DOTE 6635: Artificial Intelligence for Business Research

Double Machine Learning

Renyu (Philip) Zhang

1

Machine Learning for Causal Inference

- Using machine learning for causal inference is generally very challenging.
  - Cross-validation cannot be directly applied to hyper-parameter tuning for causal inference models (Athey and Imbens, 2016).
  - Good performance in predicting propensity scores or outcomes cannot be directly translated into good causal performance (Belloni et al., 2014).
  - Regularization in ML will introduce additional biases in causal inference (Belloni et al., 2016).
- What are the most critical issues in causal inference at large?
  - Confoundedness, non-overlapping, balance, etc.
  - AI/ML cannot magically solve these fundamental problems of causal inference.
- Double machine learning (DML) provides a framework to empower causal inference with ML.
  - Compared with other fields in business research, this is a very fast-evolving field of study.

Recursive partitioning for heterogeneous **causal** effects
S **Athey**, G **Imbens** - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… We refer to the estimators developed in this section as "**causal tree**" (CT) estimators. … for constructing **trees** for **causal** effects that allow us to do valid inference for the **causal** effects in …
☆ Save 99 Cite Cited by 2213 Related articles All 17 versions »

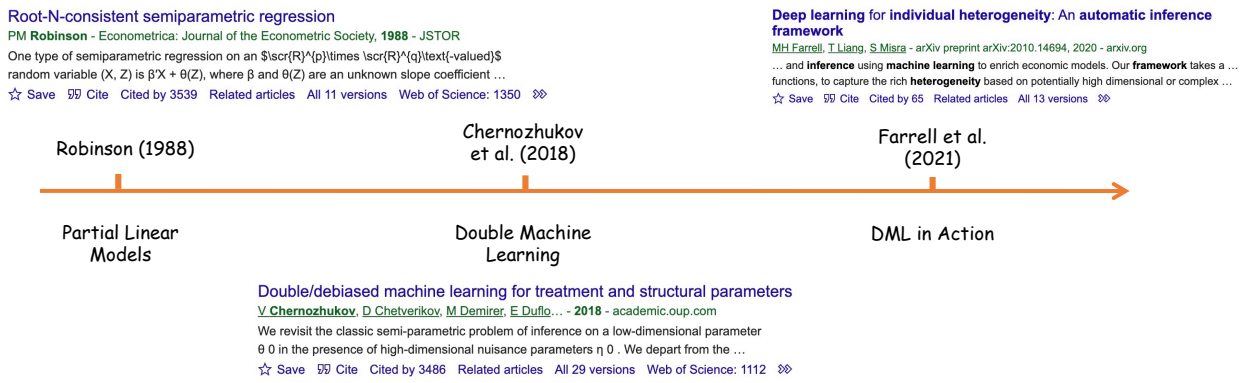
Inference on treatment effects after selection among high-dimensional controls
A **Belloni**, V **Chernozhukov**… - Review of Economic …, **2014** - academic.oup.com
We propose robust methods for inference about the effect of a treatment variable on a scalar outcome in the presence of very many regressors in a model with possibly non-Gaussian …
☆ Save 99 Cite Cited by 2040 Related articles All 32 versions »

Post-selection inference for generalized linear models with many controls
A **Belloni**, V **Chernozhukov**, Y Wei - Journal of Business & …, **2016** - Taylor & Francis
This article considers generalized linear models in the presence of many controls. We lay out a general methodology to estimate an effect of interest based on the construction of an …
☆ Save 99 Cite Cited by 233 Related articles All 11 versions Web of Science: 87 »

2

2

Today's Focus



Applied DML: A Historical Perspective

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

3

3

High-Level Takeaways of the DML Literature

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Provides a general framework, by leveraging Neyman Orthogonality, for estimating treatment effects using ML methods.
  - Causal inference usually requires estimating the. expected outcomes (and propensity scores) conditioned on covariates or confounders
  - Standard econometrics methods make strong functional form assumptions (e.g., linear models), which require strong substantive justifications; if mis-specified, causal estimates will be significantly biased.
  - DML framework automatically learns the form of conditional expectation functions from data.

- DML framework requires:
  - Some regularity conditions (recall the assumptions for AIPW)
  - ML estimators to converge, in RMSE/L2-norm at a rate of $o(n^{-1/4})$, slower than $o(n^{-1/2})$, the rate of most parametric models according to the delta method.

- DML framework outputs:
  - Root-n consistent estimators for treatment effects: Convergence to the ground-truth in probability at a rate $O(n^{-1/2})$, a property natural and common in a parametric world.
  - In frequentist perspective, root-n consistency typically means asymptotically normal, which means you can construct valid confidence intervals and do inference on your estimators.

4

4

Agenda

- Partial Linear Models
- General Double Machine Learning Framework

5

5

Let's First Look at a Simple Model

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- We use the partial linear model (PLM) to illustrate the DML framework:
  - Why ML is important and useful;
  - How the statistical theory works.

Partially Linear Model Set-up

- Y: Outcome
- D: Treatment
- X: Measured confounders
- U and V are our error terms
- We assume zero conditional mean:

$$Y = D\theta_0 + g_0(X) + U$$

$$D = m_0(X) + V$$

$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

6

6

Partial Linear Model

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

Partially Linear Model Set-up

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$

- Y: Outcome
- D: Treatment
- X: Measured confounders
- U and V are our error terms
- We assume zero conditional mean:

$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

- θ_0 is the parameter of interest.
- $g_0(.)$, the and $m_0(.)$ can take any arbitrary functional forms.
- X can be very high-dimensional, potentially dim(X) >> dim (D).
- Here, linearly additive separability is simply for illustration purposes.

7

7

Partial Linear Model

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- PLM: High-dimensional but observable confounders.
- If $g_0(.)$ and $m_0(.)$ are known, we can estimate θ_0 with a root-n consistent estimator.
- But $g_0(.)$ and $m_0(.)$ are unknown in practice, and this is where ML plays a crucial role.

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

- ML relaxes the linearity and additivity of $g_0(.)$ and $m_0(.)$, but ML alone is not sufficient (just like IPW is not sufficient).
- We can also think of PLM from the perspective of propensity scores:
  - $m_0(X)$ can be thought of as the propensity score.
  - If we can have an estimator $\hat{m}_0(.)$ for $m_0(.)$, can we construct an AIPW-type estimator for θ_0?

8

8

Partial Linear Model: Naïve Approach

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Two sources of biases: Regularization bias and overfitting bias.
- Regularization bias: ML model cannot predict the functions sufficiently fast.
  - Addressed by Neyman Orthogonality, or orthogonal moment conditions.
- Overfitting bias: The ML estimator may overfit the data with which it is trained.
  - Addressed by cross-fitting, i.e., splitting the data for training and inference.

Regularization bias. A naive approach to estimation of θ_0 using ML methods would be, for example, to construct a sophisticated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function $D\theta_0 + g_0(X)$.[2] Suppose, for the sake of clarity, that we randomly split the sample into two parts: a main part of size n, with observation numbers indexed by $i \in I$, and an auxiliary part of size $N - n$, with observations indexed by $i \in I^c$. For simplicity, we take $n = N/2$ for the moment and we turn to more general cases that cover unequal split-sizes, using more than one split, and achieving the same efficiency as if the full sample were used for estimating θ_0 in the formal development in Section 3. Suppose $\hat{g}_0$ is obtained using the auxiliary sample and that, given this $\hat{g}_0$, the final estimate of θ_0 is obtained using the main sample:

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

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

(1.3) ☹ $\Rightarrow |\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{p} \infty$

9

9

Regularization Bias of Naïve Approach

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

Regularization bias. A naive approach to estimation of θ_0 using ML methods would be, for example, to construct a sophisticated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function $D\theta_0 + g_0(X)$.[2] Suppose, for the sake of clarity, that we randomly split the sample into two parts: a main part of size n, with observation numbers indexed by $i \in I$, and an auxiliary part of size $N - n$, with observations indexed by $i \in I^c$. For simplicity, we take $n = N/2$ for the moment and we turn to more general cases that cover unequal split-sizes, using more than one split, and achieving the same efficiency as if the full sample were used for estimating θ_0 in the formal development in Section 3. Suppose $\hat{g}_0$ is obtained using the auxiliary sample and that, given this $\hat{g}_0$, the final estimate of θ_0 is obtained using the main sample:

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

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

(1.3) ☹ $\Rightarrow |\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{p} \infty$

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{:=b}$$

- Oracle estimator with known $g_0(.)$
- Asymptotically normal by linear regression or CLT.

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

If the RMSE of $\hat{g}_0(.)$ is of order $n^{-\varphi_g}$, where $\varphi_g < 1/2$, b will be of order $\sqrt{n}n^{-\varphi_g} \to \infty$.

10

10

Overcoming Regularization Bias

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

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

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{:=b}$$

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

- Orthogonality will help get ride of the bias in b.
  - Not a new idea, which was proposed in Robinson (1988) and Frisch-Waugh-Lovell (FWL) Theorem (1930).
  - Also proposed in the authors' prior work, Belloni et al. (2014).
- The basic idea is to train two ML models and double select out the biases, called double selection in Belloni et al. (2014).

Partial time regressions as compared with **individual trends**
R Frisch, FV Waugh - Econometrica: Journal of the Econometric Society, 1933 - JSTOR
... allowed to influence the **regression** coefficient between ... **partial time regression** method instead of the **individual trend** method. This is false. The possibility of determining the long **time** ...
☆ Save 99 Cite Cited by 766 Related articles All 4 versions »

Root-N-consistent semiparametric regression
PM **Robinson** - Econometrica: Journal of the Econometric Society, **1988** - JSTOR
One type of semiparametric regression on an $\scr{R}^{p}\times \scr{R}^{q}\text{-valued}$ random variable (X, Z) is β'X + θ(Z), where β and θ(Z) are an unknown slope coefficient ...
☆ Save 99 Cite Cited by 3539 Related articles All 11 versions Web of Science: 1350 »

Inference on treatment effects after selection among high-dimensional controls
A **Belloni**, V **Chernozhukov**... - Review of Economic ..., **2014** - academic.oup.com
We propose robust methods for inference about the effect of a treatment variable on a scalar outcome in the presence of very many regressors in a model with possibly non-Gaussian ...
☆ Save 99 Cite Cited by 2040 Related articles All 32 versions »

11

11

Frisch-Waugh-Lovell (FWL) Theorem

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Let's say we want to estimate the following model using OLS:
  - $Y = \beta_0 + \beta_1 D + \beta_2 X + U$
- The Frisch–Waugh–Lovell Theorem shows us that we can recover the OLS estimate of β_1 using a residuals-on-residuals OLS regression:
  1. Regress D on X using OLS
     - Let $\widehat{D}$ be the predicted values of D and let the residuals $\widehat{V} = D - \widehat{D}$
  2. Regress Y on X using OLS
     - Let $\widehat{Y}$ be the predicted values of Y and let the residuals $\widehat{W} = Y - \widehat{Y}$
  3. Regress $\widehat{W}$ on $\widehat{V}$ using OLS

→ Residual-on-residual regression.

- The estimated coefficient on $\widehat{V}$ will be the same as the estimated coefficient $\hat{\beta}_1$ from regressing Y on D and X using OLS!

For a proof of FWL Theorem, see: https://grok.com/share/bGVnYWN5_1a6a5098-2750-4ee3-8fb6-14cb5909a182

12

12

Robinson (1988)

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- The **Frisch–Waugh–Lovell** procedure:
  1. **Linear** regression of D on X
  2. **Linear** regression of Y on X
  3. **Linear** regression of the residuals from 2 on the residuals from 1
- Robinson's innovation: let's replace the **linear** regressions from 1 and 2 with some **non-parametric** regression
- **Robinson's** procedure:
  1. **Kernel** regression of D on X
  2. **Kernel** regression of Y on X
  3. **Linear** regression of the residuals from 2 on the residuals from 1

Neural nets, tree-based models, kernel regressions, etc.

13

13

Double Machine Learning (2018)

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- The idea proposed in Chernozhukov et al. (2018) is similar to Robinson (1988).

- DML using residuals-on-residuals regression:
  1. Estimate $D = \hat{m}_0(X) + \hat{V}$
  2. Estimate $Y = \hat{\ell}_0(X) + \hat{U}$
     - Note the absence of D and the switch from $g_0(\cdot)$ to $\ell_0(\cdot)$, which is essentially $E[Y \mid X]$
  3. Regress $\hat{U}$ on $\hat{V}$ using OLS for an estimate $\check{\theta}_0$

DML Estimator

$$\check{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I}\hat{V}_i\hat{V}_i\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(Y_i - \hat{\ell}_0(X_i)), \quad \ell_0(X) = E[Y|X]$$

- **Robinson's** procedure:
  1. Predict D with X using **kernel regression**
  2. Predict Y with X using **kernel regression**
  3. **Linear** regression of the residuals from 2 on the residuals from 1
- **DML residuals-on-residuals** procedure:
  1. Predict D with X using **any $n^{\frac{1}{4}}$-consistent ML model**
  2. Predict Y with X using **any $n^{\frac{1}{4}}$-consistent ML model**
  3. **Linear** regression of the residuals from 2 on the residuals from 1

Assumption: RMSE of ML models are of order $o(n^{-1/4})$.

14

14

DML: Debias using Orthogonalization

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Orthogonalization: Partialling out the effect of X from D.

Overcoming Regularization Biases using Orthogonalization. Now consider a second construction that employs an "orthogonalized" formulation obtained by directly partialling out the effect of X from D to obtain the orthogonalized regressor $V = D - m_0(X)$. Specifically, we obtain $\widehat{V} = D - \widehat{m}_0(X)$, where $\widehat{m}_0$ is an ML estimator of m_0 obtained using the auxiliary sample of observations. We are now solving an auxiliary prediction problem to estimate the conditional mean of D given X, so we are doing "double prediction" or "double machine learning".

After partialling the effect of X out from D and obtaining a preliminary estimate of g_0 from the auxiliary sample as before, we may formulate the following "debiased" machine learning estimator for θ_0 using the main sample of observations:[3]

$$\check{\theta}_0 = \Big(\frac{1}{n}\sum_{i\in I}\widehat{V}_i D_i\Big)^{-1}\frac{1}{n}\sum_{i\in I}\widehat{V}_i(Y_i - \widehat{g}_0(X_i)). \tag{1.5}$$

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

$$a^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} V_i U_i \rightsquigarrow N(0, \Sigma)$$

The remaining term c* converges to 0 if we split samples or cross-fit.

$$b^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}(\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i)) \leq \sqrt{n}n^{-(\varphi_m+\varphi_g)} \leq o(1)$$

15

15

DML vs. Naïve Plug-in

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

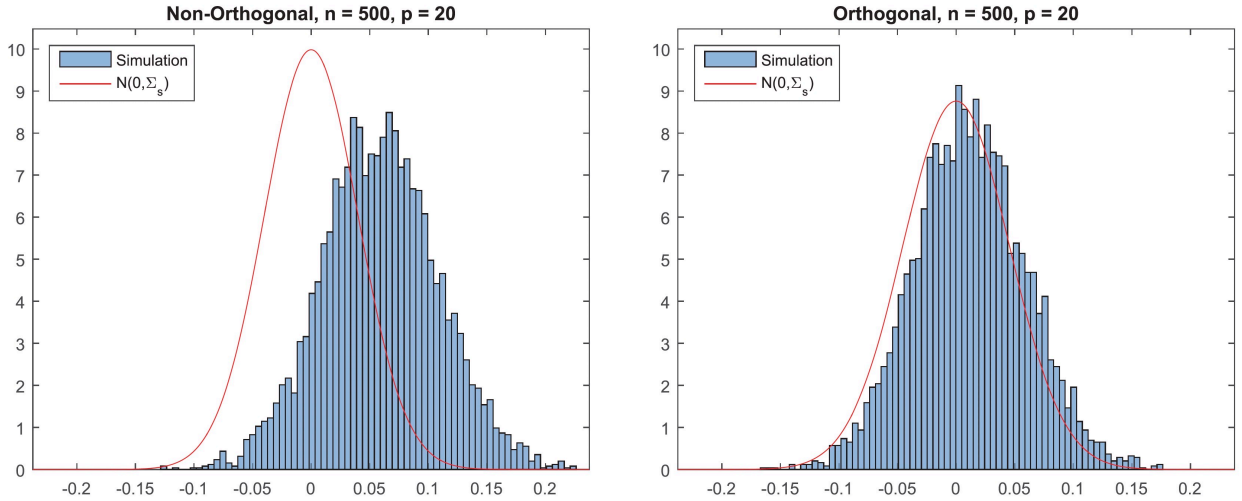


Figure Source: Chernozhukov et al. (2018), Figure 1.

16

16

Orthogonalization as Instrument

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

- Another way of thinking about orthogonalization is that you create an instrumental variable V on D that is uncorrelated with $Y - g_0(X)$ but correlated with D.
- DML estimator:

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

- Typical IV:

$$\hat{\beta}_{IV} = (Z'D)^{-1}Z'y$$

17

17

Overcoming Overfitting Bias via Sample Splitting

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

- Recall the overfitting bias.
- DML estimator: $\check{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I}\hat{V}_i D_i\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}(Y_i - \hat{g}_0(X_i))$

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

- We have shown that a* is asymptotically normal, and b* converges to 0 under common convergence rate assumptions for ML models.
- The third part c* contains the following term that only vanishes in probability if we use split samples to estimate $g_0(.)$:

$$\frac{1}{\sqrt{n}}\sum_{i\in I} V_i(\hat{g}_0(X_i) - g_0(X_i)) \qquad (1.6)$$

- With split samples, $V_i(\hat{g}_0(X_i) - g_0(X_i))$ has zero mean and $\frac{1}{n}\sum_{i\in I}(\hat{g}_0(X_i) - g_0(X_i))^2 \xrightarrow{p} 0$. (1.6) will vanish by the Chebyshev's Inequality.

18

18

Split Sample vs. Full Sample

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

ML Model:

$$\hat{g}_0(X_i) = g_0(X_i) + (Y_i - g_0(X_i))/N^{1/2-\epsilon}$$

Full Sample:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} V_i(\hat{g}_0(X_i) - g_0(X_i)) \propto N^{\epsilon} \to \infty$$

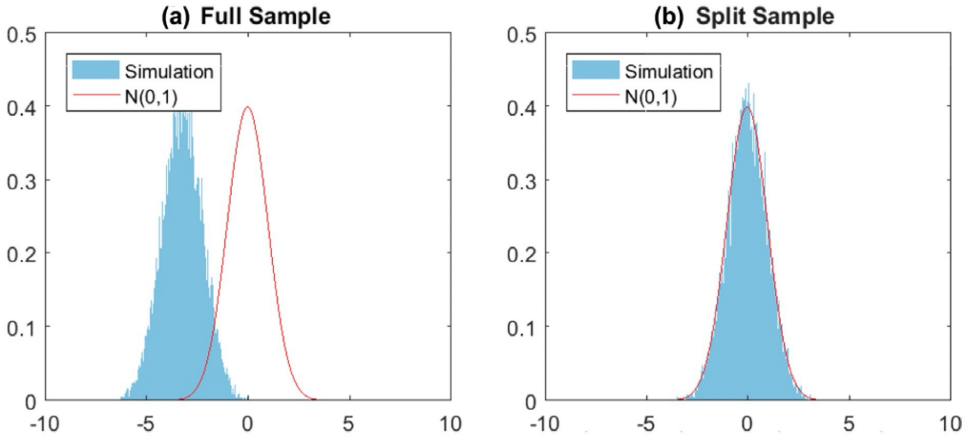


Figure Source: Chernozhukov et al. (2018), Figure 2.

19

19

Cross-Fitting to Improve Sample Efficiency

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Sample splitting reduces bias at the cost of compromised efficiency.
- Like cross-validation, we can use cross-fitting to improve sample efficiency:

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

1. Randomly partition your data into two subsets
2. Fit two ML models $\hat{g}_{0,1}$ and $\hat{m}_{0,1}$ in the first subset
3. Estimate $\check{\theta}_{0,1}$ in the second subset using the $\hat{g}_{0,1}$ and $\hat{m}_{0,1}$ functions we fit in the first subset
4. Fit two ML models $\hat{g}_{0,2}$ and $\hat{m}_{0,2}$ in the second subset
5. Estimate $\check{\theta}_{0,2}$ in the first subset using the $\hat{g}_{0,2}$ and $\hat{m}_{0,2}$ functions we fit in the second subset
6. Average our two estimates $\check{\theta}_{0,1}$ and $\check{\theta}_{0,2}$ for our final estimate $\check{\theta}_0$

→ Can be easily generalized to k-fold.

- How about the standard error and confidence interval?
- For PLM, we can just use the SE of stratified estimators; for general DML, we will talk about it later.

20

20

DML in Action

What, Why, and How: An Empiricist's Guide to Double/Debiased Machine Learning

Research Commentary

Bowen Shi
School of Economics and Management, Tsinghua University, sbw22@mails.tsinghua.edu.cn
Xiaojie Mao
School of Economics and Management, Tsinghua University, maoxj@sem.tsinghua.edu.cn
Mochen Yang
Carlson School of Management, University of Minnesota, yang3653@umn.edu
Bo Li
School of Economics and Management, Tsinghua University, libo@sem.tsinghua.edu.cn

This research commentary introduces Double/Debiased Machine Learning (DML), a novel methodological framework, to the Information Systems (IS) research community, demonstrating its power to address the challenges of empirical model specifications. DML combines the flexibility of modern machine learning (ML) techniques with the rigor of semiparametric statistical theory, enabling effective modeling of complex functions alongside valid statistical inference. The paper provides an accessible and comprehensive overview of DML's key elements—Neyman Orthogonality, cross-fitting, and high-quality ML estimation—and their roles in achieving methodological flexibility and rigor. The versatility of DML is illustrated through applications in several empirical settings common in IS research, including standard linear regression with control covariates, instrumental variable regressions, difference-in-differences, and scenarios with ML-generated covariates. Comparative simulations and real data analyses show that DML outperforms traditional parametric and semiparametric methods, and also illustrate the importance of DML's key elements. Finally, we highlight potential misconceptions and pitfalls in applying DML and offer practical advice for empirical researchers. Given the increasing complexity of data and research questions in the IS field, DML offers a timely and powerful tool for empirical researchers. By promoting a deeper understanding and appropriate use of DML, this commentary aims to empower empirical research in IS.

Key words: double/debiased machine learning, model misspecification, statistical inference, semiparametric model, empirical methods

An (IS) Empiricist's Guide to DML:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4677153

Promotion of DML in the IS community.

Demonstrations of DML in linear models: OLS, IV, DiD, etc.

Discussions of potential misconceptions and pitfalls in applying DML.

What, Why, and How: An Empiricist's Guide to Double/Debiased Machine Learning
B Shi, X Mao, M Yang, B Li - Debiased Machine Learning …, 2023 - papers.ssrn.com
… introduces Double/Debiased Machine Learning (DML), a … the flexibility of machine learning techniques with the rigor of … , and scenarios with machine learning-generated covariates. …
☆ 保存 99 引用 被引用次数：2 相关文章 »

21

21

DML to Estimate Labor Supply Elasticity

AER: Insights 2020, 2(1): 33–46
https://doi.org/10.1257/aeri.20180150

Monopsony in Online Labor Markets*

By Arindrajit Dube, Jeff Jacobs, Suresh Naidu, and Siddharth Suri*

Despite the seemingly low switching and search costs of on-demand labor markets like Amazon Mechanical Turk, we find substantial monopsony power, as measured by the elasticity of labor supply facing the requester (employer). We isolate plausibly exogenous variation in rewards using a double machine learning estimator applied to a large dataset of scraped MTurk tasks. We also reanalyze data from five MTurk experiments that randomized payments to obtain corresponding experimental estimates. Both approaches yield uniformly low labor supply elasticities, around 0.1, with little heterogeneity. Our results suggest monopsony might also be present even in putatively "thick" labor markets. (*JEL* C44, J22, J23, J42)

Monopsony in online labor markets
A Dube, J Jacobs, S Naidu, S Suri - American Economic Review …, 2020 - aeaweb.org
… **labor markets** like Amazon Mechanical Turk, we find substantial **monopsony** power, as measured by the elasticity of **labor** … Both approaches yield uniformly low **labor** supply elasticities, …
☆ Save 99 Cite Cited by 397 Related articles All 12 versions Web of Science: 95 »

Home > Information Systems Research > Vol. 35, No. 2 >
Mobile Payment Adoption: An Empirical Investigation of Alipay
Yuqian Xu, Anindya Ghose, Binqing Xiao
Published Online: 7 Jul 2023 | https://doi.org/10.1287/isre.2021.0156

https://www.aeaweb.org/articles?id=10.1257/aeri.20180150
https://pubsonline.informs.org/doi/10.1287/isre.2021.0156

Standard DML for partial linear models:

$$\ln(duration) = -\eta \ln(reward) + g_0(Z) + \epsilon, \quad E[\epsilon | Z, \ln(reward)] = 0. \tag{3}$$

$$\ln(reward) = m_0(Z) + \mu, \quad E[\mu | Z] = 0. \tag{4}$$

$$\check{\eta}^0 = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_i \hat{\xi}_i$$

$$\hat{\mu} = \ln(reward) - \hat{m}_0(Z)$$
$$\hat{\xi} = \ln(duration) - \hat{l}_0(Z)$$

Results are consistent with follow-up experiments.

Similar approach is applied to estimate the impact of Alipay on credit card adoption in a DiD specification (Xu et al., 2024).

$$Step\ 1: Outcome_{it} = g_0(\mathbf{Z}_{it}) + \tau_t + \epsilon_{it}, \tag{3}$$
$$Step\ 2: TreatGroup_i \times AfterTreat_t = m_0(\mathbf{Z}_{it}) + \tau_t + \mu_{it}, \tag{4}$$
$$Step\ 3: (Outcome_{it} - \hat{Outcome}_{it}) = \delta \times \mu_{it} + \gamma_{it}, \tag{5}$$

22

22

Evaluate Persuasive Power of Reputation

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 3, March 2024, pp. 1613–1634
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Influence via Ethos: On the Persuasive Power of Reputation in Deliberation Online

Emaad Manzoor,[a,*] George H. Chen,[b] Dokyun Lee,[c] Michael D. Smith[b]

[a] Cornell University, Ithaca, New York 14850; [b] Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [c] Boston University, Boston, Massachusetts 02215
*Corresponding author
Contact: emaadmanzoor@cornell.edu, https://orcid.org/0000-0003-3187-9719 (EM); georgechen@cmu.edu (GHC); dokyun@bu.edu, https://orcid.org/0000-0002-3186-3349 (DL); mds@cmu.edu, https://orcid.org/0000-0001-6844-7923 (MDS)

Received: May 31, 2020
Revised: December 13, 2021; May 12, 2022; August 15, 2022
Accepted: August 21, 2022
Published Online in Articles in Advance: May 8, 2023

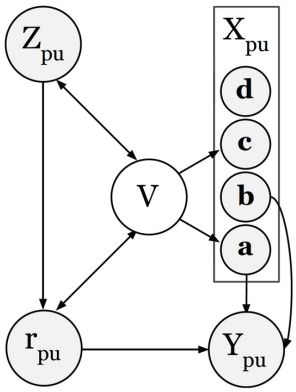
https://doi.org/10.1287/mnsc.2023.4762

Abstract. Deliberation among individuals online plays a key role in shaping the opinions that drive votes, purchases, donations, and other critical offline behavior. Yet, the determinants of opinion change via persuasion in deliberation online remain largely unexplored. Our research examines the persuasive power of *ethos*—an individual's "reputation"—using a seven-year panel of over a million debates from an argumentation platform containing explicit indicators of successful persuasion. We identify the causal effect of reputation on persuasion by constructing an instrument for reputation from a measure of past debate competition and by controlling for unstructured argument text using neural models of language in the double machine-learning framework. We find that an individual's reputation significantly impacts their persuasion rate above and beyond the validity, strength, and presentation of their arguments. In our setting, we find that having 10 additional reputation points causes a 31% increase in the probability of successful persuasion over the platform average. We also find that the impact of reputation is moderated by characteristics of the argument content, in a manner consistent with heuristic information processing under cognitive overload. We discuss managerial implications for platforms that facilitate deliberative decision making for public and private organizations online.

History: Accepted by Anandhi Bharadwaj, information systems.
Funding: This research was supported in part by the University of Wisconsin Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation, and by the Carnegie Mellon University Graduate Student Assembly and Provost's Office. Computing support was provided in part by the Social Science Computing Cooperative at the University of Wisconsin Madison.
Supplemental Material: Data and the online appendix is available at https://doi.org/10.1287/mnsc.2023.4762.

Keywords: persuasion • reputation systems • double machine-learning • causal inference with text

DML to evaluate the persuasive power of reputation:
https://pubsonline.informs.org/doi/10.1287/mnsc.2023.4762



learning (Chernozhukov et al. 2018). For our nuisance functions, we use neural networks with rectified linear unit (ReLU) activation functions (Nair and Hinton 2010), which converge at $n^{-1/4}$ rates (Farrell et al. 2018) and, thus, guarantee $\sqrt{n}$-consistent estimates of the causal effect of reputation.

Influence via ethos: On the persuasive power of reputation in deliberation online
E Manzoor, GH Chen, D Lee… - Management …, 2024 - pubsonline.informs.org
… of opinion change via persuasion in deliberation online remain largely unexplored. Our research examines the persuasive power of ethos—an individual's "reputation"—using a seven-…
☆ 保存 99 引用 被引用次数：20 相关文章 所有 7 个版本 Web of Science: 2 »

Is DML Credible in a Field Setting?

Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement

Brett R. Gordon,[a,*] Robert Moakler,[b] Florian Zettelmeyer[a,c]

[a] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; [b] Ads Research, Meta, Menlo Park, California 94025; [c] National Bureau of Economic Research, Cambridge, Massachusetts 02138
*Corresponding author
Contact: b-gordon@kellogg.northwestern.edu, https://orcid.org/0000-0001-9081-569X (BRG); rmoakler@meta.com (RM); f-zettelmeyer@kellogg.northwestern.edu (FZ)

Received: January 11, 2022
Revised: May 26, 2022; August 2, 2022
Accepted: August 16, 2022
Published Online in Articles in Advance: November 7, 2022

https://doi.org/10.1287/mksc.2022.1413

Abstract. Despite their popularity, randomized controlled trials (RCTs) are not always available for the purposes of advertising measurement. Non-experimental data are thus required. However, Facebook and other ad platforms use complex and evolving processes to select ads for users. Therefore, successful non-experimental approaches need to "undo" this selection. We analyze 663 large-scale experiments at Facebook to investigate whether this is possible with the data typically logged at large ad platforms. With access to over 5,000 user-level features, these data are richer than what most advertisers or their measurement partners can access. We investigate how accurately two non-experimental methods—double/debiased machine learning (DML) and stratified propensity score matching (SPSM)—can recover the experimental effects. Although DML performs better than SPSM, neither method performs well, even using flexible deep learning models to implement the propensity and outcome models. The median RCT lifts are 29%, 18%, and 5% for the upper, middle, and lower funnel outcomes, respectively. Using DML (SPSM), the median lift by funnel is 83% (173%), 58% (176%), and 24% (64%), respectively, indicating significant relative measurement errors. We further characterize the circumstances under which each method performs comparatively better. Overall, despite having access to large-scale experiments and rich user-level data, we are unable to reliably estimate an ad campaign's causal effect.

History: Olivier Toubia served as the senior editor for this article.
Funding: To be allowed to access the data required for this paper, B. R. Gordon and F. Zettelmeyer were part-time employees of Facebook with the title of Academic Researchers, employed for three hours per week. R. Moakler is an employee of Meta Platforms, Inc. and owns stock in the company.

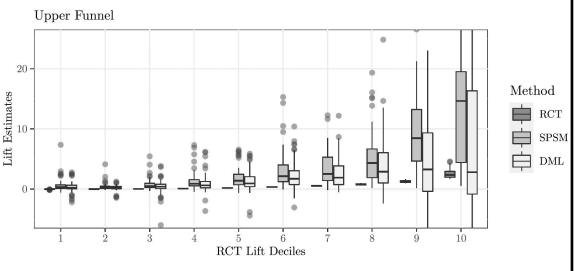
Keywords: digital advertising • field experiments • causal inference • observational methods • advertising measurement • double ML

2024 John D.C. Little Best Paper Award

DML vs. RCT to evaluate ad measurement on FB:
https://pubsonline.informs.org/doi/10.1287/mksc.2022.1413

DML and SPSC cannot recover RCT evaluation for advertising measurement.

Question: Which of the assumptions for DML are violated in the FB observational data?



Close enough? a large-scale exploration of non-experimental approaches to advertising measurement
BR Gordon, R Moakler, F Zettelmeyer - Marketing Science, 2023 - pubsonline.informs.org
… we can come "close enough" using the typical data stored on a large ad platform. To answer … These experiments were chosen to be representative of the large-scale experiments …
☆ 保存 99 引用 被引用次数：93 相关文章 所有 13 个版本 Web of Science: 15 »

How Credible is DML in a Field Setting?

Estimating Causal Effects with Double Machine Learning - A Method Evaluation

Jonathan Fuhr[1], Philipp Berens[2], and Dominik Papies[1]

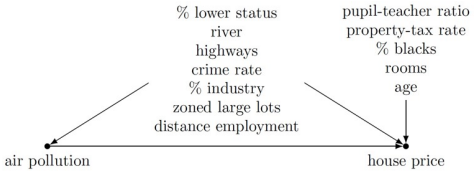
[1]*School of Business and Economics, University of Tübingen, Tübingen, Germany*
[2]*Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany*

Last edited: March 22, 2024

https://arxiv.org/abs/2403.14385

Abstract

The estimation of causal effects with observational data continues to be a very active research area. In recent years, researchers have developed new frameworks which use machine learning to relax classical assumptions necessary for the estimation of causal effects. In this paper, we review one of the most prominent methods - "double/debiased machine learning" (DML) - and empirically evaluate it by comparing its performance on simulated data relative to more traditional statistical methods, before applying it to real-world data. Our findings indicate that the application of a suitably flexible machine learning algorithm within DML improves the adjustment for various nonlinear confounding relationships. This advantage enables a departure from traditional functional form assumptions typically necessary in causal effect estimation. However, we demonstrate that the method continues to critically depend on standard assumptions about causal structure and identification. When estimating the effects of air pollution on housing prices in our application, we find that DML estimates are consistently larger than estimates of less flexible methods. From our overall results, we provide actionable recommendations for specific choices researchers must make when applying DML in practice.



DML estimates are consistently larger than estimates of less flexible methods (e.g., OLS).

Table 5: Results for the effect of air pollution on housing prices

Method	Effect estimate	Std. error	Effect at mean (%)	MSE(Y)	MSE(W)
OLS (H&R)	-0.0064	0.0011	-7.08	-	-
Simple OLS	-0.0146	0.0011	-16.17	-	-
XGBoost (naive)	-0.0137	0.0013	-15.14	-	-
OLS (raw)	-0.0058	0.0011	-6.47	-	-
OLS (flex)	-0.0071	0.0013	-7.88	-	-
OLS (DML, flex)	-0.0093	0.0006	-10.30	0.5571	1437.38
OLS (DML, raw)	-0.0059	0.0012	-6.58	0.0402	58.91
OLS (DML, H&R)	-0.0064	0.0012	-7.14	0.0375	54.88
GAMs (DML)	-0.0087	0.0015	-9.63	0.0346	42.71
Neural nets (DML)	-0.0081	0.0016	-9.00	0.0349	34.46
Lasso (DML, flex)	-0.0071	0.0015	-7.86	0.0316	33.89
XGBoost (DML)	-0.0070	0.0019	-7.73	0.0295	20.85
Random forests (DML)	-0.0075	0.0018	-8.27	0.0266	19.03

Note: MSE: mean squared error. H&R: covariate specification by Harrison and Rubinfeld (1978). raw: only using untransformed variables. flex: including squares and first-order interactions of all variables.

Hedonic housing prices and the demand for clean air
D Harrison Jr, DL Rubinfeld - Journal of environmental economics and …, 1978 - Elsevier
… the hedonic housing value function, p(h). The p(h) f unction translates a vector of housing attributes at each location into a price … of housing attributes .Q Implicit in this description of the …
☆ 保存 99 引用 被引用次数：2780 相关文章 所有 14 个版本 Web of Science: 1115 »

Estimating Causal Effects with Double Machine Learning--A Method Evaluation
J Fuhr, P Berens, D Papies - arXiv preprint arXiv:2403.14385, 2024 - arxiv.org
… necessary for the estimation of causal effects. In this paper, we review one of the most prominent methods - "double/debiased machine learning" (DML) - and empirically evaluate it by …
☆ 保存 99 引用 被引用次数：10 相关文章 所有 4 个版本 »

25

25

Agenda

- Partial Linear Models
- General Double Machine Learning Framework

26

26

A More General Framework

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Hopefully, we have developed the intuition on how DML helps remove bias and achieve faster convergence with the illustrative example, partial learning models.
- Let's now get into the real meat: The general framework of DML and Neyman Orthogonality.
- Recall what we have done to address regularization bias: Orthogonalize D w.r.t. X.
- Nuisance parameter: $\eta_0 = (g_0, m_0), \eta = (g, m)$
  - We do not really care about what g_0 and m_0 are.
  - We try to learn them to get good estimates for θ_0.
- Score function: The moment condition your estimator should satisfy, $\psi(.) = 0$.

$$\psi(W; \theta, \eta_0) = \underbrace{(D - m_0(X))}_{V} \times \underbrace{(Y - g_0(X) - (D - m_0(X))\theta)}_{U}$$

- Question: Why do we need this moment condition, score function = 0, to be satisfied?

27

27

A More General Framework

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

$$\psi(W; \theta, \eta_0) = (D - m_0(X)) \times (Y - g_0(X) - (D - m_0(X))\theta) = 0$$

- Let's look at the original model:

$$Y = V\theta_0 + g_0(X) + U$$

$$Y = (D - m_0(X))\theta_0 + g_0(X) + U$$

- The score function basically partials out the effect of X on D for the regression to obtain causal estimates.
  - The regressor is orthogonal to error.

  - $V = (D - m_0(X))$ is our regressor
  - $U = (Y - g_0(X) - (D - m_0(X))\theta)$ is our error term

28

28

Neyman Orthogonality

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Neyman Orthogonality: the score function should be robust w.r.t. small perturbations in estimation at the true value of the nuisance parameters.
- The Gateaux derivative of the score function evaluated at the true nuisance parameter η_0 is 0.
  - Gateaux derivative: https://grok.com/share/bGVnYWN5_d7a3da5b-65df-4962-8058-07b79ebf49f0

- The Neyman orthogonality condition:

$$\mathrm{D} = \partial_\eta \mathrm{E}\psi(W, \theta_0, \eta)|_{\eta=\eta_0} = \mathbf{0}$$

- Question: Why does Neyman orthogonality solves the regularization bias?

29

29

Neyman Orthogonality

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Here's, informally, why Neyman orthogonality solves the regularization bias:

- We have expansion

$$J\sqrt{n}(\hat{\theta} - \theta_0) = A_n + \sqrt{n}\mathrm{D}(\hat{\eta} - \eta_0) + C\sqrt{n}O(\|\hat{\eta} - \eta_0\|^2) + o_p(1),$$

  where the leading term A_n is well-behaved and approximately Gaussian under weak conditions, if sample-splitting is used and $\|\hat{\eta} - \eta_0\| \to 0$.
- When $\mathrm{D} \neq 0$, since $\|\hat{\eta} - \eta_0\| = O_P(n^{-\varphi})$, $0 < \varphi < 1/2$,

$$\sqrt{n}\mathrm{D}(\hat{\eta} - \eta_0) \text{ is of order } \sqrt{n}n^{-\varphi} \to \infty.$$

  and the estimator without Neyman orthogonality is not root-n consistent.

- Under Neyman orthogonality $\mathrm{D} = 0$, then

$$\sqrt{n}\mathrm{D}(\hat{\eta} - \eta) = 0,$$

  and for root-n consistency we only need,

$$C\sqrt{n}O(\|\hat{\eta} - \eta_0\|^2) \to 0,$$

  which requires $\|\hat{\eta} - \eta_0\| = o_P(n^{-1/4})$ if $C \gg 0$.

- Intuition: Under Neyman orthogonality, biases in estimating the nuisance parameters, will not ruin the moment condition, at least in the root-n consistency sense.

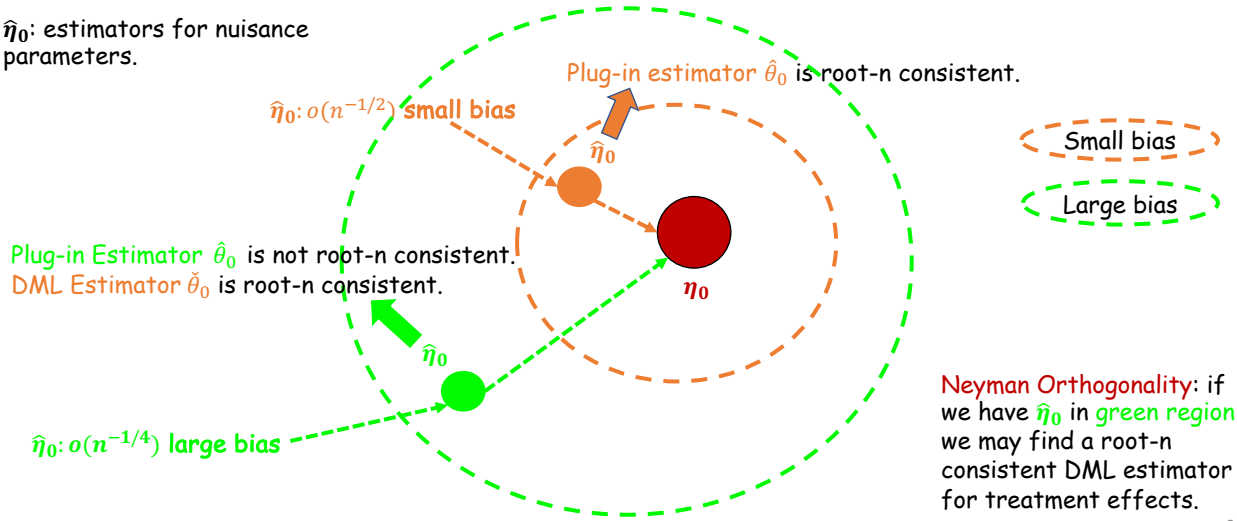
30

30

Neyman Orthogonality

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/torugf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

$\hat{\eta}_0$: estimators for nuisance parameters.



Neyman Orthogonality: if we have $\hat{\eta}_0$ in green region, we may find a root-n consistent DML estimator for treatment effects.

31

31

Double Machine Learning

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/torugf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Chernozhukov et al. (2018) prove that Neyman orthogonality is the correct condition for orthogonalizing the effect of X on D.
- We also leverage sample-splitting/cross-fitting to eliminate the overfitting bias.

1. Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = 1, ..., N$ such that the size of each fold I_k is $n = N/K$. Also, for each $k \in [K] = 1, ..., K$, define $I_k^c := 1, ..., N \setminus I_k$.
   - Create K equally sized partitions
   - I_k^c is the **complement** of I_k: if we have 100 observations and I_k is the set of observations 1–20, then I_k^c is the set of observations 21–100
2. For each $k \in [K]$, construct an ML estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

3. Construct the estimator $\tilde{\theta}_0$ ("theta-naught-tilde") as the solution to

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

   - Note that $\tilde{\theta}_0$ is **not** indexed by k, but the nuisance parameter $\hat{\eta}_{0,k}$ is
   - We're finding the $\tilde{\theta}_0$ that minimizes the average of the scores across all folds, where the scores vary by fold due to $\hat{\eta}_{0,k}$
   - This is a slightly different[2] version of the **cross-fitting** approach we talked about earlier that enables us to do sample splitting without loss of efficiency

32

32

Double Machine Learning

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Input/Condition:
  - Some regularity condition, such as unconfoundedness and overlapping.
  - ML convergence rate for the nuisance parameter at $o(n^{-1/4})$, **slower than** $o(n^{-1/2})$.
- Output:
  - Root-n consistent estimator, i.e., converging to the true estimand at rate $O(n^{-1/2})$, asymptotically normal for valid inference.
- The data generating process does not have to be a partial linear model.
- The power of the DML framework is that, as long as you can construct the Neyman orthogonal score function under these conditions, the statistical theory goes through.

33

33

What is Left?

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Construct the right Neyman orthogonal score functions to debias the estimates for different data generating processes.

Example: High-dimensional Linear Regression

EXAMPLE 2.1. (HIGH-DIMENSIONAL LINEAR REGRESSION) As an application of the construction above, consider the following linear predictive model,

$$Y = D\theta_0 + X'\beta_0 + U, \quad E_P[U(X', D)'] = 0, \tag{2.11}$$

$$D = X'\gamma_0 + V, \quad E_P[VX] = 0, \tag{2.12}$$

where, for simplicity, we assume that θ_0 is a scalar. The first equation here is the main predictive model, and the second equation only plays a role in the construction of the Neyman orthogonal scores. It is well known that θ_0 and β_0 in this model solve the optimization problem (2.5) with

$$\ell(W;\theta,\beta) = -\frac{(Y - D\theta - X'\beta)^2}{2}, \quad \theta \in \Theta = \mathbb{R}, \ \beta \in \mathcal{B} = \mathbb{R}^{d_\beta},$$

where we denote $W = (Y, D, X')'$. Hence, equations (2.6) hold with

$$\partial\ell_\theta(W;\theta,\beta) = (Y - D\theta - X'\beta)D, \quad \partial\ell_\beta(W;\theta,\beta) = (Y - D\theta - X'\beta)X,$$

and the matrix J satisfies

$$J_{\theta\beta} = -E_P[DX'], \quad J_{\beta\beta} = -E_P[XX'].$$

The Neyman orthogonal score is then given by

$$\psi(W;\theta,\eta) = (Y - D\theta - X'\beta)(D - \mu X); \qquad \eta = (\beta', \mathrm{vec}(\mu)')';$$

$$\psi(W;\theta_0,\eta_0) = U(D - \mu_0 X); \qquad \mu_0 = E_P[DX'](E_P[XX'])^{-1} = \gamma_0'. \tag{2.13}$$

If the vector of covariates X here is high-dimensional but the vectors of parameters β_0 and γ_0 are approximately sparse, we can use ℓ_1-penalized least-squares, ℓ_2-boosting, or forward selection methods to estimate β_0 and $\gamma_0 = \mu_0'$, and hence $\mu_0 = (\beta_0', \mathrm{vec}(\mu_0)')'$; see references cited in Section 1.

34

34

Beyond Partial Linear Model

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Let us go beyond patrial linear models and consider the general formulation:

$$Y = g_0(D, X) + U, \quad E_P[U \mid X, D] = 0,$$

$$D = m_0(X) + V, \quad E_P[V \mid X] = 0.$$

- Parameter of interest to estimate:
  - Average Treatment Effect (ATE)

$$\theta_0 = E_P[g_0(1, X) - g_0(0, X)]$$

  - Average Treatment Effect on Treated (ATT)

$$\theta_0 = E_P[g_0(1, X) - g_0(0, X) | D = 1]$$

35

35

Neyman Orthogonal Score Functions

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

Moment Condition

$$E_P[\psi(W; \theta_0, \eta_0)] = 0$$

Neyman Orthogonality Condition

$$\partial_\eta \mathrm{E}\psi(W, \theta_0, \eta)|_{\eta=\eta_0} = \mathbf{0}$$

$$Y = g_0(D, X) + U, \quad E_P[U \mid X, D] = 0,$$
$$D = m_0(X) + V, \quad E_P[V \mid X] = 0.$$

- Score function for ATE:

$$\psi(W; \theta, \eta) := (g(1, X) - g(0, X)) + \frac{D(Y - g(1, X))}{m(X)} - \frac{(1 - D)(Y - g(0, X))}{1 - m(X)} - \theta \qquad \text{(AIPW)}$$

$$\eta = (g, m) \qquad \eta_0 = (g_0, m_0)$$

- Score function for ATT:

$$\psi(W; \theta, \eta) = \frac{D(Y - \overline{g}(X))}{p} - \frac{m(X)(1 - D)(Y - \overline{g}(X))}{p(1 - m(X))} - \frac{D\theta}{p} \qquad \text{(AIPW for ATT)}$$

$$\eta = (\overline{g}, m, p) \quad \eta_0 = (\overline{g}_0, m_0, p_0) \quad \overline{g}_0(X) = g_0(0, X) \quad p_0 = E_P[D]$$

36

36

Bias Correction from Neyman Orthogonality

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Score function for LATE (i.e., IV estimates): $\theta_0 = \frac{E_P[\mu(1,X)] - E_P[\mu(0,X)]}{E_P[m(1,X)] - E_P[m(0,X)]}$

$$\psi(W;\theta,\eta) := \mu(1,X) - \mu(0,X) + \frac{Z(Y-\mu(1,X))}{p(X)} - \frac{(1-Z)(Y-\mu(1,X))}{1-p(X))}$$
$$- \left(m(1,X) - m(0,X) + \frac{Z(D-m(1,X))}{p(X)} - \frac{(1-Z)(D-m(0,X))}{1-p(X)} \right) \times \theta$$

$$\eta = (\mu, m, p) \qquad \eta_0 = (\mu_0, m_0, p_0)$$

$$Y = \mu_0(Z,X) + U, \quad E_P[U \mid Z, X] = 0,$$
$$D = m_0(Z,X) + V, \quad E_P[V \mid Z, X] = 0,$$
$$Z = p_0(X) + \zeta, \quad E_P[\zeta \mid X] = 0.$$

- Estimator from Neyman orthogonal score function can be viewed as the naïve plug-in estimator plus an IPW bias-correction term.

$$Y = g_0(D,X) + U, \quad E_P[U \mid X, D] = 0,$$
$$D = m_0(X) + V, \quad E_P[V \mid X] = 0.$$

$$\underbrace{(g(1,X) - g(0,X))}_{\text{Biased treatment effect estimate from ML models}} + \underbrace{\frac{D(Y-g(1,X))}{m(X)} - \frac{(1-D)(Y-g(0,X))}{1-m(X)}}_{\text{Debiasing terms}}$$

37

- AIPW = non-parametric regression + Neyman orthogonal debias
- General method to construct Neyman orthogonal score function:

The asymptotic variance of semiparametric estimators
WK Newey - Econometrica: Journal of the Econometric Society, 1994 - JSTOR
... formula for the asymptotic variance of a semiparametric estimator. A particularly ... estimates of nuisance functions. The general form of an adjustment factor for nonparametric estimates is ...
☆ 保存 99 引用 被引用次数：980 相关文章 所有 13 个版本 Web of Science: 319 »

37

DML Algorithm

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
https://www.norges-bank.no/contentassets/f2cc0752a45b4a5f8fe7eead30c0a49e/chernozhukov_slides.pdf?v=04102017101451
https://simons.berkeley.edu/sites/default/files/docs/10802/orthogonalml-slides.pdf
Applied Causal Inference Powered by ML and AI, Chapter 9: https://chapters.causalml-book.org/CausalML_book_2022.pdf

DEFINITION 3.1. (DML1) *(a) Take a K-fold random partition* $(I_k)_{k=1}^K$ *of observation indices* $[N] = \{1, \ldots, N\}$ *such that the size of each fold* I_k *is* $n = N/K$. *Also, for each* $k \in [K] = \{1, \ldots, K\}$, *define* $I_k^c := \{1, \ldots, N\} \setminus I_k$. *(b) For each* $k \in [K]$, *construct an ML estimator*

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

of η_0, *where* $\hat{\eta}_{0,k}$ *is a random element in* T, *and where randomness depends only on the subset of data indexed by* I_k^c. *(c) For each* $k \in [K]$, *construct the estimator* $\check{\theta}_{0,k}$ *as the solution of the following equation:*

$$\mathbb{E}_{n,k}[\psi(W; \check{\theta}_{0,k}, \hat{\eta}_{0,k}] = 0, \tag{3.1}$$

where ψ *is the Neyman orthogonal score, and* $E_{n,k}$ *is the empirical expectation over the kth fold of the data; that is,* $E_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$. *If achievement of exact 0 is not possible, define the estimator* $\check{\theta}_{0,k}$ *of* θ_0 *as an approximate* ϵ_N*-solution:*

$$\|E_{n,k}[\psi(W; \check{\theta}_{0,k}, \hat{\eta}_{0,k})]\| \leq \inf_{\theta \in \Theta} \|E_{n,k}[\psi(W; \theta, \hat{\eta}_{0,k})]\| + \epsilon_N, \quad \epsilon_N = o(\delta_N N^{-1/2}), \tag{3.2}$$

where $(\delta_N)_{N \geq 1}$ *is some sequence of positive constants converging to zero. (4) Aggregate the estimators:*

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \check{\theta}_{0,k}. \tag{3.3}$$

Variance for $\tilde{\theta}_0$:
$\hat{V}_{DML} = \frac{1}{N} \sum_{i=1}^{N} \hat{\psi}_i^2$

38

38

Applied Causal Inference Powered by ML and AI, Chapter 9:
https://chapters.causalml-book.org/CausalML_book_2022.pdf

DML Statistical Property

ASSUMPTION 5.1. (REGULARITY CONDITIONS FOR ATE AND ATTE ESTIMATION) *For all probability laws $P \in \mathcal{P}$ for the triple (Y, D, X) the following conditions hold: (a) equations (5.1) and (5.2) hold, with $D \in \{0, 1\}$; (b) $\|Y\|_{P,q} \leq C$; (c) $Pr_P(\varepsilon \leq m_0(X) \leq 1 - \varepsilon) = 1$; (d) $\|U\|_{P,2} \geq c$; (e) $\|E_P[U^2 \mid X]\|_{P,\infty} \leq C$; and (f) 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 conditions. With P-probability no less than $1 - \Delta_N$, $\|\hat{\eta}_0 - \eta_0\|_{P,q} \leq C$, $\|\hat{\eta}_0 - \eta_0\|_{P,2} \leq \delta_N$, $\|\hat{m}_0 - 1/2\|_{P,\infty} \leq 1/2 - \varepsilon$ and (i) for the score ψ in (5.3), where $\hat{\eta}_0 = (\hat{g}_0, \hat{m}_0)$ and the target parameter is ATE, $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{g}_0 - g_0\|_{P,2} \leq \delta_N N^{-1/2}$, and (ii) for the score ψ in (5.4), where $\hat{\eta}_0 = (\hat{\bar{g}}_0, \hat{m}_0, \hat{p}_0)$ and the target parameter is ATTE, $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{\bar{g}}_0 - \bar{g}_0\|_{P,2} \leq \delta_N N^{-1/2}$.*

THEOREM 5.1. (DML INFERENCE ON ATE AND ATTE) *Suppose that the target parameter is either ATE, $\theta_0 = E_P[g_0(1, X) - g_0(0, X)]$, and the score ψ in (5.3) is used, or ATTE, $\theta_0 = E_P[g_0(1, X) - g_0(0, X) \mid D = 1]$, and the score ψ in (5.4) is used. In addition, suppose that Assumption 5.1 holds. Then the DML1 and DML2 estimators $\tilde{\theta}_0$, constructed in Definitions 3.1 and 3.2, are first-order equivalent and obey*

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, 1), \tag{5.5}$$

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = E_P[\psi^2(W; \theta_0, \eta_0)]$. Moreover, the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2$ defined in Theorem 3.2. Consequently, confidence regions based upon the DML estimators $\tilde{\theta}_0$ have uniform asymptotic validity:

$$\lim_{N \to \infty} \sup_{P \in \mathcal{P}} |Pr_P(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1 - \alpha/2)\hat{\sigma}/\sqrt{N}]) - (1 - \alpha)| = 0.$$

The scores ψ in (5.3) and (5.4) are efficient, so both estimators are asymptotically efficient, reaching the semi-parametric efficiency bound of Hahn (1998).

SE for $\tilde{\theta}_0$:

$$\hat{\sigma}_{DML} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \hat{\psi}_i^2}$$

39

39

DML x Deep Learning

DNN for estimation and inference:
https://onlinelibrary.wiley.com/doi/full/10.3982/ECTA16901

Econometrica, Vol. 89, No. 1 (January, 2021), 181–213

DEEP NEURAL NETWORKS FOR ESTIMATION AND INFERENCE

MAX H. FARRELL
Booth School of Business, University of Chicago

TENGYUAN LIANG
Booth School of Business, University of Chicago

SANJOG MISRA
Booth School of Business, University of Chicago

We study deep neural networks and their use in semiparametric inference. We establish novel nonasymptotic high probability bounds for deep feedforward neural nets. These deliver rates of convergence that are sufficiently fast (in some cases minimax optimal) to allow us to establish valid second-step inference after first-step estimation with deep learning, a result also new to the literature. Our nonasymptotic high probability bounds, and the subsequent semiparametric inference, treat the current standard architecture: fully connected feedforward neural networks (multilayer perceptrons), with the now-common rectified linear unit activation function, unbounded weights, and a depth explicitly diverging with the sample size. We discuss other architectures as well, including fixed-width, very deep networks. We establish the nonasymptotic bounds for these deep nets for a general class of nonparametric regression-type loss functions, which includes as special cases least squares, logistic regression, and other generalized linear models. We then apply our theory to develop semiparametric inference, focusing on causal parameters for concreteness, and demonstrate the effectiveness of deep learning with an empirical application to direct mail marketing.

KEYWORDS: Deep learning, neural networks, rectified linear unit, nonasymptotic bounds, convergence rates, semiparametric inference, treatment effects, program evaluation.

Adopt DL to learn the nuisance parameters in DML.

Prove the "sufficient convergence rate" of DNN estimators.

Construct Neyman orthogonal score functions based on the specific functional forms.

Derive root-n consistent estimators for causal effects.

Deep neural networks for estimation and inference
MH Farrell, T Liang, S Misra - Econometrica, 2021 - Wiley Online Library
... We study **deep neural networks** and their use in semiparametric **inference**. We establish ... probability bounds for nonparametric **estimation** using **deep neural networks** for a large class of ...
☆ Save 99 Cite Cited by 534 Related articles All 17 versions Web of Science: 147 »

40

40

DNN for estimation and inference:
https://onlinelibrary.wiley.com/doi/full/10.3982/ECTA16901

DNN Convergence

- MLP network:

$$\widehat{f}_{\mathrm{MLP}}(\boldsymbol{x}) = \boldsymbol{W}_L\boldsymbol{\sigma}\big(\cdots\boldsymbol{\sigma}\big(\boldsymbol{W}_3\boldsymbol{\sigma}\big(\boldsymbol{W}_2\boldsymbol{\sigma}\big(\boldsymbol{W}_1\boldsymbol{\sigma}(\boldsymbol{W}_0\boldsymbol{x}+\boldsymbol{b}_0)+\boldsymbol{b}_1\big)+\boldsymbol{b}_2\big)+\boldsymbol{b}_3\big)+\cdots\big)+\boldsymbol{b}_L.$$

In sum, for a user-chosen architecture $\mathcal{F}_{\mathrm{DNN}}$, encompassing the choices $\sigma(\cdot)$, U, L, W, and the graph structure, the final estimate is computed using observed samples $z_i = (y_i, \boldsymbol{x}_i')'$, $i = 1, 2, \ldots, n$, of $\boldsymbol{Z}$, by solving

$$\widehat{f}_{\mathrm{DNN}} \in \underset{\substack{f_\theta \in \mathcal{F}_{\mathrm{DNN}} \\ \|f_\theta\|_\infty \le 2M}}{\arg\min} \sum_{i=1}^{n} \ell(f, z_i). \tag{2.4}$$

- Convergence of DNN under smoothness assumptions:

THEOREM 1—Multilayer Perceptron: *Suppose Assumptions 1 and 2 hold. Let* $\widehat{f}_{\mathrm{MLP}}$ *be the deep MLP-ReLU network estimator defined by* (2.4)*, restricted to* $\mathcal{F}_{\mathrm{MLP}}$*, for a loss function obeying* (2.1)*, with width* $H \asymp n^{\frac{d}{2(\beta+d)}} \log^2 n$ *and depth* $L \asymp \log n$*. Then with probability at least* $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$*, for n large enough,*

(a) $\|\widehat{f}_{\mathrm{MLP}} - f_*\|^2_{L_2(X)} \le C \cdot \{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\}$ *and*

(b) $\mathbb{E}_n[(\widehat{f}_{\mathrm{MLP}} - f_*)^2] \le C \cdot \{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\}$,

for a constant $C > 0$ *independent of n, which may depend on d, M, and other fixed constants.*

Remark: β is the smoothness of f_*; the total number of parameters is $W = (d+1)H + (L-1)(H^2+H) + H + 1$, which is an astronomical number.

41

41

DNN for estimation and inference:
https://onlinelibrary.wiley.com/doi/full/10.3982/ECTA16901

Score Function

- The following Neyman orthogonal score function is proposed:

Our approach to inference follows the current literature and uses sample averages of the (uncentered) influence functions. This approach yields valid inference under weaker conditions on the first step estimates (Farrell (2015), Chernozhukov et al. (2018)). Hahn (1998) showed that the influence function for a single average potential outcome is given by $\psi_t(z) - \mathbb{E}[Y(t)]$, for $t \in \{0, 1\}$ and $z = (y, t, \boldsymbol{x}')'$, where $\psi_t(z) = \mathbb{1}\{T = t\}(y - \mu_t(\boldsymbol{x}))\mathbb{P}[T = t \mid \boldsymbol{X} = \boldsymbol{x}]^{-1} + \mu_t(\boldsymbol{x})$. We estimate the unknown functions with deep learning to form

$$\widehat{\psi}_t(z_i) = \frac{\mathbb{1}\{t_i = t\}(y_i - \widehat{\mu}_t(\boldsymbol{x}_i))}{\widehat{\mathbb{P}}[T = t \mid \boldsymbol{X} = \boldsymbol{x}_i]} + \widehat{\mu}_t(\boldsymbol{x}_i), \tag{3.1}$$

(AIPW)

where $\widehat{\mathbb{P}}[T = t \mid \boldsymbol{X} = \boldsymbol{x}_i] = \widehat{p}(\boldsymbol{x}_i)$ for $t = 1$ and $1 - \widehat{p}(\boldsymbol{x}_i)$ for $t = 0$. The final estimators of τ and $\pi(s)$ are obtained by taking appropriate linear combinations:

$$\widehat{\tau} = \mathbb{E}_n\big[\widehat{\psi}_1(z_i) - \widehat{\psi}_0(z_i)\big] \quad \text{and} \quad \widehat{\pi}(s) = \mathbb{E}_n\big[s(\boldsymbol{x}_i)\widehat{\psi}_1(z_i) + \big(1 - s(\boldsymbol{x}_i)\big)\widehat{\psi}_0(z_i)\big]. \tag{3.2}$$

- The estimators are shown to be root-n consistent and asymptotically normal:

$$\sqrt{n}\widehat{\Sigma}^{-1/2}\big(\widehat{\pi}(s) - \pi(s)\big) \xrightarrow{d} \mathcal{N}(0, 1),$$

$$\text{with } \widehat{\Sigma} = \mathbb{E}_n\Big[\big(s(\boldsymbol{x}_i)\widehat{\psi}_1(z_i) + \big(1 - s(\boldsymbol{x}_i)\big)\widehat{\psi}_0(z_i)\big)^2\Big] - \widehat{\pi}(s)^2$$

42

42

Semi-parametric DML x Deep Learning

Deep Learning for Individual Heterogeneity: An Automatic Inference Framework*

Max H. Farrell Tengyuan Liang Sanjog Misra

University of Chicago, Booth School of Business

July 27, 2021

Abstract

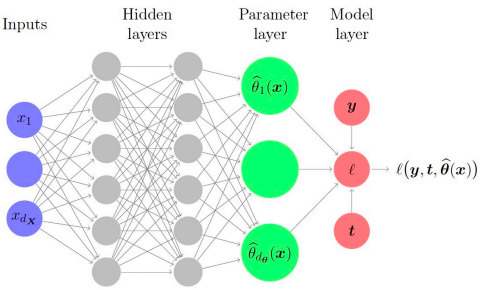
We develop methodology for estimation and inference using machine learning to enrich economic models. Our framework takes a standard economic model and recasts the parameters as fully flexible nonparametric functions, to capture the rich heterogeneity based on potentially high dimensional or complex observable characteristics. These "parameter functions" retain the interpretability, economic meaning, and discipline of classical parameters. In contrast to common implementations of machine learning in economics, these functions need not be predictions. We show that deep learning is particularly well-suited to structured modeling of heterogeneity in economics. First, we show how the network architecture can be easily designed to match the global structure of the economic model, delivering novel methodology that moves deep learning beyond prediction. Second, we prove convergence rates for the estimated parameter functions. These parameter functions are then the key input into the finite-dimensional parameter of inferential interest. We obtain valid inference based on a novel orthogonal score or influence function calculation that covers any second-stage parameter and any machine-learning-enriched model that uses a smooth per-observation loss function. No additional derivations are required and the score can be taken directly to data, using automatic differentiation if needed to obtain the components: the researcher need only define the original model and define the parameter of interest. A key insight is that we need not write down the influence function in order to evaluate it on the data. We apply this after deep learning, but our result can be used for any first-step estimator. Our framework covers, as special cases, well-known examples such as average treatment effects and partially linear models, but we also seamlessly deliver new results for such diverse examples as price elasticities, willingness-to-pay, and surplus measures in binary or multinomial choice models, average marginal and partial effects of continuous treatment variables, fractional outcome models, count data, heterogeneous production function components, and more. Across all these contexts inference can be made as automated as is currently available in special cases. We illustrate the utility of our framework with an application to a large scale advertising experiment for short-term loans. We show how economically meaningful estimates and inferences can be made that would be unavailable without our framework.

Keywords: Deep Learning, Influence Functions, Neyman Orthogonality, Heterogeneity, Structural Modeling, Semiparametric Inference

Automated Inference Framework:
https://arxiv.org/abs/2010.14694

More flexible treatment assignment, but less flexible (semiparametric) DGP.

DGP:



Neyman Orthogonal Score Function:

$$\textit{Neyman orthogonal score is } \psi(w,\theta,\Lambda)-\mu_0, \textit{ where } \quad \mu_0=\mathbb{E}\Big[H(X,\theta_0(X);t^*)\Big]$$

$$\psi(w,\theta,\Lambda)=H\left(x,\theta(x);t^*\right)-H_\theta(x,\theta(x);t^*)\Lambda(x)^{-1}\ell_\theta(y,t,\theta(x))$$

Deep learning for **individual heterogeneity**: An **automatic inference framework**
MH Farrell, T Liang, S Misra - arXiv preprint arXiv:2010.14694, 2020 - arxiv.org
... and **inference** using **machine learning** to enrich economic models. Our **framework** takes a ... functions, to capture the rich **heterogeneity** based on potentially high dimensional or complex ...
☆ Save 99 Cite Cited by 65 Related articles All 13 versions »

43

Empiricist's Guide to DML

Deep-Learning-Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence

Zikun Ye[1], Zhiqi Zhang[2], Dennis J. Zhang[2], Heng Zhang[3], Renyu Zhang[4]
[1] University of Washington, Seattle, WA
[2] Washington University in St. Louis, St. Louis, MO
[3] Arizona State University, Tempe, AZ
[4] The Chinese University of Hong Kong, Hong Kong, China
zikunye@uw.edu, z.zhiqi@wustl.edu, denniszhang@wustl.edu, hengzhang24@asu.edu, philipzhang@cuhk.edu.hk

Large-scale online platforms launch hundreds of randomized experiments (a.k.a. A/B tests) every day to iterate their operations and marketing strategies. The combinations of these treatments are typically not exhaustively tested, which triggers an important question of both academic and practical interest: Without observing the outcomes of all treatment combinations, how does one estimate the causal effect of any treatment combination and identify the optimal treatment combination? We develop a novel framework combining deep learning and doubly robust estimation to estimate the causal effect of any treatment combination for each user on the platform when observing only a small subset of treatment combinations. Our proposed framework (called debiased deep learning, DeDL) exploits Neyman orthogonality and combines interpretable and flexible structural layers in deep learning. We show theoretically that this framework yields efficient, consistent, and asymptotically normal estimators under mild assumptions, thus allowing for identifying the best treatment combination when observing only a few combinations. To empirically validate our method, we collaborated with a large-scale video-sharing platform and implemented our framework for three experiments involving three treatments where each combination of treatments is tested. When observing only a subset of treatment combinations, our DeDL approach significantly outperforms other benchmarks to accurately estimate and infer the average treatment effect of any treatment combination, and to identify the optimal treatment combination.

Key words: Deep Learning, Double Machine Learning, Causal Inference, Field Experiments, Experimentation on Online Platforms

DeDL Framework:
https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf

Operationalize and test the framework of Farrell et al. (2020) in a field setting with a large-scale RCT.

Provide a guideline for empiricists and practitioners how to do DML for causal inference.

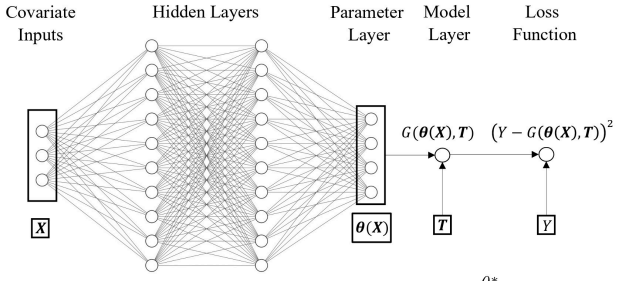
The $o(n^{-1/4})$ convergence rate assumption is impossible to verify in practice.

Deep-learning-based causal inference for **large-scale combinatorial experiments**: **Theory** and **empirical evidence**
Z Ye, Z Zhang, D Zhang, H Zhang... - Available at SSRN ..., 2023 - papers.ssrn.com
... interesting, we highlight our orthogonal-**experiment** data set is of **large scale** and high quality, delivering trustworthy **empirical evidence** on the performance of double machine learning ...
☆ Save 99 Cite Cited by 16 Related articles All 6 versions »

44

Empiricist's Guide to DML

DeDL Framework:
https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf



Step 1: Specify the DGP

$$\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\boldsymbol{t}]=G(\boldsymbol{\theta}^*(\boldsymbol{x}),\boldsymbol{t})$$

$$\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\boldsymbol{t}]=G(\boldsymbol{\theta}^*(\boldsymbol{x}),\boldsymbol{t})=\frac{\theta_4^*}{1+\exp(-(\theta_0^*(\boldsymbol{x})+\theta_1^*(\boldsymbol{x})t_1+\theta_2^*(\boldsymbol{x})t_2+\theta_3^*(\boldsymbol{x})t_3))}$$

Step 2: Train the DNN

$$\boldsymbol{\theta}^*(\cdot)\in\arg\min_{\boldsymbol{\theta}(\cdot)}\mathbb{E}[(Y-G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{T}))^2]$$

$$\hat{\boldsymbol{\theta}}(\cdot):=\underset{\boldsymbol{\theta}(\cdot)\in\mathcal{F}_{\text{DNN}}}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,\check{\boldsymbol{t}}_i,\boldsymbol{\theta}(\boldsymbol{x}_i)):=\frac{1}{n}\sum_{i=1}^{n}(y_i-G(\boldsymbol{\theta}(\boldsymbol{x}_i),\check{\boldsymbol{t}}_i))^2$$

Also need to show that:

- $\theta^*(.)$ can be nonparametrically identified in the DGP $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\boldsymbol{t}]=G(\boldsymbol{\theta}^*(\boldsymbol{x}),\boldsymbol{t})$.
- Justify the $o(n^{-1/4})$ convergence rate by citing Farrell et al. (2021).

45

45

Empiricist's Guide to DML

Step 3: Derive Neyman Orthogonal Score Function

$\boldsymbol{t}^*$ is the best treatment

DGP:

$$\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\boldsymbol{t}]=G(\boldsymbol{\theta}^*(\boldsymbol{x}),\boldsymbol{t})$$

other treatment combination. We first define the *advantage function* of the treatment combination $\boldsymbol{t}^1\in\{0,1\}^m$ over the treatment combination $\boldsymbol{t}^2\in\{0,1\}^m$ as

$$H(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t}^1,\boldsymbol{t}^2)\ :=\ G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}^1)-G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}^2).$$

Thus, the ground-truth ATE of any treatment combination $\boldsymbol{t}\in\{0,1\}^m$ can be written as

$$\mu(\boldsymbol{t})\ =\ \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t})]-\mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t}_0)]\ =\ \mathbb{E}[H(\boldsymbol{X},\boldsymbol{\theta}^*(\boldsymbol{X});\boldsymbol{t},\boldsymbol{t}_0)]. \quad (3)$$

$$\tau(\boldsymbol{t})\ :=\ \mu(\boldsymbol{t}^*)-\mu(\boldsymbol{t})\ =\ \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t}^*)]-\mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t})]\ =\ \mathbb{E}[H(\boldsymbol{X},\boldsymbol{\theta}^*(\boldsymbol{X});\boldsymbol{t}^*,\boldsymbol{t})]$$

Proposition 2 (Influence Function). *Suppose Assumptions 1, 2, 3 (in Appendix A.2), and 4 (in Appendix A.2) hold, then, the influence function for $\mu(\boldsymbol{t})$ is $\psi(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{t},\boldsymbol{t}_0)-\mu(\boldsymbol{t})$ with,*

$$\psi(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{t},\boldsymbol{t}_0)=H(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)-H_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)'\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\ell_{\boldsymbol{\theta}}(y,\check{\boldsymbol{t}},\boldsymbol{\theta}(\boldsymbol{x})), \quad (5)$$

where $\boldsymbol{z}=(y,\boldsymbol{x}',\check{\boldsymbol{t}}')'$ is observed data, $\boldsymbol{\Lambda}(\boldsymbol{x}):=2\mathbb{E}[G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{T})G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{T})'|\boldsymbol{X}=\boldsymbol{x}]$, $G_{\boldsymbol{\theta}}$ is gradient of G with respect to $\boldsymbol{\theta}$, $H_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0):=G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t})-G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}_0)$ is the gradient of H with respect to $\boldsymbol{\theta}$, and $\ell_{\boldsymbol{\theta}}(y,\check{\boldsymbol{t}},\boldsymbol{\theta}(\boldsymbol{x})):=2G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\check{\boldsymbol{t}})(G(\boldsymbol{\theta}(\boldsymbol{x}),\check{\boldsymbol{t}})-y)$ is gradient of ℓ with respect to $\boldsymbol{\theta}$.

Step 4: Estimation and Inference

Root-n Consistency

(a) For any treatment level $\boldsymbol{t}\in\{0,1\}^m$,

$$\sqrt{n}(\hat{\Psi}_{\text{DeDL}}(\boldsymbol{t};\mu))^{-1/2}(\hat{\mu}_{\text{DeDL}}(\boldsymbol{t})-\mu(\boldsymbol{t}))\to_d\mathcal{N}(0,1).$$

(b) Furthermore suppose the best treatment $\boldsymbol{t}^*:=\underset{\boldsymbol{t}\in\{0,1\}^m}{\arg\max}\,\mu(\boldsymbol{t})$ is unique. We have $\hat{\boldsymbol{t}}^*=\boldsymbol{t}^*$ with probability approaching one as the sample size goes to infinity, and for any treatment level $\boldsymbol{t}\in\{0,1\}^m$,

$$\sqrt{n}(\hat{\Psi}_{\text{DeDL}}(\boldsymbol{t};\tau))^{-1/2}(\hat{\tau}_{\text{DeDL}}(\boldsymbol{t})-\tau(\boldsymbol{t}))\to_d\mathcal{N}(0,1).$$

Algorithm 1 Implementing DeDL with Cross-fitting

1: (Cross-fitting) Split data samples into S nonoverlapping folds $\mathcal{S}_s$, $s=1,\ldots,S$.
2: (Training) For each fold s, use the complement of $\mathcal{S}_s$ to train DNN to get $\hat{\boldsymbol{\theta}}_s(\cdot)$ based on (2), and compute $\hat{\boldsymbol{\Lambda}}_s(\cdot)=2\mathbb{E}[G_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}),\boldsymbol{T})G_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}),\boldsymbol{T})'|\boldsymbol{X}=\boldsymbol{x}]$.
3: (ATE Estimation and Inference) For each $\boldsymbol{t}\in\{0,1\}^m$, leverage the influence function ψ and use data $\mathcal{S}$ to construct the ATE estimator $\hat{\mu}_{\text{DeDL}}(\boldsymbol{t})$ and variance estimator $\hat{\Psi}_{\text{DeDL}}(\boldsymbol{t};\mu)$. Conduct ATE inference based on $\hat{\mu}_{\text{DeDL}}(\boldsymbol{t})$ and $\hat{\Psi}_{\text{DeDL}}(\boldsymbol{t};\mu)$.
4: (Best Treatment Identification) Find empirical best treatment $\hat{\boldsymbol{t}}^*:=\arg\max_{\boldsymbol{t}}\hat{\mu}(\boldsymbol{t})$. Similarly, use influence function ψ and cross-fitting to construct estimators $\hat{\tau}_{\text{DeDL}}(\boldsymbol{t})$ and $\hat{\Psi}_{\text{DeDL}}(\boldsymbol{t};\tau)$ (see Appendix A.8) for the inference on best treatment identification.

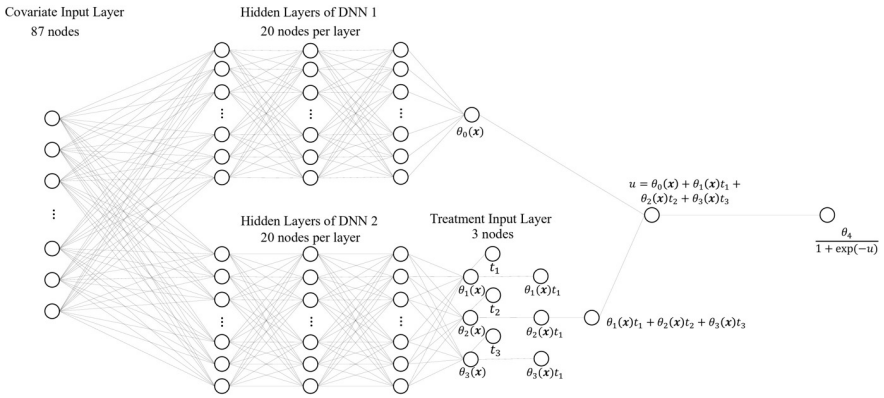
DeDL Framework: https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf 46

46

Empiricist's Guide to DML

DeDL Framework:
https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf

- What if the empirical correlation between covariates and treatment assignment deviates from the intended design?
  - Perform covariate balancing with stratified sampling.
  - Partition the covariate space into 69,111 strata, and then randomly sample the same number of users whose covariates lie within the stratum for each treatment assignment.
  - Essentially a sample reweighting/matching procedure.
- DNN for nuisance parameters should be built and trained with care.



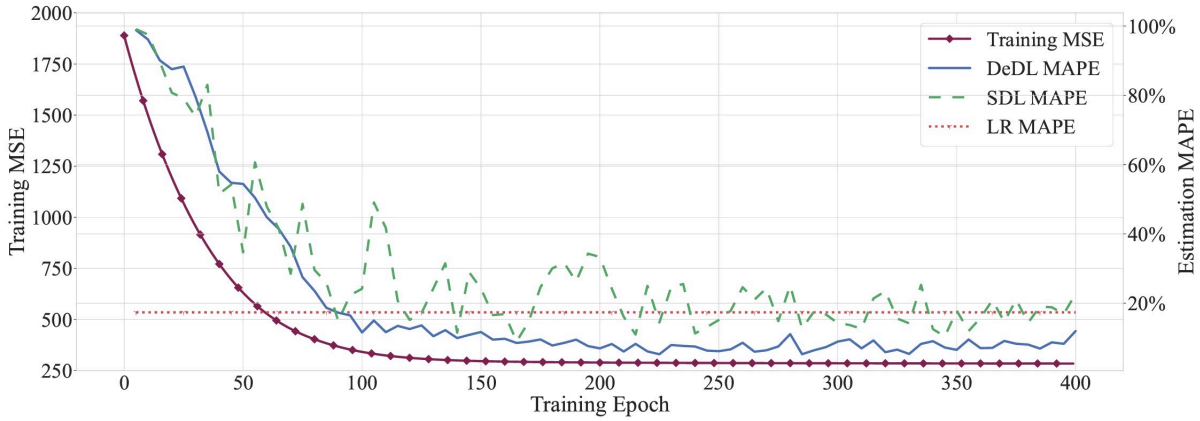
47

47

Empiricist's Guide to DML

DeDL Framework:
https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf

- Use DNN cross-validation error as the compass towards success.



48

48

DML Good and Bad News

DeDL Framework:
https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf

- Good news:
  - DML works better for more complex models and settings, i.e., more complex $G(.)$'s and a larger number of A/B tests.
  - Even if the DNN error $\mathbb{E}_X\|\widehat{\boldsymbol{\theta}}(X) - \boldsymbol{\theta}^*(X)\|_2$ is significant, DeDL still performs well as long as the link function $G(.)$ is correctly specified.

- Bad news:
  - If $G(.)$ is seriously misspecified, DeDL exacerbates the bad performances of the DNN estimators.
  - Try-and-error on $G(.)$, via the cross-validation error.
  - Generalize the framework to fully nonparametric settings (Chernozhukov et al. 2022) to automatically learn the debias term.

Parameter of Interest: $\theta_0 = \mathrm{E}[m(W, \gamma_0)]$

Conditional Mean: $\gamma_0(x) = \mathrm{E}[Y|X = x]$

Riesz Representer α_0:

$$\mathrm{E}[m(W, \gamma)] = \mathrm{E}[\alpha_0(X)\gamma(X)] \quad \text{for all } \gamma \text{ such that } \mathrm{E}[\gamma(X)^2] < \infty$$

Neyman Orthogonal Score Function:

$$\psi(w, \theta, \gamma, \alpha) = m(w, \gamma) - \theta + \alpha(x)[y - \gamma(x)]$$

Need to use base functions to estimate α_0:

$$\hat{\alpha}_\ell(x) = \frac{1}{n - n_\ell}\sum_{i \notin I_\ell} m(W_i, 1) + b(x)'\hat{\rho}_\ell,$$

$$\hat{\rho}_\ell = \arg\min_\rho \left\{-2\hat{M}_\ell'\rho + \rho'\hat{G}_\ell\rho + 2r\sum_{j=1}^{J}|\rho_j|\right\},$$

$$\hat{M}_\ell = \frac{1}{n - n_\ell}\sum_{i \notin I_\ell} m(W_i, b), \hat{G}_\ell = \frac{1}{n - n_\ell}\sum_{i \notin I_\ell} b(X_i)b(X_i)'$$

Automatic debiased machine learning of causal and structural effects
V Chernozhukov, WK Newey, R Singh - Econometrica, 2022 - Wiley Online Library
… to construct an **automatic debiased machine** learner (Auto-DML) of parameters of interest. We … **debiased machine learning** estimators for a wide variety of **effects**, including policy **effects**, …
☆ Save 99 Cite Cited by 161 Related articles All 13 versions Web of Science: 19 »

49

DML x DiD

https://academic.oup.com/restud/article/72/1/1/1581053
https://academic.oup.com/ectj/article/23/2/177/5722119

Econometrics Journal (2020), volume **23**, pp. 177–191.
doi: 10.1093/ectj/utaa001

Double/debiased machine learning for difference-in-differences models

NENG-CHIEH CHANG

*Department of Economics, University of California Los Angeles, 315 Portola Plaza, Los Angeles, CA 90095, USA.
Email: nengchiehchang@g.ucla.edu

First version received: 7 June 2019; final version accepted: 25 September 2019.

Summary: This paper provides an orthogonal extension of the semiparametric difference-in-differences estimator proposed in earlier literature. The proposed estimator enjoys the so-called Neyman orthogonality (Chernozhukov et al., 2018), and thus it allows researchers to flexibly use a rich set of machine learning methods in the first-step estimation. It is particularly useful when researchers confront a high-dimensional data set in which the number of potential control variables is larger than the sample size and the conventional nonparametric estimation methods, such as kernel and sieve estimators, do not apply. I apply this orthogonal difference-in-differences estimator to evaluate the effect of tariff reduction on corruption. The empirical results show that tariff reduction decreases corruption in large magnitude.

Keywords: *Difference-in-differences, high-dimensional data, causal inference, machine learning.*

JEL codes: *C1.*

Semiparametric difference-in-differences estimators
A Abadie - The review of economic studies, 2005 - academic.oup.com
The **difference-in-differences** (DID) **estimator** is one of the most popular tools for applied research in economics to evaluate the effects of public interventions and other treatments of …
☆ Save 99 Cite Cited by 3520 Related articles All 11 versions Web of Science: 1231 »

Double/debiased machine learning for difference-in-differences models
NC Chang - The Econometrics Journal, 2020 - academic.oup.com
… I apply this orthogonal **difference-in-differences** estimator to evaluate the effect of tariff reduction on corruption. The empirical results show that tariff reduction decreases corruption in …
☆ Save 99 Cite Cited by 81 Related articles All 5 versions Web of Science: 30 »

DGP: $Y_i(t) = \mu + X_i'\pi(t) + \tau \cdot D_i + \delta \cdot t + \alpha \cdot D_i(t) + \varepsilon_i(t)$

ATT: $\theta_0 \equiv E\left[Y_i^1(1) - Y_i^0(1) \mid D_i = 1\right]$

ASSUMPTION 2.1. $E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, D_i = 1\right] = E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, D_i = 0\right]$.

ASSUMPTION 2.2. *$P(D_i = 1) > 0$ and $P(D_i = 1|X_i) < 1$, with probability one.*

ASSUMPTION 2.3. *Conditional on $T = 0$, the data are independent and identically distributed from the distribution of (Y(0), D, X), and conditional on $T = 1$, the data are independent and identically distributed from the distribution of (Y(1), D, X).*

Case 1 (repeated outcomes): The new score function for repeated outcomes is

$$\psi_1(W, \theta_0, p_0, \eta_{10}) = \frac{Y(1) - Y(0)}{P(D=1)}\frac{D - P(D=1 \mid X)}{1 - P(D=1 \mid X)} - \theta_0 - \underbrace{\frac{D - P(D=1 \mid X)}{P(D=1)(1 - P(D=1 \mid X))}E[Y(1) - Y(0) \mid X, D=0]}_{c_1} \quad (3.1)$$

with the unknown constant $p_0 = P(D = 1)$ and the infinite-dimensional nuisance parameter

$$\eta_{10} = (P(D=1 \mid X), E[Y(1) - Y(0) \mid X, D=0]) \equiv (g_0, \ell_{10}).$$

Case 2 (repeated cross sections): The new score function for repeated cross sections is

$$\psi_2(W, \theta_0, p_0, \lambda_0, \eta_{20}) = \frac{T - \lambda_0}{\lambda_0(1 - \lambda_0)}\frac{Y}{P(D=1)}\frac{D - P(D=1 \mid X)}{1 - P(D=1 \mid X)} - \theta_0 - c_2, \quad (3.2)$$

where the adjustment term c_2 is

$$c_2 = \frac{D - P(D=1 \mid X)}{\lambda_0(1 - \lambda_0) \cdot P(D=1) \cdot (1 - P(D=1 \mid X))} \times E[(T - \lambda_0)Y \mid X, D=0].$$

The nuisance parameters are the unknown constants $p_0 = P(D = 1)$ and $\lambda_0 = P(T = 1)$ and the unknown function

$$\eta_{20} = (P(D=1 \mid X), E[(T - \lambda)Y \mid X, D=0]) \equiv (g_0, \ell_{20}).$$

50