Orthogonal Machine Learning: Power and Limitations

Lester Mackey*

Joint work with Vasilis Syrgkanis* and Ilias Zadik†

Microsoft Research New England*, Massachusetts Institute of Technology†

October 30, 2018
Vasilis: Lester, I love Double Machine Learning!

Me: What?

Vasilis: It’s a tool for accurately estimating treatment effects in the presence of many potential confounders.

Me: I have no idea what you’re talking about.

Vasilis: Let me give you an example...
Example: Estimating Price Elasticity of Demand

**Goal:** Estimate *elasticity*, the effect a change in price has on demand

- Set prices of goods and services [Chernozhukov, Goldman, Semenova, and Taddy, 2017b]
- Predict impact of tobacco tax on smoking [Wilkins, Yurekli, and Hu, 2004]

\[
\log \text{demand} = \theta_0 \times \text{elasticity} + T \times \log \text{price} + \epsilon
\]
Example: Estimating Price Elasticity of Demand

**Goal:** Estimate *elasticity*, the effect a change in price has on demand

- Set prices of goods and services [Chernozhukov, Goldman, Semenova, and Taddy, 2017b]
- Predict impact of tobacco tax on smoking [Wilkins, Yurekli, and Hu, 2004]

\[ \log \text{demand} = \theta_0 \text{elasticity} T \log \text{price} + \epsilon \text{noise} \]

**Conclusion:** Increasing price increases demand!

**Problem:** Demand increases in winter & price anticipates demand
Example: Estimating Price Elasticity of Demand

**Goal:** Estimate *elasticity*, the effect a change in price has on demand

- Set prices of goods and services [Chernozhukov, Goldman, Semenova, and Taddy, 2017b]
- Predict impact of tobacco tax on smoking [Wilkins, Yurekli, and Hu, 2004]

\[ \log \text{demand} = \theta_0 \text{elasticity} + \beta_0 \log \text{price} + \beta_0 \text{season indicator} + \epsilon \text{noise} \]

**Problem:** What if there are 100s or 1000s of potential confounders?
Example: Estimating Price Elasticity of Demand

**Goal:** Estimate *elasticity*, the effect a change in price has on demand

**Problem:** What if there are 100s or 1000s of potential confounders?

- Time of day, day of week, month, purchase and browsing history, other product prices, demographics, the weather, ...

**One option:** Estimate effect of all potential confounders really well

\[
\begin{align*}
Y &= \theta_0 \cdot T + f_0(X) + \epsilon \\
Y \quad \text{log demand} & \quad \theta_0 \quad \text{elasticity} & \quad T \quad \text{log price} & \quad f_0(X) \quad \text{effect of potential confounders} & \quad \epsilon \quad \text{noise}
\end{align*}
\]

- If nuisance function \( f_0 \) estimable at \( O(n^{-1/2}) \) rate then so is \( \theta_0 \)

**Problem:** Accurate nuisance estimates often unachievable when \( f_0 \) nonparametric or linear and high-dimensional
Example: Estimating Price Elasticity of Demand

**Problem:** What if there are 100s or 1000s of potential confounders?

**Double Machine Learning** [Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017a]

\[
Y = \theta_0 T + f_0(X) + \epsilon
\]

- Estimate nuisance \( f_0 \) somewhat poorly: \( o(n^{-1/4}) \) suffices
- Employ *Neyman orthogonal* estimator of \( \theta_0 \) robust to first-order errors in nuisance estimates; yields \( \sqrt{n} \)-consistent estimate of \( \theta_0 \)

**Questions:** Why \( o(n^{-1/4}) \)? Can we relax this? When? How?

**This talk:**

- Framework for \( k \)-th order orthogonal estimation with \( o(n^{-1/(2k+2)}) \) nuisance consistency \( \Rightarrow \sqrt{n} \)-consistency for \( \theta_0 \)
- Existence characterization and explicit construction of 2nd-order orthogonality in a popular causal inference model
**Goal:** Estimate target parameters $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ (e.g., elasticities) in the presence of unknown nuisance functions $h_0 \in \mathcal{H}$

