Coordinated Double Machine Learning

Nitai Fingerhut
Technion

Matteo Sesia
USC

Yaniv Romano
Technion

ICML 2022

Code: https://github.com/nitaifingerhut/C-DML

Data Example – 401(k) dataset

Input i.i.d. data $(\mathrm{X_1}, \mathrm{D_1}, \mathrm{Y_1}), (\mathrm{X_2}, \mathrm{D_2}, \mathrm{Y_2}), \dots, (\mathrm{X_n}, \mathrm{D_n}, \mathrm{Y_n})$

- Features $\mathrm{X_i}$: individual's age, income, and other financial attributes
- Treatment $\mathrm{D_i}$: 401(k) eligibility
- Response $\mathrm{Y_i}$: net financial assets
- $n = 2{,}000$ subjects

Goal: provide an unbiased estimate of the treatment effect

Challenges: finite samples, features and treatment are correlated

Partially Linear Regression Model

Model assumption:

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

$$Y = D \cdot \theta + g(X) + U$$

- Want to estimate the **treatment effect** $\theta \in \mathbb{R}$
- The functions $m(\cdot), g(\cdot)$ link X to D and Y correspondingly

  → Can be ANY function! possibly nonlinear ones
- U and V are independent noise terms

Double Machine Learning: Approach Chernozhukov et al. (18)

Data $(X_i, D_i, Y_i)_{i=1}^{2n}$

Training data $\mathcal{T} = (X_i, D_i, Y_i)_{i=1}^{n}$

Holdout data $\mathcal{H} = (X_i, D_i, Y_i)_{i=n+1}^{2n}$

Train 2 independent models

$$\widehat{m} = \min_{m \in \mathcal{M}} \sum_{i \in \mathcal{T}} \big(D_i - m(X_i)\big)^2$$

$$\hat{\ell} = \min_{\ell \in \mathcal{L}} \sum_{i \in \mathcal{T}} \big(Y_i - \ell(X_i)\big)^2$$

Apply $\widehat{m}$ and $\hat{\ell}$ to holdout data

$$\widehat{V} = D - \widehat{m}(X)$$

$$\widehat{U} = Y - \hat{\ell}(X)$$

Estimate $\hat{\theta} = \left(\sum_{i \in \mathcal{H}} \widehat{V}_i(X)\right)^{-1} \sum_{i \in \mathcal{H}} \widehat{V}_i \widehat{U}_i$

Double Machine Learning: Theory Chernozhukov et al. (18)

DML estimation is $\sqrt{n}$-consistent: $|\theta - \hat{\theta}| \leq \mathcal{O}(n^{-1/2})$, if either of the two cond. is met

$\hat{m}(X)$ converges at rate $\mathcal{O}(n^{-\frac{1}{2}})$

Unlikely in practice

Both $\hat{m}(X)$ and $\hat{\ell}(X)$ converge at rate $\mathcal{O}(n^{-\frac{1}{4}})$

Weaker requirement

Our new finding: there is a **third** case where DML is $\sqrt{n}$- consistent

$\hat{m}(X)$ converges at rate $\mathcal{O}(n^{-\frac{1}{4}})$ and the residuals $\Delta\hat{m}(X), \Delta\hat{\ell}(X)$ are un-correlated

DML estimation bias: $|\theta - \hat{\theta}| \leq \frac{\mathbb{E}[\Delta\hat{m}(X)\cdot\Delta\hat{\ell}(X)] - \theta\cdot\mathbb{E}[(\Delta\hat{m}(X)|)^2]}{Var[V] + \mathbb{E}[(\Delta\hat{m}(X)|)^2]} + \mathcal{O}(n^{-1/2})$

Coordinated Double Machine Learning

Data $(X_i, D_i, Y_i)_{i=1}^{2n}$

Training data $\mathcal{T} = (X_i, D_i, Y_i)_{i=1}^{n}$

Holdout data $\mathcal{H} = (X_i, D_i, Y_i)_{i=n+1}^{2n}$

Train 2 independent models

$$\widehat{m} = \min_{m \in \mathcal{M}} \sum_{i \in \mathcal{T}} \left(D_i - m(x_i)\right)^2$$

$$\hat{\ell} = \min_{\ell \in \mathcal{L}} \sum_{i \in \mathcal{T}} \left(Y_i - \ell(x_i)\right)^2$$

