The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. C1–C68.
doi: 10.1111/ectj.12097

Double/debiased machine learning for treatment and structural parameters

VICTOR CHERNOZHUKOV[†], DENIS CHETVERIKOV[‡], MERT DEMIRER[†], ESTHER DUFLO[†], CHRISTIAN HANSEN[§], WHITNEY NEWEY[†] AND JAMES ROBINS[‖]

[†]*Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139, USA.*
E-mail: vchern@mit.edu, mdemirer@mit.edu, duflo@mit.edu, wnewey@mit.edu

[‡]*University of California Los Angeles, 315 Portola Plaza, Los Angeles, CA 90095, USA.*
E-mail: chetverikov@econ.ucla.edu

[§]*University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, USA.*
E-mail: chansen1@chicagobooth.edu

[‖]*Harvard University, 677 Huntington Avenue, Boston, MA 02115, USA.*
E-mail: robins@hsph.harvard.edu

First version received: October 2016; final version accepted: June 2017

Summary We revisit the classic semi-parametric problem of inference on a low-dimensional parameter θ_0 in the presence of high-dimensional nuisance parameters η_0. We depart from the classical setting by allowing for η_0 to be so high-dimensional that the traditional assumptions (e.g. Donsker properties) that limit complexity of the parameter space for this object break down. To estimate η_0, we consider the use of statistical or machine learning (ML) methods, which are particularly well suited to estimation in modern, very high-dimensional cases. ML methods perform well by employing regularization to reduce variance and trading off regularization bias with overfitting in practice. However, both regularization bias and overfitting in estimating η_0 cause a heavy bias in estimators of θ_0 that are obtained by naively plugging ML estimators of η_0 into estimating equations for θ_0. This bias results in the naive estimator failing to be $N^{-1/2}$ consistent, where N is the sample size. We show that the impact of regularization bias and overfitting on estimation of the parameter of interest θ_0 can be removed by using two simple, yet critical, ingredients: (1) using Neyman-orthogonal moments/scores that have reduced sensitivity with respect to nuisance parameters to estimate θ_0; (2) making use of cross-fitting, which provides an efficient form of data-splitting. We call the resulting set of methods double or debiased ML (DML). We verify that DML delivers point estimators that concentrate in an $N^{-1/2}$-neighbourhood of the true parameter values and are approximately unbiased and normally distributed, which allows construction of valid confidence statements. The generic statistical theory of DML is elementary and simultaneously relies on only weak theoretical requirements, which will admit the use of a broad array of modern ML methods for estimating the nuisance parameters, such as random forests, lasso, ridge, deep neural nets, boosted trees, and various hybrids and ensembles of these methods. We illustrate the general theory by applying it to provide theoretical properties of the following: DML applied to learn the main regression parameter in a partially linear regression model; DML applied to learn the coefficient on an endogenous variable in a partially linear instrumental variables model; DML applied to learn the average treatment effect and the average treatment effect on the treated under unconfoundedness; DML applied

to learn the local average treatment effect in an instrumental variables setting. In addition to these theoretical applications, we also illustrate the use of DML in three empirical examples.

1. INTRODUCTION AND MOTIVATION

Motivation

We develop a series of simple results for obtaining root-N consistent estimation, where N is the sample size, and valid inferential statements about a low-dimensional parameter of interest, θ_0, in the presence of a high-dimensional or 'highly complex' nuisance parameter, η_0. The parameter of interest will typically be a causal parameter or treatment effect parameter, and we consider settings in which the nuisance parameter will be estimated using machine learning (ML) methods, such as random forests, lasso or post-lasso, neural nets, boosted regression trees, and various hybrids and ensembles of these methods. These ML methods are able to handle many covariates and they provide natural estimators of nuisance parameters when these parameters are highly complex. Here, highly complex formally means that the entropy of the parameter space for the nuisance parameter is increasing with the sample size in a way that moves us outside the traditional framework considered in the classical semi-parametric literature where the complexity of the nuisance parameter space is taken to be sufficiently small. The main contribution of this paper is to offer a general and simple procedure for estimating and to perform inference on θ_0 that is formally valid in these highly complex settings.

EXAMPLE 1.1. (PARTIALLY LINEAR REGRESSION) As a lead example, consider the following partially linear regression (PLR) model as in Robinson (1988):

$$Y = D\theta_0 + g_0(X) + U, \quad E[U \mid X, D] = 0, \tag{1.1}$$

$$D = m_0(X) + V, \quad E[V \mid X] = 0. \tag{1.2}$$

Here, Y is the outcome variable, D is the policy/treatment variable of interest, vector

$$X = (X_1, \ldots, X_p)$$

consists of other controls, and U and V are disturbances.[1] The first equation is the main equation, and θ_0 is the main regression coefficient that we would like to infer. If D is exogenous conditional on controls X, θ_0 has the interpretation of the treatment effect parameter or 'lift' parameter in business applications. The second equation keeps track of confounding, namely the dependence of the treatment variable on controls. This equation is not of interest per se but it is important for characterizing and removing regularization bias. The confounding factors X affect the policy variable D via the function $m_0(X)$ and the outcome variable via the function $g_0(X)$. In many applications, the dimension p of vector X is large relative to N. To capture the feature that p is not vanishingly small relative to the sample size, modern analyses then model p as *increasing* with the sample size, which causes traditional assumptions that limit the complexity of the parameter space for the nuisance parameters $\eta_0 = (m_0, g_0)$ to fail.

[1] We consider the case where D is a scalar for simplicity. Extension to the case where D is a vector of fixed, finite dimension is accomplished by introducing an equation such as (1.2) for each element of the vector.

Regularization bias. A naive approach to estimation of θ_0 using ML methods would be, for example, to construct a sophisticated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function $D\theta_0 + g_0(X)$.[2] Suppose, for the sake of clarity, that we randomly split the sample into two parts: a main part of size n, with observation numbers indexed by $i \in I$, and an auxiliary part of size $N - n$, with observations indexed by $i \in I^c$. For simplicity, we take $n = N/2$ for the moment and we turn to more general cases that cover unequal split-sizes, using more than one split, and achieving the same efficiency as if the full sample were used for estimating θ_0 in the formal development in Section 3. Suppose $\hat{g}_0$ is obtained using the auxiliary sample and that, given this $\hat{g}_0$, the final estimate of θ_0 is obtained using the main sample:

$$\hat{\theta}_0 = \Big(\frac{1}{n}\sum_{i\in I} D_i^2\Big)^{-1} \frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i)). \tag{1.3}$$

The estimator $\hat{\theta}_0$ will generally have a slower than $1/\sqrt{n}$ rate of convergence, namely,

$$|\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{p} \infty. \tag{1.4}$$

As detailed below, the driving force behind this 'inferior' behaviour is the bias in learning g_0.

To heuristically illustrate the impact of the bias in learning g_0, we can decompose the scaled estimation error in $\hat{\theta}_0$ as

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\Big(\frac{1}{n}\sum_{i\in I} D_i^2\Big)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\Big(\frac{1}{n}\sum_{i\in I} D_i^2\Big)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{:=b}.$$

The first term is well behaved under mild conditions, obeying $a \rightsquigarrow N(0, \bar{\Sigma})$ for some $\bar{\Sigma}$. Term b is the regularization bias term, which is not centred and diverges in general. Indeed, we have

$$b = (E[{D_i}^2])^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)) + o_P(1)$$

to the first order. Heuristically, b is the sum of n terms that do not have mean zero, $m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i))$, divided by $\sqrt{n}$. These terms have non-zero mean because, in high-dimensional or otherwise highly complex settings, we must employ regularized estimators – such as lasso, ridge, boosting or penalized neural nets – for informative learning to be feasible. The regularization in these estimators keeps the variance of the estimator from exploding but also necessarily induces substantive biases in the estimator $\hat{g}_0$ of g_0. Specifically, the rate of convergence of (the bias of) $\hat{g}_0$ to g_0 in the root-mean-squared error sense will typically be $n^{-\varphi_g}$ with $\varphi_g < 1/2$. Hence, we expect b to be of stochastic order $\sqrt{n}n^{-\varphi_g} \to \infty$ as D_i is centred at $m_0(X_i) \neq 0$, which then implies (1.4).

Overcoming regularization biases using orthogonalization. Now consider a second construction that employs an orthogonalized formulation obtained by directly partialling out the effect of X from D to obtain the orthogonalized regressor $V = D - m_0(X)$. Specifically, we obtain $\hat{V} = D - \hat{m}_0(X)$, where $\hat{m}_0$ is an ML estimator of m_0 obtained using the auxiliary

[2] For instance, we could use lasso if we believe g_0 is well approximated by a sparse linear combination of pre-specified functions of X. In other settings, we could, for example, use iterative methods that alternate between random forests, for estimating g_0, and least squares, for estimating θ_0.

sample of observations. We are now solving an auxiliary prediction problem to estimate the conditional mean of D given X, so we are doing 'double prediction' or 'double machine learning'.

After partialling the effect of X out from D and obtaining a preliminary estimate of g_0 from the auxiliary sample as before, we can formulate the following debiased ML (DML) estimator for θ_0 using the main sample of observations:[3]

$$\check{\theta}_0 = \Big(\frac{1}{n}\sum_{i\in I}\hat{V}_i D_i\Big)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(Y_i - \hat{g}_0(X_i)). \tag{1.5}$$

By approximately orthogonalizing D with respect to X and approximately removing the direct effect of confounding by subtracting an estimate of g_0, $\check{\theta}_0$ removes the effect of regularization bias that contaminates (1.3). The formulation of $\check{\theta}_0$ also provides direct links to both the classical econometric literature, as the estimator can clearly be interpreted as a linear instrumental variable (IV) estimator, and to the more recent literature on debiased lasso in the context where g_0 is taken to be well approximated by a sparse linear combination of pre-specified functions of X; see, e.g., Belloni et al. (2013, 2014a,b), Javanmard and Montanari (2014b), van de Geer et al. (2014) and Zhang and Zhang (2014).[4]

To illustrate the benefits of the auxiliary prediction step and the estimation of θ_0 with $\check{\theta}_0$, we sketch the properties of $\check{\theta}_0$ here. We can decompose the scaled estimation error of $\check{\theta}_0$ into three components:

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*.$$

The leading term, a^*, will satisfy

$$a^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} V_i U_i \rightsquigarrow N(0, \Sigma)$$

under mild conditions. The second term, b^*, captures the impact of regularization bias in estimating g_0 and m_0. Specifically, we will have

$$b^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}(\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i)),$$

which now depends on the product of the estimation errors in $\hat{m}_0$ and $\hat{g}_0$. Because this term depends only on the product of the estimation errors, it can vanish under a broad range of data-generating processes. Indeed, this term is upper-bounded by $\sqrt{n}n^{-(\varphi_m+\varphi_g)}$, where $n^{-\varphi_m}$ and $n^{-\varphi_g}$ are respectively the rates of convergence of $\hat{m}_0$ to m_0 and $\hat{g}_0$ to g_0; this upper bound can clearly vanish even though both m_0 and g_0 are estimated at relatively slow rates. Verifying that $\check{\theta}_0$ has good properties then requires that the remainder term, c^*, is sufficiently well behaved. Sample

[3] In Section 4, we also consider another debiased estimator, based on the partialling-out approach of Robinson (1988):

$$\check{\theta}_0 = \Big(\frac{1}{n}\sum_{i\in I}\hat{V}_i\hat{V}_i\Big)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(Y_i - \hat{\ell}_0(X_i)), \quad \ell_0(X) = E[Y|X].$$

[4] Each of these works differs in terms of detail but can be viewed through the lens of either debiasing or orthogonalization to alleviate the impact of regularization bias on subsequent estimation and inference.

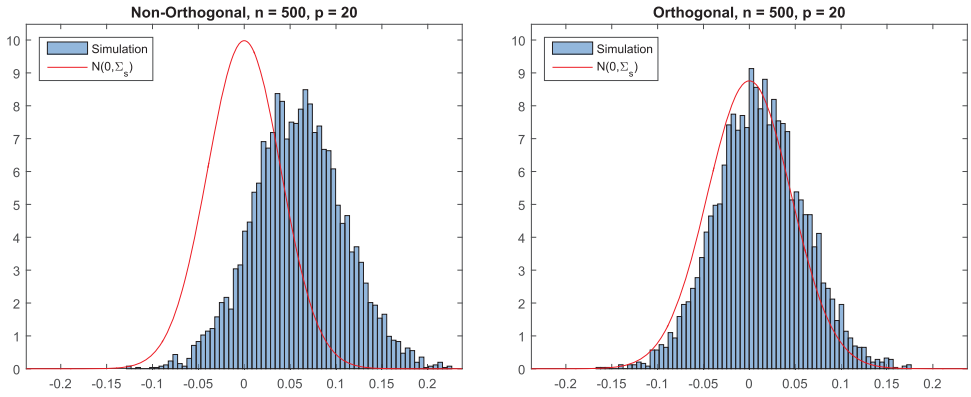

Figure 1. Comparison of the conventional and double ML estimators. [Colour figure can be viewed at wileyonlinelibrary.com]

splitting will play a key role in allowing us to guarantee that $c^* = o_P(1)$ under weak conditions as outlined below and discussed in detail in Section 3.

Figure 1 provides a numerical illustration of the negative impact of regularization bias and the benefit of orthogonalization. The left panel shows the behaviour of a conventional (non-orthogonal) ML estimator, $\hat{\theta}_0$, in the partially linear model in a simple simulation experiment where we learn g_0 using a random forest. The g_0 in this experiment is a very smooth function of a small number of variables, so the experiment is seemingly favourable to the use of random forests a priori. The histogram shows the simulated distribution of the centred estimator, $\hat{\theta}_0 - \theta_0$. The estimator is badly biased, shifted much to the right relative to the true value θ_0. The distribution of the estimator (approximated by the blue histogram) is substantively different from a normal approximation (shown by the red curve) derived under the assumption that the bias is negligible. The right panel shows the behaviour of the orthogonal, DML estimator, $\check{\theta}_0$, in the partially linear model in a simple experiment where we learn nuisance functions using random forests. Note that the simulated data are exactly the same as those underlying in the left panel. The simulated distribution of the centred estimator, $\check{\theta}_0 - \theta_0$ (given by the blue histogram) illustrates that the estimator is approximately unbiased, concentrates around θ_0, and is well approximated by the normal approximation obtained in Section 3 (shown by the red curve).

The role of sample splitting in removing bias induced by overfitting. Our analysis makes use of sample splitting, which plays a key role in establishing that remainder terms, such as c^*, vanish in probability. In the partially linear model, we find that the remainder c^* contains terms such as

$$\frac{1}{\sqrt{n}} \sum_{i \in I} V_i (\hat{g}_0(X_i) - g_0(X_i)), \tag{1.6}$$

which involve $1/\sqrt{n}$-normalized sums of products of structural unobservables from model (1.1)–(1.2) with estimation errors in learning the nuisance functions g_0 and m_0. The use of sample splitting allows simple and tight control of such terms. To see this, assume that observations are independent and recall that $\hat{g}_0$ is estimated using only observations in the auxiliary sample.

Then, conditioning on the auxiliary sample and recalling that $E[V_i|X_i] = 0$, it is easy to verify that term (1.6) has mean zero and variance of order

$$\frac{1}{n}\sum_{i\in I}(\hat{g}_0(X_i) - g_0(X_i))^2 \xrightarrow{p} 0.$$

Thus, the term (1.6) vanishes in probability by Chebyshev's inequality.

While sample splitting allows us to deal with remainder terms such as c^*, its direct application does have the drawback that the estimator of the parameter of interest only makes use of the main sample, which can result in a substantial loss of efficiency as we are only making use of a subset of the available data. However, we can flip the role of the main and auxiliary samples to obtain a second version of the estimator of the parameter of interest. By averaging the two resulting estimators, we can regain full efficiency. Indeed, the two estimators will be approximately independent, so simply averaging them offers an efficient procedure. We call this sample-splitting procedure – where we swap the roles of main and auxiliary samples to obtain multiple estimates and then average the results – 'cross-fitting'. We formally define this procedure and discuss a K-fold version of cross-fitting in Section 3.

Without sample splitting, terms such as (1.6) might not vanish and can lead to poor performance of estimators of θ_0. The difficulty arises because model errors, such as V_i, and estimation errors, such as $\hat{g}_0(X_i) - g_0(X_i)$, are generally related because the data for observation i are used in forming the estimator $\hat{g}_0$. The association can then lead to poor performance of an estimator of θ_0 that makes use of $\hat{g}_0$ as a plug-in estimator for g_0 even when this estimator converges at a very favourable rate, say $N^{-1/2+\epsilon}$.

As an artificial but illustrative example of the problems that can result from overfitting, let $\hat{g}_0(X_i) = g_0(X_i) + (Y_i - g_0(X_i))/N^{1/2-\epsilon}$ for any i in the sample used to form estimator $\hat{g}_0$, and note that the second term provides a simple model that captures overfitting of the outcome variable within the estimation sample. This estimator is excellent in terms of rates. If U_i and D_i are bounded, $\hat{g}_0$ converges uniformly to g_0 at the nearly parametric rate $N^{-1/2+\epsilon}$. Despite this fast rate of convergence, term c^* now explodes if we do not use sample splitting. For example, suppose that the full sample is used to estimate both $\hat{g}_0$ and $\check{\theta}_0$. A simple calculation then reveals that term c^* becomes

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} V_i(\hat{g}_0(X_i) - g_0(X_i)) \propto N^{\epsilon} \to \infty.$$

This bias due to overfitting is illustrated in the left panel of Figure 2. The histogram in the figure gives a simulated distribution for the studentized $\check{\theta}$ resulting from using the full sample and the contrived estimator $\hat{g}(X_i)$ given above. We can see that the histogram is shifted markedly to the left demonstrating substantial bias resulting from overfitting. The right panel of Figure 2 also illustrates that this bias is completely removed by sample splitting. The results in the right panel of Figure 2 make use of the twofold cross-fitting procedure discussed above using the estimator $\check{\theta}$ and the contrived estimator $\hat{g}(X_i)$ exactly as in the left panel. The difference is that $\hat{g}(X_i)$ is formed in one half of the sample and then $\check{\theta}$ is estimated using the other half of the sample. This procedure is then repeated, swapping the roles of the two samples, and the results are averaged. We can see that the substantial bias from the full sample estimator has been removed and that the spread of the histogram corresponding to the cross-fitting estimator is roughly the same as that of the full sample estimator, clearly illustrating the bias-reduction property and efficiency of the cross-fitting procedure.

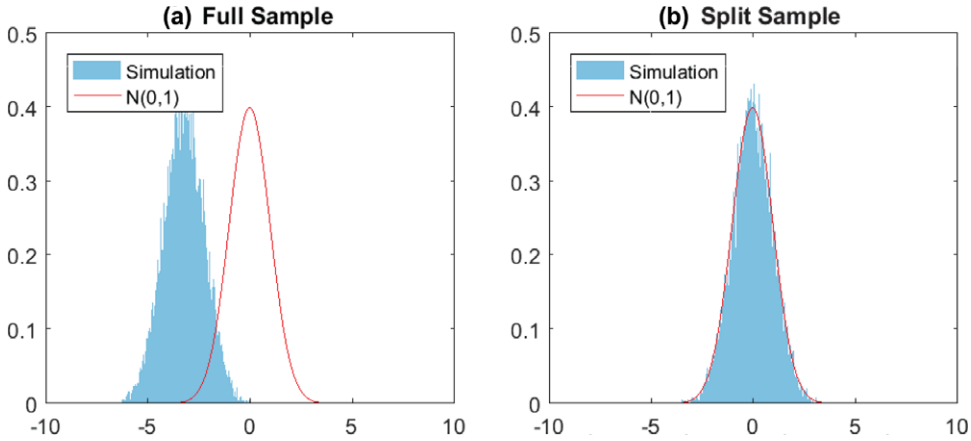

Figure 2. Comparison of full-sample and cross-fitting procedures. [Colour figure can be viewed at wileyonlinelibrary.com]

A less contrived example that highlights the improvements brought by sample splitting is the sparse high-dimensional IV model analysed in Belloni et al. (2012). Specifically, they consider the IV model

$$Y = D\theta_0 + \epsilon,$$

where $E[\epsilon|D] \neq 0$ but instruments Z exist such that $E[D|Z]$ is not a constant and $E[\epsilon|Z] = 0$. Within this model, Belloni et al. (2012) focus on the problem of estimating the optimal instrument, $\eta_0(Z) = E[D|Z]$, using lasso-type methods. If $\eta_0(Z)$ is approximately sparse in the sense that only s terms of the dictionary of series transformations $B(Z) = (B_1(Z), \ldots, B_p(Z))$ are needed to approximate the function accurately, Belloni et al. (2012) require that $s^2 \ll n$ to establish their asymptotic results when sample splitting is not used. However, they show that these results continue to hold under the much weaker requirement that $s \ll n$ if one employs sample splitting. We note that this example provides a prototypical example where Neyman orthogonality holds and ML methods can usefully be adopted to aid in learning structural parameters of interest. We also note that the weaker conditions required when using sample splitting would also carry over to sparsity-based estimators in the partially linear model cited above. We discuss this in more detail in Section 4.

While we find substantial appeal in using sample splitting, one can also use empirical process methods to verify that biases introduced due to overfitting are negligible. For example, consider the problematic term in the partially linear model described previously, $(1/\sqrt{n})\sum_{i\in I} V_i(\hat{g}_0(X_i) - g_0(X_i))$. This term is clearly bounded by

$$\sup_{g\in\mathcal{G}_N} \left| \frac{1}{\sqrt{n}} \sum_{i\in I} V_i(g(X_i) - g_0(X_i)) \right|, \tag{1.7}$$

where $\mathcal{G}_N$ is the smallest class of functions that contains estimators of g_0, $\hat{g}$, with high probability. In conventional semi-parametric statistical and econometric analysis, the complexity of $\mathcal{G}_N$ is controlled by invoking Donsker conditions, which allow verification that terms such as (1.7) vanish asymptotically. Importantly, Donsker conditions require that $\mathcal{G}_N$ has bounded complexity,

specifically a bounded entropy integral. Because of the latter property, Donsker conditions are inappropriate in settings using ML methods where the dimension of X is modelled as increasing with the sample size and estimators necessarily live in highly complex spaces. For example, Donsker conditions rule out even the simplest linear parametric model with high-dimensional regressors with parameter space given by the Euclidean ball with the unit radius:

$$\mathcal{G}_N = \{x \mapsto g(x) = x'\theta;\ \theta \in \mathbb{R}^{p_N} : \|\theta\| \leq 1\}.$$

The entropy of this model, as measured by the logarithm of the covering number, grows at the rate p_N. Without invoking Donsker conditions, one can still show that terms such as (1.7) vanish as long as the entropy of $\mathcal{G}_N$ does not increase with N too rapidly. A fairly general treatment is given by Belloni et al. (2017) who provide a set of conditions under which terms, such as c^*, can vanish making use of the full sample. However, these conditions on the growth of entropy could result in unnecessarily strong restrictions on model complexity, such as very strict requirements on sparsity in the context of lasso estimation, as demonstrated in the IV example mentioned above. Sample splitting allows one to obtain good results under very weak conditions.

Neyman orthogonality and moment conditions. Now we turn to a generalization of the orthogonalization principle above. The first 'conventional' estimator $\hat{\theta}_0$ given in (1.3) can be viewed as a solution to estimating equations

$$\frac{1}{n}\sum_{i\in I}\varphi(W;\hat{\theta}_0,\hat{g}_0) = 0,$$

where φ is a known score function and $\hat{g}_0$ is the estimator of the nuisance parameter g_0. For example, in the partially linear model above, the score function is $\varphi(W;\theta,g) = (Y - \theta D - g(X))D$. It is easy to see that this score function φ is sensitive to biased estimation of g. Specifically, the Gateaux derivative operator with respect to g does not vanish:[5]

$$\partial_g E[\varphi(W;\theta_0,g_0)][g - g_0] \neq 0.$$

The proofs of the general results in Section 3 show that this term's vanishing is a key to establishing good behaviour of an estimator for θ_0.

By contrast, the orthogonalized or double/debiased ML estimator $\check{\theta}_0$ given in (1.5) solves

$$\frac{1}{n}\sum_{i\in I}\psi(W;\check{\theta}_0,\hat{\eta}_0) = 0,$$

where $\hat{\eta}_0$ is the estimator of the nuisance parameter η_0 and ψ is an orthogonalized or debiased score function that satisfies the property that the Gateaux derivative operator with respect to η vanishes when evaluated at the true parameter values:

$$\partial_\eta E[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0. \tag{1.8}$$

We refer to property (1.8) as Neyman orthogonality and to ψ as the Neyman orthogonal score function due to the fundamental contributions in Neyman (1959, 1979), where this notion was introduced. Intuitively, the Neyman orthogonality condition means that the moment conditions used to identify θ_0 are locally insensitive to the value of the nuisance parameter, which allows one

[5] See Section 2 for the definition of the Gateaux derivative operator.

to plug in noisy estimates of these parameters without strongly violating the moment condition. In the partially linear model (1.1)–(1.2), the estimator $\check{\theta}_0$ uses the score function $\psi(W;\theta,\eta) = (Y - D\alpha - g(X))(D - m(X))$, with the nuisance parameter being $\eta = (m, g)$. It is easy to see that these score functions ψ are not sensitive to biased estimation of η_0 in the sense that (1.8) holds. The proofs of the general results in Section 3 show that this property and sample splitting are two generic keys that allow us to establish good behaviour of an estimator for θ_0.

Literature overview

Our paper builds upon two important bodies of research within the semi-parametric literature. The first is the literature on obtaining $\sqrt{N}$-consistent and asymptotically normal estimates of low-dimensional objects in the presence of high-dimensional or non-parametric nuisance functions. The second is the literature on the use of sample splitting to relax entropy conditions. We provide links to each of these bodies of literature in turn.

The problem we study is obviously related to the classical semi-parametric estimation framework, which focuses on obtaining $\sqrt{N}$-consistent and asymptotically normal estimates for low-dimensional components with nuisance parameters estimated by conventional non-parametric estimators, such as kernels or series. See, for example, the work by Levit (1975), Ibragimov and Hasminskii (1981), Bickel (1982), Robinson (1988), Newey (1990, 1994), van der Vaart (1991), Andrews (1994a), Newey et al. (1998, 2004), Robins and Rotnitzky (1995), Linton (1996), Bickel et al. (1998), Chen et al. (2003), van der Laan and Rose (2011) and Ai and Chen (2012). Neyman orthogonality (1.8), introduced by Neyman (1959), plays a key role in optimal testing theory and adaptive estimation, semi-parametric learning theory and econometrics, and, more recently, targeted learning theory. For example, Andrews (1994a), Newey (1994) and van der Vaart (1998) provide a general set of results on estimation of a low-dimensional parameter θ_0 in the presence of nuisance parameters η_0. Andrews (1994a) uses Neyman orthogonality (1.8) and Donsker conditions to demonstrate the key equicontinuity condition

$$\frac{1}{\sqrt{n}} \sum_{i \in I} \left(\psi(W_i; \theta_0, \hat{\eta}) - \int \psi(w; \theta_0, \hat{\eta}) dP(w) - \psi(W_i; \theta_0, \eta_0) \right) \xrightarrow{p} 0,$$

which reduces to (1.6) in the PLR model. Newey (1994) gives conditions on estimating equations and nuisance function estimators so that nuisance function estimators do not affect the limiting distribution of parameters of interest, providing a semi-parametric version of Neyman orthogonality. van der Vaart (1998) discusses use of semi-parametrically efficient scores to define estimators that solve estimating equations, setting averages of efficient scores to zero. He also uses efficient scores to define k-step estimators, where a preliminary estimator is used to estimate the efficient score and then updating is done to further improve estimation; see also comments below on the use of sample splitting.

There is also a related targeted maximum likelihood learning approach, introduced in Scharfstein et al. (1999) in the context of treatments effects analysis. This is substantially generalized by van der Laan and Rubin (2006), who use maximum likelihood in a least favourable direction and then perform one-step or k-step updates using the estimated scores, in an effort to better estimate the target parameter.[6] This procedure is like the least favourable direction

[6] Targeted minimum loss estimation, which shares similar properties, is also discussed in, e.g., van der Laan and Rose (2011) and van der Laan (2015).

approach in semi-parametrics; see, for example, Severini and Wong (1992). The introduction of the likelihood introduces major benefits, such as allowing simple and natural imposition of constraints inherent in the data (e.g. support restrictions when the outcome is binary or censored) and permitting the use of likelihood cross-validation to choose the nuisance parameter estimator. This data adaptive choice of the nuisance parameter has been dubbed the 'super learner' by van der Laan et al. (2007). In subsequent work, van der Laan and Rose (2011) emphasize the use of ML methods to estimate the nuisance parameters for use with the super learner. Much of this work, including recent work such as Luedtke and van der Laan (2016), Toth and van der Laan (2016) and Zheng et al. (2016), focuses on formal results under a Donsker condition, though the use of sample splitting to relax these conditions has also been advocated in the targeted maximum likelihood setting, as discussed below.

The Donsker condition is a powerful classical condition that allows rich structures for fixed function classes $\mathcal{G}$, but unfortunately it is unsuitable for high-dimensional settings. Examples of function classes where a Donsker condition holds include functions of a single variable that have total variation bounded by 1 and functions $x \mapsto f(x)$ that have $r > \dim(x)/2$ uniformly bounded derivatives. As a further example, functions composed from function classes with VC dimensions bounded by p through a fixed number of algebraic and monotonic transforms are Donsker. However, this property will no longer hold if we let $\dim(x)$ grow to infinity with the sample size as this increase in dimension would require that the VC dimension also increases with n. More generally, Donsker conditions are easily violated once dimensions become large. A major point of departure of the present work from the classical literature on semi-parametric estimation is its explicit focus on high-complexity/entropy cases. One way to analyse the problem of estimation in high-entropy cases is to see to what degree equicontinuity results continue to hold while allowing moderate growth of the complexity/entropy of $\mathcal{G}_N$. Examples of papers taking this approach in approximately sparse settings are Belloni et al. (2014b, 2016, 2017), Chernozhukov et al. (2015b), Javanmard and Montanari (2014a), van de Geer et al. (2014) and Zhang and Zhang (2014). In all of these examples, entropy growth must be limited in what can be very restrictive ways. The entropy conditions rule out the contrived overfitting example mentioned above, which does approximate realistic examples, and might otherwise place severe restrictions on the model. For example, in Belloni et al. (2010, 2012), the optimal instrument needs to be sparse of order $s \ll \sqrt{n}$.

