Introduction to Debiased Machine Learning, Lecture 3: Neyman orthogonality via "profiling" and influence functions; large sample theory.

Whitney K. Newey

2026 Nemmers Lectures
Northwestern University

CONSTRUCTING NEYMAN ORTHOGONAL MOMENT FUNCTIONS

There are two general approaches to constructing Neyman orthogonal moment functions.

One is to "profile out" unknown functions from an objective function that has expectation maximized at parameter value and function.

Differentiating the "profiled" objective function with respect to the parameter of interest gives a Neyman orthogonal moment function, as shown in Newey (1994, Econometrica) for pathwise differentiation and Chernozhukov et al. (2018, Econometrics Journal) for a functional derivative, generalizing Neyman (1959) beyond the likelihood setting.

The other approach is to add to identifying moment functions a bias correction term that is an influence function.

This is Neyman orthogonal in Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2022, Econometrica, "Locally Robust Semiparametric Estimation," LR henceforth), generalizing the approach of including a first order term from a von Mises (1947) expansion, as in Hasminski and Ibragimov (1978).

NEYMAN ORTHOGONALITY VIA PROFILING

Consider a function $q(W, \gamma, \theta)$ depending on a data observation W, an unknown function γ, and a parameter θ.

Suppose that

$$\theta_0 = \arg\max_{\theta,\gamma} E[q(W, \gamma, \theta)].$$

We can "profile out" γ by fixing the value of θ and maximizing over γ to obtain

$$\gamma(\theta) = \arg\max_{\gamma} E[q(W, \gamma, \theta)].$$

It turns out that a Neyman orthogonal moment function is given by

$$\psi(W, \gamma, \alpha, \theta) = \frac{\partial q(W, \gamma(\theta), \theta)}{\partial \theta}.$$

This generally depends on an at least one additional function α.

An example is partially linear, least squares objective function, where $q(W, \gamma, \theta) = -[Y - \theta D - \gamma(Z)]^2$.

Robinson (1988) showed Neyman orthogonality when $\gamma(Z)$ can be any function with finite second moment.

The profile objective function is

$$\begin{aligned} q(W, \gamma(\theta), \theta) &= -[Y - \theta D - \{E[Y - \theta D|Z]\}]^2 \\ &= -[Y - E[Y|Z] - \theta\{D - E[D|Z]\}]^2. \end{aligned}$$

In this objective function the conditional mean $E[\cdot|Z]$, which is the linear projection on the set of all functions of Z with finite second moment, can be replaced with the projection on any other closed linear set of functions of Z.

This replacement corresponds to profiling out with linear projections rather than conditional means, and would also have a with respect to θ that is Neyman orthogonal.

This profiling approach can also be applied to obtain Neyman orthogonal semiparametric likelihoods that can be used for efficient estimation of parameters of semiparametric models.

Consider $q(W, \gamma, \theta) = \ln f(W|\theta, \gamma)$ where W is a data observation, θ is a parameter vector, γ is a function and $f(W|\gamma, \theta)$

It follows from Newey (1994) and Theorem 2.5 of Chernozhukov et al. (2018) that for $\gamma(\theta) = \arg\max_\gamma E[\ln f(W|\gamma, \theta)]$, $\partial \ln f(W|\theta, \gamma(\theta))/\partial\theta$ is Neyman orthogonal, without requiring that $f(W|\gamma, \theta)$ be correctly specified.

Newey (1994) also showed that under correction specification, where the W has pdf $f(W|\theta_0, \gamma_0)$, the function $\partial \ln f(W|\theta_0, \gamma(\theta_0))/\partial\theta$ is the efficient score.

Thus we see that the efficient score in a correctly specified semiparametric model is Neyman orthogonal, and that profiling out the nonparametric component is a semiparametric likelihood produces a Neyman orthogonal moment function even without correct specification.