Apply $\widehat{m}$ and $\hat{\ell}$ to holdout data

$$\hat{V} = D - \widehat{m}(X)$$

$$\hat{U} = Y - \hat{\ell}(X)$$

Estimate $\hat{\theta} = \left(\sum_{i \in \mathcal{H}} \hat{V}_i(X)\right)^{-1} \sum_{i \in \mathcal{H}} \hat{V}_i \hat{U}_i$

Coordinated Double Machine Learning

Data $(X_i, D_i, Y_i)_{i=1}^{2n}$

Training data $\mathcal{T} = (X_i, D_i, Y_i)_{i=1}^{n}$

Holdout data $\mathcal{H} = (X_i, D_i, Y_i)_{i=n+1}^{2n}$

Fit C-DML

Apply $\hat{m}$ and $\hat{\ell}$ to holdout data

$$\hat{V} = D - \hat{m}(X)$$
$$\hat{U} = Y - \hat{\ell}(X)$$

Estimate $\hat{\theta} = \left(\sum_{i\in\mathcal{H}} \hat{V}_i(X)\right)^{-1} \sum_{i\in\mathcal{H}} \hat{V}_i \hat{U}_i$

C-DML loss function

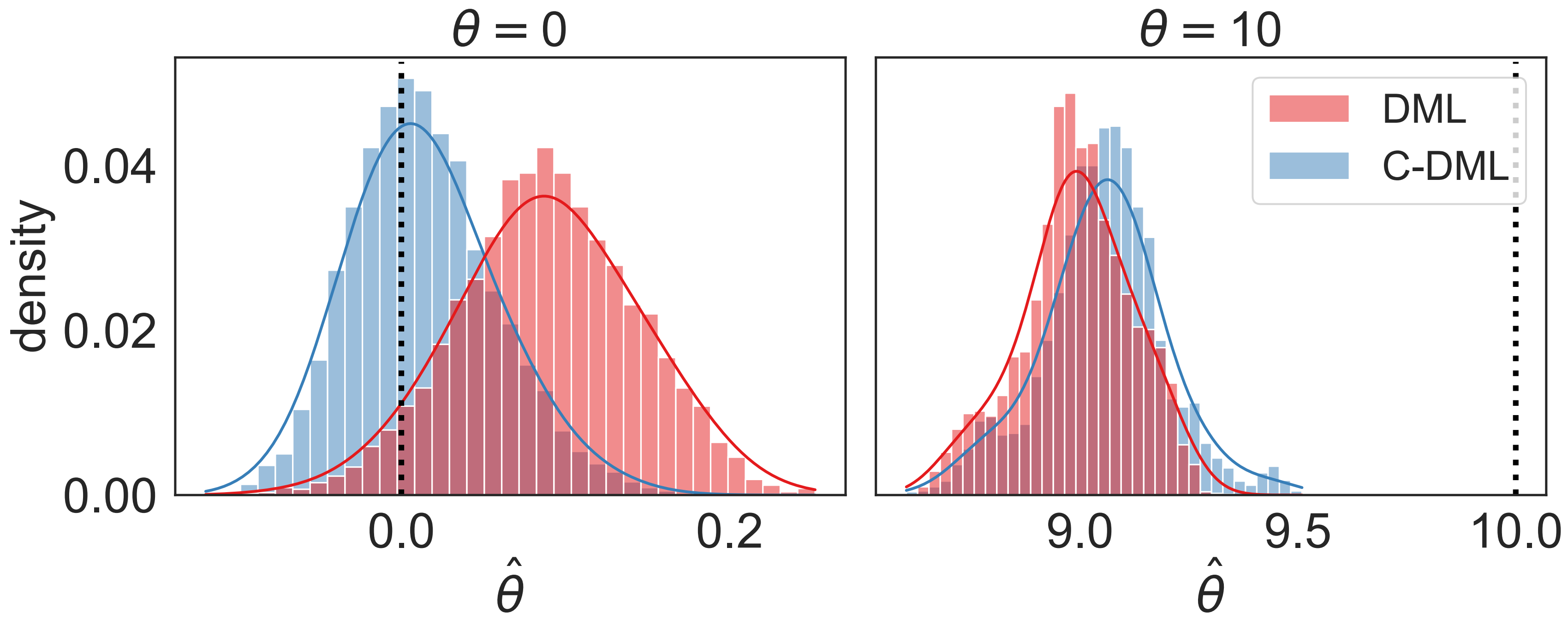
Train a **single model** with the following loss function

$$\min_{m\in\mathcal{M},\ell\in\mathcal{L}} \; \sum_{i\in\mathcal{T}}\big(D_i - m(X_i)\big)^2 + \sum_{i\in\mathcal{T}}\big(Y_i - \ell(X_i)\big)^2 +$$

$$\gamma \cdot \sum_{i\in\mathcal{T}}\left|(D_i - m(X_i)(Y_i - \ell(X_i))\right|$$

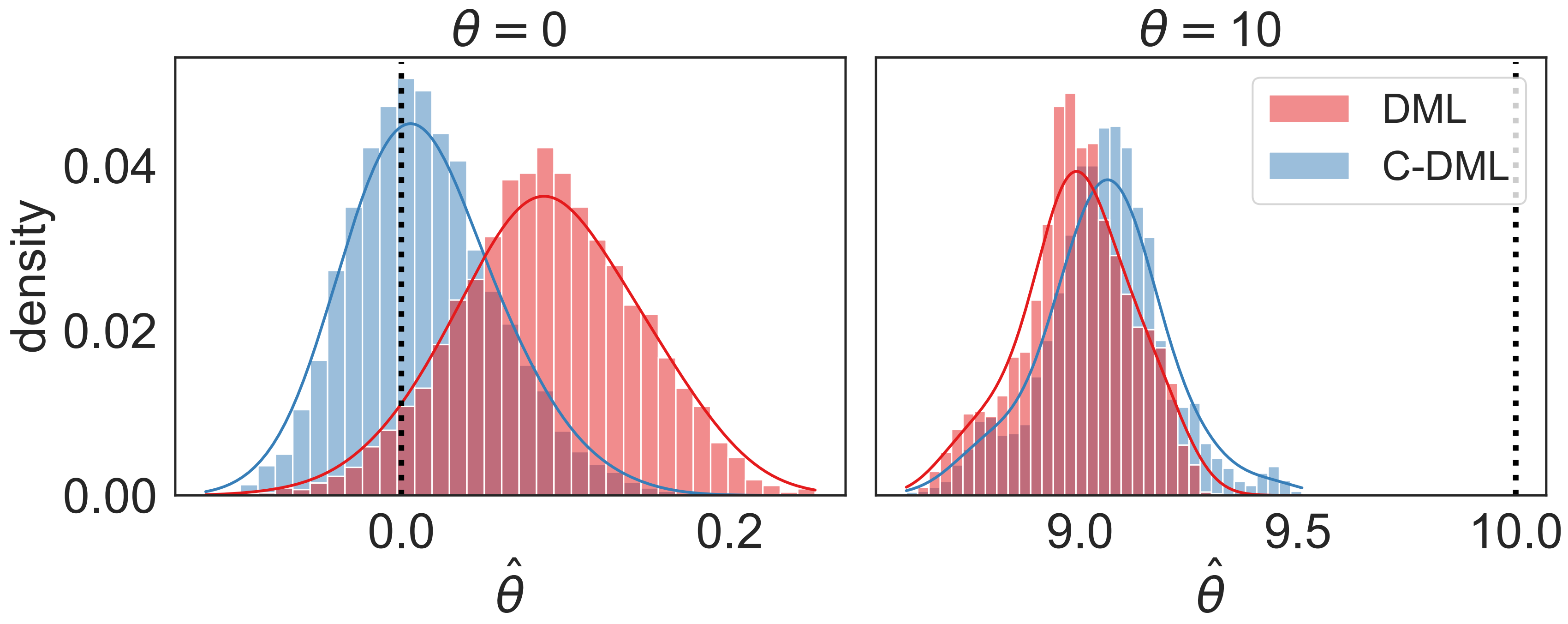
- C-DML seeks predictive models $\hat{m}(X)$ and $\hat{\ell}(X)$ with **uncorrelated residuals**
- Tuning the hyper-parameter γ via cross-validation – see paper

Results: 401(k) dataset



- The outcome is a <u>simulated</u> measure of net financial assets
- The treatment variable measures 401(k) eligibility
- 2000 examples; 5000 independent trials
- Use deep neural networks as base predictive models

Results: 401(k) dataset

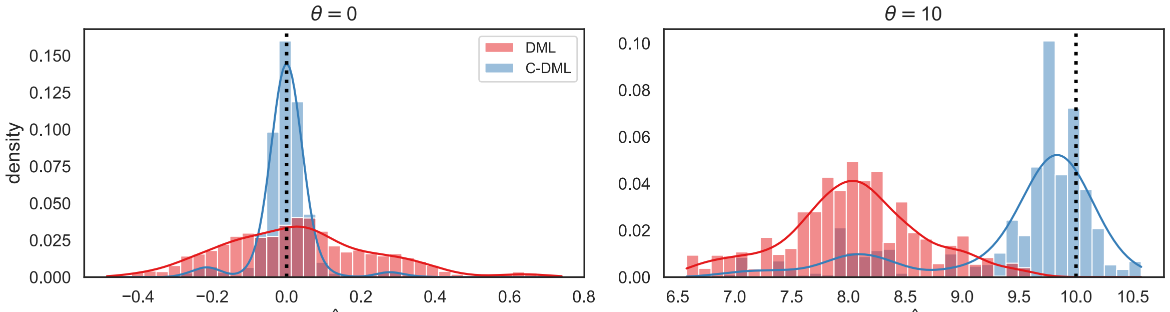


$$\mathbb{E}[\hat{\theta}_{DML} - \theta] = 0.0894$$
$$\mathbb{E}[\hat{\theta}_{C-DML} - \theta] = 0.0136$$

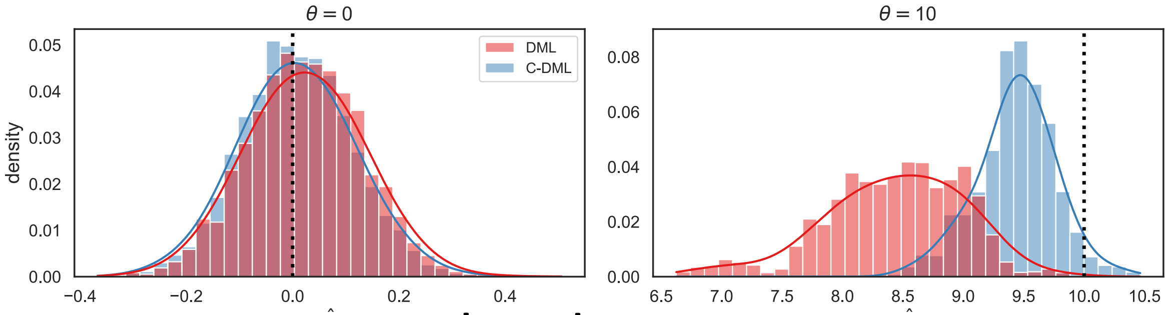
$$\mathbb{E}[\hat{\theta}_{DML} - \theta] = -1.0081$$
$$\mathbb{E}[\hat{\theta}_{C-DML} - \theta] = -0.947$$

C-DML's estimated $\hat{\theta}$ is closer to to the true θ

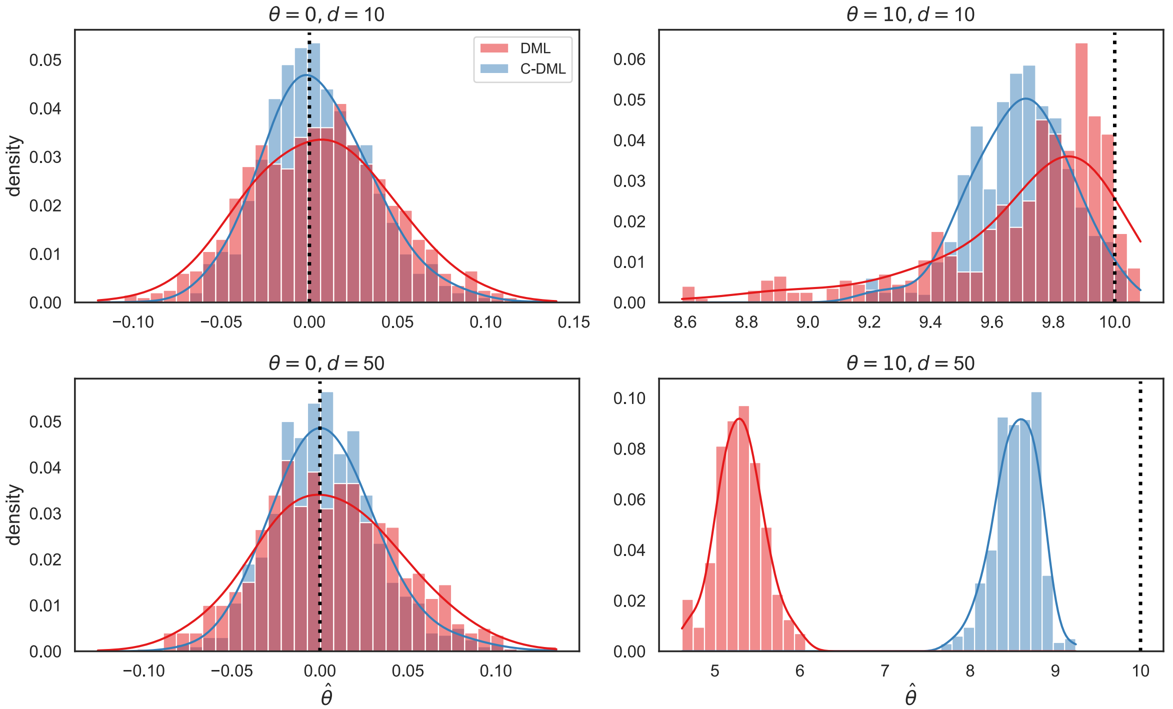
Thank you!



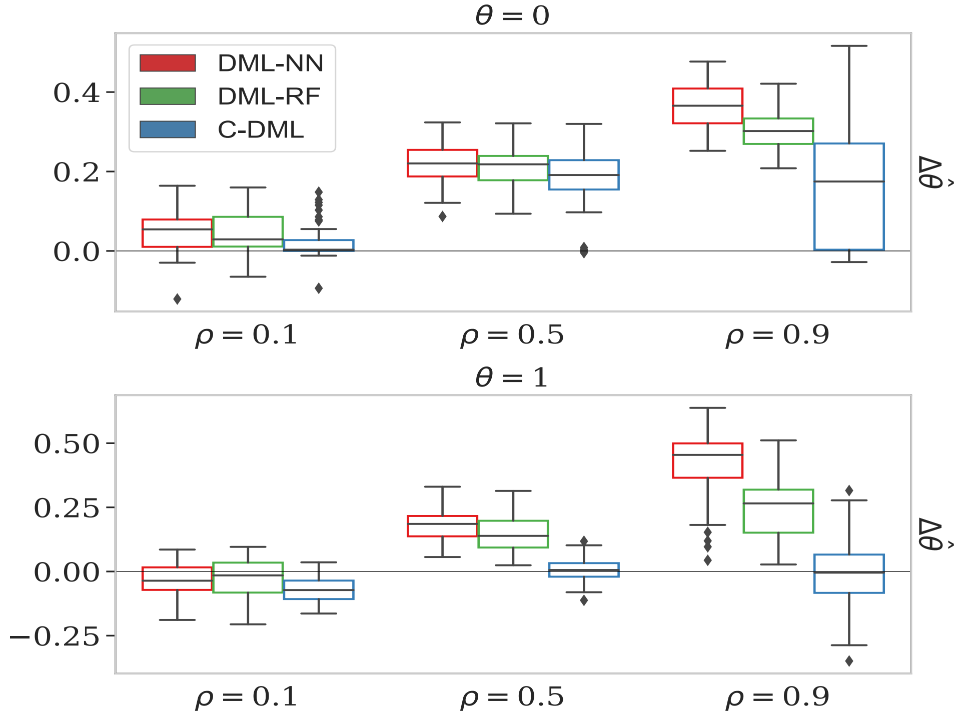
Beijing dataset



Blog dataset



CCDDHNR2018 dataset



Synthetic dataset