

ECMA 31380 - Proposal Final Project

Fernando Rocha Urbano

Autumn 2024

Which questions can we answer?

- Which model performs better for different ratio between d_c and d_a ? Meaning the amount of covariates that are relevant and irrelevant.
- Which model performs better in general?
- Which model has the best convergence rate (different sample sizes)?
- Which model performs better for different noise levels?
- Which models performs better for different levels of sparsity (amount of non-causal covariates)?
- Is the order of performance the same in terms of bias and variance?
- Which model performs better when there is misspecification?

1 Goal

The goal of this project is to evaluate the performance of different Double Machine Learning (DML) models in recovering the Average Treatment Effect (ATE) across various causal inference scenarios addressed in simulations.

Ideally, the project should provide actionable recommendations for selecting the most appropriate DML method based on the causal scenario, data structure, noise levels, and sample size, sparsity of important covariates. Additionally, we aim to create a benchmarking framework that practitioners can use to evaluate and compare causal inference estimates derived from the tested methods in different scenarios.

The DML methods compared in this study are:

- LASSO.
- Random Forest.
- Neural Networks.

The causal inference scenarios are:

- Backdoor adjustment.
- Frontdoor adjustment with and without access to mediators.
- Instrumental Variables with varying levels of instrument strength and correct vs. incorrect use of the IV.

We explore how each model implementation performs under varying conditions:

- Relationships as linear and non-linear between covariates and treatment and between covariates and target ($g(\cdot)$ and $m(\cdot)$).
- Noise levels (ε).
- Sample size (n).
- Size of high-dimensional impactful covariates (d_c).
- Size of high-dimensional irrelevant covariates (d_a).

For each of the scenario combinations, we show which of LASSO, Random Forest and Neural Networks provides more accurate ATE estimates based on the following metrics.

- Bias of $\hat{\text{ATE}}$.
- Variance of $\hat{\text{ATE}}$ across repeated simulations for a given scenario.
- Analytical Variance of the $\hat{\text{ATE}}$.
- MSE (Mean Squared Error).

2 Double Machine Learning

2.1 Problem Setup

Consider a random sample $(Y_i, T_i, X_i)_{i=1}^n$, where:

- $Y_i \in \mathbb{R}$ is the outcome variable.
- $T_i \in 0, 1$ is a binary treatment indicator.
- $X_i \in \mathbb{R}^d$ is a vector of covariates.

Our goal is to estimate the Average Treatment Effect (ATE), defined as:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)], \quad (2.1)$$

where $Y_i(t)$ denotes the potential outcome for unit i under treatment $T_i = t$.

2.2 Identification via Conditional Expectations

Under the Conditional Independence Assumption (CIA) and overlap conditions, the ATE can be identified as:

$$\tau = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)], \quad (2.2)$$

where $\mu_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i]$ is the conditional expectation of the outcome given treatment and covariates.

2.3 DML Estimator

The DML framework aims to estimate τ while controlling for high-dimensional or complex relationships between Y_i , T_i , and X_i . The key idea is to use machine learning methods to estimate the nuisance parameters and then construct an estimator for τ that is robust to estimation errors in these nuisance parameters.

2.3.1 Nuisance Parameter Estimation

We define the following nuisance functions:

$$m(X_i) = \mathbb{E}[Y_i|X_i], \quad (2.3)$$

$$g(X_i) = \mathbb{E}[T_i|X_i], \quad (2.4)$$

$$\pi(X_i) = \mathbb{P}(T_i = 1|X_i) \quad (\text{propensity score}) \quad (2.5)$$

These functions can be estimated using flexible machine learning methods such as LASSO, Random Forests, or Neural Networks.

Under mild regularity conditions and appropriate rates of convergence for the nuisance estimators, the DML estimator is root-n consistent and asymptotically normal (we make the regularity conditions explicitly in Appendix 5.1). It provides valid confidence intervals and hypothesis tests even when using complex machine learning methods for $\hat{m}(X)$ and $\hat{g}(X)$.

2.3.2 Orthogonal Score Function

To achieve robustness, we construct an orthogonal score function $\psi(Y_i, T_i, X_i; \eta)$, where $\eta = (m, g)$ represents the nuisance parameters. The orthogonal score satisfies the Neyman orthogonality condition, which ensures that small estimation errors in η have a negligible first-order impact on the estimation of τ .