A key device that we use to avoid strong entropy conditions is cross-fitting via sample splitting. Cross-fitting is a practical, efficient form of data splitting. Importantly, its use here is not simply as a device to make proofs elementary (which it does), but as a practical method to allow us to overcome the overfitting/high-complexity phenomena that commonly arise in data analysis based on highly adaptive ML methods. Our treatment builds upon the sample-splitting ideas employed in Belloni et al. (2010, 2012) who considered sample splitting in a high-dimensional sparse optimal IV model to weaken the sparsity condition mentioned in the previous paragraph to $s \ll n$. This work, in turn, was inspired by Angrist and Krueger (1995). We also build on Ayyagari (2010) and Robins et al. (2013), where ML methods and sample splitting were used in the estimation of a partially linear model of the effects of pollution while controlling for several covariates. We use the term cross-fitting to characterize our recommended procedure, partly borrowing the jargon from Fan et al. (2012), who employed a slightly different form of sample splitting to estimate the scale parameter in a high-dimensional sparse regression. Of course, the use of sample splitting to relax entropy conditions has a long history in semi-parametric estimation problems. For example, Bickel (1982) considered estimating nuisance functions using a vanishing fraction of the sample, and these results were extended to sample splitting into two

equal halves and discretization of the parameter space by Schick (1986). Similarly, van der Vaart (1998) uses two-way sample splitting and discretization of the parameter space to give weak conditions for k-step estimators using the efficient scores where sample splitting is used to estimate the updates; see also Hubbard et al. (2016). Robins et al. (2008, 2017) use sample splitting in the construction of higher-order influence function corrections in semi-parametric estimation. Some recent work in the targeted maximum likelihood literature, e.g. Zheng and van der Laan (2011), also notes the utility of sample splitting in the context of k-step updating, though this sample splitting approach is different from the cross-fitting approach we pursue.

Plan of the paper. We organize the rest of the paper as follows. In Section 2, we formally define Neyman orthogonality and provide a brief discussion that synthesizes various models and frameworks that can be used to produce estimating equations satisfying this key condition. In Section 3, we carefully define DML estimators and develop their general theory. We then illustrate this general theory by applying it to provide theoretical results for using DML to estimate and carry out inference for key parameters in the PLR model, and for using DML to estimate and carry out inference for coefficients on endogenous variables in a partially linear IV model in Section 4. In Section 5, we provide a further illustration of the general theory by applying it to develop theoretical results for DML estimation and inference for average treatment effects (ATEs) and average treatment effects on the treated (ATTEs) under unconfoundedness, and for DML estimation of local average treatment effects (LATEs) in an IV context within the potential outcomes framework; see Imbens and Rubin (2015). Finally, we apply DML in three empirical illustrations in Section 6. In the Appendix, we define additional notation and present proofs.

Notation. The symbols Pr and E denote probability and expectation operators with respect to a generic probability measure that describes the law of the data. If we need to signify the dependence on a probability measure P, we use P as a subscript in Pr_P and E_P. We use capital letters, such as W, to denote random elements and we use the corresponding lowercase letters, such as w, to denote fixed values that these random elements can take. In what follows, we use $\|\cdot\|_{P,q}$ to denote the $L^q(P)$ norm; for example, we denote $\|f\|_{P,q} := \|f(W)\|_{P,q} := (\int |f(w)|^q dP(w))^{1/q}$. We use x' to denote the transpose of a column vector x. For a differentiable map $x \mapsto f(x)$, mapping $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'} f$ to abbreviate the partial derivatives $(\partial/\partial x')f$, and we correspondingly use the expression $\partial_{x'} f(x_0)$ to mean $\partial_{x'} f(x) \mid_{x=x_0}$, etc.

2. CONSTRUCTION OF NEYMAN ORTHOGONAL SCORE/MOMENT FUNCTIONS

Here we formally introduce the model and we discuss several methods for generating orthogonal scores in a wide variety of settings, including the classical Neyman construction. We also use this as an opportunity to synthesize some recent developments in the literature.

2.1. Moment condition/estimating equation framework

We are interested in the true value θ_0 of the low-dimensional target parameter $\theta \in \Theta$, where Θ is a non-empty measurable subset of $\mathbb{R}^{d_\theta}$. We assume that θ_0 satisfies the moment conditions

$$E_P[\psi(W; \theta_0, \eta_0)] = 0, \tag{2.1}$$

where $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ is a vector of known score functions, W is a random element taking values in a measurable space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ with law determined by a probability measure $P \in \mathcal{P}_N$, and η_0 is the true value of the nuisance parameter $\eta \in T$, where T is a convex subset of some normed vector space with the norm denoted by $\|\cdot\|_T$. We assume that the score functions $\psi_j : \mathcal{W} \times \Theta \times T \to \mathbb{R}$ are measurable once we equip Θ and T with their Borel σ-fields, and we assume that a random sample $(W_i)_{i=1}^N$ from the distribution of W is available for estimation and inference.

As discussed in Section 1, we require the Neyman orthogonality condition for the score ψ. To introduce the condition, for $\widetilde{T} = \{\eta - \eta_0 : \eta \in T\}$ we define the pathwise (or the Gateaux) derivative map $\mathrm{D}_r : \widetilde{T} \to \mathbb{R}^{d_\theta}$,

$$\mathrm{D}_r[\eta - \eta_0] := \partial_r\{E_P[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))]\}, \quad \eta \in T,$$

for all $r \in [0, 1)$, which we assume to exist. For convenience, we also denote

$$\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] := \mathrm{D}_0[\eta - \eta_0], \quad \eta \in T. \tag{2.2}$$

Note that $\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))$ here is well defined because for all $r \in [0, 1)$ and $\eta \in T$,

$$\eta_0 + r(\eta - \eta_0) = (1 - r)\eta_0 + r\eta \in T,$$

as T is a convex set. In addition, let $\mathcal{T}_N \subset T$ be a nuisance realization set such that the estimators $\hat{\eta}_0$ of η_0 specified below take values in this set with high probability. In practice, we typically assume that $\mathcal{T}_N$ is a properly shrinking neighbourhood of η_0. Note that $\mathcal{T}_N - \eta_0$ is the nuisance deviation set, which contains deviations of $\hat{\eta}_0$ from η_0, $\hat{\eta}_0 - \eta_0$, with high probability. The Neyman orthogonality condition requires that the derivative in (2.2) vanishes for all $\eta \in \mathcal{T}_N$.

DEFINITION 2.1. (NEYMAN ORTHOGONALITY) *The score $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ obeys the orthogonality condition at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N \subset T$ if (2.1) holds and the pathwise derivative map $\mathrm{D}_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_N$ and vanishes at $r = 0$; namely,*

$$\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0, \quad \textit{for all } \eta \in \mathcal{T}_N. \tag{2.3}$$

We remark here that condition (2.3) holds with $\mathcal{T}_N = T$ when η is a finite-dimensional vector as long as $\partial_\eta E_P[\psi_j(W; \theta_0, \eta_0)] = 0$ for all $j = 1, \ldots, d_\theta$, where $\partial_\eta E_P[\psi_j(W; \theta_0, \eta_0)]$ denotes the vector of partial derivatives of the function $\eta \mapsto E_P[\psi_j(W; \theta_0, \eta)]$ for $\eta = \eta_0$.

Sometimes it will also be helpful to use an approximate Neyman orthogonality condition as opposed to the exact one given in Definition 2.1.

DEFINITION 2.2. (NEYMAN NEAR-ORTHOGONALITY) *The score $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ obeys the λ_N near-orthogonality condition at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N \subset T$ if (2.1) holds and the pathwise derivative map $\mathrm{D}_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_N$ and is small at $r = 0$; namely,*

$$\|\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0]\| \leq \lambda_N, \quad \textit{for all } \eta \in \mathcal{T}_N, \tag{2.4}$$

where $\{\lambda_N\}_{N \geq 1}$ is a sequence of positive constants such that $\lambda_N = o(N^{-1/2})$.

2.2. Construction of Neyman orthogonal scores

If we start with a score φ that does not satisfy the orthogonality condition above, we first transform it into a score ψ that does. Here we outline several methods for doing so.

2.2.1. Neyman orthogonal scores for likelihood and other M-estimation problems with finite-dimensional nuisance parameters. First, we describe the construction used by Neyman (1959) to derive his celebrated orthogonal score and $C(\alpha)$-statistic in a maximum likelihood setting.[7] Such a construction also underlies the concept of local unbiasedness in the construction of optimal tests in, e.g., Ferguson (1967), and it was extended to non-likelihood settings by Wooldridge (1991). The discussion of Neyman's construction here draws on Chernozhukov et al. (2015a).

To describe the construction, let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$, where $\mathcal{B}$ is a convex set, be the target and the nuisance parameters, respectively. Further, suppose that the true parameter values θ_0 and β_0 solve the optimization problem

$$\max_{\theta\in\Theta,\ \beta\in\mathcal{B}} E_P[\ell(W;\theta,\beta)], \tag{2.5}$$

where $\ell(W;\theta,\beta)$ is a known criterion function. For example, $\ell(W;\theta,\beta)$ can be the log-likelihood function associated with observation W. More generally, we refer to $\ell(W;\theta,\beta)$ as the quasi-log-likelihood function. Then, under mild regularity conditions, θ_0 and β_0 satisfy

$$E_P[\partial_\theta \ell(W;\theta_0,\beta_0)] = 0, \quad E_P[\partial_\beta \ell(W;\theta_0,\beta_0)] = 0. \tag{2.6}$$

Note that the original score function $\varphi(W;\theta,\beta) = \partial_\theta \ell(W;\theta,\beta)$ for estimating θ_0 will not generally satisfy the orthogonality condition. Now consider the new score function, which we refer to as the Neyman orthogonal score,

$$\psi(W;\theta,\eta) = \partial_\theta \ell(W;\theta,\beta) - \mu\partial_\beta \ell(W;\theta,\beta), \tag{2.7}$$

where the nuisance parameter is

$$\eta = (\beta', \mathrm{vec}(\mu)')' \in T = \mathcal{B} \times \mathbb{R}^{d_\theta d_\beta} \subset \mathbb{R}^p, \quad p = d_\beta + d_\theta d_\beta,$$

and μ is the $d_\theta \times d_\beta$ orthogonalization parameter matrix whose true value μ_0 solves the equation

$$J_{\theta\beta} - \mu J_{\beta\beta} = 0 \tag{2.8}$$

for

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta',\beta')} E_P[\partial_{(\theta',\beta')'} \ell(W;\theta,\beta)]|_{\theta=\theta_0;\,\beta=\beta_0}.$$

The true value of the nuisance parameter η is

$$\eta_0 = (\beta_0', \mathrm{vec}(\mu_0)')'; \tag{2.9}$$

[7] The $C(\alpha)$-statistic, or the orthogonal score statistic, has been explicitly used for testing and estimation in high-dimensional sparse models in Belloni et al. (2015).

and when $J_{\beta\beta}$ is invertible, (2.8) has the unique solution,

$$\mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}. \tag{2.10}$$

The following lemma shows that the score ψ in (2.7) satisfies the Neyman orthogonality condition.

LEMMA 2.1. (NEYMAN ORTHOGONAL SCORES FOR QUASI-LIKELIHOOD SETTINGS) *If (2.6) holds, J exists, and $J_{\beta\beta}$ is invertible, then the score ψ in (2.7) is Neyman orthogonal at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N = T$.*

REMARK 2.1. (ADDITIONAL NUISANCE PARAMETERS) Note that the orthogonal score ψ in (2.7) has nuisance parameters consisting of the elements of μ in addition to the elements of β, and Lemma 2.1 shows that Neyman orthogonality holds both with respect to β and with respect to μ. We will find that Neyman orthogonal scores in other settings, including infinite-dimensional ones, have a similar property.

REMARK 2.2. (EFFICIENCY) Note that in this example, μ_0 not only creates the necessary orthogonality but also creates the efficient score for inference on the target parameter θ when the quasi-log-likelihood function is the true (possibly conditional) log-likelihood, as demonstrated by Neyman (1959).

EXAMPLE 2.1. (HIGH-DIMENSIONAL LINEAR REGRESSION) As an application of the construction above, consider the following linear predictive model,

$$Y = D\theta_0 + X'\beta_0 + U, \quad E_P[U(X', D)'] = 0, \tag{2.11}$$

$$D = X'\gamma_0 + V, \quad E_P[VX] = 0, \tag{2.12}$$

where, for simplicity, we assume that θ_0 is a scalar. The first equation here is the main predictive model, and the second equation only plays a role in the construction of the Neyman orthogonal scores. It is well known that θ_0 and β_0 in this model solve the optimization problem (2.5) with

$$\ell(W; \theta, \beta) = -\frac{(Y - D\theta - X'\beta)^2}{2}, \quad \theta \in \Theta = \mathbb{R}, \ \beta \in \mathcal{B} = \mathbb{R}^{d_\beta},$$

where we denote $W = (Y, D, X')'$. Hence, equations (2.6) hold with

$$\partial\ell_\theta(W; \theta, \beta) = (Y - D\theta - X'\beta)D, \quad \partial\ell_\beta(W; \theta, \beta) = (Y - D\theta - X'\beta)X,$$

and the matrix J satisfies

$$J_{\theta\beta} = -E_P[DX'], \quad J_{\beta\beta} = -E_P[XX'].$$

The Neyman orthogonal score is then given by

$$\psi(W; \theta, \eta) = (Y - D\theta - X'\beta)(D - \mu X); \qquad \eta = (\beta', \mathrm{vec}(\mu)')';$$

$$\psi(W; \theta_0, \eta_0) = U(D - \mu_0 X); \qquad \mu_0 = E_P[DX'](E_P[XX'])^{-1} = \gamma_0'. \tag{2.13}$$

If the vector of covariates X here is high-dimensional but the vectors of parameters β_0 and γ_0 are approximately sparse, we can use ℓ_1-penalized least-squares, ℓ_2-boosting, or forward selection methods to estimate β_0 and $\gamma_0 = \mu_0'$, and hence $\mu_0 = (\beta_0', \mathrm{vec}(\mu_0)')'$; see references cited in Section 1.

If $J_{\beta\beta}$ is not invertible, (2.8) typically has multiple solutions. In this case, it is convenient to focus on a minimal norm solution,

$$\mu_0 = \arg\min \|\mu\| \text{ such that } \|J_{\theta\beta} - \mu J_{\beta\beta}\|_q = 0$$

for a suitably chosen norm $\|\cdot\|_q$ on the space of $d_\theta \times d_\beta$ matrices. With an eye on solving the empirical version of this problem, we can also consider the relaxed version of this problem,

$$\mu_0 = \arg\min \|\mu\| \text{ such that } \|J_{\theta\beta} - \mu J_{\beta\beta}\|_q \leq r_N \tag{2.14}$$

for some $r_N > 0$ such that $r_N \to 0$ as $N \to \infty$. This relaxation is also helpful when $J_{\beta\beta}$ is invertible but ill-conditioned. The following lemma shows that using μ_0 in (2.14) leads to Neyman near-orthogonal scores. The proof of this lemma can be found in the Appendix.

LEMMA 2.2. (NEYMAN NEAR-ORTHOGONAL SCORES FOR QUASI-LIKELIHOOD SETTINGS) *If (2.6) holds, J exists, the solution of the optimization problem* (2.14) *exists, and μ_0 is taken to be this solution, then the score ψ defined in* (2.7) *is Neyman λ_N near-orthogonal at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N = \{\beta \in \mathcal{B} : \|\beta - \beta_0\|_q^* \leq \lambda_N / r_N\} \times \mathbb{R}^{d_\theta d_\beta}$, where the norm $\|\cdot\|_q^*$ on $\mathbb{R}^{d_\beta}$ is defined by $\|\beta\|_q^* = \sup_A \|A\beta\|$ with the supremum being taken over all $d_\theta \times d_\beta$ matrices A such that $\|A\|_q \leq 1$.*

EXAMPLE 2.1. (CONTINUED) In the high-dimensional linear regression example above, the relaxation (2.14) is helpful when $J_{\beta\beta} = E_P[XX']$ is ill-conditioned. Specifically, if one suspects that $E_P[XX']$ is ill-conditioned, one can define μ_0 as the solution to the following optimization problem:

$$\min \|\mu\| \text{ such that } \|E_P[DX'] - \mu E_P[XX']\|_\infty \leq r_N. \tag{2.15}$$

Lemma 2.2 then shows that using this μ_0 leads to a score ψ that obeys the Neyman near-orthogonality condition. Alternatively, one can define μ_0 as the solution of the following closely related optimization problem,

$$\min_\mu (\mu E_P[XX']\mu' - \mu E_P[DX] + r_N \|\mu\|_1),$$

whose solution also obeys $\|E_P[DX] - \mu E_P[XX']\|_\infty \leq r_N$, which follows from the first-order conditions. An empirical version of either problem leads to a Lasso-type estimator of the regularized solution μ_0; see Javanmard and Montanari (2014a).

REMARK 2.3. (GIVING UP EFFICIENCY) Note that the regularized μ_0 in (2.14) creates the necessary near-orthogonality at the cost of giving up somewhat on the efficiency of the score ψ. At the same time, regularization may generate additional robustness gains as achieving full efficiency by estimating μ_0 in (2.10) may require stronger conditions.

REMARK 2.4. (CONCENTRATING-OUT APPROACH) The approach for constructing Neyman orthogonal scores described above is closely related to the following concentrating-out approach, which has been used, for example, in Newey (1994), to show Neyman orthogonality when β is infinite dimensional. For all $\theta \in \Theta$, let β_θ be the solution of the following optimization problem:

$$\max_{\beta \in \mathcal{B}} E_P[\ell(W; \theta, \beta)].$$

Under mild regularity conditions, β_θ satisfies

$$\partial_\beta E_P[\ell(W; \theta, \beta_\theta)] = 0, \quad \text{for all } \theta \in \Theta. \tag{2.16}$$

Differentiating (2.16) with respect to θ and interchanging the order of differentiation gives

$$\begin{aligned} 0 &= \partial_\theta \partial_\beta E_P[\ell(W;\theta,\beta_\theta)] = \partial_\beta \partial_\theta E_P[\ell(W;\theta,\beta_\theta)] \\ &= \partial_\beta E_P[\partial_\theta \ell(W;\theta,\beta_\theta) + [\partial_\theta \beta_\theta]' \partial_\beta \ell(W;\theta,\beta_\theta)] \\ &= \partial_\beta E_P[\psi(W;\theta,\beta,\partial_\theta\beta_\theta)]|_{\beta=\beta_\theta}, \end{aligned}$$

where we denote

$$\psi(W;\theta,\beta,\partial_\theta\beta_\theta) := \partial_\theta \ell(W;\theta,\beta) + [\partial_\theta \beta_\theta]' \partial_\beta \ell(W;\theta,\beta).$$

This vector of functions is a score with nuisance parameters $\eta = (\beta', \mathrm{vec}(\partial_\theta\beta_\theta))'$. As before, additional nuisance parameters, $\partial_\theta\beta_\theta$ in this case, are introduced when the orthogonal score is formed. Evaluating these equations at θ_0 and β_0, it follows from the previous equation that $\psi(W;\theta,\beta,\partial_\theta\beta_\theta)$ is orthogonal with respect to β and from $E_P[\partial_\beta \ell(W;\theta_0,\beta_0)] = 0$ that we have orthogonality with respect to $\partial_\theta\beta_\theta$. Thus, maximizing the expected objective function with respect to the nuisance parameters, plugging that maximum back in, and differentiating with respect to the parameters of interest produces an orthogonal moment condition. See also Section 2.2.3.

2.2.2. Neyman orthogonal scores in generalized method of moments problems. The construction in the previous section gives a Neyman orthogonal score whenever the moment conditions (2.6) hold, and, as discussed in Remark 2.2, the resulting score is efficient as long as $\ell(W;\theta,\beta)$ is the log-likelihood function. The question, however, remains about constructing the efficient score when $\ell(W;\theta,\beta)$ is not necessarily a log-likelihood function. In this section, we answer this question and describe a generalized method of moments (GMM)-based method of constructing an efficient and Neyman orthogonal score in this more general case. The discussion here is related to Lee (2005), Bera et al. (2010) and Chernozhukov et al. (2015b).

Because GMM does not require that the moment conditions (2.6) are obtained from the first-order conditions of the optimization problem (2.5), we use a different notation for the moment conditions. Specifically, we consider parameters $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$, where $\mathcal{B}$ is a convex set, whose true values, θ_0 and β_0, solve the moment conditions

$$E_P[m(W;\theta_0,\beta_0)] = 0, \tag{2.17}$$

where $m : \mathcal{W} \times \Theta \times \mathcal{B} \to \mathbb{R}^{d_m}$ is a known vector-valued function, and $d_m \geq d_\theta + d_\beta$ is the number of moment conditions. In this case, a Neyman orthogonal score function is

$$\psi(W;\theta,\eta) = \mu m(W;\theta,\beta), \tag{2.18}$$

where the nuisance parameter is

$$\eta = (\beta', \mathrm{vec}(\mu)')' \in T = \mathcal{B} \times \mathbb{R}^{d_\theta d_m} \subset \mathbb{R}^p, \quad p = d_\beta + d_\theta d_m,$$

and μ is the $d_\theta \times d_m$ orthogonalization parameter matrix whose true value is

$$\mu_0 = (A'\Omega^{-1} - A'\Omega^{-1}G_\beta(G_\beta'\Omega^{-1}G_\beta)^{-1}G_\beta'\Omega^{-1}).$$

Here

$$\begin{aligned} G_\gamma &= \partial_{\gamma'} E_P[m(W;\theta,\beta)]|_{\gamma=\gamma_0} \\ &= [\partial_{\theta'} E_P[m(W;\theta,\beta)], \partial_{\beta'} E_P[m(W;\theta,\beta)]]|_{\gamma=\gamma_0} =: [G_\theta, G_\beta], \end{aligned}$$

for $\gamma = (\theta', \beta')'$ and $\gamma_0 = (\theta_0', \beta_0')'$, A is a $d_m \times d_\theta$ moment selection matrix, Ω is a $d_m \times d_m$ positive definite weighting matrix, and both A and Ω can be chosen arbitrarily. Note that setting

$$A = G_\theta \text{ and } \Omega = \text{Var}_P(m(W; \theta_0, \beta_0)]) = E_P[m(W; \theta_0, \beta_0)m(W; \theta_0, \beta_0)']$$

leads to the efficient score in the sense of yielding an estimator of θ_0 having the smallest variance in the class of GMM estimators (Hansen, 1982), and, in fact, to the semi-parametrically efficient score; see Levit (1975), Nevelson (1977) and Chamberlain (1987). Let $\eta_0 = (\beta_0', \text{vec}(\mu_0)')'$ be the true value of the nuisance parameter $\eta = (\beta', \text{vec}(\mu)')'$. The following lemma shows that the score ψ in (2.18) satisfies the Neyman orthogonality condition.

LEMMA 2.3. (NEYMAN ORTHOGONAL SCORES FOR GMM SETTINGS) *If (2.17) holds, G_γ exists and Ω is invertible, then the score ψ in (2.18) is Neyman orthogonal at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N = T$.*

As in the quasi-likelihood case, we can also consider near-orthogonal scores. Specifically, note that one of the orthogonality conditions that the score ψ in (2.18) has to satisfy is that $\mu_0 G_\beta = 0$, which can be rewritten as

$$A'\Omega^{-1/2}(I - L(L'L)^{-1}L')L = 0,$$

where $L = \Omega^{-1/2}G_\beta$. Here, the part $A'\Omega^{-1/2}L(L'L)^{-1}L'$ can be expressed as $\gamma_0 L'$, where $\gamma_0 = A'\Omega^{-1/2}L(L'L)^{-1}$ solves the optimization problem

$$\min \|\gamma\|_o \text{ such that } \|A'\Omega^{-1/2}L - \gamma L'L\|_\infty = 0,$$

for a suitably chosen norm $\|\cdot\|_o$. When $L'L$ is close to being singular, this problem can be relaxed:

$$\min \|\gamma\|_o \text{ such that } \|A'\Omega^{-1/2}L - \gamma L'L\|_\infty \leq r_N. \tag{2.19}$$

This relaxation leads to Neyman near-orthogonal scores.

LEMMA 2.4. (NEYMAN NEAR-ORTHOGONAL SCORES FOR GMM SETTINGS) *In the set-up above, with γ_0 denoting the solution of (2.19), we have for $\mu_0 := A'\Omega^{-1} - \gamma_0 L'\Omega^{-1/2}$ and $\eta_0 = (\beta_0', vec(\mu_0)')'$ that ψ defined in (2.18) is the Neyman λ_N near-orthogonal score at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N = \{\beta \in \mathcal{B} : \|\beta - \beta_0\|_1 \leq \lambda_N / r_N\} \times \mathbb{R}^{d_\theta d_m}$.*

2.2.3. Neyman orthogonal scores for likelihood and other M-estimation problems with infinite-dimensional nuisance parameters. Here we show that the concentrating-out approach described in Remark 2.4 for the case of finite-dimensional nuisance parameters can be extended to the case of infinite-dimensional nuisance parameters. Let $\ell(W; \theta, \beta)$ be a known criterion function, where θ and β are the target and the nuisance parameters taking values in Θ and $\mathcal{B}$, respectively, and let us assume that the true values of these parameters, θ_0 and β_0, solve the optimization problem (2.5). The function $\ell(W; \theta, \beta)$ is analogous to that discussed above but now, instead of assuming that $\mathcal{B}$ is a (convex) subset of a finite-dimensional space, we assume that $\mathcal{B}$ is some (convex) set of functions, so that β is the functional nuisance parameter. For example, $\ell(W; \theta, \beta)$ could be a semi-parametric log-likelihood where β is the non-parametric part of the model. More generally, $\ell(W; \theta, \beta)$ could be some other criterion function such as the negative of a squared residual. Also, let

$$\beta_\theta = \arg\max_{\beta \in \mathcal{B}} E_P[\ell(W; \theta, \beta)] \tag{2.20}$$

be the concentrated-out non-parametric part of the model. Note that β_θ is a function-valued function. Now consider the score function

$$\psi(W;\theta,\eta) = \frac{d\ell(W;\theta,\eta(\theta))}{d\theta}, \tag{2.21}$$

where the nuisance parameter is $\eta : \Theta \to \mathcal{B}$, and its true value η_0 is given by

$$\eta_0(\theta) = \beta_\theta, \quad \text{for all } \theta \in \Theta.$$

Here, the symbol $d/d\theta$ denotes the full derivative with respect to θ, so that we differentiate with respect to both θ arguments in $\ell(W;\theta,\eta(\theta))$. The following lemma shows that the score ψ in (2.21) satisfies the Neyman orthogonality condition.

LEMMA 2.5. (NEYMAN ORTHOGONAL SCORES VIA CONCENTRATING-OUT APPROACH) *Suppose that* (2.5) *holds, and let* T *be a convex set of functions mapping* Θ *into* $\mathcal{B}$ *such that* $\eta_0 \in T$. *Also, suppose that for each* $\eta \in T$, *the function* $\theta \mapsto \ell(W;\theta,\eta(\theta))$ *is continuously differentiable almost surely. Then, under mild regularity conditions, the score* ψ *in* (2.21) *is Neyman orthogonal at* (θ_0,η_0) *with respect to the nuisance realization set* $\mathcal{T}_N = T$.

As an example, consider the partially linear model from Section 1. Let

$$\ell(W;\theta,\beta) = -\frac{1}{2}(Y - D\theta - \beta(X))^2,$$

and let $\mathcal{B}$ be the set of functions of X with finite mean square. Then

$$(\theta_0,\beta_0) = \arg\max_{\theta\in\Theta,\beta\in\mathcal{B}} E_P[\ell(W;\theta,\beta)]$$

and

$$\beta_\theta(X) = E_P[Y - D\theta|X], \quad \theta \in \Theta.$$

Hence, (2.21) gives the following Neyman orthogonal score

$$\begin{aligned}\psi(W;\theta,\beta_\theta) &= -\frac{1}{2}\frac{d\{Y - D\theta - E_P[Y - D\theta|X]\}^2}{d\theta}\\ &= (D - E_P[D|X]) \times (Y - E_P[Y|X] - (D - E_P[D|X])\theta)\\ &= (D - m_0(X)) \times (Y - D\theta - g_0(X)),\end{aligned}$$

which corresponds to the estimator θ_0 described in Section 1 in (1.5).

It is important to note that the concentrating-out approach described here gives a Neyman orthogonal score without requiring that $\ell(W;\theta,\beta)$ is the log-likelihood function. Except for the technical conditions needed to ensure the existence of derivatives and their interchangeability, the only condition that is required is that θ_0 and β_0 solve the optimization problem (2.5). If $\ell(W;\theta,\beta)$ is the log-likelihood function, however, it follows from Newey (1994, p. 1359), that the concentrating-out approach actually yields the efficient score. An alternative, but closely related, approach to derive the efficient score in the likelihood setting would be to apply Neyman's construction described above for a one-dimensional least favourable parametric submodel; see Severini and Wong (1992) and Chapter 25 of van der Vaart (1998).

REMARK 2.5. (GENERATING ORTHOGONAL SCORES BY VARYING $\mathcal{B}$) When we calculate the concentrated-out non-parametric part β_θ, we can use some other set of functions Υ instead of $\mathcal{B}$ on the right-hand side of (2.20):

$$\beta_\theta = \arg\max_{\beta \in \Upsilon} E_P[\ell(W; \theta, \beta)].$$

By replacing $\mathcal{B}$ by Υ, we can generate a different Neyman orthogonal score. Of course, this replacement might also change the true value θ_0 of the parameter of interest, which is an important consideration for the selection of Υ. For example, consider the partially linear model and assume that X has two components, X_1 and X_2. Now, consider what would happen if we replaced $\mathcal{B}$, which is the set of functions of X with finite mean square, by the set of functions Υ that is the mean square closure of functions that are additive in X_1 and X_2:

$$\Upsilon = \overline{\{h(X_1) + h(X_2)\}}.$$

Let $\bar{E}_P$ denote the least-squares projection on Υ. Then, applying the previous calculation with $\bar{E}_P$ replacing E_P gives

$$\psi(W; \theta, \beta_\theta) = (D - \bar{E}_P[D|X]) \times (Y - \bar{E}_P[Y|X] + (D - \bar{E}_P[D|X])\theta),$$

which provides an orthogonal score based on additive function of X_1 and X_2. Here, it is important to note that the solution to $E_P[\psi(W, \theta, \beta_\theta)] = 0$ will be the true θ_0 only when the true function of X in the partially linear model is additive. More generally, the solution of the moment condition would be the coefficient of D in the least-squares projection of Y on functions of the form $D\theta + h_1(X_1) + h_1(X_2)$. Note, though, that the corresponding score is orthogonal by virtue of additivity being imposed in the estimation of $\bar{E}_P[Y|X]$ and $\bar{E}_P[D|X]$.

