Introduction to Debiased Machine Learning, Lecture 4: DML for nonparametric objects; an empirical example.

Whitney K. Newey

2026 Nemmers Lectures
Northwestern University

with Victor Chernozhukov and Vasilis Syrgkanis;

https://arxiv.org/pdf/2412.18080, forthcoming in *Quantitative Economics* as "Conditional Influence Functions".

DEBIASED MACHINE LEARNING FOR NONPARAMETRIC OBJECTS

There are many nonparametric objects of interest that are functions of a conditional distribution.

A leading example is the conditional average treatment effect (CATE), conditional on a subset of the identifying covariates; for example, the CATE conditional on income.

Such conditional objects are central in policy design where the goal is to adapt treatment or policy to a subset of the observable covariates.

In economics, a related example is the average equivalent variation (exact consumer surplus) of a tax change conditional on income, which captures heterogeneity in welfare effects across the income distribution, informing economic policy.

This paper introduces the conditional influence functions (CIF) for conditional objects of interest and describes how they can be used.

A CIF, when it exists, is a derivative of a conditional parameter with respect to perturbations of the conditional distribution $F(\cdot|V)$ of a single observation given a conditioning variable V;

The CIF generalizes the classical influence function of Hampel (1974) and Huber (1981).

The CIF can be used to

–Construct conditionally Neyman-orthogonal moment functions, for estimating conditional parameters of interest that depend on high dimensional nuisance functions, where Neyman-orthogonal means that there is no first-order effect of plugging in nonparametric estimates to the conditional moments.

–Characterize effects of local misspecification on conditional parameters of interest.

–Define local policy effects.

Here we focus on constructing Neyman orthogonal estimating equations; see the paper for uses of the CIF in local misspecification analysis and conditional policy effects.

Here we only consider the nonparametric CIF, where conditional CDF is essentially unrestricted, and avoid any attempt to use the CIF for asymptotic efficiency comparisons.

Using the nonparametric CIF makes Neyman orthogonality robust to restrictions on the nonparametric estimators that are plugged in, because these restrictions are incorporated into the nonparametric CIF, similarly to Newey (1994).

There is previous work on Neyman orthogonal estimation of conditional objects of interest,

Series estimators of conditional parameters are given in Semenova and Chernozhukov (2021).

Fan, Hsu, Lieli, and Zhang (2022) give locally linear estimator of CATE with explicit bias correction,

Kernel regression estimators with automatic debiasing are given in Chernozhukov, Newey, and Singh (2022a).

Foster and Syrgkanis (2023) give conditions for oracle prediction rates when losses are estimated and Neyman orthogonal.

Optimal rates for estimation of the average treatment effect conditional on all covariates were derived in Kennedy (2023) and Kennedy et al. (2024).

Luedtke and Chung (2024) derive attainable rates of convergence and inference for conditional objects of interest in reproducing kernel spaces.

Yang, Kuchibhotla, and Tchetgen Tchetgen (2004) and Dalal, Blobaum, Kasiviswanathan, and Ramdas (2024) derive some Neyman orthogonal moment functions from conditional expectation properties.

Related work on continuous treatments and nonseparable models includes Kallus and Zhou (2018), Su, Ura, and Zhang (2019), and Colangelo and Lee (2025).

Here we study pathwise differentiability with respect to the conditional distribution itself, derive CIFs for a range of conditional objects beyond previous work, use CIFs to construct conditionally Neyman orthogonal estimating equations, and consider conditional local misspecification and policy effects.

Our contribution differs from previous work in defining the CIF in terms of perturbations of the conditional law $F(\cdot \mid V)$ and in using the CIF to construct debiased conditional moment equations for nonparametric estimation of conditional objects of interest.

We show how the CIF can be used to estimate more conditional objects than previously considered.

CONDITIONAL INFLUENCE FUNCTIONS

Recall the classical influence function.

W : One data observation in i.i.d. sequence.

F : CDF of one observation having true value F_0.

$\theta(F)$: parameter of interest having true value $\theta_0 = \theta(F_0)$.

$F_\varepsilon = (1-\varepsilon)F_0 + \varepsilon H$ for CDF H represents local variation.

The influence function (when it exists) is $\psi_0(W)$ such that for any H,

$$\frac{\partial \theta(F_\varepsilon)}{\partial \varepsilon} = E_H[\psi_0(W)] = \int \psi_0(w) H(dw), \; E[\psi_0(W)] = 0.$$

$\psi_0(W)$ is a Gateaux derivative of $\theta(F)$ at F_0 in the direction $H - F_0$.

$\psi_0(W)$ is fixed as H varies and so is a derivative of $\theta(F_\varepsilon)$ with respect of H.

