# Double/Debiased Machine Learning

Presented by: Eunkyu Seong

February 20, 2025

## 0. List of Contents

# List of Contents (2/3)

2. Neyman Orthogonality
   - Definition
   - How to construct it for M-estimators with infinite-dimensional nuisance parameter?
     - For pedagogy, we consider the finite-dimensional case
     - Concentrating out approach
   - Example: PLR
     - How the DML overcomes regularization bias and overfitting bias?

# List of Contents (3/3)

3. DML
   - Definition - algorithm
     - DML1 vs. DML2
   - Conditions for root-n convergence and asymptotic normality of the PLR-DML estimator.
4. Conclusion

# Motivation of the DML

**Example: Partially Linear Regression**

$$Y = D\theta_0 + g_0(X) + U, \quad D = m_0(X) + V \tag{1}$$

where $E[U|X, D] = 0$ and $E[V|X] = 0$.

- The first equation is the main equation, and $\theta_0$ is the main regression coefficient that we would like to infer. If $D$ is exogenous conditional on controls $X$, $\theta_0$ has the interpretation of the treatment effect parameter.
- The second equation is not of interest per se but it is important for characterizing and removing regularization bias.

# Regularization Bias (1/3)

- A naive approach to estimation of $\theta_0$ using ML methods would be to construct a sophisticated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function $D\theta_0 + g_0(X)$.
- To focus only on the regularization bias, we assume that $\hat{g}_0$ is estimated using the auxiliary sample (i.e., sample splitting).

**Naive Estimator**

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i)) \tag{2}$$

- The estimator $\hat{\theta}_0$ has a slower than $1/\sqrt{N}$ rate of convergence:

$$|\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{P} \infty. \tag{3}$$

**Decomposition of Estimation Error**

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{a}$$

$$+ \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{b} \quad (4)$$

where term $a$ is asymptotically normal

And term $b$ can be rewritten as:

$$b = (E[D_i^2])^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)) + o_P(1). \tag{5}$$

The summands have non-zero mean because, we must employ regularized estimators — such as lasso, ridge, boosting or penalized NNs. These induce substantive biases in the estimator $\hat{g}_0$ of $g_0$.

**Brief Explanation of DML**

- We estimate the functions $g_0(X)$ and $m_0(X)$ with ML estimators, separately. This is why we say *"double"*.
- There are two kinds of biases in the estimation of PLR: regularization bias and overfitting bias.
- DML overcomes these biases with:
  - *Neyman Orthogonality* and *Sample Splitting*.

# Neyman Orthogonality (1/3)

**Estimating Equation/Moment Condition**

We are interested in the true value $\theta_0$ of the low-dimensional target parameter $\theta$. We assume that $\theta_0$ satisfies the moment condition:

$$E_P[\psi(W; \theta_0, \eta_0)] = 0. \tag{6}$$

where $\psi$ is a vector of known score functions, and $W$ is a random element. $\eta_0$ is the true value of the nuisance parameter $\eta \in T$.

**An example for the score function $\psi$:**

$$Y = D\theta_0 + X'\beta_0 + U, \quad E_P[U(X', D)] = 0. \tag{7}$$

$$\partial_\theta \ell_\theta(W; \theta, \beta) = (Y - D\theta - X'\beta)D \tag{8}$$
$$\partial_\beta \ell_\beta(W; \theta, \beta) = (Y - D\theta - X'\beta)X.$$

The naive approach that we have explained is such that uses these scores and moment conditions for estimation.
Given the estimated $\hat{\beta}$, we are assuming that the moment condition holds. But this fails because of the regularization bias of $\hat{\beta}$.

**Gateaux derivative operator**
To introduce the condition, we define the pathwise derivative:

$$D_r[\eta - \eta_0] := \partial_r\{E_P[\psi(W;\theta_0, \eta_0 + r(\eta - \eta_0))]\}. \tag{9}$$

**Definition 2.1 (Neyman Orthogonality)** The score $\psi$ obeys the orthogonality condition at $\theta_0, \eta_0$ with respect to the nuisance realization set if:

$$\partial_\eta E_P[\psi(W;\theta_0, \eta_0)][\eta - \eta_0] = 0, \quad \forall \eta \in T_N. \tag{10}$$

Intuitively, small deviation from the true $\eta_0$, does not affect to estimation of $\theta_0$.

