Cross-fitting-free DML with multiway dependence

Kaicheng Chen[1] Harold D. Chiang[2]

[1]School of Economics, SUFE

[2]Department of Economics, UW-Madison

College of Economics and Finance, Hangyang University

June 25, 2026

(Preliminary–comments welcome)

Abstract

- Under multiway dependence, we show that double/debiased machine learning (DML) estimator is asymptotically normal **without cross-fitting**.

- We extend arguments for i.i.d. sampling to the multiway dependence
  - $\Rightarrow$ requires controlling the empirical processes arising from learning nuisance parameter
  - $\Rightarrow$ we develop new maximal inequalities for empirical processes with multiway dependence that may be of independent interest.

Cluster dependence in real data environments

- Real economic data often exhibit complex correlation structures, making the i.i.d. assumption invalid.
- Common multi-way clustering dependencies include
  - Regional economic data: region–industry–time
  - Firm-level panel data: firm–market–year
  - International trade data: exporter–importer–time

Challenges for DML in real data environments

- Consequences
  - ML-based inference methods built on independence lack rigorous theoretical guarantees.
  - overfitting bias and over/under rejections
  - DML based on special cross-fitting algorithms are computationally costly, less accurate, too conservative, very specific about sampling assumptions.

Double Machine Learning

- CCDDHNR (2018) propose a general methodology for estimation & inference for two-step estimation problems

$$\textbf{DML} \approx \textbf{Neyman Orthogonal Score} + \textbf{Cross-Fitting}.$$

  - The former mitigates the bias coming from slow convergence rates of ML-based learning of nuisance parameters.
  - The latter simplifies the derivations and, according to conventional wisdom, may help reduce the bias induced by overfitting (lacks justification outside certain special cases)

A Generic DML Problem

- A DML problem defined by

$$\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0$$

  where

  - W: data r.v.
  - θ_0: (scalar) parameter of interest, exactly identified[1]
  - η_0: unknown nuisance parameter (potentially nonparametric/high-dimensional)
  - ψ: score, known up to the parameters, satisfies the *Neyman Orthogonality*

$$\partial_\eta \mathbb{E}[\psi(W,\theta_0,\eta_0)](\widetilde{\eta}) := \partial_t \mathbb{E}[\psi(W,\theta_0,\eta_0 + t\widetilde{\eta})]|_{t=0} = 0, \quad \text{for all } \widetilde{\eta} \in \Gamma$$

- An estimator $\widehat{\eta} = \widehat{\eta}(\{W_i\}_{i\in I})$ for η_0 is available

[1] Over-identification case is also studied in the paper.

Example 1: Partially Linear Model

- Consider the partially linear model (Robinson, 1988)

$$Y = \theta D + m_0(X) + \varepsilon, \quad \mathbb{E}[\varepsilon \mid D, X] = 0$$

and the variable of interest D admits a first stage projection

$$D = g_0(X) + \nu, \quad \mathbb{E}[\nu \mid X] = 0$$

- Write $\eta_0 = (m_0, g_0)$ and $W = (Y, D, X')'$

$$\psi(W; \theta, \eta) = (Y - \theta D - m(X)) \cdot (D - g(X))$$

- Can verify that

$$\mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0$$

and the Neyman orthogonality condition

Example 2: AIPW Estimator

- Assume unconfoundedness and overlap:

$$(Y(1), Y(0)) \perp D \mid X, \qquad 0 < \Pr(D = 1 \mid X) < 1,$$

- The parameter of interest is the average treatment effect

$$\theta_0 = \mathbb{E}[Y(1) - Y(0)].$$

- Define the nuisance functions

$$\mu_{10}(X) = \mathbb{E}[Y \mid D = 1, X], \qquad \mu_{00}(X) = \mathbb{E}[Y \mid D = 0, X],$$

- Can vetify the AIPW score is Neyman orthogonal

$$\psi(W; \theta, \eta) = \mu_1(X) - \mu_0(X) + \frac{D}{e(X)}\{Y - \mu_1(X)\} - \frac{1 - D}{1 - e(X)}\{Y - \mu_0(X)\} - \theta.$$

Example 3: DiD Analysis through DML

- Callaway & Sant'Anna's (2021) doubly-robust group-time ATT estimand: for $t > g - \delta$

$$\mathrm{ATT}(g,t;\delta) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right) \times (Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X))\right]$$

- G_g is the indicator for the group membership and C is the indicator for the never-treated group; $\delta > 0$ is a known no-anticipation buffer window..
- Nuisance functions:
  - $p_g(X) = P(G_g = 1 \mid X, G_g + C = 1)$: generalized propensity score
  - $m_{g,t,\delta}(X) = \mathbb{E}[Y_t - Y_{g-\delta-1} \mid X, C = 1]$: outcome regression
- The corresponding score is Neyman orthogonal w.r.t. p_g and m_g, t, δ.

