Extended Notes: Double / Debiased Machine Learning Talk

These notes back each slide with more details and references to the Chernozhukov et al. (2018) paper *"Double / Debiased Machine Learning for Treatment and Structural Parameters"* and Victor Chernozhukov's talk.

Slide 2 — The problem: causal effects with many covariates

- The paper's intro sets up the problem as inference on a low-dimensional parameter θ_0 in the presence of high-dimensional or highly complex nuisance parameters η_0, estimated via ML. This is explicitly stated in the abstract and Section 1.
- The point that ML is good for prediction but not automatically for causal parameters is stressed in the talk, where Chernozhukov emphasizes that good prediction does not guarantee good estimation of a causal parameter and can even be misleading.

Slide 3 — Prediction vs. causal estimation

- The general moment-condition language ($\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0$) is introduced in Section 2 (Moment condition / estimating equations framework).
- The distinction between prediction and causal estimation is illustrated in the Introduction: ML methods are used to estimate complex nuisances, but naive plug-in leads to biased estimators of θ_0.
- The talk transcript also highlights two main points: (i) ML can predict very well, (ii) naive use produces poor estimators for causal parameters.

Slide 4 — Why econometricians love moment conditions

- The paper uses a GMM-style viewpoint: target parameters are defined by population moment conditions, then estimated by solving empirical analogues. This is formalized in Section 2.1.
- Examples like OLS, IV, and GMM are standard; the paper also references semi-parametric literature where efficient scores come from such moment conditions (e.g. Chamberlain 1987, Newey 1994).

Slide 5 — Moment conditions in the partially linear model

- The PLR model is introduced as Example 1.1 in Section 1:

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

- The naive regression-adjustment moment ($\mathbb{E}[(Y - D\theta_0 - g_0(X))D] = 0$) corresponds to treating g_0 as known and regressing $(Y - g_0(X))$ on D.
- The propensity-style moment $\mathbb{E}[(Y - D\theta_0)(D - m_0(X))] = 0$ is related to reweighting based on deviations from the propensity score. This is less central in the paper but natural in the PLR setting.
- The orthogonal score $\psi(W;\theta,g,m) = (Y - \theta D - g(X))(D - m(X))$ is explicitly given in Section 4.1 as a key example of an orthogonal moment.

Slide 7 — Naive regression adjustment: bias decomposition

- The plug-in estimator and its bias decomposition are spelled out in the Introduction. The paper considers sample splitting: estimating g_0 on an auxiliary sample and then forming

$$\hat{\theta}_0 = \left(\sum D_i^2\right)^{-1} \left(\sum D_i(Y_i - \hat{g}_0(X_i))\right).$$

- The scaled estimation error decomposes as

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a + b,$$

where

$$a = (E[D^2])^{-1} \frac{1}{\sqrt{n}} \sum D_i U_i,$$

$$b \approx (E[D^2])^{-1} \frac{1}{\sqrt{n}} \sum m_0(X_i)\{g_0(X_i) - \hat{g}_0(X_i)\}.$$

- The paper notes that ML methods yield rates like $n^{-\varphi_g}$ with $\varphi_g < 1/2$, implying that the term b diverges in general at $\sqrt{n}$-scale, so the estimator fails to be $\sqrt{n}$-consistent. This is discussed under "Regularization bias".

Slide 8 — Orthogonal (residual) score: bias decomposition

- The orthogonalized construction replaces D by its residual $V = D - m_0(X)$, estimated as $\hat{V} = D - \hat{m}(X)$, and defines a DML estimator

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

which can be rearranged into the residual regression form used in the slides.

- The decomposition

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

is sketched in the paper:

  - $a^* = (E[V^2])^{-1} n^{-1/2} \sum V_i U_i$
  - $b^* = (E[V^2])^{-1} n^{-1/2} \sum (\hat{m}(X_i) - m_0(X_i))(\hat{g}(X_i) - g_0(X_i))$
  - c^* is the remainder, controlled via sample splitting.

- The key is that b^* depends on a product of estimation errors and is of order $\sqrt{n} \cdot n^{-(\varphi_m + \varphi_g)}$, which can vanish for $\varphi_m + \varphi_g > 1/2$.

Slide 9 — Neyman orthogonality: definition & PLR example

- Definition 2.1 gives Neyman orthogonality formally: the Gateaux derivative of the moment w.r.t. the nuisance at (θ_0, η_0) is zero for all directions in a realization set.

- The notation in the paper is

$$\partial_\eta \mathbb{E}_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0.$$

- For the PLR score $\psi(W; \theta, g, m) = (Y - \theta D - g(X))(D - m(X))$, the paper notes that this orthogonality condition holds w.r.t. $\eta = (g, m)$, which accounts for the improved robustness (the derivative of the moment w.r.t. misspecification in (g, m) vanishes).

Slide 10 — Residual interpretation & IV-style view

- The residual interpretation $(\tilde{Y} = Y - \hat{g}(X))$, $(\tilde{D} = D - \hat{m}(X))$ and the estimator

$$\tilde{\theta} = \frac{\sum \tilde{D}_i \tilde{Y}_i}{\sum \tilde{D}_i^2}$$

corresponds to a regression of residualized (Y) on residualized (D).

- The paper points out that the DML estimator can be interpreted as a linear IV estimator where the instrument is the residualized treatment $(V = D - m(X))$. This is noted explicitly in the discussion about 'double prediction' and connections to optimal instruments.

- The talk also highlights this "double prediction" viewpoint: one ML problem for (Y) given (X), another for (D) given (X), then regress one residual on the other.

Slide 11 — Sample splitting and cross-fitting

- The problematic remainder terms without sample splitting involve expressions like

$$\sum V_i(\hat{g}(X_i) - g_0(X_i)),$$

  where $\hat{g}$ is estimated on the same sample, so the dependence structure can cause these terms to fail to vanish at the desired rate.

- The paper explains that with sample splitting, we estimate nuisances on an auxiliary sample, and conditionally on that sample, terms like the above have mean zero and variance determined by the squared nuisance error. The variance then goes to zero as the nuisance error shrinks.

- Cross-fitting (swapping main and auxiliary samples or using K-fold versions) restores efficiency while maintaining these favorable properties. This is formalized via DML1 and DML2 in Section 3, but in the slides we avoid naming those explicitly to keep things lighter.

Slide 12–13 — Algorithm for PLR DML

- Definitions 3.1 and 3.2 give formal DML algorithms:
  - DML1 averages fold-specific solutions of orthogonal estimating equations.
  - DML2 defines one global estimator solving a single pooled orthogonal estimating equation built from cross-fitted nuisances.
- The slides give a simplified operational version: K-fold split, fit nuisances out-of-fold, construct residuals, regress residualized outcome on residualized treatment per fold, then average.
- Variance estimation uses the empirical influence function of the orthogonal score; Theorem 3.1 and 3.2 provide asymptotic linearity and show how to estimate the asymptotic variance.

Slide 14 — What double ML delivers (high level)

- Theorem 3.1 and 3.3 show that under approximate Neyman orthogonality and appropriate rate conditions on the nuisance estimators, DML estimators are asymptotically linear and Gaussian:

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum \psi(W_i) + o_p(1),$$

  where ψ is the influence function.

- Under homoscedasticity and with efficient scores, Corollary 3.2 notes that DML achieves the semiparametric efficiency bound. For the PLR model, the orthogonal score we used is efficient under homoscedasticity.

Slide 15 — Simulation: prediction vs causal estimation

- The simulations in the paper and in the talk consider designs where g_0 is particularly suitable for random forests. Figure 1 in the paper compares the conventional (non-orthogonal) ML estimator with the DML estimator:
  - The conventional estimator's histogram is biased and poorly approximated by a normal distribution.
  - The DML estimator's histogram is centered at zero and well approximated by its normal limit.
- The talk also presents Monte Carlo results illustrating this contrast, emphasizing that prediction performance and causal performance can diverge.

Slide 16 — Application: 401(k) eligibility and savings

- Section 6.2 of the paper provides an application to 401(k) eligibility and net financial assets. The outcome and covariates are defined as in the slide:
  - Net financial assets include IRAs, 401(k), checking, bonds, stocks, mutual funds, minus non-mortgage debt.
  - Covariates include age, income, family size, education, marital status, two-earner status, DB pension status, IRA participation, home ownership.

