 + W_2 + W_3) + 2\mathbb{I}(Z \geq 0)
\end{aligned}
$$

X and W_i are nonlinearly related. All variables were taken as observable, except for the latent confounder, Z.

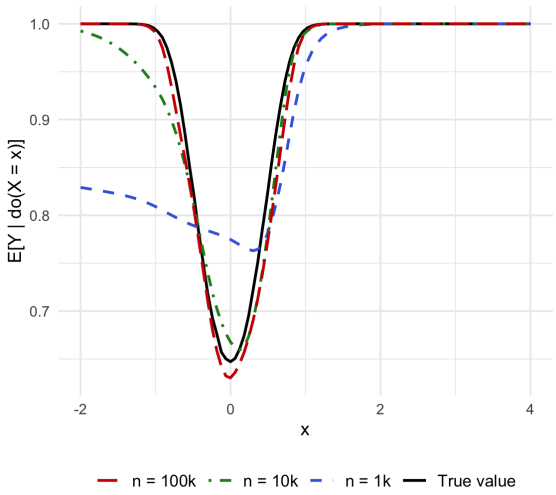
The results displayed in Figure 7 demonstrate that the estimates of `Frontdoor-S` become increasingly consistent with the true causal effect as the sample size increases, while effectively capturing complex relationships in the data.

The simulations presented and the assumptions discussed in Subsection 3.1 indicate that `Frontdoor-S` is a consistent and reliable estimator for continuous treatments, and can be particularly advantageous in scenarios with high-dimensional $\mathbb{W}$, given its reliance on estimating conditional odds of S given X, X' and $\mathbb{W}$. This framework naturally accommodates flexible classification approaches, similar to those employed for `Frontdoor-Odds`. For practical considerations on model choice and tuning strategies in this context, see James et al. [2021], Goodfellow et al. [2016].

5 CONCLUSION

We introduce three nonparametric estimators for the causal effect based on the front-door criterion: `Frontdoor-CDE`, `Frontdoor-Odds` and `Frontdoor-S`. These estima-

Figure 7: Estimated causal effect curves for different sample sizes. The black solid line represents the true interventional expectation. The dashed, dot-dashed, and long-dashed lines correspond to estimates obtained with $n = 1{,}000$, $n = 10{,}000$, and $n = 100{,}000$ samples, respectively.



tors offer flexibility, as they are obtained by plugging in arbitrary conditional density estimators or conditional odds ratio estimators into a weighted sample average. All estimators are consistent under mild nonparametric assumptions. In the binary treatment setting, simulation studies demonstrate that both `Frontdoor-CDE` and `Frontdoor-Odds` outperform parametric estimators when model assumptions are violated while remaining competitive when assumptions hold. In the continuous treatment setting, simulation results demonstrate that `Frontdoor-S` is consistent and effectively captures complex relationships in the data.

Additionally, initial studies suggest that the performance of the proposed estimators could be further improved. For example, in the current implementation, `Frontdoor-Odds` and `Frontdoor-S` estimate the conditional odds ratio through likelihood maximization. Refining this estimation method specifically for `Frontdoor-Odds` and `Frontdoor-S` may enhance their performance.

As part of future research, these nonparametric proposals might be generalized. For instance, one might also study whether `Frontdoor-CDE`, `Frontdoor-Odds` or `Frontdoor-S` can be applied to relaxations of the frontdoor criterion, particularly in cases where some mediators have confounders. These methods might also prove useful for estimating other causal quantities, such as the average treatment effect on the treated (ATT) and the complier average causal effect (CACE).

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Marc F. Bellemare, Jeffrey R. Bloem, and Noah Wexler. The paper of how: Estimating treatment effects using the front-door criterion. *Oxford Bulletin of Economics and Statistics*, 86(4):0305–9049, 2024. doi: 10.1111/obes.12598. URL `https://doi.org/10.1111/obes.12598`.

Christopher M Bishop. Mixture density networks. 1994.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Automatic debiased machine learning for dynamic treatment effects and general nested functionals. *arXiv preprint arXiv:2203.13887*, 2022. URL `https://arxiv.org/abs/2203.13887`.

