Causal Inference with Double Machine Learning

Zijian Wang

Master thesis – Final presentation

October 25, 2024

Theory of Double Machine Learning

Potential outcomes framework

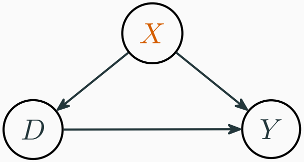
- Want to estimate the causal effect of a binary treatment D on the outcome Y
- Key idea: There are two potential outcomes $Y_i(1)$ and $Y_i(0)$
  $\Longrightarrow$ Values of Y_i if individual i did or did not receive the treatment
- Individual-level treatment effect $Y_i(1) - Y_i(0)$ is never observed
- Population parameters of interest:
  - Average treatment effect: ATE $= \mathbb{E}[Y(1) - Y(0)]$
  - Average treatment effect on the treated: ATT $= \mathbb{E}[Y(1) - Y(0) \,|\, D = 1]$

Identification under unconfoundedness

- For randomized experiments, assuming $(Y(1), Y(0)) \perp\!\!\!\perp D$ ensures identification:

$$\text{ATE} = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$$

- More reasonable assumptions for observational studies:
  - Unconfoundedness: $(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$
  - Overlap: $\mathbb{P}(D = 1 \mid X) \in (0, 1)$ a.s.



Interactive regression model

- Consider the following model:

$$
\begin{aligned}
Y &= g_0(D, X) + u, \quad \mathbb{E}[u \,|\, D, X] = 0 \\
D &= m_0(X) + v, \quad \mathbb{E}[v \,|\, X] = 0
\end{aligned}
$$

- Outcome regression function: $g_0(D, X) = \mathbb{E}[Y \,|\, D, X]$
- Propensity score: $m_0(X) = \mathbb{E}[D \,|\, X] = \mathbb{P}(D = 1 \,|\, X)$
- Causal parameters can be written as:

$$
\text{ATE} = \mathbb{E}[g_0(1, X) - g_0(0, X)], \quad \text{ATT} = \mathbb{E}[g_0(1, X) - g_0(0, X) \,|\, D = 1]
$$

Classical approaches to estimation

- Assumption: $W_i = (Y_i, D_i, X_i)$ is i.i.d. across $i = 1, \ldots, N$
- Regression estimator is derived from ATE $= \mathbb{E}[g_0(1, X) - g_0(0, X)]$
  $$\implies \hat{\tau}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{g}_0(1, X_i) - \hat{g}_0(0, X_i))$$
- Inverse propensity weighting (IPW) leverages ATE $= \mathbb{E}\left[\frac{DY}{m_0(X)} - \frac{(1-D)Y}{1-m_0(X)}\right]$
  $$\implies \hat{\tau}_{\text{ipw}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D_i Y_i}{\hat{m}_0(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{m}_0(X_i)}\right)$$
- Plug in non-parametric estimator $\hat{g}_0$ or $\hat{m}_0$ of nuisance function g_0 or m_0
  $\implies$ Estimator of ATE won't attain parametric convergence rate $N^{-1/2}$

Double machine learning

- Papers by Chernozhukov et al. [2017, 2018]
- Efficient estimation and inference procedure for ATE and ATT
- Main ingredients of double machine learning (DML):
    - Moment equation that defines an (infeasible) method-of-moments estimator, with some nuisance parameter to be estimated beforehand
    - Employs ML methods to learn the nuisance functions
    - Uses Neyman orthogonality to "remove" non-parametric convergence rates
    - Uses cross-fitting to avoid overfitting bias

Neyman orthogonal moment equations

- Generic nuisance parameter $\eta = (h_0, h_1, h_2, p)$, with the true value given by $\eta_0 = (g_0(0, \cdot), g_0(1, \cdot), m_0, \mathbb{E}[D])$
- Consider the score functions:

$$\psi^{\mathsf{ATE}}(W;\theta,\eta) = h_1(X) - h_0(X) + \frac{D(Y-h_1(X))}{h_2(X)} - \frac{(1-D)(Y-h_0(X))}{1-h_2(X)} - \theta$$
$$\psi^{\mathsf{ATT}}(W;\theta,\eta) = \frac{D(Y-h_0(X))}{p} - \frac{h_2(X)(1-D)(Y-h_0(X))}{(1-h_2(X))p} - \theta\frac{D}{p}$$

Proposition (Identification and Neyman orthogonality)

The causal parameter θ_0 is identified via the moment equation $\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0$. This moment equation is "insensitive" to small errors in the nuisance parameter, in the sense that $\partial_\eta \mathbb{E}[\psi(W;\theta_0,\eta)]\big|_{\eta=\eta_0} = 0$.