A common choice for the orthogonal score is:

$$\psi(Y_i, T_i, X_i; \eta) = \left(\frac{T_i - g(X_i)}{\pi(X_i)(1 - \pi(X_i))} \right) (Y_i - m(X_i)) + (m_1(X_i) - m_0(X_i)) - \tau, \quad (2.6)$$

where $g(X_i) = \pi(X_i)$ in case of binary treatment and $m_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i]$ for $t \in \{0, 1\}$.

2.3.3 Estimation Procedure

The estimation proceeds in several steps:

1. Estimate Nuisance Functions: Use one of the outlined ML models to obtain estimators $\hat{m}(X_i)$ and $\hat{g}(X_i)$ for the nuisance functions.
2. Compute the Score Function: Evaluate the orthogonal score $\psi(Y_i, T_i, X_i; \hat{\eta})$ using the estimated nuisance parameters.
3. Estimate τ : Solve the empirical moment condition which yields $\hat{\tau}$:

$$\frac{1}{n} \sum_{i=1}^n \psi(Y_i, T_i, X_i; \hat{\eta}) = 0. \quad (2.7)$$

The orthogonal score function mitigates the impact of errors in $\hat{m}(X)$ and $\hat{g}(X)$ on the estimation of τ while also accomodating the use of modern ML techniques, allowing complex relationships between Y , X and T .

To prevent overfitting and ensure that the estimation error in the nuisance parameters does not bias the estimator of τ we employ cross-fitting.

2.3.4 Cross-Fitting

The steps of cross-fitting are:

1. Split the Sample: Divide the data into K folds $\{\mathcal{I}_k\}_{k=1}^K$.
2. For Each Fold:
 - (a) Train Nuisance Estimators: Use data from all other folds:

$$\mathcal{I}_{-k} = \bigcup_{j \neq k} \mathcal{I}_j$$

to estimate $\hat{m}^{(-k)}(X_i)$ and $\hat{g}^{(-k)}(X_i)$.

- (b) Compute Score Function: For observations in fold \mathcal{I}_k , compute $\psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)})$ using the nuisance estimates from step (a):

$$\begin{aligned} \psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)}) &= \left(\frac{T_i - \hat{g}^{(-k)}(X_i)}{\hat{\pi}^{(-k)}(X_i)} \right) (Y_i - \hat{m}^{(-k)}(X_i)) \\ &\quad + \hat{m}_1^{(-k)}(X_i) - \hat{m}_0^{(-k)}(X_i) - \tau^{(-k)}. \end{aligned} \quad (2.8)$$

3. Aggregate: Combine the estimates from all folds:

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \hat{\tau}_i^{(-k)} = \frac{1}{K} \sum_{k=1}^K \hat{\tau}^{(-k)} \quad (2.9)$$

3 Causal Inference Scenarios

The three most relevant scenarios of causal inference are the ones that require Instrumental Variables, Backdoor Adjustment, or Front Door Adjustment. A common way to represent the causal relation related to those is through the use of DAGs.

3.1 DAG (Directed Acyclic Graph)

Directed Acyclic Graph (DAG) serves as a representation of causal assumptions and a tool for deriving statistical properties of the variables involved.

It is composed of nodes (vertices) representing random variables or features and directed edges (arrows), indicating a direct influence or causal effect of T on Y .



Figure 1: DAG Example

The acyclic characterist is due to absence of directed cycles; that is, there is no path where you can start at a node X and, by following directed edges, return to X .

3.2 Backdoor Adjustment

A backdoor path from treatment T to the outcome Y represents alternative routes through which association can flow from T to Y that are not due to the causal effect of T on Y . In the representation, the confounder C is a common cause of both T and Y , thus, a backdoor path.

In such case, the association between T and Y may be partially or entirely due to their mutual dependence on C rather than a direct causal effect, leading to biased causal estimates of the treatment if C is ignored.

The causal effect of T on Y can be expressed using the backdoor adjustment formula:

$$\mathbb{P}[Y(t)] = \sum_C \mathbb{P}[Y \mid T, C] \mathbb{P}[C], \quad (3.1)$$

Which serves a markov factorization, calculating with respect to the DAG structure.

For the backdoor adjustment to be valid, the following conditions must be satisfied:

1. No variable in the adjustment set is a descendant of the treatment T .
2. The adjustment set blocks all backdoor paths from T to Y (backdoor path is any path from T to Y that starts with an arrow into T).

