Orthogonal Machine Learning: Power and Limitations

Lester Mackey [1] **Vasilis Syrgkanis** [1] **Ilias Zadik** [1 2]

Abstract

Double machine learning provides $\sqrt{n}$-consistent estimates of parameters of interest even when high-dimensional or nonparametric nuisance parameters are estimated at an $n^{-1/4}$ rate. The key is to employ *Neyman-orthogonal* moment equations which are first-order insensitive to perturbations in the nuisance parameters. We show that the $n^{-1/4}$ requirement can be improved to $n^{-1/(2k+2)}$ by employing a k-th order notion of orthogonality that grants robustness to more complex or higher-dimensional nuisance parameters. In the partially linear regression setting, popular in causal inference, we show that we can construct second-order orthogonal moments if and only if the treatment residual is not normally distributed. Our proof relies on Stein's lemma and may be of independent interest. We conclude by demonstrating the robustness benefits of an explicit doubly-orthogonal estimation procedure for treatment effect.

1. Introduction

The increased availability of large and complex observational datasets is driving an increasing demand to conduct accurate causal inference of treatment effects in the presence of high-dimensional confounding factors. We take as our running example demand estimation from pricing and purchase data in the digital economy where many features of the world that simultaneously affect pricing decisions and demand are available in large data stores. One often appeals to modern statistical machine learning (ML) techniques to model and fit the high-dimensional or nonparametric nuisance parameters introduced by these confounders. However, most such techniques introduce bias into their estimates (e.g., via regularization) and hence yield invalid or inaccurate inferences concerning the parameters of interest (the treatment effects).

Several recent lines of have begun address the problem of debiasing ML estimators to perform accurate inference on a low dimensional component of model parameters. Prominent examples include Lasso debiasing (Zhang & Zhang; van de Geer et al., 2014; Javanmard & Montanari, 2015) and post-selection inference (Belloni et al.; Berk et al., 2013; Tibshirani et al., 2016). The recent double / debiased ML work of Chernozhukov et al. (2017) describes a general-purpose strategy for extracting valid inferences for target parameters from somewhat arbitrary and relatively inaccurate estimates of nuisance parameters.

Specifically, Chernozhukov et al. (2017) analyze a two-stage process where in the first stage one estimates nuisance parameters using arbitrary statistical ML techniques on a first stage data sample and in the second stage estimates the low dimensional parameters of interest via the generalized method of moments (GMM). Crucially, the moments in the second stage are required to satisfy a *Neyman orthogonality* condition, granting them first-order robustness to errors in the nuisance parameter estimation. A main conclusion is that the second stage estimates are $\sqrt{n}$-consistent and asymptotically normal whenever the first stage estimates are consistently estimated at a $o(n^{-1/4})$ rate.

To illustrate this result, let us consider the partially linear regression (PLR) model, popular in causal inference. In the PLR model we observe data triplets $Z = (T, Y, X)$, where $T \in \mathbb{R}$ represents a treatment or policy applied, $Y \in \mathbb{R}$ represents an outcome of interest, and $X \in \mathbb{R}^p$ is a vector of associated covariates. These observations are related via the equations

$$Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad a.s.$$
$$T = g_0(X) + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad a.s.$$

where η and ϵ represent unobserved disturbances with distributions independent of (θ_0, f_0, g_0). The first equation features the treatment effect θ_0, our object of inference. The second equation describes the relation between the treatment T and the associated covariates X. The covariates X affect the outcome Y through the nuisance function f_0 and the treatment T through the nuisance function g_0. Using the Neyman-orthogonal moment of (Chernozhukov et al.,

[1]Microsoft Research New England, USA [2]Operations Research Center, MIT, USA. Correspondence to: Lester Mackey <lmackey@microsoft.com>, Vasilis Syrgkanis <vasy@microsoft.com>, Ilias Zadik <izadik@mit.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

2017, Eq. 4.55), the authors show that it suffices to estimate the nuisance (f_0, g_0) at an $o(n^{-1/4})$ rate to construct a $\sqrt{n}$-consistent and asymptotically normal estimator of θ_0.

In this work, we provide a framework for achieving stronger robustness to first stage errors while maintaining second stage validity. In particular, we introduce a notion of higher-order orthogonality and show that if the moment is k-th order orthogonal then a first-stage estimation rate of $o(n^{-1/(2k+2)})$ suffices for $\sqrt{n}$-asymptotic normality of the second stage.

We then provide a concrete application of our approach to the case of estimating treatment effects in the PLR model. Interestingly, we show an impossibility result when the treatment residual follows a Gaussian distribution: no higher-order orthogonal moments with finite asymptotic variance exist, so first-order Neyman orthogonality appears to be the limit of robustness to first stage errors under Gaussian treatment residual. However, conversely, we also show how to construct appropriate second-order orthogonal moments whenever the treatment residual is not Gaussian. As a result, when the nuisance functions are linear in the high-dimensional confounders, our second-order orthogonal moments provide valid inferences whenever the number of relevant confounders is $o(\frac{n^{2/3}}{\log p})$; meanwhile the first-order orthogonality analyses of (Chernozhukov et al., 2017) accommodate only $o(\frac{\sqrt{n}}{\log p})$ relevant confounders.

We apply these techniques in the setting of demand estimation from pricing and purchase data, where highly non-Gaussian treatment residuals are standard. In this setting, the treatment is the price of a product, and commonly, conditional on all observable covariates, the treatment follows a discrete distribution representing random discounts offered to customers over a baseline price linear in the observables. In Figure 1 we portray the results of a synthetic demand estimation problem with dense dependence on observables. Here, the standard orthogonal moment estimation has large bias, comparable to variance, while our second-order orthogonal moments lead to nearly unbiased estimation.