2.2.4. Neyman orthogonal scores for conditional moment restriction problems with infinite-dimensional nuisance parameters. Next we consider the conditional moment restrictions framework studied in Chamberlain (1992). To define the framework, let W, R and Z be random vectors taking values in $\mathcal{W} \subset \mathbb{R}^{d_w}$, $\mathcal{R} \subset \mathbb{R}^{d_r}$ and $\mathcal{Z} \subset \mathbb{R}^{d_z}$, respectively. Assume that Z is a subvector of R and R is a subvector of W, so that $d_z \leq d_r \leq d_w$. Also, let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be a finite-dimensional parameter whose true value θ_0 is of interest, and let h be a vector-valued functional nuisance parameter taking values in a convex set of functions $\mathcal{H}$ mapping $\mathcal{Z}$ to $\mathbb{R}^{d_h}$, with the true value of h being h_0. The conditional moment restrictions framework assumes that θ_0 and h_0 satisfy the moment conditions

$$E_P[m(W; \theta_0, h_0(Z)) \mid R] = 0, \tag{2.22}$$

where $m : \mathcal{W} \times \Theta \times \mathbb{R}^{d_h} \to \mathbb{R}^{d_m}$ is a known vector-valued function. This framework is of interest because it covers a rich variety of models without having to explicitly rely on the likelihood formulation.

To build a Neyman orthogonal score $\psi(W; \theta, \eta)$ for estimating θ_0, consider the matrix-valued functional parameter $\mu : \mathcal{R} \to \mathbb{R}^{d_\theta \times d_m}$ whose true value is given by

$$\mu_0(R) = A(R)'\Omega(R)^{-1} - G(Z)\Gamma(R)'\Omega(R)^{-1}, \tag{2.23}$$

where the moment selection matrix-valued function $A : \mathcal{R} \to \mathbb{R}^{d_m \times d_\theta}$ and the weighting positive definite matrix-valued function $\Omega : \mathcal{R} \to \mathbb{R}^{d_m \times d_m}$ can be chosen arbitrarily, and the matrix-valued functions $\Gamma : \mathcal{R} \to \mathbb{R}^{d_m \times d_\theta}$ and $G : \mathcal{Z} \to \mathbb{R}^{d_\theta \times d_m}$ are given by

$$\Gamma(R) = \partial_{v'} E_P[m(W; \theta_0, v) \mid R]|_{v=h_0(Z)} \tag{2.24}$$

and

$$G(Z) = E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] \times (E_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R) \mid Z])^{-1}. \tag{2.25}$$

Note that μ_0 in (2.23) is well defined even though the right-hand side of (2.23) contains both R and Z as Z is a subvector of R. Then a Neyman orthogonal score is

$$\psi(W; \theta, \eta) = \mu(R)m(W; \theta, h(Z)), \tag{2.26}$$

where the nuisance parameter is

$$\eta = (\mu, h) \in T = \mathcal{L}^1(\mathcal{R};\ \mathbb{R}^{d_\theta \times d_m}) \times \mathcal{H}.$$

Here, $\mathcal{L}^1(\mathcal{R};\ \mathbb{R}^{d_\theta \times d_m})$ is the vector space of matrix-valued functions $f : \mathcal{R} \to \mathbb{R}^{d_\theta \times d_m}$ satisfying $E_P[\|f(R)\|] < \infty$. Also, note that even though the matrix-valued functions A and Ω can be chosen arbitrarily, setting

$$A(R) = \partial_{\theta'} E_P[m(W; \theta, h_0(Z)) \mid R]|_{\theta=\theta_0} \tag{2.27}$$

and

$$\Omega(R) = E_P[m(W; \theta_0, h_0(Z))m(W; \theta_0, h_0(Z))' \mid R] \tag{2.28}$$

leads to an asymptotic variance equal to the semi-parametric bound of Chamberlain (1992). Let $\eta_0 = (\mu_0, h_0)$ be the true value of the nuisance parameter $\eta = (\mu, h)$. The following lemma shows that the score ψ in (2.26) satisfies the Neyman orthogonality condition.

LEMMA 2.6. (NEYMAN ORTHOGONAL SCORES FOR CONDITIONAL MOMENT SETTINGS) *Suppose that (a)* (2.22) *holds, (b) the matrices* $E_P[\|\Gamma(R)\|^4]$, $E_P[\|G(Z)\|^4]$, $E_P[\|A(R)\|^2]$ *and* $E_P[\|\Omega(R)\|^{-2}]$ *are finite, and (c) for all* $h \in \mathcal{H}$, *there exists a constant* $C_h > 0$ *such that* $Pr_P(E_P[\|m(W; \theta_0, h(Z))\| \mid R] \leq C_h) = 1$. *Then the score* ψ *in* (2.26) *is Neyman orthogonal at* (θ_0, η_0) *with respect to the nuisance realization set* $\mathcal{T}_N = T$.

As an application of the conditional moment restrictions framework, let us derive Neyman orthogonal scores in the PLR example using this framework. The PLR model (1.1) is equivalent to

$$E_P[Y - D\theta_0 - g_0(X) \mid X, D] = 0,$$

which can be written in the form of the conditional moment restrictions framework (2.22) with $W = (Y, D, X')'$, $R = (D, X')'$, $Z = X$, $h(Z) = g(X)$ and $m(W; \theta, v) = Y - D\theta - v$. Hence, using (2.27) and (2.28) and denoting $\sigma(D, X)^2 = E_P[U^2 \mid D, X]$ for $U = Y - D\theta_0 - g_0(X)$, we can take

$$A(R) = -D, \quad \Omega(R) = E_P[U^2 \mid D, X] = \sigma(D, X)^2.$$

With this choice of $A(R)$ and $\Omega(R)$, we have

$$\Gamma(R) = -1, \quad G(Z) = \left(E_P\left[\frac{D}{\sigma(D, X)^2} \mid X\right]\right) \times \left(E_P\left[\frac{1}{\sigma(D, X)^2} \mid X\right]\right)^{-1},$$

and so (2.23) and (2.26) give

$$\psi(W;\theta,\eta_0) = \frac{1}{\sigma(D,X)^2}\left(D - \frac{E_P[(D/\sigma(D,X)^2) \mid X]}{E_P[(1/\sigma(D,X)^2) \mid X]}\right) \times (Y - D\theta - g_0(X)).$$

By construction, the score ψ above is efficient and Neyman orthogonal. Note, however, that using this score would require estimating the heteroscedasticity function $\sigma(D, X)^2$, which would require the imposition of some additional smoothness assumptions over this conditional variance function. Instead, if are willing to give up on efficiency to gain some robustness, we can take

$$A(R) = -D, \quad \Omega(R) = 1;$$

in which case we have

$$\Gamma(R) = -1, \quad G(Z) = E_P[D \mid X].$$

Equations (2.23) and (2.26) then give

$$\begin{aligned}\psi(W;\theta,\eta_0) &= (D - E_P[D \mid X]) \times (Y - D\theta - g_0(X)) \\ &= (D - m_0(X)) \times (Y - D\theta - g_0(X)).\end{aligned}$$

This score ψ is Neyman orthogonal and corresponds to the estimator of θ_0 described in the Introduction in (1.5). Note, however, that this score ψ is efficient only if $\sigma(X, D)$ is a constant.

2.2.5. Neyman orthogonal scores and influence functions. Neyman orthogonality is a joint property of the score $\psi(W;\theta,\eta)$, the true parameter value η_0, the parameter set T, and the distribution of W. It is not determined by any particular model for the parameter θ. Nevertheless, it is possible to use semi-parametric efficiency calculations to construct the orthogonal score from the original score as in Chernozhukov et al. (2016). Specifically, an orthogonal score can be constructed by adding to the original score the influence function adjustment for estimation of the nuisance functions that is analysed in Newey (1994). The resulting orthogonal score will be the influence function of the limit of the average of the original score.

To explain, consider the original score $\varphi(W;\theta,\beta)$, where β is some function, and let $\widehat{\beta}_0$ be a non-parametric estimator of β_0, the true value of β. Here, β is implicitly allowed to depend on θ, though we suppress that dependence for notational convenience. The corresponding orthogonal score can be formed when there is $\phi(W;\theta,\eta)$ such that

$$\int \varphi(w;\theta_0,\hat{\beta}_0)dP(w) = \frac{1}{n}\sum_{i=1}^{n}\phi(W_i;\theta_0,\eta_0) + o_P(n^{-1/2}), \qquad (*)$$

where η is a vector of nuisance functions that includes β, and $\phi(W;\theta,\eta)$ is an adjustment for the presence of the estimated function $\hat{\beta}_0$ in the original score $\varphi(W;\theta,\beta)$. The decomposition (*) typically holds when $\hat{\beta}$ is either a kernel or a series estimator with a suitably chosen tuning parameter. The Neyman orthogonal score is given by

$$\psi(W;\theta,\eta) = \varphi(W;\theta,\beta) + \phi(W;\theta,\eta). \qquad (2.29)$$

Here, $\psi(W;\theta_0,\eta_0)$ is the influence function of the limit of $n^{-1}\sum_{i=1}^{n}\varphi(W_i;\theta_0,\hat{\beta}_0)$, as analysed in Newey (1994), with the restriction $E_P[\psi(W;\theta_0,\eta_0)] = 0$ identifying θ_0.

The form of the adjustment term $\phi(W;\theta,\eta)$ depends on the estimator $\hat{\beta}_0$ and, of course, on the form of $\varphi(W;\theta,\beta)$. Such adjustment terms have been derived for various $\hat{\beta}_0$ by Newey (1994). Also, Ichimura and Newey (2015) show how the adjustment term can be computed from the limit of a certain derivative. Any of these results can be applied to a particular starting score $\varphi(W;\theta,\beta)$ and estimator $\hat{\beta}_0$ to obtain an orthogonal score.

For example, consider again the partially linear model with the original score

$$\varphi(W;\theta,\beta) = D(Y - D\theta - g_0(X)).$$

Here, $\hat{\beta}_0 = \hat{g}_0$ is a non-parametric regression estimator. From Newey (1994), we know that we obtain the influence function adjustment by taking the conditional expectation of the derivative of the score with respect to $g_0(x)$ (obtaining $-m_0(X) = -E_P[D|X]$) and multiplying the result by the non-parametric residual to obtain

$$\phi(W,\theta,\eta) = -m_0(X)\{Y - D\theta - \beta(X,\theta)\}.$$

The corresponding orthogonal score is then simply

$$\psi(W;\theta,\eta) = \{D - m_0(X)\}\{Y - D\theta - \beta(X,\theta)\},$$
$$\beta_0(X,\theta) = E_P[Y - D\theta|X], \quad m_0(X) = E_P[D|X],$$

illustrating that an orthogonal score for the partially linear model can be derived from an influence function adjustment.

Influence functions have been used to estimate functionals of non-parametric estimators by Hasminskii and Ibragimov (1979) and Bickel and Ritov (1988). Newey et al. (1998, 2004) showed that $n^{-1/2}\sum_{i=1}^{n}\psi(W_i;\theta_0,\hat{\eta}_0)$ from (2.29) will have a second-order remainder in $\hat{\eta}_0$, which is the key asymptotic property of orthogonal scores. Orthogonality of influence functions in semi-parametric models follows from van der Vaart (1991), as shown for higher-order counterparts in Robins et al. (2008, 2017). Chernozhukov et al. (2016) point out that, in general, an orthogonal score can be constructed from an original score and non-parametric estimator $\hat{\beta}_0$ by adding to the original score the adjustment term for estimation of β_0 as described above. This construction provides a way of obtaining an orthogonal score from any initial score $\varphi(W;\theta,\beta)$ and non-parametric estimator $\hat{\beta}_0$.

3. DML: POST-REGULARIZED INFERENCE BASED ON NEYMAN-ORTHOGONAL ESTIMATING EQUATIONS

3.1. Definition of DML and its basic properties

We assume that we have a sample $(W_i)_{i=1}^N$, modelled as independent and identically distributed (i.i.d.) copies of W, whose law is determined by the probability measure P on $\mathcal{W}$. Estimation will be carried out using the finite-sample analogue of the estimating equation (2.1).

We assume that the true value η_0 of the nuisance parameter η can be estimated by $\hat{\eta}_0$ using a part of the data $(W_i)_{i=1}^N$. Different structured assumptions on η_0 allow us to use different machine-learning tools for estimating η_0, for example:

(1) approximate sparsity for η_0 with respect to some dictionary calls for the use of forward selection, lasso, post-lasso, ℓ_2-boosting, or some other sparsity-based technique;

(2) well-approximability of η_0 by trees calls for the use of regression trees and random forests;
(3) well-approximability of η_0 by sparse neural and deep neural nets calls for the use of ℓ_1-penalized neural and deep neural networks;
(4) well-approximability of η_0 by at least one model mentioned in (1)–(3) above calls for the use of an ensemble/aggregated method over the estimation methods mentioned in (1)–(3).

There are performance guarantees for most of these ML methods that make it possible to satisfy the conditions stated below. Ensemble and aggregation methods ensure that the performance guarantee is approximately no worse than the performance of the best method.

We assume that N is divisible by K in order to simplify the notation. The following algorithm defines the simple cross-fitted DML as outlined in the Introduction.

DEFINITION 3.1. (DML1) *(a) Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_k^c := \{1, \ldots, N\} \setminus I_k$. (b) For each $k \in [K]$, construct an ML estimator*

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

of η_0, where $\hat{\eta}_{0,k}$ is a random element in T, and where randomness depends only on the subset of data indexed by I_k^c. (c) For each $k \in [K]$, construct the estimator $\check{\theta}_{0,k}$ as the solution of the following equation:

$$\mathbb{E}_{n,k}[\psi(W; \check{\theta}_{0,k}, \hat{\eta}_{0,k}] = 0, \tag{3.1}$$

where ψ is the Neyman orthogonal score, and $E_{n,k}$ is the empirical expectation over the kth fold of the data; that is, $E_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$. If achievement of exact 0 is not possible, define the estimator $\check{\theta}_{0,k}$ of θ_0 as an approximate ϵ_N-solution:

$$\|E_{n,k}[\psi(W; \check{\theta}_{0,k}, \hat{\eta}_{0,k})]\| \leq \inf_{\theta \in \Theta} \|E_{n,k}[\psi(W; \theta, \hat{\eta}_{0,k})]\| + \epsilon_N, \quad \epsilon_N = o(\delta_N N^{-1/2}), \tag{3.2}$$

where $(\delta_N)_{N \geq 1}$ is some sequence of positive constants converging to zero. (4) Aggregate the estimators:

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^K \check{\theta}_{0,k}. \tag{3.3}$$

This approach generalizes the 50–50 cross-fitting method mentioned in the Introduction. We now define a variation of this basic cross-fitting approach that may behave better in small samples.

DEFINITION 3.2. (DML2) *(a) Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_k^c := \{1, \ldots, N\} \setminus I_k$. (b) For each $k \in [K]$, construct an ML estimator*

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

of η_0, where $\hat{\eta}_{0,k}$ is a random element in T, and where randomness depends only on the subset of data indexed by I_k^c. (c) Construct the estimator $\tilde{\theta}_0$ as the solution to

$$\frac{1}{K} \sum_{k=1}^K E_{n,k}[\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})] = 0, \tag{3.4}$$

where ψ is the Neyman orthogonal score, and $E_{n,k}$ is the empirical expectation over the kth fold of the data; that is, $E_{n,k}[\psi(W)] = n^{-1}\sum_{i\in I_k}\psi(W_i)$. If achievement of exact 0 is not possible, define the estimator $\tilde{\theta}_0$ of θ_0 as an approximate ϵ_N-solution,

$$\left\| \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi(W;\tilde{\theta}_0,\hat{\eta}_{0,k})]] \right\| \leq \inf_{\theta\in\Theta}\left\| \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi(W;\theta_0,\hat{\eta}_{0,k})]] \right\| + \epsilon_N, \tag{3.5}$$

for $\epsilon_N = o(\delta_N N^{-1/2})$, where $(\delta_N)_{N\geq 1}$ is some sequence of positive constants converging to zero.

REMARK 3.1. (RECOMMENDATIONS) The choice of K has no asymptotic impact under our conditions but, of course, the choice of K may matter in small samples. Intuitively, larger values of K provide more observations in I_k^c from which to estimate the high-dimensional nuisance functions, which seems to be the more difficult part of the problem. We have found moderate values of K, such as 4 or 5, to work better than $K = 2$ in a variety of empirical examples and in simulations. Moreover, we generally recommend DML2 over DML1 though in some problems such as estimation of ATE in the interactive model, which we discuss later, there is no difference between the two approaches. In most other problems, DML2 is better behaved because the pooled empirical Jacobian for the equation in (3.4) exhibits more stable behaviour than the separate empirical Jacobians for the equation in (3.1).

3.2. Moment condition models with linear scores

We first consider the case of linear scores, where

$$\psi(w;\theta,\eta) = \psi^a(w;\eta)\theta + \psi^b(w;\eta), \quad \text{for all } w \in \mathcal{W}, \quad \theta \in \Theta,\ \eta \in T. \tag{3.6}$$

Let $c_0 > 0$, $c_1 > 0$, $s > 0$ and $q > 2$ be some finite constants such that $c_0 \leq c_1$, and let $\{\delta_N\}_{N\geq 1}$ and $\{\Delta_N\}_{N\geq 1}$ be some sequences of positive constants converging to zero such that $\delta_N \geq N^{-1/2}$. Also, let $K \geq 2$ be some fixed integer, and let $\{\mathcal{P}_N\}_{N\geq 1}$ be some sequence of sets of probability distributions P of W on $\mathcal{W}$.

ASSUMPTION 3.1 (LINEAR SCORES WITH APPROXIMATE NEYMAN ORTHOGONALITY) *For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold. (a) The true parameter value θ_0 obeys (2.1). (b) The score ψ is linear in the sense of (3.6). (c) The map $\eta \mapsto E_P[\psi(W;\theta,\eta)]$ is twice continuously Gateaux-differentiable on T. (d) The score ψ obeys the Neyman orthogonality or, more generally, the Neyman λ_N near-orthogonality condition at (θ_0,η_0) with respect to the nuisance realization set $\mathcal{T}_N \subset T$ for*

$$\lambda_N := \sup_{\eta\in\mathcal{T}_N}\left\| \partial_\eta E_P\psi(W;\theta_0,\eta_0)[\eta-\eta_0] \right\| \leq \delta_N N^{-1/2}.$$

(e) The identification condition holds; namely, the singular values of the matrix

$$J_0 := E_P[\psi^a(W;\eta_0)]$$

are between c_0 and c_1.

Assumption 3.1 requires scores to be Neyman orthogonal or near-orthogonal and imposes mild smoothness requirements as well as the canonical identification condition.

ASSUMPTION 3.2 (SCORE REGULARITY AND QUALITY OF NUISANCE PARAMETER ESTIMATORS) *For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold. (a) Given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^c})$ belongs to the realization set $\mathcal{T}_N$ with probability at least $1 - \Delta_N$, where $\mathcal{T}_N$ contains η_0 and is constrained by the next conditions. (b) The moment conditions hold:*

$$m_N := \sup_{\eta \in \mathcal{T}_N} (E_P[\|\psi(W; \theta_0, \eta)\|^q])^{1/q} \leq c_1;$$

$$m'_N := \sup_{\eta \in \mathcal{T}_N} (E_P[\|\psi^a(W; \eta)\|^q])^{1/q} \leq c_1.$$

(c) The following conditions on the statistical rates r_N, r'_N, and λ'_N hold:

$$r_N := \sup_{\eta \in \mathcal{T}_N} \|E_P[\psi^a(W; \eta)] - E_P[\psi^a(W; \eta_0)]\| \leq \delta_N,$$

$$r'_N := \sup_{\eta \in \mathcal{T}_N} (E_P[\|\psi(W; \theta_0, \eta) - \psi(W; \theta_0, \eta_0)\|^2])^{1/2} \leq \delta_N,$$

$$\lambda'_N := \sup_{r \in (0,1), \eta \in \mathcal{T}_N} \|\partial_r^2 E_P[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))]\| \leq \delta_N / \sqrt{N}.$$

(d) The variance of the score ψ is non-degenerate: All eigenvalues of the matrix

$$E_P[\psi(W; \theta_0, \eta_0)\psi(W; \theta_0, \eta_0)']$$

are bounded from below by c_0.

Assumptions 3.2(a)–(c) state that the estimator of the nuisance parameter belongs to the realization set $\mathcal{T}_N \subset T$, which is a shrinking neighbourhood of η_0, that contracts around η_0 with the rate determined by the 'statistical' rates r_N, r'_N and λ'_N. These rates are not given in terms of the norm $\|\cdot\|_T$ on T, but rather are the intrinsic rates that are most connected to the statistical problem at hand. However, in smooth problems, as discussed below this translates, in the worst cases, to the crude requirement that the nuisance parameters are estimated at the rate $o(N^{-1/4})$.

The conditions in Assumption 3.2 embody refined requirements on the quality of nuisance parameter estimators. In many applications, where $(\theta, \eta) \mapsto \psi(W; \theta, \eta)$ is smooth, we can bound

$$r_N \lesssim \varepsilon_N, \quad r'_N \lesssim \varepsilon_N, \quad \lambda'_N \lesssim \varepsilon_N^2, \tag{3.7}$$

where ε_N is the upper bound on the rate of convergence of $\hat{\eta}_0$ to η_0 with respect to the norm $\|\cdot\|_T = \|\cdot\|_{P,2}$:

$$\|\hat{\eta}_0 - \eta\|_T \lesssim \varepsilon_N.$$

Note that $\mathcal{T}_N$ can be chosen as the set of η that is within a neighbourhood of size ε_N of η_0, possibly with other restrictions, in this case. If only (3.7) holds, Assumption 3.2, particularly $\lambda'_N = o(N^{-1/2})$, imposes the (crude) rate requirement

$$\varepsilon_N = o(N^{-1/4}). \tag{3.8}$$

This rate is achievable for many ML methods under structured assumptions on the nuisance parameters. Among many others, see Bickel et al. (2009), Bühlmann and van de Geer (2011), Belloni et al. (2011, 2012) and Belloni and Chernozhukov (2011, 2013) for ℓ_1-penalized and related methods in a variety of sparse models, Kozbur (2016) for forward selection in sparse

models, Luo and Spindler (2016) for L_2-boosting in sparse linear models, Wager and Walther (2016) for concentration results for a class of regression trees and random forests, and Chen and White (1999) for a class of neural nets.

However, the presented conditions allow for more refined statements than (3.8). We note that many important structured problems – such as estimation of parameters in PLR models, estimation of parameters in partially linear structural equation models, and estimation of ATEs under unconfoundedness – are such that some cross-derivatives vanish, allowing more refined requirements than (3.8). This feature allows us to require much finer conditions on the quality of the nuisance parameter estimators than the crude bound (3.8). For example, in many problems

$$\lambda'_N = 0, \tag{3.9}$$

because the second derivatives vanish,

$$\partial_r^2 E_P[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))] = 0.$$

This occurs in the following important examples:

(1) the optimal instrument problem (see Belloni et al., 2012);
(2) the PLR model when $m_0(X) = 0$ or is otherwise known (see Section 4);
(3) the treatment effect examples when the propensity score is known, which includes randomized control trials as an important special case (see Section 5).

If both (3.7) and (3.9) hold, Assumption 3.2, particularly $r_N = o(1)$ and $r'_N = o(1)$, imposes the weakest possible rate requirement:

$$\varepsilon_N = o(1).$$

We note that similar refined rates have appeared in the context of estimation of treatment effects in high-dimensional settings under sparsity; see Farrell (2015) and Athey et al. (2016) and the related discussion in Remark 5.2. Our refined rate results complement this work by applying to a broad class of estimation contexts, including estimation of ATEs, and to a broad set of ML estimators.

THEOREM 3.1. (PROPERTIES OF THE DML) *Suppose that Assumptions 3.1 and 3.2 hold. In addition, suppose that $\delta_N \geq N^{-1/2}$ for all $N \geq 1$. Then the DML1 and DML2 estimators $\tilde{\theta}_0$ concentrate in a $1/\sqrt{N}$ neighbourhood of θ_0 and are approximately linear and centred Gaussian,*

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\psi}(W_i) + O_P(\rho_N) \rightsquigarrow N(0, \mathrm{I}_d), \tag{3.10}$$

uniformly over $P \in \mathcal{P}_N$, where the size of the remainder term obeys

$$\rho_N := N^{-1/2} + r_N + r'_N + N^{1/2}\lambda_N + N^{1/2}\lambda'_N \lesssim \delta_N. \tag{3.11}$$

Here, $\bar{\psi}(\cdot) := -\sigma^{-1} J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$ is the influence function, and the approximate variance is

$$\sigma^2 := J_0^{-1} E_P[\psi(W; \theta_0, \eta_0)\psi(W; \theta_0, \eta_0)'](J_0^{-1})'.$$

The result establishes that the estimator based on the orthogonal scores achieves the root-N rate of convergence and is approximately normally distributed. It is noteworthy that this convergence result, both the rate of concentration and the distributional approximation, holds

uniformly with respect to P varying over an expanding class of probability measures $\mathcal{P}_N$. This means that the convergence holds under any sequence of probability distributions $(P_N)_{N\geq 1}$ with $P_N \in \mathcal{P}_N$ for each N, which in turn implies that the results are robust with respect to perturbations of a given P along such sequences. The same property can be shown to fail for methods not based on orthogonal scores.

THEOREM 3.2. (VARIANCE ESTIMATOR FOR DML) *Suppose that Assumptions 3.1 and 3.2 hold. In addition, suppose that* $\delta_N \geq N^{-[(1-2/q)\wedge 1/2]}$ *for all* $N \geq 1$. *Consider the following estimator of the asymptotic variance matrix of* $\sqrt{N}(\tilde{\theta}_0 - \theta_0)$*:*

$$\hat{\sigma}^2 = \hat{J}_0^{-1} \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})'](\hat{J}_0^{-1})',$$

where

$$\hat{J}_0 = \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi^a(W; \hat{\eta}_{0,k})],$$

and $\tilde{\theta}_0$ *is either the DML1 or the DML2 estimator. This estimator concentrates around the true variance matrix* σ^2*,*

$$\hat{\sigma}^2 = \sigma^2 + O_P(\varrho_N), \quad \varrho_N := N^{-[(1-2/q)\wedge 1/2]} + r_N + r'_N \lesssim \delta_N.$$

Moreover, $\hat{\sigma}^2$ *can replace* σ^2 *in the statement of Theorem 3.1 with the size of the remainder term updated as* $\rho_N = N^{-[(1-2/q)\wedge 1/2]} + r_N + r'_N + N^{1/2}\lambda_N + N^{1/2}\lambda'_N$.

Theorems 3.1 and 3.2 can be used for standard construction of confidence regions, which are uniformly valid over a large, interesting class of models:

COROLLARY 3.1. (UNIFORMLY VALID CONFIDENCE BANDS) *Under the conditions of Theorem 3.2, suppose we are interested in the scalar parameter* $\ell'\theta_0$ *for some* $d_\theta \times 1$ *vector* ℓ. *Then the confidence interval*

$$\text{CI} := \left(\ell'\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\sqrt{\ell'\hat{\sigma}^2\ell/N}\right)$$

obeys

$$\sup_{P\in\mathcal{P}_N} |Pr_P(\ell'\theta_0 \in \text{CI}) - (1-\alpha)| \to 0.$$

Indeed, the above theorem implies that CI obeys $Pr_{P_N}(\ell'\theta_0 \in \text{CI}) \to (1-\alpha)$ under any sequence $\{P_N\} \in \mathcal{P}_N$, which implies that these claims hold uniformly in $P \in \mathcal{P}_N$. For example, one may choose $\{P_N\}$ such that, for some $\epsilon_N \to 0$

$$\sup_{P\in\mathcal{P}_N} |Pr_P(\ell'\theta_0 \in \text{CI}) - (1-\alpha)| \leq |Pr_{P_N}(\ell'\theta_0 \in \text{CI}) - (1-\alpha)| + \epsilon_N \to 0.$$

Next we note that the estimators need not be semi-parametrically efficient, but under some conditions they can be.

COROLLARY 3.2. (CASES WITH SEMI-PARAMETRIC EFFICIENCY) *Under the conditions of Theorem 3.1, if the score* ψ *is efficient for estimating* θ_0 *at a given* $P \in \mathcal{P} \subset \mathcal{P}_N$, *in the semi-parametric sense as defined in van der Vaart (1998), then the large sample variance* σ_0^2 *of* $\tilde{\theta}_0$ *reaches the semi-parametric efficiency bound at this* P *relative to the model* $\mathcal{P}$.

3.3. Models with non-linear scores

Let $c_0 > 0$, $c_1 > 0$, $a > 1$, $v > 0$, $s > 0$ and $q > 2$ be some finite constants, and let $\{\delta_N\}_{N\geq 1}$, $\{\Delta_N\}_{N\geq 1}$ and $\{\tau_N\}_{N\geq 1}$ be some sequences of positive constants converging to zero. To derive the properties of the DML estimator, we will use the following assumptions.