DML with Cross-Fitting

Score Nuisance

$\mathbb{E}_{n,1}[\psi(W;\theta,\widehat{\eta}_1)]$

Nuisance Score

$\mathbb{E}_{n,2}[\psi(W;\theta,\widehat{\eta}_2)]$

DML with Cross-Fitting

- Randomly partition $\{1, ..., n\}$ into S splits $\{I_1, ..., I_S\}$
- For each $s \in \{1, ..., S\}$, obtain an estimate

$$\widehat{\eta}_s = \widehat{\eta}\left((W_i)_{i \in \{1,...,N\} \setminus I_s}\right)$$

  using only the subsample with $i \in \{1, ..., N\} \setminus I_s$
- Define $\widetilde{\theta}$, a DML estimator for θ_0, as the solution to

$$\sum_{s=1}^{S} \mathbb{E}_{n,s}[\psi(W; \theta, \widehat{\eta}_s)] = 0,$$

  where $\mathbb{E}_{n,s}[f(W)] = |I_s|^{-1} \sum_{i \in I_s} f(W_i)$ denotes the subsample empirical mean using only data with $i \in I_s$

Multiway Dependence

- Suppose the data follow two-way dependence:

$$W_{i_1 i_2} \perp W_{j_1 j_2}$$

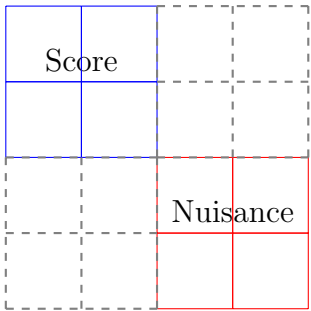
  unless $i_1 = j_1$ or $i_2 = j_2$

- Examples include
  - market share data (i_1 = market, i_2 = product): e.g. BLP (1995)
  - bipartite network data (i_1 = book, i_2 = reader): e.g. Graham (2024)
  - scanner data (i_1 = store, i_2 = product): e.g. Nielsen retail
  - matched employer-employee data (i_1 = e, i_2 = e)
  - matched student-teacher data (i_1 = s, i_2 = t)

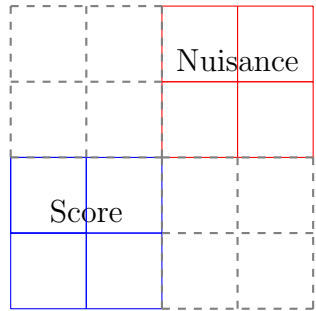
  ...

- Chiang, Kato, Ma, Sasaki (2023): multiway cross-fitting

DML with Multiway Cross-Fitting

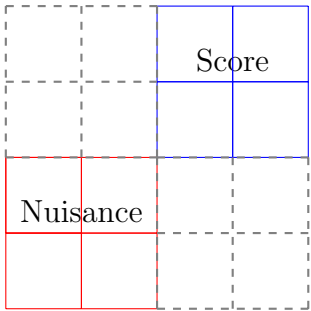


$\mathbb{E}_{n,11}[\psi(W;\theta,\widehat{\eta}_{11})]$

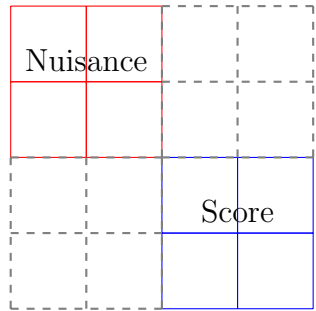


$\mathbb{E}_{n,21}[\psi(W;\theta,\widehat{\eta}_{21})]$

DML with Multiway Cross-Fitting



$\mathbb{E}_{n,12}[\psi(W;\theta,\widehat{\eta}_{12})]$



$\mathbb{E}_{n,22}[\psi(W;\theta,\widehat{\eta}_{22})]$

DML with Multiway Cross-Fitting

- Dataset $\{W_{i_1 i_2} : 1 \leqslant i_1 \leqslant N_1,\ 1 \leqslant i_2 \leqslant N_2\}$
- Randomly partition $\{1, ..., N_1\}$ into $\{I_1, ..., I_S\}$ and $\{1, ..., N_2\}$ into $\{J_1, ..., J_S\}$
- For each $(s, t) \in \{1, ..., S\}^2$, obtain a nuisance parameter estimate

$$\widehat{\eta}_{st} = \widehat{\eta}\left((W_{i_1 i_2})_{(i_1,i_2)\in(\{1,...,N_1\}\setminus I_s)\times(\{1,...,N_2\}\setminus J_t)}\right)$$

- Multiway DML estimator $\widetilde{\theta}$ is obtained by solving

$$\sum_{s=1}^{S}\sum_{t=1}^{S}\mathbb{E}_{n,st}[\psi(W;\theta,\widehat{\eta}_{st})] = 0,$$

where $\mathbb{E}_{n,st}[f(W)] = (|I_s||J_t|)^{-1}\sum_{(i_1,i_2)\in I_s\times J_t} f(W_{i_1 i_2})$