Notational conventions For each $n \in \mathbb{N}$, we introduce the shorthand $[n]$ for $\{1, \dots, n\}$. We let $\xrightarrow{p}$ and $\xrightarrow{d}$ represent convergence in probability and convergence in distribution respectively. When random variables A and B are independent, we use $\mathbb{E}_A[g(A,B)] \triangleq \mathbb{E}[g(A,B) \mid B]$ to represent expectation only over the variable A. For a sequence of random vectors $(X_n)_{n=1}^{\infty}$ and a deterministic sequence of scalars $(a_n)_{n=1}^{\infty}$, we write $X_n = O_P(a_n)$ to mean X_n/a_n is stochastically bounded, i.e., for any $\epsilon > 0$ there is $R_\epsilon, N_\epsilon > 0$ with $\Pr(\|X_n/a_n\| > R_\epsilon) \leq \epsilon$ for all $n > N_\epsilon$. We let $N(\mu, \Sigma)$ represent a multivariate Gaussian distribution with mean μ and covariance Σ.

2. Z-Estimation with Nuisance Functions and Orthogonality

Our aim is to estimate an unknown target parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ given access to independent replicates $(Z_t)_{t=1}^{2n}$ of a random data vector $Z \in \mathbb{R}^\rho$ drawn from a distribution satisfying d moment conditions,

$$\mathbb{E}[m(Z, \theta_0, h_0(X))|X] = 0 \quad a.s. \tag{1}$$

Here, $X \in \mathbb{R}^p$ is a sub-vector of the observed data vector Z, $h_0 \in \mathcal{H} \subseteq \{h : \mathbb{R}^p \to \mathbb{R}^\ell\}$ is a vector of ℓ unknown nuisance functions, and $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is a vector of d known moment functions. We assume that these moment conditions exactly identify the parameter θ_0, and we allow for the data to be high-dimensional, with ρ and p potentially growing with the sample size n. However, the number of parameters of interest d and the number of nuisance functions ℓ are assumed to be constant.

We will analyze a two-stage estimation process where we first estimate the nuisance parameters using half of our sample[1] and then form a Z-estimate of the target parameter θ_0 using the remainder of the sample and our first-stage estimates of the nuisance. This *sample-splitting* procedure proceeds as follows.

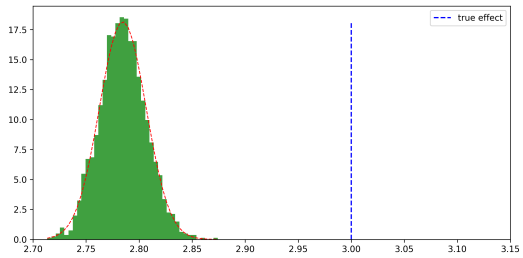
1. *First stage.* Form an estimate $\hat{h} \in \mathcal{H}$ of h_0 using $(Z_t)_{t=n+1}^{2n}$ (e.g., by running a nonparametric or high-dimensional regression procedure).
2. *Second stage.* Compute a Z-estimate $\hat{\theta}^{SS} \in \Theta$ of θ_0 using an empirical version of the moment conditions (1) and $\hat{h}$ as a plug-in estimate of h_0:

$$\hat{\theta}^{SS} \quad \text{solves} \quad \tfrac{1}{n}\textstyle\sum_{t=1}^{n} m(Z_t, \theta, \hat{h}(X_t)) = 0. \tag{2}$$

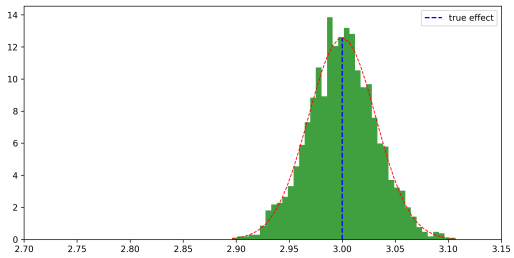
Relegating only half of the sample to each stage represents a statistically inefficient use of data and, in many applications, detrimentally impacts the quality of the first-stage estimate $\hat{h}$. A form of repeated sample splitting called *K-fold cross-fitting* (see, e.g., Chernozhukov et al., 2017) addresses both of these concerns. K-fold cross-fitting partitions the index set of the datapoints $[2n]$ into K subsets $I_1, \dots, I_K$ of cardinality $\frac{2n}{K}$ (assuming for simplicity that K divides $2n$) and produces the following two-stage estimate:

1. *First stage.* For each $k \in [K]$, form an estimate $\hat{h}_k \in \mathcal{H}$ of h_0 using only the datapoints $(Z_t)_{t \in I_k^c}$ corresponding to $I_k^c = [2n] \setminus I_k$.
2. *Second stage.* Compute a Z-estimate $\hat{\theta}^{SS} \in \Theta$ of θ_0 using an empirical version of the moment conditions

[1] Unequal divisions of the sample can also be used; we focus on an equal division for simplicity of presentation.



(a) Orthogonal estimates ($\hat{\theta} = 2.78$, $\hat{\sigma} = .022$)



(b) Second-order orthogonal estimates ($\hat{\theta} = 3.$, $\hat{\sigma} = .032$)

Figure 1. We portray the distribution of estimates based on orthogonal moments and second-order orthogonal moments. The true treatment effect $\theta_0 = 3$. Sample size $n = 5000$, dimension of confounders $d = 1000$, support size of sparse linear nuisance functions $s = 100$. The details of this experiment can be found in Section 5.

and $(\hat{h}_k)_{k\in[K]}$ as plug-in estimators of h_0:

$$\hat{\theta}^{CF} \text{ solves } \frac{1}{2n}\sum_{k=1}^{K}\sum_{t\in I_k} m(Z_t, \theta, \hat{h}_k(X_t)) = 0. \quad (3)$$

Throughout, we assume K is a constant independent of all problem dimensions. As we will see in Theorem 1, a chief advantage of cross-fitting over sample splitting is improved relative efficiency with an asymptotic variance that reflects the use of the full dataset in estimating θ.