ASSUMPTION 3.3. (NON-LINEAR MOMENT CONDITION PROBLEM WITH APPROXIMATE NEYMAN ORTHOGONALITY) *For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold. (a) The true parameter value θ_0 obeys (2.1), and Θ contains a ball of radius $c_1 N^{-1/2} \log N$ centred at θ_0. (b) The map $(\theta, \eta) \mapsto E_P[\psi(W; \theta, \eta)]$ is twice continuously Gateaux-differentiable on $\Theta \times T$. (c) For all $\theta \in \Theta$, the identification relation*

$$2\|E_P[\psi(W; \theta, \eta_0)]\| \geq \|J_0(\theta - \theta_0)\| \wedge c_0$$

is satisfied, for the Jacobian matrix

$$J_0 := \partial_{\theta'}\{E_P[\psi(W; \theta, \eta_0)]\}|_{\theta=\theta_0}$$

having singular values between c_0 and c_1. (d) The score ψ obeys the Neyman orthogonality or, more generally, the Neyman near-orthogonality with $\lambda_N = \delta_N N^{-1/2}$ for the set $\mathcal{T}_N \subset T$.

Assumption 3.3 is mild and rather standard in moment condition problems. Assumption 3.3(a) requires θ_0 to be sufficiently separated from the boundary of Θ. Assumption 3.3(b) only requires differentiability of the function $(\theta, \eta) \mapsto E_P[\psi(W; \theta, \eta)]$ and does not require differentiability of the function $(\theta, \eta) \mapsto \psi(W; \theta, \eta)$. Assumption 3.3(c) implies sufficient identifiability of θ_0. Assumption 3.3(d) is the orthogonality condition that has already been extensively discussed.

ASSUMPTION 3.4. (SCORE REGULARITY AND REQUIREMENTS ON THE QUALITY OF ESTIMATION OF NUISANCE PARAMETERS) *Let K be a fixed integer. For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold. (a) Given a random subset I of $\{1, \ldots, N\}$ of size $n = N/K$, we have that the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i\in I^c})$ belongs to the realization set $\mathcal{T}_N$ with probability at least $1 - \Delta_N$, where $\mathcal{T}_N$ contains η_0 and is constrained by conditions given below. (b) The parameter space Θ is bounded and for each $\eta \in \mathcal{T}_N$, the function class $\mathcal{F}_{1,\eta} = \{\psi_j(\cdot, \theta, \eta) : j = 1, \ldots, d_\theta, \theta \in \Theta\}$ is suitably measurable and its uniform covering entropy obeys*

$$\sup_Q \log N(\epsilon \|F_{1,\eta}\|_{Q,2}, \mathcal{F}_{1,\eta}, \|\cdot\|_{Q,2}) \leq v \log(a/\epsilon), \quad \text{for all } 0 < \epsilon \leq 1, \tag{3.12}$$

where $F_{1,\eta}$ is a measurable envelope for $\mathcal{F}_{1,\eta}$ that satisfies $\|F_{1,\eta}\|_{P,q} \leq c_1$. (c) The following conditions on the statistical rates r_N, r'_N, and λ'_N hold:

$$r_N := \sup_{\eta\in\mathcal{T}_N, \theta\in\Theta} \|E_P[\psi(W; \theta, \eta) - E_P[\psi(W; \theta, \eta_0)]\| \leq \delta_N \tau_N,$$

$$r'_N := \sup_{\eta\in\mathcal{T}_N, \|\theta-\theta_0\|\leq\tau_N} (E_P[\|\psi(W; \theta, \eta) - \psi(W; \theta_0, \eta_0)\|^2])^{1/2} \text{ and } r'_N \log^{1/2}(1/r'_N) \leq \delta_N,$$

$$\lambda'_N := \sup_{r\in(0,1), \eta\in\mathcal{T}_N, \|\theta-\theta_0\|\leq\tau_N} \|\partial_r^2 E_P[\psi(W; \theta_0, \eta_0 + r(\theta - \theta_0) + r(\eta - \eta_0))]\| \leq \delta_N N^{-1/2}.$$

(d) The variance of the score is non-degenerate. All eigenvalues of the matrix

$$E_P[\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)']$$

are bounded from below by c_0.

Assumptions 3.3(a)–(c) state that the estimator of the nuisance parameter belongs to the realization set $\mathcal{T}_N \subset T$, which is a shrinking neighbourhood of η_0 that contracts at the statistical rates r_N, r'_N and λ'_N. These rates are not given in terms of the norm $\|\cdot\|_T$ on T, but rather are intrinsic rates that are connected to the statistical problem at hand. In smooth problems, these conditions translate to the crude requirement that nuisance parameters are estimated at the $o(N^{-1/4})$ rate, as discussed previously in the case with linear scores. However, these conditions can be refined as, for example, when $\lambda'_N = 0$ or when some cross-derivatives vanish in λ'_N; see the linear case in the previous subsection for further discussion. Suitable measurability and pointwise entropy conditions, required in Assumption 3.4(b), are mild regularity conditions that are satisfied in all practical cases. The assumption of a bounded parameter space Θ in Assumption 3.4(b) is embedded in the entropy condition, but we state it separately for clarity. This assumption was not needed in the linear case, and it can be removed in the non-linear case with the imposition of more complicated Huber-like regularity conditions. Assumption 3.4(c) is a set of mild growth conditions.

REMARK 3.2. (RATE REQUIREMENTS ON NUISANCE PARAMETER ESTIMATORS) Similar to the discussion in the linear case, the conditions in Assumption 3.4 are very flexible and embody refined requirements on the quality of the nuisance parameter estimators. The conditions essentially reduce to the previous conditions in the linear case, with the exception of compactness, which is imposed to make the conditions easy to verify in non-linear cases.

THEOREM 3.3. (PROPERTIES OF THE DML FOR NON-LINEAR SCORES) *Suppose that Assumptions 3.3 and 3.4 hold. In addition, suppose that* $\delta_N \geq N^{-1/2+1/q}\log N$ *and that* $N^{-1/2}\log N \leq \tau_N \leq \delta_N$ *for all* $N \geq 1$. *Then the DML1 and DML2 estimators* $\tilde{\theta}_0$ *concentrate in a* $1/\sqrt{N}$ *neighbourhood of* θ_0, *and are approximately linear and centred Gaussian,*

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{\psi}(W_i) + O_P(\rho_N) \rightsquigarrow N(0, \mathrm{I}),$$

uniformly over $P \in \mathcal{P}_N$, *where the size of the remainder term obeys*

$$\rho_N := N^{-1/2+1/q}\log N + r'_N \log^{1/2}(1/r'_N) + N^{1/2}\lambda_N + N^{1/2}\lambda'_N \lesssim \delta_N,$$

$\bar{\psi}(\cdot) := -\sigma_0^{-1}J_0^{-1}\psi(\cdot,\theta_0,\eta_0)$ *is the influence function, and the approximate variance is*

$$\sigma^2 := J_0^{-1}E_P[\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)'](J_0^{-1})'.$$

Moreover, in the statement above, σ^2 *can be replaced by a consistent estimator* $\hat{\sigma}^2$, *obeying* $\hat{\sigma}^2 = \sigma^2 + o_P(\varrho_N)$ *uniformly in* $P \in \mathcal{P}_N$, *with the size of the remainder term updated as* $\rho_N = \rho_N + \varrho_N$. *Furthermore, Corollaries 3.1 and 3.2 continue to hold under the conditions of this theorem.*

3.4. Finite-sample adjustments to incorporate uncertainty induced by sample splitting

The estimation technique developed in this paper relies on subsamples obtained by randomly partitioning the sample: an auxiliary sample for estimating the nuisance functions and a main

sample for estimating the parameter of interest. Although the specific sample partition has no impact on estimation results asymptotically, the effect of the particular random split on the estimate can be important in finite samples. To make the results more robust with respect to the partitioning, we propose to repeat the DML estimator S times, obtaining the estimates

$$\tilde{\theta}_0^s, \quad s = 1, \ldots, S.$$

Features of these estimates may then provide insight into the sensitivity of results to the sample splitting, and we can report results that incorporate features of this set of estimates that should be less driven by any particular sample-splitting realization.

DEFINITION 3.3. (INCORPORATING THE IMPACT OF SAMPLE SPLITTING USING MEAN AND MEDIAN METHODS) *For point estimation, we define*

$$\tilde{\theta}_0^{mean} = \frac{1}{S}\sum_{s=1}^{S} \tilde{\theta}_0^s \quad or \quad \tilde{\theta}_0^{median} = median\{\tilde{\theta}_0^s\}_{s=1}^S,$$

where the median operation is applied coordinate-wise. To quantify and incorporate the variation introduced by sample splitting, we consider variance estimators,

$$\hat{\sigma}^{2,mean} = \frac{1}{S}\sum_{s=1}^{S} \left(\hat{\sigma}_s^2 + (\hat{\theta}_s - \tilde{\theta}^{mean})(\hat{\theta}_s - \tilde{\theta}^{mean})'\right), \tag{3.13}$$

and a more robust version,

$$\hat{\sigma}^{2,median} = median\{\hat{\sigma}_s^2 + ((\hat{\theta}_s - \tilde{\theta}^{median})(\hat{\theta}_s - \tilde{\theta}^{median})'\}_{s=1}^S, \tag{3.14}$$

where the median picks out the matrix with median operator norm, which preserve non-negative definiteness.

We recommend using medians, reporting $\tilde{\theta}_0^{\text{median}}$ and $\hat{\sigma}^{2,median}$, as these quantities are more robust to outliers.

COROLLARY 3.3. *If S is fixed, as $N \to \infty$ and maintaining either Assumptions 3.1 and 3.2 or Assumptions 3.3 and 3.4 as appropriate, $\tilde{\theta}_0^{mean}$ and $\tilde{\theta}_0^{median}$ are first-order equivalent to $\tilde{\theta}_0$ and obey the conclusions of Theorems 3.1 and 3.2 or of Theorem 3.3. Moreover, $\hat{\sigma}^{2,median}$ and $\hat{\sigma}^{2,mean}$ can replace $\hat{\sigma}$ in the statement of the appropriate theorems.*

It would be interesting to investigate the behaviour under the regime where $S \to \infty$ as $N \to \infty$.

4. INFERENCE IN PARTIALLY LINEAR MODELS

4.1. Inference in partially linear regression models

Here we revisit the PLR model

$$Y = D\theta_0 + g_0(X) + U, \quad E_P[U \mid X, D] = 0, \tag{4.1}$$

$$D = m_0(X) + V, \quad E_P[V \mid X] = 0. \tag{4.2}$$

The parameter of interest is the regression coefficient θ_0. If D is conditionally exogenous (as good as randomly assigned conditional on covariates), then θ_0 measures the average causal/treatment effect of D on potential outcomes.

The first approach to inference on θ_0, which we described in the Introduction, is to employ the DML method using the score function

$$\psi(W;\theta,\eta) := \{Y - D\theta - g(X)\}(D - m(X)), \quad \eta = (g,m), \tag{4.3}$$

where $W = (Y, D, X)$, and g and m are P-square-integrable functions mapping the support of X to $\mathbb{R}$. It is easy to see that θ_0 satisfies the moment condition $E_P\psi(W;\theta_0,\eta_0) = 0$, and also the orthogonality condition $\partial_\eta E_P\psi(W;\theta_0,\eta_0)[\eta - \eta_0] = 0$ where $\eta_0 = (g_0, m_0)$.

A second approach employs the Robinson-style 'partialling-out' score function

$$\psi(W;\theta,\eta) := \{Y - \ell(X) - \theta(D - m(X))\}(D - m(X)), \quad \eta = (\ell,m), \tag{4.4}$$

where $W = (Y, D, X)$ and ℓ and m are P-square-integrable functions mapping the support of X to $\mathbb{R}$. This gives an alternative parametrization of the score function above, and using this score is first-order equivalent to using the previous score. It is easy to see that θ_0 satisfies the moment condition $E_P\psi(W;\theta_0,\eta_0) = 0$, and also the orthogonality condition $\partial_\eta E_P\psi(W;\theta_0,\eta_0)[\eta - \eta_0] = 0$, for $\eta_0 = (\ell_0, m_0)$, where $\ell_0(X) = E_P[Y|X]$.

In the partially linear model, the DML approach complements Belloni et al. (2013, 2014a,b, 2015), Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014b), all of which consider estimation and inference for parameters within the partially linear model using lasso-type methods without cross-fitting. By relying upon cross-fitting, the DML approach allows for the use of a much broader collection of ML methods for estimating the nuisance functions and also allows relaxation of sparsity conditions in the case where lasso or other sparsity-based estimators are used. Both the DML approach and the approaches taken in the aforementioned papers can be seen as heuristically debiasing the score function $(Y - D\theta - g(X))D$, which does not possess the orthogonality property unless $m_0(X) = 0$.

Let $(\delta_N)_{n=1}^\infty$ and $(\Delta_N)_{n=1}^\infty$ be sequences of positive constants approaching 0 as before. Also, let c, C and q be fixed strictly positive constants such that $q > 4$, and let $K \geq 2$ be a fixed integer. Moreover, for any $\eta = (\ell_1, \ell_2)$, where ℓ_1 is a function ℓ_1 and ℓ_2 are functions mapping the support of X to $\mathbb{R}$, denote $\|\eta\|_{P,q} = \|\ell_1\|_{P,q} \vee \|\ell_2\|_{P,q}$. For simplicity, assume that N/K is an integer.

ASSUMPTION 4.1. (REGULARITY CONDITIONS FOR PARTIALLY LINEAR REGRESSION MODEL) *Let $\mathcal{P}$ be the collection of probability laws P for the triple $W = (Y, D, X)$ such that (a) (4.1) and (4.2) hold; (b) $\|Y\|_{P,q} + \|D\|_{P,q} \leq C$; (c) $\|UV\|_{P,2} \geq c^2$ and $E_P[V^2] \geq c$; (d) $\|E_P[U^2 \mid X]\|_{P,\infty} \leq C$ and $\|E_P[V^2 \mid X]\|_{P,\infty} \leq C$; and (e) given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat\eta_0 = \hat\eta_0((W_i)_{i\in I^c})$ obeys the following conditions for all $n \geq 1$. With P-probability no less than $1 - \Delta_N$, $\|\hat\eta_0 - \eta_0\|_{P,q} \leq C$, $\|\hat\eta_0 - \eta_0\|_{P,2} \leq \delta_N$ and (i) for the score ψ in (4.3), where $\hat\eta_0 = (\hat g_0, \hat m_0)$, $\|\hat m_0 - m_0\|_{P,2} \times \|\hat g_0 - g_0\|_{P,2} \leq \delta_N N^{-1/2}$, and (ii) for the score ψ in (4.4), where $\hat\eta_0 = (\hat\ell_0, \hat m_0)$, $\|\hat m_0 - m_0\|_{P,2} \times (\|\hat m_0 - m_0\|_{P,2} + \|\hat\ell_0 - \ell_0\|_{P,2}) \leq \delta_N N^{-1/2}$.*

REMARK 4.1. (RATE CONDITIONS FOR ESTIMATORS OF NUISANCE PARAMETERS) The only non-primitive condition here is the assumption on the rate of estimating the nuisance parameters. These rates of convergence are available for most often used ML methods and are case-specific, so we do not restate conditions that are needed to reach these rates.

The following theorem follows as a corollary to the results in Section 3 by verifying Assumptions 3.1 and 3.2, and it will be proven as a special case of Theorem 4.2 below.

THEOREM 4.1. (DML INFERENCE ON REGRESSION COEFFICIENTS IN THE PARTIALLY LINEAR REGRESSION MODEL) *Suppose that Assumption 4.1 holds. Then the DML1 and DML2 estimators $\tilde{\theta}_0$ constructed in Definitions 3.1 and 3.2 using the score in either* (4.3) *or* (4.4) *are first-order equivalent and obey*

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, 1),$$

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = (E_P[V^2])^{-1}E_P[V^2U^2](E_P[V^2])^{-1}$. Moreover, the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2$ defined in Theorem 3.2. Consequently, confidence regions based upon the DML estimators $\tilde{\theta}_0$ have uniform asymptotic validity:

$$\lim_{N\to\infty} \sup_{P\in\mathcal{P}} |Pr_P(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}]) - (1-\alpha)| = 0.$$

REMARK 4.2. (ASYMPTOTIC EFFICIENCY UNDER HOMOSCEDASTICITY) Under conditional homoscedasticity, i.e. $E[U^2|Z] = E[U^2]$, the asymptotic variance σ^2 reduces to $E[V^2]^{-1}E[U^2]$, which is the semi-parametric efficiency bound for θ.

REMARK 4.3. (TIGHTNESS OF CONDITIONS UNDER CROSS-FITTING) The conditions in Theorem 4.1 are sharp, though they are simplified for ease of presentation. The sharpness can be understood by examining the case where the regression function g_0 and the propensity function m_0 are sparse with sparsity indices $s^g \ll N$ and $s^m \ll N$. They are estimated by ℓ_1-penalized estimators $\hat{g}_0$ and $\hat{m}_0$ that have sparsity indices of orders s^g and s^m and converge to g_0 and m_0 at the rates $\sqrt{s^g/N}$ and $\sqrt{s^m/N}$ (ignoring logs). The rate conditions in Assumption 4.1 then require (ignoring logs) that

$$\sqrt{s^g/N}\sqrt{s^m/N} \ll N^{-1/2} \Leftrightarrow s^g s^m \ll N,$$

which is much weaker than the condition

$$(s^g)^2 + (s^m)^2 \ll N$$

(ignoring logs) required without sample splitting. For example, if the propensity function m_0 is very sparse (low s_m), then the regression function is allowed to be quite dense (high s_g), and vice versa. If the propensity function is known ($s^m = 0$) or can be estimated at the $N^{-1/2}$ rate, then only consistency for $\hat{g}_0$ is needed. Such comparisons also extend to approximately sparse models.

4.2. Inference in partially linear IV models

Here we extend the PLR model studied in Section 4.1 to allow for IV identification. Specifically, we consider the model

$$Y = D\theta_0 + g_0(X) + U, \quad E_P[U \mid X, Z] = 0, \tag{4.5}$$

$$Z = m_0(X) + V, \quad E_P[V \mid X] = 0, \tag{4.6}$$

where Z is the instrumental variable. As before, the parameter of interest is θ and its true value is θ_0. If $Z = D$, the model (4.5)–(4.6) coincides with (4.1)–(4.2) but is otherwise different.

To estimate θ_0 and to perform inference on it, we use the score

$$\psi(W;\theta,\eta) := (Y - D\theta - g(X))(Z - m(X)), \quad \eta = (g,m), \tag{4.7}$$

where $W = (Y, D, X, Z)$ and g and m are P-square-integrable functions mapping the support of X to $\mathbb{R}$. Alternatively, we can use the Robinson-style score

$$\psi(W;\theta,\eta) := (Y - \ell(X) - \theta(D - r(X)))(Z - m(X)), \quad \eta = (\ell,m,r), \tag{4.8}$$

where $W = (Y, D, X, Z)$ and ℓ, m and r are P-square-integrable functions mapping the support of X to $\mathbb{R}$. It is straightforward to verify that both scores satisfy the moment condition $E_P[\psi(W;\theta_0,\eta_0)] = 0$ and also the orthogonality condition $\partial_\eta E_P[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0$, for $\eta_0 = (g_0, m_0)$ in the former case and $\eta_0 = (\ell_0, m_0, r_0)$ for ℓ_0 and r_0 defined by $\ell_0(X) = E_P[Y \mid X]$ and $r_0(X) = E_P[D \mid X]$, respectively, in the latter case.[8]

Note that the score in (4.8) has a minor advantage over the score in (4.7) because all of its nuisance parameters are conditional mean functions, which can be directly estimated by the ML methods. If one prefers to use the score in (4.7), one has to construct an estimator of g_0 first. To do so, one can first obtain a DML estimator of θ_0 based on the score in (4.8), say $\tilde{\theta}_0$. Then, using the fact that $g_0(X) = E_P[Y - D\theta_0 \mid X]$, one can construct an estimator $\hat{g}_0$ by applying an ML method to regress $Y - D\tilde{\theta}_0$ on X. Alternatively, one can use assumption-specific methods to directly estimate g_0, without using the score (4.8) first. For example, if g_0 can be approximated by a sparse linear combination of a large set of transformations of X, one can use the methods of Gautier and Tsybakov (2014) to obtain an estimator of g_0.

Let $(\delta_N)_{n=1}^\infty$ and $(\Delta_N)_{n=1}^\infty$ be sequences of positive constants approaching 0 as before. Also, let c, C and q be fixed strictly positive constants such that $q > 4$, and let $K \geq 2$ be a fixed integer. Moreover, for any $\eta = (\ell_j)_{j=1}^l$ mapping the support of X to $\mathbb{R}^l$, denote $\|\eta\|_{P,q} = \max_{1\leq j\leq l} \|\ell_j\|_{P,q}$. For simplicity, assume that N/K is an integer.

ASSUMPTION 4.2. (REGULARITY CONDITIONS FOR PARTIALLY LINEAR IV MODEL) *For all probability laws $P \in \mathcal{P}$ for the quadruple $W = (Y, D, X, Z)$ the following conditions hold: (a) equations (4.5) and (4.6) hold; (b) $\|Y\|_{P,q} + \|D\|_{P,q} + \|Z\|_{P,q} \leq C$; (c) $\|UV\|_{P,2} \geq c^2$ and $|E_P[DV]| \geq c$; (d) $\|E_P[U^2 \mid X]\|_{P,\infty} \leq C$ and $\|E_P[V^2 \mid X]\|_{P,\infty} \leq C$; and (e) given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i\in I^c})$ obeys the following conditions. With P-probability no less than $1 - \Delta_N$, $\|\hat{\eta}_0 - \eta_0\|_{P,q} \leq C$, $\|\hat{\eta}_0 - \eta_0\|_{P,2} \leq \delta_N$ and (i) for the score ψ in (4.7), where $\hat{\eta}_0 = (\hat{g}_0, \hat{m}_0)$, $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{g}_0 - g_0\|_{P,2} \leq \delta_N N^{-1/2}$, and (ii) for the score ψ in (4.8), where $\hat{\eta}_0 = (\hat{\ell}_0, \hat{m}_0, \hat{r}_0)$, $\|\hat{m}_0 - m_0\|_{P,2} \times \left(\|\hat{r}_0 - r_0\|_{P,2} + \|\hat{\ell}_0 - \ell_0\|_{P,2}\right) \leq \delta_N N^{-1/2}$.*

The following theorem follows as a corollary to the results in Section 3 by verifying Assumptions 3.1 and 3.2.

THEOREM 4.2. (DML INFERENCE ON REGRESSION COEFFICIENTS IN THE PARTIALLY LINEAR IV MODEL) *Suppose that Assumption 4.2 holds. Then the DML1 and DML2 estimators*

[8] It is interesting to note that the methods for constructing Neyman orthogonal scores described in Section 2 may give scores that are different from those in (4.7) and (4.8). For example, applying the method for conditional moment restriction problems in Section 2.2.4 with $\Omega(R) = 1$ gives the score $\psi(W;\theta,\eta) = (Y - D\theta - g(X))(r(Z,X) - f(X))$, where the true values of $r(Z,X)$ and $f(X)$ are $r_0(Z,X) = E_P[D \mid Z,X]$ and $f_0(X) = E_P[D \mid X]$, respectively. It may be interesting to compare properties of the DML estimators $\tilde{\theta}_0$ based on this score with those based on (4.7) and (4.8) in future work.

$\tilde{\theta}_0$ constructed in Definitions 3.1 and 3.2 using the score in either (4.7) *or* (4.8) *are first-order equivalent and obey $\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0,1)$ uniformly over $P \in \mathcal{P}$, where $\sigma^2 = (E_P[DV])^{-1}E_P[V^2U^2](E_P[DV])^{-1}$. Moreover, the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2$ defined in Theorem 3.2. Consequently, confidence regions based upon the DML estimators $\tilde{\theta}_0$ have uniform asymptotic validity:*

$$\lim_{N\to\infty} \sup_{P\in\mathcal{P}} |Pr_P(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}]) - (1-\alpha)| = 0.$$

5. INFERENCE ON TREATMENT EFFECTS IN THE INTERACTIVE MODEL

5.1. Inference on ATE and ATTE

In this section, we specialize the results of Section 3 to estimate treatment effects under the unconfoundedness assumption of Rosenbaum and Rubin (1983). Within this setting, there is a large classical literature focused on low-dimensional settings that provides methods for adjusting for confounding variables including regression methods, propensity score adjustment methods, matching methods, and 'doubly-robust' combinations of these methods; see, e.g. Robins and Rotnitzky (1995), Hahn (1998), Hirano et al. (2003) and Abadie and Imbens (2006) as well as the textbook overview provided in Imbens and Rubin (2015). In this section, we present results that complement this important classic work as well as the rapidly expanding body of work on estimation under unconfoundedness using ML methods; see, among others, Athey et al. (2016), Belloni et al. (2014a, 2017), Farrell (2015) and Imai and Ratkovic (2013).

We specifically consider estimation of ATEs when treatment effects are fully heterogeneous and the treatment variable is binary, $D \in \{0, 1\}$. We consider vectors (Y, D, X) such that

$$Y = g_0(D, X) + U, \quad E_P[U \mid X, D] = 0, \tag{5.1}$$

$$D = m_0(X) + V, \quad E_P[V \mid X] = 0. \tag{5.2}$$

Because D is not additively separable, this model is more general than the partially linear model for the case of binary D. A common target parameter of interest in this model is the ATE:[9]

$$\theta_0 = E_P[g_0(1, X) - g_0(0, X)].$$

Another common target parameter is the ATTE:

$$\theta_0 = E_P[g_0(1, X) - g_0(0, X)|D = 1].$$

The confounding factors X affect the policy variable via the propensity score $m_0(X)$ and the outcome variable via the function $g_0(D, X)$. Both of these functions are unknown and potentially complicated, and we can employ ML methods to learn them.

[9] Without unconfoundedness/conditional exogeneity, these quantities measure association, and could be referred to as average predictive effect (APE) and average predictive effect for the exposed (APEX). Inferential results for these objects would follow immediately from Theorem 5.1.

We proceed to set up moment conditions with scores obeying orthogonality conditions. For estimation of the ATE, we employ

$$\psi(W;\theta,\eta) := (g(1,X) - g(0,X)) + \frac{D(Y - g(1,X))}{m(X)} - \frac{(1-D)(Y - g(0,X))}{1 - m(X)} - \theta, \quad (5.3)$$

where the nuisance parameter $\eta = (g, m)$ consists of P-square-integrable functions g and m mapping the support of (D, X) to $\mathbb{R}$ and the support of X to $(\varepsilon, 1-\varepsilon)$, respectively, for some $\varepsilon \in (0, 1/2)$. The true value of η is $\eta_0 = (g_0, m_0)$. This orthogonal moment condition is based on the influence function for the mean for missing data from Robins and Rotnitzky (1995).

For estimation of the ATTE, we use the score

$$\psi(W;\theta,\eta) = \frac{D(Y - \overline{g}(X))}{p} - \frac{m(X)(1-D)(Y - \overline{g}(X))}{p(1 - m(X))} - \frac{D\theta}{p}, \quad (5.4)$$

where the nuisance parameter $\eta = (\overline{g}, m, p)$ consists of P-square-integrable functions $\overline{g}$ and m mapping the support of X to $\mathbb{R}$ and to $(\varepsilon, 1-\varepsilon)$, respectively, and a constant $p \in (\varepsilon, 1-\varepsilon)$, for some $\varepsilon \in (0, 1/2)$. The true value of η is $\eta_0 = (\overline{g}_0, m_0, p_0)$, where $\overline{g}_0(X) = g_0(0, X)$ and $p_0 = E_P[D]$. Note that estimating ATTE does not require estimating $g_0(1, X)$. Note also that because p is a constant, it does not affect the DML estimators $\tilde{\theta}_0$ based on the score ψ in (5.4), but having p simplifies the formula for the variance of $\tilde{\theta}_0$.

Using their respective scores, it can be easily seen that true parameter values θ_0 for ATE and ATTE obey the moment condition $E_P[\psi(W;\theta_0,\eta_0)] = 0$, and also that the orthogonality condition $\partial_\eta E_P[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0$ holds.

Let $(\delta_N)_{n=1}^\infty$ and $(\Delta_N)_{n=1}^\infty$ be sequences of positive constants approaching 0. Also, let c, ε, C and q be fixed strictly positive constants such that $q > 2$, and let $K \geq 2$ be a fixed integer. Moreover, for any $\eta = (\ell_1, \ldots, \ell_l)$, denote $\|\eta\|_{P,q} = \max_{1 \leq j \leq l} \|\ell_j\|_{P,q}$. For simplicity, assume that N/K is an integer.

ASSUMPTION 5.1. (REGULARITY CONDITIONS FOR ATE AND ATTE ESTIMATION) *For all probability laws $P \in \mathcal{P}$ for the triple (Y, D, X) the following conditions hold: (a) equations (5.1) and (5.2) hold, with $D \in \{0, 1\}$; (b) $\|Y\|_{P,q} \leq C$; (c) $Pr_P(\varepsilon \leq m_0(X) \leq 1 - \varepsilon) = 1$; (d) $\|U\|_{P,2} \geq c$; (e) $\|E_P[U^2 \mid X]\|_{P,\infty} \leq C$; and (f) given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^c})$ obeys the following conditions. With P-probability no less than $1 - \Delta_N$, $\|\hat{\eta}_0 - \eta_0\|_{P,q} \leq C$, $\|\hat{\eta}_0 - \eta_0\|_{P,2} \leq \delta_N$, $\|\hat{m}_0 - 1/2\|_{P,\infty} \leq 1/2 - \varepsilon$ and (i) for the score ψ in (5.3), where $\hat{\eta}_0 = (\hat{g}_0, \hat{m}_0)$ and the target parameter is ATE, $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{g}_0 - g_0\|_{P,2} \leq \delta_N N^{-1/2}$, and (ii) for the score ψ in (5.4), where $\hat{\eta}_0 = (\hat{\overline{g}}_0, \hat{m}_0, \hat{p}_0)$ and the target parameter is ATTE, $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{\overline{g}}_0 - \overline{g}_0\|_{P,2} \leq \delta_N N^{-1/2}$.*

REMARK 5.1. The only non-primitive condition here is the assumption on the rate of estimating the nuisance parameters. These rates of convergence are available for the most often used ML methods and are case-specific, so we do not restate conditions that are needed to reach these rates. The conditions are not the tightest possible, but offer a set of simple conditions under which Theorem 5.1 follows as a special case of the general theorem provided in Section 3. We could obtain more refined conditions by doing customized proofs.

