Orthogonal Machine Learning: Power and Limitations

Ilias Zadik\textsuperscript{1}, joint work with Lester Mackey\textsuperscript{2}, Vasilis Syrgkanis\textsuperscript{2}

\textsuperscript{1}Massachusetts Institute of Technology (MIT) and
\textsuperscript{2}Microsoft Research New England (MSRNE)

35th International Conference on Machine Learning (ICML) 2018
Main Application: Pricing a product in the digital economy!
Main Application: Pricing a product in the digital economy!

Simple: Plot Demand and Price and run Linear Regression:
**Introduction**

*Challenge:* What if the bottom points are from the Summer and upper from the Winter? Then new Linear Regression:

![Graph showing linear regression between demand and price]
Introduction

Challenge: What if the bottom points are from the Summer and upper from the Winter? Then new Linear Regression:

In current reality, thousands of confounders like seasonality simultaneously affect price and demand; How do we price correctly?
The Partially Linear Regression Problem (PLR)

Definition (Partially Linear Regression (PLR))

Let \( p \in \mathbb{N}, \theta_0 \in \mathbb{R}, f_0, g_0 : \mathbb{R}^p \rightarrow \mathbb{R} \).

- \( T \in \mathbb{R} \) treatment or policy applied [e.g. price],
- \( Y \in \mathbb{R} \) outcome of interest [e.g. demand],
- \( X \in \mathbb{R}^p \) vector of associated covariates [e.g. seasonality...].

Related by

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

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

where \( \eta, \epsilon \) represent noise variables.
The Partially Linear Regression Problem (PLR)

**Definition (Partially Linear Regression (PLR))**

Let $p \in \mathbb{N}$, $\theta_0 \in \mathbb{R}$, $f_0, g_0 : \mathbb{R}^p \to \mathbb{R}$.

- $T \in \mathbb{R}$ treatment or policy applied [e.g. price],
- $Y \in \mathbb{R}$ outcome of interest [e.g. demand],
- $X \in \mathbb{R}^p$ vector of associated covariates [e.g. seasonality..].

Related by

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

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

where $\eta, \epsilon$ represent noise variables.

**Goal:** Given $n$ iid samples of $(Y_i, T_i, X_i), i = 1, \ldots, n$ find a $\sqrt{n}$-consistent asymptotically normal ($\sqrt{n}$-a.n.) estimator of $\theta_0$; $\sqrt{n}(\hat{\theta}_0 - \theta_0) \to \mathcal{N}(0, \sigma^2)$.
Hence, n samples for learning \( \theta_0 \), **but** ...
Hence, n samples for learning $\theta_0$, \textbf{but} ...

**Challenge 1:** Except $\theta_0$ we also need to learn $f_0, g_0$!
Hence, n samples for learning $\theta_0$, but ...

**Challenge 1:** Except $\theta_0$ we also need to learn $f_0, g_0$!

**Challenge 2:** And we do not really want to spend too many samples learning them (more than necessary to estimate $\theta_0$!)
Hence, \( n \) samples for learning \( \theta_0 \), **but** ... 

**Challenge 1:** Except \( \theta_0 \) we also need to learn \( f_0, g_0 \)!

**Challenge 2:** And we do not really want to spend too many samples learning them (more than necessary to estimate \( \theta_0 \! \))

**Main Question:** What is the optimal learning rate of the nuisance functions \( f_0, g_0 \) so that we get a \( \sqrt{n} \)-a.n. estimator of \( \theta_0 \)?

*Maximum \( a_n \) so that \( \| \hat{f}_0 - f_0 \|, \| \hat{g}_0 - g_0 \| = o(a_n) \) suffices.*
• Trivial Rate, learn $f_0, g_0$ at $n^{-\frac{1}{2}}$-rate.
• Trivial Rate, learn \( f_0, g_0 \) at \( n^{-\frac{1}{2}} \)-rate.