INFLUENCE FUNCTIONS

An *asymptotically linear* estimator $\hat{\theta}$ is one where there is function $\psi(w)$ with $E[\psi(W)] = 0$, $E[\psi(W)^2] < \infty$, and

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sum_{i=1}^{n} \psi(W_i)/\sqrt{n} + o_p(1).$$

The function $\psi(w)$ is referred to as the *influence function*.

An equivalent property that motivates this terminology is

$$\hat{\theta} = \theta_0 + \sum_{i=1}^{n} \psi(W_i)/n + o_p(1/\sqrt{n}).$$

Thus, to first order the influence of a single observation on $\hat{\theta}$ is $\psi(W_i)$.

A simple example is the mean where $\theta_0 = E[a(W_i)]$ for some known function $a(W_i)$.

Here

$$\psi(w) = a(w) - E[a(W_i].$$

$$\hat{\theta} = \theta_0 + \sum_{i=1}^{n} \psi(W_i)/n + o_p(1/\sqrt{n}).$$

Yet another way to write this equation: Think of $\hat{\theta} = \theta(\hat{F})$ where $\hat{F}$ is the empirical distribution,

$$\begin{aligned}\hat{\theta} &= \theta_0 + \int \psi(w)\hat{F}(dw) + o_p(1/\sqrt{n}) \\ &= \theta_0 + \int \psi(w)\{\hat{F} - F_0\}(dw) + o_p(1/\sqrt{n}),\end{aligned}$$

where F_0 denotes the true distribution of the data and the second equality follows by $E[\psi(W)] = \int \psi(w)F_0(dw) = 0$.

Thus we can think of top equation as a first-order expansion of $\hat{\theta}$ in the empirical distribution $\hat{F}$.

It is helpful to think of the true parameter value itself as function of the true distribution F_0 of the data, say $\theta_0 = \theta(F_0)$ for some function $\theta(F)$ of the distribution F of a single observation, and the estimator $\hat{\theta}$ as the same function the data empirical distribution, which $\hat{\theta}$ often is, so $\hat{\theta} = \theta(\hat{F})$ for some function $\theta(F)$.

Then the equation we started with

$$\hat{\theta} = \theta_0 + \sum_{i=1}^{n} \psi(W_i)/n + o_p(1/\sqrt{n}),$$

becomes

$$\theta(\hat{F}) = \theta(F_0) + \int \psi(w)\{\hat{F} - F_0\}(dw) + o_p(1/\sqrt{n}).$$

Thus we can think of the equation that defines the influence function as a first-order expansion of an estimator in the empirical distribution of the data, with the influence function characterizing the first term.

It turns out that across a wide class of distributions, including for complicated $\theta(F)$, this expansion continues to hold, with

$$\theta(F) = \theta(F_0) + \int \psi(w)\{F - F_0\}(dw) + R_2(F).$$

where F is any distribution obeying some regularity conditions and R_2 is a second order remainder.

$$\theta(F) = \theta(F_0) + \int \psi(w)\{F - F_0\}(dw) + R_2(F)$$
$$= \theta(F_0) + \int \psi(w)F(dw) + R_2(F).$$

Thus we see that the influence function gives the first-order effect of F on $\theta(F)$; this is often referred to as a generalized Von Mises expansion, with Von Mises (1947) focusing on this expansion when F is the empirical CDF.

This are two important uses for t is expansion:

1. We can use the behavior of $\theta(F)$ as F varies to obtain the influence function.

2. The linear (second) term in the expansion F is first order, so when some $\tilde{F}$ that is biased (e.g. kernel, NN, etc) is used the expectation of the second term $E[\int \psi(w)\tilde{F}(dw)]$ can be a large source of bias; this bias can be reduced by an influence correction term $\int \tilde{\psi}(w)\hat{F}(dw) = \sum_{i=1}^{n} \tilde{\psi}(W_i)/n$ to form a debiased estimator $\tilde{\theta} = \theta(\tilde{F}) + \int \tilde{\psi}(w)\hat{F}(dw)$ that will be Neyman orthogonal under certain conditions; we discuss this influence function based approach to Neyman orthogonality in the next Section.