The following theorem follows as a corollary to the results in Section 3 by verifying Assumptions 3.1 and 3.2.

THEOREM 5.1. (DML INFERENCE ON ATE AND ATTE) *Suppose that the target parameter is either ATE, $\theta_0 = E_P[g_0(1, X) - g_0(0, X)]$, and the score ψ in* (5.3) *is used, or ATTE, $\theta_0 = E_P[g_0(1, X) - g_0(0, X) \mid D = 1]$, and the score ψ in* (5.4) *is used. In addition, suppose that Assumption 5.1 holds. Then the DML1 and DML2 estimators $\tilde{\theta}_0$, constructed in Definitions 3.1 and 3.2, are first-order equivalent and obey*

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, 1), \tag{5.5}$$

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = E_P[\psi^2(W; \theta_0, \eta_0)]$. Moreover, the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2$ defined in Theorem 3.2. Consequently, confidence regions based upon the DML estimators $\tilde{\theta}_0$ have uniform asymptotic validity:

$$\lim_{N\to\infty} \sup_{P\in\mathcal{P}} |Pr_P(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1 - \alpha/2)\hat{\sigma}/\sqrt{N}]) - (1 - \alpha)| = 0.$$

The scores ψ in (5.3) *and* (5.4) *are efficient, so both estimators are asymptotically efficient, reaching the semi-parametric efficiency bound of Hahn (1998).*

REMARK 5.2. (TIGHTNESS OF CONDITIONS) The conditions in Assumption 5.1 are sharp though simplified for ease of presentation. The sharpness can be understood by examining the case where the regression function g_0 and the propensity function m_0 are sparse with sparsity indices $s^g \ll N$ and $s^m \ll N$. They are estimated by ℓ_1-penalized estimators $\hat{g}_0$ and $\hat{m}_0$ that have sparsity indices of orders s^g and s^m and converge to g_0 and m_0 at the rates $\sqrt{s^g/N}$ and $\sqrt{s^m/N}$ (ignoring logs). Then the rate conditions in Assumption 5.1 require

$$\sqrt{s^g/N}\sqrt{s^m/N} \ll N^{-1/2} \Leftrightarrow s^g s^m \ll N$$

(ignoring logs), which is much weaker than the condition $(s^g)^2 + (s^m)^2 \ll N$ (ignoring logs) required without sample splitting. For example, if the propensity score m_0 is very sparse, then the regression function is allowed to be quite dense with $s^g > \sqrt{N}$, and vice versa. If the propensity score is known ($s^m = 0$), then only consistency for $\hat{g}_0$ is needed. Such comparisons also extend to approximately sparse models. We note that similar refined rates appeared in Farrell (2015), who considers estimation of treatment effects in a setting where an approximately sparse model holds for both the regression and propensity score functions. In interesting related work, Athey et al. (2016) show that $\sqrt{N}$-consistent estimation of an ATE is possible under very weak conditions on the propensity score – allowing for the possibility that the propensity score may not be consistently estimated – under strong sparsity of the regression function such that $s_g \ll \sqrt{N}$. Thus, the approach taken in this context and by Athey et al. (2016) are complementary, and either might be preferred, depending on whether or not the regression function can be estimated extremely well based on a sparse method.

5.2. Inference on local average treatment effects

In this section, we consider estimation of LATEs with a binary treatment variable, $D \in \{0, 1\}$, and a binary instrument, $Z \in \{0, 1\}$.[10] As before, Y denotes the outcome variable, and X is the vector of covariates.

[10] Similar results can be provided for the local average treatment effect on the treated (LATT) by adapting the following arguments to use the orthogonal scores for the LATT; see, e.g. Belloni et al. (2017).

Consider the functions μ_0, m_0 and p_0, where μ_0 maps the support of (Z, X) to $\mathbb{R}$, and m_0 and p_0, respectively, map the support of (Z, X) and X to $(\varepsilon, 1-\varepsilon)$ for some $\varepsilon \in (0, 1/2)$, such that

$$Y = \mu_0(Z, X) + U, \quad E_P[U \mid Z, X] = 0, \tag{5.6}$$

$$D = m_0(Z, X) + V, \quad E_P[V \mid Z, X] = 0, \tag{5.7}$$

$$Z = p_0(X) + \zeta, \quad E_P[\zeta \mid X] = 0. \tag{5.8}$$

We are interested in estimating

$$\theta_0 = \frac{E_P[\mu(1, X)] - E_P[\mu(0, X)]}{E_P[m(1, X)] - E_P[m(0, X)]}.$$

Under the assumptions of Imbens and Angrist (1994) and Frölich (2007), θ_0 is the LATE – the average treatment effect for compliers that are observations that would have $D = 1$ if Z were 1 and $D = 0$ if Z were 0. To estimate θ_0, we use the score

$$\psi(W; \theta, \eta) := \mu(1, X) - \mu(0, X) + \frac{Z(Y - \mu(1, X))}{p(X)} - \frac{(1 - Z)(Y - \mu(1, X))}{1 - p(X))}$$
$$- \left(m(1, X) - m(0, X) + \frac{Z(D - m(1, X))}{p(X)} - \frac{(1 - Z)(D - m(0, X))}{1 - p(X)} \right) \times \theta,$$

where $W = (Y, D, X, Z)$ and the nuisance parameter $\eta = (\mu, m, p)$ consists of P-square-integrable functions μ, m and p, with μ mapping the support of (Z, X) to $\mathbb{R}$ and m and p, respectively, mapping the support of (Z, X) and X to $(\varepsilon, 1 - \varepsilon)$ for some $\varepsilon \in (0, 1/2)$. It is easy to verify that this score satisfies the moment condition $E_P[\psi(W; \theta_0, \eta_0)] = 0$ and also the orthogonality condition $\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0$ for $\eta_0 = (\mu_0, m_0, p_0)$.

Let $(\delta_N)_{n=1}^{\infty}$ and $(\Delta_N)_{n=1}^{\infty}$ be sequences of positive constants approaching 0. Also, let c, C and q be fixed strictly positive constants such that $q > 4$, and let $K \geq 2$ be a fixed integer. Moreover, for any $\eta = (\ell_1, \ell_2, \ell_3)$, where ℓ_1 is a function mapping the support of (Z, X) to $\mathbb{R}$ and ℓ_2 and ℓ_3 are functions respectively mapping the support of (Z, X) and X to $(\varepsilon, 1 - \varepsilon)$ for some $\varepsilon \in (0, 1/2)$, we denote $\|\eta\|_{P,q} = \|\ell_1\|_{P,q} \vee \|\ell_2\|_{P,2} \vee \|\ell_3\|_{P,q}$. For simplicity, assume that N/K is an integer.

ASSUMPTION 5.2. (REGULARITY CONDITIONS FOR LATE ESTIMATION) *For all probability laws $P \in \mathcal{P}$ for the quadruple $W = (Y, D, X, Z)$ the following conditions hold: (a) equations (5.6)–(5.8) hold, with $D \in \{0, 1\}$ and $Z \in \{0, 1\}$; (b) $\|Y\|_{P,q} \leq C$; (c) $Pr_P(\varepsilon \leq p_0(X) \leq 1 - \varepsilon) = 1$; (d) $E_P[m_0(1, X) - m_0(0, X)] \geq c$, (e) $\|U - \theta_0 V\|_{P,2} \geq c$; (f) $\|E_P[U^2 \mid X]\|_{P,\infty} \leq C$; and (g) given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^c})$ obeys the following conditions. With P-probability no less than $1 - \Delta_N$, $\|\hat{\eta}_0 - \eta_0\|_{P,q} \leq C$, $\|\hat{\eta}_0 - \eta_0\|_{P,2} \leq \delta_N$, $\|\hat{p}_0 - 1/2\|_{P,\infty} \leq 1/2 - \varepsilon$ and $\|\hat{p}_0 - p_0\|_{P,2} \times (\|\hat{\mu}_0 - \mu_0\|_{P,2} + \|\hat{m}_0 - m_0\|_{P,2}) \leq \delta_N N^{-1/2}$.*

The following theorem follows as a corollary to the results in Section 3 by verifying Assumptions 3.1 and 3.2.

THEOREM 5.2. (DML INFERENCE ON LATE) *Suppose that Assumption 5.2 holds. Then the DML1 and DML2 estimators $\tilde{\theta}_0$ constructed in Definitions 3.1 and 3.2 and based on the score ψ above are first-order equivalent and obey*

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, 1), \tag{5.9}$$

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = (E_P[m(1, X) - m(0, X)])^{-2} E_P[\psi^2(W; \theta_0, \eta_0)]$. Moreover, the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2$ defined in Theorem 3.2. Consequently, confidence regions based upon the DML estimators $\tilde{\theta}_0$ have uniform asymptotic validity:

$$\lim_{N\to\infty} \sup_{P\in\mathcal{P}} |Pr_P(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1 - \alpha/2)\hat{\sigma}/\sqrt{N}]) - (1 - \alpha)| = 0.$$

6. EMPIRICAL EXAMPLES

To illustrate the methods developed in the preceding sections, we consider three empirical examples. The first example reexamines the Pennsylvania Reemployment Bonus experiment, which used a randomized control trial to investigate the incentive effect of unemployment insurance. In the second, we use the DML method to estimate the effect of 401(k) eligibility, the treatment variable, and 401(k) participation, a self-selected decision to receive the treatment that we instrument for with assignment to the treatment state, on accumulated assets. In this example, the treatment variable is not randomly assigned and we aim to eliminate the potential biases due to the lack of random assignment by flexibly controlling for a rich set of variables. In the third, we revisit the IV estimation by Acemoglu et al. (2001) of the effects of institutions on economic growth by estimating a partially linear IV model.

6.1. Effect of unemployment insurance bonus on unemployment duration

In this example, we reanalyse the Pennsylvania Reemployment Bonus experiment, which was conducted by the US Department of Labor in the 1980s to test the incentive effects of alternative compensation schemes for unemployment insurance (UI). This experiment has been previously studied by Bilias (2000) and Bilias and Koenker (2002). In these experiments, UI claimants were randomly assigned either to a control group or to one of five treatment groups.[11] In the control group, the standard rules of the UI system applied. Individuals in the treatment groups were offered a cash bonus if they found a job within some pre-specified period of time (qualification period), provided that the job was retained for a specified duration. The treatments differed in the level of the bonus, the length of the qualification period, and whether the bonus was declining over time in the qualification period; see Bilias and Koenker (2002) for further details.

In our empirical example, we focus only on the most generous compensation scheme, treatment 4, and we drop all individuals who received other treatments. In this treatment, the bonus amount is high and the qualification period is long compared to other treatments, and claimants are eligible to enroll in a workshop. Our treatment variable, D, is an indicator variable for being assigned treatment 4, and the outcome variable, Y, is the log of duration of unemployment for the UI claimants. The vector of covariates, X, consists of age group dummies,

[11] There are six treatment groups in the experiments. Following Bilias (2000). we merge groups 4 and 6.

gender, race, the number of dependents, quarter of the experiment, location within the state, existence of recall expectations and type of occupation.

We report results based on five simple methods for estimating the nuisance functions used in forming the orthogonal estimating equations. We consider three tree-based methods, labelled 'Random forest', 'Reg. tree' and 'Boosting', one ℓ_1-penalization based method, labelled 'Lasso' and a neural network method, labelled 'Neural network'. For Reg. tree, we fit a single CART tree to estimate each nuisance function with penalty parameter chosen by tenfold cross-validation. The results in the Random forest column are obtained by estimating each nuisance function with a random forest that averages over 1000 trees. The results in Boosting are obtained using boosted regression trees with regularization parameters chosen by tenfold cross-validation. To estimate the nuisance functions using the neural networks, we use two neurons and a decay parameter of 0.02, and we set activation function as logistic for classification problems and as linear for regression problems.[12] Lasso estimates an ℓ_1-penalized linear regression model using the data-driven penalty parameter selection rule developed in Belloni et al. (2012). For Lasso, we use a set of 96 potential control variables formed from the raw set of covariates and all second-order terms (i.e. all squares and first-order interactions). For the remaining methods, we use the raw set of covariates as features.

We also consider two hybrid methods labelled 'Ensemble' and 'Best'. Ensemble optimally combines four of the ML methods listed above by estimating the nuisance functions as weighted averages of estimates from Lasso, Boosting, Random forest and Neural network. The weights are restricted to sum to one and are chosen so that the weighted average of these methods gives the lowest average mean squared out-of-sample prediction error estimated using fivefold cross-validation. The final column in Table 1 (Best) reports results that combine the methods in a different way. After obtaining estimates from the five simple methods and Ensemble, we select the best methods for estimating each nuisance function based on the average out-of-sample prediction performance for the target variable associated with each nuisance function obtained from each of the previously described approaches. As a result, the reported estimate in the last column uses different ML methods to estimate different nuisance functions. Note that if a single method outperformed all the others in terms of prediction accuracy for all nuisance functions, the estimate in the Best column would be identical to the estimate reported under that method.

Table 1 presents DML2 estimates of the ATE on unemployment duration using the median method described in Section 3.4. We report results for the heterogeneous effect model in Panel A and for the partially linear model in Panel B. Because the treatment is randomly assigned, we use the fraction of treated as the estimator of the propensity score in forming the orthogonal estimating equations.[13] For both the partially linear model and the interactive model, we report estimates obtained using twofold cross-fitting and fivefold cross-fitting. All results are based on taking 100 different sample splits. We summarize results across the sample splits using the median method. For comparison, we report two different standard errors: in brackets, we report the median standard errors from across the 100 splits; in parentheses, we report standard errors adjusted for variability across the sample splits using the median method in parentheses.

[12] We also experimented with Deep Learning methods from which we obtained similar results for some tuning parameters. However, we ran into stability and computational issues and we have chosen not to report these results in the empirical section.

[13] We also estimated the effects using non-parametric estimates of the conditional propensity score obtained from the ML procedures given in the column labels. As expected due to randomization, the results are similar to those provided in Table 1 and are not reported for brevity.

Table 1. Estimated effect of cash bonus on unemployment duration.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
Panel A: interactive regression model							
ATE	−0.081	−0.084	−0.074	−0.079	−0.073	−0.079	−0.078
(twofold)	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
ATE	−0.081	−0.085	−0.074	−0.077	−0.073	−0.078	−0.077
(fivefold)	[0.036]	[0.036]	[0.036]	[0.035]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Panel B: partially linear regression model							
ATE	−0.080	−0.084	−0.077	−0.076	−0.074	−0.075	−0.075
(twofold)	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)
ATE	−0.080	−0.084	−0.077	−0.074	−0.073	−0.075	−0.074
(fivefold)	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)

Note: Estimated ATE and standard errors from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard errors across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

The estimation results are consistent with the findings of previous studies that have analysed the Pennsylvania Bonus Experiment. The ATE on unemployment duration is negative and significant across all estimation methods at the 5% level regardless of the standard error estimator used. Interestingly, we see that there is no practical difference across the two different standard errors in this example.

6.2. Effect of 401(k) eligibility and participation on net financial assets

The key problem in determining the effect of 401(k) eligibility is that working for a firm that offers access to a 401(k) plan is not randomly assigned. To overcome the lack of random assignment, we follow the strategy developed in Poterba et al. (1994a,b). In these papers, the authors use data from the 1991 Survey of Income and Program Participation and they argue that eligibility for enrolling in a 401(k) plan in these data can be taken as exogenous after conditioning on a few observables – of which the most important for their argument is income. The basic idea of their argument is that, at least around the time 401(k) initially became available, people were unlikely to be basing their employment decisions on whether an employer offered a 401(k) but would instead focus on income and other aspects of the job. Following this argument, whether one is eligible for a 401(k) may then be taken as exogenous after appropriately conditioning on income and other control variables related to job choice.

A key component of the argument underlying the exogeneity of 401(k) eligibility is that eligibility can only be taken as exogenous after conditioning on income and other variables related to job choice that might correlate with whether a firm offers a 401(k). Poterba et al. (1994a,b) and many subsequent papers adopt this argument but control only linearly for a small

Table 2. Estimated effect of 401(k) eligibility on net financial assets.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
Panel A: interactive regression model							
ATE	6830	7713	7770	7806	7764	7702	7546
(twofold)	[1282]	[1208]	[1276]	[1159]	[1328]	[1149]	[1360]
	(1530)	(1271)	(1363)	(1202)	(1468)	(1170)	(1533)
ATE	7170	7993	8105	7713	7788	7839	7753
(fivefold)	[1201]	[1198]	[1242]	[1155]	[1238]	[1134]	[1237]
	(1398)	(1236)	(1299)	(1177)	(1293)	(1148)	(1294)
Panel B: partially linear regression model							
ATE	7717	8709	9116	8759	8950	9010	9125
(twofold)	[1346]	[1363]	[1302]	[1339]	[1335]	[1309]	[1304]
	(1749)	(1427)	(1377)	(1382)	(1408)	(1344)	(1357)
ATE	8187	8871	9247	9110	9038	9166	9215
(fivefold)	[1298]	[1358]	[1295]	[1314]	[1322]	[1299]	[1294]
	(1558)	(1418)	(1328)	(1328)	(1355)	(1310)	(1312)

Note: Estimated ATE and standard errors from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard errors across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

number of terms. One might wonder whether such specifications are able to adequately control for income and other related confounds. At the same time, the power to learn about treatment effects decreases as one allows more flexible models. The principled use of flexible ML tools offers one resolution to this tension. The results presented below thus complement previous results that rely on the assumption that confounding effects can adequately be controlled for by a small number of variables chosen *ex ante* by the researcher.

In the example in this paper, we use the same data as in Chernozhukov and Hansen (2004). We use net financial assets – defined as the sum of IRA balances, 401(k) balances, checking accounts, US saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks, and mutual funds less non-mortgage debt – as the outcome variable, Y, in our analysis. Our treatment variable, D, is an indicator for being eligible to enroll in a 401(k) plan. The vector of raw covariates, X, consists of age, income, family size, years of education, a married indicator, a two-earner status indicator, a defined benefit pension status indicator, an IRA participation indicator, and a home-ownership indicator.

In Table 2, we report DML2 estimates of ATE of 401(k) eligibility on net financial assets both in the partially linear model as in (1.1) and allowing for heterogeneous treatment effects using the interactive model outlined in Section 5.1. To reduce the disproportionate impact of extreme propensity score weights in the interactive model, we trim the propensity scores at 0.01 and 0.99. We present two sets of results based on sample splitting as discussed in Section 3 using twofold cross-fitting and fivefold cross-fitting. As in the previous section, we consider 100 different sample partitions and summarize the results across different sample splits using the median method. For comparison, we report two different standard errors: in brackets, we report the median standard errors from across the 100 splits; in parentheses, we report standard errors

Table 3. Estimated effect of 401(k) participation on net financial assets.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
LATE	8978	11073	11384	11329	11094	11119	10952
(twofold)	[2192]	[1749]	[1832]	[1666]	[1903]	[1653]	[1657]
	(3014)	(1849)	(1993)	(1718)	(2098)	(1689)	(1699)
LATE	8944	11459	11764	11133	11186	11173	11113
(fivefold)	[2259]	[1717]	[1788]	[1661]	[1795]	[1641]	[1645]
	(3307)	(1786)	(1893)	(1710)	(1890)	(1678)	(1675)

Note: Estimated LATE based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard errors across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

adjusted for variability across the sample splits using the median method. We consider the same methods with the same tuning choices for estimating the nuisance functions as in the previous example, with one exception, and so we do not repeat details for brevity. The one exception is that we implement neural networks with eight neurons and a decay parameter of 0.01 in this example.

Turning to the results, it is first worth noting that the estimated ATE of 401(k) eligibility on net financial assets is $19,559 with an estimated standard error of 1413 when no control variables are used. Of course, this number is not a valid estimate of the causal effect of 401(k) eligibility on financial assets if there are neglected confounding variables as suggested by Poterba et al. (1994a,b). When we turn to the estimates that flexibly account for confounding reported in Table 2, we see that they are substantially attenuated relative to this baseline that does not account for confounding, suggesting much smaller causal effects of 401(k) eligibility on financial asset holdings. It is interesting and reassuring that the results obtained from the different flexible methods are broadly consistent with each other. This similarity is consistent with the theory that suggests that results obtained through the use of orthogonal estimating equations and any sensible method of estimating the necessary nuisance functions should be similar. Finally, it is interesting that these results are also broadly consistent with those reported in the original work of Poterba et al. (1994a,b), who used a simple intuitively motivated functional form, suggesting that this intuitive choice was sufficiently flexible to capture much of the confounding variation in this example.

As a further illustration, we also report the LATE in this example where we take the endogenous treatment variable to be participating in a 401(k) plan. Even after controlling for features related to job choice, it seems likely that the actual choice of whether to participate in an offered plan would be endogenous. Of course, we can use eligibility for a 401(k) plan as an instrument for participation in a 401(k) plan under the conditions that were used to justify the exogeneity of eligibility for a 401(k) plan provided above in the discussion of estimation of the ATE of 401(k) eligibility.

We report DML2 results of estimating the LATE of 401(k) participation using 401(k) eligibility as an instrument in Table 3. We employ the procedure outlined in Section 5.2 using the same ML estimators to estimate the quantities used to form the orthogonal estimating equation as we employed to estimate the ATE of 401(k) eligibility outlined previously, so we omit the details for brevity. Looking at the results, we see that the estimated causal effect of 401(k) participation on net financial assets is uniformly positive and statistically significant across all of

the considered methods. As when looking at the ATE of 401(k) eligibility, it is reassuring that the results obtained from the different flexible methods are broadly consistent with each other. It is also interesting that the results based on flexible ML methods are broadly consistent with, though somewhat attenuated relative to, those obtained by applying the same specification for controls as used in Poterba et al. (1994a,b) and using a linear IV model, which returns an estimated effect of participation of $13,102 with estimated standard error of (1922). The mild attenuation may suggest that the simple intuitive control specification used in the original baseline specification is too simplistic.

Looking at Tables 2 and 3, there are other interesting observations that can provide useful insights into understanding the finite sample properties of the DML estimation method. First, the standard errors of the estimates obtained using fivefold cross-fitting are lower than those obtained from twofold cross-fitting for all methods across all cases. This fact suggests that having more observations in the auxiliary sample may be desirable. Specifically, the fivefold cross-fitting estimates use more observations to learn the nuisance functions than twofold cross-fitting and thus are likely learn them more precisely. This increase in precision in learning the nuisance functions may then translate into more precisely estimated parameters of interest. While intuitive, we note that this statement does not seem to be generalizable, in that there does not appear to be a general relationship between the number of folds in cross-fitting and the precision of the estimate of the parameter of interest (see the next example). Secondly, we also see that the standard errors of the Lasso estimates after adjusting for variation due to sample splitting are noticeably larger than the standard errors coming from the other ML methods. We believe that this is due to the fact that the out-of-sample prediction errors from a linear model tend to be larger when there is a need to extrapolate. In our framework, if the main sample includes observations that are outside of the range of the observations in the auxiliary sample, the model has to extrapolate to those observations. The fact that the standard errors are lower in fivefold cross-fitting than in twofold cross-fitting for the Lasso estimations also supports this hypothesis, because the higher number of observations in the auxiliary sample reduces the degree of extrapolation. We also see that there is a noticeable increase in the standard errors that account for variability due to sample splitting relative to the simple unadjusted standard errors in this case, though these differences do not qualitatively change the results.

6.3. Effect of institutions on economic growth

To demonstrate DML estimation of partially linear structural equation models with instrumental variables, we consider estimation of the effect of institutions on aggregate output following the work of Acemoglu et al. (2001, hereafter AJR). Estimating the effect of institutions on output is complicated by the clear potential for simultaneity between institutions and output. Specifically, better institutions may lead to higher incomes, but higher incomes may also lead to the development of better institutions. To help overcome this simultaneity, AJR use mortality rates for early European settlers as an instrument for institution quality. The validity of this instrument hinges on the arguments that settlers set up better institutions in places where they are more likely to establish long-term settlements, that where they are likely to settle for the long term is related to settler mortality at the time of initial colonization, and that institutions are highly persistent. The exclusion restriction for the instrumental variable is then motivated by the argument that GDP, while persistent, is unlikely to be strongly influenced by mortality in the previous century, or earlier, except through institutions.

Table 4. Estimated effect of institutions on output.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
Twofold	0.85	0.81	0.84	0.77	0.94	0.80	0.83
	[0.28]	[0.42]	[0.38]	[0.33]	[0.32]	[0.35]	[0.34]
	(0.22)	(0.29)	(0.30)	(0.27)	(0.28)	(0.30)	(0.29)
Fivefold	0.77	0.95	0.90	0.73	1.00	0.83	0.88
	[0.24]	[0.46]	[0.41]	[0.33]	[0.33]	[0.37]	[0.41]
	(0.17)	(0.45)	(0.40)	(0.27)	(0.30)	(0.34)	(0.39)

Note: Estimated coefficient from a linear IV model based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard errors across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

In their paper, AJR note that their IV strategy will be invalidated if other factors are also highly persistent and related to the development of institutions within a country and to the country's GDP. A leading candidate for such a factor, as they discuss, is geography. AJR address this by assuming that the confounding effect of geography is adequately captured by a linear term in distance from the equator and a set of continent dummy variables. Using DML allows us to relax this assumption and to replace it by a weaker assumption that geography can be sufficiently controlled by an unknown function of distance from the equator and continent dummies, which can be learned by ML methods.

We use the same set of 64 country-level observations as AJR. The data set contains measurements of GDP, settler morality, an index that measures protection against expropriation risk and geographic information. The outcome variable, Y, is the logarithm of GDP per capita and the endogenous explanatory variable, D, is a measure of the strength of individual property rights that is used as a proxy for the strength of institutions. To deal with endogeneity, we use an instrumental variable Z, which is mortality rates for early European settlers. Our raw set of control variables, X, include distance from the equator and dummy variables for Africa, Asia, North America and South America.

We report results from applying DML2 following the procedure outlined in Section 4.2 in Table 4. The considered ML methods and tuning parameters are the same as the previous examples except for the Ensemble method, from which we exclude Neural network, as the small sample size causes stability problems in training the neural network. We use the raw set of covariates and all second-order terms when doing Lasso estimation, and we simply use the raw set of covariates in the remaining methods. As in the previous examples, we consider 100 different sample splits and report the median estimates of the coefficient and two different standard error estimates. In brackets, we report the median standard errors from across the 100 splits, and we report standard errors adjusted for variability across the sample splits using the median method in parentheses. Finally, we report results from both twofold cross-fitting and fivefold cross-fitting as in the other examples.

In this example, we see uniformly large and positive point estimates across all procedures considered, and estimated effects are statistically significant at the 5% level. As in the second example, we see that adjusting for variability across sample splits leads to noticeable increases in estimated standard errors but does not result in qualitatively different conclusions. Interestingly, we see that the estimated standard errors based on fivefold cross-fitting are larger than those

based on twofold cross-fitting in all procedures except lasso, which differs from the finding in the 401(k) example. Further understanding these differences and the impact of the number of folds on inference for objects of interest seems to be an interesting question for future research. Finally, although the estimated coefficients are somewhat smaller than the baseline estimates reported in AJR – an estimated coefficient of 1.10 with estimated standard error of 0.46 (see AJR, Table 4, Panel A, column 7) – the results are qualitatively similar, indicating a strong and positive effect of institutions on output.

6.4. Comments on empirical results

Before closing this section, we want to emphasize some important conclusions that can be drawn from these empirical examples. First, the choice of the ML method used in estimating nuisance functions does not substantively change the conclusion in any of the examples, and we have obtained broadly consistent results regardless of which method we employ. The robustness of the results to the different methods is implied by the theory assuming that all of the employed methods are able to deliver sufficiently high-quality approximations to the underlying nuisance functions. Secondly, the incorporation of uncertainty due to sample splitting using the median method increases the standard errors relative to a baseline that does not account for this uncertainty, though these differences do not alter the main results in any of the examples. This lack of variation suggests that the parameter estimates are robust to the particular sample split used in the estimation in these examples.

ACKNOWLEDGEMENTS

We would like to acknowledge research support from the National Science Foundation. We also thank participants of the MIT Stochastics and Statistics seminar, the Kansas Econometrics conference, the Royal Economic Society Annual Conference, the Hannan Lecture at the Australasian Econometric Society meeting, the Econometric Theory lecture at the EC^2 meetings 2016 in Toulouse, the CORE 50th Anniversary Conference, the Becker–Friedman Institute Conference on Machine Learning and Economics, the INET conferences on Big Data at the University of Southern California in Los Angeles, the World Congress of Probability and Statistics 2016, the Joint Statistical Meetings 2016, the New England Day of Statistics Conference, CEMMAP's Masterclass on Causal Machine Learning, and St Gallen's summer school on Big Data, for many useful comments and questions. We would like to thank Susan Athey, Peter Aronow, Jin Hahn, Guido Imbens, Mark van der Laan and Matt Taddy for constructive comments. We thank Peter Aronow for pointing us to the literature on targeted learning on which we build, along with prior works of Neyman, Bickel, and the many other contributions to semi-parametric learning theory.

REFERENCES

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*, 235–67.

Acemoglu, D., S. Johnson and J. A. Robinson (2001). The colonial origins of comparative development: an empirical investigation. *American Economic Review 91* (5), 1369–401 (AJR).

Ai, C. and X. Chen (2012). The semi-parametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics 170*, 442–57.

Andrews, D. W. K. (1994a). Asymptotics for semi-parametric econometric models via stochastic equicontinuity. *Econometrica 62*, 43–72.

Andrews, D. W. K. (1994b). Empirical process methods in econometrics. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, *Volume IV*, Chapter 37, 2247–94. Amsterdam: Elsevier.

Angrist, J. D. and A. B. Krueger (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and Economic Statistics 13*, 225–35.

Athey, S., G. Imbens, and S. Wager (2016). Approximate residual balancing: de-biased inference of average treatment effects in high-dimensions. Preprint (arXiv:1604.07125v3).

Ayyagari, R. (2010). Applications of influence functions to semi-parametric regression models. PhD Thesis, Harvard School of Public Health, Harvard University.

Belloni, A. and V. Chernozhukov (2011). ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics 39*, 82–130.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*, 521–47.

Belloni, A., V. Chernozhukov and C. Hansen (2010). Lasso methods for Gaussian instrumental variables models. Preprint (arXiv:1012.1297).