**Given**
- Independent replicates $(Z_t)_{t=1}^{2n}$ of a data vector $Z = (T, Y, X)$

**Example (Partially Linear Regression (PLR))**
- $T \in \mathbb{R}$ represents a treatment or policy applied (e.g., log price)
- $Y \in \mathbb{R}$ represents an outcome of interest (e.g., log demand)
- $X \in \mathbb{R}^p$ is a vector of associated covariates (e.g., seasonality)

These observations satisfy

$$
Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon | X, T] = 0 \quad a.s.
$$

$$
T = g_0(X) + \eta, \quad \mathbb{E}[(\eta | X] = 0 \quad a.s., \quad \text{Var}(\eta) > 0
$$

for noise $\eta$ and $\epsilon$, target parameter $\theta_0$, and nuisance $h_0 = (f_0, g_0)$.  

Two-stage $Z$-estimation with Sample Splitting

**Goal:** Estimate target parameters $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ (e.g., elasticities) in the presence of unknown nuisance functions $h_0 \in \mathcal{H}$

**Given**

- Independent replicates $(Z_t)_{t=1}^{2n}$ of a data vector $Z = (T, Y, X)$
- Moment functions $m$ that identify the target parameters $\theta_0$:
  \[
  \mathbb{E}[m(Z, \theta_0, h_0(X))|X] = 0 \quad \text{a.s. and} \quad \mathbb{E}[m(Z, \theta, h_0(X))] \neq 0 \quad \text{if} \quad \theta \neq \theta_0
  \]
- PLR model example: $m(Z, \theta, h_0(X)) = (Y - \theta T - f_0(X))^T$

**Two-stage $Z$-estimation with sample splitting**

1. Fit estimate $\hat{h} \in \mathcal{H}$ of $h_0$ using $(Z_t)_{t=n+1}^{2n}$ (e.g., via nonparametric or high-dimensional regression)

2. $\hat{\theta}^{SS}$ solves
  \[
  \frac{1}{n} \sum_{t=1}^{n} m(Z_t, \theta, \hat{h}(X_t)) = 0
  \]

**Con:** Splitting statistically inefficient, possible detriment in first stage
Two-stage $Z$-estimation with Cross Fitting

**Goal:** Estimate target parameters $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ (e.g., elasticities) in the presence of unknown nuisance functions $h_0 \in \mathcal{H}$

**Given**
- Independent replicates $(Z_t)_{t=1}^{2n}$ of a data vector $Z = (T, Y, X)$
- Moment functions $m$ that identify the target parameters $\theta_0$:
  \[
  \mathbb{E}[m(Z, \theta_0, h_0(X))|X] = 0 \text{ a.s. and } \mathbb{E}[m(Z, \theta, h_0(X))] \neq 0 \text{ if } \theta \neq \theta_0
  \]
- PLR model example: $m(Z, \theta, h_0(X)) = (Y - \theta T - f_0(X))T$

**Two-stage $Z$-estimation with cross fitting**

[Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017a]

0. Split data indices into $K$ batches $I_1, \ldots, I_K$
1. For $k \in \{1, \ldots, K\}$, fit estimate $\hat{h}_k \in \mathcal{H}$ of $h_0$ excluding $I_k$
2. $\hat{\theta}^{CF}$ solves $\frac{1}{n} \sum_{k=1}^{K} \sum_{t \in I_k} m(Z_t, \theta, \hat{h}_k(X_t)) = 0$

**Pro:** Repairs sample splitting deficiencies
Goal: $\sqrt{n}$-Asymptotic Normality

**Two-stage $Z$-estimators**

- $\hat{\theta}^{SS}$ solves $\frac{1}{n} \sum_{t=1}^{n} m(Z_t, \theta, \hat{h}(X_t)) = 0$
- $\hat{\theta}^{CF}$ solves $\frac{1}{n} \sum_{k=1}^{K} \sum_{t \in I_k} m(Z_t, \theta, \hat{h}_k(X_t)) = 0$