• [Chernozhukov et al, 2017]: It suffices to learn \( f_0, g_0 \) at \( n^{-\frac{1}{4}} \)-rate to construct a \( \sqrt{n} \)-a.n. estimator of \( \theta_0 \).
Literature Review

- Trivial Rate, learn $f_0, g_0$ at $n^{-\frac{1}{2}}$-rate.
- [Chernozhukov et al, 2017]: It suffices to learn $f_0, g_0$ at $n^{-\frac{1}{4}}$-rate to construct a $\sqrt{n}$-a.n. estimator of $\theta_0$.

The technique is based on

- Generalized Method of Moments (Z-estimation)
- with a “First Order Orthogonal Moment”.
Choose \( m \) such that \( \mathbb{E}[m(Y, T, f_0(X), g_0(X), \theta_0) | X] = 0 \), a.s..
Z-Estimation for PLR

Choose \( m \) such that \( \mathbb{E} [ m(Y, T, f_0(X), g_0(X), \theta_0) | X] = 0 \), a.s..

Given \( n \) samples \( Z_i = (X_i, T_i, Y_i) \),

- **(Stage 1)** Use \( Z_{n+1}, \ldots, Z_{2n} \) samples to form \( \hat{f}_0, \hat{g}_0 \sim f_0, g_0 \).
- **(Stage 2)** Use \( Z_1, \ldots, Z_n \) to find \( \hat{\theta}_0 \) by solving

\[
\frac{1}{n} \sum_{t=1}^{n} m(T_t, Y_t, \hat{f}_0(X_t), \hat{g}_0(X_t), \hat{\theta}_0) = 0.
\]

Chernozhukov et al, 2017 suggests a simple first-order orthogonal moment

\( m(Y, T, f(X), g(X), \theta) = (Y - \theta T - f(X))(T - g(X)) \)

for PLR. For this choice \( n - \frac{1}{4} \) first stage error suffices!
Z-Estimation for PLR

Choose \( m \) such that \( \mathbb{E} [m(Y, T, f_0(X), g_0(X), \theta_0)|X] = 0 \), a.s..

Given \( n \) samples \( Z_i = (X_i, T_i, Y_i) \),

- **(Stage 1)** Use \( Z_{n+1}, \ldots, Z_{2n} \) samples to form \( \hat{f}_0, \hat{g}_0 \sim f_0, g_0 \).
- **(Stage 2)** Use \( Z_1, \ldots, Z_n \) to find \( \hat{\theta}_0 \) by solving

\[
\frac{1}{n} \sum_{t=1}^{n} m(T_t, Y_t, \hat{f}_0(X_t), \hat{g}_0(X_t), \hat{\theta}_0) = 0.
\]

[Chernozhukov et al, 2017] suggests a **simple first-order orthogonal moment**

\[
m(Y, T, f(X), g(X), \theta) = (Y - \theta T - f(X)) (T - g(X))
\]

for PLR. For this choice \( n^{-\frac{1}{4}} \) first stage error suffices!
Definition (First-Order Orthogonality)

A moment \( m : \mathbb{R}^p \to \mathbb{R} \) is first-order orthogonal with respect to the nuisance function if

\[
\mathbb{E} \left[ \nabla_{\gamma} m(Y, T, \gamma, \theta_0) \big|_{\gamma=(f_0(X),g_0(X))} \big| X \right] = 0.
\]
Definition (First-Order Orthogonality)

A moment $m : \mathbb{R}^p \rightarrow \mathbb{R}$ is first-order orthogonal with respect to the nuisance function if

$$\mathbb{E} \left[ \nabla_{\gamma} m(Y, T, \gamma, \theta_0) \big|_{\gamma=(f_0(X),g_0(X))} \middle| X \right] = 0.$$ 

**Intuition:** Low sensitivity to first stage errors because of Taylor Expansion!

**Question 1:** Can we generalize to higher order orthogonality? Will this improve the first stage error we can tolerate?