Issues with Cross-Fitting

- Estimator is random given the data, especially for small S
- Nuisance parameter is poorly estimated, especially when S is small
- Computationally demanding, especially for large S or complicated first stage

$\Rightarrow$ these difficulties are exacerbated under multiway dependence

Illustrating Example

- Consider $\eta_0 = \mathbb{E}[W_{i_1 i_2}]$ and

$$W_{i_1 i_2} = a_{i_1} + \varepsilon_{i_1 i_2}, \qquad a_{i_1}, \varepsilon_{i_1 i_2} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

- Suppose $n = N_1 = N_2$, for sample average estimator
  - **Full-sample estimator** :

$$\mathrm{Var}(\widehat{\eta}_{\text{full}}) = \frac{1}{n} + \frac{1}{n^2} \sim \frac{1}{n}$$

  - 2^2**-fold cross-fitted estimator**:

$$\mathrm{Var}(\widehat{\eta}_{st}) = \frac{2}{n} + \frac{4}{n^2} \sim \frac{2}{n}$$

Cross-Fitting-Free DML

- Sometimes (doubly) cross-fitting is beneficial: Newey and Robins (2018) – "complex" first stage, large sample
- Classical semiparametrics avoid cross-fitting using stochastic equicontinuity:
  - Robinson (1988), Ichimura (1993), Andrews (1994), ...etc
- DML without cross-fitting in special settings: Belloni, Chernozhukov, Hansen (2012), Farrell (2015), Belloni et al (2016), Chen (2025)
- Recent DML literature also avoids cross-fitting with new techniques:
  - i.i.d. sampling: Belloni, Chernozhukov, Kato (2015), Chen, Syrgkanis, Austern (2022)
  - Weak dependence (β-mixing): Cao and Leung (2025)
  - Multiway dependence: **THIS PAPER**

Background and Motivation

Simulation Evidence

Asymptotics of DML without Cross-Fitting

Maximal Inequalities under Multiway Cluster Dependence

DGPs

- A partial linear IV model:

$$
\begin{aligned}
Y_{ij} &= \theta D_{ij} + g(X_{ij}) + U_{ij}, \quad \theta = 1 \\
g(X_{ij}) &= \sin(X_{ij,1} + X_{ij,2}) + 0.5X_{ij,3} + \cdots + 0.5X_{ij,d} \\
D_{ij} &= 0.5Z_{ij} + 0.2(X_{ij,1} + X_{ij,2})^2 + V_{ij} \\
Z_{ij} &= X_{ij,2}X_{ij,3} + \varepsilon_{ij} \\
X_{ij,k} &= \alpha_{i,k} + \gamma_{j,k} + \epsilon_{ij,k}, \quad k = 1, \ldots, d \\
U_{ij} &= \alpha_{i,u} + \gamma_{j,u} + \epsilon_{ij,u} \\
V_{ij} &= \alpha_{i,v} + \gamma_{j,v} + \epsilon_{ij,v} \\
\varepsilon_{ij} &= \alpha_{i,\varepsilon} + \gamma_{j,\varepsilon} + \epsilon_{ij,\varepsilon}
\end{aligned}
$$

Machine Learners Tuning

Table: Tuning choices for nuisance-function learners

Learner	Features	Main tuning choices	Tuning rule
Ridge spline	cubic natural spline, 3 knots, raw dimension $(q+2)^d$	5-fold CV	`lambda.min`
LASSO spline, MPB	Same spline basis	Initial $\widetilde{\lambda} = \log(n)\sqrt{\log(p)/n}$; MPB with $B = 200$, $\alpha = 0.10$, $c = 1.10$; final $\lambda = 2cq_{1-\alpha}$	fixed
Elastic net spline	Same spline basis	$\alpha = 0.50$; 5-fold CV	`lambda.min`
Random forest	Raw X	500 trees; $mtry = \max\{1, \lfloor\sqrt{d_X}\rfloor\}$; min node size 5	fixed
Neural net	Raw X	Dense network $(64, 32)$ with ReLU; adaptive SGD (Adam) with learning rate 10^{-3}; 30 epochs; batch size 32; validation split 0.2 with MSE loss and patience 5 for early stop.	fixed

Notes: Each learner is used separately for the nuisance functions of Y, D, and Z. For spline learners, knots, centering, and scaling are computed using the training sample.