DML algorithm

1. Form a K-fold partition of $\{1, \ldots, N\}$ into $\{I_k\}_{k=1}^{K}$ each of the size $n := N/K$, and define $I_k^c := \{1, \ldots, N\} \setminus I_k$.
2. For each k, use $\{W_i\}_{i \in I_k^c}$ to compute an ML estimator of η_0:
$$\hat{\eta}_0(I_k^c) := \big(\hat{g}_0(0, \cdot\,; I_k^c),\ \hat{g}_0(1, \cdot\,; I_k^c),\ \hat{m}_0(\,\cdot\,; I_k^c),\ \bar{D}(I_k^c)\big)$$
3. For each k, define $\check{\theta}_{0,k}$ as the root θ of:
$$\frac{1}{n} \sum_{i \in I_k} \psi(W_i; \theta, \hat{\eta}_0(I_k^c)) = 0$$
4. Obtain the final estimator of θ_0 by averaging:
$$\hat{\theta}_0 := \frac{1}{K} \sum_{k=1}^{K} \check{\theta}_{0,k}$$

Asymptotics of DML estimator

Required model regularity:

- Uniform overlap: $m_0(X) \in (\varepsilon, 1-\varepsilon)$ for some $\varepsilon > 0$

High-level assumptions on the ML estimator:

- Convergence: $\|\hat{g}_0(d, X; I_k^c) - g_0(d, X)\|_{L^2} + \|\hat{m}_0(X; I_k^c) - m_0(X)\|_{L^2} = o_P(1)$
- Conv. rates: $\|\hat{g}_0(d, X; I_k^c) - g_0(d, X)\|_{L^2} \cdot \|\hat{m}_0(X; I_k^c) - m_0(X)\|_{L^2} = o_P(N^{-1/2})$
- $\hat{m}_0(X; I_k^c) \in (\varepsilon, 1-\varepsilon)$ with probability approaching 1

Theorem (Asymptotic normality)

The DML estimator $\hat{\theta}_0$ is asymptotically normally distributed and $\sqrt{N}$-consistent:

$$\sqrt{N}(\hat{\theta}_0 - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Asymptotics of DML estimator (cont.)

Proof sketch for ATT case.

We have the following representation for the auxiliary estimation error:

$$
\begin{aligned}
&\frac{\bar{D}(I_k)}{\bar{D}(I_k^c)}\sqrt{n}(\check{\theta}_{0,k} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i \in I_k} \psi(W_i; \theta_0, \hat{\eta}_0(I_k^c)) \\
= &\frac{1}{\sqrt{n}} \sum_{i \in I_k} \psi(W_i; \theta_0, \hat{\eta}_0(I_k^c)) - \sqrt{n}\,\mathbb{E}\big[\psi(W; \theta_0, \hat{\eta}_0(I_k^c)) \,\big|\, \{W_i\}_{i \in I_k^c}\big] \\
&+ \sqrt{n}\,\mathbb{E}\big[\psi(W; \theta_0, \hat{\eta}_0(I_k^c)) \,\big|\, \{W_i\}_{i \in I_k^c}\big] \\
= &\frac{1}{\sqrt{n}} \sum_{i \in I_k} \psi(W_i; \theta_0, \eta_0) - \sqrt{n}\,\mathbb{E}\big[\psi(W; \theta_0, \eta_0) \,\big|\, \{W_i\}_{i \in I_k^c}\big] + o_P(1)
\end{aligned}
$$

This implies that $\sqrt{n}(\check{\theta}_{0,k} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i \in I_k} \psi(W_i; \theta_0, \eta_0) + o_P(1)$.

Asymptotics of DML estimator (cont.)

Proof sketch for ATT case (cont.)

By the CLT and Slutsky's theorem, we finally obtain:

$$\sqrt{N}(\hat{\theta}_0 - \theta_0) = \sum_{k=1}^{K} \frac{\sqrt{n}(\check{\theta}_{0,k} - \theta_0)}{\sqrt{K}} = \sum_{k=1}^{K} \frac{1}{\sqrt{nK}} \sum_{i \in I_k} \psi(W_i; \theta_0, \eta_0) + o_P(1)$$
$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(W_i; \theta_0, \eta_0) + o_P(1) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

$\square$

- Asymptotic variance: $\sigma^2 = \mathbb{E}[\psi(W; \theta_0, \eta_0)^2]$
- Consistent estimation via $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \psi(W_i; \hat{\theta}_0, \hat{\eta}_0(I^c_{k(i)}))^2$
- Approximate $(1-\alpha) \cdot 100\%$ CI: $\left[\hat{\theta}_0 \pm z_{1-\alpha/2} \cdot \hat{\sigma}/\sqrt{N}\right]$

Numerical experiments

Simple model with 3 confounders

- Confounding variables $X = (X_1, X_2, X_3)'$:

$$(X_1, X_2)' \sim \mathcal{N}\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1 & -0.2\\ -0.2 & 0.5\end{pmatrix}\right), \quad X_3 \sim \mathcal{U}(0,1)$$

- Treatment $D = \mathbb{1}_{X'\beta+\xi\geq 0}$, with $\beta = (1, 2, -1)'$ and error term $\xi \sim F_{\text{log}}$, $\xi \perp\!\!\!\perp X$
  $\implies$ Propensity score $m_0(X) = F_{\text{log}}(X'\beta)$
- Outcome $Y = g_0(D, X) + u$, with regression function
  $g_0(D, X) = DX_1 + F_{\text{log}}(X_2) - 2X_3^2$ and error term $u \sim \mathcal{N}(0, X_3^2)$
- Ground truth: $\theta_0^{\text{ATE}} = \mathbb{E}[g_0(1, X) - g_0(0, X)] = \mathbb{E}[X_1] = 1$

Non-parametric rates of ML methods

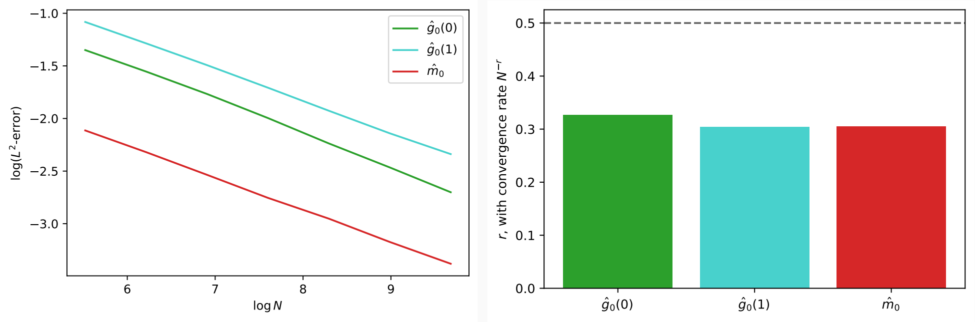


Figure: Convergence rates of nuisance estimators, here XGBoost

DML vs. "classical" estimators

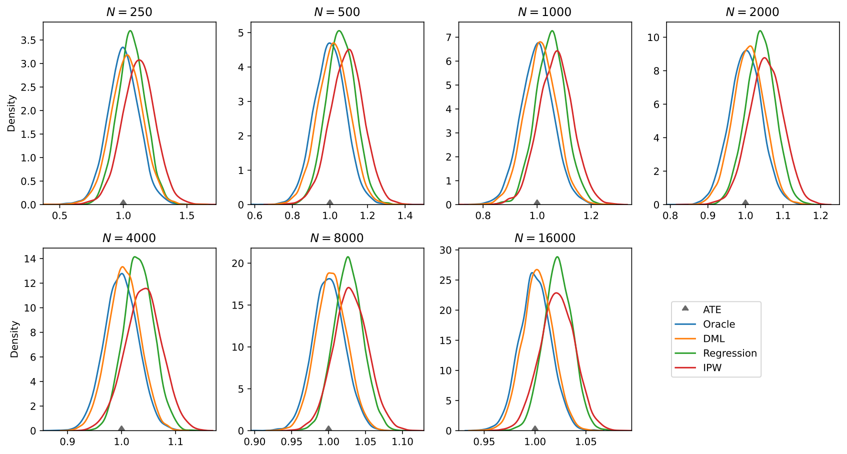


Figure: Sampling distributions

DML vs. “classical” estimators (cont.)

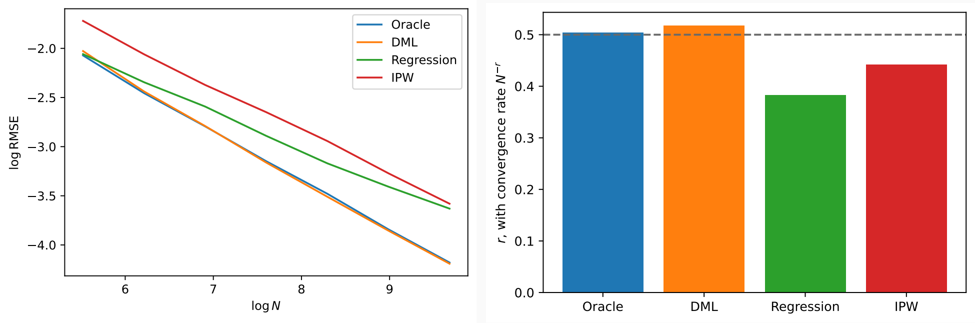


Figure: Convergence rates of ATE estimators

$\Longrightarrow$ Neyman orthogonality eliminates non-parametric convergence behavior

Effect of cross-fitting

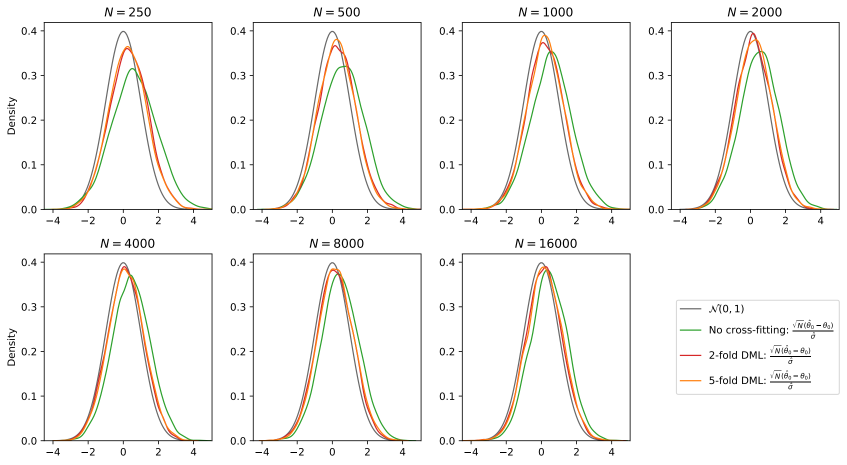


Figure: Assessing asymptotic normality

Effect of cross-fitting (cont.)

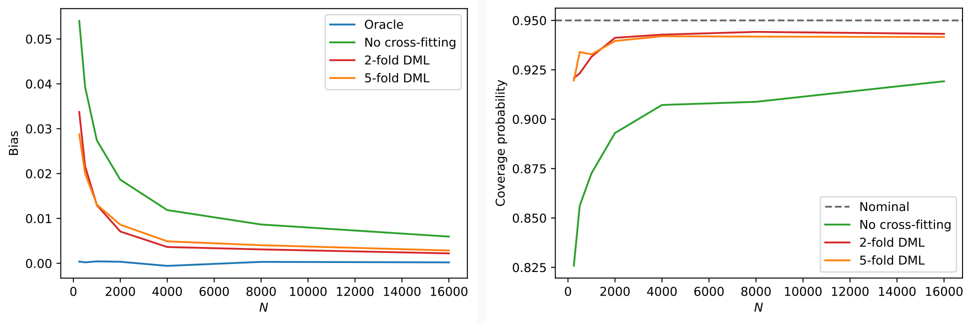


Figure: Biases and coverage probabilities of CIs

$\Longrightarrow$ Bias reduction through cross-fitting (removal of "overfitting bias")