**Goal:** Establish conditions under which $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$ enjoy $\sqrt{n}$-asymptotic normality ($\sqrt{n}$-a.n.), that is

$$\sqrt{n}(\hat{\theta}^{SS} - \theta_0) \xrightarrow{d} N(0, \Sigma) \text{ and } \sqrt{2n}(\hat{\theta}^{CF} - \theta_0) \xrightarrow{d} N(0, \Sigma)$$

- Asymptotically valid confidence intervals for $\theta_0$ based on Gaussian or Student’s t quantiles
- Asymptotically valid association tests, like the Wald test
First-order Orthogonality

Definition (First-order Orthogonal Moments)

[Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017a]

Moments $m$ are first-order orthogonal w.r.t. the nuisance $h_0(X)$ if

$$
\mathbb{E}\left[ \nabla_\gamma m(Z, \theta_0, \gamma)|_{\gamma=h_0(X)} \mid X \right] = 0.
$$

- Principle dates back to early work of [Neyman, 1979]
- Grants first-order insensitivity to errors in nuisance estimates
  - Annihilates first-order term in Taylor expansion around nuisance
  - Recall: $m$ is 0-th order orthogonal, $\mathbb{E}[m(Z, \theta_0, h_0(X)) \mid X] = 0$
- Not satisfied by $m(Z, \theta, h(X)) = (Y - \theta T - f(X))T$
- Satisfied by $m(Z, \theta, h(X)) = (Y - \theta T - f(X))(T - g(X))$

Main result of Chernozhukov et al. [2017a]: under 1st-order orthogonality, $\hat{\theta}^{SS}, \hat{\theta}^{CF}$ \(\sqrt{n}\)-a.n. when $\|\hat{h}_i - h_{0,i}\| = o_p(n^{-1/4}), \forall i$
Higher-order Orthogonality

**Definition ($k$-Orthogonal Moments)**

Moments $m$ are $k$-orthogonal, if for all $\alpha \in \mathbb{N}^\ell$ with $\|\alpha\|_1 \leq k$:

$$
\mathbb{E} \left[ D^\alpha m(Z, \theta_0, \gamma) \big| \gamma = h_0(X) \right] = 0.
$$

where

$$
D^\alpha m(Z, \theta, \gamma) = \nabla_{\gamma_1}^{\alpha_1} \nabla_{\gamma_2}^{\alpha_2} \ldots \nabla_{\gamma_\ell}^{\alpha_\ell} m(Z, \theta, \gamma)
$$

and the $\gamma_i$’s are the coordinates of the $\ell$ nuisance functions.

- Grants $k$-th-order insensitivity to errors in nuisance estimates
- Annihilates terms with order $\leq k$ in Taylor expansion around nuisance
Asymptotic Normality from $k$-Orthogonality

**Theorem ([Mackey, Syrgkanis, and Zadik, 2018])**

Under $k$-orthogonality and standard identifiability and regularity assumptions, $\|\hat{h}_i - h_{0,i}\| = o_p(n^{-1/(2k+2)})$ for all $i$ suffices for $\sqrt{n}$-a.n. of $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$ with $\Sigma = J^{-1}VJ^{-1}$ for $J = \mathbb{E}[\nabla_\theta m(Z, \theta_0, h_0(X))]$ and $V = \text{Cov}(m(Z, \theta_0, h_0(X)))$.

- Actually suffices to have **product** of nuisance function errors decay $(n^{1/2} \cdot \sqrt{\mathbb{E}[\prod_{i=1}^\ell |\hat{h}_i(X) - h_{0,i}(X)|^{2\alpha_i} | \hat{h}] \rightarrow 0}$ for $\|\alpha\|_1 = k + 1$): if one is more accurately estimated, another can be estimated more crudely.
- We prove similar results for non-uniform orthogonality.
- $o_p(n^{-1/(2k+2)})$ rate holds the promise of coping with more complex or higher-dimensional nuisance functions.

**Question:** How do we construct $k$-orthogonal moments in practice?
**Second-order Orthogonality for PLR: Limitations**

**Question:** Can we construct $k$-orthogonal moments in practice?

\[
Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad a.s.
\]

\[
T = g_0(X) + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad a.s., \quad \text{Var}(\eta) > 0
\]

**Theorem ([Mackey, Syrgkanis, and Zadik, 2018])**

*Suppose the conditional distribution of $\eta$ given $X$ is a.s. Gaussian.*

*Then no 2-orthogonal twice differentiable $m$ yields $\sqrt{n}$-consistency.*

- We use Stein’s lemma (\(\mathbb{E}[q'(Z)] = \mathbb{E}[Zq(Z)]\) for \(Z \sim N(0, 1)\)) to show 2-orthogonality implies \(\mathbb{E}[\nabla_\theta m(Z, \theta_0, h_0(X))] = 0\) and hence infinite asymptotic variance for the $Z$-estimator.

- Sad, but non-Gaussian residuals are common in pricing where $T = \log$ price, and $\eta$ is a random log percentage discount (25% off now through Sunday!) over the log baseline price $g_0(X)$. 
Question: How do we construct $k$-orthogonal moments in practice?

$$Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad a.s.$$  

$$T = g_0(X) + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad a.s., \quad \text{Var}(\eta) > 0$$

**Exploit non-Gaussianity:** $\eta$ conditionally Gaussian given $X \Leftrightarrow \mathbb{E}[\eta^{r+1} \mid X] = r\mathbb{E}[\eta^2 \mid X]\mathbb{E}[\eta^{r-1} \mid X]$ for all $r \in \mathbb{N}$

**Theorem ([Mackey, Syrgkanis, and Zadik, 2018])**

Suppose that, for some $r \in \mathbb{N}$, $\mathbb{E}[\eta^{r+1}] \neq r\mathbb{E}[\mathbb{E}[\eta^2 \mid X]\mathbb{E}[\eta^{r-1} \mid X]]$. If we know $\mathbb{E}[\eta^r \mid X]$, then the 2-orthogonal moments

$$m(Z, \theta, q(X), g(X), \mu_{r-1}(X))$$

$$\triangleq (Y - q(X) - \theta(T - g(X)))$$

$$\times ((T - g(X))^r - \mathbb{E}[\eta^r \mid X] - r(T - g(X))\mu_{r-1}(X))$$

satisfy our standard identifiability and regularity conditions.

- $o(n^{-1/6})$ nuisance estimation error suffices for $\sqrt{n}$-a.n.
Question: How do we construct $k$-orthogonal moments in practice?

\[ Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad \text{a.s.} \]

\[ T = g_0(X) + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad \text{a.s.,} \quad \text{Var}(\eta) > 0 \]

**Exploit non-Gaussianity:** \( \eta \) conditionally Gaussian given \( X \) \( \iff \)

\[ \mathbb{E}[\eta^{r+1} \mid X] = r \mathbb{E}[\eta^2 \mid X] \mathbb{E}[\eta^{r-1} \mid X] \quad \text{for all} \quad r \in \mathbb{N} \]

**Theorem ([Mackey, Syrgkanis, and Zadik, 2018])**

Suppose that, for some \( r \in \mathbb{N} \), \( \mathbb{E}[\eta^{r+1}] \neq r \mathbb{E}[\mathbb{E}[\eta^2 \mid X] \mathbb{E}[\eta^{r-1} \mid X]] \).