DAG 3.2 illustrates the backdoor path involving the confounder C , treatment T , outcome Y , and features with non-causal association X_a which would not be present in a typical backdoor adjustment DAG, but play a relevant role in the simulations proposed.

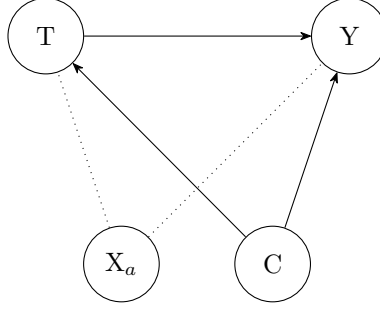


Figure 2: DAG of Backdoor Path

3.2.1 Backdoor Adjustment Scenario in DML

In Backdoor Adjustment with DML, the ideal $m(X_i) = \mathbb{E}[Y_i|X_i]$ and $g(X_i) = \mathbb{E}[T_i|X_i]$ from Equations (2.3) and (2.4) are:

$$m_{BA}(C_i) = \mathbb{E}[Y_i|C_i], \quad (3.2)$$

$$g_{BA}(C_i) = \mathbb{E}[T_i|C_i], \quad (3.3)$$

In our simulations we also test the habit of the ML models to estimate ATE under the following scenarios. In these scenarios, the superscript w_i on the nuisance functions denotes the i -th scenario which involves misspecifications of the nuisance functions.

Inclusion of C_i and $X_{a,i}$ on every nuisance function

$$m_{BA}^{w_1}(C_i, X_{a,i}) = \mathbb{E}[Y_i|C_i, X_{a,i}], \quad (3.4)$$

$$g_{BA}^{w_1}(C_i, X_{a,i}) = \mathbb{E}[T_i|C_i, X_{a,i}], \quad (3.5)$$

Equations (3.4) and (3.5) represent a scenario in which one would not be aware that X_a does not cause T and Y .

Only partial inclusion of C_i and inclusion of $X_{a,i}$ on every nuisance function

$$m_{BA}^{w_2}(C_i^p, X_{a,i}) = \mathbb{E}[Y_i|C_i^p, X_{a,i}], \quad (3.6)$$

$$g_{BA}^{w_2}(C_i^p, X_{a,i}) = \mathbb{E}[T_i|C_i^p, X_{a,i}], \quad (3.7)$$

Equations (3.6) and (3.7) also represent a scenario in which one would not be aware that X_a does not cause T and Y and also does not include all causal cofounders. C_i^p is subset of C_i included, where $C_i \in \mathbb{R}^{d_c}$ and $C_i^p \in \mathbb{R}^{d_{cp}}$ and $d_c > d_{cp}$.

3.3 Instrumental Variable

Instrumental variable (IV) estimation is a method used to estimate the causal effect of a treatment T_i on an outcome Y_i when there is unmeasured confounding U_i that cannot be addressed using

backdoor or frontdoor adjustments. This method leverages an instrument Z_i , which influences the treatment T_i but has no direct effect on the outcome Y_i except through T_i , and is independent of any unmeasured confounders U_i affecting both T_i and Y_i .

For the instrumental variable method to be valid, the following conditions must be satisfied:

1. Relevance: The instrument Z is associated with the treatment T (i.e., $\text{Cov}(Z, T) \neq 0$ or $Z \not\perp T$).
2. Exclusion Restriction: The instrument Z affects the outcome Y only through its effect on the treatment T (i.e., there is no direct effect of Z on Y and no other pathways from Z to Y except through T).
3. Independence (Ignorability): The instrument Z is independent of any unmeasured confounders U that affect both T and Y (i.e., $Z \perp U$).

DAG 3.3 illustrates the instrumental variable setup involving the unobserved confounder U , instrument Z , treatment T , outcome Y , observed confounder C , and features with non-causal associations X_a . Again, the last two would not be present in a typical instrumental variable DAG, but are included due to their relevance in the proposed simulations.