We finish this Section by showing how to do 1., i.e. how use the Von Mises expanion to obtain the influence function of an estimator.

First, $\theta(F)$ is simply the probability limit of $\hat{\theta}$ when F is a general distribution, i.e. F is unrestricted except for regularity conditions (Newey, 1994).

Consider $F_\tau = (1-\tau)F_0 + \tau H = F_0 + \tau(H - F_0)$.

Then from first-order expansion

$$\begin{aligned}\frac{d\theta(F_\tau)}{d\tau} &= \frac{d}{\delta\tau}\int \psi(w)\tau\{H - F_0\}(dw) = \frac{d}{\delta\tau}\int \psi(w)\tau H(dw) \\ &= \int \psi(w)H(dw).\end{aligned}$$

Get $\psi(\bar{w})$ at any $w = \bar{w}$ by taking H to be distribution with probability 1 at $\bar{w}$.

Justified in Ichimura and Newey (2022); references to influence function formulae given below.

NEYMAN ORTHOGONALITY VIA THE INFLUENCE FUNCTION

$$\theta(F) = \theta(F_0) + \int \psi(w)F(dw) + R_2(F).$$

** Additional explanation) Consider again the expansion displayed immediately above.

Let $\tilde{F}$ be some distribution estimator and $\tilde{\psi}(w)$ be an associated estimator of the influence function satisfying

$$\int \tilde{\psi}(w)\tilde{F}(dw) = 0.$$

Let the influence function bias corrected estimator mentioned above be

$$\tilde{\theta} = \theta(\tilde{F}) + \int \tilde{\psi}(w)\hat{F}(dw).$$

Applying the expnasion to this gives

$$\begin{aligned}
\tilde{\theta} - \theta_0 &= \int \psi(w)\tilde{F}(dw) + \int \tilde{\psi}(w)\hat{F}(dw) + R_2(\tilde{F}) \\
&= \int \psi(w)\hat{F}(dw) - \int \psi(w)\hat{F}(dw) + \int \psi(w)\tilde{F}(dw) + \int \tilde{\psi}(w)\hat{F}(dw) + R_2(\hat{F}) \\
&= \int \psi(w)\hat{F}(dw) - \int \psi(w)\{\hat{F} - \tilde{F}\}(dw) + \int \tilde{\psi}(w)\{\hat{F} - \tilde{F}\}(dw) + R_2(\tilde{F}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \psi(W_i) + \tilde{R}_2 + R_2(\tilde{F}),\ \tilde{R}_2 = \int \{\tilde{\psi}(w) - \psi(w)\}(\hat{F} - \tilde{F})(dw).
\end{aligned}$$

Summarizing,

$$\tilde{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \psi(W_i) + \tilde{R}_2 + R_2(\tilde{F}),\ \tilde{R}_2 = \int \{\tilde{\psi}(w) - \psi(w)\}(\hat{F} - \tilde{F})(dw).$$

Both $\tilde{R}_2$ and $R_2(\tilde{F})$ are second order, so there is no second-order effect of the nonparametric estimation, i.e. we have Neyman orthogonality.

We can see how adding the bias correction term $\int \tilde{\psi}(w)\hat{F}(dw)$ helps with this.

Assuming $\int \tilde{\psi}(w)\hat{F}(dw)$ is constructed with cross-fitting

$$E[\int \tilde{\psi}(w)\hat{F}(dw)|\tilde{\psi}(w)] = \int \tilde{\psi}(w)F_0(dw) = -\int \tilde{\psi}(w)\{\tilde{F} - F_0\}(dw).$$

Thus, in expectation the bias correction term $\int \tilde{\psi}(w)\hat{F}(dw)$ term estimates the negative of the first order term $\int \psi(w)\{\tilde{F} - F_0\}(dw)$ that appears in the expansion.