Then, except for the \((q(X), \mu_r(X))\) and \((g(X), \mu_r(X))\) pairings,

\[
m(Z, \theta, q(X), g(X), \mu_{r-1}(X), \mu_r(X))
\]

\[
\triangleq (Y - q(X) - \theta(T - g(X))) \times ((T - g(X))^r - \mu_r(X) - r(T - g(X))\mu_{r-1}(X))
\]

is 2-orthogonal and satisfies our standard conditions.

- \( o(n^{-1/3}) \) error for \( \mu_r(X) \) and \( o(n^{-1/6}) \) for rest suffice for \( \sqrt{n} \)-a.n.
High-dimensional Linear Nuisance Setting

\[ Y = \theta_0 T + \langle X, \beta_0 \rangle + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad a.s. \]
\[ T = \langle X, \gamma_0 \rangle + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad a.s., \quad \text{Var}(\eta) > 0 \]

- \( \beta_0, \gamma_0 \in \mathbb{R}^p \) are s-sparse, \((\eta, \epsilon, X)\) independent, \( q_0 = \theta_0 \beta_0 + \gamma_0 \)

How many relevant confounders (non-zeros) can we tolerate?

- Lasso can estimate \( \beta_0, \gamma_0 \) with \( O(\sqrt{s \log p/n}) \) error
- Zeroth-order orthogonality rate \( O(n^{-1/2}) \): \( s = O(1/\log p) \)
  - \( m = (Y - \theta T - \langle X, \beta \rangle)T \)
- First-order orthogonality rate \( o(n^{-1/4}) \): \( s = o(n^{1/2}/\log p) \)
  [Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017a]
  - \( m = (Y - \theta T - \langle X, \beta \rangle)(T - \langle X, \gamma \rangle) \)
  - \( m = (Y - \langle X, q \rangle - \theta(T - \langle X, \gamma \rangle))(T - \langle X, \gamma \rangle) \)
PLR with High-dimensional Linear Nuisance

High-dimensional Linear Nuisance Setting

\[ Y = \theta_0 T + \langle X, \beta_0 \rangle + \epsilon, \quad E[\epsilon \mid X, T] = 0 \quad a.s. \]

\[ T = \langle X, \gamma_0 \rangle + \eta, \quad E[\eta \mid X] = 0 \quad a.s., \quad \text{Var}(\eta) > 0 \]

- \( \beta_0, \gamma_0 \in \mathbb{R}^p \) are \( s \)-sparse, \( (\eta, \epsilon, X) \) independent, \( q_0 = \theta_0 \beta_0 + \gamma_0 \)

---

**Theorem ([Mackey, Syrgkanis, and Zadik, 2018])**

Suppose \( E[\eta^4] \neq 3E[\eta^2]^2 \), \( X \) has i.i.d. \( N(0, 1) \) entries, \( \epsilon \) and \( \eta \) are bounded by \( C \), and \( \theta_0 \in [-M, M] \). If \( s = o(n^{2/3}/\log p) \), and we

(a) estimate \( q_0, \gamma_0 \) via Lasso with \( \lambda_n = 2CM \sqrt{3 \log(p)/n} \) and

(b) estimate \( E[\eta^2] \) and \( E[\eta^3] \) using \( \hat{\eta}_t \triangleq T'_t - \langle X'_t, \hat{\gamma} \rangle \),

\[ \hat{\mu}_2 = \frac{1}{n} \sum_{t=1}^{n} \hat{\eta}_t^2, \quad \text{and} \quad \hat{\mu}_3 = \frac{1}{n} \sum_{t=1}^{n} (\hat{\eta}_t^3 - 3\hat{\mu}_2 \hat{\eta}_t), \]

for \( (T'_t, X'_t)_{t=1}^n \) an i.i.d. sample independent of \( \hat{\gamma} \),

then the moments \( m = (Y - \langle X, q \rangle - \theta(T - \langle X, \gamma \rangle)) \times ((T - \langle X, \gamma \rangle)^3 - \mu_3 - 3(T - \langle X, \gamma \rangle)\mu_2) \) yield \( \sqrt{n} \)-a.n.
High-dimensional Linear Nuisance Setting

\[ Y = \theta_0 T + \langle X, \beta_0 \rangle + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad a.s. \]
\[ T = \langle X, \gamma_0 \rangle + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad a.s., \quad \text{Var}(\eta) > 0 \]

- \( \beta_0, \gamma_0 \in \mathbb{R}^p \) are \( s \)-sparse, \( (\eta, \epsilon, X) \) independent, \( q_0 = \theta_0 \beta_0 + \gamma_0 \)
- Mimic price elasticity of demand setting: \( T \) represents log price and \( \eta \) drawn from discrete distribution representing random (log) discounts over baseline price
1st (top) vs. 2nd order, $s = 100$, $n = 5000$, $p = 1000$, $\theta_0 = 3$. 
1st vs. 2nd order, $n = 5000$, $p = 1000$, $\theta_0 = 3$.  

![Bias and STD plots](image)
High-dimensional PLR: Varying Sparsity

1st vs. 2nd order, $n = 5000$, $p = 1000$, $\theta_0 = 3$. 

![Graph showing MSE and First Stage L2 error with varying support size.](image)
High-dimensional PLR: MSE for Varying $n, p, s$

$n = 10000, p = 1000$ and $n = 5000, p = 2000$
High-dimensional PLR: Varying Noise Level

\[ n = 5000, p = 1000 \]

\[ \sigma_\epsilon = 10 \]

\[ \sigma_\epsilon = 20 \]
Recap

What have we accomplished?

1. Introduced a notion of $k$-orthogonality for two-stage $Z$-estimation with nuisance, generalizing Neyman orthogonality
2. Showed that $o(n^{-\frac{1}{2k+2}})$ nuisance estimate error suffices for $\sqrt{n}$-asymptotic normality of target parameters
3. Established that non-normality of $\eta|X$ necessary for the existence of useful 2-orthogonal moments in PLR model
4. Derived explicit 2-orthogonal moments for PLR given knowledge of non-normality
5. Used 2-orthogonal moments to tolerate $o\left(\frac{n^{\frac{3}{2}}}{\log p}\right)$ sparsity in high-dimensional PLR
6. Showed benefits over standard $o\left(\frac{n^{\frac{1}{2}}}{\log p}\right)$ first-order orthogonal moments in synthetic demand estimation experiments
Many opportunities for future development

1. Second-order orthogonality
   - How to select optimal / improved double orthogonal moments
   - How to construct moments for other causal inference models

2. $k$-th order orthogonality for $k > 2$
   - When are $k$-th order orthogonal moments available and useful?
   - How do we construct them explicitly?

3. Lower bounds: (non-Gaussian) examples where first-order orthogonality provably worse than second-order orthogonality


5. Applications to problems with non-Gaussian treatment residuals


Experiment Specification

- \( \eta \) is drawn from a discrete distribution with values \( \{0.5, 0, -1.5, -3.5\} \) taken with probabilities (.65, .2, .1, .05).
- \( \epsilon \) is drawn independently from a uniform \( U(-\sigma_\epsilon, \sigma_\epsilon) \) distribution.
- Importantly, the coordinates of the \( s \) non-zero entries of the coefficient \( \beta_0 \) are the same as the coordinates of the \( s \) non-zero entries of \( \gamma_0 \).
- Each non-zero coefficient was generated independently from a uniform \( U(0, 5) \) distribution.
- The regularization parameter \( \lambda_n \) of each Lasso was \( \sqrt{\log(p)/n} \).
- For each instance of the problem, i.e., each random realization of the coefficients, we generated 2000 independent datasets to estimate the bias and standard deviation of each estimator. We repeated this process over 100 randomly generated problem instances, each time with a different draw of the coefficients \( \gamma_0 \) and \( \beta_0 \), to evaluate variability across different realizations of the nuisance functions.