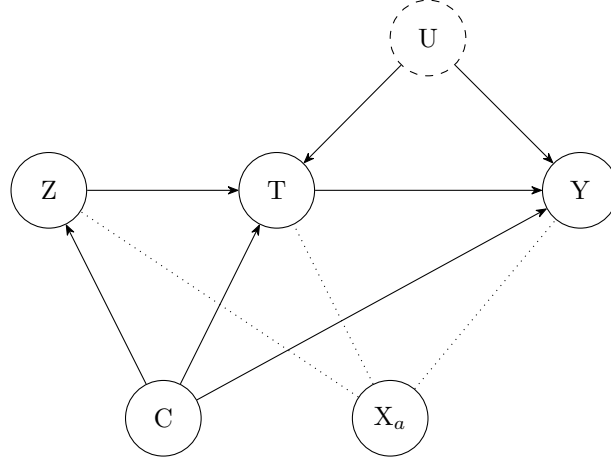


Figure 3: DAG of Instrumental Variable

3.3.1 Instrumental Variable Scenario in DML

To adapt the DML framework to the IV setting, we need to define appropriate nuisance functions and construct an orthogonal score function suitable for the IV context.

The ideal nuisance functions in this scenario are:

$$m_{\text{IV}}(C_i) = \mathbb{E}[Y_i \mid C_i], \quad (3.8)$$

$$q_{\text{IV}}(C_i) = \mathbb{E}[Z_i \mid C_i], \quad (3.9)$$

$$g_{\text{IV}}(C_i) = \mathbb{E}[T_i \mid C_i], \quad (3.10)$$

Where $m_{IV}(C_i)$ is the outcome model, capturing the expected outcome given covariates, $q_{IV}(Z_i, C_i)$ is the instrument propensity score, representing the expected treatment given the instrument and covariates, and $g_{IV}(C_i)$ is the treatment model, representing the expected treatment given covariates.

The orthogonal score function for the IV scenario is different from the standard DML orthogonal score function. An appropriate orthogonal score function in the linear IV context is:

$$\psi_{IV}(W_i; \tau, \eta) = (Y_i - m_{IV}(C_i) - \tau[T_i - g_{IV}(C_i)])(Z_i - q_{IV}(C_i)), \quad (3.11)$$

In our simulations we also test the habit of the ML models to estimate ATE under the following scenarios. In these scenarios, the superscript w_i on the nuisance functions denotes the i -th scenario which involves misspecifications of the nuisance functions.

Inclusion of C_i and $X_{a,i}$ on every nuisance function

$$m_{IV}^{w_1}(C_i) = \mathbb{E}[Y_i | C_i, X_{a,i}], \quad (3.12)$$

$$q_{IV}^{w_1}(C_i) = \mathbb{E}[Z_i | C_i, X_{a,i}], \quad (3.13)$$

$$g_{IV}^{w_1}(C_i) = \mathbb{E}[T_i | C_i, X_{a,i}], \quad (3.14)$$

Equations (3.12), (3.13), and (3.14) represent a scenario in which one would not be aware that X_a does not cause T , Z and Y .

Only partial inclusion of C_i and inclusion of $X_{a,i}$ on every nuisance function

$$m_{IV}^{w_2}(C_i^p, X_{a,i}) = \mathbb{E}[Y_i | C_i, X_{a,i}], \quad (3.15)$$

$$q_{IV}^{w_2}(C_i^p, X_{a,i}) = \mathbb{E}[Z_i | C_i, X_{a,i}], \quad (3.16)$$

$$g_{IV}^{w_2}(C_i^p, X_{a,i}) = \mathbb{E}[T_i | C_i, X_{a,i}], \quad (3.17)$$

Equations (3.15), (3.16), and (3.17) also represent a scenario in which one would not be aware that X_a does not cause T and Y and also does not include all causal cofounders. C_i^p is subset of C_i included, where $C_i \in \mathbb{R}^{d_c}$ and $C_i^p \in \mathbb{R}^{d_{cp}}$ and $d_c > d_{cp}$.

Z_i is treated as a normal cofounder

$$m_{IV}^{w_3}(C_i, Z_i) = \mathbb{E}[Y_i | C_i, Z_i], \quad (3.18)$$

$$g_{IV}^{w_3}(C_i, Z_i) = \mathbb{E}[T_i | C_i, Z_i], \quad (3.19)$$

Equations (3.20) and (3.21) represent a scenario in which one would treat Z as a normal covariate. In this case, the orthogonal score function would be the same as in the backdoor adjustment scenario of (2.6).