Complex model with 10 confounders

- Confounding variables $X = (X_1, \dots, X_{10})'$:

$$(X_1, \dots, X_8)' \sim \mathcal{N}_8(\mu, \Sigma), \quad (X_9, X_{10})' \sim \mathcal{U}([0,1]^2)$$

- Treatment $D = \mathbb{1}_{X'\beta + 0.25X_8^2 - X_9X_{10} + \xi \geq 0}$, with error term $\xi \sim F_t$, $\xi \perp\!\!\!\perp X$
$\implies$ Propensity score $m_0(X) = F_t(X'\beta + 0.25X_8^2 - X_9X_{10})$

- Outcome $Y = g_0(D, X) + u$, with regression function $g_0(D, X) =$
$X_1 + 2X_2 + 2X_3 + 3X_4 + (D+1)X_5 + F_{\log}(X_6)X_7^2 - X_9(X_{10}^{1/2} + 2X_7) + DX_3X_9^{3/2}$
and error term $u \sim \mathcal{N}(0, \sigma^2(X))$, $\sigma(X) = \frac{1}{10}\sum_{i=1}^{10} |X_i|$

- Ground truth: $\theta_0^{\mathsf{ATE}} = \mathbb{E}[g_0(1, X) - g_0(0, X)] = \mathbb{E}[X_5] + \mathbb{E}[X_3]\mathbb{E}[X_9^{3/2}] = 0.5$

Testing different nuisance learners

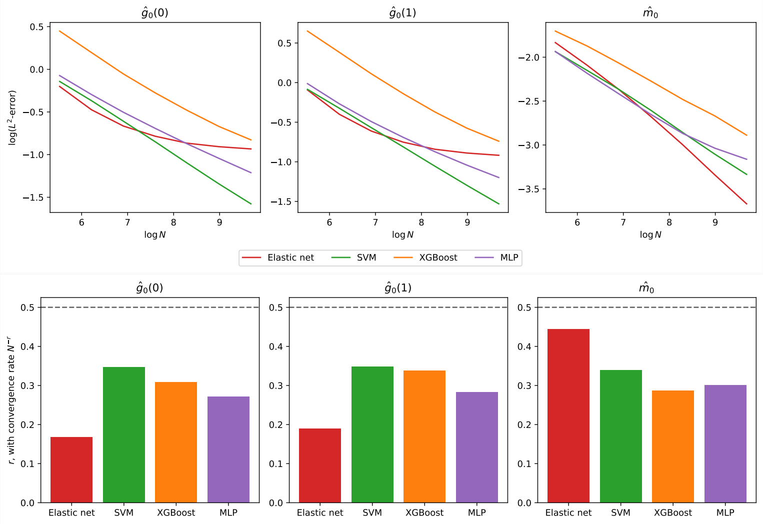


Figure: Approx. errors and convergence rates of nuisance estimators

Testing different nuisance learners (cont.)

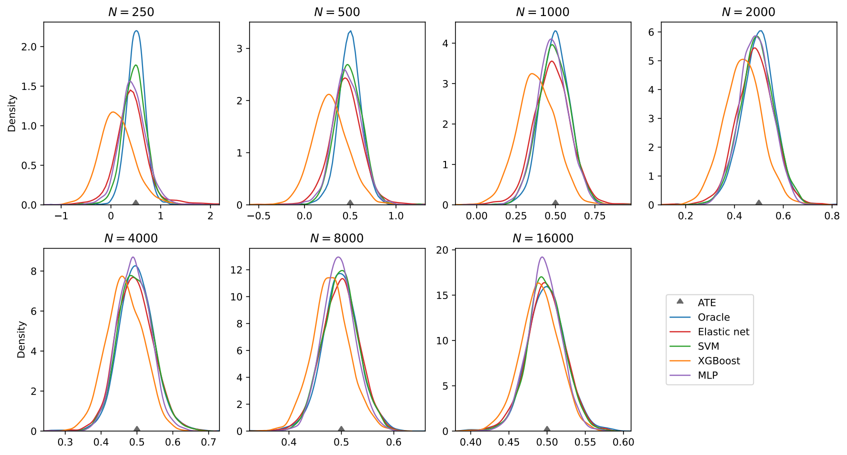


Figure: Sampling distributions

Testing different nuisance learners (cont.)

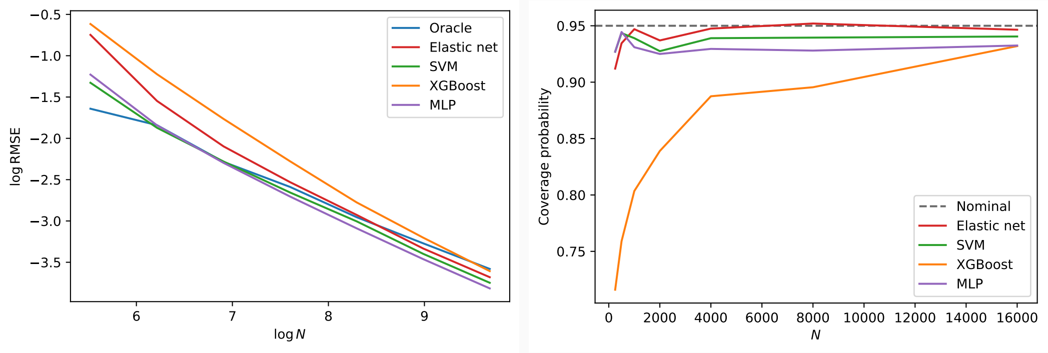


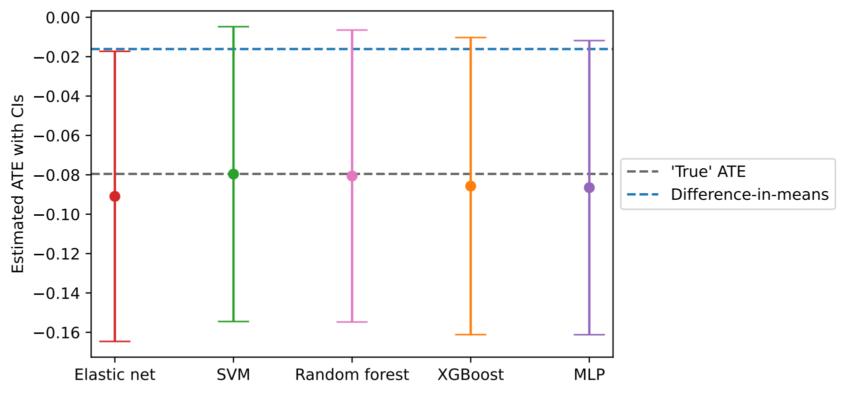
Figure: RMSE decay and coverage probabilities of CIs

$\Longrightarrow$ Should compute DML estimator using different ML methods, allowing us to assess the robustness of estimation results

Inducing confounding in reemployment bonus data

- RCT to study whether offering a cash bonus to claimants of the unemployment insurance who manage to find a job within a certain period of time can reduce the unemployment duration [Bilias, 2000]
- ATE of bonus on log unemployment duration can be consistently estimated by the difference in mean outcomes: $\widehat{\mathsf{ATE}}_{\mathsf{full}} \approx -0.0795$
- Crafted a covariates-based selection mechanism to extract a sub-dataset on which difference-in-means severely underestimates the reduction in log unemployment duration: $\widehat{\mathsf{ATE}}_{\mathsf{biased}} \approx -0.0161$

DML applied to reemployment bonus data



$\Longrightarrow$ DML overcomes confounding and correctly recovers the “true” ATE

References

Y. Bilias. Sequential testing of duration data: The case of the Pennsylvania 'Reemployment Bonus' experiment. *Journal of Applied Econometrics*, 15(6):575–594, 2000. ISSN 08837252, 10991255. URL http://www.jstor.org/stable/2678561.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, May 2017. doi: 10.1257/aer.p20171038. URL https://www.aeaweb.org/articles?id=10.1257/aer.p20171038.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.