Main Question. Our primary inferential goal is to establish conditions under which the estimators $\hat{\theta}^{SS}$ in (2) and $\hat{\theta}^{CF}$ (3) enjoy $\sqrt{n}$-asymptotic normality, that is

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

for some constant covariance matrix Σ. Coupled with a consistent estimator of Σ, asymptotic normality enables the construction of asymptotically valid confidence intervals for θ based on Gaussian or Student's t quantiles and asymptotically valid hypothesis tests, like the Wald test, based on chi-squared limits.

2.1. Higher-order Orthogonality

We would like our two-stage procedures to produce accurate estimates of θ_0 even when the first stage nuisance estimates are relatively inaccurate. With this goal in mind, Chernozhukov et al. (2017) defined the notion of Neyman-orthogonal moments, inspired by the early work of Neyman (1979). In our setting, the orthogonality condition of (Chernozhukov et al., 2017) is implied by the following condition, which we will call *first-order orthogonality*:

Definition 1 (First-order Orthogonal Moments)**.** *A vector of moments $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is* first-order orthogonal *with respect to the nuisance $h_0(X)$ if*

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

Here, $\nabla_\gamma m(Z, \theta_0, \gamma)$ is the gradient of the vector of moments with respect to its final ℓ arguments.

Intuitively, first-order orthogonal moments are insensitive to small perturbations in the nuisance parameters and hence robust to small errors in estimates of these parameters. A main result of (Chernozhukov et al., 2017) is that, if the moments m are first-order orthogonal, then $o(n^{-1/4})$ error rates[2] in the first stage estimation of h_0 are sufficient for $\sqrt{n}$-asymptotic normality of the estimates $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$.

Our aim is to accommodate slower rates of convergence in the first stage of estimation by designing moments robust to larger nuisance estimation errors. To achieve this, we will introduce a generalized notion of orthogonality that requires higher-order nuisance derivatives of m to be conditionally mean zero. We will make use of the following higher-order differential notation:

Definition 2 (Higher-order Differentials)**.** *Given a vector of moments $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ and a vector $\alpha \in \mathbb{N}^\ell$ we denote by $D^\alpha m(Z, \theta, \gamma)$ the α-differential of m with respect to its final ℓ arguments:*

$$D^\alpha m(Z, \theta, \gamma) = \nabla_{\gamma_1}^{\alpha_1} \nabla_{\gamma_2}^{\alpha_2} \dots \nabla_{\gamma_\ell}^{\alpha_\ell} m(Z, \theta, \gamma) \quad (4)$$

We are now equipped to define our notion of *S-orthogonal moments*:

Definition 3 (S-Orthogonal Moments)**.** *A vector of moments $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is* S-orthogonal *with respect to the nuisance $h_0(X)$ for some* orthogonality set *$S \subseteq \mathbb{N}^\ell$, if for any $\alpha \in S$:*

$$\mathbb{E}\left[D^\alpha m(Z, \theta_0, h_0(X)) \mid X\right] = 0. \quad (5)$$

We will often be interested in the special case of Definition 3 in which S is comprised of all vectors $\alpha \in \mathbb{N}^\ell$ with $\|\alpha\|_1 \leq$

[2] In the sense of root mean squared error: $n^{1/4}\sqrt{\mathbb{E}[\|h_0(X) - \hat{h}(X)\|_2^2 \mid \hat{h}]} \xrightarrow{p} 0$.

k. This implies that all mixed nuisance derivatives of the moment of order k or less are conditionally mean zero. We will refer to this special case as *k-orthogonality* or *k-th order orthogonality*.

Definition 4 (k-Orthogonal Moments)**.** *A vector of moments $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is* k-orthogonal *if it is S_k-orthogonal for the* k-orthogonality set*, $S_k \triangleq \{\alpha \in \mathbb{N}^\ell : \|\alpha\|_1 \leq k\}$.*

The general notion of S-orthogonality allows for our moments to be more robust to errors in some nuisance functions and less robust to errors in others. This is particularly valuable when some nuisance functions are easier to estimate than others; we will encounter such an example in Section 4.2.

3. Higher-order Orthogonality and Root-n Consistency

We will now show that S-orthogonality together with appropriate consistency rates for the first stage estimates of the nuisance functions imply $\sqrt{n}$-consistency and asymptotic normality of the two-stage estimates $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$. Beyond orthogonality and consistency, our main Assumption 1 demands identifiability, non-degeneracy, and regularity of the moments m, all of which are standard for establishing the asymptotic normality of Z-estimators.

Assumption 1. *For a non-empty orthogonality set $S \subseteq \mathbb{N}^\ell$ and $k \triangleq \max_{\alpha \in S} \|\alpha\|_1$, we assume the following:*

1. **S-Orthogonality.** *The moments m are S-orthogonal.*
2. **Identifiability.** $\mathbb{E}[m(Z, \theta, h_0(X))] \neq 0$ *when* $\theta \neq \theta_0$.
3. **Non-degeneracy.** *The matrix* $\mathbb{E}\left[\nabla_\theta m(Z, \theta_0, h_0(X))\right]$ *is invertible.*
4. **Smoothness.** $\nabla^k m$ *exists and is continuous.*
5. **Consistency of First Stage.** *The first stage estimates satisfy*

   $$\mathbb{E}[\textstyle\prod_{i=1}^\ell |\hat{h}_i(X) - h_{0,i}(X)|^{4\alpha_i} \mid \hat{h}] \xrightarrow{p} 0, \quad \forall \alpha \in S,$$

   where the convergence in probability is with respect to the auxiliary data set used to fit $\hat{h}$.
6. **Rate of First Stage.** *The first stage estimates satisfy*

   $$n^{1/2} \cdot \sqrt{\mathbb{E}[\textstyle\prod_{i=1}^\ell |\hat{h}_i(X) - h_{0,i}(X)|^{2\alpha_i} \mid \hat{h}]} \xrightarrow{p} 0,$$