**How to Construct Neyman Orthogonal Scores?**
The approach is based on the concentrating-out method.
For all $\theta \in \Theta$, let $\beta_\theta$ be the solution of the following:

$$\max_{\beta \in B} E_P[\ell(W; \theta, \beta)]. \tag{11}$$

$\beta_\theta$ satisfies

$$\partial_\beta E_P[\ell(W; \theta, \beta_0)] = 0, \quad \text{for all } \theta \in \Theta.$$

# Construction of Orthogonal Scores (2/2)

Differentiating with respect to $\theta$ gives:

$$
\begin{aligned}
0 &= \partial_\theta \partial_\beta E_P[\ell(W;\theta,\beta_0)] = \partial_\beta \partial_\theta E_P[\ell(W;\theta,\beta_0)] && (12)\\
&= \partial_\beta E_P[\partial_\theta \ell(W;\theta,\beta_0) + [\partial_\theta \beta_0] \partial_\beta \ell(W;\theta,\beta_0)] && (13)\\
&= \partial_\beta E_P[\psi(W;\theta,\beta,\partial_\theta \beta_0)]\Big|_{\beta=\beta_0} && (14)
\end{aligned}
$$

where

$$
\psi(W;\theta,\beta,\partial_\theta \beta_0) := \partial_\theta \ell(W;\theta,\beta) + [\partial_\theta \beta_0]^T \partial_\beta \ell(W;\theta,\beta). \qquad (15)
$$

**Example: Partially Linear Regression**

Overcoming regularization biases using orthogonalization: Applying (15) to PLR we obtain

$$\psi(W;\theta,\beta_0) = (D - m_0(X)) \times (Y - D\theta - g_0(X)). \tag{16}$$

Then, if we use this score for estimation:

$$\hat{V} = D - \hat{m}_0(X). \tag{17}$$

$$\tilde{\theta}_0 = \left(\sum \hat{V}_i D_i\right)^{-1} \sum \hat{V}_i (Y_i - \hat{g}_0(X_i)). \tag{18}$$

Decomposing the estimation error:

$$\sqrt{n}(\tilde{\theta}_0 - \theta_0) = a^* + b^* + c^*. \tag{19}$$

# Example: Partially Linear Regression (2/4)

The first term, $a^*$ satisfies

$$a^* = E[VD]^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i U_i \sim N(0, \Sigma) \tag{20}$$

The second term captures the impact of regularization bias in estimating $g_0$ and $m_0$.

$$b^* = E[VD]^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} (\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i)) \tag{21}$$

This term is upper-bounded by $\sqrt{n} n^{-(\phi_m + \phi_g)}$, where $n^{-\phi_m}$ and $n^{-\phi_g}$ are respectively the rates of convergence of $\hat{m}_0$ and $\hat{g}_0$; this upper bound vanishes. Intuitively, even if each estimator converges slowly, the product of the two estimators can vanish fast enough.

# Example: Partially Linear Regression (3/4)

$$c^* = E[VD]^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i (g_0(X_i) - \hat{g}_0(X_i)) + \tag{22}$$

$$E[VD]^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} U_i (m_0(X_i) - \hat{m}_0(X_i)) \tag{23}$$

Let's give an attention to the following term in $c^*$:

$$\frac{1}{\sqrt{n}} \sum_{i \in I} V_i (\hat{g}_0(X_i) - g_0(X_i)). \tag{24}$$

# Example: Partially Linear Regression (4/4)

Conditioning on the auxiliary sample, we know that it has mean zero

$$E[E[V_i(\hat{g}_0(X_i) - g_0(X_i))|X_i, i \in I^c]] =$$
$$= E[(\hat{g}_0(X_i) - g_0(X_i))]E[V_i|X_i, i \in I^c] = 0.$$

and variance of order

$$\frac{1}{n}\sum_{i \in I}(\hat{g}_0(X_i) - g_0(X_i))^2 \xrightarrow{p} 0. \tag{25}$$

Thus this term vanishes in probability by Chebyshev's inequality. Without Sample Splitting the model error $V$ and the estimation error $\hat{g}_0(X_i) - g_0(X_i)$ are correlated, thus do not vanish.

# Simulation Results (1/2)

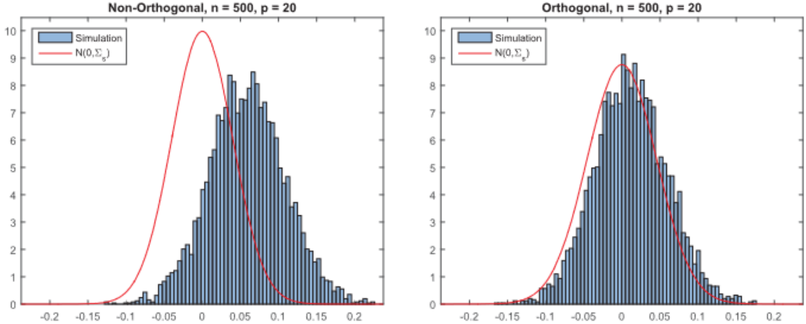


**Figure 1.** Comparison of the conventional and double ML estimators. [Colour figure can be viewed at wileyonlinelibrary.com]

# Simulation Results (2/2)

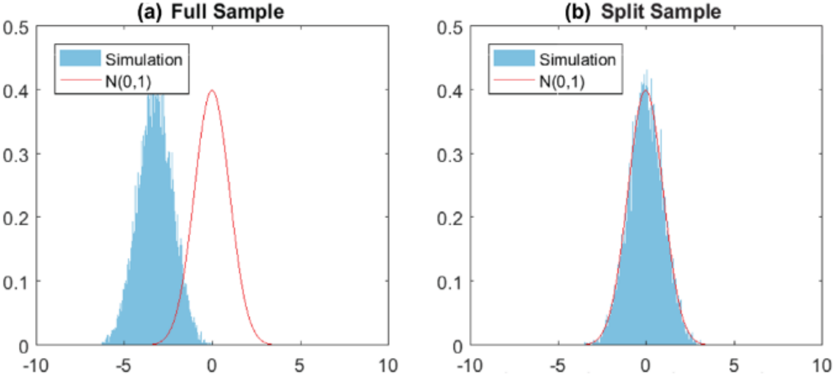


**Figure 2.** Comparison of full-sample and cross-fitting procedures. [Colour figure can be viewed at wileyonlinelibrary.com]

**Algorithm**

**Definition 3.1 (DML1)**

Take a $K$-fold random partition $(I_k)_{k=1}^K$ of observation indices such that:

- Each fold $I_k$ is of size $n = N/K$.
- Define $I_k^c := \{1, \ldots, N\} \setminus I_k$.
- Construct an ML estimator $\hat{\eta}_{0,k}$ using data indexed by $I_k^c$.
- Construct the estimator $\tilde{\theta}_{0,k}$ as the solution to:

$$E_{n,k}[\psi(W; \tilde{\theta}_{0,k}, \hat{\eta}_{0,k})] = 0. \tag{26}$$

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \tilde{\theta}_{0,k}. \tag{27}$$

# Double/Debiased Machine Learning (2/2)

**Definition 3.2 (DML2)**
Take a $K$-fold random partition $(I_k)_{k=1}^K$ of observation indices such that:

- Define $I_k^c := \{1, \ldots, N\} \setminus I_k$.
- Construct an ML estimator $\hat{\eta}_{0,k}$ using data indexed by $I_k^c$.
- Construct the estimator $\tilde{\theta}_0$ as the solution to:

$$\frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})] = 0. \tag{28}$$

**Assumption 4.1: Regularity Conditions for PLR Model**
Let $\mathcal{P}$ be the collection of probability laws $P$ for the triple $W = (Y, D, X)$ such that:

1. (4.1) and (4.2) hold;
2. $\|Y\|_{p,q} + \|D\|_{p,q} \leq C$;
3. $\|UV\|_{p,2} \geq c^2$ and $E_P[V^2] \geq c$;
4. $E_P[U^2 \mid X]_{p,\infty} \leq C$ and $E_P[V^2 \mid X]_{p,\infty} \leq C$;
5. Given a random subset $I$ of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^c})$ obeys the following condition for all $n \geq 1$. With $P$-probability no less than $1 - \Delta_N$:
    1. $\|\hat{\eta}_0 - \eta_0\|_{p,q} \leq C$, $\|\hat{\eta}_0 - \eta_0\|_{p,2} \leq \delta_N, \delta_N \geq N^{-1/2}$;
    2. For the score $\psi$ in (4.3), where $\hat{\eta}_0 = (\hat{g}_0, \hat{m}_0)$,

$$\|\hat{m}_0 - m_0\|_{p,2} \times \|\hat{g}_0 - g_0\|_{p,2} \leq \delta_N N^{-1/2}. \tag{29}$$

# DML Inference in PLR Model (2/3)

**Theorem 4.1: DML Inference on Regression Coefficients**
Suppose that Assumption 4.1 holds. Then the DML1 and DML2 estimators using the score in (4.3) are first-order equivalent and obey

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \to_d N(0,1),$$

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = E_P[V^2]^{-1} E_P[V^2 U^2] E_P[V^2]^{-1}$. Moreover, the result continues to hold if $\sigma^2$ is replaced by $\hat{\sigma}^2$ defined in Theorem 3.2

# DML Inference in PLR Model (3/3)

Suppose that Assumptions 3.1 and 3.2 hold. In addition, suppose that $\delta_N \geq N^{-[max\{(1-2/q),1/2\}]}$ for all $N \geq 1$. Consider the following estimator of the asymptotic variance matrix of $\sqrt{N}(\tilde{\theta}_0 - \theta_0)$:

$$\hat{\sigma}^2 = \hat{J}_0^{-1} \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})'](\hat{J}_0^{-1})'$$

where

$$\hat{J}_0 = \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi^a(W; \hat{\eta}_{0,k})]$$

This estimator concentrates around the true variance matrix $\sigma^2$.