To describe a conditional influence function, let $V = T(W)$ be some random vector, and define

F^v : CDF of one observation conditional on $V = v$,
with true conditional CDF F_0^v;
$\theta(F^v)$: parameter of interest, having true value $\theta_0(v) = \theta(F_0^v)$;
$F_\varepsilon^v = (1-\varepsilon)F_0^v + \varepsilon H^v$ for alternate conditional CDF H^v.

The conditional influence function (when it exists) is $\psi_0^v(W)$ such that for any H^v,

$$\frac{\partial \theta(F_\varepsilon^v)}{\partial \varepsilon} = \int \psi_0^v(W) dH^v, \; E[\psi_0^v(W)|V] = 0.$$

Running Example: Conditional average potential outcome, given some function of covariates.

There is an outcome variable $Y = DY(1) + (1 - D)Y(0)$ where $D \in \{0, 1\}$ is treatment, $Y(1)$, $Y(0)$ are potential outcomes with and without treatment, are mean independent of treatment conditional on covariates Z, and $\pi_0(Z) = \Pr(D = 1|Z) > 0$ with probability one.

Let $V = t(Z)$ some function of the covariates (e.g. a single covariate), $X = (D, Z)$, $W = (Y, X)$, and $\gamma_0(X) = E[Y|X]$.

An object of interest is

$$\theta_0(v) = E[Y(1)|V = v] = E[\gamma_0(1, Z)|V = v].$$

The CIF of $\theta_0(V)$ is shown in the paper to be

$$\gamma_0(1, Z) - \theta_0(V) + \alpha_0(X)\{Y - \gamma_0(X)\},\ \alpha_0(X) = \frac{D}{\pi_0(Z)}.$$

LINEAR FUNCTIONS OF A CONDITIONAL MEAN

Generalize running example to an object of interest

$$\theta_0(V) = E[m(W, \gamma_0)|V],\ \gamma_0(X) = E[Y|X],\ V = T(X).$$

In running example, $m(W, \gamma) = \gamma(1, Z)$, for $X = (D, Z)$.

The key condition for existence of a conditional influence function for $\theta_0(V)$ is there is $\alpha_0(X)$ such that

$E[m(W, \gamma)|V] = E[\alpha_0(X)\gamma(X)|V]$ for all $\gamma(X)$ with $E[\gamma(X)^2|V] < \infty$ w.p.1.

Above equation is a conditional Riesz representation and $\alpha_0(X)$ a conditional Riesz representer.

In the running example, where $m(W, \gamma) = \gamma(1, Z)$, we have $\alpha_0(X) = D/\pi_0(Z)$.

Another example is conditional average derivative, which is conditional ATE for continuous treatment; see Imbens and Newey (2009).

$$\theta_0(V) = E[\frac{\partial}{\partial d}\gamma_0(D, Z)|V],\ V = T(Z).$$

Here, if $V = T(Z)$, for $m(W, \gamma) = \partial\gamma(D, Z)/\partial d$, iterated expectations and integration by parts conditional on Z give

$$E[m(W, \gamma)|V] = E[E[\partial\gamma(D, Z)/\partial d|Z]|V] = E[\alpha_0(D, Z)\gamma(D, Z)|V],$$
$$\alpha_0(D, Z) = -\partial \ln f_0(D|Z)/\partial d,$$

where $f_0(D|Z)$ is conditional pdf of D conditional on Z.

Thus, the key condition is satisfied for the average derivative conditional on a function V of the covariates Z.

Key condition requires that V cannot be a function of variables in X that $m(W, \gamma)$ operates on.

In general the CIF of $\theta_0(V) = E[m(W, \gamma_0)|V]$ of $\gamma_0(X) = E[Y|X]$ is

$$m(W, \gamma_0) - \theta_0(V) + \alpha_0(X)[Y - \gamma_0(X)].$$

The derivation of this as the CIF is in the paper; we can show directly that this is conditionally Neyman orthogonal.

$$\begin{aligned}
&E[m(W, \gamma) + \{Y - \gamma(X)\}|V] - \theta_0(V) \\
&= E[m(W, \gamma)|V] - E[m(W, \gamma_0)|V] + E[\alpha(X)\{Y - \gamma(X)\}|V] \\
&= E[\alpha_0(X)\{\gamma(X) - \gamma_0(X)\}|V] + E[\alpha(X)\{Y - \gamma(X)\}|V] \\
&= E[\alpha_0(X)\{\gamma(X) - \gamma_0(X)\}|V] + E[\alpha(X)\{\gamma_0(X) - \gamma(X)\}|V] \\
&= -E[\{\alpha(X) - \alpha_0(X)\}\{\gamma(X) - \gamma_0(X)\}|V].
\end{aligned}$$

Thus we see that the conditional expectation of CIF evaluated at any γ and α differs from zero only by a second-order term.