   $\forall \alpha \in \{a \in \mathbb{N}^\ell : \|a\|_1 \leq k+1\} \setminus S$*, where the convergence in probability is with respect to the auxiliary data set used to fit $\hat{h}$.*
7. **Regularity of Moments.** *There exists an $r > 0$ such that the following regularity conditions hold:*
   - *(a)* $\mathbb{E}[\sup_{\theta \in \mathcal{B}_{\theta_0, r}} \|\nabla_\theta m(Z, \theta, h_0(X))\|_F] < \infty$ *for* $\mathcal{B}_{\theta_0, r} \triangleq \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq r\}$.
   - *(b)* $\sup_{h \in \mathcal{B}_{h_0, r}} \mathbb{E}[\sup_{\theta \in \mathcal{B}_{\theta_0, r}} \|\nabla_\gamma \nabla_\theta m(Z, \theta, h(X))\|^2] < \infty$ *for* $\mathcal{B}_{h_0, r} \triangleq \{h \in \mathcal{H} : \max_{\alpha : \|\alpha\|_1 \leq k+1} \mathbb{E}[\prod_{i=1}^\ell |h_i(X) - h_{0,i}(X)|^{2\alpha_i}] \leq r\}$.
   - *(c)* $\max_{\alpha : \|\alpha\|_1 \leq k+1} \sup_{h \in \mathcal{B}_{h_0, r}} \mathbb{E}\left[|D^\alpha m(Z, \theta_0, h(X))|^4\right] \leq \lambda_*(\theta_0, h_0) < \infty$.
   - *(d)* $\mathbb{E}[\sup_{\theta \in A, h \in \mathcal{B}_{h_0, r}} \|m(Z, \theta, h(X))\|_2] < \infty$, *for any compact $A \subseteq \Theta$,*
   - *(e)* $\sup_{\theta \in A, h \in \mathcal{B}_{h_0, r}} \mathbb{E}[\|\nabla_\gamma m(Z, \theta, h(X))\|^2] < \infty$, *for any compact $A \subseteq \Theta$.*

We are now ready to state our main theorem on the implications of S-orthogonality for second stage $\sqrt{n}$-asymptotic normality. The proof can be found in Section A.

Theorem 1 (Main Theorem)**.** *Under Assumption 1, if $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$ are consistent, then*

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

where $\Sigma = J^{-1} V J^{-1}$ for $J = \mathbb{E}\left[\nabla_\theta m(Z, \theta_0, h_0(X))\right]$ and $V = \mathtt{Cov}(m(Z, \theta_0, h_0(X)))$.

A variety of standard sufficient conditions guarantee the consistency of $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$. Our next result, proved in Section B, establishes consistency under either of two commonly satisfied assumptions.

Assumption 2. *One of the following sets of conditions is satisfied:*

1. **Compactness conditions:** Θ *is compact.*
2. **Convexity conditions:** Θ *is convex, θ_0 is in the interior of Θ, and, with probability approaching 1, the mapping $\theta \mapsto \frac{1}{n} \sum_{t=1}^n m(Z_t, \theta, \hat{h}(X_t))$ is the gradient of a convex function.*

Remark A continuously differentiable vector-valued function $\theta \mapsto F(\theta)$ on a convex domain Θ is the gradient of a convex function whenever the matrix $\nabla_\theta F(\theta)$ is symmetric and positive semidefinite for all θ.

Theorem 2 (Consistency)**.** *If Assumptions 1 and 2 hold, then $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$ are consistent.*

3.1. Sufficient Conditions for First Stage Rates

Our assumption on the first stage estimation rates, i.e., that $\forall \alpha \in \{a \in \mathbb{N}^\ell : \|a\|_1 \leq k+1\} \setminus S$

$$n^{1/2} \cdot \sqrt{\mathbb{E}\left[\textstyle\prod_{i=1}^\ell |\hat{h}_i(X) - h_{0,i}(X)|^{2\alpha_i} \mid \hat{h}\right]} \xrightarrow{p} 0$$

may seem complex, as it involves the interaction of the errors of multiple nuisance function estimates. In this section we give sufficient conditions that involve only the rates of individual nuisance function estimates and which imply our first stage rate assumptions. In particular, we are interested in formulating consistency rate conditions for each nuisance function h_i with respect to an $\mathcal{L}^p$ norm,

$$\|\hat{h}_i - h_{0,i}\|_p = \mathbb{E}[\|\hat{h}_i(X) - h_{0,i}(X)\|_p^p \mid \hat{h}]^{1/p}.$$

We will make use of these sufficient conditions when applying our main theorem to the partially linear regression model in Section 4.2.

Lemma 3. *Let $k = \max_{a \in S} \|a\|_1$. Then*

(1) Assumption 1.6 holds if any of the following holds $\forall \alpha \in \{a \in \mathbb{N}^\ell : \|a\|_1 \leq k+1\} \setminus S$:

- $\sqrt{n} \prod_{i=1}^{\ell} \|\hat{h}_i - h_{0,i}\|_{2\|\alpha\|_1}^{\alpha_i} \xrightarrow{p} 0$ (6)
- $\forall i, \quad n^{\frac{1}{\kappa_i \|\alpha\|_1}} \|\hat{h}_i - h_{0,i}\|_{2\|\alpha\|_1} \xrightarrow{p} 0$ (7)
  for some $\kappa_i \in (0,2]$ where $\frac{1}{\|\alpha\|_1} \sum_{i=1}^{\ell} \frac{\alpha_i}{\kappa_i} \geq \frac{1}{2}$
- $\forall i, \quad n^{\frac{1}{\kappa_i \|\alpha\|_1}} \|\hat{h}_i - h_{0,i}\|_{2\|\alpha\|_1} \xrightarrow{p} 0$ (8)
  for some $\kappa_i \in (0,2]$.