**Question 2:** Does higher order orthogonal moments exist for PLR?
**Definition (First-Order Orthogonality)**

A moment $m : \mathbb{R}^p \rightarrow \mathbb{R}$ is first-order orthogonal with respect to the nuisance function if

$$\mathbb{E} \left[ \nabla_{\gamma} m(Y, T, \gamma, \theta_0) \mid \gamma = (f_0(X), g_0(X)) \mid X \right] = 0.$$  

*Intuition:* Low sensitivity to first stage errors because of Taylor Expansion!

**Question 1:** Can we generalize to higher order orthogonality? Will this improve the first stage error we can tolerate?
Z-Estimation for PLR: Comments

Definition (First-Order Orthogonality)

A moment $m: \mathbb{R}^p \rightarrow \mathbb{R}$ is first-order orthogonal with respect to the nuisance function if

$$E \left[ \nabla_\gamma m(Y, T, \gamma, \theta_0) \big| \gamma = (f_0(X), g_0(X)) \big| X \right] = 0.$$

Intuition: Low sensitivity to first stage errors because of Taylor Expansion!

Question 1: Can we generalize to higher order orthogonality? Will this improve the first stage error we can tolerate?

Question 2: Does higher order orthogonal moments exist for PLR?
Definition: Higher-Order Orthogonality

Let \( k \in \mathbb{N} \).

**Definition (k-Orthogonal Moment)**

The moment condition is called \( k \)-orthogonal, if for any \( \alpha \in \mathbb{N}^2 \) with \( \alpha_1 + \alpha_2 \leq k \):

\[
\mathbb{E} [D^\alpha m(Y, T, f_0(X), g_0(X), \theta_0)|X] = 0.
\]

where

\[
D^\alpha = \nabla^{\alpha_1} \gamma_1 \nabla^{\alpha_2} \gamma_2
\]

and \( \gamma_i \)'s are the coordinates of the nuisance \( f_0, g_0 \).
Main Result on k-Orthogonality: $n^{-\frac{1}{2k+2}}$ rate suffices!

**Theorem (informal)**

Let $m$ be a moment which is $k$-orthogonal and satisfies certain identifiability and smoothness assumptions. Then if the Stage 1 error of estimating $f_0, g_0$ is

$$o(n^{-\frac{1}{2k+2}}),$$

the solution to the Stage 2 equation $\hat{\theta}_0$ is a $\sqrt{n}$-a.n. estimator of $\theta_0$. 

**Comments:**

- The existence of a smooth $k$-orthogonal moment implies $n^{-\frac{1}{2k+2}}$ nuisance error suffices!
- The proof is based on a careful higher-order Taylor Expansion argument.
- The original Theorem deals with a much more general case of GMM than PLR (Come to Poster for details!)
Main Result on k-Orthogonality: $n^{-\frac{1}{2k+2}}$ rate suffices!

**Theorem (informal)**

Let $m$ be a moment which is $k$-orthogonal and satisfies certain identifiability and smoothness assumptions. Then if the Stage 1 error of estimating $f_0, g_0$ is

$$o(n^{-\frac{1}{2k+2}}),$$

the solution to the Stage 2 equation $\hat{\theta}_0$ is a $\sqrt{n}$-a.n. estimator of $\theta_0$.

**Comments:**

- The existence of a smooth $k$-orthogonal moment implies $n^{-\frac{1}{2k+2}}$ nuisance error suffices!
- The proof is based on a careful higher-order Taylor Expansion argument.
- The original Theorem deals with a much more general case of GMM than PLR (Come to Poster for details!)
Question: Can we construct a 2-orthogonal moment for PLR?
2-orthogonal moment for PLR: A Gaussianity Issue

**Question:** Can we construct a 2-orthogonal moment for PLR?

**Gist of the Result:**
Yes **if and only if** the treatment residual $\eta|X$ is **not** normally distributed!
Limitation: No if \( \eta | X \) is normally distributed!
Limitation: **No** if $\eta|X$ is normally distributed!

**Theorem (informal)**

Assume $\eta|X$ is normally distributed. Then there is no $m$ which is
- 2-orthogonal
- satisfies certain identifiability and smoothness assumptions and,
- the solution of Stage 2 satisfies $\hat{\theta}_0 - \theta_0 = O_P(\frac{1}{\sqrt{n}})$. 