Belloni, A., V. Chernozhukov and L. Wang (2011). Square-root-lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*, 791–806.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*, 2369–429.

Belloni, A., V. Chernozhukov and C. Hansen (2013). Inference for high-dimensional sparse econometric models. In D. Acemoglu, M. Arellano and E. Dekel (Eds.), *Advances in Economics and Econometrics: Tenth World Congress of Econometric Society*, *Volume III*, 245–95. Cambridge: Cambridge University Press.

Belloni, A., V. Chernozhukov and C. Hansen (2014a). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies 81*, 608–50.

Belloni, A., V. Chernozhukov and L. Wang (2014b). Pivotal estimation via square-root Lasso in nonparametric regression. *Annals of Statistics 42*, 757–88.

Belloni, A., V. Chernozhukov and K. Kato (2015). Uniform post selection inference for LAD regression models and other z-estimators. *Biometrika 102*, 77–94.

Belloni, A., V. Chernozhukov and Y. Wei (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business and Economic Statistics 34*, 606–19.

Belloni, A., V. Chernozhukov, I. Fernández-Val and C. Hansen (2017). Program evaluation with high-dimensional data. *Econometrica 85*, 233–98.

Bera, A., G. Montes-Rojas and W. Sosa-Escudero (2010). General specification testing with locally misspecified models. *Econometric Theory 26*, 1838–45.

Bickel, P. J. (1982). On adaptive estimation. *Annals of Statistics 10*, 647–71.

Bickel, P. and Y. Ritov (1988). Estimating integrated squared density derivatives. *Sankhya A-50*, 381–93.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov and J. A. Wellner (1998). *Efficient and Adaptive Estimation for Semi-Parametric Models*. Berlin: Springer.

Bickel, P. J., Y. Ritov and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics 37*, 1705–32.

Bilias, Y. (2000). Sequential testing of duration data: the case of the Pennsylvania 'reemployment bonus' experiment. *Journal of Applied Econometrics 15*, 575–94.

Bilias, Y. and R. Koenker (2002). Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. In B. Fitzenberger, R. Koenker and J. A. Machado (Eds.), *Studies in Empirical Economics: Economic Applications of Quantile Regression*, 199–220. Heidelberg: Physica-Verlag.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data*, Springer Series in Statistics. Berlin: Springer.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*, 305–34.

Chamberlain, G. (1992). Efficiency bounds for semi-parametric regression. *Econometrica 60*, 567–96.

Chen, X. and H. White (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory 45*, 682–91.

Chen, X., O. Linton and I. van Keilegom (2003). Estimation of semi-parametric models when the criterion function is not smooth. *Econometrica 71*, 1591–608.

Chernozhukov, V. and C. Hansen (2004). The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and Statistics 86*, 735–51.

Chernozhukov, V., D. Chetverikov and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics 42*, 1564–97.

Chernozhukov, V., J. Escanciano, H. Ichimura, W. Newey and J. Robins (2016). Locally robust semi-parametric estimation. Preprint (arXiv:1608.00033).

Chernozhukov, V., C. Hansen and M. Spindler (2015a). Post-selection and post-regularization inference in linear models with very many controls and instruments. *Americal Economic Review: Papers and Proceedings 105*, 486–90.

Chernozhukov, V., C. Hansen and M. Spindler (2015b). Valid post-selection and post-regularization inference: an elementary, general approach. *Annual Review of Economics 7*, 649–88.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer Texts in Statistics. Berlin: Springer.

Fan, J., S. Guo and K. Yu (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society*, Series B *74*, 37–65.

Farrell, M. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 174*, 1–23.

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York, NY: Academic Press.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics 139*, 35–75.

Gautier, E. and A. Tsybakov (2014). High-dimensional instrumental variables regression and confidence sets. Preprint (arXiv:1105.2454v4).

Hahn, J. (1998). On the role of the propensity score in efficient semi-parametric estimation of average treatment effects. *Econometrica 66*, 315–31.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–54.

Hasminskii, R. and I. Ibragimov (1979). On the nonparametric estimation of functionals. In P. Mandl and M. Hušková (Eds.), *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 41–51. Amsterdam: North-Holland.

Hirano, K., G. W. Imbens and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–89.

Hubbard, A. E., S. Kherad-Pajouh and M. J. van der Laan (2016). Statistical inference for data adaptive target parameters. *International Journal of Biostatistics 12*, 3–19.

Ibragimov, I. A. and R. Z. Hasminskii (1981). *Statistical Estimation: Asymptotic Theory*. New York, NY: Springer-Verlag.

Ichimura, H. and W. Newey (2015). The influence function of semi-parametric estimators. Preprint (arXiv:1508.01378).

Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics 7*, 443–70.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–475.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*, 2869–909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory 60*, 6522–54.

Kozbur, D. (2016). Testing-based forward model selection. Preprint (arXiv:1512.02666).

Lee, L. (2005). A $c(\alpha)$-type gradient test in the GMM approach. Working paper, Ohio State University.

Levit, B. Y. (1975). On the efficiency of a class of nonparametric estimates. *Theory of Probability and Its Applications 20*, 723–40.

Linton, O. (1996). Edgeworth approximation for MINPIN estimators in semi-parametric regression models. *Econometric Theory 12*, 30–60.

Luedtke, A. R. and M. J. van der Laan (2016). Optimal individualized treatments in resource-limited settings. *International Journal of Biostatistics 12*, 283–303.

Luo, Y. and M. Spindler (2016). High-dimensional l_2 boosting: rate of convergence. Preprint (arXiv:1602.08927).

Nevelson, M. (1977). On one informational lower bound. *Problemy Peredachi Informatsii 13*, 26–31.

Newey, W. (1990). Semi-parametric efficiency bounds. *Journal of Applied Econometrics 5*, 99–135.

Newey, W. (1994). The asymptotic variance of semi-parametric estimators. *Econometrica 62*, 1349–82.

Newey, W. K., F. Hsieh and J. Robins (1998). Undersmoothing and bias corrected functional estimation. Working paper, MIT Economics Department (http://economics.mit.edu/files/11219).

Newey, W. K., F. Hsieh and J. M. Robins (2004). Twicing kernels and a small bias property of semi-parametric estimators. *Econometrica 72*, 947–62.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander (Ed.), *Probability and Statistics*, 416–44. New York, NY: Wiley.

Neyman, J. (1979). $c(\alpha)$ tests and their use. *Sankhya*, 1–21.

Poterba, J. M., S. F. Venti and D. A. Wise (1994a). 401(k) plans and tax-deferred savings. In D. Wise (Ed.), *Studies in the Economics of Aging*, 105–42. Chicago, IL: University of Chicago Press.

Poterba, J. M., S. F. Venti, and D. A. Wise (1994b). Do 401(k) contributions crowd out other personal saving? *Journal of Public Economics 58*, 1–32.

Robins, J. and A. Rotnitzky (1995). Semi-parametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–29.

Robins, J., L. Li, E. Tchetgen and A. van der Vaart (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In D. Nolan and T. Speed (Eds.), *Probability and Statistics: Essays in Honor of David A. Freedman*, 335–421. Beachwood, OH: Institute of Mathematical Statistics.

Robins, J., P. Zhang, R. Ayyagari, R. Logan, E. Tchetgen, L. Li, A. Lumley and A. van der Vaart (2013). New statistical approaches to semi-parametric regression with application to air pollution research. Research Report 175, Health Effects Institute.

Robins, J., L. Li, R. Mukherjee, E. Tchetgen and A. van der Vaart (2017). Minimax estimation of a functional on a structured high dimensional model. Forthcoming in *Annals of Statistics*.

Robinson, P. M. (1988). Root-N-consistent semi-parametric regression. *Econometrica 56*, 931–54.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Scharfstein, D. O., A. Rotnitzky and J. M. Robins (1999). Rejoinder to "adjusting for non-ignorable drop-out using semi-parametric non-response models". *Journal of the American Statistical Association 94*, 1135–46.

Schick, A. (1986). On asymptotically efficient estimation in semi-parametric models. *Annals of Statistics 14*, 1139–51.

Severini, T. A. and W. H. Wong (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics 20*, 1768–802.

Toth, B. and M. J. van der Laan (2016). TMLE for marginal structural models based on an instrument. Working Paper 350, UC Berkeley Division of Biostatistics Working Paper Series.

van de, Geer, S., P. Bühlmann, Y. Ritov and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics 42*, 1166–202.

van der Laan, M. J. (2015). A generally efficient targeted minimum loss based estimator. Working Paper 343, UC Berkeley Division of Biostatistics Working Paper Series.

van der Laan, M. J. and S. Rose (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Berlin: Springer.

van der Laan, M. and D. Rubin (2006). Targeted maximum likelihood learning. Working Paper 213, UC Berkeley Division of Biostatistics Working Paper Series.

van der Laan, M. J., E. C. Polley and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*.

van der Vaart, A. W. (1991). On differentiable functionals. *Annals of Statistics 19*, 178–204.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wager, S. and G. Walther (2016). Adaptive concentration of regression trees, with application to random forests. Preprint (arXiv:1503.06388).

Wooldridge, J. (1991). Specification testing and quasi-maximum-likelihood estimation. *Journal of Econometrics 48*, 29–55.

Zhang, C. and S. Zhang (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society*, Series B *76*, 217–42.

Zheng, W. and M. J. van der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In M. J. van der Laan and S. Rose (Eds.), *Targeted Learning*, 459–74. Berlin: Springer.

Zheng, W., Z. Luo, and M. J. van der Laan (2016). Marginal structural models with counterfactual effect modifiers. Working Paper 348, UC Berkeley Division of Biostatistics Working Paper Series.

APPENDIX: PROOFS OF RESULTS

In this appendix, we use C to denote a strictly positive constant that is independent of n and $P \in \mathcal{P}_N$. The value of C may change at each appearance. Also, the notation $a_N \lesssim b_N$ means that $a_N \leq Cb_N$ for all n and some C. The notation $a_N \gtrsim b_N$ means that $b_N \lesssim a_N$. Moreover, the notation $a_N = o(1)$ means that there exists a sequence $(b_N)_{n\geq 1}$ of positive numbers such that (a) $|a_N| \leq b_N$ for all n, (b) b_N is independent of $P \in \mathcal{P}_N$ for all n and (c) $b_N \to 0$ as $n \to \infty$. Finally, the notation $a_N = O_P(b_N)$ means that for all $\epsilon > 0$, there exists C such that $Pr_P(a_N > Cb_N) \leq 1 - \epsilon$ for all n. Using this notation allows us to avoid repeating 'uniformly over $P \in \mathcal{P}_N$' many times in the proofs.

Define the empirical process $\mathbb{G}_n(\psi(W))$ as a linear operator acting on measurable functions $\psi : \mathcal{W} \to \mathbb{R}$ such that $\|\psi\|_{P,2} < \infty$ via

$$\mathbb{G}_n(\psi(W)) := \mathbb{G}_{n,I}(\psi(W)) := \frac{1}{\sqrt{n}} \sum_{i \in I} \psi(W_i) - \int \psi(w) dP(w).$$

Analogously, we defined the empirical expectation as

$$E_n[\psi(W)] := E_{n,I}[\psi(W)] := \frac{1}{n} \sum_{i \in I} \psi(W_i).$$

The following lemma is useful particularly in the sample-splitting contexts.

LEMMA 6.1. (CONDITIONAL CONVERGENCE IMPLIES UNCONDITIONAL) *Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random vectors. (a) If, for $\epsilon_m \to 0$, $Pr(\|X_m\| > \epsilon_m \mid Y_m) \to_{Pr} 0$, then $Pr(\|X_m\| > \epsilon_m) \to 0$. In particular, this occurs if $E[\|X_m\|^q/\epsilon_m^q \mid Y_m] \to_{Pr} 0$ for some $q \geq 1$, by Markov's inequality. (b) Let $\{A_m\}$ be a sequence of positive constants. If $\|X_m\| = O_P(A_m)$ conditional on Y_m, namely, that for any $\ell_m \to \infty$, $Pr(\|X_m\| > \ell_m A_m \mid Y_m) \to_{Pr} 0$, then $\|X_m\| = O_P(A_m)$ unconditionally, namely, that for any $\ell_m \to \infty$, $Pr(\|X_m\| > \ell_m A_m) \to 0$.*

Proof: (a) For any $\epsilon > 0$ $Pr(\|X_m\| > \epsilon_m) \leq E[Pr(\|X_m\| > \epsilon_m \mid Y_m)] \to 0$, as the sequence $\{Pr(\|X_m\| > \epsilon_m \mid Y_m)\}$ is uniformly integrable. To show the second part note that $Pr(\|X_m\| > \epsilon_m \mid Y_m) \leq E[\|X_m\|^q/\epsilon_m^q \mid Y_m] \vee 1 \xrightarrow{p} 0$ by Markov's inequality. (b) This follows from (a). □

Let $(W_i)_{i=1}^n$ be a sequence of independent copies of a random element W taking values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ according to a probability law P. Let $\mathcal{F}$ be a set of suitably measurable functions $f : \mathcal{W} \to \mathbb{R}$, equipped with a measurable envelope $F : \mathcal{W} \to \mathbb{R}$.

LEMMA 6.2. (MAXIMAL INEQUALITY, CHERNOZHUKOV ET AL. (2014)) *Work with the set-up above. Suppose that $F \geq \sup_{f \in \mathcal{F}} |f|$ is a measurable envelope for $\mathcal{F}$ with $\|F\|_{P,q} < \infty$ for some $q \geq 2$. Let $M = \max_{i \leq n} F(W_i)$ and $\sigma^2 > 0$ be any positive constant such that $\sup_{f \in \mathcal{F}} \|f\|_{P,2}^2 \leq \sigma^2 \leq \|F\|_{P,2}^2$. Suppose that there exist constants $a \geq e$ and $v \geq 1$ such that*

$$\log \sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leq v \log(a/\epsilon), \ 0 < \epsilon \leq 1.$$

Then

$$E_P[\|\mathbb{G}_n\|_{\mathcal{F}}] \leq K \left(\sqrt{v \sigma^2 \log\Big(\frac{a\|F\|_{P,2}}{\sigma}\Big)} + \frac{v \|M\|_{P,2}}{\sqrt{n}} \log\Big(\frac{a\|F\|_{P,2}}{\sigma}\Big) \right),$$

where K is an absolute constant. Moreover, for every $t \geq 1$, with probability $> 1 - t^{-q/2}$, $\|\mathbb{G}_n\|_{\mathcal{F}} \leq (1+\alpha) E_P[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q) \Big((\sigma + n^{-1/2} \|M\|_{P,q}) \sqrt{t} + \alpha^{-1} n^{-1/2} \|M\|_{P,2} t \Big), \quad \forall \alpha > 0$, where $K(q) > 0$ is a constant depending only on q. In particular, setting $a \geq n$ and $t = \log n$, with probability $> 1 - c(\log n)^{-1}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leq K(q, c) \left(\sigma \sqrt{v \log\Big(\frac{a\|F\|_{P,2}}{\sigma}\Big)} + \frac{v \|M\|_{P,q}}{\sqrt{n}} \log\Big(\frac{a\|F\|_{P,2}}{\sigma}\Big) \right), \tag{A.1}$$

where $\|M\|_{P,q} \leq n^{1/q} \|F\|_{P,q}$ and $K(q, c) > 0$ is a constant depending only on q and c.

Proof of Lemma 2.1: Because J exists and $J_{\beta\beta}$ is invertible, (2.8) has the unique solution μ_0 given in (2.10), and so we have by (2.6) that $E[\psi(W; \theta_0, \eta_0)] = 0$ for η_0 given in (2.9). Moreover,

$$\partial_{\eta'} E_P \psi(W; \theta_0, \eta_0) = \Big([J_{\theta\beta} - \mu_0 J_{\beta\beta}], E[\partial_{\beta'} \ell(W; \theta_0, \beta_0)] \otimes I_{d_\theta \times d_\theta} \Big) = 0,$$

where $I_{d_\theta \times d_\theta}$ is the $d_\theta \times d_\theta$ identity matrix and $\otimes$ is the Kronecker product. Hence, the asserted claim holds by the remark after Definition 2.1. □

Proof of Lemma 2.2: The proof is similar to that of Lemma 2.1, except that now we have to verify (2.4) intead of (2.3). To do so, take any $\beta \in \mathcal{B}$ such that $\|\beta - \beta_0\|_q^* \leq \lambda_N / r_N$ and any $d_\theta \times d_\beta$ matrix μ. Denote $\eta = (\beta', \text{vec}(\mu)')'$. Then

$$\begin{aligned}\|\partial_\eta E_P \psi(W, \theta_0, \eta_0)[\eta - \eta_0]\| &= \|(J_{\theta\beta} - \mu_0 J_{\beta\beta})(\beta - \beta_0)\| \\ &\leq \|J_{\theta\beta} - \mu_0 J_{\beta\beta}\|_q \times \|\beta - \beta_0\|_q^* \leq r_n \times (\lambda_N / r_N) = \lambda_N.\end{aligned}$$

This completes the proof of the lemma. □

Proof of Lemma 2.3: The proof is similar to that of Lemma 2.1, except that now we have

$$\partial_{\eta'} E_P \psi(W, \theta_0, \eta_0) = [\mu_0 G_\beta, E_P m(W, \theta_0, \beta_0)' \otimes I_{d_\theta \times d_\theta}] = 0,$$

where $I_{d_\theta \times d_\theta}$ is the $d_\theta \times d_\theta$ identity matrix and $\otimes$ is the Kronecker product. □

Proof of Lemma 2.4: The proof is similar to that of Lemma 2.2, except that now for any $\beta \in \mathcal{B}$ such that $\|\beta - \beta_0\|_1 \leq \lambda_N / r_N$, any $d_\theta \times k$ matrix μ, and $\eta = (\beta', \text{vec}(\mu)')'$, we have

$$\begin{aligned}\|\partial_\eta E_P \psi(W, \theta_0, \eta_0)[\eta - \eta_0]\| &= \|\mu_0 G_\beta (\beta - \beta_0)\| \\ &\leq \|A' \Omega^{-1/2} L - \gamma_0 L' L\|_\infty \times \|\beta - \beta_0\|_1 \\ &\leq r_n \times (\lambda_N / r_N) = \lambda_N.\end{aligned}$$

□

Proof of Lemma 2.5: Take any $\eta \in T$, and consider the function

$$Q(W; \theta, r) := \ell(W; \theta, \eta_0(\theta) + r(\eta(\theta) - \eta_0(\theta))), \quad \theta \in \Theta, \ r \in [0, 1].$$

Then

$$\psi(W; \theta, \eta_0 + r(\eta - \eta_0)) = \partial_\theta Q(W; \theta, r),$$

and so

$$\begin{aligned}\partial_r E_P[\psi(W; \theta, \eta_0 + r(\eta - \eta_0))] &= \partial_r E_P[\partial_\theta Q(W; \theta, r)] \\ &= \partial_r \partial_\theta E_P[Q(W; \theta, r)] = \partial_\theta \partial_r E_P[Q(W; \theta, r)] \\ &= \partial_\theta \partial_r E_P[\ell(W; \theta, \eta_0(\theta) + r(\eta(\theta) - \eta_0(\theta)))]. \qquad \text{(A.2)}\end{aligned}$$

Hence,

$$\partial_r E_P[\psi(W; \theta, \eta_0 + r(\eta - \eta_0))]|_{r=0} = 0$$

because

$$\partial_r E_P[\ell(W; \theta, \eta_0(\theta) + r(\eta(\theta) - \eta_0(\theta)))]|_{r=0} = 0, \quad \text{for all } \theta \in \Theta,$$

as $\eta_0(\theta) = \beta_\theta$ solves the optimization problem

$$\max_{\beta \in \mathcal{B}} E_P[\ell(W; \theta, \beta)], \quad \text{for all } \theta \in \Theta.$$

Here, the regularity conditions are needed to make sure that we can interchange E_P and ∂_θ and also ∂_θ and ∂_r in (A.2). □

Proof of Lemma 2.6: First, we demonstrate that $\mu_0 \in \mathcal{L}^1(\mathcal{R};\ \mathbb{R}^{d_\theta \times d_m})$. Indeed,

$$\begin{aligned} E_P[\|\mu_0(R)\|] &\leq E_P[\|A(R)'\Omega(R)^{-1}\|] + E_P[\|G(Z)\Gamma(R)\Omega(R)^{-1}\|] \\ &\leq E_P[\|A(R)\| \times \|\Omega(R)\|^{-1}] + E_P[\|G(Z)\| \times \|\Gamma(R)\| \times \|\Omega(R)\|^{-1}] \\ &\leq (E_P[\|A(R)\|^2] \times E_P[\|\Omega(R)\|^{-2}])^{1/2} \\ &\quad + (E_P[\|G(Z)\|^2 \times \|\Gamma(R)\|^2] \times E_P[\|\Omega(R)\|^{-2}])^{1/2}, \end{aligned}$$

which is finite by assumptions of the lemma as

$$E_P[\|G(Z)\|^2 \times \|\Gamma(R)\|^2] \leq (E_P[\|G(Z)\|^4] \times E_P[\Gamma(R)\|^4])^{1/2} < \infty.$$

Next, we demonstrate that

$$E_P[\|\psi(W, \theta_0, \eta)\|] < \infty \quad \text{for all } \eta \in T.$$

Indeed, for all $\eta \in T$, there exist $\mu \in \mathcal{L}^1(\mathcal{R};\ \mathbb{R}^{d_\theta \times d_m})$ and $h \in \mathcal{H}$ such that $\eta = (\mu, h)$, and so

$$\begin{aligned} E_P[\|\psi(W, \theta_0, \eta)\|] &= E_P[\|\mu(X)m(W, \theta_0, h(Z))\|] \\ &\leq E_P[\|\mu(R)\| \times \|m(W, \theta_0, h(Z))\|] \\ &= E_P[\|\mu(R)\| \times E_P[\|m(W, \theta_0, h(Z)) \mid R]] \leq C_h E[\|\mu(R)\|], \end{aligned}$$

which is finite by assumptions of the lemma. Further, (2.1) holds because

$$\begin{aligned} E_P[\psi(W, \theta_0, \eta_0)] &= E_P[\mu_0(R)m(W, \theta_0, h_0(Z))] \\ &= E_P[\mu_0(R)E_P[m(W, \theta_0, h_0(Z)) \mid R]] = 0, \end{aligned} \tag{A.3}$$

where the last equality follows from (2.22).

Finally, we demonstrate that (2.3) holds. To do so, take any $\eta = (\mu, h) \in \mathcal{T}_N = T$. Then

$$\begin{aligned} &E_P[\psi(W, \theta_0, \eta_0 + r(\eta - \eta_0)] \\ &\quad = E_P[(\mu_0(R) + r(\mu(R) - \mu_0(R)))m(W, \theta_0, h_0(Z) + r(h(Z) - h_0(Z)))], \end{aligned}$$

and so

$$\partial_\eta E_P[\psi(W, \theta_0, \eta_0)][\eta - \eta_0] = \mathcal{I}_1 + \mathcal{I}_2,$$

where

$$\begin{aligned} \mathcal{I}_1 &= E_P[(\mu(R) - \mu_0(R))m(W, \theta_0, h_0(Z)], \\ \mathcal{I}_2 &= E_P[\mu_0(R)\partial_{v'} m(W, \theta_0, v)|_{v=h_0(Z)}(h(Z) - h_0(Z))]. \end{aligned}$$

Here, $\mathcal{I}_1 = 0$ by the same argument as that in (A.3) and $\mathcal{I}_2 = 0$ because

$$\begin{aligned} \mathcal{I}_2 &= E_P[\mu_0(R)E_P[\partial_{v'} m(W, \theta_0, v)|_{v=h_0(Z)} \mid X](h(Z) - h_0(Z))] \\ &= E_P[\mu_0(R)\Gamma(X)(h(Z) - h_0(Z))] = E_P[E_P[\mu_0(R)\Gamma(R) \mid Z](h(Z) - h_0(Z))] = 0 \end{aligned}$$

because

$$\begin{aligned} E_P[\mu_0(R)\Gamma(X) \mid Z] &= E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] - E_P[G(Z)\Gamma(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] \\ &= E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] - G(Z)E_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] \\ &= E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] - E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] \\ &\quad \times \Big(E_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R) \mid Z]\Big)^{-1} \times E_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] \\ &= E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] - E_P[A(R)'\Omega(R)^{-1}\Gamma(R) \mid Z] = 0. \end{aligned}$$

□

Proof of Theorem 3.1 (DML2 case): To start with, note that (3.11) follows immediately from the assumptions. Hence, it suffices to show that (3.10) holds uniformly over $P \in \mathcal{P}_N$.

Fix any sequence $\{P_N\}_{N\geq 1}$ such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, to show that (3.10) holds uniformly over $P \in \mathcal{P}_N$, it suffices to show that

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{\psi}(W_i) + O_{P_N}(\rho_N) \rightsquigarrow N(0, \mathrm{I}_d). \tag{A.4}$$

To do so, we proceed in five steps. Step 1 shows the main argument, and Steps 2–5 present auxiliary calculations. In the proof, it will be convenient to denote by $\mathcal{E}_N$ the event that $\hat{\eta}_{0,k} \in \mathcal{T}_N$ for all $k \in [K]$. Note that by Assumption 3.2 and the union bound, $Pr_{P_N}(\mathcal{E}_N) \geq 1 - K\Delta_n = 1 - o(1)$ as $\Delta_n = o(1)$.

STEP 1. Denote

$$\hat{J}_0 := \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi^a(W; \hat{\eta}_{0,k})], \quad R_{N,1} := \hat{J}_0 - J_0,$$

$$R_{N,2} := \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi(W; \theta_0, \hat{\eta}_{0,k})] - \frac{1}{N}\sum_{i=1}^{N}\psi(W_i; \theta_0, \eta_0).$$

In Steps 2–5 respectively, we will show that

$$\|R_{N,1}\| = O_{P_N}(N^{-1/2} + r_N), \tag{A.5}$$

$$\|R_{N,2}\| = O_{P_N}(N^{-1/2}r'_N + \lambda_N + \lambda'_N), \tag{A.6}$$

$$\left\|N^{-1/2}\sum_{i=1}^{N}\psi(W_i; \theta_0, \eta_0)\right\| = O_{P_N}(1), \tag{A.7}$$

$$\|\sigma^{-1}\| = O_{P_N}(1). \tag{A.8}$$

Because $N^{-1/2} + r_N \leq \rho_N = o(1)$ and all singular values of J_0 are bounded below from zero by Assumption 3.1, it follows from (A.5) that with P_N-probability $1 - o(1)$, all singular values of $\hat{J}_0$ are bounded below from zero as well. Therefore, with the same P_N-probability,

$$\tilde{\theta}_0 = -\hat{J}_0^{-1}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi^b(W; \hat{\eta}_{0,k})]$$

and

$$\sqrt{N}(\tilde{\theta}_0 - \theta_0) = -\sqrt{N}\hat{J}_0^{-1}\left(\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi^b(W;\hat{\eta}_{0,k})] + \hat{J}_0\theta_0\right)$$

$$= -\sqrt{N}\hat{J}_0^{-1}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi(W;\theta_0,\hat{\eta}_{0,k})]$$

$$= -\left(J_0 + R_{N,1}\right)^{-1} \times \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0) + \sqrt{N}R_{N,2}\right). \tag{A.9}$$

In addition, given that

$$(J_0 + R_{N,1})^{-1} - J_0^{-1} = (J_0 + R_{N,1})^{-1}(J_0 - (J_0 + R_{N,1}))J_0^{-1}$$
$$= -(J_0 + R_{N,1})^{-1}R_{N,1}J_0^{-1},$$

it follows from (A.5) that

$$\|(J_0 + R_{N,1})^{-1} - J_0^{-1}\| \le \|(J_0 + R_{N,1})^{-1}\| \times \|R_{N,1}\| \times \|J_0^{-1}\|$$
$$= O_{P_N}(1)O_{P_N}(N^{-1/2} + r_N)O_{P_N}(1) = O_{P_N}(N^{-1/2} + r_N). \tag{A.10}$$

Moreover, because $r'_N + \sqrt{N}(\lambda_N + \lambda'_N) \le \rho_N = o(1)$, it follows from (A.6) and (A.7) that

$$\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0) + \sqrt{N}R_{N,2}\right\| \le \left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0)\right\| + \|\sqrt{N}R_{N,2}\|$$
$$= O_{P_N}(1) + o_{P_N}(1) = O_{P_N}(1). \tag{A.11}$$

Combining (A.10) and (A.11) gives

$$\left\|((J_0 + R_{N,1})^{-1} - J_0^{-1}) \times \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0) + \sqrt{N}R_{N,2}\right)\right\|$$

$$\le \left\|(J_0 + R_{N,1})^{-1} - J_0^{-1}\right\| \times \left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0) + \sqrt{N}R_{N,2}\right\|$$

$$= O_{P_N}(N^{-1/2} + r_N).$$

Now, substituting the last bound into (A.9) yields

$$\sqrt{N}(\tilde{\theta}_0 - \theta_0) = -J_0^{-1} \times \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0) + \sqrt{N}R_{N,2}\right) + O_{P_N}(N^{-1/2} + r_N)$$

$$= -J_0^{-1} \times \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(W_i;\theta_0,\eta_0) + O_{P_N}(\rho_N),$$

where in the second line we used (A.6) and the definition of ρ_N. Combining this with (A.8) gives

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{\psi}(W_i) + O_{P_N}(\rho_N) \tag{A.12}$$

by the definition of $\bar{\psi}$ given in the statement of the theorem. In turn, because $\rho_N = o(1)$, combining (A.12) with the Lindeberg–Feller central limit theorem (CLT) and the Cramer–Wold device yields (A.4). To complete the proof of the theorem, it remains to establish the bounds (A.5)–(A.8). We do so in the following four steps.