(2) Assumption 1.5 holds if $\forall i$, $\|\hat{h}_i - h_{0,i}\|_{4k} \xrightarrow{p} 0$.

A simpler description of the sufficient conditions arises under k-orthogonality (Definition 4), since the set $\{a \in \mathbb{N}^\ell : \|a\|_1 \leq k+1\} \setminus S_k$ contains only vectors α with $\|\alpha\| = k+1$.

Corollary 4. *If S is the canonical k-orthogonality set S_k (Definition 4), then Assumption 1.6 holds whenever*

$$\forall i, \quad n^{\frac{1}{2(k+1)}} \|\hat{h}_i - h_{0,i}\|_{2(k+1)} \xrightarrow{p} 0,$$

and Assumption 1.5 holds whenever $\forall i$, $\|\hat{h}_i - h_{0,i}\|_{4k} \xrightarrow{p} 0$.

In the case of first-order orthogonality, Corollary 4 requires that the first stage nuisance functions be estimated at a $o(n^{-1/4})$ rate with respect to the $\mathcal{L}^4$ norm. This is almost but not exactly the same as the condition presented in (Chernozhukov et al., 2017), which require $o(n^{-1/4})$ consistency rates with respect to the $\mathcal{L}^2$ norm. Ignoring the expectation over X, the two conditions are equivalent.[3] Moreover, in the case of k-orthogonality, Corollary 4 requires $o(n^{-1/2(k+1)})$ rates with respect to the $\mathcal{L}^{2(k+1)}$ norm. More generally, S-orthogonality allows for some functions to be estimated slower than others as we will see in the case of the sparse linear model.

[3] We would recover the exact condition in (Chernozhukov et al., 2017) if we replaced Assumption 1.7c with the more stringent assumption that $|D^\alpha m(Z, \theta, h(X))| \leq \lambda_*$ a.s.

4. Second-order Orthogonality for Partially Linear Regression

When second-order orthogonal moments satisfying Assumption 1 are employed, Corollary 4 implies that an $o(n^{-1/6})$ rate of nuisance parameter estimation is sufficient for $\sqrt{n}$-consistency of $\hat{\theta}^{SS}$ and $\hat{\theta}^{CF}$. This asymptotic improvement over first-order orthogonality holds the promise of accommodating more complex and higher-dimensional nuisance parameters. In this section, we detail both the limitations and the power of this approach in the partially linear regression (PLR) model setting popular in causal inference (see, e.g, Chernozhukov et al., 2017).

Definition 5 (Partially Linear Regression (PLR))**.** *In the partially linear regression model of observations $Z = (T, Y, X)$, $T \in \mathbb{R}$ represents a treatment or policy applied, $Y \in \mathbb{R}$ represents an outcome of interest, and $X \in \mathbb{R}^p$ is a vector of associated covariates. These observations are related via the equations*

$$Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0 \quad a.s.$$
$$T = g_0(X) + \eta, \quad \mathbb{E}[\eta \mid X] = 0 \quad a.s.$$

where η and ϵ represent unobserved noise variables with distributions independent of (θ_0, f_0, g_0).

4.1. Limitations: the Gaussian Treatment Barrier

Our first result shows that, under the PLR model, if the treatment noise, η, is conditionally Gaussian given X, then no second-order orthogonal moment can satisfy Assumption 1, because every twice continuously differentiable 2-orthogonal moment has $\mathbb{E}\left[\nabla_\theta m(Z, \theta_0, h_0(X))\right] = 0$ (a violation of Assumption 1.3). The proof in Section D relies on Stein's lemma.

Theorem 5. *Under the PLR model, suppose that η is conditionally Gaussian given X (a.s. X). If a twice differentiable moment function m is second-order orthogonal with respect to the nuisance parameters $(f_0(X), g_0(X))$, then it must satisfy $\mathbb{E}\left[\nabla_\theta m(Z, \theta_0, h_0(X))\right] = 0$ and hence violate Assumption 1.3. Therefore no second-order orthogonal moment satisfies Assumption 1.*

In the following result, proved in Section E, we establish that under mild conditions Assumption 1.3 is necessary for the $\sqrt{n}$-consistency of $\hat{\theta}^{SS}$ in the PLR model.

Proposition 6. *Under the PLR model, suppose that $|\Theta| \geq 2$ and that the conditional distribution of (ϵ, η) given X has full support on $\mathbb{R}^2$ (a.s. X). Then no moment function m simultaneously satisfies*

1. *Assumption 1, except for Assumption 1.3,*
2. *$\mathbb{E}\left[\nabla_\theta m(Z, \theta_0, h_0(X))\right] = 0$, and*
3. *$\hat{\theta}^{SS} - \theta_0 = O_P(1/\sqrt{n})$.*

4.2. Power: Second-order Orthogonality under Non-Gaussian Treatment

We next show that, inversely, second-order orthogonal moments are available whenever the conditional distribution of treatment noise given X is not a.s. Gaussian. Our proofs rely on a standard characterization of a Gaussian distribution, proved in Section F:

Lemma 7. *If $\mathbb{E}[\eta|X] = 0$ a.s., the conditional distribution of η given X is a.s. Gaussian if and only if for all $r \in \mathbb{N}, r \geq 2$ it holds that, $\mathbb{E}\left[\eta^{r+1}|X\right] = r\mathbb{E}[\eta^2|X]\mathbb{E}\left[\eta^{r-1}|X\right]$ a.s.*

We will focus on estimating the nuisance functions $q_0 = f_0 + \theta_0 g_0$ and g_0 instead of the nuisance functions f_0 and g_0, since the former task is more practical in many applications. This is because estimating q_0 can be accomplished by carrying out an arbitrary non-parametric regression of Y onto X. In contrast, estimating f_0 typically involves regressing Y onto (X, T), where T is constrained to enter linearly. The latter might be cumbersome when using arbitrary ML regression procedures.