We now turn to bias correcting in GMM where first step machine learners or other nonparametric estimators appear in the moment function. We follow closely Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2022, Locally Robust Semiparametric Estimation").

Let $g(W, \gamma.\theta)$ denote a vector of identifying moment functions, θ_0 the true parameter vector, and γ_0 the probability limit of some machine learner when the distribution of a single observation W is the true distribution F_0.

We assume that θ_0 is the unique solution tos

$$E[g(W, \gamma_0, \theta)] = 0,$$

i.e. that the moment functions functions identify θ_0.

An example from Lecture 2 is $g(W, \gamma, \theta) = m(W, \gamma) - \theta$.

A general way to construct an orthogonal moment function is to add to $g(W, \gamma, \theta)$ a bias correction term that we denote by $\phi(W, \gamma, \alpha, \theta)$, where α denotes an additional, unknown bias correction function.

The orthogonal moment function has the form

$$\psi(W, \gamma, \alpha, \theta) = g(W, \gamma, \theta) + \phi(W, \gamma, \alpha, \theta).$$

To find $\phi(W, \gamma, \alpha, \theta)$ we use influence functions.

Let $\gamma(F)$ denote the probability limit of the first step machine learner when F is any true distribution satisfying some regularity conditions (e.g. F is a continuous distribution with certain finite moments).

In Lecture 3 where $\gamma_0(X) = E[Y|X]$ we have $\gamma(F)(X) = E_F[Y|X]$.

For a scalar $\tau \in [0, 1]$ and a general distribution H let $F_\tau = (1 - \tau)F_0 + \tau H$.

The function $\phi(W, \gamma, \alpha, \theta)$ is the first step influence function (FSIF) such that for every H,

$$\left.\frac{\partial E[g(W, \gamma(F_\tau), \theta)]}{\partial \tau}\right|_{\tau=0} = \int \phi(w, \gamma_0, \alpha_0, \theta) H(dw),$$
$$E[\phi(w, \gamma_0, \alpha_0, \theta)] = 0.$$

** Note that ϕ is the influence function for γ in $E[g(W, \gamma, \theta)]$, treating the distribution in this expectation as known.

**This works because the estimation of the distributioin in the expectation is already accounted for in the presence of the sample average of $g(W_i, \gamma, \theta)$.

$$\left.\frac{\partial E[g(W,\gamma(F_\tau),\theta)]}{\partial \tau}\right|_{\tau=0} = \int \phi(w,\gamma_0,\alpha_0,\theta)H(dw),$$
$$E[\phi(w,\gamma_0,\alpha_0,\theta)] = 0.$$

The $\phi(W,\gamma,\alpha,\theta)$ satisfying these conditions is known in statistics as as the influence function of $E[g(W,\gamma(F),\theta)]$.

It can be shown that in the setting of Lecture 2,

$$\phi(W,\gamma,\alpha,\theta) = \alpha(X)\{Y-\gamma(X)\},$$

so that the orthogonal moment function is

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

as we showed in Lecture 2.

Many formulae for $\phi(W, \gamma, \alpha, \theta)$ are available.

Newey (1994) gives formulae when γ is

–conditional means

–nonparametric least squares projections

–conditional pdfs.

Ichimura and Newey (2022, "The Influence Function of Semiparamtric Estimators", Quantitative Economics) gives formulae when γ is

–conditional quantiles and conditional expectiles

–other conditional location and scale measures

–projection versions of conditional location and scale

–solutions to conditional moment restrictions.

More available from recent working papers, e.g. for parameters that depend on nonparametric function identified by a conditional moment restriction.

There is a quite simple proof of Neyman orthogonality.

Consider H in $F_\tau = (1-\tau)F_0 + \tau H$ such that the value of α_0 does not vary with τ.

Zero expectation property of influence function implies

$$0 \equiv E_{F_\tau}[\phi(W, \gamma(F_\tau), \alpha_0, \theta)].$$

Differentiating this identity with respect to τ at $\tau = 0$ and applying the chain rule gives

$$\begin{aligned} 0 &= \int \phi(w, \gamma_0, \alpha_0, \theta) H(dw) + \left.\frac{\partial E[\phi(W, \gamma(F_\tau), \alpha_0, \theta)]}{\partial \tau}\right|_{\tau=0} \\ &= \left.\frac{\partial E[g(W, \gamma(F_\tau), \theta)]}{\partial \tau}\right|_{\tau=0} + \left.\frac{\partial E[\phi(W, \gamma(F_\tau), \alpha_0, \theta)]}{\partial \tau}\right|_{\tau=0} \\ &= \left.\frac{\partial E[\psi(W, \gamma(F_\tau), \alpha_0, \theta)]}{\partial \tau}\right|_{\tau=0} \end{aligned}$$

Orthogonality with respect to variation in $\gamma(F_\tau)$; debiasing function $\phi(W, \gamma, \alpha_0, \theta)$ cancels out effect of γ on $E[g(W, \gamma, \theta)]$.

The function $\phi(W, \gamma, \alpha, \theta)$ provides no additional identifying information beyond $g(W, \gamma, \theta)$ because $E[\phi(W, \gamma_0, \alpha_0, \theta_0)] = 0$ for all true distributions due to the zero mean property of an influence function.

Consequently adding $\phi(W, \gamma, \alpha, \theta)$ to $g(W, \gamma, \theta)$ does not change asymptotic variance of estimator of θ.

Adding $\phi(W, \gamma, \alpha, \theta)$ does debias.

This is a nonparametric bias correction, so that Neyman orthogonality does not depend on any restrictions on the distribution of the data.

General properties of $\psi(W, \gamma, \alpha, \theta)$ are

A) Orthogonality with respect to γ: for $\delta \in \mathbb{R}$ and any allowed γ,

$$\left.\frac{\partial E[\psi(W, \gamma_0 + \delta(\gamma - \gamma_0), \alpha_0, \theta)]}{\partial \delta}\right|_{\delta=0} = 0.$$

B) Robustness with respect to α: For all α,

$$E[\psi(W, \gamma_0, \alpha, \theta_0)] = 0.$$

C) Also robust to γ (called doubly robust) if and only if $E[\psi(W, \gamma, \alpha_0, \theta_0)]$ is linear (affine) in γ, so

$$E[\psi(W, \gamma, \alpha_0, \theta_0)] = 0.$$

A) - C) are shown in Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2016, "Locally Robust Semiparametric Estimation," *Econometrica*, 2022; LR).

A stronger orthogonality property is generally useful for asymptotic theory, namely for a function norm $\|\gamma\|$,

$$|E[\psi(W, \gamma, \alpha_0, \theta_0)]| \leq C \|\gamma - \gamma_0\|^2.$$

** Proof of B) robustness with respect to α, that is, for all α, $E[\psi(W, \gamma_0, \alpha, \theta_0)] = 0$.

Suppose that there is F with $\alpha(F) = \alpha$. We take F to be the true distribution and choose a path where the deviation is F_0 and assume that $\gamma((1-\tau)F + \tau F_0) = \gamma_0$ for all τ close enough to zero.

Then by the definition of ϕ,

$$\begin{aligned} E[\phi(W, \gamma_0, \alpha, \theta_0)] &= \int \phi(w, \gamma_0, \alpha, \theta_0) F_0(dw) \\ &= \frac{\partial}{\partial \tau} E[g(W, \gamma((1-\tau)F + \tau F_0), \theta_0)] \\ &= \frac{\partial}{\partial \tau} E[g(W, \gamma_0, \theta_0)] = 0. \end{aligned}$$