$N = M$	d_X	K	Learner	Bias	SD	RMSE	Coverage	Runtime/MPB	CF/No-CF
50	4	0	MPB LASSO	0.014	0.149	0.150	0.933	1.00×	Ref.
			Ridge	0.000	0.144	0.143	0.930	10.41×	Ref.
			Elastic net	-0.005	0.144	0.144	0.932	29.25×	Ref.
			Random forest	-0.023	0.125	0.127	0.933	5.20×	Ref.
			Neural network	0.016	0.146	0.147	0.926	26.63×	Ref.
50	4	2	MPB LASSO	0.007	1.585	1.584	0.924	1.00×	1.12×
			Ridge	-0.033	0.118	0.122	0.966	4.62×	0.50×
			Elastic net	-0.059	0.421	0.425	0.958	8.78×	0.34×
			Random forest	-0.027	0.105	0.109	0.973	3.77×	0.81×
			Neural network	0.023	0.126	0.128	0.937	66.83×	2.82×
50	4	4	MPB LASSO	-0.003	0.167	0.167	0.990	1.00×	7.17×
			Ridge	-0.023	0.150	0.152	0.974	11.36×	7.82×
			Elastic net	-0.038	0.293	0.296	0.986	40.25×	9.87×
			Random forest	-0.025	0.116	0.119	0.988	5.52×	7.61×
			Neural network	0.026	0.134	0.136	0.977	95.23×	25.65×

$N = M$	d_X	K	Learner	Bias	SD	RMSE	Coverage	Runtime/MPB	CF/No-CF
25	4	0	MPB LASSO	0.086	0.265	0.278	0.899	1.00×	Ref.
			Ridge	-0.019	0.165	0.166	0.949	3.82×	Ref.
			Elastic net	0.001	0.223	0.223	0.920	6.85×	Ref.
			Random forest	-0.035	0.151	0.155	0.952	3.22×	Ref.
			Neural network	0.033	0.196	0.198	0.897	42.23×	Ref.
25	4	2	MPB LASSO	0.075	1.047	1.049	0.954	1.00×	1.67×
			Ridge	-0.002	0.569	0.569	0.921	3.53×	1.54×
			Elastic net	0.119	13.084	13.078	0.899	2.34×	0.57×
			Random forest	-0.027	0.122	0.125	0.974	1.91×	0.99×
			Neural network	0.041	0.211	0.215	0.941	108.49×	4.28×
25	4	4	MPB LASSO	-0.334	8.643	8.645	0.984	1.00×	8.87×
			Ridge	-0.031	0.177	0.179	0.977	4.21×	9.78×
			Elastic net	-0.071	1.680	1.680	0.975	3.03×	3.93×
			Random forest	-0.031	0.137	0.140	0.996	2.63×	7.25×
			Neural network	0.030	0.176	0.178	0.990	182.14×	38.27×

Thank you!

Questions and comments welcome.

Background and Motivation

Simulation Evidence

Asymptotics of DML without Cross-Fitting

Maximal Inequalities under Multiway Cluster Dependence

Asymptotics of DML

- Let's derive asymptotics of $\widehat{\theta}$ together!
- Suppose the data $\{W_i\}_{i=1}^n$ are i.i.d. copies of W
- Given a nuisance parameter estimator $\widehat{\eta}$, an estimator $\widehat{\theta}$ for θ_0 can be defined as the solution of

$$\mathbb{E}_n[\psi(W;\theta,\widehat{\eta})] := \frac{1}{n}\sum_{i=1}^{n}\psi(W_i;\theta,\widehat{\eta}) = 0$$

- By definition and linearisation

$$\begin{aligned} 0 =& \mathbb{E}_n[\psi(W;\widehat{\theta},\widehat{\eta})] \\ =& \mathbb{E}_n[\psi(W;\theta_0,\widehat{\eta})] + \underbrace{\mathbb{E}_n[\partial_\theta \psi(W;\theta_0,\widehat{\eta})]}_{\text{(assume } \neq 0)}(\widehat{\theta}-\theta_0) \\ &+ O_p(\|\widehat{\theta}-\theta_0\|^2) \end{aligned}$$

- Thus, with consistency of $\widehat{\theta}$ and $\|\widehat{\eta}-\eta_0\|_{P,2} = o_p(n^{-1/4})$[2], then

$$\begin{aligned} \widehat{\theta}-\theta_0 =& -\underbrace{(\mathbb{E}_n[\partial_\theta \psi(W;\theta_0,\widehat{\eta})])^{-1}}_{\xrightarrow{p}\mathbb{E}[\partial_\theta \psi(W;\theta_0,\eta_0)]^{-1}} \mathbb{E}_n[\psi(W;\theta_0,\widehat{\eta})] \\ &+ o_p(n^{-1/2}) \end{aligned}$$

- It suffices to establish asymptotics for

$$\mathbb{E}_n[\psi(W;\theta_0,\widehat{\eta})] = (*)$$

[2] $\|f\|_{P,2} = \sqrt{\mathbb{E}[f(W)^2]}$ is the $L^2(P)$-norm.