Our first result, established in Section G, produces finite-variance 2-orthogonal moments when an appropriate moment of the treatment noise η is known.

Theorem 8. *Under the PLR model, suppose that we know $\mathbb{E}[\eta^r|X]$ 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 the conditional distribution of η given X is **not** a.s. Gaussian. Then the moments*

$$\begin{aligned} &m\left(Z, \theta, q(X), g(X), \mu_{r-1}(X)\right) \\ &\triangleq \left(Y - q(X) - \theta\left(T - g(X)\right)\right) \\ &\quad \times \left(\left(T - g(X)\right)^r - \mathbb{E}[\eta^r|X] - r\left(T - g(X)\right)\mu_{r-1}(X)\right) \end{aligned}$$

satisfy each of the following properties

- **2-orthogonality** *with respect to the nuisance $h_0(X) = (q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])$,*
- **Identifiability:** *When $\theta \neq \theta_0$,*

$$\mathbb{E}[m(Z, \theta, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])] \neq 0,$$

- **Non-degeneracy:**

$$\mathbb{E}[\nabla_\theta m\left(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X]\right)] \neq 0,$$

- **Smoothness:** *$\nabla^k m$ is continuous for all $k \in \mathbb{N}$.*

Our next result, proved in Section H, addresses the more realistic setting in which we do not have exact knowledge of $\mathbb{E}\left[\eta^r|X\right]$. We introduce an additional nuisance parameter and still satisfy an orthogonality condition with respect to these parameters.

Theorem 9. *Under the PLR model, suppose that $\mathbb{E}[\eta^{r+1}] \neq r\mathbb{E}[\mathbb{E}[\eta^2|X]\mathbb{E}[\eta^{r-1}|X]]$ for $r \in \mathbb{N}$, so that the conditional distribution of η given X is **not** a.s. Gaussian. Then, if*

$$S \triangleq \{\alpha \in \mathbb{N}^4 : \|\alpha\|_1 \leq 2\} \setminus \{(1,0,0,1), (0,1,0,1)\},$$

the moments

$$\begin{aligned} &m\left(Z, \theta, q(X), g(X), \mu_{r-1}(X), \mu_r(X)\right) \\ &\triangleq \left(Y - q(X) - \theta\left(T - g(X)\right)\right) \\ &\quad \times \left(\left(T - g(X)\right)^r - \mu_r(X) - r\left(T - g(X)\right)\mu_{r-1}(X)\right) \end{aligned}$$

satisfy each of the following properties

- *S-**orthogonality** with respect to the nuisance $h_0(X) = (q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X], \mathbb{E}[\eta^r|X])$,*
- **Identifiability:** *When $\theta \neq \theta_0$,*

$$\mathbb{E}[m(Z, \theta, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X], \mathbb{E}[\eta^r|X])] \neq 0,$$

- **Non-degeneracy:**

$$\mathbb{E}[\nabla_\theta m\left(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X], \mathbb{E}[\eta^r|X]\right)] \neq 0,$$

- **Smoothness:** *$\nabla^k m$ continuous for all $k \in \mathbb{N}$.*

In words, S-orthogonality here means that m satisfies the orthogonality condition for all mixed derivatives of total order at most 2 with respect to the four nuisance parameters except the mixed derivatives with respect to $(q_0(X), \mathbb{E}[\eta^r|X])$ and $(g_0(X), \mathbb{E}[\eta^r|X])$.

4.3. Application to High-dimensional Linear Nuisance Functions

We now consider deploying the PLR model in the high-dimensional linear regression setting, where $f_0(X) = \langle X, \beta_0\rangle$ and $g_0(X) = \langle X, \gamma_0\rangle$ for two s-sparse vectors $\beta_0, \gamma_0 \in \mathbb{R}^p$, p tends to infinity as $n \to \infty$, and (η, ϵ, X) are mutually independent. Define $q_0 = \theta_0\beta_0 + \gamma_0$. In this high-dimensional regression setting, Chernozhukov et al. (2017, Rem. 4.3) showed that two-stage estimation with first-order orthogonal moments

$$\begin{aligned} &m\left(Z, \theta, \langle X, q\rangle, \langle X, \gamma\rangle\right) = \\ &\left(Y - \langle X, q\rangle - \theta\left(T - \langle X, \gamma\rangle\right)\right)\left(T - \langle X, \gamma\rangle\right) \end{aligned} \tag{9}$$

and Lasso estimates of the nuisance provides a $\sqrt{n}$-asymptotically normal estimator of θ_0 when $s = o(n^{\frac{1}{2}}/\log p)$. Our next result, established in Appendix I, shows that we can accommodate $s = o(n^{\frac{2}{3}}/\log p)$ with an explicit set of higher-order orthogonal moments.