Z_i is treated as a normal cofounder and inclusion of $X_{a,i}$ on every nuisance function

$$m_{IV}^{w_4}(C_i, Z_i, X_{a,i}) = \mathbb{E}[Y_i | C_i, Z_i, X_{a,i}], \quad (3.20)$$

$$g_{IV}^{w_4}(C_i, Z_i, X_{a,i}) = \mathbb{E}[T_i | C_i, Z_i, X_{a,i}], \quad (3.21)$$

Equations (3.20) and (3.21) represent a scenario in which one would treat Z as a normal covariate and is not aware that X_a does not cause T , Z and Y . In this case, the orthogonal score function would also be the same as in the backdoor adjustment scenario of (2.6).

4 Simulation Methodology

4.1 Data Generating Process

4.1.1 Cofounders

The cofounders are generated with multivariate normal distribution using a sparse covariance matrix Σ_d and a vector of means $\boldsymbol{\mu}_d = \mathbf{0}$.

$\Sigma_d \in \mathbb{R}^{d \times d}$, where $d := d_a + d_c$. d_a is the number of non-causal cofounders previously represented as X_a and d_c is the number of causal cofounders previously represented as C .

$$X \sim \mathcal{N}_d(\boldsymbol{\mu}_d, \Sigma_d) \quad (4.1)$$

The sparsity of the covariance matrix of the cofounders Σ_d is determined by α_d , which represents the probability that any covariance coefficient is zero. Larger α_d 's result in sparser covariance matrices. To provide intuition on it, we report the average absolute correlation $|\rho_d|$ in the matrix.

After generating the data from the multivariate normal distribution defined as $X \in \mathbb{R}^{n \times d}$ we randomly separate the features of X in X_a and C . Following the logic specified above, $X_a \in \mathbb{R}^{n \times d_a}$ and $C \in \mathbb{R}^{n \times d_c}$.

In (3.6) and (3.7) “Only partial inclusion of C_i and inclusion of $X_{a,i}$ on every nuisance function”, after generating cofounders, treatment, and target, we select a subset of the causal cofounders C_i^p to be used in the nuisance functions. This would represent a scenario of unobserved cofounders U_i . We define which causal cofounders are not included in the nuisance function randomly, where we use p_u , which represents the percentage of causal cofounders not included in the nuisance functions. Meaning that $C_i^p \in \mathbb{R}^{n \times d_{cp}}$, where $d_{cp} = d_c - \lceil p_u \cdot d_c \rceil$.

4.1.2 Treatment

The probability of treatment is generated from a logic:

$$\mathbb{P}[T_i] = \frac{1}{1 + e^{(-f(C_i) + \varepsilon_{t,i})}} \quad (4.2)$$

where $f(C_i)$ is a function of the causal cofounders C_i and $\varepsilon_{t,i}$ is a random error term generated from a normal distribution with mean 0 and variance σ_t .

$f(C_i)$ is a generic non-linear function explained in Section ??.

Treatment is later generated from a Bernoulli distribution with the probability of treatment from above in (4.2).

$$T_i \sim \text{Bernoulli}(\mathbb{P}[T_i]) \quad (4.3)$$

4.1.3 Treatment Effect

The true treatment effect is generated from:

$$\tau(C_i) = f(C_i) \quad (4.4)$$

Again, $f(C_i)$ is a generic non-linear function explained in Section 4.1.5.

Meaning that there is heterogeneity in the treatment effect across the causal confounders C_i .

The true ATE (Average Treatment Effect) is the average of the treatment effect across the population:

$$\tau = \mathbb{E}[\tau(C_i)] = \frac{1}{n} \sum_{i=1}^n \tau(C_i) \quad (4.5)$$

4.1.4 Outcome

The observed outcome is:

$$Y_i = Y_{i,0} + \tau(C_i) \cdot T_i + \varepsilon_{y,i} \quad (4.6)$$

where $Y_{i,0}$ is the potential outcome if the treatment T_i was not applied, $\tau(C_i)$ is the treatment effect, and $\varepsilon_{y,i}$ is a random error term generated from a normal distribution with mean 0 and variance σ_y .

$$Y_{i,0} = f(C_i) \quad (4.7)$$

Once more, $f(C_i)$ is a generic non-linear function explained in Section 4.1.5.

The potential outcome for treatment and control are:

$$Y_i(1) = Y_{i,0} + \tau(C_i), \quad Y_i(0) = Y_{i,0}, \quad (4.8)$$

which can not be directly observed.