Bias & Empirical Process Decomposition

- Recall that $\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0$
- By adding and subtracting terms

$$(*) = \underbrace{\mathbb{E}_n[\psi(W;\theta_0,\eta_0)]}_{=I} + \underbrace{\mathbb{E}[\psi(W;\theta_0,\widehat{\eta})]}_{=II} + \underbrace{n^{-1/2}\mathbb{G}_n[\psi(W;\theta_0,\widehat{\eta}) - \psi(W;\theta_0,\eta_0)]}_{=III}$$

where $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n \big(f(W_i) - \mathbb{E}[f(W_i)]\big)$ for a given f

Limiting Distribution (Standard)

- Term I consists of i.i.d. centred summands, by a standard CLT

$$\sqrt{n}I = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i; \theta_0, \eta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\psi(W; \theta_0, \eta_0)))$$

- For non-independent cases, so long as a CLT is available, term I remains asymptotically normal

  $\Rightarrow$ Under multiway dependence, CLTs are available from, e.g. Davezies, D'Haultfoeuille, Guyonvarch (2021), Chiang, Kato, Sasaki (2023), ...

Controlling Bias (Standard)

- II is a bias term in which the dependence does NOT play a direct role
- Thanks to orthogonality and regularity conditions,

$$\begin{aligned}&\mathbb{E}[\psi(W;\theta_0,\widehat{\eta})]\\ =&\mathbb{E}[\psi(W;\theta_0,\eta_0)] + \partial_\eta \mathbb{E}[\psi = (W;\theta_0,\eta_0)](\widehat{\eta}) + O(\|\widehat{\eta}-\eta_0\|_{P,2}^2)\\ =&0 + 0 + O(\|\widehat{\eta}-\eta_0\|_{P,2}^2)\end{aligned}$$

- If $\|\widehat{\eta}-\eta_0\|_{P,2} = o_p(n^{-1/4})$, then $\sqrt{n}|II| = o(1)$

Controlling Empirical Process with Cross-Fitting

- Suppose $\widehat{\eta} \perp \{W_i\}_{i=1}^n$
- Conditionally on $\widehat{\eta}$, its summands are i.i.d. with zero mean.

$$
\begin{aligned}
n^{-1/2}|III| &= |\mathbb{G}_n[\psi(W;\theta_0,\widehat{\eta}) - \psi(W;\theta_0,\eta_0)]| \\
&= \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big\{ \psi(W_i;\theta_0,\widehat{\eta}) - \psi(W_i;\theta_0,\eta_0) - \mathbb{E}[\psi(W_i;\theta_0,\widehat{\eta}) - \psi(W_i;\theta_0,\eta_0)] \Big\} \right| \\
&\lesssim_p \sqrt{\mathbb{E}\left[\Big(\psi(W;\theta_0,\widehat{\eta}) - \psi(W;\theta_0,\eta_0) \Big)^2 \right]} \lesssim_p \|\widehat{\eta} - \eta_0\|_{P,2}^{\nu}
\end{aligned}
$$

for some $\nu > 0$, under regularity conditions

$\Rightarrow$ $\|\widehat{\eta} - \eta_0\|_{P,2} = o_p(1)$ suffices for $\sqrt{n}|III| = o_p(1)$

Controlling Empirical Process without Cross-Fitting

- Without cross-fitting: $\widehat{\eta} \not\perp \{W_i\}_{i=1}^n$
- Suppose for some $\mathcal{H}_n \ni \eta_0$, we have $\widehat{\eta} \in \mathcal{H}_n$ with probability $1 - o(1)$. Define

$$\mathcal{F}_n = \{w \mapsto \psi(w; \theta_0, \eta) - \psi(w; \theta_0, \eta_0) : \eta \in \mathcal{H}_n\}$$

- Under i.i.d., $\mathbb{G}_n(f)$, $f \in \mathcal{F}_n$ consists of i.i.d. summands
- Can apply a maximal inequality to control these "localised" empirical processes

$$\sqrt{n}|III| \lesssim_p \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)| \lesssim_p \frac{\{\text{"Size" \& "Complexity" of } \mathcal{F}_n\}}{\sqrt{n}}$$

- Similar to the classical semiparametrics (relies on stochastic equicontinuity over a fixed $\mathcal{F}$)

Localisation of Empirical Process

- Machine learners have growing complexities
  $\Rightarrow$ tackle the growing "Complexity" with shrinking "Size"
- "Size" of $\mathcal{H}_n$ (hence $\mathcal{F}_n$) can be chosen to contract to η_0 sharply if a good rate of $\|\widehat{\eta} - \eta_0\|_{P,2}$ is known

  $\Rightarrow$ typically so since the rate is needed for orthogonality
- $\sqrt{n}|III| = o_p(1)$ so long as the "Complexity" do not grow too fast with n

  $$\{\text{"Size" \& "Complexity" of } \mathcal{F}_n\} = o(\sqrt{n})$$

  can be obtained by using a maximal inequality
- However, such a maximal inequality is **NOT available** under multiway dependence
  $\Rightarrow$ we develop our own

Summary & What Else Is in This Paper?

