Locally Robust Semiparametric Estimation[*]

Victor Chernozhukov
MIT

Juan Carlos Escanciano
Universidad Carlos III de Madrid

Hidehiko Ichimura
University of Arizona and University of Tokyo

Whitney K. Newey
MIT and NBER

James M. Robins
Harvard University

November 2021

Abstract

Many economic and causal parameters depend on nonparametric or high dimensional first steps. We give a general construction of locally robust/orthogonal moment functions for GMM, where first steps have no effect, locally, on average moment functions. Using these orthogonal moments reduces model selection and regularization bias, as is important in many applications, especially for machine learning first steps. Also, associated standard errors are robust to misspecification when there is the same number of moment functions as parameters of interest.

We use these orthogonal moments and cross-fitting to construct debiased machine learning estimators of functions of high dimensional conditional quantiles and of dynamic discrete choice parameters with high dimensional state variables. We show that additional first steps needed for the orthogonal moment functions have no effect, globally, on average orthogonal moment functions. We give a general approach to estimating those additional first steps. We characterize double robustness and give a variety of new doubly robust moment functions. We give general and simple regularity conditions for asymptotic theory.

Keywords: Local robustness, orthogonal moments, double robustness, semiparametric estimation, bias, GMM. JEL Classification: C13; C14; C21; D24

[*] Escanciano acknowledges research support by Spanish grant PGC 2018-096732-B-100 (MCI/AEI/FEDER, UE) and Newey by NSF Grant 1757140. Helpful comments were provided by a co-editor, three referees, M. Cattaneo, X. Chen, B. Deaner, J. Hahn, M. Jansson, S. Klosin, Z. Liao, O. Linton, R. Moon, A. Pakes, A. de Paula, P. Phillips, V. Semenova, Y. Zhao and participants in seminars at Cambridge, Columbia, Cornell, Harvard-MIT, UCL, USC, Yale, and Xiamen. B. Deaner provided superb research assistance.

1 Introduction

Many economic and causal parameters depend on nonparametric or high dimensional first steps, denoted as γ, such as conditional expectations. Examples include dynamic discrete choice, games, average exact consumer surplus, and treatment effects. This paper shows how to construct moment functions for GMM estimators that are locally robust, referred to henceforth as orthogonal. Orthogonal moment functions are those where γ has no local effect on average moment functions, i.e. the average moment functions are insensitive to small perturbations in γ around the probability limit (plim) γ_0 of an estimator $\hat{\gamma}$. We show that such moment functions can be constructed by adding to identifying moment functions the first step influence function (FSIF) that gives the effect of γ on average identifying moment functions under general misspecification. Such orthogonal moments reduce bias from estimation of γ and lead to standard errors for parameters of interest that are robust to misspecification when there is the same number of moment functions as parameters. In constructing sample moments we also cross-fit, a form of sample splitting where the moment function for each observation is evaluated at $\hat{\gamma}$ that only use other observations, which further reduces bias. A GMM estimator based on orthogonal moment functions with cross-fitting is referred to here as debiased GMM.

Debiased GMM has several advantages over plug-in GMM where only the identifying moment functions are used. First, standard confidence intervals for debiased GMM are valid under local alternatives when $\hat{\gamma}$ uses model selection, as in Chernozhukov, Hansen, and Spindler (2015), but for plug-in GMM are typically biased and invalid, similar to Leeb and Potscher (2005). Second, for a regularized first step $\hat{\gamma}$ debiased GMM is typically much less biased than plug-in GMM. Thus, debiased GMM is preferred over plug-in GMM in the many applications where $\hat{\gamma}$ involves model selection and/or regularization. Third, orthogonal moment functions will be doubly robust when they are linear or affine in γ, meaning that average moments do not depend on γ in those cases. We give this double robustness characterization and use it to derive new classes of doubly robust moment functions. Fourth, in important settings debiased GMM has faster remainder rates than plug-in GMM, e.g. Newey and Robins (2017). In addition, regularity conditions for debiased GMM are general and simple relative to those for plug-in GMM. We show asymptotic normality for debiased GMM for any $\hat{\gamma}$ where certain mean square consistency conditions hold and either one (under double robustness) or two (more generally) mean-square rates hold and that these conditions are generally not sufficient for plug-in GMM. Debiased GMM is computationally more complicated than plug-in GMM in requiring estimation of a first step α in addition to γ, although estimation of α is required anyway for standard errors for plug-in GMM, as in Newey (1994a).

Machine learning is useful for estimating economic and causal models where there are high dimensional covariates or state variables, e.g. as in Belloni et al. (2012), Robins et al. (2013), Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), Belloni et al. (2017), and Athey, Imbens, and Wager (2018). Machine learning methods that are useful for these purposes include Lasso, Dantzig, boosting, neural nets, random forests, and others. Orthogonal moment functions reduce model selection and/or regularization biases which are common for machine learning first steps. Cross-fitting for debiased GMM also reduces bias and avoids the need for Donsker conditions, which are not known to hold for many machine learning first steps. The large sample theory given here imposes only mean-square convergence properties which will hold for a variety of machine learning first steps. The advantages of debiased GMM make it preferred to plug-in GMM for many machine learning first steps.

The orthogonal moment functions we give open the way to debiased GMM estimation of many objects of interest. We illustrate by constructing debiased GMM estimators for functionals of quantile regressions and for parameters of dynamic discrete choice models. The quantile regression estimator allows for high dimensional regressors. The dynamic discrete choice estimator is based on machine learners of conditional choice probabilities allowing for high dimensional state variables. The estimator incorporates a novel Lasso estimator where the left-hand side variable is a function of a machine learner. The estimator and the results we give provide a prototype for using machine learning for dynamic structural models. The relationship of the orthogonal moments in this paper to previous literature is discussed in Section 4.

Debiased GMM is more robust to the additional first step α than previously recognized in the literature. We show that the average of orthogonal moment functions does not depend on α when γ is equal to the plim γ_0 of $\hat{\gamma}$. Consequently, estimators of α are not required to converge faster than $n^{-1/4}$, where n is the sample size. We also give automatic estimators of α that use orthogonality in their construction and generalize Chernozhukov, Newey, and Robins (2018) and Chernozhukov, Newey, and Singh (2018) to general moment functions and a larger set of first steps.

Doubly robust moment functions have been constructed by Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (2001), Graham (2011), and Firpo and Rothe (2019). This paper innovates by characterizing double robustness and deriving large classes of new doubly robust moment functions, including affine functionals of least squares regressions and other first steps obtained from orthogonality of residuals and instrumental variables.

Targeted maximum likelihood, Van der Laan and Rubin (2006), based on machine learners has been considered by Van der Laan and Rose (2011). Here we focus on debiased GMM.

Recent work on debiased machine learning by Chernozhukov et al. (2018), Chernozhukov, Newey, and Robins (2018), and Chernozhukov, Newey, and Singh (2018) is partly based on and is also generalized by this paper. The construction of orthogonal moments given here was described in Chernozhukov et al. (2018), which cited this paper for that construction and contains no results from this paper. The asymptotic theory in this paper uses the orthogonal moment construction here to improve on the asymptotic theory of Chernozhukov et al. (2018), as described in Section 6. The doubly robust moment conditions considered in Chernozhukov, Newey, and Robins (2018) and Chernozhukov, Newey, and Singh (2018) were derived in the first version of this paper, Chernozhukov et al. (2016), and the asymptotic theory in those other papers uses theory given in this paper. The automatic machine learner given here for the additional unknown functions α generalizes that in Chernozhukov, Newey, and Singh (2018). In addition Newey and Robins (2017) and Hirshberg and Wager (2018) are concerned with linear functions of a regression that are formulated here. Furthermore, Bonhomme and Weidner (2018) have shown the importance of orthogonal moment functions in specification analysis, Foster and Srygkanis (2019) in deriving rates of convergence for machine learners, Semenova (2019) for machine learning for partially identified models, and Singh and Sun (2019) for machine learning of complier effects

There are other biases arising from nonlinearity of moment conditions in the first step $\hat{\gamma}$. Cattaneo and Jansson (2018) and Cattaneo, Jansson, and Ma (2018) give useful bootstrap and jackknife methods that reduce nonlinearity bias. Newey and Robins (2017) show that one can also remove this bias by cross fitting in some settings. We use cross-fitting in this paper.

Section 2 describes orthogonal moment functions and debiased GMM. Section 3 gives the quantile and dynamic discrete choice examples. Section 4 shows and discusses orthogonality. Section 5 gives novel classes of doubly robust moment functions and characterizes double robustness. Section 6 provides general and simple asymptotic theory for debiased GMM. Proofs and precise bias results for plug-in estimators are given in Appendices.

2 Debiased GMM

This Section describes the orthogonal moment functions, cross-fitting, and automatic estimators of α we use for debiased GMM. We begin with an example.

Example 1: An illustrative example has data observation $W = (Y, X, Z)$ and parameter of interest $\theta_0 = E[Z\gamma_0(X)] = E[\alpha_0(X)\gamma_0(X)]$ for $\gamma_0(X) = E[Y|X]$ and $\alpha_0(X) = E[Z|X]$. This example is of interest in its own right as the component of the expected conditional covariance $E[Cov(Z, Y|X)] = E[ZY] - \theta_0$ that depends on unknown functions. This covariance is useful for the analysis of covariance and for estimation of a partially linear model, Robinson (1988). We specify the identifying moment function as $z\gamma(x) - \theta$ and the plim of $\hat{\gamma}$ to be $E[Y|X]$, so that $\hat{\gamma}$ is a nonparametric regression estimator. Model selection and/or regularization of $\hat{\gamma}$ will typically lead to large biases in a plug-in estimator $\tilde{\theta} = \sum_{i=1}^{n} Z_i\hat{\gamma}(X_i)/n$ as previously mentioned, for n data observations W_i. An orthogonal moment function can be constructed by adding the FSIF, which for the identifying moment function $z\gamma(x) - \theta$ was shown to be $\alpha(x)[y - \gamma(x)]$ in Proposition 4 of Newey (1994a). The orthogonal moment function is the sum of the identifying moment function and the FSIF, given by

$$\psi(w,\gamma,\alpha,\theta):=z\gamma(x)-\theta+\alpha(x)[y-\gamma(x)].$$

In this example it follows by iterated expectations and subtracting and adding θ_0 that

$$\begin{aligned}E[\psi(W,\gamma,\alpha,\theta)] &= E[\alpha_0(X)\gamma(X)]-\theta+E[\alpha(X)\{\gamma_0(X)-\gamma(X)\}] \\ &= \theta_0-\theta-E[\{\alpha(X)-\alpha_0(X)\}\{\gamma(X)-\gamma_0(X)\}].\end{aligned} \tag{2.1}$$

Here departures of (γ,α) from (γ_0,α_0) have only a second-order effect on the average orthogonal moment function, leading to small bias from first step estimation.

2.1 Constructing Orthogonal Moment Functions

To describe debiased GMM in general let θ denote a finite dimensional parameter vector of interest, γ be the unknown first step function from the Introduction, and W a data observation with unknown cumulative distribution function (CDF) F_0. We assume that there is a vector $g(w,\gamma,\theta)$ of known functions of a possible realization w of W, γ, and θ such that

$$E[g(W,\gamma_0,\theta_0)]=0,$$

where $E[\cdot]$ is the expectation under F_0 and γ_0 is the probability limit (plim) under F_0 of a first step estimator $\hat{\gamma}$. Here we assume that θ_0 is identified by these moments, i.e. that θ_0 is the unique solution to $E[g(W,\gamma_0,\theta)]=0$ over θ in some set Θ.

The identifying moment functions $g(w,\gamma,\theta)$ and $\hat{\gamma}$ can be used to estimate the parameter of interest θ_0. Let $W_1,...,W_n$ be a sample of i.i.d. data observations. Estimated sample moment functions can be formed by plugging the first step estimator $\hat{\gamma}$ into $g(W_i,\gamma,\theta)$ and averaging over data observations to obtain $\sum_{i=1}^n g(W_i,\hat{\gamma},\theta)/n$. One could form a "plug-in" GMM estimator by minimizing a quadratic form in these estimated sample moments, but such an estimator will be highly biased by first step model selection and/or regularization as discussed in the Introduction. This bias can be reduced by using orthogonal moment functions.

The orthogonal moment functions we give are based on influence functions. To describe them we need a few additional concepts and notation. Let F denote a possible CDF for a data observation W and suppose that the plim of $\hat{\gamma}$ is $\gamma(F)$ when F is the true distribution of a data observation W. Here $\gamma(F)$ is the plim of $\hat{\gamma}$ under general misspecification, similar to Newey (1994a), where F is unrestricted except for regularity conditions such as existence of $\gamma(F)$ or the expectation of certain functions of the data. For example, if $\hat{\gamma}(x)$ is a nonparametric estimator of $E[Y|X=x]$ then $\gamma(F)(x)=E_F[Y|X=x]$ is the conditional expectation function when F is the true distribution of W, which is well defined under the regularity condition that $E_F[|Y|]$ is finite. We assume that $\gamma(F_0) = \gamma_0$, consistent with γ_0 being the plim of $\hat{\gamma}$ when F_0 is the CDF of W.

Next, let F_0 again denote the true distribution of W, H be some alternative distribution that is unrestricted except for regularity conditions, and $F_\tau = (1-\tau)F_0 + \tau H$ for $\tau \in [0,1]$. We assume that H is chosen so that $\gamma(F_\tau)$ exists for τ small enough and possibly other regularity conditions are satisfied. We make the key assumption that there exists a function $\phi(w,\gamma,\alpha,\theta)$ such that

$$\frac{d}{d\tau}E[g(W,\gamma(F_\tau),\theta)] = \int \phi(w,\gamma_0,\alpha_0,\theta)H(dw), \qquad (2.2)$$
$$E[\phi(W,\gamma_0,\alpha_0,\theta)] = 0,\ E[\phi(W,\gamma_0,\alpha_0,\theta)^2] < \infty,$$

for all H and all θ. Here α is an unknown function, additional to γ, on which only $\phi(w,\gamma,\alpha,\theta)$ depends, α_0 is the α such that equation (2.2) is satisfied, and $d/d\tau$ is the derivative from the right (i.e. for nonnegative values of τ) at $\tau = 0$. This equation is a well known Gateaux derivative characterization of the influence function $\phi(w,\gamma_0,\alpha_0,\theta)$ of the functional $\mu(F) = E[g(W,\gamma(F),\theta)]$, as in Von Mises (1947), Hampel (1974), and Huber (1981). The restriction that $\gamma(F_\tau)$ exists allows $\phi(w,\gamma_0,\alpha_0,\theta)$ to be the influence function when $\gamma(F)$ is only well defined for certain types of distributions, such as when $\gamma(F)$ is a conditional expectation or pdf. Also $\phi(w,\gamma_0,\alpha_0,\theta)$ is unique because we are not restricting H except for regularity conditions. Here γ_0 and α_0 can depend on θ, equation (2.2) is assumed to hold for each $\theta \in \Theta$, and F_0 and H do not depend on θ.

We refer to $\phi(w,\gamma,\alpha,\theta)$ as the *first step influence function* (FSIF) because it characterizes the local effect of the first step plim $\gamma(F)$ on the average moment function $\mu(F)$ as F_τ varies away from F_0 in any direction H. The FSIF can be calculated from the derivative in equation (2.2) by evaluating it where H is a point mass, e.g. as shown in Ichimura and Newey (2021), or as a solution to equation (3.9) of Newey (1994a). Equation (2.1) (and equation (3.9) of Newey, 1994a) is a strong condition with existence of $\phi(w,\gamma,\alpha,\theta)$ generally equivalent to $E[g(W,\gamma(F),\theta)]$ having a finite semiparametric variance bound. This $\phi(w,\gamma,\alpha,\theta)$ coincides with the "adjustment term" of Newey (1994a) that accounts for the nonparametric estimator $\hat{\gamma}$ in the asymptotic variance of plug-in GMM when model selection and regularization biases are small. Equation (2.2) is essentially equivalent to equation (3.9) of Newey (1994a), with the requirement that equation (2.2) be satisfied for all H in a regular, rich enough class replacing the requirement that equation (3.9) of Newey (1994a) be satisfied for all regular parametric submodels, as further explained in Ichimura and Newey (2021).

Orthogonal moment functions can be constructed by adding the FSIF to the identifying moment functions to obtain

$$\psi(W,\gamma,\alpha,\theta) = g(W,\gamma,\theta) + \phi(W,\gamma,\alpha,\theta). \qquad (2.3)$$

This vector of moment functions has two key orthogonality properties. The first property is that, for the set Γ of possible directions of departure of $\gamma(F)$ from γ_0, which we assume to be linear,

$$\frac{d}{dt}E[\psi(W,\gamma_0+t\delta,\alpha_0,\theta)]=0 \text{ for all } \delta\in\Gamma \text{ and } \theta\in\Theta, \tag{2.4}$$

where t is a scalar and the derivative is evaluated at $t=0$. Here δ represents a possible direction of deviation of $\gamma(F)$ from γ_0 and t the size of a deviation. This property means that varying γ away from γ_0 has no effect, locally, on $E[\psi(W,\gamma,\alpha_0,\theta)]$. The second property is that for the set $\mathcal{A}$ of α_0 such that equation (2.2) is satisfied for some F_0,

$$E[\phi(W,\gamma_0,\alpha,\theta)]=0 \text{ for all } \theta\in\Theta \text{ and } \alpha\in\mathcal{A}. \tag{2.5}$$

Consequently varying α will have no effect, globally, on $E[\psi(W,\gamma_0,\alpha,\theta)] = E[g(W,\gamma_0,\theta)] + E[\phi(W,\gamma_0,\alpha,\theta)] = E[g(W,\gamma_0,\theta)]$. These properties are shown in Section 4.

Example 1: (continued) Equation (2.1) gives $E[\psi(W,\gamma_0+t\delta,\alpha_0,\theta)]=\theta_0-\theta$ and iterated expectations gives $E[\phi(W,\gamma_0,\alpha,\theta)]=E[\alpha(X)\{Y-\gamma_0(X)\}]=0$, so that both of equations (2.4) and (2.5) are satisfied.

Constructing orthogonal moment functions is greatly facilitated by the wide variety of known $\phi(W,\gamma,\alpha,\theta)$. For first step least squares projections (including conditional expectations), density weighted conditional means, and their derivatives $\phi(W,\gamma,\alpha,\theta)$ is given in Newey (1994a). Hahn (1998) and Hirano, Imbens, and Ridder (2003) used those results to obtain $\phi(W,\gamma,\alpha,\theta)$ for treatment effect estimators. Bajari et al. (2010) and Bajari et al. (2009) derived $\phi(W,\gamma,\alpha,\theta)$ for some first steps used in structural estimation. Hahn and Ridder (2013, 2019) derived $\phi(W,\gamma,\alpha,\theta)$ for generated regressors that depend on first step conditional expectations. Chen and Liao (2015) derived $\phi(W,\gamma,\alpha,\theta)$ for a first step that approximately minimizes the sample average of a function of a data observation and γ. Ai and Chen (2007, p. 40) and Ichimura and Newey (2021) give $\phi(W,\gamma,\alpha,\theta)$ for first step estimators of functions satisfying conditional moment and orthogonality conditions respectively. Semenova (2018) derived $\phi(W,\gamma,\alpha,\theta)$ for support functions used in partial identification. This wide variety of known $\phi(W,\gamma,\alpha,\theta)$ can be used to construct orthogonal moment functions in many settings.

2.2 Cross-Fitting

We combine orthogonal moment functions with cross-fitting, a form of sample splitting, to construct debiased sample moments; e.g. see Bickel (1982), Schick (1986), Klaassen (1987), and Chernozhukov et al. (2018). Partition the observation indices $(i=1,...,n)$ into L groups I_ℓ, $(\ell=1,...,L)$. Consider $\hat{\gamma}_\ell$, $\hat{\alpha}_\ell$, and an initial estimator $\tilde{\theta}_\ell$ that are constructed using all observations *not* in I_ℓ. Debiased sample moment functions are

$$\hat{\psi}(\theta)=\hat{g}(\theta)+\hat{\phi},\ \hat{g}(\theta)=\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}g(W_i,\hat{\gamma}_\ell,\theta),\ \hat{\phi}=\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\phi(W_i,\hat{\gamma}_\ell,\hat{\alpha}_\ell,\tilde{\theta}_\ell). \tag{2.6}$$

Large sample properties are given in Section 6 for fixed L. The choice of $L=5$ works well based on a variety of empirical examples and in simulations, for medium sized data sets of a few thousand observations; see Chernozhukov et al. (2018). The choice $L=10$ works well for small data sets with the larger L providing more observations for construction of $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$.

The cross-fitting used here, where $\hat{\psi}(\theta)$ is an average over observations not used to form $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$, eliminates bias due to averaging over observations that are used to construct the first step. Eliminating such "own observation" bias helps remainders converge faster to zero, e.g. as in Newey and Robins (2017), and can be important in practice, e.g. as in the jackknife instrumental variables estimators of Angrist and Krueger (1995) and Blomquist and Dahlberg (1999). It also eliminates the need for Donsker conditions for $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$, which is important for many machine learning first steps that are not known to satisfy such conditions, as discussed in Chernozhukov et al. (2018).

A debiased GMM estimator is

$$\hat{\theta}=\arg\min_{\theta\in\Theta}\hat{\psi}(\theta)'\hat{\Upsilon}\hat{\psi}(\theta), \tag{2.7}$$

where $\hat{\Upsilon}$ is a positive semi-definite weighting matrix and Θ is the set of parameter values. A choice of $\hat{\Upsilon}$ that minimizes the asymptotic variance of $\hat{\theta}$ will be $\hat{\Upsilon}=\hat{\Psi}^{-1}$, for

$$\hat{\Psi}=\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}\hat{\psi}_{i\ell}',\ \hat{\psi}_{i\ell}=g(W_i,\hat{\gamma}_\ell,\tilde{\theta}_\ell)+\phi(W_i,\hat{\gamma}_\ell,\hat{\alpha}_\ell,\tilde{\theta}_\ell).$$

There is no need to account for the presence of $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ in $\hat{\psi}_{i\ell}$ because of the orthogonality of $\psi(w,\gamma,\alpha,\theta)$. An estimator $\hat{V}$ of the asymptotic variance of $\sqrt{n}(\hat{\theta}-\theta_0)$ is

$$\hat{V}=(\hat{G}'\hat{\Upsilon}\hat{G})^{-1}\hat{G}'\hat{\Upsilon}\hat{\Psi}\hat{\Upsilon}\hat{G}(\hat{G}'\hat{\Upsilon}\hat{G})^{-1},\ \hat{G}=\frac{\partial\hat{g}(\hat{\theta})}{\partial\theta}. \tag{2.8}$$

The initial estimator $\tilde{\theta}_\ell$ can be based on only the identifying moment conditions and constructed as

$$\tilde{\theta}_\ell=\arg\min_{\theta\in\Theta}\hat{g}_\ell(\theta)'\hat{\Upsilon}_\ell\hat{g}_\ell(\theta),\ \hat{g}_\ell(\theta)=\frac{1}{n-n_\ell}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}g(W_i,\tilde{\gamma}_{\ell\ell'},\theta),$$

where $\hat{\Upsilon}_\ell$ uses only observations not in I_ℓ, n_ℓ is the number of observations in I_ℓ, and $\tilde{\gamma}_{\ell\ell'}$ uses observations not in I_ℓ and not in $I_{\ell'}$. One could iterate on the initial estimator $\tilde{\theta}_\ell$ in the debiased moments $\hat{\psi}(\theta)$ and/or $\hat{\Psi}$ by calculating $\hat{\theta}$ and/or $\hat{\Psi}$ a second time with $\tilde{\theta}_\ell$ being a debiased GMM estimator obtained from a prior iteration. Because of the orthogonality condition (2.5) the use of $\tilde{\theta}_\ell$ in constructing $\hat{\phi}$ will not affect the asymptotic distribution of $\hat{\theta}$.

Example 1: (continued) Here $\psi(w, \gamma, \theta) = z\gamma(x) - \theta + \alpha(x)[y - \gamma(x)]$. Forming $\hat{\psi}(\theta)$ as we have described and solving $\hat{\psi}(\hat{\theta}) = 0$ for $\hat{\theta}$ gives the debiased GMM estimator

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{Z_i \hat{\gamma}_\ell(X_i) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\} .$$

The efficiency of debiased GMM is entirely determined by the choice of moment functions, first step, and weighting matrix. The matrix $\hat{\Psi}^{-1}$ is an optimal choice of weighting matrix as usual for GMM. The presence of $\hat{\phi}$ in the orthogonal moment functions $\hat{\psi}(\theta)$ does not affect identification of θ. The FSIF has mean zero for all possible distributions of W so that $\hat{\phi}$ will converge in probability to zero. The sole purpose of including $\hat{\phi}$ is to remove the local effect of $\hat{\gamma}_\ell$ on average moment functions.

2.3 Automatic Estimation of α_0

The debiased moments require a first step estimator $\hat{\alpha}_\ell$ with plim α_0. When the form of α_0 is known one can plug-in nonparametric estimators of unknown components of α_0 to form $\hat{\alpha}_\ell$. We can also use the orthogonality of $\psi(w, \gamma, \alpha_0, \theta)$ with respect to γ in equation (2.4) to construct estimators of α_0 without knowing the form of α_0. This approach is "automatic" in only requiring the orthogonal moment function $\psi(W, \gamma, \alpha, \theta)$ and data for construction of $\hat{\alpha}_\ell$.

Equation (2.4) can be thought of as a population moment condition for α_0 for each δ. We can form a corresponding sample moment function by replacing the expectation by a sample average and γ_0 and θ_0 by cross-fit estimators to obtain

$$\hat{\psi}_\gamma(\delta, \alpha) = \frac{d}{dt} \frac{1}{n - n_\ell} \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} \psi(W_i, \tilde{\gamma}_{\ell\ell'} + t\delta, \alpha, \tilde{\theta}_{\ell\ell'}) \Bigg|_{t=0}, \ \delta \in \Gamma, \tag{2.9}$$

where $\tilde{\gamma}_{\ell\ell'}$ and $\tilde{\theta}_{\ell\ell'}$ do not depend on observations in I_ℓ or $I_{\ell'}$ and we assume that $\psi(W_i, \tilde{\gamma}_{\ell\ell'} + t\delta, \alpha, \tilde{\theta}_{\ell\ell'})$ is differentiable in t. We can then replace α by a sieve (i.e. parametric approximation) and estimate the sieve parameters using these sample moments for a variety of choices of δ. We can also regularize to allow for a high dimensional specification for α. The sample moments in equation (2.9) depend only on observations not in I_ℓ so that the resulting $\hat{\alpha}_\ell$ will also, as required for the cross-fitting in debiased GMM.

Example 1: (continued) Here α_0 is a function of X that has finite second moment and

$$\begin{aligned}\hat{\psi}_\gamma(\delta,\alpha) &= \frac{d}{dt}\frac{1}{n-n_\ell}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}\left\{Z_i[\hat{\gamma}_{\ell\ell'}(X_i)+t\delta(X_i)]+\alpha(X_i)[Y_i-\hat{\gamma}_{\ell\ell'}(X_i)-t\delta(X_i)]-\tilde{\theta}_{\ell\ell'}\right\}\Bigg|_{t=0}\\ &= \frac{1}{n-n_\ell}\sum_{i\notin I_\ell}[Z_i-\alpha(X_i)]\,\delta(X_i).\end{aligned}$$

This is a sample moment corresponding to the population moment condition $E[\{Z-\alpha_0(X)\}\,\delta(X)] = 0$, which holds by $\alpha_0(X) = E[Z|X]$. If $\alpha(X)$ was replaced by a linear combination $\rho' b(x)$ of a dictionary $b(x) = (b_1(x), ..., b_p(x))'$ of functions and $\delta(X)$ chosen to be one element $b_j(X)$ of the dictionary then the sample moment function is

$$\hat{\psi}_\gamma(b_j,\rho' b) = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell}[Z_i-\rho' b(X_i)]\,b_j(X_i).$$

The collection of sample moment conditions $-\hat{\psi}_\gamma(b_j,\rho' b) = 0$ $(j = 1, ..., p)$ are the first order conditions for minimizing the least squares objective function for the regression of Z_i on $b(X_i)$. Adding an $L1$ penalty to this objective function and minimizing leads to the Lasso least squares estimator

$$\hat{\alpha}_\ell(x) = \hat{\rho}' b(x), \hat{\rho} = \arg\min_\rho \sum_{i\notin I_\ell}[Z_i-\rho' b(X_i)]^2/2 + r\sum_{j=1}^p|\rho_j|$$

The construction of $\hat{\alpha}_\ell$ in Example 1 can be generalized to a wide class of regression settings that will be useful for many additional examples. This generalization builds on Example 1 to enable estimation of more complicated objects such as a functional of quantile regression in Example 2 below. We generalize γ from being a conditional expectation to an object γ that is restricted to be an element of a linear set Γ. In Example 1 where the conditional expectation is an unknown function of X the Γ is all functions of X with finite second moment. Instead Γ could, for example, be restricted to be a linear combination of a sequence $(b_1(X), b_2(X), ...)$, corresponding a high dimensional regression. We also generalize $\gamma(F)$ from a conditional expectation to a function satisfying an orthogonality condition. For a scalar residual $\lambda(w,\gamma(x))$ we consider $\gamma(F) \in \Gamma$ satisfying

$$E_F[\delta(X)\lambda(W,\gamma(F)(X))] = 0 \text{ for all } \delta\in\Gamma. \tag{2.10}$$

The $\gamma(F)(X) = E_F[Y|X]$ of Example 1 satisfies this equation for $\lambda(W,\gamma) = Y-\gamma(X)$ and Γ equal to all functions of X with finite second moment. A high dimensional quantile regression in Example 2 will satisfy this equation for $\lambda(w,\gamma(x)) = 1(y<\gamma(x))-\zeta$ for $0<\zeta<1$ and Γ equal to a linear combinations of $(b_1(X), b_2(X), ...)$. The FSIF for $\gamma(F)$ in equation (2.10) is

$$\phi(w,\gamma,\alpha,\theta) = \alpha(x,\theta)\lambda(w,\gamma(x)),$$

as in Ai and Chen (2007, p. 40) for conditional moments when Γ is unrestricted and Ichimura and Newey (2021) in general, where the formula for $\alpha(x,\theta)$ is given.

Example 1 can be further generalized to consider a general vector of identifying moment functions other than $z\gamma(x)-\theta$. Here there will be one $\alpha(x,\theta)$ for each component of $g(w,\gamma,\theta)$. Let $(b_1(X), b_2(X), ...)$ "span" Γ, meaning any $\delta \in \Gamma$ can be well approximated in mean-square by a finite linear combination of $b_j(X)$. We will describe an estimator for the k^{th} component $\alpha_k(x,\theta)$ of $\alpha(x,\theta)$ corresponding to the k^{th} component $g_k(w,\gamma,\theta)$ of $g(w,\gamma,\theta)$. Let $\hat{\lambda}_{i\gamma} = d\lambda(W_i, \hat{\gamma}_{\ell\ell'}(X_i)+t)/dt|_{t=0}$ for $i \in I_{\ell'}$ and let e_j denote the j^{th} column of a $p-$dimensional identity matrix. Then for $b(X) = (b_1(X), ..., b_p(X))'$,

$$\hat{\psi}_{k\gamma}(b_j, \rho' b) = \hat{M}_{jk\ell} + e_j'\hat{Q}_\ell \rho, \;\; \hat{M}_{jk\ell} = \frac{1}{n-n_\ell}\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} \frac{d}{dt} g_k(W_i, \hat{\gamma}_{\ell\ell'} + tb_j, \tilde{\theta}_{\ell\ell'}),$$
$$\hat{Q}_\ell = \frac{1}{n-n_\ell}\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} \hat{\lambda}_{\gamma i} b(X_i) b(X_i)',$$

corresponding to the k^{th} orthogonal moment function $\psi_k(w,\gamma,\alpha) = g_k(w,\gamma,\theta) + \alpha_k(x,\theta)\lambda(w,\gamma(x))$ and the j^{th} component b_j of $b(X)$. Let $\hat{M}_{k\ell} = (\hat{M}_{1k\ell}, ..., \hat{M}_{pk\ell})'$ so that

$$\hat{\psi}_{k\gamma}(b_j, \rho' b) = \frac{\partial}{\partial \rho_j}\{\hat{M}_{k\ell}'\rho + \rho'\hat{Q}_\ell\rho/2\}.$$

The collection of sample moment conditions $-\hat{\psi}_{k\gamma}(b_j, \rho' b) = 0$, $(j = 1, ..., p)$ are the first order conditions for minimizing $-\hat{M}_{k\ell}'\rho - \rho'\hat{Q}_\ell\rho/2$. Here we assume that we can normalize so that $v_\rho(X) := dE[\lambda(W, \gamma_0(X)+t)|X]/dt|_{t=0} < 0$ so that $-\hat{Q}_\ell$ is positive semi-definite asymptotically. Adding an $L1$ penalty to this objective function and minimizing leads to the Lasso minimum distance estimator

$$\hat{\alpha}_{k\ell}(x) = \hat{\rho}_{k\ell}' b(x), \;\; \hat{\rho}_{k\ell} = \arg\min_\rho \left\{ -\hat{M}_{k\ell}'\rho - \rho'\hat{Q}_\ell\rho/2 + r\sum_{j=1}^p |\rho_j| \right\}. \tag{2.11}$$

As usual for Lasso we assume that each element $b_j(x)$ of the dictionary has been standardized to have standard deviation 1. One choice of r would be $\hat{r} = \arg\min_r \sum_{\ell=1}^L \{-\hat{M}_{k\ell}'\hat{\rho}_{k\ell}^r - \hat{\rho}_{k\ell}'^r \hat{Q}_\ell \hat{\rho}_{k\ell}^r/2\}$ that minimizes a cross-validation criteria where $\hat{\rho}_{k\ell}^r$ is from equation (2.11) for a given r and the minimization is over a grid of r values.

This estimator $\hat{\alpha}_{k\ell}(x)$ of $\alpha_{k0}(x,\theta_0)$ generalizes that of Chernozhukov, Newey, and Singh (2018) to allow for a residual $\lambda(w,\gamma(x))$ other than $y-\gamma(x)$ and a general moment function. The nested sample splitting used for $\hat{\gamma}_{\ell\ell'}(X_i)$ requires that the first step learner be computed for L^2 subsamples. Use of $\hat{\gamma}_\ell(x)$ as a starting value for computation of each $\hat{\gamma}_{\ell\ell'}(X_i)$ may aid in this computation. The nested cross-fitting allows for a very general first step that need only have a mean-square convergence rate.

Orthogonality was used to estimate unknown components of doubly robust moment functions for the average treatment effect by Vermeulen and Vansteelandt (2015), Tan (2020), and Avagyan and Vansteelandt (2021) in order to obtain standard errors that are robust to misspecification. We use orthogonality here to estimate the unknown α_0 in the FSIF for a general identifying moment function and first step. The resulting standard errors are robust to misspecification because the FSIF takes full account of the plim of $\hat{\gamma}$ under general misspecification.

It would be interesting to use the moment functions (2.9) to construct $\hat{\alpha}$ for first steps other than the generalized linear regression $\gamma(F)$ in equation (2.10). That is beyond the scope of this paper and is reserved to future work, including identification of α_0 and asymptotic theory for $\hat{\alpha}$.

This approach to estimating the FSIF uses its form $\phi(w, \gamma, \alpha, \theta)$ to construct an estimator of α_0. This approach is parsimonious in estimating only unknown parts of $\phi(w, \gamma, \alpha, \theta)$ rather than the whole function. It is also possible to estimate the entire FSIF using just the first step and the identifying moments. Such estimators are available for first step series and kernel estimation. For first step series estimation an estimator of $\phi(w, \gamma, \alpha, \theta)$ can be constructed by treating the first step estimator as if it were parametric and applying a standard formula for parametric two-step estimators, e.g. as in Newey (1994a), Ackerberg, Chen, and Hahn (2012), and Chen and Liao (2015). For parametric maximum likelihood the resulting orthogonal moment functions are the basis of Neyman's (1959) C-alpha test. For first step kernel estimation one can use the numerical influence function estimator of Newey (1994b) to estimate $\phi(w, \gamma, \alpha, \theta)$, as suggested in a previous version of this paper and shown to work in a low dimensional nonparametric setting in Bravo, Escanciano, and van Keilegom (2020). The idea is to differentiate with respect to the effect of the i^{th} observation on sample moments. Kernel estimators do not seem well suited to high dimensional settings so we do not consider them here.

It is also possible to estimate the FSIF using a numerical derivative version of equation (2.2). This approach has been given in Carone, Luedtke, and van der Laan (2016) and Bravo, Escanciano, and van Keilegom (2020) for construction of orthogonal moment functions. We focus here on the more parsimonious approach of estimating α_0 when there is a known form $\phi(w, \gamma, \alpha, \theta)$ for the FSIF.

3 Further Examples of Debiased GMM

In this Section we give two fully worked out examples of debiased GMM estimators, functionals of quantile regressions and structural parameters of a dynamic discrete choice model.

3.1 Example 2: Functional of a Quantile Regression

The object of interest in this example is an expected linear function of a quantile regression

$$\theta_0 = E[m(W, \gamma_0)],\ \gamma(F) = \arg\min_{\gamma \in \Gamma} E_F[v(Y - \gamma(X))],\ v(u) = [\zeta - 1(u < 0)]u,\ 0 < \zeta < 1, \quad (3.1)$$

where $m(w, \gamma)$ is a linear functional of γ, Y is a dependent variable of interest, Γ is a linear set of functions of x (such as all functions of X with finite second moment), and we assume the minimum $\gamma(F)$ exists. An example of $m(w, \gamma)$ is a weighted average derivative of γ where $m(w, \gamma) = \int \omega(x)[\partial\gamma(x)/\partial x_1]dx$ for a weight $\omega(x)$. Here the identifying moment function is $g(w, \gamma, \theta) = m(w, \gamma) - \theta$. The first order condition for this $\gamma(F)$ is equation (2.10) with $\lambda(w, \gamma(x)) = 1(y < \gamma(x)) - \zeta$, so the FSIF has the form $\phi(w, \gamma, \alpha) = \alpha(x)\lambda(y, \gamma(x))$, as in Section 2.3.

In this example the automatic estimator $\hat{\alpha}_\ell$ of equation (2.11) does not exist because $\lambda(y, \gamma(x)+t)$ is not continuous in γ and hence not differentiable. We address this complication by using kernel weighting in the construction of a $\hat{Q}_\ell$ to use in equation (2.11). Let $\hat{\gamma}_\ell(x)$ be a learner of γ_0, computed from observations not in I_ℓ, and $\hat{\gamma}_{\ell\ell'}(x)$ be computed from observations not in I_ℓ or $I_{\ell'}$. Also let $K(u)$ be a bounded, univariate kernel, with $\int K(u)du = 1$ and $\int K(u)udu = 0$, h a bandwidth, and $b(x)$ be a $p \times 1$ vector of functions of x. We specify that

$$\hat{Q}_\ell = \frac{-1}{n - n_\ell} \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} \frac{1}{h} K(\frac{Y_i - \hat{\gamma}_{\ell\ell'}(X_i)}{h}) b(X_i) b(X_i)'.$$

The role of the kernel term in this $\hat{Q}_\ell$ is to smooth over the discontinuity in $\lambda(w, \gamma(x)) = 1(y < \gamma(x)) - \zeta$ at $\gamma(x) = y$. This $\hat{Q}_\ell$ estimates $E[f(0|X)b(X)b(X)']$ where $f(0|X)$ is the conditional pdf of $Y - \gamma_0(X)$ evaluated 0.

A debiased GMM estimator of θ_0 can be formed from any learner $\hat{\gamma}_\ell$ of γ_0 as

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} [m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)\lambda(W_i, \hat{\gamma}_\ell(X_i))],\ \hat{V} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell}^2,$$

$$\hat{\alpha}_\ell(x) = b(x)'\hat{\rho}_\ell,\ \hat{\rho}_\ell = \arg\min_\rho \{-2\hat{M}_\ell'\rho - \rho'\hat{Q}_\ell\rho + 2r\sum_{j=1}^{p} |\rho_j|\}, \quad (3.2)$$

$$\hat{M}_{\ell j} = \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} m(W_i, b_j),\ \hat{M}_\ell = (\hat{M}_{\ell 1}, ..., \hat{M}_{\ell p})',$$

where $\hat{\psi}_{i\ell} = m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)\lambda(W_i, \hat{\gamma}_\ell(X_i)) - \hat{\theta}$. The asymptotic theory for this estimator is given in Section 6. That theory requires that the regularization constant r for $\hat{\alpha}_\ell$ goes to zero slower than the conventional Lasso rate of $\sqrt{\ln(p)/n}$, in order to accommodate the presence of kernel weighting and $\hat{\gamma}_{\ell\ell'}(X_i)$ in $\hat{Q}_\ell$.

Previously Chaudhuri, Doksum, and Samarov (1997) gave a plug-in kernel estimator of the average derivative of a conditional quantile and the FSIF for that functional. Ackerberg et al. (2014) also gave an expression for the FSIF for functionals of conditional quantiles other than the weighted average derivative. Example 2 innovates in giving a debiased GMM estimator for a general linear functional of a quantile regression. Also, the $\hat{\alpha}_\ell(x)$ here is a linear combination of a dictionary of functions rather than a ratio of a function of X and an estimator of the conditional pdf of $Y - \gamma_0(X)$ as was used previously for estimating asymptotic variances.

3.2 Example 3: Dynamic Discrete Choice

For dynamic discrete choice with high dimensional state variables we estimate structural parameters via learners of conditional choice probabilities. This approach replaces computation of expected value functions with nonparametric estimation as suggested in Hotz and Miller (1993). For simplicity we focus this example on binary choice, providing methods and results that will be available for the more complicated models employed more widely in practice. In particular we provide a Lasso estimator of conditional value function differences where the dependent variable is a function of estimated future choice probabilities. In Section 6 we provide convergence rate results for this estimator. We also give the FSIF for this kind of first step. This analysis provides a prototype for estimation of dynamic structural models with high dimensional state variables.

In dynamic binary choice individuals choose between two alternatives $j = 1$ and $j = 2$ in T time periods to maximize the expected present discounted value of per period utility $U_{jt} = D_j(X_t)'\theta_0 + \varepsilon_{jt}$, $(j = 1, 2; t = 1, ..., T)$, where ε_{jt} is i.i.d. with known CDF, independent of the entire history $\{X_t\}_{t=1}^{\infty}$ of a state variable vector X_t, and X_t is Markov of order 1 and stationary. The parameter of interest is θ_0. We develop an estimator that allows for high dimensional X_t.

We assume that choice 1 is a renewal choice where the conditional distribution of X_{t+1} given X_t and choice 1 does not depend on X_t. We also assume that $D_1(X_t) = (-1, 0')'$ and the first element of $D_2(X_t)$ equals 0 so that the first element in θ is a binary choice constant. Let Y_{jt} equal a dummy variable equal to 1 when choice j is made and $\gamma_{10}(X_t) = \Pr(Y_{2t} = 1|X_t)$ be the conditional choice probability of alternative 2. Also let $V(X_t)$ denote the expected value function. As in Hotz and Miller (1993) there is a known function $H(p)$ such that for $\gamma_{20}(X_t) = E[H(\gamma_{10}(X_{t+1}))|X_t, Y_{2t} = 1]$ and $\gamma_{30} = E[H(\gamma_{10}(X_{t+1}))|Y_{1t} = 1]$,

$$E[V(X_{t+1})|X_t, Y_{2t} = 1] - E[V(X_{t+1})|Y_{1t} = 1] = \gamma_{20}(X_t) - \gamma_{30}. \tag{3.3}$$

For example when ε_{1t} and ε_{2t} are independent Type I extreme value this equation is satisfied for $H(p) = .5227 - \ln(1 - p)$. Then for the CDF $\Lambda(a)$ of $\varepsilon_{t1} - \varepsilon_{t2}$, $D(X_t) = D_2(X_t) - D_1(X_t)$, and δ the discount factor the conditional choice probability for $j = 2$ is

$$\Pr(Y_{2t} = 1|X_t) = \Lambda(a(X_t, \theta_0, \gamma_{20}, \gamma_{30})),\ a(x, \theta, \gamma_2, \gamma_3) = D(x)'\theta + \delta\{\gamma_2(x) - \gamma_3\}. \tag{3.4}$$

We consider data of i.i.d. observations on individuals each followed for T time periods, that also includes the $T+1$ observation X_{T+1} of the state variables, where $W = (X_1', Y_{21}, ..., X_T', Y_{2T}, X_{T+1}')'$. An estimator of θ_0 can be obtained by constructing first step estimators $\hat{\gamma}_2(x)$ and $\hat{\gamma}_3$, substituting these estimators for γ_2 and γ_3 in $a(x, \theta, \gamma_2, \gamma_3)$ in equation (3.4), and then maximizing a binary choice log-likelihood as if $\hat{\gamma}_2(x)$ and $\hat{\gamma}_{30}$ were true. We specify as identifying moment functions the derivative of the pseudo log-likelihood associated with the binary choice probability in equation (3.4) with respect to θ,

$$g(W, \gamma, \theta) = \frac{1}{T}\sum_{t=1}^{T} D(X_t)\pi(a(X_t, \theta, \gamma_2, \gamma_3))[Y_{2t} - \Lambda(a(X_t, \theta, \gamma_2, \gamma_3))], \tag{3.5}$$

$$\pi(a) = \frac{\Lambda_a(a)}{\Lambda(a)[1 - \Lambda(a)]},\ \Lambda_a(a) = \frac{d\Lambda(a)}{da}.$$

Estimators of $\gamma_{10}(x)$, $\gamma_{20}(x)$, and γ_{30} are needed as a first step $\hat{\gamma}_\ell$ for the identifying moment function. We will consider any $\hat{\gamma}_{1\ell}(x)$ that converges sufficiently quickly in mean-square. For example $\hat{\gamma}_{1\ell}$ could be logit Lasso or a linear Lasso estimator with dependent variable Y_{2t}. We use Lasso to construct $\hat{\gamma}_{2\ell}(x)$ in order to control for estimation error that results from an estimated dependent variable. Let $\hat{\gamma}_{1\ell\ell'}(x)$ be an estimator of the conditional choice probability computed from observations not in I_ℓ or $I_{\ell'}$. Let $b(x)$ denote a $p \times 1$ dictionary of functions of the state variables x. We form $\hat{\gamma}_{2\ell}(x)$ and $\hat{\gamma}_{3\ell}$ as

$$\hat{\gamma}_{2\ell}(x) = b(x)'\hat{\beta}_{2\ell},\ \hat{\beta}_{2\ell} = \arg\min_{\beta}\{-2\hat{M}_{2\ell}'\beta - \beta'\hat{Q}_{2\ell}\beta + 2r\sum_{j=1}^{p}|\beta_j|\}, \tag{3.6}$$

$$\hat{M}_{2\ell} = \frac{1}{(n-n_\ell)T}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}\sum_{t=1}^{T} Y_{2it}b(X_{it})H\left(\hat{\gamma}_{1\ell\ell'}(X_{i,t+1})\right),\ \hat{Q}_{2\ell} = \frac{-1}{(n-n_\ell)T}\sum_{i\notin I_\ell}\sum_{t=1}^{T} Y_{2it}b(X_{it})b(X_{it})',$$

$$\hat{\gamma}_{3\ell} = \frac{1}{\hat{P}_1(n-n_\ell)T}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}\sum_{t=1}^{T} Y_{1it}H\left(\hat{\gamma}_{1\ell\ell'}(X_{i,t+1})\right),\ \hat{P}_1 = \frac{1}{(n-n_\ell)T}\sum_{i\notin I_\ell}\sum_{t=1}^{T} Y_{1it}.$$

Here $\hat{\gamma}_{2\ell}(x)$ is Lasso with left hand side variable $H\left(\hat{\gamma}_{1\ell\ell'}(X_{i,t+1})\right)$ and right-hand side variables $b(X_{it})Y_{2it}$ and $\hat{\gamma}_{3\ell}$ is a sample mean conditional on $Y_{1it} = 1$. Here $\hat{\gamma}_{2\ell}$, and $\hat{\gamma}_{3\ell}$ use all observations with $i \notin I_\ell$. The nested sample splitting in $\hat{\gamma}_{1\ell\ell'}(x)$ is useful in only requiring mean-square convergence rates for conditional choice probabilities although it is somewhat complicated. When L is of moderate size (e.g. $L = 5$) there may be many $\hat{\gamma}_{1\ell\ell'}(x)$ to compute (e.g. 25), although an estimate for a particular ℓ and ℓ' could provide a good starting value for other splits.

This example is more complicated than Examples 1 and 2 in having multiple first steps. When there are multiple first steps the FSIF is the sum of the FSIF's for each first step that is obtained while holding the other first steps fixed at their true value, as in Newey (1994a, p. 1357). Here there are three first steps, $\hat{\gamma}_{1\ell}(x)$, $\hat{\gamma}_{2\ell}(x)$, and $\hat{\gamma}_{3\ell}$. Consequently the FSIF is the sum of three terms, one term for each first step, given by

$$\begin{aligned}
\phi(W,\gamma,\alpha,\theta) &= \phi_1(W,\gamma,\alpha,\theta)+\phi_2(W,\gamma,\alpha,\theta)+\phi_3(W,\gamma,\alpha),\\
\phi_1(W,\gamma,\alpha,\theta) &= \frac{1}{T}\sum_{t=1}^{T}\alpha_1(X_t,\theta)[Y_{2t}-\gamma_1(X_t)],\ \phi_3(W,\gamma,\alpha)=\alpha_3\frac{1}{T}\sum_{t=1}^{T}Y_{2t}\{H(\gamma_1(X_{t+1}))-\gamma_3\},\\
\phi_2(W,\gamma,\alpha,\theta) &= \frac{1}{T}\sum_{t=1}^{T}\alpha_2(X_t,Y_{2t},\theta)[H(\gamma_1(X_{t+1}))-\gamma_2(X_t)].
\end{aligned}$$

Viewing each of $\hat{\gamma}_{1\ell}(x)$, $\hat{\gamma}_{2\ell}(x)$, and $\hat{\gamma}_{3\ell}$ as nonparametric regressions, the form of each $\phi_j(w,\gamma,\alpha,\theta)$ follows from Proposition 4 of Newey (1994a). The true functions α_{10}, α_{20}, and α_{30} are

$$\begin{aligned}
\alpha_{10}(x,\theta_0) &= \{E[\alpha_{20}(X_t,Y_{2t},\theta_0)|X_{t+1}=x]+\alpha_{30}E[Y_{1t}|X_{t+1}=x]\}H_p(\gamma_{10}(x)), \qquad (3.7)\\
\alpha_{20}(x,y_2,\theta_0) &= -\delta D(x)\pi(a(x))\frac{\Lambda_a(a(x))y_2}{\Lambda(a(x))},\ a(x)=a(x,\theta_0,\gamma_{20},\gamma_{30}),\\
\alpha_{30} &= -E[\alpha_{20}(X_t,Y_{2t},\theta_0)]/P_{10},\ P_{10}=E[Y_{1t}].
\end{aligned}$$

In this example we construct $\hat{\alpha}_{2\ell}$, $\hat{\alpha}_{3\ell}$, and $\hat{\alpha}_{1\ell}$ using the known form of the FSIF given here. We obtain an initial estimator $\tilde{\theta}_\ell$ from binary choice pseudo maximum likelihood over $i \notin I_\ell$, $t \le T$ with γ_2 and γ_3 replaced by $\hat{\gamma}_{2\ell}$ and $\hat{\gamma}_{3\ell}$ respectively in the choice probability formula. Also let $\hat{a}_{it}=a(X_{it},\hat{\theta}_\ell,\hat{\gamma}_{2\ell},\hat{\gamma}_{3\ell})$ and construct $\hat{\alpha}_{2\ell}(X_{it},Y_{2it},\tilde{\theta}_\ell)$ by substituting $\hat{a}_{it}$ for $a(x)$, X_{it} for x, Y_{2it} for y_2, and $\tilde{\theta}_\ell$ for θ_0 in the formula for $\alpha_{20}(x,y_2,\theta_0)$ in equation (3.7). Next obtain $\hat{\alpha}_{3\ell}$ by replacing $\alpha_{20}(X_t,Y_{2t},\theta_0)$ by $\hat{\alpha}_2(X_{it},Y_{2it},\tilde{\theta}_\ell)$ and population expectations by sample averages over $i \notin I_\ell, t \le T$ in the third line of equation (3.7). Also, obtain $\hat{\alpha}_{1\ell}(x,\theta)$ by replacing α_{30} and $\gamma_{10}(x)$ by $\hat{\alpha}_{3\ell}$ and $\hat{\gamma}_{1\ell}(x)$ respectively in the first line of equation (3.7) and by replacing the two conditional expectations by the predicted values from Lasso regressions over $i \notin I_\ell, t \le T$ with regressors $b(X_{i,t+1})$, dependent variables equal to each element of $\hat{\alpha}_{2\ell}(X_{it},Y_{2it},\tilde{\theta}_\ell)$ and Y_{1it} respectively, and regularization factors r_2 and r_3 respectively, analogously to $\hat{\gamma}_{2\ell}(x)$. Finally substitute $\hat{\alpha}_{1\ell}$, $\hat{\alpha}_{2\ell}$, and $\hat{\alpha}_{3\ell}$ for α_{10}, α_{20}, and α_{30} and $\hat{\gamma}_{1\ell}$, $\hat{\gamma}_{2\ell}$ and $\hat{\gamma}_{3\ell}$ for γ_1, γ_2, and γ_3 in the formulas for ϕ_1, ϕ_2, and ϕ_3 and construct a debiased GMM estimator as in Section 2.1.

To explore the finite sample properties of this estimator we carried out a Monte Carlo study for a model similar to that of Rust (1987). The state variables consisted of a positive variable x_1 (mileage) and other variables $x_2, ..., x_6$ with transition

$$\begin{aligned}
X_{1,t+1} &= 1+1(Y_{2t}=1)X_{1t}+S_{t+1}^2,\ S_{t+1}|X_t \sim N(.2+\sum_{k=1}^{5}c_k X_{t,k+1},1),\\
c &= (.1,.025,.0111,.0063,.004);
\end{aligned}$$

where $(X_{2t},...,X_{6t})$ is i.i.d. over t, X_{2t}, X_{4t}, and X_{6t} are chi-squared with one degree of freedom and X_{3t} and X_{5t} are binary with $\Pr(X_{kt}=1)=1/2$, $k=3,5$. We specified that $D(x)$ is two dimensional with $D_1(x)=(-1,0)'$ and $D_2(x)=(0,\sqrt{x_1})'$ and that ε_{1t}, ε_{2t} are independent Type I extreme value, so that $\Lambda(a)=e^a/(1+e^a)$ corresponds to binary logit.

To generate the data we solved the Bellman equation on a finite grid using the fact that the state space has a two dimensional structure in terms of x_1 and $\sum_{k=1}^{5} c_k x_{k+1}$, with linear interpolation between grid points. We did not enforce this index structure in estimation, so that the estimation treated the state space as dimension six. We carry out 500 Monte Carlo replications for $T=10$ and $n=100$, 300, 1000, and $10,000$. We specified five fold cross fitting where $L=5$. We consider three specifications of the vector $b(x)$ used by Lasso, consisting of a) the elements of x, b) those from a) and squares of elements of x; c) those from b) and all products of two elements of x. The conditional choice probability estimators $\hat{\gamma}_{1\ell,\ell'}(x)$ and $\hat{\gamma}_{1\ell}(x)$ were logit Lasso trimmed to be between .0001 and .9999. We used the MATLAB Lasso and logit Lasso procedures for computation. The regularization values r_1, r_2, and r_3 for each Lasso were chosen by two fold cross-validation. Although we do not know whether the resulting $r's$ satisfy the conditions in the asymptotic theory of Section 6, we do this so that the estimator in the Monte Carlo is based on an "off the shelf" machine learner of unknown functions.

The results are reported in Tables 1, 2, and 3. The PI labels the plug-in GMM estimator based only on identifying moment functions, DB the debiased GMM, Bias is the absolute value of bias, Med SE denotes the median of the estimated standard errors corresponding to equation (2.8), SD denotes standard deviation, and Cvg denotes coverage probability of a nominal 95 percent confidence interval.

Table 1: $b(X)$ Linear

		PI Bias	DB Bias	Med SE	PI SD	DB SD	PI Cvg	DB Cvg
$n=100$	θ_2	.35	.00	.13	.11	.10	.17	.99
	θ_1	.61	.04	.21	.21	.18	.17	.96
$n=300$	θ_2	.33	.01	.08	.07	.06	.00	.98
	θ_1	.58	.05	.12	.13	.11	.00	.94
$n=1000$	θ_2	.33	.01	.04	.04	.03	.00	.98
	θ_1	.58	.06	.07	.07	.06	.00	.87
$n=10000$	θ_2	.32	.01	.01	.01	.01	.00	.93
	θ_1	.57	.05	.02	.02	.02	.00	.21

Table 2: $b(X)$ Linear, Squares

		PI Bias	DB Bias	Med SE	PI SD	DB SD	PI Cvg	DB Cvg
$n = 100$	θ_2	.24	.03	.13	.11	.33	.61	.98
	θ_1	.41	.03	.21	.20	.38	.57	.97
$n = 300$	θ_2	.24	.03	.08	.07	.07	.08	.98
	θ_1	.41	.02	.12	.13	.13	.11	.97
$n = 1000$	θ_2	.24	.02	.04	.04	.03	.00	.98
	θ_1	.42	.01	.07	.07	.06	.00	.97
$n = 10000$	θ_2	.24	.02	.01	.01	.01	.00	.81
	θ_1	.42	.01	.02	.02	.02	.00	.95

Table 3: $b(X)$ Linear, Squares, and Interactions

		PI Bias	DB Bias	Med SE	PI SD	DB SD	PI Cvg	DB Cvg
$n = 100$	θ_2	.16	.01	.13	.11	.38	.85	.98
	θ_1	.26	.03	.21	.19	.29	.82	.96
$n = 300$	θ_2	.15	.02	.07	.07	.07	.51	.98
	θ_1	.24	.01	.12	.12	.12	.52	.97
$n = 1000$	θ_2	.14	.01	.03	.03	.03	.04	.99
	θ_1	.23	.01	.07	.07	.06	.07	.97
$n = 10000$	θ_2	.13	.01	.01	.01	.01	.00	.98
	θ_1	.23	.01	.02	.02	.02	.00	.94

In all cases debiased GMM has much smaller bias than the plug-in estimator. For the richest dictionary $b(X)$ of Table 3, coverage probabilities are quite close to the nominal value though conservative. For the less rich dictionaries of Tables 1 and 2 enough bias remains relative to variance that coverage probabilities are far from their nominal values for the largest sample size. In contrast plug-in GMM has large bias and confidence interval coverage probabilities that are far from their nominal values in all cases. Remarkably, for larger sample sizes or smaller dimensional $b(x)$, debiased GMM is no more variable than plug-in GMM, and in several cases is less variable. Overall, the performance of the debiased GMM estimator in this example with an "off the shelf" machine learner suggests that debiased GMM for dynamic discrete choice and other structural models could be useful in practice.

An alternative debiased GMM estimator could be obtained by taking the plim's of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ to be least squares projections that solve equation (2.10) for corresponding residuals and using automatic debiasing like that described in Section 2.3. Monte Carlo examples in other settings, such as Singh and Sun (2019), find that approach deliver estimators and standard errors with good finite sample properties. We intend to consider this approach to debiased GMM for dynamic discrete choice in future work.

4 Orthogonality

To show the orthogonality conditions of equations (2.4) and (2.5) we begin with a simpler, more basic result. We assume that equation (2.2) is satisfied for all F_0 in some set $\mathcal{F}$ and for $\theta \in \Theta$. For $F \in \mathcal{F}$ let $\alpha(F)$ denote α_0 such that equation (2.2) is satisfied when $F_0 = F$, i.e. when F is the CDF of W. Note that $\alpha_0 = \alpha(F_0)$. Also let

$$\bar{\psi}(\gamma, \alpha, \theta) := E[\psi(W, \gamma, \alpha, \theta)].$$

Because $\phi(w, \gamma, \alpha, \theta)$ is the FSIF it will satisfy the mean zero condition in equation (2.2) identically in F, so that $0 \equiv E_F[\phi(W, \gamma(F), \alpha(F), \theta)]$. Substituting $F_\tau = (1-\tau)F_0 + \tau H$ for F and differentiating this identity with respect to τ at $\tau = 0$ gives

$$\begin{aligned} 0 &= \int \phi(w, \gamma_0, \alpha_0, \theta) H(dw) + \frac{\partial}{\partial \tau} E[\phi(W, \gamma(F_\tau), \alpha(F_\tau), \theta)] \qquad (4.1) \\ &= \frac{\partial}{\partial \tau} E[g(W, \gamma(F_\tau), \theta)] + \frac{\partial}{\partial \tau} E[\phi(W, \gamma(F_\tau), \alpha(F_\tau), \theta)] = \frac{\partial}{\partial \tau} \bar{\psi}(\gamma(F_\tau), \alpha(F_\tau), \theta), \end{aligned}$$

where the first and third equalities follow by the chain rule and the second equality follows from the influence function formula in equation (*2.2*). This equation shows that the functions γ and α have no first order effect on $\bar{\psi}(\gamma, \alpha, \theta)$ along the path $(\gamma(F_\tau), \alpha(F_\tau))$. The second equality shows how the presence of $E[\phi(W, \gamma, \alpha, \theta)]$ "partials out" the effect of varying τ on $E[g(W, \gamma(F_\tau), \theta)]$. The zero mean property of the FSIF implies that the local effect of γ on $E[g(W, \gamma, \theta_0)]$ along the path $\gamma(F_\tau)$ is cancelled, or "partialled out," by the effect of varying γ and α on $E[\phi(W, \gamma, \alpha, \theta)]$ along the path $(\gamma(F_\tau), \alpha(F_\tau))$. The following result gives precise conditions for equation (4.1):

THEOREM 1: *If for any $\theta \in \Theta$ i) equation (2.2) is satisfied; ii) $\int \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta) F_\tau(dw) = 0$ for all $\tau \in [0, \bar{\tau})$ with $\bar{\tau} > 0$, and iii) $\int \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta) F_0(dw)$ and $\int \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta) H(dw)$ are continuous at $\tau = 0$ then equation (4.1) is satisfied.*

The proofs of this result, Theorems 2-7, and Lemma 8 are given in Appendix A.

EXAMPLE 1: (continued) Here $g(w, \gamma, \theta) = z\gamma(x) - \theta$ and $\phi(w, \gamma, \alpha) = \alpha(x)[y - \gamma(x)]$ so that by equation (2.1)

$$\begin{aligned} \bar{\psi}(\gamma, \alpha, \theta) &= E[Z\gamma(X)] - \theta + E[\alpha(X)\{Y - \gamma(X)\}] \\ &= \theta_0 - \theta - E[\{\alpha(X) - \alpha_0(X)\}\{\gamma(X) - \gamma_0(X)\}]. \end{aligned}$$

When $\alpha(X) = \alpha_0(X)$ the presence of γ in the identifying moment $E[g(W, \gamma, \theta)]$ is exactly cancelled, or partialled out, by the presence of γ in $E[\phi(W, \gamma, \alpha_0)] = E[\alpha_0(X)\{\gamma_0(X) - \gamma(X)\}]$, so that variation in γ has no effect.

Equation (*4.1*) is a zero total derivative condition for joint variation in (γ, α) along the path $(\gamma(F_\tau), \alpha(F_\tau))$. In many cases $\gamma(F)$ and $\alpha(F)$ are distinct objects so that it is possible to choose F_τ so that $\alpha(F_\tau)$ varies with τ and $\gamma(F_\tau) = \gamma_0$ remains equal to its true value. For example $\gamma(F)$ and $\alpha(F)$ may be determined by the distributions of different random variables and so be distinct objects. In such cases equation (2.2) implies the orthogonality property of equation (2.5).

THEOREM 2: *For any $\alpha \in \mathcal{A}$ and $\theta \in \Theta$, if i) there is F_α such that $\alpha(F_\alpha) = \alpha$ and $\gamma(F_\tau^\alpha) = \gamma_0$ for $F_\tau^\alpha = (1-\tau)F_\alpha + \tau F_0$ and all τ small enough; ii) $d\int g(w, \gamma(F_\tau^\alpha), \theta)F_\alpha(dw)/d\tau = \int \phi(w, \gamma_0, \alpha, \theta)F_0(dw)$ then*

$$E[\phi(W, \gamma_0, \alpha, \theta)] = 0. \tag{4.2}$$

Note that hypothesis ii) is just the characterization of the FSIF in equation (2.2) when the true distribution is F_α and H is set equal to F_0. Thus, Theorem 2 shows that equation (2.5) is satisfied at every $\theta \in \Theta$ and every α such that there is some distribution F_α such that $\alpha = \alpha(F_\alpha)$ and $\gamma(F_\tau^\alpha) = \gamma_0$ for all τ close enough to zero. In all the examples of which we are aware equation (4.2) is easy to confirm by inspection of $\phi(W, \gamma_0, \alpha, \theta)$.

EXAMPLE 1: (continued) Here $E[\phi(W, \gamma_0, \alpha, \theta)] = E[\alpha(X)\{Y - \gamma_0(X)\}] = 0$ for any $\alpha(X)$ by $\gamma_0(X) = E[Y|X]$ and iterated expectations.

Theorem 2 shows that equation (2.5) is a general property of the FSIF and is not confined to a particular set of examples.

The orthogonality property of equation (2.4) will follow from Theorem 1 when F_τ can be specified so that $\alpha(F_\tau) = \alpha_0$ and some regularity conditions are satisfied. In that case equation (4.1) implies $\partial\bar{\psi}(\gamma(F_\tau), \alpha_0, \theta)/\partial\tau = 0$. As in the discussion preceding Theorem 2 choosing F_τ in this way will generally be possible when $\gamma(F)$ and $\alpha(F)$ are distinct objects depending on distinct features of F. Equation (2.4) then will follow if the set of pathwise derivatives $d\gamma(F_\tau)/d\tau$ is well defined and rich enough and the chain rule can be applied, as it can under Hadamard differentiability of $\bar{\psi}(\gamma, \alpha_0, \theta)$ in γ and $\gamma(F_\tau)$ in τ.

THEOREM 3: *If there is a norm $\|\gamma\|$, a linear set Γ, and a set $\mathcal{H}$ such that for all $H \in \mathcal{H}$; i) $\alpha(F_\tau) = \alpha_0$ for all τ small enough and equation (4.1) is satisfied; ii) $\bar{\psi}(\gamma, \alpha_0, \theta)$ is Hadamard differentiable at γ_0 tangentially to Γ; iii) $\gamma(F_\tau)$ is Hadamard differentiable at $\tau = 0$; iv) the closure of $\{\partial\gamma(F_\tau)/\partial\tau : H \in \mathcal{H}\}$ is Γ then*

$$\left.\frac{\partial}{\partial t}\bar{\psi}(\gamma_0 + t\delta, \alpha_0, \theta)\right|_{t=0} = 0 \text{ for all } \delta \in \Gamma. \tag{4.3}$$

Furthermore, if $\bar{\psi}(\gamma, \alpha_0, \theta_0)$ is twice continuously Frechet differentiable in a neighborhood of γ_0 then there is $C > 0$ such that for $\|\gamma - \gamma_0\|$ small enough

$$\left\|\bar{\psi}(\gamma, \alpha_0, \theta_0)\right\| \leq C \left\|\gamma - \gamma_0\right\|^2. \tag{4.4}$$

Hadamard and Frechet differentiability are defined and discussed e.g. in van der Vaart (1998, 20.2). The first conclusion of this result is the orthogonality condition of equation (2.4). Thus, Theorems 2 and 3 combined show that adding the FSIF $\phi(w, \gamma, \alpha, \theta)$ to identifying moment functions $g(w, \gamma, \theta)$ is a construction such that $\psi(w, \gamma, \alpha, \theta) = g(w, \gamma, \theta)+ \ \phi(w, \gamma, \alpha, \theta)$ satisfies the orthogonality conditions of equations (2.4) and (2.5).

The second conclusion of Theorem 3, given as equation (4.4), bounds the departure from zero of the expected moment functions as just γ varies. This bound is useful for formulating regularity conditions for root-n consistency of debiased GMM when $\bar{\psi}(\gamma, \alpha_0, \theta_0)$ is nonlinear in γ, as explained in Section 6. When formulating regularity conditions for particular moment functions and first step estimators it may be simpler to directly confirm equation (4.4). In many cases equation (4.4) will be satisfied under specific regularity conditions when $\|a\|$ is a mean-square norm $\|a\| = \sqrt{E[a(W)^2]}$. Equation (4.4) with a mean-square norm $\|\gamma - \gamma_0\|$ has wide applicability to machine learning first steps where mean-square convergence rates are available.

If γ were taken to be the limit of a nonparametric estimator $\hat{\gamma}$ for fixed bandwidth, number of series terms, or regularization then we can think of $\|\gamma - \gamma_0\|$ as bias in γ from nonparametric estimation. Frechet differentiability of $\bar{\psi}(\gamma, \alpha_0, \theta_0)$ in γ and equation (4.4) then imply that $\bar{\psi}(\gamma, \alpha_0, \theta_0) = o(\|\gamma - \gamma_0\|)$, so that the orthogonal moments $\bar{\psi}(\gamma, \alpha_0, \theta_0)$ shrink to zero faster than the nonparametric bias. Thus the orthogonal moment functions constructed here have the small bias property considered in Newey, Hsieh, and Robins (1998, 2004). As usual for GMM the estimator $\hat{\theta}$ will inherit this property of the orthogonal moment functions.

The construction of orthogonal moment functions we consider has antecedents in the literature on functional estimation, where the identifying moment conditions are $g(w, \gamma, \theta) = m(\gamma) - \theta$ for some explicit functional of $m(\gamma)$ of γ. Here $\phi(w, \gamma, \alpha)$ is the influence function of $m(\gamma(F))$ and $\psi(w, \gamma, \alpha, \theta) = m(\gamma) - \theta + \phi(w, \gamma, \alpha)$. Such orthogonal moment functions when γ is a pdf were given by Hasminskii and Ibragimov (1978), Pfanzagl and Wefelmeyer (1981), and Bickel and Ritov (1988). Newey, Hsieh, and Robins (1998, 2004) generalized this construction to allow $m(w, \gamma)$ to depend on w, to γ that is a product of a pdf and conditional mean, and to $\gamma = F$. Robins et al. (2008) gave similar constructions. The construction given here for any γ was also described in Chernozhukov et al. (2018) and Bravo, Escanciano, and van Keilegom (2020) and originated in the joint research for this paper.

Equation (4.1) is similar to Theorem 2.2 of Robins et al. (2008) but different in applying directly to the standard influence function characterization in equation (2.2) without specifying a score function (derivative of the log-likelihood of a model). Proceeding in this way allows us to show orthogonality of $\psi(w, \gamma, \alpha, \theta)$ using equation (4.1). To the best of our knowledge equation (4.1) has not appeared in this form previously. Also, Theorem 2 and Theorem 3, for any first step γ, were not shown in any previous work of which we are aware. These results are distinguished from Newey, Hsieh, and Robins (1998, 2004) and Robins et al. (2008, 2009) in the conclusion of Theorem 2 and in providing, for *any* first step γ, key regularity conditions used in the asymptotic theory of Section 6.

The orthogonalization given here is estimator based rather than model based, relying only on $g(w, \gamma, \theta)$ and the plim $\gamma(F)$ of $\hat{\gamma}$ under general misspecification and not on the specification of a model. The behavior of the sample moments under misspecification is fully accounted for by using the plim $\gamma(F)$ of $\hat{\gamma}$ under general misspecification in the construction of the orthogonal moment functions. Consequently, standard errors will be robust to misspecification when the number of moment functions and parameters of interest is the same.

There are also model based approaches to construction of orthogonal moment functions. Efficient scores and influence functions for semiparametric models are known to be orthogonal for first steps corresponding to those in the model, as in Bickel et al (1993) and Van der Vaart (1998). More generally the residual from the projection of identifying moment functions on the tangent set is orthogonal for first steps corresponding to those in the model, as in Newey (1990). The orthogonality of such moment functions is sensitive to model misspecification unless the model is nonparametric. Also, orthogonality is sensitive to using first step estimators of precisely the objects specified in the model, as pointed out by van der Laan (2014) and Vermeulen and Vansteelandt (2015).

The construction we give bypasses the semiparametric efficiency framework and is based on the simpler influence function characterization in equation (2.2) and the limit $\gamma(F)$ for any semiparametric estimator as in Newey (1994a). This construction highlights the distinct roles of $g(w, \gamma, \theta)$ as identifying moments and $\phi(w, \gamma, \alpha, \theta)$ as a bias correction that does not affect identification and is entirely determined by $g(w, \gamma, \theta)$ and $\gamma(F)$. The distinct role of $\phi(w, \gamma, \alpha, \theta)$ leads directly to Theorem 2 and motivates the automatic estimation of α_0 in Section 2. For these reasons we describe the orthogonal moment construction as adding the FSIF to identifying moments rather than finding an efficient influence function.

The orthogonal moment functions we give are nonparametric, efficient influence functions for a special object, the plim of the identifying moment functions evaluated at $\gamma(F)$, that is $E_F[g(W, \gamma(F), \theta)]$. This object is nonparametric and so it has a unique influence function, as in Van der Vaart (1991) and Newey (1994a). By differentiating $E_{F_\tau}[g(W, \gamma(F_\tau), \theta_0)]$ with respect to τ, applying the chain rule, and using (2.2) it follows that the influence function of this object at F_0 is $\psi(W, \gamma_0, \alpha_0, \theta_0)$ under the moment condition $E[g(W, \gamma_0, \theta_0)] = 0$. The sample average of the debiased moment function $\hat{\psi}(\theta_0)$ is a nonparametric estimator of $E_F[g(W, \gamma(F), \theta_0)]$. If $\hat{\psi}(\theta_0)$ is asymptotically equivalent to a sample average and locally regular, meaning that for H

and data that are i.i.d. with CDF $F_{\tau_n} = (1-\tau_n)F_0 + \tau_n H$ and $\tau_n = O(1/\sqrt{n})$ the limiting distribution of $\sqrt{n}\{\hat{\psi}(\theta_0) - E_{F_{\tau_n}}[g(W,\gamma(F_{\tau_n}),\theta_0)]\}$ does not depend on τ_n, then uniqueness of the influence function for a nonparametric object implies

$$\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[g(W_i,\hat{\gamma}_\ell,\theta_0) + \phi(W_i,\hat{\gamma}_\ell,\hat{\alpha}_\ell,\tilde{\theta}_\ell)] = \hat{\psi}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\psi(W_i,\gamma_0,\alpha_0,\theta_0) + o_p(n^{-1/2}), \quad (4.5)$$

That is, the *only* sample average of a function of the data that $\hat{\psi}(\theta_0)$ can be asymptotically equivalent to, and also be locally regular, is $\sum_{i=1}^{n}\psi(W_i,\gamma_0,\alpha_0,\theta_0)/n$. Equation (4.5) is precisely an asymptotic version of orthogonality where the sample average $\hat{\psi}(\theta_0)$ is asymptotically equivalent to the sample average of the same function with estimators $\hat{\gamma}_\ell$, $\hat{\alpha}_\ell$, and $\tilde{\theta}_\ell$ replaced by their limits. In this standard way orthogonality of $\psi(W_i,\gamma_0,\alpha_0,\theta_0)$ can be understood as resulting from $\psi(W_i,\gamma_0,\alpha_0,\theta_0)$ being the efficient influence function of the nonparametric object $E_F[g(W,\gamma(F),\theta)]$ evaluated at the plim $\gamma(F)$ of the first step.

In some cases the identifying moment functions $g(w,\gamma,\theta)$ may already be orthogonal, so that $\phi(w,\gamma,\alpha,\theta) = 0$. An important class of orthogonal moment functions are those where $g(w,\gamma,\theta)$ is the derivative with respect to θ of an objective function where nonparametric parts have been concentrated out. This class of moment functions includes those of Robinson (1988), Ichimura (1993), and various partially linear regression models where ζ represents a conditional expectation. It also includes the efficient score for a semiparametric model when the nonparametric component estimates the maximizer of the expected log likelihood; see Newey (1994a, pp. 1358-1359), and van der Vaart (1998, pp. 391-396). In general, suppose that there is a function $q(w,\theta,\zeta)$ such that $g(w,\gamma,\theta) = \partial q(w,\theta,\zeta(\theta))/\partial\theta$ and $\zeta(\theta) = \arg\max_\zeta E[q(W,\theta,\zeta)]$, where γ includes $\zeta(\theta)$ and possibly additional functions. Proposition 2 of Newey (1994a) and Lemma 2.5 of Chernozhukov et al. (2018) then imply that $g(w,\gamma,\theta)$ is orthogonal.

5 Double Robustness

The zero derivative condition in equation (2.4) is an appealing robustness property. This condition can be interpreted as local insensitivity of the average moments $\bar{\psi}(\gamma,\alpha_0,\theta) = E[\psi(W,\gamma,\alpha_0,\theta)]$ to γ, with $\bar{\psi}(\gamma,\alpha_0,\theta)$ varying little as γ varies away from γ_0. Because it is difficult to get unknown functions exactly right, especially in high dimensional settings, this property is an appealing one.

Such robustness considerations, well explained in Robins and Rotnitzky (2001), have motivated the development of doubly robust moment functions that have expectation that does not depend on one first step if the other is set equal to its plim. Doubly robust moment conditions allow two chances for identifying moment conditions to hold, an appealing robustness feature. Also, doubly robust moment conditions have simpler conditions for asymptotic normality than general debiased GMM, as discussed in Section 6.

In this Section we characterize double robustness and derive several novel classes of doubly robust moment functions. We construct doubly robust moment functions by adding to identifying moment functions the FSIF. In this way the derivation of new doubly robust moment functions is aided by the construction of orthogonal moment functions in Section 2.

5.1 Characterizing Double Robustness

Double robustness is that for all $\gamma \in \Gamma$, $\alpha \in \mathcal{A}$, and $\theta \in \Theta$,

$$\bar{\psi}(\gamma, \alpha_0, \theta) = \bar{\psi}(\gamma_0, \alpha_0, \theta) = \bar{\psi}(\gamma_0, \alpha, \theta), \tag{5.1}$$

where Γ is the set of possible plim's $\gamma(F)$ that we continue to assume is linear. The second equality of equation (5.1) already follows from the conclusion of Theorem 2 that implies

$$\bar{\psi}(\gamma_0, \alpha, \theta) = E[g(W, \gamma_0, \theta)] + E[\phi(W, \gamma_0, \alpha, \theta)] = E[g(W, \gamma_0, \theta)] = \bar{\psi}(\gamma_0, \alpha_0, \theta).$$

The first conclusion of Theorem 3 gives a local version of the first equality in equation (5.1). If $\bar{\psi}(\gamma, \alpha_0, \theta)$ is affine in γ then this local property becomes global so that double robustness holds. Clearly doubly robust moment conditions have $\bar{\psi}(\gamma, \alpha_0, \theta)$ that is constant in γ, and therefore are affine, so that $\bar{\psi}(\gamma, \alpha_0, \theta)$ being affine in γ and satisfying the first conclusion of Theorem 3 is a complete characterization of double robustness.

THEOREM 4: *$\psi(w, \gamma, \alpha, \theta)$ is doubly robust if and only if $\bar{\psi}(\gamma, \alpha_0, \theta)$ is affine in γ and*

$$\left.\frac{d\bar{\psi}(\gamma_0 + t\delta, \alpha_0, \theta)}{dt}\right|_{t=0} = 0 \textit{ for all } \delta \in \Gamma.$$

This characterization can be used to construct doubly robust moment functions. A good place to start is an identifying moment function $g(w, \gamma, \theta)$ that is affine in γ. If the associated FSIF $\phi(w, \gamma, \alpha, \theta)$ is also affine in γ then the orthogonal moment function $\psi(w, \gamma, \alpha, \theta) = g(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta)$ will be affine in γ and so be doubly robust, since the second condition of Theorem 4 is satisfied by Theorem 3. In the remainder of this Section we use Theorem 4 to construct a variety of doubly robust moment functions.

Robins and Rotnitzky (2001) gave conditions for the existence of doubly robust moment functions in semiparametric models. Theorem 4 is complementary to those results in giving a characterization of doubly robust moment functions.

5.2 Double Robustness for Regression First Steps

An important set of examples are those with $\theta_0 = E[m(W, \gamma_0)]$ where $m(W, \gamma)$ is linear in γ,

$$\gamma(F) = \arg\min_{\gamma \in \Gamma} E_F[\{Y - \gamma(X)\}^2],$$

and Γ is a linear set of functions that is closed in mean-square (meaning $\gamma \in \Gamma$ if for every $\varepsilon > 0$ there exists $\gamma_\varepsilon \in \Gamma$ with $E[\{\gamma(X) - \gamma_\varepsilon(X)\}^2] < \varepsilon$). Here the identifying moment function is $g(w, \gamma, \theta) = m(w, \gamma) - \theta$, which is affine in γ. Also by Proposition 4 of Newey (1994a), if there is $\alpha_0(X) \in \Gamma$ with finite second moment such that $E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)]$ for all $\gamma \in \Gamma$, the FSIF is

$$\phi(w, \gamma, \alpha) = \alpha(x)[y - \gamma(x)],$$

for $\alpha \in \Gamma$. Here the set $\mathcal{A}$ of possible α is Γ. By $m(w, \gamma)$ and $\phi(w, \gamma, \alpha)$ both affine in γ Theorem 4 gives

THEOREM 5: *If $m(w, \gamma)$ is linear in $\gamma \in \Gamma$ and there is $\alpha_0 \in \Gamma$ such that $E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)]$ for all $\gamma \in \Gamma$ then $\psi(w, \gamma, \alpha, \theta) = m(w, \gamma) - \theta + \alpha(x)[y - \gamma(x)]$ is doubly robust.*

Existence of $\alpha_0(x)$ in Theorem 5 is guaranteed by the Riesz representation theorem if $E[m(W, \gamma)]$ is mean-square continuous in γ, so we refer to $\alpha_0(x)$ as the Riesz representer. In Example 1 note that $\psi(w, \gamma, \alpha, \theta) = Z\gamma(X) + Y\alpha(X) - \alpha(X)\gamma(X) - \theta$ is doubly robust by Theorem 5. Many important doubly robust moment functions are special cases of Theorem 5, including average treatment effects, policy effects, and average derivatives, as discussed in Newey and Robins (2017), Chernozhukov, Newey, and Robins (2018), Hirshberg and Wager (2018), and Chernozhukov, Newey, and Singh (2018). In these papers a variety of other doubly robust moment functions are also formulated based on Theorem 5.

5.3 Double Robustness for Nonparametric IV

A larger set of important examples are those where the first step minimizes a population two-stage least squares objective function based on an affine residual. Let $\lambda(W, \gamma)$ denote a scalar residual function that is affine in γ, $\mathcal{B} = \{b(X)\}$ denote a linear set of possible instrumental variables $b(X)$ that is closed in mean-square, and $proj_F(a(W)|\mathcal{S})$ denote the least squares projection of $a(W)$ on a set $\mathcal{S}$ that is linear and closed in mean-square when F is the true CDF. We consider debiased GMM where

$$\gamma(F) = \arg\min_{\gamma \in \Gamma} E_F[proj_F(\lambda(W, \gamma)|\mathcal{B})^2]. \tag{5.2}$$

This $\gamma(F)$ is the plim of the nonparametric 2SLS estimator of Newey and Powell (2003), Newey (1991), and Ai and Chen (2007). It follows from Ai and Chen (2007, p. 40), and Ichimura and Newey (2021) that when $E[\lambda(W, \gamma_0)b(X)] = 0$ for all $b \in \mathcal{B}$, which we assume, and certain other conditions are satisfied, then the FSIF has the form

$$\phi(w, \gamma, \alpha, \theta) = \alpha(x, \theta)\lambda(w, \gamma), \ \alpha(x, \theta) \in \mathcal{B}.$$

It will follow from Theorem 4 that if the identifying moment function $g(w, \gamma, \theta)$ and the residual $\lambda(w, \gamma)$ are affine in γ on a linear set Γ that is closed in mean-square then the orthogonal moment function

$$\psi(w, \gamma, \alpha, \theta) = g(w, \gamma, \theta) + \alpha(x, \theta)\lambda(w, \gamma)$$

will be affine also, and so will be doubly robust.

A special case, that generalizes the linear regression functionals of Theorem 5 to allow for endogeneity, has $\lambda(W, \gamma) = Y - \gamma(Z)$. Let $\mathcal{A}$ denote the mean-square closure of $\{proj_{F_0}(\gamma(Z)|\mathcal{B}) : \gamma \in \Gamma\}$

Theorem 6: *If i) $m(w, \gamma)$ is linear in γ and there is $v_m(Z) \in \Gamma$ such that $E[m(W, \gamma)] = E[v_m(Z)\gamma(Z)]$ for all $\gamma \in \Gamma$; ii) there exists $b_m \in \mathcal{B}$ such that $proj_{F_0}(b_m(X)|\Gamma) = v_m(Z)$; iii) $E[b(X)\{Y - \gamma_0(Z)\}] = 0$ for all $b \in \mathcal{B}$ then $\psi(w, \gamma, \alpha, \theta) = m(w, \gamma) - \theta + \alpha(x)[y - \gamma(z)]$ is doubly robust for $\alpha_0(X) = proj_{F_0}(b_m(X)|\mathcal{A})$.*

Condition i) is like the hypothesis that $E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)]$ in Theorem 5, condition ii) is similar to the Severini and Tripathi (2012) condition for root-n consistent estimability of θ_0, and iii) is a correct specification condition.

Example 4: Many novel doubly robust moment functions can be constructed based on Theorem 6, including policy effects and average derivatives. A weighted average derivative example has $m(w, \gamma) = \bar{v}(z)\partial\gamma(z)/\partial z_1$ for some known weight $\bar{v}(z)$ where Γ is all functions of Z with finite second moment. Integration by parts gives

$$E[m(W, \gamma)] = E[v_m(Z)\gamma(Z)],\ v_m(Z) = -\frac{\partial\{f_0(Z)\bar{v}(Z)\}/\partial z_1}{f_0(Z)},$$

where $f_0(z)$ is the marginal pdf of Z. Then by Theorem 6 a doubly robust moment function is

$$\psi(w, \gamma, \alpha, \theta) = \bar{v}(z)\frac{\partial\gamma(z)}{\partial z_1} - \theta + \alpha(x)[y - \gamma(z)],$$

when there exists $b_m(X)$ with $v_m(Z) = E[b_m(X)|Z]$. This is a doubly robust moment function that could be used to construct a doubly robust version of the plug-in estimator of Ai and Chen (2007).

Using Theorem 4 to construct doubly robust moment functions can depend on specifying γ to make $g(w, \gamma, \theta)$ and $\phi(w, \gamma, \alpha, \theta)$ affine in γ. We illustrate with a well known example.

Example 5: Suppose that the object of interest is $\theta_0 = E[Y^*]$ where $Y = 1(D = 1)Y^*$ is observed for an observed completed data indicator $D \in \{0, 1\}$ and the data are missing at random with $E[Y^*|X, D = 1] = E[Y^*|X]$ for observed covariates X. Inverse probability weighting gives $\theta_0 = E[DY/P_0(X)] = E[P_0(X)^{-1}E[DY|X]]$, which is nonlinear in the unknown propensity score $P_0(X) = \Pr(D = 1|X)$. A corresponding affine in γ identifying moment function is $g(w, \gamma, \theta) = g(w, \gamma, \theta) = \gamma(x)dy - \theta$ with true first step $\gamma_0(X) = P_0(X)^{-1}$. This γ_0 minimizes the objective function in equation (5.2) at $F = F_0$ for $\lambda(w, \gamma) = 1 - \gamma(x)d$ that is affine in γ and $\mathcal{B}$ equal to the set of all functions of X with finite second moment. Also, for $\alpha_0(X) = E[Y|X, D = 1] = E[DY|X]\gamma_0(X)$ we have

$$\begin{aligned} E[g(W, \gamma, \theta_0)] &= E[E[DY|X]\{\gamma(X) - \gamma_0(X)\}] = E[\alpha_0(X)\gamma_0(X)^{-1}\{\gamma(X) - \gamma_0(X)\}] \\ &= E[\alpha_0(X)\{\gamma(X)P_0(X) - 1\}] = E[\alpha_0(X)\{\gamma(X)D - 1\}] = -E[\alpha_0(X)\lambda(W, \gamma)]. \end{aligned}$$

The doubly robust moment function from equation (5.3) is then $\psi(w, \gamma, \alpha, \theta) = d\gamma(x)y - \theta + \alpha(x)(1 - \gamma(x)d)$, which is the doubly robust moment function of Robins, Rotnitzky, and Zhao (1994). This example shows how a classic doubly robust moment function is a special case of Theorem 4, with moment condition that is affine in a first step γ for γ_0 equal to the inverse propensity score.

5.4 Double Robustness for Probability Density First Step

Another interesting class of doubly robust moment conditions are those where the first step γ is a pdf of a function X of the data observation W. By Proposition 5 of Newey (1994a), the first step influence function is

$$\phi(w, \gamma, \alpha, \theta) = \alpha(x, \theta) - \int \alpha(u, \theta)\gamma(u)du,$$

which is affine in γ. When the identifying moment function is affine adding this FSIF gives a doubly robust moment function. For $g(w, \gamma, \theta) = m(w, \gamma) - \theta$ a double robustness result is:

Theorem 7: *If $m(W, \gamma)$ is linear in γ and there exists $\alpha_0(x)$ with $\int \alpha_0(u)^2 du < \infty$ and $E[m(W, \gamma)] = \int \alpha_0(u)\gamma(u)du$ for all γ with $\int \gamma(u)^2 du < \infty$ then $\psi(W, \gamma, \alpha, \theta) = m(W, \gamma) - \theta + \alpha(X) - \int \alpha(u)\gamma(u)du$ is doubly robust.*

Here $\alpha_0(x)$ is the Riesz representer of Proposition 5 of Newey (1994a) for the Lebesgue inner product.

Example 6: An example is the density weighted average derivative of Powell, Stock, and Stoker (1989), where $g(w, \gamma, \theta) = -2y \cdot \partial\gamma(x)/\partial x - \theta$ and $\alpha_0(x) = \partial\{E[Y|X = x]\gamma_0(x)\}/\partial x$. Because $g(w, \gamma, \theta)$ is affine in γ Theorem 7 implies

$$\psi(W, \gamma, \alpha, \theta) = -2Y\frac{\partial\gamma(X)}{\partial x} - \theta + \alpha(X) - \int \alpha(u)\gamma(u)du,$$

is doubly robust. Double robustness of this moment function seems to be a novel result.

Robustness results for multiple first steps can be obtained from simple extensions of Theorems 2-4. When γ is a $S \times 1$ vector of first steps then there will be one FSIF $\phi_s(w, \gamma, \alpha_s, \theta)$ for each distinct component γ_s of γ, $(s = 1, ..., S)$. Each FSIF will satisfy $E[\phi_s(W, \gamma_0, \alpha_s, \theta)] = 0$ identically in α_s and θ by Theorem 2. Consider varying γ_s while $\gamma_{\tilde{s}} = \gamma_{\tilde{s}0}$ for each $\tilde{s} \neq s$. Then $g(w, \gamma, \theta) + \phi_s(w, \gamma, \alpha_{s0}, \theta)$ will satisfy Theorem 3 as γ_s varies and by Theorem 4 $E[g(W, \gamma, \theta) + \phi_s(W, \gamma, \alpha_{s0}, \theta)]$ does not vary with γ_s if and only if $E[g(W, \gamma, \theta) + \phi_s(W, \gamma, \alpha_{s0}, \theta)]$ is affine in γ_s. These multiple robustness features generalize those of Tchetgen Tchetgen (2009) to the orthogonal moment functions we consider.

Doubly robust moment functions also can be used for identification of the parameter of interest θ_0, e.g. as in Escanciano and Li (2021).

6 Asymptotic Theory

In this Section we give simple and general asymptotic theory for debiased GMM. We begin with conditions for the key property

$$\sqrt{n}\hat{\psi}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \theta_0, \gamma_0, \alpha_0) + o_p(1). \tag{6.1}$$

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

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

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

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

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

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

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

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

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

This condition imposes a rate condition on the interaction remainder $\hat{\Delta}_\ell(w)$, that its average must go to zero faster than $1/\sqrt{n}$. Condition iii) is a minimal regularity condition that is used in Newey and Robins (2017).

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

Assumption 3 incorporates Theorem 2 in i) and doubly robust moment functions through ii), in which case Assumption 3 imposes no conditions additional to Assumptions 1 and 2. Conditions iii) and iv) are alternative small bias conditions that are only required to hold for $\hat{\gamma}_\ell$, and not for $\hat{\alpha}_\ell$. Condition iii) requires a faster than $n^{-1/4}$ rate for $\hat{\gamma}$ as is familiar from the semiparametric estimation literature. In many cases iii) will be satisfied for the mean-square norm $\|a\| = \sqrt{E[a(W)^2]}$ so that Assumptions 1-3 will only require mean-square convergence rates.

Lemma 8: *If Assumptions 1-3 are satisfied then equation (6.1) is satisfied.*

This key asymptotic result differs from previous results of e.g. Andrews (1994) and Newey (1994a) in requiring no Donsker conditions. This feature is made possible by the use of cross-fitting in the moment conditions. Avoiding Donsker conditions is important for machine learning first steps that generally do not, or are not known to, satisfy Donsker conditions, as previously discussed in Chernozhukov et al. (2018).

The absence of Donsker conditions helps the conditions of Lemma 8 be much simpler than previous results. The use of orthogonal moment conditions also makes the conditions simpler because they avoid the need to show that

$$\frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{i=1}^{n} \phi(W_i, \hat{\gamma}_\ell, \alpha_0, \theta_0) \xrightarrow{p} 0, \tag{6.2}$$

which is required for plug-in estimators, as discussed in Appendix C. Showing that this condition is satisfied typically involves substantial calculation that is specific to the estimator $\hat{\gamma}_\ell$.

At the same time Lemma 8 is more general than previous results in applying to any first step estimator, rather than a specific one like the early results of e.g. Robinson (1988), Powell, Stock, and Stoker (1989), and others. Furthermore, all that is required of the first-step estimator is mean-square consistency as in Assumption 1 and mean-square rates as in Assumptions 2 and 3. A wide variety of first step estimators will satisfy these conditions, notably machine learners where mean-square consistency and rates have been obtained.

The debiased GMM estimator $\hat{\theta}$ and the Lemma 8 is more complicated in involving construction of and properties for $\hat{\alpha}_\ell$. Some such $\hat{\alpha}_\ell$ is needed in any case for asymptotic variance estimation for a plug-in estimator, which will have the same form as $\hat{V}$, so that this feature does not make debiased GMM more complicated than a plug-in estimator for large sample inference.

Lemma 8 shares this generality and simplicity with the asymptotic theory of Chernozhukov et al. (2018). Lemma 8 improves on Chernozhukov et al. (2018) in allowing $\hat{\alpha}_\ell$ to converge slower than $n^{-1/4}$ in general, in Assumption 1 applying separately to $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$, and having weaker conditions for terms that involve both $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ in Assumption 2. These improvements result from Theorem 2 and the structure of orthogonal moments we consider, as the sum of identifying moment functions and the FSIF.

The next condition is useful for consistency of $\hat{\Psi}$ that appears in the debiased GMM asymptotic variance estimator of equation (2.8).

Assumption 4: *$\int \left\| g(w,\hat{\gamma}_\ell,\tilde{\theta}_\ell) - g(w,\hat{\gamma}_\ell,\theta_0)\right\|^2 F_0(dw) \xrightarrow{p} 0$ and $\int \left\|\hat{\Delta}_\ell(w)\right\|^2 F_0(dw) \xrightarrow{p} 0$ for each $(\ell = 1,...,L)$.*

It is also important to have conditions for convergence of the Jacobian of the identifying sample moments $\partial\hat{g}(\bar{\theta})/\partial\theta \xrightarrow{p} G = E[\partial g(W,\gamma_0,\theta_0)/\partial\theta]$ for any $\bar{\theta} \xrightarrow{p} \theta_0$. To that end we impose the following condition:

Assumption 5: *G exists and there is a neighborhood $\mathcal{N}$ of θ_0 and $\|\cdot\|$ such that i) for each ℓ, $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$;* ii) for all $\|\gamma-\gamma_0\|$ small enough *$g(W,\gamma,\theta)$ is differentiable in θ on $\mathcal{N}$ with probability approaching 1 and there is $C>0$ and $d(W,\gamma)$ such that for $\theta \in \mathcal{N}$ and $\|\gamma-\gamma_0\|$ small enough*

$$\left\|\frac{\partial g(W,\gamma,\theta)}{\partial\theta} - \frac{\partial g(W,\gamma,\theta_0)}{\partial\theta}\right\| \le d(W,\gamma)\left\|\theta-\theta_0\right\|^{1/C};\ E[d(W,\gamma)] < C.$$

iii) For each $\ell = 1,...,L$, j, and k, $\int |\partial g_j(w,\hat{\gamma}_\ell,\theta_0)/\partial\theta_k - \partial g_j(w,\gamma_0,\theta_0)/\partial\theta_k| F_0(dw) \xrightarrow{p} 0.$

With these conditions in place the asymptotic normality of semiparametric GMM follows in a standard way.

Theorem 9: *If Assumptions 1-3 and 5 are satisfied, $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\Upsilon} \xrightarrow{p} \Upsilon$, and $G'\Upsilon G$ is nonsingular, then*

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,V),\ V = (G'\Upsilon G)^{-1}G'\Upsilon\Psi\Upsilon G(G'\Upsilon G)^{-1}.$$

If also Assumption 4 is satisfied then $\hat{V} = (\hat{G}'\hat{\Upsilon}\hat{G})^{-1}\hat{G}'\hat{\Upsilon}\hat{\Psi}\hat{\Upsilon}\hat{G}(\hat{G}'\hat{\Upsilon}\hat{G})^{-1} \xrightarrow{p} V$.

This result and Theoems 10-12 to follow are proven in the online appendix. This result assumes consistency of $\hat{\theta}$. In Appendix F we give primitive conditions for consistency that are like those of this Section in using mean-square consistency of first steps. When θ and $g(w,\gamma,\theta)$ have the same dimension these conditions allow for misspecification where θ_0 is interpreted as the unique solution to $E[g(W,\gamma_0,\theta)] = 0$.

6.1 Functionals of NPIV

Functionals of a first step satisfying an orthogonality condition $E[b(X)\lambda(W,\gamma_0)] = 0$ for all $b \in \mathcal{B}$, similar to Section 5.3, are of general interest. We give conditions for an identifying moment function $m(w,\gamma) - \theta$ that includes Example 2. As in Ichimura and Newey (2021) the FSIF is $\alpha(x)\lambda(w,\gamma)$. Debiased GMM is

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)\lambda(W_i,\hat{\gamma}_\ell)]\,,\ \hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}^2,$$

where $\hat{\psi}_{i\ell} = m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)\lambda(W_i,\hat{\gamma}_\ell) - \hat{\theta}$. Let $\bar{\lambda}(X,\gamma) = E[\lambda(W,\gamma)|X]$, $V = Var(m(W,\gamma_0) + \alpha_0(X)\lambda(W,\gamma_0))$, and $\|a\| = \sqrt{\int a(w)^2 F_0(w)}$ denote the mean-square norm.

Theorem 10: *If i) $E[\lambda(W,\gamma_0)b(X)] = 0$ for all $b \in \mathcal{B}$ and $\hat{\alpha}_\ell \in \mathcal{B}$ with probability approaching one; ii) $\alpha_0(X)$ and $E[\lambda(W,\gamma_0)^2|X]$ are bounded and $E[m(W,\gamma_0)^2] < \infty$; for $(\ell = 1,...,L)$, iii) $\int[m(w,\hat{\gamma}) - m(w,\gamma_0)]^2F_0(dw) \xrightarrow{p} 0$, $\int[\lambda(w,\hat{\gamma}_\ell) - \lambda(w,\gamma_0)]^2F_0(dw) \xrightarrow{p} 0$, $\|\hat{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$; iv) either a) $\int[\hat{\alpha}_\ell(x) - \alpha_0(x)]^2[\lambda(w,\hat{\gamma}_\ell) - \lambda(w,\gamma_0)]^2F_0(dw) \xrightarrow{p} 0$ and $\sqrt{n}\,\|\hat{\alpha}_\ell - \alpha_0\|\,\left\|\bar{\lambda}(\hat{\gamma}_\ell) - \bar{\lambda}(\gamma_0)\right\| \xrightarrow{p} 0$, or b) $\sqrt{n}\,\|\hat{\alpha}_\ell - \alpha_0\|\,\|\lambda(\hat{\gamma}_\ell) - \lambda(\gamma_0)\| \xrightarrow{p} 0$ and $\hat{\alpha}_\ell(x)$ in $\hat{\psi}_{i\ell}$ is replaced by $\bar{\alpha}_\ell(x) = \hat{\alpha}_\ell(x)1(|\hat{\alpha}_\ell(x)| \leq M) + sgn(\hat{\alpha}_\ell(x))M1(|\hat{\alpha}_\ell(x)| > M)$ and $M\,\|\lambda(\hat{\gamma}_\ell) - \lambda(\gamma_0)\| \xrightarrow{p} 0$; v) $\sqrt{n}\bar{\psi}(\hat{\gamma}_\ell,\alpha_0,\theta_0) \xrightarrow{p} 0$; then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,V),\ \hat{V} \xrightarrow{p} V.$$

6.2 Example 2: Functionals of Quantile Regression

We will specify conditions that allow us to apply Theorem 10 to this problem. The next condition will be sufficient for condition iv) of Theorem 10 for $\lambda(w,\gamma)$ in Section 3.1.

Assumption 6: *i) There exists bounded $v_m(x)$ such that $E[m(W,\gamma)] = E[v_m(X)\gamma(X)]$ for all $\gamma(X)$ with $E[\gamma(X)^2] < \infty$; ii) $U = Y - \gamma_0(X)$ is continuously distributed and there is $C > 0$ such that the conditional pdf $f(u|X)$ of U conditional on X satisfies $C^{-1} \leq f(0|X) \leq C$ and is twice continuously differentiable in u with probability one with $|\partial^j f(u|X)/\partial u^j| \leq C$, $(j = 1,2)$; iii) Γ is closed in mean-square.*

This condition specifies that $E[m(W,\gamma)]$ is a mean-square continuous linear functional of γ with Riesz representer $v_m(X)$ and imposes some restrictions on the conditional pdf of U given X. Here $\alpha_0(X) = \arg\min_{\alpha\in\Gamma} E[f(0|X)\{-f(0|X)^{-1}v_m(X) - \alpha(X)\}^2]$ is the linear projection of $-f(0|X)^{-1}v_m(X)$ on Γ weighted by the conditional pdf $f(0|X)$. The $\hat{\alpha}_\ell(X)$ given in Section 3 will estimate $\alpha_0(X)$ because weighting by $f(0|X)$ is incorporated in the kernel weighting included in $\hat{Q}_\ell$. This weighting allows us to avoid inverting an estimator of $f(0|X)$. We obtain a mean-square convergence rate for this $\hat{\alpha}_\ell(x)$ by extending the results of Chernozhukov, Newey, and

Singh (2018) to allow kernel weighting in $\hat{Q}_\ell$. Because this paper is focused on the properties of $\hat{\theta}$ we reserve the full conditions to Appendix D, only stating here the conditions required of the kernel $K(u)$, the bandwidth h, and the regularization factor r in the Lasso minimum distance estimator in equation (3.2). This condition will require that r shrinks slower than the usual Lasso rate.

Assumption 7: i) *$K(u)$ is a symmetric bounded kernel of order κ with bounded support; ii) $h\sqrt{n} \longrightarrow \infty$; iii) for each ℓ, ℓ', $\|\hat{\gamma}_{\ell\ell'} - \gamma_0\| = O_p(n^{-d_\gamma})$; iv) $\sqrt{\ln(p)/(hn)} + h^2 + n^{-d_\gamma} = o(r)$; v) $r \longrightarrow 0$; vi) there is $C > 0$ such that $|m(W, b)| \leq C\,|b(X)|$ for all $b \in \Gamma$.*

The following result gives conditions for asymptotic inference for the estimator of a linear functional of a regression quantile estimator.

Theorem 11: *If i) Assumptions 6 and 7 are satisfied; ii) $E[m(W, \gamma_0)^2] < \infty$ and $\int[m(w, \hat{\gamma}) - m(w, \gamma_0)]^2 F_0(dw) \xrightarrow{p} 0$; iii) $\|\hat{\gamma}_\ell - \gamma_0\| = O_p(n^{-d_\gamma})$ for $1/4 < d_\gamma < 1/2$; either iv) Assumptions D1 and D2 are satisfied and $\sqrt{n}\sqrt{r}n^{-d_\gamma} \longrightarrow 0$ or v) Assumptions D1-D4 are satisfied and $\sqrt{n}r^{2\xi/(1+2\xi)}n^{-d_\gamma} \longrightarrow 0$; and vi) $v_m(X)$ and $\alpha_0(X)$ are bounded then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),\ \hat{V} \xrightarrow{p} V.$$

This result depends on the regression quantile estimator converging at a mean-square rate that is faster than $n^{-1/4}$. Such a rate for an L_1 regularized regression quantile estimator is derived by Belloni and Chernozhukov (2011).

6.3 Example 3: Dynamic Discrete Choice

A result that is important for the properties of $\hat{\theta}$ for dynamic discrete choice and more generally for economic structural model is a convergence rate for the estimator $\hat{\gamma}_2(x)$ of the value function term in the choice probability. We maintain independence of observations across i but allow arbitrary dependence across t.

Assumption 8: *i) There is $\varepsilon > 0$ such that $\gamma_{10}(X) \in [\varepsilon, 1-\varepsilon]$, for all ℓ, ℓ', $\hat{\gamma}_{1\ell\ell'}(X_t) \in [\varepsilon, 1-\varepsilon]$, and $H(p)$ is twice continuously differentiable on $[\varepsilon, 1-\varepsilon]$; ii) For all ℓ, ℓ', $\|\hat{\gamma}_{1\ell\ell'} - \gamma_0\| = O_p(n^{-d_1})$, $0 < d_1 < 1/2$; iii) Assumptions D1, D3, and D5 are satisfied with $\alpha_0(x) = \gamma_{20}(x)$ and sparse approximation rate $\xi_1 > 1/2$; iv) $n^{-d_1} = o(r_1)$ and $r_1 = O(n^{-d_1}\ln(n))$;* v) *$\gamma_{20}(X) = \sum_{j=1}^{\infty}\beta_{j0}b_j(X)$ with $\sum_{j>p}|\beta_{j0}| = O_p(n^{-d_1(2\xi_1-1)/(2\xi_1+1)}\ln(n))$.*

In practice condition i) requires fixed trimming where $\hat{\gamma}_{1\ell\ell'}(X_t)$ is censored below by ε and above by $1-\varepsilon$, with ε being known. Here and in the Theorem 12 below we impose tighter restrictions on the penalty sizes r_1, r_2, and r_3 than needed in order to allow smaller sparse approximation rates, e.g. ξ_1 in Assumption 8.

THEOREM 12: *If i) Assumption 8 is satisfied, ii) $\Lambda(a) > 0$ for all $a \in \Re$, $\ln \Lambda_a(a)$ is concave, $\Lambda(a)$ is twice differentiable with uniformly bounded derivatives, $D(x)$ is bounded, $E[D(X)D(X)']$ is nonsingular; iii) Assumptions D1 and D2 are satisfied for $\alpha_0(x)$ equal to each element of $E[D(X_t)\pi(a_0(X_t))\Lambda_a(a(X_t)Y_{2t}/\Lambda(a(X_t))|X_{t+1} = x]$* with sparse approximation rate ξ_2 and for $E[Y_{1t}|X_{t+1} = x]$ *with sparse approximation rate ξ_3; and iv) $d_1 > 1/4$; v) $1 + [(2\xi_1 - 1)/(2\xi_1 + 1)]2\xi_2/[2\xi_2 + 1] > 1/2d_1$, $n^{-d_1(2\xi_1-1)/(2\xi_1+1)} = o(r_2)$, and $r_2 = O(n^{-d_1(2\xi_1-1)/(2\xi_1+1)}\ln(n))$; vi) $\xi_3/[2\xi_3+1]+d_1 > 1/2$, $\sqrt{\ln(p)/n} = o(r_3)$, and $r_3 = O(\sqrt{\ln(p)/n}\ln(n))$; vii) $(4\xi_1-1)/(2\xi_1+1) > 1/2d_1$; then for $V = G^{-1}E[\psi_0(W)\psi_0(W)']G^{-1}$*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V), \ \hat{V} \xrightarrow{p} V.$$

7 Appendix A: Proofs of Key Results

Proof of Theorem 1: Let $\phi(w, F_\tau) := \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta)$. By ii),

$$0 = (1 - \tau)\int \phi(w, F_\tau)F_0(dw) + \tau \int \phi(w, F_\tau)H(dw).$$

Dividing by τ and solving gives

$$\frac{1}{\tau}\int \phi(w, F_\tau)F_0(dw) = -\int \phi(w, F_\tau)H(dw) + \int \phi(w, F_\tau)F_0(w).$$

Taking limits as $\tau \longrightarrow 0$, $\tau > 0$ it follows by iii) that

$$\frac{1}{\tau}\int \phi(w, F_\tau)F_0(dw) \longrightarrow -\int \phi(w, F_0)H(dw) + 0 = -\int \phi(w, F_0)H(dw). \tag{7.1}$$

By ii) we have $\int \phi(w, F_0)F_0(dw) = 0$, so that $\int \phi(w, F_\tau)F_0(dw)$ is differentiable in τ from the right at $\tau = 0$ and

$$\frac{d}{d\tau}\int \phi(w, F_\tau)F_0(dw) = -\int \phi(w, F_0)H(dw) = -\frac{d}{d\tau}\int g(w, \gamma(F_\tau), \theta)F_0(dw),$$

where the last equality follows by i). Adding $d\int g(w, \gamma(F_\tau), \theta)F_0(dw)/d\tau$ to both sides of this equation gives equation (4.1). *Q.E.D.*

Proof of Theorem 2: Since $\gamma(F_\tau^\alpha) = \gamma_0$ is a constant hypothesis ii) implies that

$$E[\phi(W, \gamma_0, \alpha, \theta)] = \int \phi(w, \gamma_0, \alpha, \theta)F_0(dw) = d\int g(w, \gamma(F_\tau^\alpha), \theta)F_\alpha(dw)/d\tau$$
$$= d\int g(w, \gamma_0, \theta)F_\alpha(dw)/d\tau = 0. \ Q.E.D.$$

Proof of Theorem 3: By ii), the chain rule for Hadamard derivatives (see 20.9 of Van der Vaart, 1998), and by eq. *(4.1)* it follows that for $\delta_H = d\gamma(F_\tau)/d\tau$,

$$\bar{\psi}_\gamma(\delta_H, \alpha_0, \theta_0) = \frac{\partial \bar{\psi}(\gamma(F_\tau), \alpha_0, \theta)}{\partial \tau} = 0.$$

Equation (*4.3*) follows by $\bar{\psi}_\gamma(\delta, \alpha_0, \theta_0)$ being a continuous linear function and iii). Equation (4.4) follows by Proposition 7.3.3 of Luenberger (1969). *Q.E.D.*

Proof of Theorem 4: Suppose that $\psi(w, \gamma, \alpha, \theta)$ is doubly robust. Then for any $\delta \in \Gamma$ and any t it follows that $\gamma_0 + t\delta \in \Gamma$ so that

$$\bar{\psi}(\gamma_0 + t\delta, \alpha_0, \theta) = \bar{\psi}(\gamma_0, \alpha_0, \theta),$$

identically in t. Differentiating with respect to t gives $d\bar{\psi}(\gamma_0 + t\delta, \alpha_0, \theta)/dt = 0$. Also, $\bar{\psi}(\gamma, \alpha_0, \theta)$ is constant as a function of γ and so is affine. Now suppose that $\bar{\psi}(\gamma, \alpha_0, \theta)$ is affine in γ. Then for any t,

$$\bar{\psi}(\gamma_0 + t(\gamma - \gamma_0), \alpha_0, \theta) = \bar{\psi}((1-t)\gamma_0 + t\gamma, \alpha_0, \theta) =$$
$$(1-t)\bar{\psi}(\gamma_0, \alpha_0, \theta) + t\bar{\psi}(\gamma, \alpha_0, \theta) = \bar{\psi}(\gamma_0, \alpha_0, \theta) + t[\bar{\psi}(\gamma, \alpha_0, \theta) - \bar{\psi}(\gamma_0, \alpha_0, \theta)].$$

It then follows from $d\bar{\psi}(\gamma_0 + t\delta, \alpha_0, \theta)/dt = 0$ for $\delta = \delta - \delta_0 \in \Gamma$ that

$$0 = \frac{d}{dt}\bar{\psi}(\gamma_0 + t(\gamma - \gamma_0), \alpha_0, \theta) = \bar{\psi}(\gamma, \alpha_0, \theta) - \bar{\psi}(\gamma_0, \alpha_0, \theta),$$

i.e. that $\bar{\psi}(\gamma, \alpha_0, \theta) = \bar{\psi}(\gamma_0, \alpha_0, \theta)$. *Q.E.D.*

Proof of Theorem 5: By orthogonality of least squares residuals $E[\alpha_0(X)\{Y - \gamma_0(X)\}] = 0$ so that

$$\begin{aligned}\bar{\psi}(\gamma, \alpha_0, \theta) &= E[m(W, \gamma)] - \theta + E[\alpha_0(X)\{Y - \gamma(X)\}] \\ &= E[\alpha_0(X)\gamma(X)] - \theta + E[\alpha_0(X)\gamma_0(X)] - E[\alpha_0(X)\gamma(X)] = \theta_0 - \theta. \ Q.E.D.\end{aligned}$$

Proof of Theorem 6: By i), ii), iii), iterated expectations, and $proj_{F_0}(\gamma(Z)|\mathcal{B}) \in \mathcal{A}$, for all $\gamma \in \Gamma$,

$$\begin{aligned}E[m(W, \gamma)] &= E[v_m(Z)\gamma(Z)] = E[proj_{F_0}(b_m(X)|\Gamma)\gamma(Z)] = E[b_m(X)proj_{F_0}(\gamma(Z)|\mathcal{B})] \\ &= E[\alpha_0(X)proj_{F_0}(\gamma(Z)|\mathcal{B})] = E[\alpha_0(X)\gamma(Z)].\end{aligned}$$

Therefore for $\tilde{\gamma} = \gamma - \gamma_0$ it follows by $E[\alpha_0(X)\{Y - \gamma(Z)\}] = -E[\alpha_0(X)\tilde{\gamma}(Z)]$ that

$$\begin{aligned}\bar{\psi}(\gamma, \alpha_0, \theta) &= E[m(W, \gamma)] - \theta + E[\alpha_0(X)\{Y - \gamma(X)\}] \\ &= E[m(W, \tilde{\gamma})] - \theta + \theta_0 - E[\alpha_0(X)\tilde{\gamma}(Z)] = -\theta + \theta_0. \ Q.E.D.\end{aligned}$$

Proof of Theorem 7: For $\tilde{\gamma} = \gamma - \gamma_0$,

$$\bar{\psi}(\gamma, \alpha_0, \theta) = E[m(W, \gamma)] - \theta + E[\alpha_0(X)] - \int \alpha_0(u)\gamma(u)du$$
$$= \theta_0 - \theta + \int \alpha_0(u)\{\gamma(u) - \gamma_0(u)\}du + \int \alpha_0(u)\{\gamma_0(u) - \gamma(u)\}du = \theta_0 - \theta. \ Q.E.D.$$

Proof of Lemma 8: Define

$$\hat{R}_{1\ell i} := g(W_i, \hat{\gamma}_\ell, \theta_0) - g(W_i, \gamma_0, \theta_0), \ \hat{R}_{2\ell i} := \phi(W_i, \hat{\gamma}_\ell, \alpha_0, \theta_0) - \phi(W_i, \gamma_0, \alpha_0, \theta_0), \tag{7.2}$$
$$\hat{R}_{3\ell i} := \phi(W_i, \gamma_0, \hat{\alpha}_\ell, \hat{\theta}_\ell) - \phi(W_i, \gamma_0, \alpha_0, \theta_0), \ i \in I_\ell.$$

Then for $\hat{\Delta}_\ell(W_i)$ as in Assumption 2 we have

$$g(W_i, \hat{\gamma}_\ell, \theta_0) + \phi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \psi(W_i, \gamma_0, \alpha_0, \theta_0) = \hat{R}_{1\ell i} + \hat{R}_{2\ell i} + \hat{R}_{3\ell i} + \hat{\Delta}_\ell(W_i). \tag{7.3}$$

Let $\mathcal{W}_\ell^c$ denote the observations not in I_ℓ, so that $\hat{\gamma}_\ell$, $\hat{\alpha}_\ell$, and $\hat{\theta}_\ell$ depend only on $\mathcal{W}_\ell^c$. Therefore by $0 = E[g(W, \gamma_0, \theta_0)] = E[\phi(W, \gamma_0, \alpha_0, \theta_0)]$ and Assumption 3 i) (which comes form Theorem 2),

$$E[\hat{R}_{1\ell i} + \hat{R}_{2\ell i}|\mathcal{W}_\ell^c] = \int [g(w, \hat{\gamma}_\ell, \theta_0) + \phi(w, n\hat{\gamma}_\ell, \alpha_0, \theta_0)]F_0(dw) = \bar{\psi}(\hat{\gamma}_\ell, \alpha_0, \theta_0), \tag{7.4}$$
$$E[\hat{R}_{3\ell i}|\mathcal{W}_\ell^c] = \int \phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell)F_0(dw) = 0.$$

Then by Assumption 3,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_\ell} E[\hat{R}_{1\ell i} + \hat{R}_{2\ell i} + R_{3\ell i}|\mathcal{W}_\ell^c] \right\| = \frac{n_\ell}{\sqrt{n}} \left\| \bar{\psi}(\hat{\gamma}_\ell, \alpha_0, \theta_0) \right\| \leq \sqrt{n} \left\| \bar{\psi}(\hat{\gamma}_\ell, \alpha_0, \theta_0) \right\| \xrightarrow{p} 0. \tag{7.5}$$

Also by observations in I_ℓ mutually independent conditional on $\mathcal{W}_\ell^c$ and Assumption 1,

$$E[\left\{ \frac{1}{\sqrt{n}} \sum_{i \in I_\ell} (\hat{R}_{j\ell i} - E[\hat{R}_{j\ell i}|\mathcal{W}_\ell^c]) \right\}^2 |\mathcal{W}_\ell^c] = \frac{n_\ell}{n} Var(\hat{R}_{j\ell i}|\mathcal{W}_\ell^c) \leq E[\hat{R}_{j\ell i}^2|\mathcal{W}_\ell^c] \xrightarrow{p} 0, \ (j = 1, 2, 3).$$

Then by the triangle and conditional Markov inequalities,

$$\frac{1}{\sqrt{n}} \sum_{i \in I_\ell} (\hat{R}_{1\ell i} + \hat{R}_{2\ell i} + \hat{R}_{3\ell i} - E[\hat{R}_{1\ell i} + \hat{R}_{2\ell i} + \hat{R}_{3\ell i}|\mathcal{W}_\ell^c]) \xrightarrow{p} 0. \tag{7.6}$$

By equations (7.5) and (7.6) and the triangle inequality $\sum_{i \in I_\ell}(\hat{R}_{1\ell i} + \hat{R}_{2\ell i} + \hat{R}_{3\ell i})/\sqrt{n} \xrightarrow{p} 0$. It also follows from Assumption 2 i) similarly to equation (7.6), or Assumption 2 ii) by the triangle inequality, or just by Assumption 2 iii) that $\sum_{i \in I_\ell} \hat{\Delta}_{\ell i}(W_i)/\sqrt{n} \xrightarrow{p} 0$. The conclusion then follows by the triangle inequality and equations (7.3), $Q.E.D.$

REFERENCES

Ackerberg, D., X. Chen, and J. Hahn (2012): "A Practical Asymptotic Variance Estimator for Two-step Semiparametric Estimators," *Review of Economics and Statistics* 94: 481–498.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-Step GMM," *The Review of Economic Studies* 81: 919–943.

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica* 71, 1795-1843.

Ai, C. and X. Chen (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics* 141, 5–43.

Andrews, D.W.K. (1994): "Asymptotics for Semiparametric Models via Stochastic Equicontinuity," *Econometrica* 62, 43-72.

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, 225-235.

Athey, S., G. Imbens, and S. Wager (2018): "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society, Series B,* 80, 597-623.

Avagyan, V. and S. Vansteelandt (2017): "High-dimensional Inference for the Average Treatment Effect Under Model Misspecification Using Penalized Bias-Reduced Doubly-Robust Estimation," *Biostatistics and Epidemiology*, DOI: 10.1080/24709360.2021.1898730

Bajari, P., V. Chernozhukov, H. Hong, and D. Nekipelov (2009): "Nonparametric and Semiparametric Analysis of a Dynamic Discrete Game," working paper, Stanford.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010): "Estimating Static Models of Strategic Interactions," *Journal of Business and Economic Statistics* 28, 469-482.

Belloni, A., and V. Chernozhukov (2011): "ℓ1-Penalized Regression in High-Dimensional Sparse Models," *Annals of Statistics* 9, 82-130.

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, 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies* 81, 608–650.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica* 85, 233-298.

Bickel, P.J. (1982): "On Adaptive Estimation," *Annals of Statistics* 10, 647–671.

Bickel, P.J. and Y. Ritov (1988): "Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates," *Sankhyā: The Indian Journal of Statistics, Series A* 238, 381-393.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag, New York.

Blomquist, S. and M. Dahlberg (1999): "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments," *Journal of Applied Econometrics* 14, 69-88.

Bonhomme, S., and M. Weidner (2018): "Minimizing Sensitivity to Misspecification," arxiv.

Bravo, F., J.C. Escanciano, and I. van Keilegom (2020): "Two-step Semiparametric Likelihood Inference," *Annals of Statistics* 48, 1-26.

Carone, M., A.R. Luedtke, and M.J. van der Laan (2019): "Toward Computerized Efficient Estimation in Infinite Dimensional Models," *Journal of the American Statistical Association* 114, 1174-1190.

Cattaneo, M.D., and M. Jansson (2018): "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica* 86, 955–995.

Cattaneo, M.D., M. Jansson, and X. Ma (2018): "Two-step Estimation and Inference with Possibly Many Included Covariates," *Review of Economic Studies* 86, 1095–1122.

Chaudhuri, P.K. Doksum, and A. Samarov (1997): "On Average Derivative Quantile Regression," *Annals of Statistics* 25, 715-744.

Chen, X., O.B. Linton, and I. van Keilegom (2003): “Estimation of Semiparametric Models when the Criterion Function Is Not Smooth,” *Econometrica* 71, 1591-1608.

Chen, X., and Z. Liao (2015): “Sieve Semiparametric GMM Under Weak Dependence," *Journal of Econometrics* 189, 163-186.

Chernozhukov, V., J.C. Escanciano, H. Ichimura, W.K. Newey (2016): "Locally Robust Semiparametric Estimation," arxiv.org/pdf/1608.00033v1.pdf.

Chernozhukov, V., C. Hansen, and M. Spindler (2015): "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics* 7: 649–688.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins (2018): "Debiased/Double Machine Learning for Treatment and Structural Parameters,*Econometrics Journal* 21, C1-C68.

Chernozhukov, V., W.K. Newey, and J. Robins (2018): "Double/De-Biased Machine Learning Using Regularized Riesz Representers," arxiv.org/abs/1802.08667v1.

Chernozhukov, V., W.K. Newey, and R. Singh (2018): "Learning L2-Continuous Regression Functionals via Regularized Riesz Representers," https://arxiv.org/pdf/1809.05224v1.pdf.

Escanciano, J-C. and W. Li (2021): "Optimal Linear Instrumental Variables Approximations," *Journal of Econometrics* 221, 223-246.

Farrell, M. (2015): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *Journal of Econometrics* 189, 1–23.

Firpo, S. and C. Rothe (2019): “Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically,” *Econometric Theory* 35, 1048–1087.

Foster, D.F. and V. Syrgkanis (2019): "Orthogonal Statistical Learning," arxiv.

Graham, B.W. (2011): "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica* 79, 437–452.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66, 315-331.

Hahn, J. and G. Ridder (2013): "Asymptotic Variance of Semiparametric Estimators With Generated Regressors," *Econometrica* 81, 315-340.

Hahn, J. and G. Ridder (2019): “Three-stage Semi-Parametric Inference: Control Variables and Differentiability," *Journal of Econometrics* 211, 262-293.

Hampel, F.R. (1974): "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association* 69, 383-393.

Hasminskii, R.Z. and I.A. Ibragimov (1978): "On the Nonparametric Estimation of Functionals," *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics*, 41-51.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71: 1161–1189.

Hirshberg, D.A. and S. Wager (2019): "Augmented Minimax Linear Estimation," *Annals of Statistics*, forthcoming.

Hotz, V.J. and R.A. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies* 60, 497-529.

Huber, P. (1981): *Robust Statistics,* New York: Wiley.

Ichimura, H. (1993): "Estimation of Single Index Models," *Journal of Econometrics* 58, 71-120.

Ichimura, H. and W.K. Newey (2021): "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, forthcoming.

Klaassen, C.A.J. (1987): "Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators," *Annals of Statistics* 15, 1548-1562.

Klein, R. and R.H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387-421.

Leeb, H. and B.M. Potscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21, 21-59.

Luenberger, D.G. (1969): *Optimization by Vector Space Methods*, New York: Wiley.

Newey, W.K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* 5, 99-135.

Newey, W.K. (1991): ”Uniform Convergence in Probability and Stochastic Equicontinuity,” *Econometrica* 59, 1161-1167.

Newey, W.K. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

Newey, W.K. (1994b): ”Kernel Estimation of Partial Means and a General Variance Estimator,” *Econometric Theory* 10, 233-253.

Newey, W.K., and J.L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica* 71, 1565-1578.

Newey, W.K., F. Hsieh, and J.M. Robins (1998): “Undersmoothing and Bias Corrected Functional Estimation," MIT Dept. of Economics working paper 72, 947-962.

Newey, W.K., F. Hsieh, and J.M. Robins (2004): “Twicing Kernels and a Small Bias Property of Semiparametric Estimators,” *Econometrica* 72, 947-962.

Newey, W.K., and J. Robins (2017): "Cross Fitting and Fast Remainder Rates for Semiparametric Estimation," CEMMAP Working paper WP41/17.

Neyman, J. (1959): “Optimal Asymptotic Tests of Composite Statistical Hypotheses,” *Probability and Statistics, the Harald Cramer Volume*, ed., U. Grenander, New York, Wiley.

Pfanzagl, J., and W. Wefelmeyer (1982): *Contributions to a General Asymptotic Statistical Theory,* Springer Lecture Notes in Statistics.

Powell, J.L., J.H. Stock, and T.M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

Robins, J.M., A. Rotnitzky, and L.P. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89: 846–866.

Robins, J.M.,and A. Rotnitzky (2001): Comment on “Semiparametric Inference: Question and an Answer,” by P.A. Bickel and J. Kwon, *Statistica Sinica* 11, 863-960.

Robins, J.M., L. Li, E. Tchetgen, and A. van der Vaart (2008): "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," *IMS Collections Probability and Statistics: Essays in Honor of David A. Freedman, Vol 2,* 335-421.

Robins, J., P. Zhang, R. Ayyagari, R. Logan, E. Tchetgen, L. Li, A. Lumley, and A. van der Vaart (2013): "New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research," Research Report Health E Inst.

Robinson, P.M. (1988): "'Root-N-consistent Semiparametric Regression," *Econometrica* 56, 931-954.

Rust, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica* 55, 999-1033.

Schick, A. (1986): "On Asymptotically Efficient Estimation in Semiparametric Models," *Annals of Statistics* 14, 1139-1151.

Semenova, V. (2018): "Machine Learning for Set-Identified Linear Models," arxiv.

Severini, T. and G. Tripathi (2012): "Efficiency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors," *Journal of Econometrics* 170, 491-498.

Singh, R., and L. Sun (2019): "De-biased Machine Learning in Instrumental Variable Models for Treatment Effects," https://arxiv.org/pdf/1909.05244.pdf.

Tan, Z. (2020): "Model-Assisted Inference for Treatment Effects Using Regularized Calibrated Estimation with High-Dimensional Data," *Annals of Statistics* 48, 811-837.

Tchetgen Tcehtgen, E.J. (2009): "A Commentary on G. Molenburgh's Review of Missing Data Methods," *Drug Information Journal* 43, 433-435.

van der Laan, M. J. (2014), "Targeted Estimation of Nuisance Parameters to Obtain Valid Statistical Inference," *International Journal of Biostatistics* 10, 29–57.

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

van der Laan, M.J. and D. Rubin (2006): "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics* 2.

van der Vaart, A.W. (1991): "On Differentiable Functionals," *The Annals of Statistics,* 19, 178-204.

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

Vermeulen, K. and S. Vansteelandt (2015): "Bias-Reduced Doubly Robust Estimation," *Journal of the American Statistical Association* 110, 1024-1036.

Von Mises, R. (1947), "On the Asymptotic Distribution of Differentiable Statistical Functions," *Annals of Mathematical Statistics* 18, 309-34.