The proof is based on Stein's Lemma: $E[q'(Z)] = E[Zq(Z)]$ for $Z \sim N(0, 1)$, which allows us to algebraically connect 2-orthogonality with the asymptotic variance of $\hat{\theta}_0$!
**Limitation:** No if $\eta|X$ is normally distributed!

**Theorem (informal)**

Assume $\eta|X$ is normally distributed. Then there is no $m$ which is

- 2-orthogonal
- satisfies certain identifiability and smoothness assumptions and,
- the solution of Stage 2 satisfies $\hat{\theta}_0 - \theta_0 = O_P(\frac{1}{\sqrt{n}})$. 

The proof is based on Stein’s Lemma: $E[q'(Z)] = E[Zq(Z)]$ for $Z \sim N(0, 1)$, which allows us to **connect algebraically** 2-orthogonality with the asymptotic variance of $\hat{\theta}_0$!
2-orthogonal moment for PLR? Power!

*Power:* Yes if $\eta|X$ is not normally distributed!
2-orthogonal moment for PLR? Power!

*Power:* Yes if $\eta|X$ is not normally distributed!

Technical Detail before Theorem: We need to change nuisance from $f_0, g_0$ to $q_0 = \theta_0 g_0 + f_0, g_0$ for our positive result.
**2-orthogonal moment for PLR? Power!**

*Power:* **Yes** if \( \eta|X \) is **not** normally distributed!

Technical Detail before Theorem: We need to change nuisance from \( f_0, g_0 \) to \( q_0 = \theta_0 g_0 + f_0, g_0 \) for our positive result.

---

**Theorem**

*Under the PLR model, suppose that we know \( \mathbb{E}[\eta^r|X], \mathbb{E}[\eta^{r-1}|X] \) and that \( \mathbb{E}[\eta^{r+1}] \neq r\mathbb{E}\mathbb{E}[\eta^2|X]\mathbb{E}[\eta^{r-1}|X] \) for some \( r \in \mathbb{N} \), so that \( \eta|X \) is **not** a.s. Gaussian. Then the moments*

\[
m (X, Y, T, \theta, q(X), g(X)) := (Y - q(X) - \theta (T - g(X))) \\
\times \left( (T - g(X))^r - \mathbb{E}[\eta^r|X] - r (T - g(X)) \mathbb{E}[\eta^{r-1}|X] \right)
\]

*are 2-orthogonal and satisfy identifiability and smoothness assumptions.*
Comments:

(1) 2-orthogonal moment exist under non-Gaussianity of $\eta|X$!
Comments:

(1) 2-orthogonal moment exist under non-Gaussianity of $\eta|X$!

(2) Non-Gaussianity is standard in pricing (random discounts of a baseline price)
Comments:

(1) 2-orthogonal moment exist under non-Gaussianity of $\eta|X$!

(2) Non-Gaussianity is standard in pricing (random discounts of a baseline price)

(3) Proof: Reverse Engineer The Limitation Theorem.
2-orthogonal moment for PLR? Power (comments)

Comments:

1. 2-orthogonal moment exist under non-Gaussianity of $\eta|X$!
2. Non-Gaussianity is standard in pricing (random discounts of a baseline price)
3. Proof: Reverse Engineer The Limitation Theorem.
4. More general result in the paper without knowing the conditional moments.
Suppose \( f_0(X) = \langle X, \beta_0 \rangle, g_0(X) = \langle X, \gamma_0 \rangle \) for s-sparse \( \beta_0, \gamma_0 \in \mathbb{R}^p \).
Suppose $f_0(X) = \langle X, \beta_0 \rangle$, $g_0(X) = \langle X, \gamma_0 \rangle$ for s-sparse $\beta_0, \gamma_0 \in \mathbb{R}^p$.

How high sparsity can we tolerate with the suggested methods? (Stage 1 Error $\Leftrightarrow$ Bounds on sparsity)
Suppose $f_0(X) = <X, \beta_0>$, $g_0(X) = <X, \gamma_0>$ for s-sparse $\beta_0, \gamma_0 \in \mathbb{R}^p$.

How high sparsity can we tolerate with the suggested methods? (Stage 1 Error $\Leftrightarrow$ Bounds on sparsity)

LASSO can learn s-sparse linear $f_0, g_0$ with error $\sqrt{\frac{s \log p}{n}}$. How does this compare to the error we can tolerate?
Suppose \( f_0(X) = \langle X, \beta_0 \rangle, \ g_0(X) = \langle X, \gamma_0 \rangle \) for s-sparse \( \beta_0, \gamma_0 \in \mathbb{R}^p \).

How high sparsity can we tolerate with the suggested methods? (Stage 1 Error \( \Leftrightarrow \) Bounds on sparsity)

LASSO can learn s-sparse linear \( f_0, g_0 \) with error \( \sqrt{\frac{s \log p}{n}} \). How does this compare to the error we can tolerate?

Literature:

- Trivial Rate \( o\left( \frac{1}{\sqrt{n}} \right) \) - No \( s \) works.
- First-Order Orthogonal Rate \( o\left( n^{-\frac{1}{4}} \right); \ s = o\left( \frac{n^{\frac{1}{2}}}{\log p} \right) \) works.
Theorem

Suppose that

- \( \mathbb{E}[\eta^3] \neq 0 \)
- \( X \) has i.i.d. mean-zero standard Gaussian entries,
- \( \epsilon, \eta \) are almost surely bounded by the known value \( C \),
- and \( \theta_0 \in [-M, M] \) for known \( M \).

If

\[
s = o\left(\frac{n^{2/3}}{\log p}\right),
\]

and in the first stage of estimation we use LASSO with
\( \lambda_n = 2CM\sqrt{3\log(p)/n} \) then, using the 2-orthogonal moments \( m \) for \( r = 2 \)
the solutions of Stage 2 equations is \( \sqrt{n} \)-a.n. estimator of \( \theta_0 \).
Experiments 1: Fixed Sparsity

We consider $s = 100$, $n = 5000$, $p = 1000$, $\theta_0 = 3$.

**Figure:** Histogram for First Order Orthogonal.

**Figure:** Histogram for Second Order Orthogonal.

First Order Orthogonal: Bias Order of Magnitude Bigger than Variance!
Experiments 2: Varying Sparsity

We consider $n = 5000$, $p = 1000$, $\theta_0 = 3$.

Figure: 1st vs 2nd Order Orthogonal: BIAS, STD, MSE, Stage 1 $L_2$-error.
Experiments 3: MSE for Varying $n$, $p$, $s$

Figure:

- $n=2000, p=2000$
- $n=2000, p=5000$
- $n=5000, p=1000$
Summary

- We introduced the notion of k-orthogonality for GMM. Suffices to have $n^{-\frac{1}{2k+2}}$ first stage error for them to work. [Come to Poster for the general result!]
- We established that non-normality of $\eta|X$ is sufficient and necessary for the existence of useful 2-orthogonal moments for PLR.
- We used 2-orthogonal moment to tolerate $o\left(\frac{n^{\frac{2}{3}}}{\log p}\right)$ sparsity, much larger than state-of-art tolerance.
- We made synthetic experiments that support our claims.
Future Work

• How fundamental is the impossibility result when \( \eta | X \) is normally distributed? Can we establish a general lower bound?

• How fundamental is the sparsity \( o\left(\frac{n^2}{\log p}\right) \) barrier?

• Can we construct useful higher orthogonal moments for PLR?
Future Work

• How fundamental is the impossibility result when $\eta|X$ is normally distributed? Can we establish a general lower bound?

• How fundamental is the sparsity $o\left(\frac{n^{3/2}}{\log p}\right)$ barrier?

• Can we construct useful higher orthogonal moments for PLR?

Thank you!!