**Double robustness is equivalent to

$$E[\psi(W, \gamma, \alpha_0, \theta_0)] = 0.$$

for all γ, since we showed that generally $E[\psi(W, \gamma_0, \alpha, \theta_0)] = 0$ for all α. Thus, double robustness impli $[\psi(W, \gamma, \alpha_0, \theta_0)]$ is linear (affine) in γ, since a constant function is linear. Also we can show that if $E[\psi(W, \gamma, \alpha_0, \theta_0)]$ is linear (affine) in γ, then above equation is satisfied. Linearity condition A) above, which is Neyman orthogonality with respect to γ, imply above equation. Thus, linearity and Neyman orthogonality are necessary and sufficient for double robustness. For example, do not have double robustness when $\theta_0 = E[m(W, \gamma_0)]$ and $m(W, \gamma)$ is nonlinear in γ or $\phi(W, \gamma, \alpha_0, \theta)$ is nonlinear in γ, such as for γ a nonparametric quantile regression.

Another characterization of double robustness is given in Rotnitzky, A., E. Smucler, and J.M. Robins (2021): "Characterization of Parameters with a Mixed Bias Property," *Biometrika* 108, 231-238. This paper gives two examples that they claim is not covered in LR paper; turns out all examples are covered by LR paper, as detailed in private correspondence.

**Need estimator $\hat{\alpha}_\ell$ of the debiasing function α_0.

One approach is to figure out a formula for α_0 and plug in machine learner.

Better approach is use Neyman orthogonality with respect to γ, property A), to construct an *automatic* learner $\hat{\alpha}_\ell$.

In Lecture 2 we described how to do that for $\psi(W, \gamma, \alpha_0, \theta_0) = m(W, \gamma) - \theta + \alpha(X)[Y - \gamma(X)]$.

**In general we only have moment conditions implied by A), one moment condition for each possible direction $\gamma - \gamma_0$ away from γ_0; there are cases known for NPIV where these moment conditions cannot be turned into objective functioin that can be minimized to get α_0.

REGULARITY CONDITIONS FOR DML

Give here conditions from LR for orthogonal moment function constructed from adding an influence function, bias correction to identifying moment function.

Have relatively simple regularity conditions for

$$\sqrt{n}\hat{\psi}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \gamma_0, \alpha_0, \theta_0) + o_p(1). \tag{1}$$

This means first step estimation of $\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\theta}$ does not affect asymptotic distribution.

Assumption 1: $E[\|\psi(W, \theta_0, \gamma_0, \alpha_0)\|^2] < \infty$ *and*

$$i) \int \|g(w, \hat{\gamma}_\ell, \theta_0) - g(w, \gamma_0, \theta_0)\|^2 F_0(dw) \xrightarrow{p} 0;$$

$$ii) \int \|\phi(w, \hat{\gamma}_\ell, \alpha_0, \theta_0) - \phi(w, \gamma_0, \alpha_0, \theta_0)\|^2 F_0(dw) \xrightarrow{p} 0,$$

$$iii) \int \left\|\phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \phi(w, \gamma_0, \alpha_0, \theta_0)\right\|^2 F_0(dw) \xrightarrow{p} 0.$$

These are mild mean square consistency conditions for $\hat{\gamma}_\ell$ and $(\hat{\alpha}_\ell, \tilde{\theta}_\ell)$ separately.

Let

$$\hat{\Delta}_\ell(w) = \phi(w, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \phi(w, \hat{\gamma}_\ell, \alpha_0, \theta_0) + \phi(w, \gamma_0, \alpha_0, \theta_0).$$

This is a second order term in $\hat{\gamma}, \hat{\alpha}$; for example if $\phi(w, \gamma, \alpha) = \alpha(x)[y - \gamma(x)]$ then $\hat{\Delta}_\ell(w) = -[\hat{\alpha}(x) - \alpha_0(x)][\hat{\gamma}(x) - \gamma_0(x)]$.

Assumption 2: *For each* $\ell = 1, ..., L$*, either i)*

$$\sqrt{n} \int \hat{\Delta}_\ell(w) F_0(dw) \xrightarrow{p} 0, \quad \int \left\| \hat{\Delta}_\ell(w) \right\|^2 F_0(dw) \xrightarrow{p} 0,$$

or ii) $\sum_{i \in I_\ell} \left\| \hat{\Delta}_\ell(W_i) \right\| / \sqrt{n} \xrightarrow{p} 0$*, or iii)* $\sum_{i \in I_\ell} \hat{\Delta}_\ell(W_i) / \sqrt{n} \xrightarrow{p} 0$.

This condition imposes a rate condition on the interaction remainder $\hat{\Delta}_\ell(w)$, that its average must go to zero faster than $1/\sqrt{n}$.

When $\phi(w, \gamma, \alpha) = \alpha(x)[y - \gamma(x)]$, it is implied by

$$\sqrt{n} \left\| \hat{\gamma}_\ell - \gamma_0 \right\| \left\| \hat{\alpha}_\ell - \alpha_0 \right\| \xrightarrow{p} 0,$$

where $\|a\| = \left(E[a(W)^2] \right)^{1/2}$ is mean square norm.

This sometimes referred to as "doubly robust rate condition",

Assumption 3: *For each* $\ell = 1, ..., L$*, i)* $\int \phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell) F_0(dw) = 0$ *with probability approaching one*; *and either ii)* $\bar{\psi}(\gamma, \alpha_0, \theta_0) = E[\psi(W, \gamma, \alpha_0, \theta_0)]$ *is affine in* γ; *or iii)* $\|\hat{\gamma}_\ell - \gamma_0\| = o_p(n^{-1/4})$ *and* $\left\|\bar{\psi}(\gamma, \alpha_0, \theta_0)\right\| \leq C \|\gamma - \gamma_0\|^2$ *for all* γ *with* $\|\gamma - \gamma_0\|$ *small enough; or* iv)$\sqrt{n}\bar{\psi}(\hat{\gamma}_\ell, \alpha_0, \theta_0) \xrightarrow{p} 0$.

Assumption 3 incorporates $E[\phi(w, \gamma_0, \alpha, \theta)] = 0$.

Doubly robust moment allowed for through ii), in which case Assumption 3 imposes no conditions additional to Assumptions 1 and 2.

Conditions iii) and iv) are alternative small bias conditions that are only required to hold for $\hat{\gamma}_\ell$, and not for $\hat{\alpha}_\ell$.

Assumption 3: *For each* $\ell = 1, ..., L$, *i)* $\int \phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell) F_0(dw) = 0$ *with probability approaching one*; *and either ii)* $\bar{\psi}(\gamma, \alpha_0, \theta_0) = E[\psi(W, \gamma, \alpha_0, \theta_0)]$ *is affine in* γ; *or iii)* $\|\hat{\gamma}_\ell - \gamma_0\| = o_p(n^{-1/4})$ *and* $\left\|\bar{\psi}(\gamma, \alpha_0, \theta_0)\right\| \leq C \|\gamma - \gamma_0\|^2$ *for all* γ *with* $\|\gamma - \gamma_0\|$ *small enough; or* iv)$\sqrt{n}\bar{\psi}(\hat{\gamma}_\ell, \alpha_0, \theta_0) \xrightarrow{p} 0$.

Condition iii) requires a faster than $n^{-1/4}$ rate for $\hat{\gamma}$ as is needed for the nonlinear remainder to be smaller than $1/\sqrt{n}$.

In many cases iii) will be satisfied for the mean square norm $\|a\| = \sqrt{E[a(W)^2]}$ so that Assumptions 1-3 will only require mean-square convergence rates.

This is important for complicated machine learners where mean square rates are often known but not much else is known about large sample properties.

Faster than $n^{-1/4}$ convergence is never required for $\hat{\alpha}$.

Theorem: *(Lemma 8 of LR) If Assumptions 1-3 are satisfied then equation (1) is satisfied.*