William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

Niccolò Dalmasso, Taylor Pospisil, Ann B Lee, Rafael Izbicki, Peter E Freeman, and Alex I Malz. Conditional density estimation tools in Python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, page 100362, 2020.

Niccolò Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B. Lee. Likelihood-free frequentist inference: Bridging classical statistics and machine learning for reliable simulator-based inference. *arXiv preprint arXiv:2107.03920*, 2021. URL `https://arxiv.org/abs/2107.03920`. Submitted on 8 Jul 2021 (v1), last revised 26 Nov 2024 (v10).

Biprateep Dey, David Zhao, Brett H. Andrews, Jeffrey A. Newman, Rafael Izbicki, and Ann B. Lee. Towards instance-wise calibration: Local amortized diagnostics and reshaping of conditional densities (ladar), 2024. URL `https://arxiv.org/abs/2205.14568`.

Alan H. Dorfman and Richard Valliant. The hájek estimator revisited. In *Proceedings of the Section on Survey Research Methods*, pages 760–765, Alexandria, VA, 1997. American Statistical Association. URL `https://www.asasrms.org/Proceedings/papers/1997_130.pdf`.

Isabel R. Fulcher, Ilya Shpitser, Stella Marealle, and Eric J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: The generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):199–214, 2020. doi: 10.1111/rssb.12345. URL https://doi.org/10.1111/rssb.12345.

Madelyn Glymour, Judea Pearl, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Sander Greenland and James M Robins. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.

Anna Guo, David Benkeser, and Razieh Nabi. Targeted machine learning for average causal effect estimation using the front-door functional. *arXiv preprint arXiv:2312.10234*, 2023. URL https://arxiv.org/abs/2312.10234.

Shantanu Gupta, Zachary C. Lipton, and David Childers. Estimating treatment effects with observed confounders and mediators. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *PMLR*, pages 982–991. Proceedings of Machine Learning Research, 2021. URL https://proceedings.mlr.press/v161/gupta21b.html.

R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational & Graphical Statistics*, 5:315–336, 1996.

T. Ichimura and D. Fukuda. A fast algorithm for computing least-squares cross-validations for nonparametric conditional kernel density functions. *Computational Statistics Data Analysis*, 54(12):3404–3410, 2010.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Rafael Izbicki and Ann B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 2017.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2nd edition, 2021. URL https://www.statlearning.com/.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: Dose, heterogeneous, and incremental response curves. *Biometrika*, 111 (2):497–516, 2024. doi: 10.1093/biomet/asad042. URL https://academic.oup.com/biomet/article/111/2/497/7219715.

I. Takeuchi, K. Nomura, and T. Kanamori. Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, 21(2):533–559, 2009.

Nonparametic Estimation of the Average Causal Effect using the Front-door Criterion (Supplementary Material)

Ana P. L. Cavalcante[1] **Rafael Izbicki**[2] **Rafael B. Stern**[1]

[1]Department of Statistics, University of São Paulo, São Paulo, BRA
[2]Department of Statistics, Federal University of São Carlos, São Paulo, BRA

A THE FRONT-DOOR CRITERION

The following theorem can be found in Glymour et al. [2016].

Theorem A.1. If W satisfies the front-door criterion relative to (X, Y) and if $\mathbb{P}(x, w) > 0$, then the causal effect of X on Y is identifiable and given by:

$$\mathbb{P}(y \mid \text{do}(x)) = \sum_w \mathbb{P}(w \mid x) \sum_{x'} \mathbb{P}(y \mid x', w)\mathbb{P}(x')$$

A.1 PROOF OF THEOREM 1.1

Proof of Theorem 1.1. If W satisfies the front-door criterion (Definition 1.1) for estimating the causal effect of X on Y, then, using Theorem A.1, we have

$$\begin{aligned}
\mathbb{E}[Y \mid \text{do}(X = x)] &= \sum_y y \cdot f(y \mid \text{do}(x)) \\
&= \sum_y \sum_w \sum_{x'} y \cdot f(w \mid x) f(y \mid x', w) f(x') \\
&= \sum_y \sum_w \sum_{x'} y \frac{f(w \mid x)}{f(w \mid x')} f(y, w, x') \\
&= \mathbb{E}\left[Y \frac{f(W \mid x)}{f(W \mid X)} \right]
\end{aligned}$$

The continuous and discrete case are analogous.

B PROOF OF THEORETICAL PROPERTIES

Lemma B.1. If $(W_n)_{n \in \mathbb{N}}$ is a sequence of random variables such that $\mathbb{E}[|W_n|] = o(1)$, then $W_n \xrightarrow{\mathbb{P}} 0$.

Proof of Lemma B.1.

$$\begin{aligned}
\mathbb{P}(|W_n| > \epsilon) &\leq \frac{\mathbb{E}[|W_n|]}{\epsilon} && \text{Markov's inequality} \\
&= o(1)
\end{aligned}$$

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

B.1 PROOF OF THEOREM 3.1.A

Proof of Theorem 3.1.a. By the Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i f(W_i|x)}{f(W_i|X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{Y\cdot f(W|x)}{f(W|X)}\right]$. Under the conditions of Theorem 1.1, $\mathbb{E}\left[\frac{Y\cdot f(W|x)}{f(W|X)}\right] = \mathbb{E}[Y \mid \text{do}(X=x)]$. Therefore, using Lemma B.1, it is sufficient to prove that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i f(W_i \mid x)}{f(W_i \mid X_i)}\right|\right] = o(1).$$

Consider that Assumption 1 is valid.

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i f(W_i \mid x)}{f(W_i \mid X_i)}\right|\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{Y_i\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} - \frac{Y_i f(W_i \mid x)}{f(W_i \mid X_i)}\right|\right] \\
&= \mathbb{E}\left[\left|\frac{Y_1\hat{f}(W_1 \mid x)}{\hat{f}(W_1 \mid X_1)} - \frac{Y_1 f(W_1 \mid x)}{f(W_1 \mid X_1)}\right|\right] && \text{Assumption 3.a} \\
&= \mathbb{E}\left[\left|Y_1 \cdot \frac{\hat{f}(W_1 \mid x)f(W_1 \mid X_1) - f(W_1 \mid x)\hat{f}(W_1 \mid X_1)}{\hat{f}(W_1 \mid X_1)f(W_1 \mid X_1)}\right|\right] \\
&\leq \delta^{-2}\mathbb{E}\left[\left|Y_1 \cdot \left(\hat{f}(W_1 \mid x)f(W_1 \mid X_1) - f(W_1 \mid x)\hat{f}(W_1 \mid X_1)\right)\right|\right] && \text{Assumption 6} \\
&= \delta^{-2}\mathbb{E}\left[\left|\hat{f}(W_1 \mid x)f(W_1 \mid X_1) - f(W_1 \mid x)\hat{f}(W_1 \mid X_1)\right| \cdot \mathbb{E}\left[|Y_1|\middle|W_1, X_1\right]\right] \\
&\leq M\delta^{-2}\mathbb{E}\left[\left|\hat{f}(W_1 \mid x)f(W_1 \mid X_1) - f(W_1 \mid x)\hat{f}(W_1 \mid X_1)\right|\right] && \text{Assumption 2} \\
&\leq M\delta^{-2}\left(\mathbb{E}\left[\left|\hat{f}(W_1 \mid x)\left(f(W_1 \mid X_1) - \hat{f}(W_1 \mid X_1)\right)\right|\right] + \mathbb{E}\left[\left|\hat{f}(W_1 \mid X_1)\left(f(W_1 \mid x) - \hat{f}(W_1 \mid x)\right)\right|\right]\right) \\
&\leq M\delta^{-2}\left(\mathbb{E}\left[\hat{f}^2(W_1|x)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(f(W_1 \mid X_1) - \hat{f}(W_1 \mid X_1)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\hat{f}^2(W_1|X_1)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(f(W_1 \mid x) - \hat{f}(W_1 \mid x)\right)^2\right]^{\frac{1}{2}}\right) \\
&\leq MK\delta^{-2}\left(\mathbb{E}\left[\left(f(W_1 \mid X_1) - \hat{f}(W_1 \mid X_1)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\left(f(W_1 \mid x) - \hat{f}(W_1 \mid x)\right)^2\right]^{\frac{1}{2}}\right) && \text{Assumption 5.a} \\
&= o(1) && \text{Assumption 4.a}
\end{aligned}$$

This proves that $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\hat{f}(W_i|x)}{\hat{f}(W_i|X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{Y\cdot f(W|x)}{f(W|X)}\right]$.

Again using the Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^{n}\frac{f(W_i|x)}{f(W_i|X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{f(W|x)}{f(W|X)}\right] = 1$. Using Lemma B, we will now prove that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{f(W_i \mid x)}{f(W_i \mid X_i)}\right|\right] = o(1).$$

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{f(W_i \mid x)}{f(W_i \mid X_i)}\right|\right]$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{\hat{f}(W_i \mid x)}{\hat{f}(W_i \mid X_i)} - \frac{f(W_i \mid x)}{f(W_i \mid X_i)}\right|\right] \\
&= \mathbb{E}\left[\left|\frac{\hat{f}(W_1 \mid x)}{\hat{f}(W_1 \mid X_1)} - \frac{f(W_1 \mid x)}{f(W_1 \mid X_1)}\right|\right] && \text{Assumption 3.a} \\
&= \mathbb{E}\left[\left|\frac{\hat{f}(W_1 \mid x)f(W_1 \mid X_1) - f(W_1 \mid x)\hat{f}(W_1 \mid X_1)}{\hat{f}(W_1 \mid X_1)f(W_1 \mid X_1)}\right|\right] \\
&\leq \delta^{-2}\mathbb{E}\left[\left|\hat{f}(W_1 \mid x)f(W_1 \mid X_1) - f(W_1 \mid x)\hat{f}(W_1 \mid X_1)\right|\right] && \text{Assumption 6} \\
&\leq \delta^{-2}\left(\mathbb{E}\left[\left|\hat{f}(W_1 \mid x)\left(f(W_1 \mid X_1) - \hat{f}(W_1 \mid X_1)\right)\right|\right] + \mathbb{E}\left[\left|\hat{f}(W_1 \mid X_1)\left(f(W_1 \mid x) - \hat{f}(W_1 \mid x)\right)\right|\right]\right) \\
&\leq \delta^{-2}\left(\mathbb{E}\left[\hat{f}^2(W_1|x)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(f(W_1 \mid X_1) - \hat{f}(W_1 \mid X_1)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\hat{f}^2(W_1|X_1)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(f(W_1 \mid x) - \hat{f}(W_1 \mid x)\right)^2\right]^{\frac{1}{2}}\right) \\
&\leq K\delta^{-2}\left(\mathbb{E}\left[\left(f(W_1 \mid X_1) - \hat{f}(W_1 \mid X_1)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\left(f(W_1 \mid x) - \hat{f}(W_1 \mid x)\right)^2\right]^{\frac{1}{2}}\right) && \text{Assumption 5.a} \\
&= o(1) && \text{Assumption 4.a}
\end{aligned}
$$

As $\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}(W_i|x)}{\hat{f}(W_i|X_i)} \xrightarrow{\mathbb{P}} 1$ and $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\hat{f}(W_i|x)}{\hat{f}(W_i|X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)]$, by using Slutsky's theorem, we prove that $\widehat{\mathbb{E}}_f[Y \mid \text{do}(X = x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)]$.

□

B.2 PROOF OF THEOREM 3.1.B

Proof of Theorem 3.1.b. By the Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^{n} Y_i O(W_i, X_i, x) \xrightarrow{\mathbb{P}} \mathbb{E}\left[Y \cdot O(X, W, x)\right]$. Under the conditions of Theorem 1.1, $\mathbb{E}\left[Y \cdot O(W, X, x)\right] = \mathbb{E}[Y \mid \text{do}(X = x)]$. Therefore, using Lemma B.1, it is sufficient to prove that

$$
\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{O}(W_i, X_i, x) - \frac{1}{n}\sum_{i=1}^{n} Y_i O(W_i, X_i, x)\right|\right] = o(1).
$$

Consider that Assumption 1 is valid.

$$
\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{O}(W_i, X_i, x) - \frac{1}{n}\sum_{i=1}^{n} Y_i O(W_i, X_i, x)\right|\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|Y_i\hat{O}(W_i, X_i, x) - Y_i O(W_i, X_i, x)\right|\right] \\
&\leq \mathbb{E}\left[\left|Y_1 \cdot \left(\hat{O}(W_1, X_1, x) - O(W_1, X_1, x)\right)\right|\right] && \text{Assumption 3.b} \\
&= \mathbb{E}\left[\left|\hat{O}(W_1, X_1, x) - O(W_1, X_1, x)\right| \cdot \mathbb{E}\left[|Y_1|\middle|W_1, X_1\right]\right] \\
&\leq M\mathbb{E}\left[\left|\hat{O}(W_1, X_1, x) - O(W_1, X_1, x)\right|\right] && \text{Assumption 2}
\end{aligned}
$$

$$
\begin{aligned}
&\leq M\mathbb{E}\left[\left|\hat{O}_1(W_1,X_1,x)\hat{O}_2(X_1,x) - O_1(W_1,X_1,x)O_2(X_1,x)\right|\right] \\
&\leq M\left(\mathbb{E}\left[\left|\hat{O}_1(W_1,X_1,x)\left(\hat{O}_2(X_1,x) - O_2(X_1,x)\right)\right|\right] + \mathbb{E}\left[\left|O_2(X_1,x)\left(\hat{O}_1(W_1,X_1,x) - O_1(W_1,X_1,x)\right)\right|\right]\right) \\
&\leq M\left(\mathbb{E}\left[\hat{O}_1^2(W_1,X_1,x)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\hat{O}_2(X_1,x) - O_2(X_1,x)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[O_2^2(X_1,x)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\hat{O}_1(W_1,X_1,x) - O_1(W_1,X_1,x)\right)^2\right]^{\frac{1}{2}}\right) \\
&\leq MK\left(\mathbb{E}\left[\left(\hat{O}_2(X_1,x) - O_2(X_1,x)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\left(\hat{O}_1(W_1,X_1,x) - O_1(W_1,X_1,x)\right)^2\right]^{\frac{1}{2}}\right) && \text{Assumption 5.b} \\
&= o(1) && \text{Assumption 4.b}
\end{aligned}
$$

This proves that $\frac{1}{n}\sum_{i=1}^n Y_i\hat{O}(W_i,X_i,x) \xrightarrow{\mathbb{P}} \mathbb{E}\left[Y\cdot O(W,X,x)\right]$.

Again using the Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^n O(W_i,X_i,x) \xrightarrow{\mathbb{P}} \mathbb{E}\left[O(W,X,x)\right] = 1$. Using Lemma B, we will now prove that

$$
\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \hat{O}(W_i,X_i,x) - \frac{1}{n}\sum_{i=1}^n O(W_i,X_i,x)\right|\right] = o(1).
$$

$$
\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \hat{O}(W_i,X_i,x) - \frac{1}{n}\sum_{i=1}^n O(W_i,X_i,x)\right|\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left|\hat{O}(W_i,X_i,x) - O(W_i,X_i,x)\right|\right] \\
&\leq \mathbb{E}\left[\left|\hat{O}(W_1,X_1,x) - O(W_1,X_1,x)\right|\right] && \text{Assumption 3.b} \\
&= \mathbb{E}\left[\left|\hat{O}_1(W_1,X_1,x)\hat{O}_2(X_1,x) - O_1(W_1,X_1,x)O_2(X_1,x)\right|\right] \\
&\leq \mathbb{E}\left[\left|\hat{O}_1(W_1,X_1,x)\left(\hat{O}_2(X_1,x) - O_2(X_1,x)\right)\right|\right] + \mathbb{E}\left[\left|O_2(X_1,x)\left(\hat{O}_1(W_1,X_1,x) - O_1(W_1,X_1,x)\right)\right|\right] \\
&\leq \left(\mathbb{E}\left[\hat{O}_1^2(W_1,X_1,x)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\hat{O}_2(X_1,x) - O_2(X_1,x)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[O_2^2(X_1,x)\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\hat{O}_1(W_1,X_1,x) - O_1(W_1,X_1,x)\right)^2\right]^{\frac{1}{2}}\right) \\
&\leq K\left(\mathbb{E}\left[\left(\hat{O}_2(X_1,x) - O_2(X_1,x)\right)^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\left(\hat{O}_1(W_1,X_1,x) - O_1(W_1,X_1,x)\right)^2\right]^{\frac{1}{2}}\right) && \text{Assumption 5.b} \\
&= o(1) && \text{Assumption 4.b}
\end{aligned}
$$

As $\frac{1}{n}\sum_{i=1}^n \hat{O}(W_i,X_i,x) \xrightarrow{\mathbb{P}} 1$ and $\frac{1}{n}\sum_{i=1}^n Y_i\hat{O}(W_i,X_i,x) \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X=x)]$, by using Slutsky's theorem, we prove that $\widehat{\mathbb{E}}_o[Y \mid \text{do}(X=x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X=x)]$.

□

B.3 PROOF OF THEOREM 3.1.C

Proof of Theorem 3.1.c. By the Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^n Y_i\lambda(W_i,X_i,x) \xrightarrow{\mathbb{P}} \mathbb{E}\left[Y\cdot\lambda(X,W,x)\right]$. Under the conditions of Theorem 1.1, $\mathbb{E}\left[Y\cdot\lambda(W,X,x)\right] = \mathbb{E}[Y \mid \text{do}(X=x)]$. Therefore, using Lemma B.1, it is sufficient to prove

that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{\lambda}(W_i, X_i, x) - \frac{1}{n}\sum_{i=1}^{n} Y_i\lambda(W_i, X_i, x)\right|\right] = o(1).$$

Consider that Assumption 1 is valid.

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{\lambda}(W_i, X_i, x) - \frac{1}{n}\sum_{i=1}^{n} Y_i\lambda(W_i, X_i, x)\right|\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|Y_i\hat{\lambda}(W_i, X_i, x) - Y_i\lambda(W_i, X_i, x)\right|\right] \\
&\leq \mathbb{E}\left[\left|Y_1 \cdot \left(\hat{\lambda}(W_1, X_1, x) - \lambda(W_1, X_1, x)\right)\right|\right] && \text{Assumption 3.c} \\
&= \mathbb{E}\left[\left|\hat{\lambda}(W_1, X_1, x) - \lambda(W_1, X_1, x)\right| \cdot \mathbb{E}\left[|Y_1|\middle|W_1, X_1\right]\right] \\
&\leq M\mathbb{E}\left[\left|\hat{\lambda}(W_1, X_1, x) - \lambda(W_1, X_1, x)\right|\right] && \text{Assumption 2} \\
&= o(1) && \text{Assumption 4.c}
\end{aligned}$$

This proves that $\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{\lambda}(W_i, X_i, x) \xrightarrow{\mathbb{P}} \mathbb{E}\left[Y \cdot \lambda(W, X, x)\right]$.

Again using the Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^{n} \lambda(W_i, X_i, x) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\lambda(W, X, x)\right] = 1$. Using Lemma B, we will now prove that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}(W_i, X_i, x) - \frac{1}{n}\sum_{i=1}^{n} \lambda(W_i, X_i, x)\right|\right] = o(1).$$

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}(W_i, X_i, x) - \frac{1}{n}\sum_{i=1}^{n} \lambda(W_i, X_i, x)\right|\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|\hat{\lambda}(W_i, X_i, x) - \lambda(W_i, X_i, x)\right|\right] \\
&\leq \mathbb{E}\left[\left|\hat{\lambda}(W_1, X_1, x) - \lambda(W_1, X_1, x)\right|\right] && \text{Assumption 3.c} \\
&= o(1) && \text{Assumption 4.c}
\end{aligned}$$

As $\frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}(W_i, X_i, x) \xrightarrow{\mathbb{P}} 1$ and $\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{\lambda}(W_i, X_i, x) \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)]$, by using Slutsky's theorem, we prove that $\widehat{\mathbb{E}}_s[Y \mid \text{do}(X = x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y \mid \text{do}(X = x)]$.

□