Theorem 10. *In the high-dimensional linear regression setting, suppose that either* $\mathbb{E}[\eta^3] \neq 0$ *(non-zero skewness) or* $\mathbb{E}[\eta^4] \neq 3\mathbb{E}[\eta^2]^2$ *(excess kurtosis), that* X *has i.i.d. mean-zero standard Gaussian entries, that* ϵ *and* η *are almost surely bounded by the known value* C*, and that* $\theta_0 \in [-M, M]$ *for known* M*. If* $s = o(n^{2/3}/\log p)$*, and in the first stage of estimation we*

(a) create estimates $\hat{q}, \hat{\gamma}$ *of* q_0, γ_0 *via Lasso regression of* Y *on* X *and* T *on* X *respectively, with regularization parameter* $\lambda_n = 2CM\sqrt{3\log(p)/n}$ *and*

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

$$\hat{\mu}_2 = \tfrac{1}{n}\textstyle\sum_{t=1}^n \hat{\eta}_t^2, \text{ and } \hat{\mu}_3 = \tfrac{1}{n}\textstyle\sum_{t=1}^n (\hat{\eta}_t^3 - 3\hat{\mu}_2\hat{\eta}_t),$$

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

then, using the moments m *of Theorem 9 with* $r = 2$ *in the case of non-zero skewness or* $r = 3$ *in the case of excess kurtosis,* $\hat{\theta}^{SS}$ *and* $\hat{\theta}^{CF}$ *are* $\sqrt{n}$*-asymptotically normal estimators of* θ_0*.*

5. Experiments

We perform an experimental analysis of the second order orthogonal estimator of Theorem 10 with $r = 3$ for the case of estimating treatment effects in the PLR model with high-dimensional sparse linear nuisance functions. We compare our estimator with the double ML estimator (labeled 'dml' in our figures) based on the first-order orthogonal moments (9) of (Chernozhukov et al., 2017). Our experiments are designed to simulate demand estimation from pricing and purchase data, where non-Gaussian treatment residuals are standard. Here, our covariates X correspond to all collected variables that may affect a pricing policy. A typical randomized experiment in a pricing policy takes the form of random discounts from a baseline price as a company offers random discounts to customers periodically to gauge demand level. In this case, the treatment residual – the unexplained fluctuation in price – is decidedly non-Gaussian distribution and rather follows a discrete distribution over a small number of price points. Python code recreating all experiments is available at `https://github.com/IliasZadik/double_orthogonal_ml`.

Experiment Specification We generated n independent replicates of outcome Y, treatment T, and confounding covariates X. The confounders X have dimension p and have independent components from the $N(0, 1)$ distribution. The treatment is a sparse linear function of X, $T = \langle \gamma_0, X\rangle + \eta$, where only s of the p coefficients of γ_0 are non-zero. The x-axis on each plot is the number of non-zero coefficients s. Moreover, η is drawn from a discrete distribution with values $\{0.5, 0, -1.5, -3.5\}$ taken respectively with probabilities $(.65, .2, .1, .05)$. Here, the treatment represents the price of a product or service, and this data generating process simulates random discounts over a baseline price. Finally, the outcome is generated by a linear model, $Y = \theta_0 T + \langle \beta_0, X\rangle + \epsilon$, where $\theta_0 = 3$ is the treatment effect, β_0 is another sparse vector with only s non-zero entries, and ϵ is drawn independently from a uniform $U(-\sigma_\epsilon, \sigma_\epsilon)$ distribution. Importantly, the coordinates of the s non-zero entries of the coefficient β_0 are the same as the coordinates of the s non-zero entries of γ_0. The latter ensures that variables X create a true endogeneity problem, i.e., that X affects both the treatment and the outcome directly. In such settings, controlling for X is important for unbiased estimation.

To generate an instance of the problem, the common support of both γ_0 and β_0 was generated uniformly at random from the set of all coordinates, and each non-zero coefficient was generated independently from a uniform $U(0, 5)$ distribution. The first stage nuisance functions were fitted for both methods by running the Lasso on a subsample of $n/2$ sample points. For the first-order method all remaining $n/2$ points were used for the second stage estimation of θ_0. For the second-order method, the moments $\mathbb{E}[\eta^2]$ and $\mathbb{E}[\eta^3]$ were estimated using a subsample of $n/4$ points as described in Theorem 10, and the remaining $n/4$ sample points were used for the second stage estimation of θ_0. For each method we performed cross-fitting across the first and second stages, and for the second-order method we performed nested cross-fitting between the $n/4$ subsample used for the $\mathbb{E}[\eta^2]$ and $\mathbb{E}[\eta^3]$ estimation and the $n/4$ subsample used for the second stage estimation. The regularization parameter λ_n of each Lasso was chosen to be $\sqrt{\log(p)/n}$.

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 γ_0 and β_0, to evaluate variability across different realizations of the nuisance functions.

Distribution of Errors with Fixed Sparsity In Figure 1, we display the distribution of estimates based on orthogonal moments and second-order orthogonal moments for a particular sparsity level $s = 100$ and for $n = 5000$ and $p = 1000$. We observe that both estimates are approximately normally distributed, but the orthogonal moment estimation exhibits significant bias, an order of magnitude larger than the variance.

Bias-Variance Tradeoff with Varying Sparsity Figure 2 portrays the median quantities (solid lines) and maximum and minimum of these quantities (error bars) across the

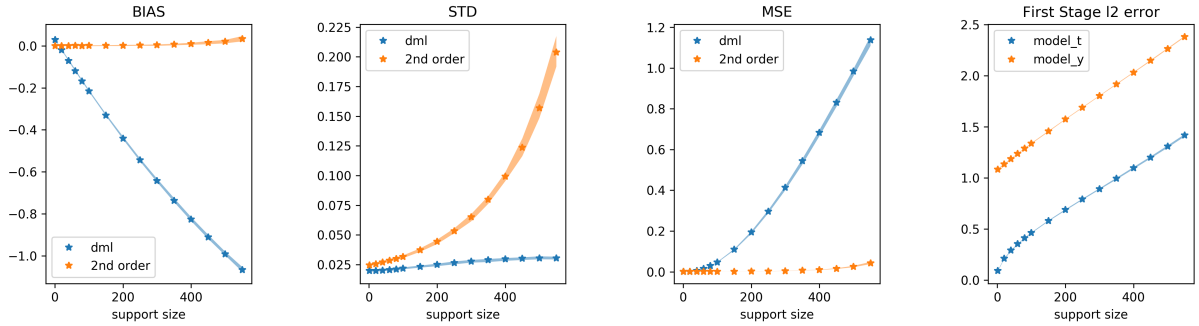
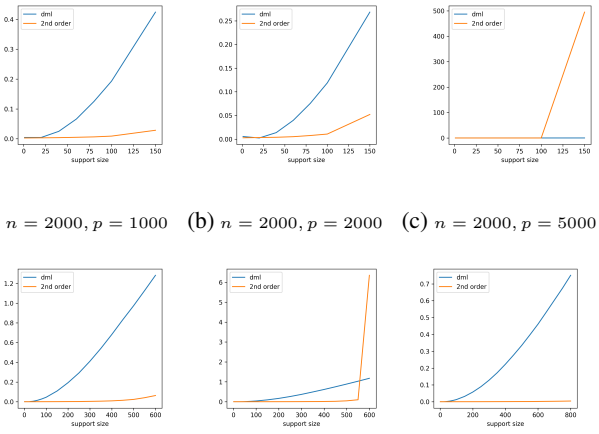


Figure 2. Comparison of estimates $\hat{\theta}^{CF}$ based on orthogonal moments and second order orthogonal moments under the PLR model as a function of the number of non-zero coefficients in the nuisance vectors γ_0 and β_0. See Section 5 for more details. The parameters used for this figure were $n = 5000$, $p = 1000$, $\sigma_\epsilon = 1$. The fourth figure displays the ℓ_2 error in the coefficients discovered by the first stage estimates for each of the nuisance functions: model_t is the model for $\mathbb{E}[T|X]$ and model_y is the model for $\mathbb{E}[Y|X]$.

100 different nuisance function draws as a function of the support size for $n = 5000$, $p = 1000$, and $\sigma_\epsilon = 1$.

Varying n and p In Figure 3, we display how performance varies with n and p. Due to computational considerations, for this parameter exploration, we only used a single problem instance for each (n, p, s) triplet rather than 100 instances as in the exploration above. We note that for $n = 2000, p = 5000$ the breaking point of our method is around $s = 100$, while for $n = 5000, p = 2000$ it is around $s = 550$. For $n = 10000, p = 1000$ our method performs exceptionally well even until $s = 800$.

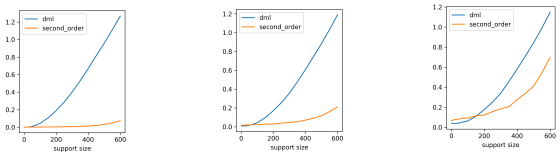


(a) $n = 2000, p = 1000$ (b) $n = 2000, p = 2000$ (c) $n = 2000, p = 5000$

(d) $n = 5000, p = 1000$ (e) $n = 5000, p = 2000$ (f) $n = 10000, p = 1000$

Figure 3. MSE of both estimators as the sparsity varies for different sample size and dimension pairs (n, p). Note that the range of the support sizes is larger for larger n. $\sigma_\epsilon = 1$.

Varying σ_ϵ Finally Figure 4 displays performance as the variance σ_ϵ of the noise ϵ grows.



(a) $\sigma_\epsilon = 3$ (b) $\sigma_\epsilon = 10$ (c) $\sigma_\epsilon = 20$

Figure 4. MSE of both estimators as the sparsity varies for different variance parameters σ_ϵ. $n = 5000, p = 1000$.

6. Conclusion

We considered the problem of treatment effect estimation and inference in the presence of high-dimensional or non-parametric nuisance. We first introduced the notion of k-th order orthogonal moments for two-stage Z-estimation, generalizing the first-order Neyman orthogonality studied in (Chernozhukov et al., 2017). Assuming the existence of such a moment, we established that learning the nuisance parameters at a $o(n^{-1/(2k+2)})$ rate suffices for the $\sqrt{n}$-consistent and asymptotic normal estimates of the parameters of interest. We then focused on the PLR model popular in causal inference and established that a second-order orthogonal moment exists if and only if the treatment residual is not normally distributed. When the nuisance functions of the PLR model are linear but high-dimensional, this allowed us to tolerate significantly denser nuisance vectors than those accommodated by (Chernozhukov et al., 2017). We complemented our results with various synthetic demand estimation experiments showing the benefits of second-order orthogonal moments over first-order orthogonal moments.

References

Belloni, A., Chernozhukov, V., Val, I. F., and Hansen, C. Program evaluation and causal inference with high dimensional data. *Econometrica*, 85(1):233–298.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 04 2013. doi: 10.1214/12-AOS1077.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, May 2017.

Durrett, R. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, USA, 4th edition, 2010. ISBN 0521765390, 9780521765398.

Flanders, H. Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6):615–627, 1973.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Javanmard, A. and Montanari, A. De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. *ArXiv e-prints*, August 2015.

Newey, W. and McFadden, D.l. Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111 – 2245, 1994. ISSN 1573-4412. doi: http://dx.doi.org/10.1016/S1573-4412(05)80005-4.

Neyman, J. C() tests and their use. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 41(1/2):1–21, 1979. ISSN 0581572X.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 11 1981.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

van de Geer, S., Buhlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014. doi: 10.1214/14-AOS1221.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

Zhang, C. H. and Zhang, S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242. doi: 10.1111/rssb.12026.