Even have conditional double robustness, that conditional expectation of pseudo outcome is $\theta_0(V)$ if either $\gamma = \gamma_0$ or $\alpha = \alpha$.

ESTIMATING LINEAR FUNCTIONS OF CONDITIONAL MEAN

We can use the CIF to construct an estimator for $\theta_0(V)$.

Recall that the conditional mean of CIF is zero, so taking conditional expectations and solving

$$\theta_0(V) = E[m(W, \gamma_0) + \alpha_0(X)\{Y - \gamma_0(X)\}|V].$$

We can form an estimator of $\theta_0(V)$ by plugging in nonparametric estimators $\hat{\gamma}$ and $\hat{\alpha}$ to form a pseudo outcome

$$\hat{S}(W) = m(W, \hat{\gamma}) + \hat{\alpha}(X)\{Y - \hat{\gamma}(X)\},$$

and then estimate $\theta_0(V)$ from a nonparametric regression of $\hat{S}(W)$ on V.

Second term $\hat{\alpha}(X)[Y - \hat{\gamma}(X)]$ acts as a bias correction.

Pseudo-outcome does not depend on choice of conditioning variable V, so can use same pseudo outcomes for all conditioning variables.

To estimate $\theta_0(V)$ from observations $W_1, ..., W_n$ we can use the debiased pseudo outcomes

$$\hat{S}_i = m(W_i, \hat{\gamma}) + \hat{\alpha}(X_i)[Y_i - \hat{\gamma}(X_i)].$$

We can run a nonparametric regression of the pseudo outcome $\hat{S}_i$ on V_i at v to estimate $\theta_0(v) = E[m(W, \gamma_0)|V = v]$.

To avoid overfitting we use cross fitting where partition observation indices $i = 1, ..., n$ into distinct sets $I_\ell, (\ell = 1, ..., L)$, and let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ be computed from observations not in I_ℓ.

Let $\hat{S}_i = m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]$, $i \in I_\ell$.

We run a nonparametric regression of $\hat{S}_i$ on V_i at v to estimate $\theta_0(v)$.

Estimation is based on pseudo outcomes $\hat{S}_i = m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]$.

Get $\hat{\gamma}_\ell(X_i)$ from nonparametric regression of Y_i on X_i and also need $\hat{\alpha}_\ell(X_i)$.

Conditional Riesz representation implies unconditional one with $E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)]$ for all γ with $E[\gamma(X)^2] < \infty$.

Can use Riesz regression as in Chernozhukov, Newey, Quintas-Martinez, and Syrgkanis (2021).

The needed $\hat{\alpha}_\ell(X_i)$ can be obtained as minimum of Riesz loss

$$\hat{\alpha}_\ell = \arg\min_{\alpha \in \mathcal{A}} \sum_{i \notin I_\ell} [-2m(W_i, \alpha) + \alpha(X_i)^2],$$

for a set $\mathcal{A}$ of approximating functions, with penalty possibly added.

Can minimize over neural net and other specifications for $\mathcal{A}$.

These are automatic methods for constructing $\hat{\alpha}$, requiring only $m(W, \alpha)$ and not using a specific formula for α_0.

To describe a locally linear regression estimator of $\theta_0(V)$ based on the debiased outcomes $\hat{S}_i$, let $K_h(u) = h^{-r}K(u/h)$.

A locally linear estimator is

$$\hat{\theta}(v) = \arg\min_{\theta,\beta} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} (\hat{S}_i - \theta - (v - V_i)'\beta)^2 K_h(v - V_i). \tag{1}$$

This $\hat{\theta}(v)$ can be computed as the estimator of the coefficient of 1 in a weighted least squares regression of $\hat{S}_i$ on $(1, v - V_i)$, with weight $K_h(v - V_i)$ for the i^{th} observation.

That is, for

$$\mathcal{X}_v = \left[\begin{pmatrix} 1 \\ v - V_1 \end{pmatrix}, ..., \begin{pmatrix} 1 \\ v - V_n \end{pmatrix} \right]', \mathcal{W}_v = diag(K_h(v - V_i)),\ \hat{S} = (\hat{S}_1, ..., \hat{S}_n)',$$

the estimator is

$$(\hat{\theta}(v), \hat{\beta}(v))' = (\mathcal{X}_v'\mathcal{W}_v\mathcal{X}_v)^{-1}\mathcal{X}_v'\mathcal{W}_v\hat{S}.$$

The asymptotic variance of $\hat{\theta}(v)$ can be estimated by ignoring the presence of $\hat{\alpha}$ and $\hat{\gamma}$ and forming the usual locally linear regression variance estimator

$$\widehat{Var}(\hat{\theta}(v)) = e_1'(\mathcal{X}_v'\mathcal{W}_v\mathcal{X}_v)^{-1}\mathcal{X}_v'\mathcal{W}_v\hat{\Omega}_v\mathcal{W}_v\mathcal{X}_v(\mathcal{X}_v'\mathcal{W}_v\mathcal{X}_v)^{-1}e_1,$$
$$\hat{\Omega}_v = diag(\hat{u}_i(v)^2),\ \hat{u}_i(v) = \hat{S}_i - \hat{\theta}(v) - (v - V_i)'\hat{\beta}(v).$$

Straightforward to give conditions for this estimator to have same limiting distribution as estimator where $\hat{S}_i$ is replaced by $S_i = m(W_i, \gamma_0) + \alpha_0(X_i)[Y_i - \gamma_0(X_i)]$.

Assumption 1: $\int K(u)du = 1$ *and* $K(u)$ *and* $K(u)u^2$ *are bounded.*

We also assume that $m(W, \gamma)$ is a mean square continuous function of γ and that the Riesz representer and $Var(Y|X)$ are bounded.

Assumption 2: $E[m(W, \gamma)^2] \leq CE[\gamma(X)^2]$ *for some* $C > 0$ *and* $\alpha_0(X)$ *and* $Var(Y|X)$ *are bounded*.

For any function $h(X)$ let $\|h\| = \sqrt{E[h(X)^2]}$ denote the mean-square norm.

Our main regularity conditions are mean square rates of convergence for $\hat{\gamma}$ and $\hat{\alpha}$ relative to the bandwidth h of the kernel estimator.

Assumption 3: i)$\sqrt{nh^r} \longrightarrow \infty$, $\|\hat{\gamma} - \gamma_0\| = o_p(h^{r/2})$, $\|\hat{\alpha} - \alpha_0\| = o_p(h^{r/2})$; ii) $\sqrt{n}\,\|\hat{\gamma} - \gamma_0\|\,\|\hat{\alpha} - \alpha_0\| = o_p(h^{r/2})$.

Theorem: *If there is a conditional Riesz representation with bounded* $\alpha_0(X)$ *and Assumptions 1 - 3 are satisfied then*

$$\sqrt{nh^r}\{\hat{\theta}(v) - \tilde{\theta}(v)\} \xrightarrow{p} 0.$$

Assumption 3: i)$\sqrt{nh^r} \longrightarrow \infty$, $\|\hat{\gamma} - \gamma_0\| = o_p(h^{r/2})$, $\|\hat{\alpha} - \alpha_0\| = o_p(h^{r/2})$; ii) $\sqrt{n}\,\|\hat{\gamma} - \gamma_0\|\,\|\hat{\alpha} - \alpha_0\| = o_p(h^{r/2})$.

The rate conditions of Assumption 3 are stronger than those for estimating an unconditional parameter at the root-n rate.

The mean square convergence rates of $\hat{\gamma}$ and $\hat{\alpha}$ must be faster than $h^{r/2}$ in Assumption 3 i) and ii).

These rates are needed here because we only rely on mean square rates for $\hat{\gamma}$ and $\hat{\alpha}$ in order to allow $\hat{\gamma}$ and $\hat{\alpha}$ to be machine learners that are only known to have mean-square rates of convergence, as is the case for many machine learners.

Assumption 3 could be weakened; if L_4 convergence rates for $\hat{\alpha}$ and $\hat{\gamma}$ were available and use L_4 norm then could replace $h^{r/2}$ with $h^{r/4}$; similar to Foster and Syrgkanis (2023).

If uniform rates were available, can drop h terms.

An important question is whether the oracle rate for estimating $\theta_0(v)$ when γ_0 and α_0 are known can be attained.

The answer is yes for CATE when γ_0 and α_0 are smooth enough and dimension of V is smaller than dimension of Z for CATE.

To attain the oracle rate it will suffice that number of derivatives that exist and are bounded, of γ_0 and α_0, are greater than dimension.

OTHER THINGS IN THE PAPER

A general construction of conditional estimating equations for objects of interest that are conditionally Neyman orthogonal.

Generalize from γ_0 being a conditional mean to

$$\gamma_0 = \arg\min_{\gamma(X)} E[Q(W, \gamma)|V],$$

for some objective function $Q(W, v)$, such as $\gamma_0(X)$ a conditional quantile.

Give a conditional influence function for weighted average potential outcome for continuous treatment.

SUMMARY

We have given

–Conditional influence function.

–Debiased machine learners of conditional objects of interest that are linear functionals of conditional means.

–Asymptotic theory for locally linear nonparametric estimators.

See paper for extensions mentioned here.