STEP 2. In this step, we establish (A.5). Because K is a fixed integer, which is independent of N, it suffices to show that for any $k \in [K]$,

$$\|E_{n,k}[\psi^a(W;\hat{\eta}_{0,k})] - E_{P_N}[\psi^a(W;\eta_0)]\| = O_{P_N}(N^{-1/2} + r_N). \tag{A.13}$$

To do so, fix any $k \in [K]$ and observe that by the triangle inequality,

$$\|E_{n,k}[\psi^a(W;\hat{\eta}_{0,k})] - E_{P_N}[\psi^a(W;\eta_0)]\| \leq \mathcal{I}_{1,k} + \mathcal{I}_{2,k}, \tag{A.14}$$

where

$$\mathcal{I}_{1,k} := \|E_{n,k}[\psi^a(W;\hat{\eta}_{0,k})] - E_{P_N}[\psi^a(W;\hat{\eta}_{0,k}) \mid (W_i)_{i\in I_k^c}]\|,$$
$$\mathcal{I}_{2,k} := \|E_{P_N}[\psi^a(W;\hat{\eta}_{0,k}) \mid (W_i)_{i\in I_k^c}] - E_{P_N}[\psi^a(W;\eta_0)]\|.$$

To bound $\mathcal{I}_{2,k}$, note that on the event $\mathcal{E}_N$, which holds with P_N-probability $1 - o(1)$,

$$\mathcal{I}_{2,k} \leq \sup_{\eta\in\mathcal{T}_N} \|E_{P_N}[\psi^a(W;\eta)] - E_{P_N}[\psi^a(W;\eta_0)]\| = r_N,$$

and so $\mathcal{I}_{2,k} = O_{P_N}(r_N)$. To bound $\mathcal{I}_{1,k}$, note that conditional on $(W_i)_{i\in I_k^c}$, the estimator $\hat{\eta}_{0,k}$ is non-stochastic, and so on the event $\mathcal{E}_N$,

$$\begin{aligned} E_{P_N}[\mathcal{I}_{1,k}^2 \mid (W_i)_{i\in I_k^c}] &\leq n^{-1} E_{P_N}[\|\psi^a(W;\hat{\eta}_{0,k})\|^2 \mid (W_i)_{i\in I_k^c}] \\ &\leq \sup_{\eta\in\mathcal{T}_N} n^{-1} E_{P_N}[\|\psi^a(W;\eta)\|^2] \leq c_1^2/n, \end{aligned}$$

where the last inequality holds by Assumption 3.2. Hence, $\mathcal{I}_{1,k} = O_{P_N}(N^{-1/2})$ by Lemma 6.1. Combining the bounds $\mathcal{I}_{1,k} = O_{P_N}(N^{-1/2})$ and $\mathcal{I}_{2,k} = O_{P_N}(r_N)$ with (A.14) gives (A.13).

STEP 3. In this step, we establish (A.6). This is the step where we invoke the Neyman orthogonality (or near-orthogonality) condition. Again, because K is a fixed integer, which is independent of N, it suffices to show that for any $k \in [K]$,

$$E_{n,k}[\psi(W;\theta_0,\hat{\eta}_{0,k})] - \frac{1}{n}\sum_{i\in I_k} \psi(W_i;\theta_0,\eta_0) = O_{P_N}(N^{-1/2} r_N' + \lambda_N + \lambda_N'). \tag{A.15}$$

To do so, fix any $k \in [K]$ and introduce the following additional empirical process notation,

$$\mathbb{G}_{n,k}[\phi(W)] = \frac{1}{\sqrt{n}} \sum_{i\in I_k} \left(\phi(W_i) - \int \phi(w) dP_N \right),$$

where ϕ is any P_N-integrable function on $\mathcal{W}$. Then observe that by the triangle inequality,

$$\left\| E_{n,k}[\psi(W;\theta_0,\hat{\eta}_{0,k})] - \frac{1}{n}\sum_{i\in I_k} \psi(W_i;\theta_0,\eta_0) \right\| \leq \frac{\mathcal{I}_{3,k} + \mathcal{I}_{4,k}}{\sqrt{n}}, \tag{A.16}$$

where

$$\mathcal{I}_{3,k} := \|\mathbb{G}_{n,k}[\psi(W;\theta_0,\hat{\eta}_{0,k})] - \mathbb{G}_{n,k}[\psi(W;\theta_0,\eta_0)]\|,$$
$$\mathcal{I}_{4,k} := \sqrt{n}\|E_{P_N}[\psi(W;\theta_0,\hat{\eta}_{0,k}) \mid (W_i)_{i\in I_k^c}] - E_{P_N}[\psi(W;\theta_0,\eta_0)]\|.$$

To bound $\mathcal{I}_{3,k}$, note that, as above, conditional on $(W_i)_{i\in I_k^c}$, the estimator $\hat{\eta}_{0,k}$ is non-stochastic, and so on the event $\mathcal{E}_N$,

$$\begin{aligned} E_{P_N}[\mathcal{I}_{3,k}^2 \mid (W_i)_{i\in I_k^c}] &= E_{P_N}[\|\psi(W;\theta_0,\hat{\eta}_{0,k}) - \psi(W;\theta_0,\eta_0)\|^2 \mid (W_i)_{i\in I_k^c}] \\ &\leq \sup_{\eta\in\mathcal{T}_N} E_{P_N}[\|\psi(W;\theta_0,\eta) - \psi(W;\theta_0,\eta_0)\|^2 \mid (W_i)_{i\in I_k^c}] \\ &\leq \sup_{\eta\in\mathcal{T}_N} E_{P_N}[\|\psi(W;\theta_0,\eta) - \psi(W;\theta_0,\eta_0)\|^2] = (r_N')^2 \end{aligned}$$

by the definition of r_N' in Assumption 3.2. Hence, $\mathcal{I}_{3,k} = O_{P_N}(r_N')$ by Lemma 6.1. To bound $\mathcal{I}_{4,k}$, introduce the function

$$f_k(r) := E_{P_N}[\psi(W;\theta_0,\eta_0 + r(\hat{\eta}_{0,k} - \eta_0)) \mid (W_i)_{i\in I_k^c}] - E_{P_N}[\psi(W;\theta_0,\eta_0)], \quad r \in [0,1].$$

Then, by Taylor expansion,

$$f_k(1) = f_k(0) + f_k'(0) + f_k''(\tilde{r})/2, \quad \text{for some } \tilde{r} \in (0,1).$$

However, $\|f_k(0)\| = 0$ because

$$E_{P_N}[\psi(W;\theta_0,\eta_0) \mid (W_i)_{i\in I_k^c}] = E_{P_N}[\psi(W;\theta_0,\eta_0)].$$

In addition, on the event $\mathcal{E}_N$, by the Neyman λ_N near-orthogonality condition imposed in Assumption 3.1,

$$\|f_k'(0)\| = \|\partial_\eta E_{P_N}\psi(W;\theta_0,\eta_0)[\hat{\eta}_{0,k} - \eta_0]\| \leq \lambda_N.$$

Moreover, on the event $\mathcal{E}_N$,

$$\|f_k''(\tilde{r})\| \leq \sup_{r\in(0,1)} \|f_k''(r)\| \leq \lambda_N'$$

by the definition λ_N' in Assumption 3.2. Hence,

$$\mathcal{I}_{4,k} = \sqrt{n}\|f_k(1)\| = O_{P_N}(\sqrt{n}(\lambda_N + \lambda_N')).$$

Combining the bounds on $\mathcal{I}_{3,k}$ and $\mathcal{I}_{4,k}$ with (A.16) and using the fact that $n^{-1} = O(N^{-1})$ gives (A.15).

STEP 4. To establish (A.7), note that

$$E_{P_N}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \psi(W_i;\theta_0,\eta_0)\right\|^2\right] = E_{P_N}\left[\|\psi(W;\theta_0,\eta_0)\|^2\right] \leq c_1^2$$

by Assumption 3.2. Combining this with the Markov inequality gives (A.7).

STEP 5. Here we establish (A.8). Note that all eigenvalues of the matrix

$$\sigma^2 = J_0^{-1}E_P[\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)'](J_0^{-1})'$$

are bounded from below by c_0/c_1^2 because all singular values of J_0 are bounded from above by c_1 by Assumption 3.1 and all eigenvalues of $E_P[\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)']$ are bounded from below by c_0 by Assumption 3.2. Hence, given that $\|\sigma^{-1}\|$ is the largest eigenvalue of σ^{-1}, it follows that $\|\sigma^{-1}\| = c_1/\sqrt{c_0}$. This gives (A.8) and completes the proof of the theorem. □

Proof of Theorem 3.1 (DML1 case): As in the case of the DML2 version, note that (3.11) follows immediately from the assumptions, and so it suffices to show that (3.10) holds uniformly over $P \in \mathcal{P}_N$.

Fix any sequence $\{P_N\}_{N\geq 1}$ such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, to show that (3.10) holds uniformly over $P \in \mathcal{P}_N$, it suffices to show that

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\psi}(W_i) + O_{P_N}(\rho_N) \rightsquigarrow N(0, \mathrm{I}_d). \tag{A.17}$$

To do so, for all $k \in [K]$, denote

$$\hat{J}_{0,k} := E_{n,k}[\psi^a(W; \hat{\eta}_{0,k})], \quad R_{N,1,k} := \hat{J}_{0,k} - J_0,$$
$$R_{N,2,k} := E_{n,k}[\psi(W; \theta_0, \hat{\eta}_{0,k})] - \frac{1}{n} \sum_{i \in I_k} \psi(W_i; \theta_0, \eta_0).$$

Because K is a fixed integer, which is independent of n, it follows by the same arguments as those in Steps 2–5 in the Proof of Theorem 3.1 (DML2 case) that

$$\max_{k\in[K]} \|R_{N,1,k}\| = O_{P_N}(N^{-1/2} + r_N), \tag{A.18}$$

$$\max_{k\in[K]} \|R_{N,2,k}\| = O_{P_N}(N^{-1/2} r'_N + \lambda_N + \lambda'_N), \tag{A.19}$$

$$\max_{k\in[K]} \|n^{-1/2} \textstyle\sum_{i\in I_k} \psi(W_i; \theta_0, \eta_0)\| = O_{P_N}(1), \tag{A.20}$$

$$\|\sigma^{-1}\| = O_{P_N}(1). \tag{A.21}$$

Because $N^{-1/2} + r_N \leq \rho_N = o(1)$ and all singular values of J_0 are bounded below from zero by Assumption 3.1, it follows from (A.18) that for all $k \in [K]$, with P_N-probability $1 - o(1)$, all singular values of $\hat{J}_{0,k}$ are bounded below from zero, and so with the same P_N-probability,

$$\check{\theta}_{0,k} = -\hat{J}_{0,k}^{-1} E_{n,k}[\psi^b(W; \hat{\eta}_{0,k})].$$

Hence, by the same arguments as those in Step 1 in the Proof of Theorem 3.1 (DML2 case), it follows from the bounds (A.18)–(A.21) that for all $k \in [K]$,

$$\sqrt{n}\sigma^{-1}(\check{\theta}_{0,k} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i \in I_k} \bar{\psi}(W_i) + O_{P_N}(\rho_N).$$

Therefore,

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \sqrt{N}\sigma^{-1}\Big(\frac{1}{K} \sum_{k=1}^{K} \check{\theta}_{0,k} - \theta_0\Big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\psi}(W_i) + O_{P_N}(\rho_N). \tag{A.22}$$

In turn, because $\rho_N = o(1)$, combining (A.22) with the Lindeberg–Feller CLT and the Cramer–Wold device yields (A.17) and completes the proof of the theorem. □

Proof of Theorem 3.2: In this proof, all bounds hold uniformly in $P \in \mathcal{P}_N$ for $N \geq 3$, and we do not repeat this qualification throughout. Also, the second asserted claim follows immediately from the first one and Theorem 3.1. Hence, it suffices to prove the first asserted claim.

In the Proof of Theorem 3.1 (DML2 case), we established that $\|\hat{J}_0 - J_0\| = O_P(r_N + N^{-1/2})$. Hence, because $\|J_0^{-1}\| \leq c_0^{-1}$ by Assumption 3.1 and

$$\|E_P[\psi(W; \theta_0, \eta_0)\psi(W; \theta_0, \eta_0)']\| \leq E_P[\|\psi(W; \theta_0, \eta_0)\|^2] \leq c_1^2$$

by Assumption 3.2, it suffices to show that

$$\Big\| \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{n,k}[\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})'] - E_P[\psi(W; \theta_0, \eta_0)\psi(W; \theta_0, \eta_0)'] \Big\| = O_P(\varrho_N).$$

Moreover, as both K and d_θ, the dimension of ψ, are fixed integers, which are independent of N, the last bound will follow if we show that for all $k \in [K]$ and all $j, k \in [d_\theta]$,

$$\mathcal{I}_{kjl} := |\mathbb{E}_{n,k}[\psi_j(W; \tilde{\theta}_0, \hat{\eta}_{0,k})\psi_l(W; \tilde{\theta}_0, \hat{\eta}_{0,k})] - E_P[\psi_j(W; \theta_0, \eta_0)\psi_l(W; \theta_0, \eta_0)]|$$

satisfies

$$\mathcal{I}_{kjl} = O_P(\varrho_N). \tag{A.23}$$

To do so, observe that by the triangle inequality,

$$\mathcal{I}_{kjl} \leq \mathcal{I}_{kjl,1} + \mathcal{I}_{kjl,2}, \tag{A.24}$$

where

$$\mathcal{I}_{kjl,1} := |E_{n,k}[\psi_j(W; \tilde{\theta}_0, \hat{\eta}_{0,k})\psi_l(W; \tilde{\theta}_0, \hat{\eta}_{0,k})] - E_{n,k}[\psi_j(W; \theta_0, \eta_0)\psi_l(W; \theta_0, \eta_0)]|,$$
$$\mathcal{I}_{kjl,2} := |\mathbb{E}_{n,k}[\psi_j(W; \theta_0, \eta_0)\psi_l(W; \theta_0, \eta_0)] - E_P[\psi_j(W; \theta_0, \eta_0)\psi_l(W; \theta_0, \eta_0)]|.$$

We bound $\mathcal{I}_{kjl,2}$ first. If $q \geq 4$, then

$$\begin{aligned} E_P[\mathcal{I}_{kjl,2}^2] &\leq n^{-1} E_P\Big[(\psi_j(W; \theta_0, \eta_0)\psi_l(W; \theta_0, \eta_0))^2\Big] \\ &\leq n^{-1}\Big(E_P[\psi_j^4(W; \theta_0, \eta_0)] \times E_P[\psi_l^4(W; \theta_0, \eta_0)]\Big)^{1/2} \\ &\leq n^{-1} E_P[\|\psi(W; \theta_0, \eta_0)\|^4] \leq c_1^4, \end{aligned}$$

where the second line holds by the Hölder inequality, and the third by Assumption 3.2. Hence, $\mathcal{I}_{kjl,2} = O_P(N^{-1/2})$. If $q \in (2, 4)$, we apply the following von Bahr–Esseen inequality with $p = q/2$: if $X_1, \ldots, X_n$ are independent random variables with mean zero, then for any $p \in [1, 2]$,

$$E\Big[\Big|\sum_{i=1}^n X_i\Big|^p\Big] \leq \Big(2 - \frac{1}{n}\Big)\sum_{i=1}^n E[|X_i|^p];$$

see DasGupta (2008, p. 650). This gives

$$\begin{aligned} E_P[\mathcal{I}_{kjl,2}^{q/2}] &\lesssim n^{-q/2+1} E_P[(\psi_j(W; \theta_0, \eta_0)\psi_l(W; \theta_0, \eta_0))^{q/2}] \\ &\leq n^{-q/2+1} E_P[\|\psi(W; \theta_0, \eta_0)\|^q] \lesssim n^{-q/2+1} \end{aligned}$$

by Assumption 3.2. Hence, $\mathcal{I}_{kjl,2} = O_P(N^{2/q-1})$. We conclude that

$$\mathcal{I}_{kjl,2} = O_P(N^{-[(1-2/q)\wedge(1/2)]}). \tag{A.25}$$

Next, we bound $\mathcal{I}_{kjl,1}$. To do so, observe that for any numbers a, b, δa and δb such that $|a| \vee |b| \leq c$ and $|\delta a| \vee |\delta b| \leq r$, we have

$$|(a + \delta a)(b + \delta b) - ab| \leq 2r(c + r).$$

Denoting

$$\psi_{hi} := \psi_h(W_i; \theta_0, \eta_0) \text{ and } \hat{\psi}_{hi} := \psi_h(W_i; \tilde{\theta}_0, \hat{\eta}_{0,k}), \text{ for } (h, i) \in \{j, l\} \times I_k,$$

and applying the inequality above with $a := \psi_{ji}$, $b := \psi_{li}$, $a + \delta a := \hat{\psi}_{ji}$, $b + \delta b := \hat{\psi}_{li}$, $r := |\hat{\psi}_{ji} - \psi_{ji}| \vee |\hat{\psi}_{li} - \psi_{li}|$ and $c := |\psi_{ji}| \vee |\psi_{li}|$ gives

$$\mathcal{I}_{kjl,1} = \left| \frac{1}{n} \sum_{i \in I_k} \hat{\psi}_{ji} \hat{\psi}_{li} - \psi_{ji} \psi_{li} \right| \leq \frac{1}{n} \sum_{i \in I_k} |\hat{\psi}_{ji} \hat{\psi}_{li} - \psi_{ji} \psi_{li}|$$

$$\leq \frac{2}{n} \sum_{i \in I_k} (|\hat{\psi}_{ji} - \psi_{ji}| \vee |\hat{\psi}_{li} - \psi_{li}|) \times (|\psi_{ji}| \vee |\psi_{li}| + |\hat{\psi}_{ji} - \psi_{ji}| \vee |\hat{\psi}_{li} - \psi_{li}|)$$

$$\leq \left(\frac{2}{n} \sum_{i \in I_k} (|\hat{\psi}_{ji} - \psi_{ji}|^2 \vee |\hat{\psi}_{li} - \psi_{li}|^2) \right)^{1/2}$$

$$\times \left(\frac{2}{n} \sum_{i \in I_k} (|\psi_{ji}| \vee |\psi_{li}| + |\hat{\psi}_{ji} - \psi_{ji}| \vee |\hat{\psi}_{li} - \psi_{li}|)^2 \right)^{1/2}.$$

In addition, the expression in the last line above is bounded by

$$\left(\frac{2}{n} \sum_{i \in I_k} |\psi_{ji}|^2 \vee |\psi_{li}|^2 \right)^{1/2} + \left(\frac{2}{n} \sum_{i \in I_k} |\hat{\psi}_{ji} - \psi_{ji}|^2 \vee |\hat{\psi}_{li} - \psi_{li}|^2 \right)^{1/2},$$

and so

$$\mathcal{I}_{kjl,1}^2 \lesssim R_N \times \left(\frac{1}{n} \sum_{i \in I_k} \|\psi(W_i; \theta_0, \eta_0)\|^2 + R_N \right),$$

where

$$R_N := \frac{1}{n} \sum_{i \in I_k} \|\psi(W_i; \tilde{\theta}_0, \hat{\eta}_{0,k}) - \psi(W_i; \theta_0, \eta_0)\|^2.$$

Moreover,

$$\frac{1}{n} \sum_{i \in I_k} \|\psi(W_i; \theta_0, \eta_0)\|^2 = O_P(1),$$

by the Markov inequality because

$$E_P \left[\frac{1}{n} \sum_{i \in I_k} \|\psi(W_i; \theta_0, \eta_0)\|^2 \right] = E_P[\|\psi(W; \theta_0, \eta_0\|^2] \leq c_1^2$$

by Assumption 3.2. It remains to bound R_N. We have

$$R_N \lesssim \frac{1}{n} \sum_{i \in I_k} \|\psi^a(W_i; \hat{\eta}_{0,k})(\tilde{\theta}_0 - \theta_0)\|^2 + \frac{1}{n} \sum_{i \in I_k} \|\psi(W_i; \theta_0, \hat{\eta}_{0,k}) - \psi(W_i; \theta_0, \eta_0)\|^2. \tag{A.26}$$

The first term on the right-hand side of (A.26) is bounded from above by

$$\left(\frac{1}{n} \sum_{i \in I_k} \|\psi^a(W_i; \hat{\eta}_{0,k})\|^2 \right) \times \|\tilde{\theta}_0 - \theta_0\|^2 = O_P(1) \times O_P(N^{-1}) = O_P(N^{-1}),$$

and the conditional expectation of the second term given $(W_i)_{i \in I_k^c}$ on the event that $\hat{\eta}_{0,k} \in \mathcal{T}_N$ is equal to

$$E_P[\|\psi(W; \theta_0, \hat{\eta}_{0,k}) - \psi(W; \theta_0, \eta_0)\|^2 \mid (W_i)_{i \in I_k^c}]$$

$$\leq \sup_{\eta \in \mathcal{T}_N} E_P[\|\psi(W; \theta_0, \eta) - \psi(W; \theta_0, \eta_0)\|^2 \mid (W_i)_{i \in I_k^c}] = (r'_N)^2.$$

Because the event that $\hat{\eta}_{0,k} \in \mathcal{T}_N$ holds with probability $1 - \Delta_N = 1 - o(1)$, it follows that $R_N = O_P(N^{-1} + (r'_N)^2)$, and so

$$\mathcal{I}_{kjl,1} = O_P(N^{-1/2} + r'_N). \tag{A.27}$$

Combining the bounds (A.25) and (A.27) with (A.24) gives (A.23) and completes the proof of the theorem. □

Proof of Theorem 3.3: We only consider the case of the DML1 estimator and note that the DML2 estimator can be treated similarly.

With the help of Lemma 6.3, which establishes approximate linearity of the subsample DML estimators $\check{\theta}_{0,k}$ and is presented below, the proof is the same as that in the linear case presented in the Proof of Theorem 3.1 (DML1 case).

LEMMA 6.3. (LINEARIZATION FOR SUBSAMPLE DML IN NON-LINEAR PROBLEMS) *Under the conditions of Theorem 3.3, for any $k = 1, \ldots, K$, the estimator $\check{\theta}_0 = \check{\theta}_{0,k}$ defined by (3.2) obeys*

$$\sqrt{n}\sigma_0^{-1}(\check{\theta}_0 - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i \in I} \bar{\psi}(W_i) + O_P(\rho'_n) \tag{A.28}$$

uniformly over $P \in \mathcal{P}_N$, where $\rho'_n = n^{-1/2} + r_N + r'_N + n^{1/2}\lambda_N + n^{1/2}\lambda'_N \lesssim \delta_N$ and where $\bar{\psi}(\cdot) := -\sigma_0^{-1} J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$.

Proof: Fix any $k = 1, \ldots, K$ and any sequence $\{P_N\}_{N \geq 1}$ such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. To prove the asserted claim, it suffices to show that the estimator $\check{\theta}_0 = \check{\theta}_{0,k}$ satisfies (A.28) with P replaced by P_N. To do so, we split the proof into four steps. In the proof, we use E_n, $\mathbb{G}_n$, I and $\hat{\eta}_0$ instead of $E_{n,k}$, $\mathbb{G}_{n,k}$, I_k and $\hat{\eta}_{0,k}$, respectively.

STEP 1. (Preliminary rate result) We claim that with P_N-probability $1 - o(1)$,

$$\|\check{\theta}_0 - \theta_0\| \leq \tau_N. \tag{A.29}$$

To show this claim, note that the definition of $\check{\theta}_0$ implies that

$$\|E_n[\psi(W; \check{\theta}_0, \hat{\eta}_0)]\| \leq \|E_n[\psi(W; \theta_0, \hat{\eta}_0)]\| + \epsilon_N,$$

which in turn implies via the triangle inequality that, with P_N-probability $1 - o(1)$,

$$\|E_{P_N}[\psi(W; \theta, \eta_0)]|_{\theta = \check{\theta}_0}\| \leq \epsilon_N + 2\mathcal{I}_1 + 2\mathcal{I}_2, \tag{A.30}$$

where

$$\mathcal{I}_1 := \sup_{\theta \in \Theta, \eta \in \mathcal{T}_N} \|E_{P_N}[\psi(W; \theta, \eta)] - E_{P_N}[\psi(W; \theta, \eta_0)]\|,$$

$$\mathcal{I}_2 := \max_{\eta \in \{\eta_0, \hat{\eta}_0\}} \sup_{\theta \in \Theta} \|E_n[\psi(W; \theta, \eta)] - E_{P_N}[\psi(W; \theta, \eta)]\|.$$

Here $\epsilon_N = o(\tau_N)$ because $\epsilon_N = o(\delta_N N^{-1/2})$, $\delta_N = o(1)$ and $\tau_N \geq c_0 N^{-1/2} \log n$. Also, $\mathcal{I}_1 = r_N \leq \delta_N \tau_N = o(\tau_N)$ by Assumption 3.4(c). Moreover, applying Lemma 6.2 to the function class $\mathcal{F}_{1,\eta}$ for $\eta = \eta_0$ and $\eta = \hat{\eta}_0$ defined in Assumption 3.4, conditional on $(W_i)_{i \in I^c}$ and I^c, so that $\hat{\eta}_0$ is fixed after conditioning, shows that with P_N-probability $1 - o(1)$,

$$I_2 \lesssim N^{-1/2}(1 + N^{-1/2+1/q} \log n) \lesssim N^{-1/2} = o(\tau_N).$$

Hence, it follows from (A.30) and Assumption 3.3 that with P_N-probability $1 - o(1)$,

$$\|J_0(\check{\theta}_0 - \theta_0)\| \wedge c_0 \leq \|E_{P_N}[\psi(W; \theta, \eta_0)]|_{\theta = \check{\theta}_0}\| = o(\tau_N). \tag{A.31}$$

Combining this bound with the fact that the singular values of J_0 are bounded away from zero, which holds by Assumption 3.3, gives the claim of this step.

STEP 2. (Linearization) Here we prove the claim of the lemma. First, by definition of $\check{\theta}_0$, we have

$$\sqrt{n}\|E_n[\psi(W;\check{\theta}_0,\hat{\eta}_0)]\| \leq \inf_{\theta\in\Theta}\sqrt{n}\|E_n[\psi(W;\theta,\hat{\eta}_0)]\| + \epsilon_N\sqrt{n}. \tag{A.32}$$

Also, it will be shown in Step 4 that

$$\begin{aligned}\mathcal{I}_3 &:= \inf_{\theta\in\Theta}\sqrt{n}\|E_n[\psi(W;\theta,\hat{\eta}_0)]\| \\ &= O_{P_N}(n^{-1/2+1/q}\log n + r'_N\log^{1/2}(1/r'_N) + \lambda_N\sqrt{n} + \lambda'_N\sqrt{n}).\end{aligned} \tag{A.33}$$

Moreover, for any $\theta\in\Theta$ and $\eta\in\mathcal{T}_N$, we have

$$\begin{aligned}\sqrt{n}E_n[\psi(W;\theta,\eta)] = \sqrt{n}E_n[\psi(W;\theta_0,\eta_0)] + \mathbb{G}_n[\psi(W;\theta,\eta)-\psi(W;\theta_0,\eta_0)] \\ +\sqrt{n}(E_{P_N}[\psi(W;\theta,\eta)),\end{aligned} \tag{A.34}$$

where we are using the fact that $E_{P_N}[\psi(W;\theta_0,\eta_0)]=0$. Finally, by Taylor expansion of the function $r\mapsto E_{P_N}[\psi(W;\theta_0+r(\theta-\theta_0),\eta_0+r(\eta-\eta_0))]$, which vanishes at $r=0$,

$$\begin{aligned}E_{P_N}[\psi(W;\theta,\eta)] = J_0(\theta-\theta_0) + \partial_\eta E_{P_N}\psi(W;\theta_0,\eta_0)[\eta-\eta_0] \\ +\int_0^1 2^{-1}\partial_r^2 E_{P_N}[W;\theta_0+r(\theta-\theta_0),\eta_0+r(\eta-\eta_0)]dr.\end{aligned} \tag{A.35}$$

Therefore, because $\|\check{\theta}_0-\theta_0\|\leq\tau_N$ and $\eta\in\mathcal{T}_N$ with P_N-probability $1-o(1)$, and because by Neyman λ_N-near orthogonality,

$$\|\partial_\eta E_{P_N}[\psi(W;\theta_0,\eta_0)][\hat{\eta}_0-\eta_0]\|\leq\lambda_N,$$

applying (A.34) with $\theta=\check{\theta}_0$ and $\eta=\hat{\eta}_0$, we have with P_N-probability $1-o(1)$,

$$\sqrt{n}\Big\|E_n[\psi(W;\theta_0,\eta_0)] + J_0(\check{\theta}_0-\theta_0)\Big\| \leq \lambda_N\sqrt{n} + \epsilon_N\sqrt{n} + \mathcal{I}_3 + \mathcal{I}_4 + \mathcal{I}_5,$$

where by Assumption 3.4,

$$\mathcal{I}_4 := \sqrt{n}\sup_{\|\theta-\theta_0\|\leq\tau_N,\eta\in\mathcal{T}_N}\Big\|\int_0^1 2^{-1}\partial_r^2 E_{P_N}[W;\theta_0+r(\theta-\theta_0),\eta_0+r(\eta-\eta_0)]dr\Big\| \leq \lambda'_N\sqrt{n},$$

and by Step 3 below, with P_N-probability $1-o(1)$,

$$\begin{aligned}\mathcal{I}_5 &:= \sup_{\|\theta-\theta_0\|\leq\tau_N}\|\mathbb{G}_n(\psi(W;\theta,\hat{\eta}_0)-\psi(W;\theta_0,\eta_0))\| \\ &\leq r'_N\log^{1/2}(1/r'_N) + n^{-1/2+1/q}\log n.\end{aligned} \tag{A.36}$$

Therefore, because all singular values of J_0 are bounded below from zero by Assumption 3.3(d), it follows that

$$\begin{aligned}\|J_0^{-1}\sqrt{n}E_n[\psi(W;\theta_0,\eta_0)] + \sqrt{n}(\check{\theta}_0-\theta_0)\| \\ = O_{P_N}(n^{-1/2+1/q}\log n + r'_N\log^{1/2}(1/r'_N) + (\epsilon_N+\lambda_N+\lambda'_N)\sqrt{n}.\end{aligned}$$

The asserted claim now follows by multiplying both parts of the display by $\Sigma_0^{-1/2}$ (under the norm on the left-hand side) and noting that singular values of Σ_0 are bounded below from zero by Assumptions 3.3 and 3.4.