4.1.5 $f(C_i)$ Non-Linear Transformation

The function $f(C_i)$, used to calculate $Y_{i,0}$ and $\mathbb{P}[T_i]$ is a generic function composed non-linear relationships.

More specifically $f(C_i)$ is the weighted sum of different “ q ” transformations in the causal confounders C_i :

$$f(C) = \sum_{j=1}^{d_c} (\beta_j \cdot q_j(C_j)) \quad (4.9)$$

Where C_j is the the vector of the j -th causal confounder, β_j is a random coefficient drawn from a uniform distribution $U(-1, 1)$, and $q_j(C_j)$ is a random transformation of the j -th causal confounder. More specifically, the random transformation $q_j(C_j)$ is drawn from the following set of possible transformations:

- i. Linear: $q_j(C_{i,j}) = C_{i,j}$
- ii. Quadratic: $q_j(C_{i,j}) = C_{i,j}^2$
- iii. Cubic: $q_j(C_{i,j}) = C_{i,j}^3$
- iv. Logarithmic: $q_j(C_{i,j}) = \log(|C_{i,j}| + 1)$
- v. Exponential: $q_j(C_{i,j}) = \exp(\frac{1}{5}C_{i,j})$
- vi. Sine: $q_j(C_{i,j}) = \sin(C_{i,j})$
- vii. Cosine: $q_j(C_{i,j}) = \cos(C_{i,j})$
- viii. Indicator: $q_j(C_{i,j}) = \mathbb{I}\{C_{i,j} > 0\} - \mathbb{I}\{C_{i,j} \leq 0\}$
- ix. Piecewise: $q_j(C_{i,j}) = 2\mathbb{I}\{C_{i,j} < 0\} + \mathbb{I}\{0 \leq C_{i,j} < 1\} + \frac{1}{2}\mathbb{I}\{C_{i,j} > 1\}$

For each simulation $f(C)$ is calculated with some of the specified transformation sets and β 's three times: one for the calculation of $Y_{i,0}$, one for the calculation of $\mathbb{P}[T_i]$, and one for the calculation of $\tau(C_i)$.

For some simulations, we only allow for linear transformations in $f(C)$, aiming to compare performance of the DML models in a simpler setting.

5 Appendix

5.1 Mild Regularity Conditions for Double Machine Learning

- i. Smoothness of the Target Function: The functions $m(X)$ and $g(X)$, representing the outcome and treatment models, should belong to a sufficiently smooth function class (e.g., Hölder class or Sobolev space). This ensures that they can be approximated well by machine learning methods.
- ii. Boundedness and Regularity: The outcome Y , treatment T , and covariates X should have bounded support or satisfy moment conditions, such as:

$$\mathbb{E}[|Y|^2] < \infty, \quad \mathbb{E}[|T|^2] < \infty.$$

Additionally, $m(X)$ and $g(X)$ should have bounded derivatives in some cases.

- iii. Sparsity or Complexity Control: The nuisance estimators $\hat{m}(X)$ and $\hat{g}(X)$ should satisfy complexity restrictions, such as sparsity in high-dimensional settings or controlled VC dimensions, to ensure valid estimation and inference.
- iv. Consistency and Rates of Convergence: The nuisance estimators $\hat{m}(X)$ and $\hat{g}(X)$ must converge to the true $m(X)$ and $g(X)$ at sufficiently fast rates (typically faster than $n^{-1/4}$ in terms of mean squared error).
- v. Independence or Weak Dependence: Observations should be independent and identically distributed (i.i.d.) or satisfy weak dependence conditions, such as mixing properties for time-series data.
- vi. Overlap Condition (Positivity): The propensity score $\pi(X) = \mathbb{P}(T = 1|X)$ must be bounded away from 0 and 1:

$$0 < c \leq \pi(X) \leq 1 - c \quad \text{for some } c > 0.$$
- vii. Orthogonality of the Moment Function: The estimating equation or moment function used in DML should satisfy a doubly-robust property, meaning that small estimation errors in $\hat{m}(X)$ and $\hat{g}(X)$ do not affect the asymptotic distribution of the estimator.
- viii. Cross-Fitting: Proper cross-fitting should be employed to ensure the orthogonality of the estimating equations and to avoid overfitting. This mitigates potential overfitting by using disjoint data splits for estimating nuisance parameters and evaluating the final moment function.