- The identification strategy follows Poterba et al. (1994): conditional on these covariates, eligibility can be treated as exogenous.
- The paper reports DML estimates of the ATE of eligibility on assets, with magnitudes in the several thousand dollars range, robust across a variety of learners. The exact numbers depend on the specification but are in the \$7–\$9k ballpark.

Slide 17 — Beyond PLR: ATE, ATTE, PLIV, LATE

- **ATE/ATTE**: Section 5.1 develops orthogonal scores for average treatment effects and average treatment effects on the treated under unconfoundedness. These scores combine outcome regression and propensity score and are designed to be orthogonal.
- **Partially linear IV model**: Section 4.2 analyzes the model

$$Y = D\theta_0 + g_0(X) + U, \quad \mathbb{E}[U \mid X, Z] = 0; \quad D = m_0(X, Z) + V, \quad \mathbb{E}[V \mid X, Z] = 0.$$

  This enforces the exclusion restriction: conditional on (X), (Z) only affects (Y) through (D). Orthogonal scores use residualized outcome and residualized instrument / endogenous variable.
- **LATE**: Section 5.2 provides orthogonal scores for local average treatment effects with binary (D) and (Z). Nuisances include conditional means $\mu_0(z, x) = \mathbb{E}[Y \mid Z = z, X = x]$, $m_0(z, x) = \mathbb{E}[D \mid Z = z, X = x]$, and the instrument propensity $p_0(X) = \mathbb{E}[Z \mid X]$.

Slide 18 — What DML does not fix

- The paper repeatedly stresses that DML works under the same identification assumptions as the underlying causal models:
  - PLR: conditional exogeneity given (X).
  - ATE/ATTE: unconfoundedness given (X).
  - PLIV/LATE: valid instruments with appropriate structural assumptions.
- DML does not address unobserved confounding; it only allows flexible nonparametric estimation of the observed-nuisance structure. The need for good research design and careful control selection is highlighted especially in the applications section.

Slide 19 — Evidence from method evaluation studies

- While the specific 'Estimating Causal Effects with Double Machine Learning – A Method Evaluation' paper is not in the uploaded files, its findings are consistent with broader simulation evidence:
  - DML with rigid learners (e.g. plain lasso in misspecified designs) can inherit bias similar to OLS.
  - DML with flexible learners tends to reduce bias in nonlinear confounding scenarios.
- The DML paper's empirical and simulation results show that choice of learner inside DML affects performance: see the 401(k) example where different ML methods give similar but not identical estimates, and the simulation figures where non-orthogonal estimators perform poorly even with strong predictors.
- This supports the message that DML is not a magic bullet: ML choice heavily influences finite-sample performance.

Slide 20 — Checklist & take-home message

- The general abstract theory in Sections 3 and 5 shows that as long as nuisance estimators achieve certain rates and orthogonality holds, DML yields valid asymptotic inference for low-dimensional θ_0.
- In practice, the workflow 'identify estimand, derive orthogonal score, choose ML, cross-fit, do sensitivity checks' is exactly what the paper suggests via its repeated pattern across models (PLR, PLIV, ATE/ATTE, LATE).
- Victor Chernozhukov's talk also emphasizes this as a reusable template: two prediction problems for nuisances, one orthogonal estimating equation, sample splitting and averaging.

Meta: talk-structure choices vs. guidelines

- The structure of the talk (early statement of key idea, emphasis on a single central idea, use of one running model with one main algorithm) follows typical advice for research talks:
  - Communicate one key idea; do not try to present all technical details.
  - Prefer depth over breadth, and use examples (PLR, 401(k)) to ground intuition.
  - Keep must-have parts (motivation, key result, main example) within the allotted time, and be ready to truncate advanced material if needed.

These notes should put you in a good position to answer deeper questions about:

- why naive plug-in fails ($\sqrt{n}$-bias decomposition),
- how orthogonality is defined and checked,
- why cross-fitting is necessary,
- how the method extends beyond PLR,
- and what the empirical implications are.