- Multiway DML without Cross-Fitting:
  Localisation → Hoeffding-Type Decomposition→ Maximal Inequalities
- What else is in the paper?
  - Generic **debiased GMM** estimators under general multiway dependence
  - Asymptotic normality holds **without** cross-fitting under general high-level complexity conditions (VC-type) as well as smoothness boundedness of the score function.
  - High-level conditions are satisfied for some **regularised GLMs** (p and sparsity), **tree-based learners** (p and leaves), and **neural networks** (p, layers, hidden units, total parameters)
  - **Multiway-cluster robust** consistent variance estimator

Background and Motivation

Simulation Evidence

Asymptotics of DML without Cross-Fitting

Maximal Inequalities under Multiway Cluster Dependence

Complexity

A class $\mathcal{F}$ is of VC-type if for an envelope F[3] and some positive characteristics (A, v)

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2}) \leqslant \left(\frac{A}{\varepsilon}\right)^v \quad \text{for all } 0 < \varepsilon \leqslant 1$$

- $N(T, \|\cdot\|, \varepsilon)$: covering number for a set T w.r.t. a pseudonorm $\|\cdot\|$, $\varepsilon > 0$
- Supremum over all finite discrete measures Q
- Examples: a finite class, a class of indicator functions, a class indexed by a finite-dimensional vector space, a class of smooth functions indexed by finite-dimensional parameters, their Lipschitz and/or bounded-variation transformations, and finite union/intersection/sum/product/division of these classes...

[3] i.e. $|f(x)| \leqslant F(x)$ for all x and $f \in \mathcal{F}$.

Maximal Inequality under i.i.d.

- Suppose $\{W_i\}_{i=1}^n$ are i.i.d. From the literature (Pollard, 1990; van der Vaart and Wellner, 2010),
  - *Global maximal inequality*

$$\left(\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)|^q\right]\right)^{1/q} \lesssim \|F\|_{P,q\vee 2}\sqrt{v\log(A\vee n)}, \quad q\in[1,\infty)$$

  - *Local maximal inequality*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)|\right] \lesssim \sigma\sqrt{v\log(A\vee n)} + \frac{\|\max_{i=1,\ldots,n}F(W_i)\|_{P,2}}{\sqrt{n}}v\log(A\vee n).$$

    where $\sigma^2 = \sup_{f\in\mathcal{F}}\mathbb{E}f^2$, $\|f\|_{P,q} = (\mathbb{E}|f|^q)^{1/q}$.

- Both are absent under multiway-cluster dependence.

Representations for Multiway Dependnce

- Suppose the data $\{W_{\boldsymbol{i}} : \boldsymbol{i} = (i_1, \ldots, i_K) \in \mathbb{N}^K\}$ exhibit K-way dependence
- Assume that the data r.v.'s are separately exchangeable and dissociated, or equivalently, they admit the *Aldous–Hoover–Kallenberg representation*:

$$W_{\boldsymbol{i}} \stackrel{d}{=} \tau\Big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}}\}_{\boldsymbol{e} \in \{0,1\}^K \setminus \{0\}}\Big), \qquad \textbf{(AHK)}$$

where $\{U_{\boldsymbol{i} \odot \boldsymbol{e}}\}_{\boldsymbol{e} \in \{0,1\}^K \setminus \{0\},\, \boldsymbol{i} \in [\boldsymbol{N}]}$ are i.i.d. r.v.'s, τ is Borel-measurable, and $\odot$ is the pointwise product[4].

[4]For $\boldsymbol{i} = (i_1, \ldots, i_K)$, $\boldsymbol{j} = (j_1, \ldots, j_K)$, $\boldsymbol{i} \odot \boldsymbol{j} = (i_1 \cdot j_1, \ldots, i_K \cdot j_K)$

Example: $K = 2$

- Suppose $K = 2$, i_1 a market, i_2 a product
- The AHK representation implies existence of the representation

$$W_{i_1 i_2} \stackrel{d}{=} \tau(U_{i_1 0}, U_{0\, i_2}, U_{i_1 i_2}),$$

where

- $U_{i_1 0}$: market specific shock
- $U_{0 i_2}$: a product specific shock
- $U_{i_1 i_2}$: interactive shock

and the shocks are all mutually independent

$\Rightarrow$ $W_{i_1 i_2}$ and $W_{i_1' i_2'}$ are possibly dependent if either $i_1 = i_1'$ or $i_2 = i_2'$

Empirical Processes with Multiway Dependnce

- Given a data $\{W_{\boldsymbol{i}} : \boldsymbol{i} \in I_{\mathbf{N}} := \{1, ..., N_1\} \times ... \times \{1, ..., N_K\}\}$
- Can accommodate heterogeneous number of observations per $\boldsymbol{i}$ cell with a little rewriting–theory remains unchanged
- $n = \min\{N_1, ..., N_K\}$
- Denote for f

$$\mathbb{G}_n(f) = \frac{\sqrt{n}}{|I_{\mathbf{N}}|} \sum_{\boldsymbol{i} \in I_{\mathbf{N}}} \left(f(W_{\boldsymbol{i}}) - \mathbb{E}[f(W_{\boldsymbol{i}})] \right)$$

- Bounding the whole empirical processes is difficult due to dependence
  $\Rightarrow$ Use orthogonal projections or *Hoeffding-type decomposition* (Hoeffding 1948; Chiang, Kato, Sasaki, 2023)

Multiway Interactions

- For each $k \in \{1, \ldots, K\}$, define

$$\mathcal{E}_k = \left\{ \boldsymbol{e} = (e_1, ..., e_K) \in \{0,1\}^K : \sum_{j=1}^{K} e_j = k \right\}$$

  Interpretation: $\mathcal{E}_k$ consists of all the possible k-way interactions and $\boldsymbol{e} \in \mathcal{E}_k$ indicates which k-way interactions are turned on

- $\{0,1\}^K = \bigcup_{k=0}^{K} \mathcal{E}_k$
- Let $\boldsymbol{e}_k \in \mathcal{E}_1$ be a vector with all zero elements except for the k-th entry.

Example: $K = 2$, i_1 a market, i_2 a product

- $\mathcal{E}_1 = \{(1,0), (0,1)\} = \{e_1, e_2\}$ (market or product-specific effects)
- $\mathcal{E}_2 = \{(1,1)\}$ (interactive effect)

Hoeffding-Type Decomposition

- For $\boldsymbol{a} = (a_1, \ldots, a_K)$, $\boldsymbol{b} = (b_1, \ldots, b_K)$, write $\boldsymbol{a} \leqslant \boldsymbol{b}$ if $a_j \leqslant b_j$ for all $1 \leqslant j \leqslant K$
- WLOG $\mathbb{E}[f(W_{\boldsymbol{i}})] = 0$ and define for each $\boldsymbol{e} \in \{0,1\}^K$

$$(P_{\boldsymbol{e}} f)\Big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\Big) = \mathbb{E}\Big[f(W_{\boldsymbol{i}}) \,\Big|\, \{U_{\boldsymbol{i} \odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\Big]$$

- Hoeffding-type projections: for each unit vector $\boldsymbol{e} \in \mathcal{E}_1$ that

$$(\pi_{\boldsymbol{e}} f)(U_{\boldsymbol{i} \odot \boldsymbol{e}}) = (P_{\boldsymbol{e}} f)(U_{\boldsymbol{i} \odot \boldsymbol{e}}),$$

and for each $\boldsymbol{e} \in \bigcup_{k=2}^{K} \mathcal{E}_k$ define recursively

$$(\pi_{\boldsymbol{e}} f)\Big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\Big) = (P_{\boldsymbol{e}} f)\Big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\Big) - \sum_{\substack{\boldsymbol{e}' \leqslant \boldsymbol{e} \\ \boldsymbol{e}' \neq \boldsymbol{e}}} (\pi_{\boldsymbol{e}'} f)\Big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}''}\}_{\boldsymbol{e}'' \leqslant \boldsymbol{e}'}\Big)$$

Example

- $K = 2$, i_1 a market, i_2 a product
- Hoeffding-type projections:

$$
\begin{aligned}
(\pi_{(1,0)} f)(U_{i_1 0}) &= \mathbb{E}[f(W_{i_1 i_2}) \mid U_{i_1 0}] && \text{(market-specific effect)} \\
(\pi_{(0,1)} f)(U_{0\, i_2}) &= \mathbb{E}[f(W_{i_1 i_2}) \mid U_{0\, i_2}] && \text{(product-specific effect)} \\
(\pi_{(1,1)} f)(U_{i_1 0}, U_{0\, i_2}, U_{i_1 i_2}) &= f(W_{i_1 i_2}) - \mathbb{E}[f(W_{i_1 i_2}) \mid U_{i_1 0}] - \mathbb{E}[f(W_{i_1 i_2}) \mid U_{0\, i_2}] \\
& && \text{(interactive effect)}
\end{aligned}
$$

- Hence by telescoping

$$
\begin{aligned}
f(W_{i_1 i_2}) =& (\pi_{(1,0)} f)(U_{i_1 0}) + (\pi_{(0,1)} f)(U_{0\, i_2}) + (\pi_{(1,1)} f)(U_{i_1 0}, U_{0\, i_2}, U_{i_1 i_2}) \\
=& \sum_{k=1}^{2} \sum_{\boldsymbol{e} \in \mathcal{E}_k} (\pi_{\boldsymbol{e}} f)\big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\big)
\end{aligned}
$$

- This yields the Hoeffding-type decomposition

$$\mathbb{G}_n(f) = \sqrt{n} \sum_{k=1}^{K} \sum_{\boldsymbol{e} \in \mathcal{E}_k} H_{\boldsymbol{N}}^{\boldsymbol{e}}(f),$$

where for each $\boldsymbol{e} \in \bigcup_{k=1}^{K} \mathcal{E}_k$, the corresponding *Hoeffding-type projection*

$$H_{\boldsymbol{N}}^{\boldsymbol{e}}(f) = \frac{1}{|I_{\boldsymbol{N},\boldsymbol{e}}|} \sum_{\boldsymbol{i} \in I_{\boldsymbol{N},\boldsymbol{e}}} (\pi_{\boldsymbol{e}} f)\big(\{U_{\boldsymbol{i} \odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\big),$$

$I_{\boldsymbol{N},\boldsymbol{e}} = \{\boldsymbol{i} \odot \boldsymbol{e} : \boldsymbol{i} \in I_{\mathbf{N}}\}$

$\Rightarrow$ To control $\mathbb{G}_n(f)$, it suffices to control $H_{\boldsymbol{N}}^{\boldsymbol{e}}(f)$ for each $\boldsymbol{e} \in \{0,1\}^K$

Local Maximal Inequality

Theorem For each $\boldsymbol{e} \in \mathcal{E}_k$, let $\sigma_{\boldsymbol{e}}$ be a constant such that

$$\sup_{f\in\mathcal{F}} \|P_{\boldsymbol{e}} f\|_{P,2} \leqslant \sigma_{\boldsymbol{e}} \leqslant \|P_{\boldsymbol{e}} F\|_{P,2}, \qquad M_{\boldsymbol{e}} = \max_{t=1,\ldots,n} P_{\boldsymbol{e}} F\left(\{U_{(t,\ldots,t)\odot \boldsymbol{e}'}\}_{\boldsymbol{e}' \leqslant \boldsymbol{e}}\right).$$

If $\mathcal{F}$ is VC-type,

$$|I_{\boldsymbol{N},\boldsymbol{e}}|^{1/2}\mathbb{E}\left[\sup_{f\in\mathcal{F}} |H^{\boldsymbol{e}}_{\boldsymbol{N}}(f)|\right]$$
$$\lesssim \sigma_{\boldsymbol{e}}\{v\log(A\|P_{\boldsymbol{e}}F\|_{P,2}/\sigma_{\boldsymbol{e}})\}^{k/2} + \frac{\|M_{\boldsymbol{e}}\|_{P,2}}{\sqrt{n}}\{v\log(A\|P_{\boldsymbol{e}}F\|_{P,2}/\sigma_{\boldsymbol{e}})\}^{k}$$

- The first term: $\sigma_{\boldsymbol{e}}$ ("Size") typically decreasing as the $v\log(...)$ ("Complexity") increases
- The second term: often negligible due to the extra $1/\sqrt{n}$
- Can be used to show $\sqrt{n}|III| = o_p(1)$ and thus DML without cross-fitting

Learning Nuisance Parameter under Multiway Dependence

- Our maximal inequalities are also useful for establishing convergence rate $\|\widehat{\eta} - \eta_0\|_{P,2}$ under multiway dependence

- e.g. for LASSO with a p-dimensional predictor, we need a bound in probability for

$$\max_{1 \leqslant j \leqslant p} \left| \sum_{\boldsymbol{i} \in I_{\mathbf{N}}} X_{\boldsymbol{i},j} \varepsilon_{\boldsymbol{i}} \right|$$

- Under multiway dependence, this can again be controlled by using maximal inequality + Hoeffding-type decomposition

Remarks on the Proofs

- The proofs of both maximal inequalities rely on symmetrisation (Chiang, Kato, Sasaki, 2023), the maximal inequality for Rademacher chaos (de la Peñ a and Giné, 1999), and, in the case of local maximal inequality, Hoffmann–Jørgensen inequality

- The related maximal inequality for U-processes is obtained using the classical Hoeffding-averaging technique to allow for an application of Hoffmann–Jørgensen m

  $\Rightarrow$ Hoeffding-averaging is not applicable for multiway means

- To overcome this difficulty, we show that each $\{\boldsymbol{i} \odot \boldsymbol{e} : \boldsymbol{i} \in [\boldsymbol{N}]\}$ can be partitioned into groups such that the sum within each group consists of n i.i.d. random variables, allowing the application of Hoffmann–Jørgensen

Bonus: Maximal Score and Non-Smooth M-Estimators

The maximal inequality is also the key to:

"Gaussian Approximation for Maximum Score and Non-Smooth M-Estimators with Multiway Dependence" (2026) w Ahnaf Rafi (UVA)

- Maximum score estimator (Manski, 1975) exhibits non-Gaussian $n^{1/3}$-rate asymptotics under i.i.d. (Kim and Pollard, 1990)
- Under multiway dependence, the estimator instead attains asymptotic normality at a parametric rate
- General non-smooth M-estimation theory under multiway clustering
- Provide a valid bootstrap for inference

Summary

- Existing DML methods reply on **cross-fitting**, but under multiway dependence it is **complicated** and **causes extra issues**.
- To study DML methods without cross-fitting, we first develop **maximal inequalities under multiway-cluster dependence**.
  - Also of independent interest.
- Applying our results, we characterize the complexity conditions of the score class indexed by the nuisance parameter, under which **DML without cross-fitting is valid**.
  - We verify that common machine learner satisfy such complexity conditions.
- In simulation, we illustrate the **issue with cross-fitted DML** as well as the **substantial computational advantage** with DML without cross-fitting.