STEP 3. Here we derive a bound on $\mathcal{I}_5$ in (A.36). We have

$$\mathcal{I}_5 \lesssim \sup_{f\in\mathcal{F}_2} |\mathbb{G}_n(f)|, \quad \mathcal{F}_2 = \{\psi_j(\cdot,\theta,\hat{\eta}_0) - \psi_j(\cdot,\theta_0,\eta_0) : \ j = 1,\dots,d_\theta, \ \|\theta-\theta_0\| \le \tau_n\}.$$

To bound $\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|$, we apply Lemma 6.2 conditional on $(W_i)_{i\in I^c}$ and I^c so that $\hat{\eta}_0$ can be treated as fixed. Observe that with P_N-probability $1-o(1)$, $\sup_{f\in\mathcal{F}_2}\|f\|_{P_N,2} \lesssim r'_N$ where we used Assumption 3.4. Thus, an application of Lemma 6.2 to the empirical process $\{\mathbb{G}_n(f), f\in\mathcal{F}_2\}$ with an envelope $F_2 = F_{1,\hat{\eta}_0} + F_{1,\eta_0}$ and $\sigma = Cr'_N$ for sufficiently large constant C conditional on $(W_i)_{i\in I^c}$ and I^c yields that with P_N-probability $1-o(1)$,

$$\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)| \lesssim r'_N \log^{1/2}(1/r'_N) + n^{-1/2+1/q}\log n. \tag{A.37}$$

This follows because $\|F_2\|_{P,q} = \|F_{1,\hat{\eta}_0} + F_{1,\eta_0}\|_{P,q} \le 2C_1$ by Assumption 3.4(b) and the triangle inequality, and

$$\log\sup_Q N(\epsilon\|F_2\|_{Q,2}, \mathcal{F}_2, \|\cdot\|_{Q,2}) \le 2v\log(2a/\epsilon), \quad \text{for all } 0<\epsilon\le 1,$$

because $\mathcal{F}_2 \subset \mathcal{F}_{1,\hat{\eta}_0} - \mathcal{F}_{1,\eta_0}$ for $\mathcal{F}_{1,\eta}$ defined in Assumption 3.4(b), and

$$\begin{aligned}
&\log\sup_Q N(\epsilon\|F_{1,\hat{\eta}_0} + F_{1,\eta_0}\|_{Q,2}, \mathcal{F}_{1,\hat{\eta}_0} - \mathcal{F}_{1,\eta_0}, \|\cdot\|_{Q,2}) \\
&\quad\le \log\sup_Q N((\epsilon/2)\|F_{1,\hat{\eta}_0}\|_{Q,2}, \mathcal{F}_{1,\hat{\eta}_0}, \|\cdot\|_{Q,2}) \\
&\qquad + \log\sup_Q N((\epsilon/2)\|F_{1,\eta_0}\|_{Q,2}, \mathcal{F}_{1,\eta_0}, \|\cdot\|_{Q,2})
\end{aligned}$$

by the proof of Theorem 3 in Andrews (1994b). The claim of this step follows.

STEP 4. Here we derive a bound on $\mathcal{I}_3$ in (A.33). Let $\bar{\theta}_0 = \theta_0 - J_0^{-1}E_n[\psi(W;\theta_0,\eta_0)]$. Then $\|\bar{\theta}_0 - \theta_0\| = O_{P_N}(1/\sqrt{n}) = o_{P_N}(\tau_n)$ as $E_{P_N}[\|\sqrt{n}E_n[\psi(W;\theta_0,\eta_0)]\|]$ is bounded and the singular values of J_0 are bounded below from zero by Assumption 3.3(d). Therefore, $\bar{\theta}_0 \in \Theta$ with P_N-probability $1-o(1)$ by Assumption 3.3(a). Hence, with the same probability,

$$\inf_{\theta\in\Theta}\sqrt{n}\|E_n[\psi(W;\theta,\hat{\eta}_0)]\| \le \sqrt{n}\|E_n[\psi(W;\bar{\theta}_0,\hat{\eta}_0)]\|,$$

and so it suffices to show that with P_N-probability $1-o(1)$,

$$\sqrt{n}\|E_n[\psi(W;\bar{\theta}_0,\hat{\eta}_0)]\| = O(n^{-1/2+1/q}\log n + r'_N\log^{1/2}(1/r'_N) + \lambda_N\sqrt{n} + \lambda'_N\sqrt{n}).$$

To prove it, substitute $\theta = \bar{\theta}_0$ and $\eta = \widehat{\eta}_0$ into (A.34) and use the Taylor expansion in (A.35). This shows that with P_N-probability $1-o(1)$,

$$\begin{aligned}
\sqrt{n}\|E_n[\psi(W;\bar{\theta}_0,\hat{\eta}_0)]\| &\le \sqrt{n}\|E_n[\psi(W;\theta_0,\eta_0)] + J_0(\bar{\theta}_0-\theta_0)\| + \lambda_N\sqrt{n} + \mathcal{I}_4 + \mathcal{I}_5 \\
&= \lambda_N\sqrt{n} + \mathcal{I}_4 + \mathcal{I}_5.
\end{aligned}$$

Combining this with the bounds on $\mathcal{I}_4$ and $\mathcal{I}_5$ derived above gives the claim of this step and completes the proof of the lemma. □

Proof of Theorems 4.1 and 4.2: Because Theorem 4.1 is a special case of Theorem 4.2 (with $Z = D$), it suffices to prove the latter. Also, we only consider the DML estimators based on the score (4.7) and note that the estimators based on the score (4.8) can be treated similarly.

Observe that the score ψ in (4.7) is linear in θ:

$$\begin{aligned}
\psi(W;\theta,\eta) &= (Y - D\theta - g(X))(Z - m(X)) = \psi^a(W;\eta)\theta + \psi^b(W;\eta); \\
\psi^a(W;\eta) &= D(m(X) - Z), \quad \psi^b(W;\eta) = (Y - g(X))(Z - m(X)).
\end{aligned}$$

Therefore, all asserted claims of Theorem 4.2 follow from Theorems 3.1 and 3.2 and Corollary 3.1 as long as we can verify Assumptions 3.1 and 3.2, which we do here. We do so with $\mathcal{T}_N$ being the set of all $\eta = (g, m)$ consisting of P-square-integrable functions g and m such that

$$\|\eta - \eta_0\|_{P,q} \leq C,$$
$$\|\eta - \eta_0\|_{P,2} \leq \delta_N,$$
$$\|m - m_0\|_{P,2} \times \|g - g_0\|_{P,2} \leq \delta_N N^{-1/2}.$$

Also, we replace the constant q and the sequence $(\delta_N)_{N\geq 1}$ in Assumptions 3.1 and 3.2 by $q/2$ and $(\delta'_N)_{N\geq 1}$ with $\delta'_N = (C + 2\sqrt{C} + 2)(\delta_N \vee N^{-((1-4/q)\wedge(1/2))})$ for all N (recall that we assume that $q > 4$, and the analysis in Section 3 only requires that $q > 2$; also, δ'_N satisfies $\delta'_N \geq N^{-1/2}$, which is required in Theorems 3.1 and 3.2). We proceed in five steps. All bounds in the proof hold uniformly over $P \in \mathcal{P}$ but we omit this qualifier for brevity).

STEP 1. We first verify Neyman orthogonality. We have that $E_P[\psi(W; \theta_0, \eta_0)] = 0$ by definition of θ_0 of η_0. Also, for any $\eta = (g, m) \in \mathcal{T}_N$, the Gateaux derivative in the direction $\eta - \eta_0 = (g - g_0, m - m_0)$ is given by

$$\begin{aligned}\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = {} & E_P[(g(X) - g_0(X))(m_0(X) - Z)] \\ & + E_P[(m_0(X) - m(X))(Y - D\theta_0 - g_0(X))] = 0,\end{aligned}$$

by the law of iterated expectations, as $V = Z - m_0(X)$ and $U = (Y - D\theta_0 - g_0(X))$ obey $E_P[V|X] = 0$ and $E_P[U|Z, X] = 0$. This gives Assumption 3.1(d) with $\lambda_N = 0$.

STEP 2. Note that

$$|J_0| = |E_P[\psi^a(W; \eta_0)]| = |E_P[D(m_0(X) - Z)]| = |E_P[DV]| \geq c > 0$$

by Assumption 4.2(c). In addition,

$$\begin{aligned}|E_P[\psi^a(W; \eta_0)]| &= |E_P[D(m_0(X) - Z)]| \leq \|D\|_{P,2}\|m_0(X)\|_{P,2} + \|D\|_{P,2}\|Z\|_{P,2} \\ &\leq 2\|D\|_{P,2}\|Z\|_{P,2} \leq 2\|D\|_{P,q}\|Z\|_{P,q} \leq 2C^2\end{aligned}$$

by the triangle inequality, the Hölder inequality, the Jensen inequality and Assumption 4.2(b). This gives Assumption 3.1(e). Hence, given that Assumptions 3.1(a)–(c) hold trivially, Steps 1 and 2 together show that all conditions of Assumption 3.1 hold.

STEP 3. Note that Assumption 3.2(a) holds by construction of the set $\mathcal{T}_N$ and Assumption 4.2(e). Also, note that $\psi(W; \theta_0, \eta_0) = UV$, and so

$$E_P[\psi(W; \theta_0, \eta_0)\psi(W; \theta_0, \eta_0)'] = E_P[U^2V^2] \geq c^4 > 0,$$

by Assumption 4.2(c), which gives Assumption 3.2(d).

STEP 4. Here we verify Assumption 3.2(b). For any $\eta = (g, m) \in \mathcal{T}_N$, we have

$$\begin{aligned}(E_P[\|\psi^a(W; \eta)\|^{q/2}])^{2/q} &= \|\psi^a(W; \eta)\|_{P,q/2} = \|D(m(X) - Z)\|_{P,q/2} \\ &\leq \|D(m(X) - m_0(X))\|_{P,q/2} + \|Dm_0(X)\|_{P,q/2} + \|DZ\|_{P,q/2} \\ &\leq \|D\|_{P,q}\|m(X) - m_0(X)\|_{P,q} + \|D\|_{P,q}\|m_0(X)\|_{P,q} + \|D\|_{P,q}\|Z\|_{P,q} \\ &\leq C\|D\|_{P,q} + 2\|D\|_{P,q}\|Z\|_{P,q} \leq 3C^2\end{aligned}$$

by Assumption 4.2(b), which gives the bound on m'_N in Assumption 3.2(b). Also, because

$$|E_P[(D - r_0(X))(Z - m_0(X))]| = |E_P[DV]| \geq c$$

by Assumption 4.2(c), it follows that θ_0 satisfies

$$\begin{aligned}|\theta_0| &= \frac{|E_P[(Y-\ell_0(X))(Z-m_0(X))]|}{|E_P[(D-r_0(X))(Z-m_0(X))]|}\\ &\le c^{-1}(\|Y\|_{P,2}+\|\ell_0(X)\|_{P,2})(\|Z\|_{P,2}+\|m_0(X)\|_{P,2})\\ &\le 4c^{-1}\|Y\|_{P,2}\|Z\|_{P,2}\le 4C^2/c.\end{aligned}$$

Hence,

$$\begin{aligned}(E_P[\|\psi(W;\theta_0,\eta)\|^{q/2}])^{2/q} &= \|\psi(W;\theta_0,\eta)\|_{P,q/2}\\ &= \|(Y-D\theta_0-g(X))(Z-m(X))\|_{P,q/2}\\ &\le \|U(Z-m(X))\|_{P,q/2}+\|(g(X)-g_0(X))(Z-m(X))\|_{P,q/2}\\ &\le \|U\|_{P,q}\|Z-m(X)\|_{P,q}+\|g(X)-g_0(X)\|_{P,q}\|Z-m(X)\|_{P,q}\\ &\le (\|U\|_{P,q}+C)\|Z-m(X)\|_{P,q}\\ &\le (\|Y-D\theta_0\|_{P,q}+\|g_0(X)\|_{P,q}+C)\\ &\quad\times(\|Z\|_{P,q}+\|m_0(X)\|_{P,q}+\|m(X)-m_0(X)\|_{P,q})\\ &\le (2\|Y-D\theta_0\|_{P,q}+C)(2\|Z\|_{P,q}+C)\\ &\le (2\|Y\|_{P,q}+2\|D\|_{P,q}|\theta_0|+C)(2\|Z\|_{P,q}+C)\\ &\le 3C(3C+8C^3/c),\end{aligned}$$

where we used the fact that because $g_0(X)=E_P[Y-D\theta_0\mid X]$, $\|g_0(X)\|_{P,q}\le\|Y-D\theta_0\|_{P,q}$ by the Jensen inequality. This gives the bound on m_N in Assumption 3.2(b). Hence, Assumption 3.2(b) holds.

STEP 5. Finally, we verify Assumption 3.2(c). For any $\eta=(g,m)\in\mathcal{T}_N$, we have

$$\begin{aligned}\|E_P[\psi^a(W;\eta)]-E_P[\psi^a(W;\eta_0)]\| &= |E_P[\psi^a(W;\eta)-\psi^a(W;\eta_0)]|\\ &= |E_P[D(m(X)-m_0(X))]|\\ &\le \|D\|_{P,2}\|m(X)-m_0(X)\|_{P,2}\le C\delta_N\le\delta_N',\end{aligned}$$

which gives the bound on r_N in Assumption 3.2(c). Further,

$$\begin{aligned}(E_P[\|\psi(W;\theta_0,\eta)-\psi(W;\theta_0,\eta_0)\|^2])^{1/2} &= \|\psi(W;\theta_0,\eta)-\psi(W;\theta_0,\eta_0)\|_{P,2}\\ &= \|(U+g_0(X)-g(X))(Z-m(X))-U(Z-m_0(X))\|_{P,2}\\ &\le \|U(m(X)-m_0(X))\|_{P,2}+\|(g(X)-g_0(X))(Z-m(X))\|_{P,2}\\ &\le \sqrt{C}\|m(X)-m_0(X)\|_{P,2}+\|V(g(X)-g_0(X))\|_{P,2}\\ &\quad+\|(g(X)-g_0(X))(m(X)-m_0(X))\|_{P,2}\\ &\le \sqrt{C}\|m(X)-m_0(X)\|_{P,2}+\sqrt{C}\|g(X)-g_0(X)\|_{P,2}\\ &\quad+\|g(X)-g_0(X)\|_{P,2}\times\|m(X)-m_0(X)\|_{P,2}\\ &\le (2\sqrt{C}+N^{-1/2})\delta_N\le(2\sqrt{C}+1)\delta_N\le\delta_N',\end{aligned}$$

which gives the bound on r'_N in Assumption 3.2(c). Finally, let

$$f(r) := E_P[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0)], \quad r \in (0, 1).$$

Then, for any $r \in (0, 1)$,

$$f(r) = E_P[(U - r(g(X) - g_0(X)))(V - r(m(X) - m_0(X)))],$$

and so

$$\begin{aligned}\partial f(r) &= -E_P[(g(X) - g_0(X))(V - r(m(X) - m_0(X)))] \\ &\quad -E_P[(U - r(g(X) - g_0(X)))(m(X) - m_0(X))], \\ \partial^2 f(r) &= 2E_P[(g(X) - g_0(X))(m(X) - m_0(X))].\end{aligned}$$

Hence,

$$|\partial^2 f(r)| \le 2\|g(X) - g_0(X)\|_{P,2} \times \|m(X) - m_0(X)\|_{P,2} \le 2\delta_N N^{-1/2} \le \delta'_N N^{-1/2},$$

which gives the bound on λ'_N in Assumption 3.2(c). Thus, all conditions of Assumptions 3.1 are verified. This completes the proof. □

Proof of Theorems 5.1 and 5.2: The proof of Theorem 5.2 is similar to that of Theorem 5.1 and is therefore omitted. In turn, regarding Theorem 5.1, we show the proof for the case of ATE and note that the proof for the case of ATTE is similar.

Observe that the score ψ in (5.3) is linear in θ:

$$\begin{aligned}\psi(W; \theta, \eta) &= \psi^a(W; \eta)\theta + \psi^b(W; \eta), \qquad \psi^a(W; \eta) = -1, \\ \psi^b(W; \eta) &= (g(1, X) - g(0, X)) + \frac{D(Y - g(1, X))}{m(X)} - \frac{(1 - D)(Y - g(0, X))}{1 - m(X)}.\end{aligned}$$

Therefore, all asserted claims of Theorem 5.1 follow from Theorems 3.1 and 3.2 and Corollary 3.1 as long as we can verify Assumptions 3.1 and 3.2, which we do here. We do so with $\mathcal{T}_N$ being the set of all $\eta = (g, m)$ consisting of P-square-integrable functions g and m such that

$$\begin{aligned}\|\eta - \eta_0\|_{P,q} &\le C, \\ \|\eta - \eta_0\|_{P,2} &\le \delta_N, \\ \|m - 1/2\|_{P,\infty} &\le 1/2 - \varepsilon; \\ \|m - m_0\|_{P,2} \times \|g - g_0\|_{P,2} &\le \delta_N N^{-1/2}.\end{aligned}$$

Also, we replace the sequence $(\delta_N)_{N \ge 1}$ in Assumptions 3.1 and 3.2 by $(\delta'_N)_{N \ge 1}$ with $\delta'_N = C_\varepsilon(\delta_N \vee N^{-((1-4/q)\wedge(1/2))})$ for all N, where C_ε is a sufficiently large constant that depends only on ε and C (note that δ'_N satisfies $\delta'_N \ge N^{-((1-4/q)\wedge(1/2))}$, which is required in Theorems 3.1 and 3.2). We proceed in five steps. All bounds in the proof hold uniformly over $P \in \mathcal{P}$ but we omit this qualifier for brevity.

STEP 1. We first verify Neyman orthogonality. We have that $E[\psi(W; \theta_0, \eta_0)] = 0$ by definition of θ_0 and η_0. Also, for any $\eta = (g, m) \in \mathcal{T}_N$, the Gateaux derivative in the direction $\eta - \eta_0 = (g - g_0, m - m_0)$ is given by

$$\begin{aligned}\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] &= E_P[g(1, X) - g_0(1, X)] - E_P[g(0, X) - g_0(0, X)] \\ -E_P\Big[\frac{D(g(1, X) - g_0(1, X))}{m_0(X)}\Big] &+ E_P\Big[\frac{(1 - D)(g(0, X) - g_0(0, X))}{1 - m_0(X)}\Big]\end{aligned}$$

$$-E_P\Big[\frac{D(Y-g_0(1,X))(m(X)-m_0(X))}{m_0^2(X)}\Big]$$
$$-E_P\Big[\frac{(1-D)(Y-g_0(0,X))(m(X)-m_0(X))}{(1-m_0(X))^2}\Big],$$

which is 0 by the law of iterated expectations, as

$$E_P[D \mid X] = m_0(X), \qquad E_P[1-D \mid X] = 1-m_0(X),$$
$$E_P[D(Y-g_0(1,X)) \mid X] = 0, \qquad E_P[(1-D)(Y-g_0(0,X)) \mid X] = 0.$$

This gives Assumption 3.1(d) with $\lambda_N = 0$.

STEP 2. Note that $J_0 = -1$, and so Assumption 3.1(e) holds trivially. Hence, given that Assumptions 3.1(a)–(c) hold trivially as well, Steps 1 and 2 together show that all conditions of Assumption 3.1 hold.

STEP 3. Note that Assumption 3.2(a) holds by construction of the set $\mathcal{T}_N$ and Assumption 5.1(f). Also,

$$\begin{aligned}
E_P[\psi^2(W;\theta_0,\eta_0)] &= E_P[E_P[\psi^2(W;\theta_0,\eta_0) \mid X]] \\
&= E_P\Big[E_P[(g_0(1,X)-g_0(0,X)-\theta_0)^2 \mid X] \\
&\quad + E_P\Big[\Big(\frac{D(Y-g_0(1,X))}{m_0(X)} - \frac{(1-D)(Y-g_0(0,X))}{1-m_0(X)}\Big)^2 \mid X\Big]\Big] \\
&\geq E_P\Big[\Big(\frac{D(Y-g_0(1,X))}{m_0(X)} - \frac{(1-D)(Y-g_0(0,X))}{1-m_0(X)}\Big)^2\Big] \\
&= E_P\Big[\frac{D^2(Y-g_0(1,X))^2}{m_0(X)^2} + \frac{(1-D)^2(Y-g_0(0,X))^2}{(1-m_0(X))^2}\Big] \\
&\geq \frac{1}{(1-\varepsilon)^2} E_P[D^2(Y-g_0(1,X))^2 + (1-D)^2(Y-g_0(0,X))^2] \\
&= \frac{1}{(1-\varepsilon)^2} E_P[DU^2 + (1-D)U^2] = \frac{1}{(1-\varepsilon)^2} E_P[U^2] \geq \frac{c^2}{(1-\varepsilon)^2}.
\end{aligned}$$

This gives Assumption 3.2(d).

STEP 4. Here we verify Assumption 3.2(b). We have

$$\begin{aligned}
\|g_0(D,X)\|_{P,q} &= (E_P[|g_0(D,X)|^q])^{1/q} \\
&\geq (E_P[|g_0(1,X)|^q Pr_P(D=1 \mid X) + |g_0(0,X)|^q Pr_P(D=0 \mid X)])^{1/q} \\
&\geq \varepsilon^{1/q}(E_P[|g_0(1,X)|^q] + E_P[|g_0(0,X)|^q])^{1/q} \\
&\geq \varepsilon^{1/q}(E_P[|g_0(1,X)|^q] \vee E_P[|g_0(0,X)|^q])^{1/q} \\
&\geq \varepsilon^{1/q}(\|g_0(1,X)\|_{P,q} \vee \|g_0(0,X)\|_{P,q}),
\end{aligned}$$

where in the third line, we used the facts that $Pr_P(D=1 \mid X) = m_0(X) \geq \varepsilon$ and $Pr_P(D=0 \mid X) = 1-m_0(X) \geq \varepsilon$. Hence, given that $\|g_0(D,X)\|_{P,q} \leq \|Y\|_{P,q} \leq C$ by the Jensen inequality and Assumption 5.1(b), it follows that

$$\|g_0(1,X)\|_{P,q} \leq C/\varepsilon^{1/q} \quad \text{and} \quad \|g_0(0,X)\|_{P,q} \leq C/\varepsilon^{1/q}.$$

Similarly, for any $\eta \in (g,m) \in \mathcal{T}_N$,

$$\|g(1,X)-g_0(1,X)\|_{P,q} \leq C/\varepsilon^{1/q} \quad \text{and} \quad \|g(0,X)-g_0(0,X)\|_{P,q} \leq C/\varepsilon^{1/q}$$

as $\|g(D, X) - g_0(D, X)\|_{P,q} \leq C$. In addition,

$$|\theta_0| = |E_P[g_0(1, X) - g_0(0, X)]| \leq \|g_0(1, X)\|_{P,2} + \|g_0(0, X)\|_{P,2} \leq 2C/\varepsilon^{1/q}.$$

Therefore, for any $\eta = (g, m) \in \mathcal{T}_N$, we have

$$\begin{aligned}
(E_P[|\psi(W;\theta_0,\eta)|^q])^{1/q} &= \|\psi(W;\theta_0,\eta)\|_{P,q} \\
&\leq (1+\varepsilon^{-1})\Big(\|g(1,X)\|_{P,q} + \|g(0,X)\|_{P,q}\Big) + 2\|Y\|_{P,q}/\varepsilon + |\theta_0| \\
&\leq (1+\varepsilon^{-1})\Big(\|g(1,X) - g_0(1,X)\|_{P,q} + \|g(0,X) - g_0(0,X)\|_{P,q}\Big) \\
&\quad + (1+\varepsilon^{-1})\Big(\|g_0(1,X)\|_{P,q} + \|g_0(0,X)\|_{P,q}\Big) + 2C/\varepsilon + 2C/\varepsilon^{1/q} \\
&\leq 4C(1+\varepsilon^{-1})/\varepsilon^{1/q} + 2C/\varepsilon + 2C/\varepsilon^{1/q}.
\end{aligned}$$

This gives the bound on m_N in Assumption 3.2(b). Also, we have

$$(E_P[|\psi^a(W;\eta)|^q])^{1/q} = 1.$$

This gives the bound on m'_N in Assumption 3.2(b). Hence, Assumption 3.2(b) holds.

STEP 5. Finally, we verify Assumption 3.2(c). For any $\eta = (g, m) \in \mathcal{T}_N$, we have

$$\|E_P[\psi^a(W;\eta) - \psi^a(W;\eta_0)]\| = |1 - 1| = 0 \leq \delta'_N,$$

which gives the bound on r_N in Assumption 3.2(c). Further, by the triangle inequality,

$$\begin{aligned}
(E_P[\|\psi(W;\theta_0,\eta) - \psi(W;\theta_0,\eta_0)\|^2])^{1/2} &= \|\psi(W;\theta_0,\eta) - \psi(W;\theta_0;\eta_0)\|_{P,2} \\
&\leq \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3,
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{I}_1 &:= \|g(1,X) - g_0(1,X)\|_{P,2} + \|g(0,X) - g_0(0,X)\|_{P,2}, \\
\mathcal{I}_2 &:= \Big\| \frac{D(Y - g(1,X))}{m(X)} - \frac{D(Y - g_0(1,X))}{m_0(X)} \Big\|_{P,2}, \\
\mathcal{I}_3 &:= \Big\| \frac{(1-D)(Y - g(0,X))}{1 - m(X)} - \frac{(1-D)(Y - g_0(0,X))}{1 - m_0(X)} \Big\|_{P,2}.
\end{aligned}$$

To bound $\mathcal{I}_1$, note that by the same argument as that used in Step 4,

$$\|g(1,X) - g_0(1,X)\|_{P,2} \leq \delta_N/\varepsilon^{1/2} \quad \text{and} \quad \|g(0,X) - g_0(0,X)\|_{P,2} \leq \delta_N/\varepsilon^{1/2}, \quad \text{(A.38)}$$

and so $\mathcal{I}_1 \leq 2\delta_N/\varepsilon^{1/2}$. To bound $\mathcal{I}_2$, we have

$$\begin{aligned}
\mathcal{I}_2 &\leq \varepsilon^{-2}\|Dm_0(X)(Y - g(1,X)) - Dm(X)(Y - g_0(1,X))\|_{P,2} \\
&\leq \varepsilon^{-2}\|m_0(X)(g_0(1,X) + U - g(1,X)) - m(X)U\|_{P,2} \\
&\leq \varepsilon^{-2}(\|m_0(X)(g(1,X) - g_0(1,X))\|_{P,2} + \|(m(X) - m_0(X))U\|_{P,2}) \\
&\leq \varepsilon^{-2}(\|g(1,X) - g_0(1,X)\|_{P,2} + \sqrt{C}\|m(X) - m_0(X)\|_{P,2}) \\
&\leq \varepsilon^{-2}(\varepsilon^{-1/2} + \sqrt{C})\delta_N.
\end{aligned}$$

Here, the first inequality follows from the bounds $\varepsilon \leq m_0(X) \leq 1 - \varepsilon$ and $\varepsilon \leq m(X) \leq 1 - \varepsilon$, the second from the facts that $D \in \{0, 1\}$ and for $D = 1$, $Y = g_0(1, X) + U$, the third from the

triangle inequality, the fourth from the facts that $m_0(X) \leq 1$ and $E_P[U^2 \mid X] \leq C$, and the fifth from (A.38). Similarly, $\mathcal{I}_3 \leq \varepsilon^{-2}(\varepsilon^{-1/2} + \sqrt{C})\delta_N$. Combining these inequalities shows that

$$(E_P[\|\psi(W;\theta_0,\eta) - \psi(W;\theta_0,\eta_0)\|^2])^{1/2} \leq 2(\varepsilon^{-1/2} + \varepsilon^{-5/2} + \sqrt{C}\varepsilon^{-2})\delta_N \leq \delta'_N,$$

as long as C_ε in the definition of δ'_N satisfies $C_\varepsilon \geq 2(\varepsilon^{-1/2} + \varepsilon^{-5/2} + \sqrt{C}\varepsilon^{-2})$. This gives the bound on r'_N in Assumption 3.2(c).
Finally, let

$$f(r) := E_P[\psi(W;\theta_0,\eta_0 + r(\eta - \eta_0))], \quad r \in (0,1).$$

Then for any $r \in (0,1)$,

$$\begin{aligned}
\partial^2 f(r) = {} & E_P\Big[\frac{D(g(1,X) - g_0(1,X))(m(X) - m_0(X))}{(m_0(X) + r(m(X) - m_0(X)))^2}\Big] \\
& + E_P\Big[\frac{(1-D)(g(0,X) - g_0(0,X))(m(X) - m_0(X))}{(1 - m_0(X) - r(m(X) - m_0(X)))^2}\Big] \\
& + E_P\Big[\frac{(g(1,X) - g_0(1,X))(m(X) - m_0(X))}{(m_0(X) + r(m(X) - m_0(X)))^2}\Big] \\
& + 2E_P\Big[\frac{D(Y - g_0(1,X) - r(g(1,X) - g_0(1,X)))(m(X) - m_0(X))^2}{(m_0(X) + r(m(X) - m_0(X)))^3}\Big] \\
& + E_P\Big[\frac{(g(0,X) - g_0(0,X))(m(X) - m_0(X))}{(1 - m_0(X) - r(m(X) - m_0(X)))^2}\Big] \\
& - 2E_P\Big[\frac{(1-D)(Y - g_0(0,X) - r(g(0,X) - g_0(0,X)))(m(X) - m_0(X))^2}{(1 - m_0(X) - r(m(X) - m_0(X)))^3}\Big],
\end{aligned}$$

and so, given that

$$\begin{aligned}
D(Y - g_0(1,X)) &= DU, \qquad (1-D)(Y - g_0(0,X)) = (1-D)U, \\
E_P[U \mid D, X] &= 0, \qquad |m(X) - m_0(X)| \leq 2,
\end{aligned}$$

it follows that for some constant C'_ε that depends only on ε and C,

$$|\partial^2 f(r)| \leq C'_\varepsilon \|m - m_0\|_{P,2} \times \|g - g_0\|_{P,2} \leq \delta'_N N^{-1/2},$$

as long as the constant C_ε in the definition of δ'_N satisfies $C_\varepsilon \geq C'_\varepsilon$. This gives the bound on λ'_N in Assumption 3.2(c). Thus, all conditions of Assumptions 3.1 are verified. This completes the proof. □

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication files