Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation

Whitney K. Newey* James M. Robins†
MIT *Harvard*

January 2018

Abstract

There are many interesting and widely used estimators of a functional with finite semiparametric variance bound that depend on nonparametric estimators of nuisance functions. We use cross-fitting (i.e. sample splitting) to construct novel estimators with fast remainder rates. We give cross-fit doubly robust estimators that use separate subsamples to estimate different nuisance functions. We obtain general, precise results for regression spline estimation of average linear functionals of conditional expectations with a finite semiparametric variance bound. We show that a cross-fit doubly robust spline regression estimator of the expected conditional covariance is semiparametric efficient under minimal conditions. Cross-fit doubly robust estimators of other average linear functionals of a conditional expectation are shown to have the fastest known remainder rates for the Haar basis or under certain smoothness conditions. Surprisingly, the cross-fit plug-in estimator also has nearly the fastest known remainder rate, but the remainder converges to zero slower than the cross-fit doubly robust estimator. As specific examples we consider the expected conditional covariance, mean with randomly missing data, and a weighted average derivative.

Keywords: Semiparametric estimation, semiparametric efficieny, bias, smoothness.

*Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: wnewey@mit.edu.

†Harvard School of Public Health, 677 Huntington Ave, Kresge Building Room 823, Boston, MA 02115 U.S.A E-mail: robins@hsph.harvard.edu.

1 Introduction

There are many interesting and widely used estimators of a functional with finite semi-parametric variance bound that depend on the estimation, in a first step, of nuisance functions, such as conditional expectations or densities. Examples include estimators of the mean with data missing at random, the average treatment effect, the expected conditional covariance, partially linear models, and weighted average derivatives. Because the nuisance functions can often be high dimensional it is desirable to minimize the impact of estimating these functions. By using cross-fitting (i.e. sample splitting) to estimate the nuisance functions we obtain novel estimators whose second order remainders converge to zero as fast as known possible. In particular, such estimators are often root-n consistent under minimal smoothness conditions. Furthermore, such estimators may have higher order mean square error that converges to zero as fast as known possible.

Bias reduction is key to constructing semiparametric estimators with fast remainder rates. The rates at which the variance of remainders goes to zero are quite similar for different semi-parametric estimators but the bias rates differ greatly. We use cross-fitting for bias reduction. We show how fast remainder rates can be attained by using different parts of an i.i.d. sample to estimate different components of an estimator.

In this paper we consider regression spline estimation of average linear functionals of conditional expectations with a finite semiparametric variance bound, as we have been able to obtain general, precise results for functionals in this class. The class includes the five examples mentioned above.

We define a cross fit (CR) plug-in estimator to be one where we estimate the functional by simply replacing the unknown conditional expectation by a nonparametric estimator from a separate part of the sample. Cross-fitting eliminates an "own observation" bias term, thereby decreasing the size of the remainder. Functionals in our class have doubly robust influence functions that depend on two unknown functions. This implies there exists an estimator depending on both unknown functions that has exact bias zero if the unknown functions are replaced by fixed functions, at least one of which is equal to the truth. Here we use double cross-fitting where the two unknown functions are themselves estimated from separate subsamples, so that the final estimator depends on three separate subsamples. Surprisingly, single cross fitting in which both unknown functions are estimated from the same subsample has a remainder that can converge even slower than CF plug-in estimators. In contrast, doubly robust estimators with double cross fitting improve on cross-fit plug-in estimators in the sense that remainder terms can converge at faster rates. We also show how multiple cross-fitting could be used to reduce bias for any semiparametric estimator that is a polynomial in first step spline estmators of unknown functions.

We construct cross-fit (CF) plug-in and doubly cross-fit doubly robust (DCDR) estimators that are semiparametrically efficient under minimal conditions when the nuisance functions are in a Holder class of order less than or equal to one. When a nuisance function is Holder of order exceeding one, we propose DCDR estimators that have remainders that converge no slower and often faster than the CF plug-in estimator. In the special case of the expected conditional covariance functional, the DCDR estimator is always semiparametric efficient under minimal conditions. For other functionals in our class the CF plug-in and DCDR estimator are semiparametric efficient under minimal conditions, provided the conditional expectation is Holder of order greater than or equal to one-half the regressor dimension; furthermore, in this case, the remainder goes to zero as fast as known possible for both CF plug-in and DCDR estimators. When the conditional expectation is Holder of order less than or equal to one-half the regressor dimension but greater than or equal to one, the remainder for the DCDR has a remainder that converges faster than the CF plug-in estimator.

In the case where the conditional expectation is Holder of order no less than one but less than one-half the regressor dimension, we show semiparametric efficiency under minimal conditions for the expected conditional covariance, but not for other functionals. The higher order influence function (HOIF) estimators of Robins et al. (2008, 2017) and Mukherjee, Newey, and Robins (2017) will be semiparametric efficient under minimal conditions for these other functionals, including the mean with data missing at random and the average treatment effect.

CF plug-in estimators have been considered by Bickel (1982) in the context of adaptive semiparametric efficient estimation, Powell, Stock, and Stoker (1989) for density weighted average derivatives, and by many others. Kernel and series CF plug-in estimators of the integrated squared density and certain other functionals of a density have been shown to be semiparametric efficient under minimal conditions by Bickel and Ritov (1988), Laurent (1996), Newey, Hsieh, and Robins (2004), and Gine and Nickl (2008). Our DCDR estimator appears to be novel as does the fact that a CF plug-in estimator can be semiparametric efficient under minimal conditions. Ayyagari (2010), Robins et al. (2013), Kandasamy et. al. (2015), Firpo and Rothe (2016), and Chernozhukov et al.(2017) have considered doubly robust estimators that eliminate own observation terms. Double cross-fitting in double robust estimation appears not to have been analyzed before.

Our results for splines make use of the Rudelson (1999) law of large numbers for matrices similarly to Belloni et al.(2015). The results for the CF plug-in estimator for general splines extend those of Ichimura and Newey (2017) to sample averages of functionals. The double robustness of the influence function for the functionals we consider is shown in Chernozhukov et al.(2016), where the doubly robust estimators of Scharfstein, Rotnitzky, and Robins (1999), Robins, Rotnitzky, and van der Laan (2000), Robins et. al. (2008), and Firpo and Rothe (2016) are extended to a wide class of average linear functionals of expectations.

The DCDR estimator for the mean with missing data and average treatment effect uses a spline approximation to the reciprocal of the propensity score rather than the reciprocal of a propensity score estimator. The reciprocal of a propensity score estimator has been used in much of the previous literature on plug in and doubly robust estimation, including Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Hahn (1998), and Hirano, Imbens, and Ridder (2003). Estimators based on approximating the reciprocal of the propensity score have been considered by Robins et al. (2007), Athey, Imbens, and Wager (2017), and recently in independent work by Hirschberg and Wager (2017).

Other approaches to bias reduction for semiparametric estimators have been proposed. Robins et al.(2008, 2017) and Mukherjee, Newey, and Robins (2017) develop higher order influence function (HOIF) estimators with smaller bias. In Section 2 we will discuss the relationship of this paper to HOIF. Cattaneo and Jansson (2017) propose promising bootstrap confidence intervals for plug-in kernel estimators that include bias corrections. Also, Cattaneo, Jansson, and Ma (2017) show that the jackknife can be used to reduce bias of plug-in series estimators. For the class of functionals in this paper cross-fitting removes bias so that there is no need for bootstrap or jackknife bias corrections in order to attain the fastest remainder rates.

In Section 2 we will describe the cross-fitting approach to bias reduction and show how it relates to HOIF. Section 3 describes the linear functionals and regression spline estimators we consider. Sections 4, 5, and 6 give results for the CF plug-in estimator, the DCDR expected conditional covariance estimator, and DCDR estimators of other linear functionals, respectively.

Before explaining the results of this paper it is helpful to be more specific about our goal. We will consider i.i.d. data $z_1, ..., z_n$. We are interested in an asymptotically linear semiparametric estimator $\hat{\beta}$ satisfying

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(z_i\right)+O_p\left(\Delta_n\right), \Delta_n \longrightarrow 0, \tag{1.1}$$

where $\psi(z)$ is the influence function of $\hat{\beta}$ and Δ_n characterizes the size of the remainder. Our goal is to find estimators where Δ_n converges to zero at the fastest known rate.

For the integrated squared density, Bickel and Ritov (1988) gave a kernel based estimator where the rate for Δ_n is fast enough that $\hat{\beta}$ is semiparametric efficient under minimal conditions.

To motivate our candidate for the optimal rate the remainder can converge to zero for series estimators of an average linear functional of a conditional expectation with positive information bound, we consider the series estimator of the coefficients of a partially linear regression in Donald and Newey (1994). The model there is $E[y_i|a_i, x_i] = a_i^T\beta_0 + \lambda_0(x_i)$ where $\lambda_0(x_i)$ is an unknown function of an $r \times 1$ vector x_i. Consider the estimator $\hat{\beta}$ obtained from regressing y_i on a_i and a $K \times 1$ vector $p(x_i)$ of power series or regression splines in an i.i.d. sample of size n. Assume that the functions $\lambda_0(x)$ and $\alpha_0(x) = E[a_i|x_i = x]$ are each members of a Holder class of order s_λ and s_α respectively. Define

$$\Delta_n^* = \sqrt{n}K^{-(s_\gamma+s_\alpha)/r} + K^{-s_\gamma/r} + K^{-s_\alpha/r} + \sqrt{\frac{K}{n}}.$$

Donald and Newey (1994) showed that under regularity conditions, including $K/n \longrightarrow 0$, equation (1.1) is satisfied with $\Delta_n = \Delta_n^*$. Here $\sqrt{n}K^{-(s_\gamma+s_\alpha)/r}$ gives the rate at which the bias of $\sqrt{n}(\hat{\beta} - \beta_0)$ goes to zero. Also, $K^{-s_\gamma/r}$ and $K^{-s_\alpha/r}$ are stochastic equicontinuity bias terms, and $\sqrt{K/n}$. that accounts for stochastic equicontinuity and degenerate U-statistic variance terms. Furthermore, there exists $K = K_n$ satisfying $K_n/n \longrightarrow 0$ such that $\Delta_n^* \longrightarrow 0$ if and only if $s_\gamma + s_\alpha > r/2$. However the Donald and Newey (1994) result used the fact that the partially linear model implies $y_i - a_i^T\beta_0$ is mean independent of a_i given x_i and thus is not a locally nonparametric model. A model is said to be locally nonparametric if, at each law P in the model, the tangent space is all of $L_2(P)$ .Henceforth in this paper, we shall always assume a locally nonparametric model.

Robins et al. (2009) showed that the condition $s_\gamma + s_\alpha > r/2$ is necessary and sufficient for the existence of a semparametric efficient estimator of

$$\xi_0 = E\left[cov\left(a_i, y_i|x_i\right)\right]/E\left[var(a_i|x_i)\right],$$

Note ξ_0 is the probability limit of the Donald and Newey (1994) estimator regardless of whether the partially linear model holds. That is, ξ_0 is the coefficient b in the population linear projection of y_i on all functions of the form $a_i b + \lambda(x_i)$. Robins et al. (2008) proved sufficiency using a higher order influence function estimator of ξ_0, which is a U-statistic whose order increases as $\ln(n)$. In contrast, the aforementioned estimator of Donald and Newey (1994), although much simpler, is not semiparametric efficient for ξ_0 in a locally nonparametric model under the minimal condition $s_\gamma + s_\alpha > r/2$. The current paper was thus motivated by the question whether one could construct a simple efficient estimator of ξ_0 whose remainder Δ_n will go to zero as fast as Δ_n^*, the fastest rate known to be possible. In summary, our goal is to construct estimators that are much simpler than the HOIF estimators and yet satisfy equation (1.1) with $\Delta_n = \Delta_n^*$.

2 Cross-Fitting and Fast Remainder Rates

To explain how cross-fitting can help achieve fast remainder rates we consider estimation of the expected conditional covariance

$$\beta_0 = E[Cov(a_i, y_i|x_i)] = E[a_i\{y_i - \gamma_0(x_i)\}],$$

where $\gamma_0(x_i) = E[y_i|x_i]$. This object is useful in the estimation of weighted average treatment effects as further explained below. We assume that the functions $\gamma_0(x)$ and $\alpha_0(x) = E[a_i|x_i = x]$ are each members of a Holder class of order s_γ and s_α respectively.

One way to construct an estimator of β_0 is the "plug-in" method where a nonparametric estimator $\hat{\gamma}$ is substituted for γ_0 and a sample average for the expectation to form

$$\bar{\beta} = \frac{1}{n}\sum_{i=1}^{n} a_i\{y_i - \hat{\gamma}(x_i)\}.$$

This estimator generally suffers from an "own observation" bias that is of order $K/\sqrt{n}$ when $\hat{\gamma}$ is a spline regression estimator, which converges to zero slower than Δ_n^*. This bias can be eliminated by replacing $\hat{\gamma}(x)$ with an estimator $\hat{\gamma}_{-i}(x)$ that does not use z_i in its construction. The resulting estimator of β_0 is

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} a_i\{y_i - \hat{\gamma}_{-i}(x_i)\}.$$

This estimator is a cross-fit (CF) plug-in estimator in the sense that $\hat{\gamma}_{-i}$ uses a subsample that does not include i. The cross-fitting eliminates the own observation bias. The remainder rate Δ_n for $\hat{\beta}$ will be often be faster than for $\bar{\beta}$, sometimes as fast as Δ_n^* as explained below. This approach to eliminating own observation bias when the first step is a density estimator has been used by Bickel (1982), Bickel and Ritov (1988), Powell, Stock, and Stoker (1989), Laurent (1996), and others. Here we obtain the novel result that, for a spline regression first step, a CF plug-in estimator can have the fastest rate Δ_n^* even when the usual plug-in estimator does not.

Doubly robust estimators have another source of bias that can also be eliminated by double cross-fitting. To explain we consider a single cross-fit doubly robust estimator of the expected conditional covariance. Let $\hat{\gamma}_{-i}(x)$ and $\hat{\alpha}_{-i}(x)$ be nonparametric estimators of $\gamma_0(x_i) = E[y_i|x_i]$ and $\alpha_0(x_i) = E[a_i|x_i]$ that do not depend on the i^{th} observation. Consider the estimator

$$\check{\beta} = \frac{1}{n}\sum_{i=1}^{n}[a_i - \hat{\alpha}_{-i}(x_i)][y_i - \hat{\gamma}_{-i}(x_i)].$$

This estimator is doubly robust in the sense of Scharfstein, Rotnitzky, and Robins (1999) and Robins, Rotnitzky, and van der Laan (2000), being consistent if either $\hat{\alpha}_{-i}$ or $\hat{\gamma}_{-i}$ are consistent. It uses cross-fitting to eliminate own observation bias. This estimator does have a nonlinearity bias since $\hat{\alpha}_{-i}(x_i)$ and $\hat{\gamma}_{-i}(x_i)$ are constructed from the same data in single crossfitting. That bias is of the same order $K/\sqrt{n}$ as the own observation bias for a spline regression plug-in estimator. This bias can be thought of as arising from nonlinearity of $\check{\beta}$ in the two nonparametric estimators $\hat{\alpha}_{-i}(x_i)$ and $\hat{\gamma}_{-i}(x_i)$.

One can remove the nonlinearity bias in the doubly robust estimator by using different parts of the data to construct the two nonparametric estimators. Let $\hat{\gamma}_{-i}(x_i)$ be constructed from a subset of the observations that does not include observation i and let $\tilde{\alpha}_{-i}(x_i)$ be constructed from a subset of the observations that does not include i or any observations used to form $\hat{\gamma}_{-i}$.

A doubly cross-fit doubly robust estimator (DCDR) is

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n}[a_i - \tilde{\alpha}_{-i}(x_i)][y_i - \hat{\gamma}_{-i}(x_i)].$$

This estimator uses cross-fitting to remove both the own observation and the nonlinearity biases. We will show that $\Delta_n^* = \Delta_n$ when $\tilde{\alpha}_{-i}(x_i)$ and $\hat{\gamma}_{-i}(x_i)$ are spline regression estimators for a $K \times 1$ vector of multivariate splines of at least order $\max\{s_\gamma, s_a\} - 1$ with evenly spaced knots. Consequently, this estimator will be root-n consistent and semiparametric efficient when $s_\gamma + s_\alpha > r/2$ and K is chosen appropriately, which is the minimal condition of Robins et al. (2009).

Remarkably, the doubly robust estimator $\check{\beta}$ where $\hat{\alpha}_{-i}(x_i)$ and $\hat{\gamma}_{-i}(x_i)$ use the same data may have a slower remainder rate than the CF plug-in estimator $\hat{\beta}$. The use of the same data for $\hat{\alpha}_{-i}(x_i)$ and $\hat{\gamma}_{-i}(x_i)$ introduces a bias term of size $K/\sqrt{n}$. Such a term is not present in the CF plug-in estimator. The $K/\sqrt{n}$ term is eliminated for the doubly robust estimator by forming $\tilde{\alpha}_{-i}(x_i)$ and $\hat{\gamma}_{-i}(x_i)$ from different samples. We find that the DCDR estimator $\tilde{\beta}$ improves on the CF plug in estimator by increasing the rate at which a certain part of Δ_n goes to zero. Specifics will be given below.

We note that the own observation bias can also be thought of as nonlinearity bias. The parameter β_0 has the form

$$\beta_0 = \int a\{y - \gamma_0(x)\}F_0(dz),$$

where F_0 denotes the distribution of $z = (y, a, x)$. This object is quadratic in γ_0 and F_0 jointly. The own observation bias can be thought of as a quadratic bias resulting from using all the data to simultaneously estimate γ_0 and the distribution F_0 of a single observation. The CF plug in estimator $\hat{\beta}$ eliminates this nonlinearity bias. Also, the doubly robust estimator can be thought of as estimating $\int[a - \alpha_0(x)][y - \gamma_0(x)]F_0(dz)$, which is cubic in α_0, γ_0, and F_0 jointly. The DCDR estimator can be thought of as eliminating the cubic bias by estimating each of $\alpha_0(x)$, $\gamma_0(x)$, and F_0 from distinct groups of observations.

One potential concern about DCDR estimators is that each of the nonparametric components $\hat{\gamma}$ and $\tilde{\alpha}$ only use a fraction of the data because they are each based on subsamples that the other does not use. For example, they only use less than half the data if they are based on approximately the same subsample size. This does not affect remainder rates but could affect small sample efficiency. One might be able to improve small sample efficiency by averaging over DCDR estimators that use different sample splits to construct $\hat{\gamma}$ and $\tilde{\alpha}$, though that is beyond the scope of this paper. Our concern in this paper is remainder rates for asymptotically efficient estimation.

Cross-fitting can be applied to eliminate bias terms for any estimator that depends on powers of nonparametric estimators. Such cross-fitting would replace each power by a product of nonparametric estimators that are computed from distinct subsamples of the data, analogously to the DCDR estimators above.

We now provide a more quantitative version of our results. Let $p(x)$ be a vector of multivariate regression splines of dimension K with evenly spaced knots. We will always take $K = K_n$ to satisfy $K \ln(K)/n \to 0$. Suppose that $\hat{\gamma}_{-i}(x) = p(x)^T[\Sigma_{j\in\mathcal{I}_\ell} p(x_j)p(x_j)^T]^{-1}\Sigma_{j\in\mathcal{I}_\ell} p(x_j)y_j$ is a series estimator from regressing y_j on $p(x_j)$ in a subsample of observations indexed by $\mathcal{I}_\ell$, where $\{\mathcal{I}_\ell\}_{\ell=1}^L$ is a partition of $\{1,...,n\}$, $i \notin \mathcal{I}_\ell$, L is fixed and the number of elements of each $\mathcal{I}_\ell$ is of order n. Suppose that for the doubly robust estimator $\tilde{\alpha}(x_i)$ is constructed analogously from a separate subsample.

When $s_\gamma \le 1$ and $s_\alpha \le 1$ and $p(x)$ is a Haar basis of dummy variables that are indicator functions of cubes partitioning the support of x_i we show that the CF plug-in estimator has $\Delta_n = \Delta_n^* + \ln(n)K^{-s_\gamma/r}$ and the DCDR doubly robust estimator has $\Delta_n = \Delta_n^*$. Hence the DCDR estimator has the fast remainder rate. Further the CF plug-in estimator has the fast remainder Δ_n^*, except at those laws where $K^{-s_\gamma/r}$ is the dominating term in Δ_n^*. At such laws, the DCDR estimator improves on the CF plug-in but only by a factor of $\ln(n)$. We also show that these results extend to the entire class of average linear functionals of a conditional expectation with finite semiparametric variance bound.

When s_γ and s_α are any positive numbers and $p(x)$ is a spline basis of order at least $\max\{s_\gamma, s_\alpha\} - 1$ we show that the CF plug in estimator of the expected conditional covariance has $\Delta_n = \Delta_n^* + \sqrt{K\ln(K)/n}K^{1/2-s_\gamma/r}$ and the DCDR estimator has $\Delta_n = \Delta_n^*$.Here the plug-in estimator has the fast remainder $\Delta_n = \Delta_n^*$ for $s_\gamma > r/2$ and the doubly robust estimator has $\Delta_n = \Delta_n^*$ for all s_γ. For other functionals in our class we show that the DCDR estimator has $\Delta_n = \Delta_n^* + \sqrt{K^3\ln(K)^2/n^3}K^{1/2-s_\gamma/r}$, which has $\Delta_n = \Delta_n^*$ when $[K\ln(K)/n]K^{1/2-s_\gamma/r} \longrightarrow 0$. Thus the DCDR estimator has remainder that can converge to zero at a faster rate that of the CF plug-in estimator.

We note that the source of the term in Δ_n that is added to Δ_n^* in each case can be attributed to estimators of the second moment matrix $\Sigma = E[p(x_i)p(x_i)^T]$ of the regression splines. If each $\hat{\Sigma}_\ell$ were replaced by Σ in the estimators then the resulting objects would all have $\Delta_n = \Delta_n^*$.

For brevity, we demonstrate this only for plug-in estimator. Consider the plug-in object $\dot{\beta}$ having the same formula as $\hat{\beta}$ except that $\hat{\gamma}_{-i}(x)$ is replaced by $\dot{\gamma}_{-i}(x) = p(x)^T\Sigma^{-1}\sum_{j\in\mathcal{I}_\ell} p(x_i)y_i/n_\ell$. Let $\bar{\alpha}(x) = p(x)^T\Sigma^{-1}E[p(x_i)\alpha_0(x_i)]$. Standard approximation properties of splines give the approximation rates $\{E[\{\gamma_0(x_i) - \bar{\gamma}(x_i)\}^2]\}^{1/2} = O(K^{-s_\gamma/r})$ and $\{E[\{\alpha_0(x_i) - \bar{\alpha}(x_i)\}^2]\}^{1/2} = O(K^{-s_\alpha/r})$. By the Cauchy-Schwartz inequality

$$\begin{aligned}\sqrt{n}E[\{\alpha_0(x_i) - \bar{\alpha}(x_i)\}\{\gamma_0(x_i) - \bar{\gamma}(x_i)\}] &\le \sqrt{n}\{E[\{\alpha_0(x_i) - \bar{\alpha}(x_i)\}^2]\}^{1/2}\{E[\{\gamma_0(x_i) - \bar{\gamma}(x_i)\}^2]\}^{1/2} \\ &= O(\sqrt{n}K^{-(s_\gamma+s_\alpha)/r}).\end{aligned}$$

Note also that $E[\dot{\gamma}_{-i}(x)] = \bar{\gamma}(x) = p(x)^T\Sigma^{-1}E[p(x_i)\gamma_0(x_i)]$. Then the root-n normalized bias of

$\dot{\beta}$ is

$$
\begin{aligned}
E\left[\sqrt{n}\left(\dot{\beta}-\beta_0\right)\right] &= \sqrt{n}\int a\{y-E\left[\dot{\gamma}_{-i}(x)\right]\}F_0\left(dz\right)-E[a_i\{y_i-\gamma_0(x_i)\}] \\
&= \sqrt{n}E[a_i\{\gamma_0(x_i)-\bar{\gamma}(x_i)\}]=\sqrt{n}E[\alpha_0(x_i)\{\gamma_0(x_i)-\bar{\gamma}(x_i)\}] \qquad (2.1)\\
&= \sqrt{n}E[\{\alpha_0(x_i)-\bar{\alpha}(x_i)\}\{\gamma_0(x_i)-\bar{\gamma}(x_i)\}]=O(\sqrt{n}K^{-(s_\gamma+s_\alpha)/r}),
\end{aligned}
$$

which has our desired Δ_n^* rate. Also, there will be stochastic equicontinuity bias terms of order $K^{-s_\gamma/r}$ and $K^{-s_\alpha/r}$ and stochastic equicontinuity variance and degenerate U-statistic variance terms of order $\sqrt{K/n}$. Overall the remainder for $\dot{\beta}$ will satisfy $\Delta_n=\Delta_n^*$. Thus, a CF plug-in object $\dot{\beta}$ where Σ replaces each $\hat{\Sigma}_\ell$ will have the fast remainder rate.

We note that the bias in equation (2.1) depends on the product $K^{-(s_\gamma+s_\alpha)/r}$ of the approximation rate $K^{-s_\gamma/r}$ for $\gamma_0(x)$ and the approximation rate $K^{-s_\alpha/r}$ for $\alpha_0(x)$, rather than just the bias rate $K^{-s_\gamma/r}$ for the nonparametric estimator being plugged-in. This product form results from the fact that the parameter of interest β_0 has a finite semiparametric variance bound. The product bias form in equation (2.1) for plug-in series estimators was shown in Newey (1994).

It is interesting to compare our estimators with HOIF estimators. We continue to focus on the average conditional covariance. The HOIF estimator of that β_0 can depend on initial estimators $\hat{\gamma}(x)$ and $\hat{\alpha}(x)$ of $\gamma_0(x)$ and $\alpha_0(x)$ obtained from a training subsample. For a vector of spline regressors $p(x)$ let $\hat{\Sigma}$ be the sample second moment matrix of $p(x)$ from the training sample. Let $\hat{B}(x)=\hat{\Sigma}^{-1}[p(x)p(x)^T-\hat{\Sigma}]$ and

$$
\begin{aligned}
\hat{\beta}_H &= \frac{1}{n}\sum_{i=1}^{n}[a_i-\hat{\alpha}(x_i)][y_i-\hat{\gamma}(x_i)]-\frac{1}{n(n-1)}\sum_{i\neq j}[a_i-\hat{\alpha}(x_i)]p(x_i)^T\hat{\Sigma}^{-1}p(x_j)[y_j-\hat{\gamma}(x_j)]\\
&+\sum_{q=1}^{Q}\frac{(-1)^{q+1}(n-2-q)!}{n!}\sum_{i\neq j}[a_i-\hat{\alpha}(x_i)]p(x_i)^T\left[\sum_{\ell_1\neq\cdots\neq\ell_q\neq i\neq j}\Pi_{r=1}^{q}\hat{B}(x_{\ell_r})\right]\hat{\Sigma}^{-1}p(x_j)[y_j-\hat{\gamma}(x_j)],
\end{aligned}
$$

where all the sums are over an estimation subsample that does not overlap with the training sample. This $\hat{\beta}_H$ is the empirical HOIF estimator of Mukherjee, Newey, and Robins (2017) of order $Q+2$. By Theorem 3 of Mukherjee, Newey, and Robins (2017) the bias of $\sqrt{n}(\hat{\beta}_H-\beta_0)$ conditional on the training sample has order

$$
\sqrt{n}\left\|\hat{\alpha}-\alpha_0\right\|_2\left\|\hat{\gamma}-\gamma_0\right\|_2\left(\frac{K\ln(K)}{n}\right)^{Q/2}=\left\|\hat{\alpha}-\alpha_0\right\|_2\left\|\hat{\gamma}-\gamma_0\right\|_2K\ln(K)\left(\frac{K\ln(K)}{n}\right)^{(Q-1)/2}.
$$

where $\left\|\delta\right\|_2=\{E[\delta(x_i)^2]\}^{1/2}$. The order of this bias will be smaller than $\sqrt{K/n}$ as long as K grows no faster than $n^{1-\varepsilon}$ for some $\varepsilon>0$, although that is not needed for semiparametric efficiency. As shown in Mukherjee, Newey, and Robins (2017), if Q grows like $\sqrt{\ln(n)}$, K like $n/\ln(n)^3$, and other regularity conditions are satisfied then $\hat{\beta}_H$ will be semiparametric efficient under the minimal condition $s_\gamma+s_\alpha>r/2$ of Robins et al.(2009).

We can explain the different properties of HOIF and series estimators by comparing the CF plug-in estimator with the HOIF when the training sample estimators $\hat{\gamma}$ and $\hat{\alpha}$ are set equal to zero. In that case the HOIF estimator is

$$\hat{\beta}_H = \frac{1}{n}\sum_{i=1}^{n} a_i y_i - \frac{1}{n(n-1)}\sum_{i\neq j} a_i p(x_i)^T \hat{\Sigma}^{-1} p(x_j) y_j$$
$$+\sum_{q=1}^{Q}\frac{(-1)^{q+1}(n-2-q)!}{n!}\sum_{i\neq j} a_i p(x_i)^T\left[\sum_{\ell_1\neq\cdots\neq\ell_q\neq i\neq j}\Pi_{r=1}^{q}\hat{B}(x_{\ell_r})\right]\hat{\Sigma}^{-1}p(x_j)y_j.$$

Consider $\check{\gamma}_{-i}(x) = p(x)^T\hat{\Sigma}^{-1}\sum_{j\neq i} p(x_j)y_j/(n-1)$. This is an estimator of $\gamma_0(x)$ that is like a series estimator except the inverse second moment matrix $\hat{\Sigma}^{-1}$ comes from the training sample and the cross-moments $\sum_{j\neq i} p(x_j)y_j/(n-1)$ from the estimation subsample. The first two terms of the HOIF estimator can then be written as

$$\check{\beta} = \frac{1}{n}\sum_{i=1}^{n} a_i[y_i - \check{\gamma}_{-i}(x_i)].$$

Let T denote the training sample. Then we have

$$\begin{aligned} E[\check{\beta}-\beta_0|T] &= E[\alpha_0(x_i)\{\gamma_0(x_i)-\check{\gamma}_{-i}(x_i)\}] = E[\alpha_0(x_i)\gamma_0(x_i)] - E[\alpha_0(x_i)p(x_i)^T]\hat{\Sigma}^{-1}E[p(x_i)\gamma_0(x_i)] \\ &= E[\alpha_0(x_i)\gamma_0(x_i) - \bar{\alpha}(x_i)\bar{\gamma}(x_i)] + E[\alpha_0(x_i)p(x_i)^T](\Sigma^{-1}-\hat{\Sigma}^{-1})E[p(x_i)\gamma_0(x_i)] \\ &= O(K^{-(s_\gamma+s_\alpha)/r}) + \Lambda(\hat{\Sigma},\Sigma), \Lambda(\hat{\Sigma},\Sigma) = E[\alpha_0(x_i)p(x_i)^T](\Sigma^{-1}-\hat{\Sigma}^{-1})E[p(x_i)\gamma_0(x_i)]. \end{aligned}$$

Thus the bias of $\check{\beta}$ is the sum of the approximation bias $K^{-(s_\gamma+s_\alpha)/r}$ and $\Lambda(\hat{\Sigma},\Sigma)$. The rest of the HOIF estimator, i.e. $\hat{\beta}_H - \check{\beta}$, can be thought of as a bias correction for $\Lambda(\hat{\Sigma},\Sigma)$. Note that

$$E[\hat{\beta}_H - \check{\beta}|T] = \sum_{q=1}^{Q}\frac{(-1)^{q+1}(n-2-q)!}{n!}E[\alpha_0(x_i)p(x_i)]^T\left[\hat{\Sigma}^{-1}(\Sigma-\hat{\Sigma})\right]^q\hat{\Sigma}^{-1}E[p(x_i)\gamma_0(x_i)].$$

Here we see that $E[\hat{\beta}_H - \check{\beta}|T]$ is the negative of a Taylor expansion to order Q of $\Lambda(\hat{\Sigma},\Sigma)$ in $\hat{\Sigma}$ around Σ. Therefore, it will follow that

$$E[\hat{\beta}_H - \beta_0|T] = O(K^{-(s_\gamma+s_\alpha)/r}) + O(\left\|\hat{\Sigma}-\Sigma\right\|_{op}^{Q}) = O(K^{-(s_\gamma+s_\alpha)/r}) + O(\left(\frac{K\ln(K)}{n}\right)^{Q/2}),$$

where $\|\cdot\|_{op}$ is the operator norm for a matrix and the second equality follows by the Rudelson (1999) matrix law of large numbers. This equation is similar to the conclusion of Theorem 3 of Mukherjee, Newey, and Robins (2017).

In comparison with the HOIF estimator the CF plug-in series estimator has a remainder rate from estimating Σ that is $\ln(n)K^{-s_\gamma/r}$ for $s_\gamma, s_\alpha \le 1$ and Haar splines and $\sqrt{K\ln(K)/n}K^{1/2-s_\gamma/r}$ more generally, without any higher order U-statistic correction for the presence of $\hat{\Sigma}^{-1}$. The

DCDR estimator has $\Delta_n = \Delta_n^*$, also without the need to rely on any higher-order U-statistics. The key difference between the HOIF and these other estimators is that the plug-in and doubly robust estimators use spline regression in their construction and the HOIF estimator uses $\hat{\Sigma}^{-1}$ from a training subsample.

Previously the HOIF estimator was the only known method of obtaining an semiparametric efficient estimator of the expected conditional covariance under the minimal conditions of Robins et al.(2009). We find here that the CF plug-in estimator with a Haar basis can do this for $s_\gamma, s_\alpha \leq 1$ and for a general spline basis with $s_\gamma \geq r/2$. We also find that the DCDR estimator can do this for all s_γ and s_α. These estimators are simpler than the HOIF estimator in not requiring the higher order U-statistic terms. It would be interesting to compare the size of constants in respective remainder terms where HOIF could have an advantage by virtue of its higher order influence function interpretation. That comparison is beyond the scope of this paper.

The HOIF estimator remains the only known estimator that is semiparametric efficient under the Robins et al.(2009) minimal conditions for the mean with missing data over all s_γ and s_α. We expect that property of HOIF to extend to all the linear average functionals we are considering in this paper.

In summary, cross-fitting can be used to reduce bias of estimators and obtain faster remainder rates. If cross fitting is not used for either the plug-in or the doubly robust estimator there would be an additional $K/\sqrt{n}$ bias term in the remainder. This extra term can increase the bias of the estimator significantly for large K. It is well known to be very important in some settings, such as instrumental variables estimation as shown by Blomquist and Dahlberg (1999) and Imbens, Angrist, and Krueger (1999). Also, its presence prevents the plug-in estimator from attaining root-n consistency under minimal conditions. Cross-fitting eliminates this large remainder for the linear functionals we consider and results in plug-in and doubly robust estimators with remainders that converge to zero as fast as known possible for $s_\gamma, s_\alpha \leq 1$, for $s_\gamma > r/2$, and for any s_α and s_γ for a doubly robust estimator of the expected conditional covariance.

3 Estimators of Average Linear Functionals

We will analyze estimators of functionals of a conditional expectation

$$\gamma_0(x) = E[y_i | x_i = x],$$

where y_i is a scalar component and x_i a subvector of z_i. Let γ represent a possible conditional expectation function and $m(z, \gamma)$ denote a function of γ and a possible realization z of a data observation. We consider

$$\beta_0 = E\left[m(z_i, \gamma_0)\right],$$

where $m(z, \gamma)$ is an affine functional of γ for every z, meaning $m(z, \gamma) - m(z, 0)$ is linear in γ.

There are many important examples of such an object. One of these is the expected conditional covariance we consider in Section 2. There $m(z, \gamma) = a[y - \gamma(x)]$. This object shows up in different forms in the numerator and denominator of

$$\xi_0 = \frac{E[Cov(a_i, y_i|x_i)]}{E[Var(a_i|x_i)]}.$$

Here δ_0 is the coefficient of a_i in the population least squares projection of y_i on functions of the form $a_i\delta + g(x_i)$. Under an ignorability assumption this object δ_0 can be interpreted as a weighted average of conditional average treatment effects when a_i is a binary indicator for treatment and x_i are covariates.

Another important example is the mean when data are missing at random. The object of interest is $\beta_0 = E[Y_i]$ where Y_i is a latent variable that is not always observed. Let a_i be an observed binary indicator where $a_i = 1$ if Y_i is observed. Suppose that there are observed covariates w_i such that Y_i is mean independent of a_i conditional on w_i, i.e. $E[Y_i|a_i = 1, w_i] = E[Y_i|w_i]$. Then for the observed variable $y_i = a_iY_i$ we have

$$E[E[y_i|a_i = 1, w_i]] = E[E[Y_i|a_i = 1, w_i]] = E[E[Y_i|w_i]] = \beta_0.$$

Let $x = (a, w)$ and $\gamma_0(x_i) = E[y_i|x_i]$. Then for $m(z, \gamma) = \gamma(1, w)$ we have $\beta_0 = E[m(z_i, \gamma_0)]$.

A third example is a weighted average derivative, where the object of interest is

$$\beta_0 = \int v(x) \left[\partial\gamma_0(x)/\partial x_1\right] dx,$$

for some weight function $v(x)$, with x_1 continuously distributed and $\int v(x)dx = 1$. This object is proportional to β_{10} in a conditional mean index model where $E[y_i|x_i] = \tau(x_i^T\beta_0)$ for some unknown function $\tau(\cdot)$, as in Stoker (1986). This object is included in the framework of this paper for $m(z, \gamma) = \int v(x) \left[\partial\gamma(x)/\partial x_1\right] dx$. Assuming that $v(x)$ is zero at the boundary, integration by parts gives

$$m(z, \gamma) = m(\gamma) = \int \omega(x)\gamma(x)dx, \omega(x) = -\partial v(x)/\partial x_1.$$

Throughout we will focus on the case where estimators of β_0 have a finite semiparametric variance bound and so should be root-n consistently estimable under sufficient regularity conditions. As discussed in Newey (1994), this corresponds to $E[m(z_i, \gamma)]$ being mean square continuous as a function of γ, so that by the Riesz representation theorem the following condition is satisfied:

ASSUMPTION 1: *There is* $\alpha_0(x)$ *with* $E[\alpha_0(x_i)^2] < \infty$ *and for all* γ with $E[\gamma(x_i)^2] < \infty$,

$$E\left[m(z_i, \gamma) - m(z_i, 0)\right] = E\left[\alpha_0(x_i)\gamma(x_i)\right]. \tag{3.1}$$

The function $\alpha_0(x)$ has an important role in the asymptotic theory. The bias in a series estimator of β_0 will depend on the expected product of biases in approximating $\gamma_0(x)$ and $\alpha_0(x)$. Consequently there will be a trade-off in conditions that can be imposed on $\gamma_0(x)$ and $\alpha_0(x)$ so that the estimators of β_0 have good properties.

To help explain this condition we give the form of $\alpha_0(x)$ in each of the examples. In the expected conditional covariance example iterated expectations gives

$$\begin{aligned} E\left[m\left(z_i,\gamma\right)-m(z_i,0)\right] &= -E[a_i\gamma(x_i)] = -E[E[a_i|x_i]\gamma(x_i)] = E[\alpha_0(x_i)\gamma(x_i)], \\ \alpha_0(x_i) &= -E[a_i|x_i]. \end{aligned} \tag{3.2}$$

In the missing data example, for the propensity score $\Pr(a_i=1|w_i)=\pi_0(w_i)$, iterated expectations gives

$$\begin{aligned} E\left[m\left(z_i,\gamma\right)-m(z_i,0)\right] &= E[\gamma(1,w_i)] = E[\frac{\pi_0(w_i)}{\pi_0(w_i)}\gamma(1,w_i)] = E[\frac{a_i}{\pi_0(w_i)}\gamma(1,w_i)] \\ &= E[\frac{a_i}{\pi_0(w_i)}\gamma(x_i)] = E[\alpha_0(x_i)\gamma(x_i)], \alpha_0(x_i) = \frac{a_i}{\pi_0(w_i)}. \end{aligned} \tag{3.3}$$

In the average derivative example, multiplying and dividing by the pdf $f_0(x)$ of x_i gives

$$\begin{aligned} E\left[m\left(z_i,\gamma\right)-m(z_i,0)\right] &= \int \omega(x)\gamma(x)dx = \int \frac{\omega(x)}{f_0(x)}\gamma(x)f_0(x)dx = E[\frac{\omega(x_i)}{f_0(x_i)}\gamma(x_i)] \\ &= E[\alpha_0(x_i)\gamma(x_i)], \alpha_0(x_i) = \frac{\omega(x_i)}{f_0(x_i)}. \end{aligned} \tag{3.4}$$

Our estimators of β_0 will be based on a nonparametric estimator $\hat{\gamma}$ of γ_0 and possibly on a nonparametric estimator $\tilde{\alpha}$ of α_0. The CF plug-in estimator is given by

$$\hat{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} m(z_i,\hat{\gamma}_\ell),$$

where $I_\ell, (\ell=1,...,L)$ is a partition of the observation index set $\{1,...,n\}$ into L distinct subsets of about equal size and $\hat{\gamma}_\ell$ only uses observations *not* in I_ℓ. We will consider a fixed number of groups L in the asymptotics. It would be interesting to consider results where the number of groups grows with the sample size, even "leave one out" estimators where I_ℓ only includes one observation, but theory for those estimators is more challenging and we leave it to future work.

The DCDR estimator makes use of $\tilde{\alpha}_\ell$ that may be constructed from different observations than $\hat{\gamma}_\ell$. The doubly robust estimator is

$$\tilde{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left\{m(z_i,\hat{\gamma}_\ell)+\tilde{\alpha}_\ell(x_i)[y_i-\hat{\gamma}_\ell(x_i)]\right\}.$$

This estimator has the form of a plug-in estimator plus the sample average of $\tilde{\alpha}_\ell(x_i)[y_i-\hat{\gamma}_\ell(x_i)]$, which is an estimator of the influence function of $\int m(z,\hat{\gamma}_\ell)F_0(dz)$. The addition of $\tilde{\alpha}_\ell(x_i)[y_i -$

$\hat{\gamma}_\ell(x_i)]$ will mean that the nonparametric estimators $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$ do not affect the asymptotic distribution of $\tilde{\beta}$, i.e. the limit distribution would be the same if $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$ were replaced by their true values and $\Delta_n \longrightarrow 0$. This estimator allows for full cross-fitting where $\tilde{\alpha}$ and $\hat{\gamma}$ may be based on distinct subsamples.

The cross-fit estimator $\tilde{\beta}$ is doubly robust in the sense that $\tilde{\beta}$ will be consistent as long as either $\hat{\gamma}_\ell$ or $\tilde{\alpha}_\ell$ is consistent, as shown by Chernozhukov et al.(2016) for this general class of functionals. When $\hat{\gamma}(x)$ is a series estimator like that described above the CF plug-in estimator $\hat{\beta}$ is also doubly robust in a more limited sense. It will be consistent with fixed $p(x)$ if either $\gamma_0(x)$ or $\alpha_0(x)$ is a linear combination of $p(x)$, as shown for the mean with missing data in Robins et al.(2007) and in Chernozhukov et al.(2016) for the general linear function case we are considering.

Throughout the paper we assume that each data point z_i is used for estimation for some group ℓ and that the number of observations in group ℓ, the number used to form $\hat{\gamma}_\ell$, and the number used to form $\tilde{\alpha}_\ell$ grow at the same rate as the sample size. To make this condition precise let $\bar{n}_\ell$ be the number of elements in I_ℓ, $\hat{n}_\ell$ be the number used to form $\hat{\gamma}_\ell$, and $\tilde{n}_\ell$ be the number of observations used to form $\tilde{\alpha}_\ell$. We will assume throughout that all the observations are used for each ℓ, i.e. that either $\bar{n}_\ell + \hat{n}_\ell = n$ or $\bar{n}_\ell + \hat{n}_\ell + \tilde{n}_\ell = n$ if different observations are used for $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$.

Assumption 2: *There is a constant $C > 0$ such that either $\bar{n}_\ell + \hat{n}_\ell = n$ and $\min_\ell\{\bar{n}_\ell, \hat{n}_\ell\} \geq Cn$ or $\bar{n}_\ell + \hat{n}_\ell + \tilde{n}_\ell = n$ and $\min_\ell\{\bar{n}_\ell, \hat{n}_\ell, \tilde{n}_\ell\} \geq Cn$. For the plug-in estimator groups are as close as possible to being of equal size.*

The assumption that the group sizes are as close to equal as possible for the plug-in estimator is made for simplicity but could be relaxed.

We turn now to conditions for the regression spline estimators of $\gamma_0(x)$ and $\alpha_0(x)$. We continue to consider regression spline first steps where $p(x)$ is a $K \times 1$ vector of regression splines. The nonparametric estimator of $\gamma_0(x)$ will be a series regression estimator where

$$\hat{\gamma}_\ell(x) = p(x)^T \hat{\delta}_\ell,\ \hat{\delta}_\ell = \hat{\Sigma}_\ell^- \hat{h}_\ell,\ \hat{\Sigma}_\ell = \frac{1}{\hat{n}_\ell} \sum_{i \in \hat{I}_\ell} p(x_i) p(x_i)^T,\ \hat{h}_\ell = \frac{1}{\hat{n}_\ell} \sum_{i \in \hat{I}_\ell} p(x_i) y_i,$$

where a T superscript denotes the transpose, $\hat{I}_\ell$ is the index set for observations used to construct $\hat{\gamma}_\ell(x)$, and A^- denotes any generalized inverse of a positive semi-definite matrix A. Under conditions given below $\hat{\Sigma}_\ell$ will be nonsingular with probability approaching one so that $\hat{\Sigma}_\ell^- = \hat{\Sigma}_\ell^{-1}$ for each ℓ.

The DCDR estimator $\tilde{\beta}$ uses an estimator of $\alpha_0(x)$. The function $\alpha_0(x)$ cannot generally be interpreted as a conditional expectation and so cannot generally be estimated by a linear regression. Instead we use Assumption 1 and equation (3.1) to construct an estimator. Let

$v(z) = (m(z,p_1) - m(z,0), ..., m(z,p_K) - m(z,0))^T$. Then by Assumption 1,

$$E[v(z_i)] = E[p(x_i)\alpha_0(x_i)],$$

so that $\tilde{h}_{\ell\alpha} = \sum_{i\in\tilde{I}_\ell} v(z_i)/\tilde{n}_\ell$ is an unbiased estimator of $E[p(x_i)\alpha_0(x_i)]$. A series estimator of $\alpha_0(x)$ is then

$$\tilde{\alpha}_\ell(x) = p(x)^T\tilde{\delta}_{\ell\alpha}, \tilde{\delta}_{\ell\alpha} = \tilde{\Sigma}_\ell^- \tilde{h}_{\ell\alpha}, \tilde{\Sigma}_\ell = \frac{1}{\tilde{n}_\ell}\sum_{i\in\tilde{I}_\ell} p(x_i)p(x_i)^T.$$

Here $\tilde{\delta}_{\ell\alpha}$ is an estimator of the coefficients of the population regression of $\alpha_0(x)$ on $p(x)$, but $\tilde{\delta}_{\ell\alpha}$ is not obtained from a linear regression. This type of estimator of $\alpha_0(x)$ was used to construct standard errors for functionals of series estimators in Newey (1994).

Now that we have specified the form of the estimators $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$ we can give a complete description of the estimators in each of the examples. For the expected conditional covariance recall that $m(z,\gamma) = a[y - \gamma(x)]$. Therefore the CF plug-in estimator will be

$$\hat{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} a_i[y_i - \hat{\gamma}_\ell(x_i)]. \tag{3.5}$$

Also, as discussed above, for the expected conditional covariance $\alpha_0(x) = -E[a_i|x_i = x]$ and $v(z_i) = -a_i p(x_i)$, so that $\tilde{\alpha}_\ell(x) = -\tilde{\gamma}_{a\ell}(x)$ where $\tilde{\gamma}_{a\ell}(x) = p(x)^T\tilde{\Sigma}_\ell^- \sum_{i\in\tilde{I}_\ell} p(x_i)a_i/\tilde{n}_\ell$ is the regression of a_i on $p(x_i)$ for the observations indexed by $\tilde{I}_\ell$. Then the DCDR estimator is

$$\tilde{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i + \tilde{\alpha}_\ell(x_i)][y_i - \hat{\gamma}_\ell(x_i)] = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{a_i - \tilde{E}[a_i|x_i]\}[y_i - \hat{\gamma}_\ell(x_i)], \tag{3.6}$$

where $\tilde{E}[a_i|x_i] = -\tilde{\alpha}_\ell(x_i)$ is the predicted value from the regression of a_i on $p(x_i)$. This estimator is the average of the product of two nonparametric regression residuals, where the average and each of the nonparametric estimators can be constructed from different samples.

For the missing data example the estimators are based on series estimation of $E[y_i|a_i = 1, w_i]$. Let $q(w)$ denote a $K\times 1$ vector of splines, $x = (a, w^T)^T$, and $p(x) = (aq(w)^T, (1-a)q(w)^T)^T$. The predicted value $\hat{\gamma}(1,w)$ will be the same as from a linear regression of y_i on $q(w_i)$ for observations with $a_i = 1$. That is, $\hat{\gamma}(1,w) = q(w)^T\hat{\delta}_\ell$ where

$$\hat{\delta}_\ell = \hat{\Sigma}_\ell^- \hat{h}_\ell,\ \hat{\Sigma}_\ell = \frac{1}{\hat{n}_\ell}\sum_{i\in\hat{I}_\ell} a_i q(w_i)q(w_i)^T,\ \hat{h}_\ell = \frac{1}{\hat{n}_\ell}\sum_{i\in\hat{I}_\ell} a_i q(w_i)y_i.$$

The CF plug-in estimator is

$$\hat{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} q(w_i)^T\hat{\delta}_\ell.$$

The DCDR estimator is based on an estimator of the inverse propensity score $\pi_0(w_i)^{-1} = 1/\pi_0(w_i)$ given by

$$\widetilde{\pi(w_i)}_\ell^{-1} = q(w_i)^T\tilde{\delta}_\ell^\alpha, \tilde{\delta}_\ell^\alpha = \tilde{\Sigma}_\ell^- \tilde{h}_\ell^\alpha, \tilde{\Sigma}_\ell = \frac{1}{\tilde{n}_\ell}\sum_{i\in\tilde{I}_\ell} a_i q(w_i)q(w_i)^T,\ \tilde{h}_\ell^\alpha = \frac{1}{\tilde{n}_\ell}\sum_{i\in\tilde{I}_\ell} q(w_i),$$

where $\tilde{n}_\ell$ is the number of observation indices in $\tilde{I}_\ell$. This estimator of the inverse propensity score is a version of one discussed in Robins et al.(2007). The DCDR estimator is

$$\tilde{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left\{q(w_i)^T\hat{\delta}_\ell + a_i\widetilde{\pi(w_i)}_\ell^{-1}[y_i - q(w_i)^T\hat{\delta}_\ell]\right\}.$$

This has the usual form for a doubly robust estimator of the mean with data missing at random. It differs from previous estimators in having the full CF form where the nonparametric estimators are based on distinct subsamples of the data.

For the average derivative example $m(z,\gamma) = \int\omega(x)\gamma(x)dx$ does not depend on z so we can use all the data in the construction of the plug-in estimator. That estimator is given by

$$\hat{\beta} = \int\omega(x)\hat{\gamma}(x)dx = v^T\hat{\delta},\ v = \int\omega(x)p(x)dx, \hat{\delta} = [\sum_{i=1}^{n}p(x_i)p(x_i)^T]^-\sum_{i=1}^{n}p(x_i)y_i. \tag{3.7}$$

As shown in equation (3.4), $\alpha_0(x) = f_0(x)^{-1}\omega(x)$, where $f_0(x)$ is the pdf of x. Also here $v(z) = v$ so the estimator of $\alpha_0(x)$ is $p(x)^T\tilde{\Sigma}_\ell^- v$. The DCDR estimator is then

$$\tilde{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{\int\omega(x)\hat{\gamma}_\ell(x)dx + \left[p(x_i)^T\tilde{\Sigma}_\ell^- v\right][y_i - \hat{\gamma}_\ell(x_i)]\}. \tag{3.8}$$

Both the plug-in and the DCDR estimators depend on the integral $v = \int\omega(x)p(x)dx$. Generally this vector of integrals will not exist in closed form so that construction of these estimators will require numerical computation or estimation of v, such as by simulation.

We now impose some specific conditions on $p(x)$.

Assumption 3: *$p(x) = aq(w)$ where i) the support of w_i is $[0,1]^r$, w_i is continuously distributed with bounded pdf that is bounded away from zero; ii) $q(w)$ are tensor product b-splines of order κ with knot spacing approximately proportional to the number of knots; iii) $q(w)$ is normalized so that $\lambda_{\min}(E[q(w_i)q(w_i)^T]) \geq C > 0$ and $\sup_{w\in[0,1]^r}\|q(w)\| \leq C\sqrt{K}$; iv) a_i is bounded and $E[a_i^2|w_i]$ is bounded away from zero.*

Under condition i) it is known that there is a normalization such that condition iii) is satisfied, e.g. as in Newey (1997). To control the bias of the estimator we require that the true regression function $\gamma_0(x)$ and the auxiliary function $\alpha_0(x)$ each be in a Holder class of functions. We define a function $g(x)$ to be Holder of order s if there is a constant C such that $g(x)$ is continuously differentiable of order $\bar{s} = int[s]$ and each of its $\bar{s}$ partial derivatives $\nabla^{\bar{s}} g(x)$ satisfies $|\nabla^{\bar{s}} g(\tilde{x}) - \nabla^{\bar{s}} g(x)| \leq C \left\| \tilde{x} - x \right\|^{s-\bar{s}}$.

Assumption 4: *$\gamma_0(x)$ and $\alpha_0(x)$ are Holder of order s_γ and s_α respectively.*

This condition implies that the population least squares approximations to $\gamma_0(x)$ and $\alpha_0(x)$ converge at certain rates. Let $\zeta_\gamma = \min\{1+\kappa, s_\gamma\}/r$, $\zeta_\alpha = \min\{1+\kappa, s_a\}/r$, $\Sigma = E[p(x_i)p(x_i)^T]$, $\delta = \Sigma^{-1}E[p(x_i)\gamma_0(x_i)]$, $\gamma_K(x) = p(x)^T\delta$, $\delta_\alpha = \Sigma^{-1}E[p(x_i)\alpha_0(x_i)]$, $\alpha_K(x) = p(x)^T\delta_a$. Then standard approximation theory for splines gives

$$\begin{aligned} &E[\{\gamma_0(x_i) - \gamma_K(x_i)\}^2] = O(K^{-2\zeta_\gamma}), \sup_{x\in[0,1]^r} |\gamma_0(x) - \gamma_K(x)| = O(K^{-\zeta_\gamma}), \\ &E[\{\alpha_0(x_i) - \alpha_K(x_i)\}^2] = O(K^{-2\zeta_\alpha}). \end{aligned}$$

We will use these results to derive the rates at which certain remainders converge to zero.

We also impose the following condition:

Assumption 5: $Var(y_i|x_i) \leq C$, $K \longrightarrow \infty$, *and* $K\ln(K)/n \longrightarrow 0$.

These are standard conditions for series estimators of conditional expectations. A bounded conditional variance for y_i helps bound the variance of series estimators. The upper bound on the rate at which K grows is slightly stronger than $K/n \longrightarrow 0$. This upper bound on K allows us to apply the Rudelson (1999) law of large numbers for symmetric matrices to show that the various second moment matrices of $p(x)$ converge in probability. Another condition we impose is:

Assumption 6: $\lambda_{\max}(E[v(z_i)v(z_i)^T]) \leq Cd_K$ *and* $\{E[\{m(z_i,\gamma_K) - m(z_i,\gamma_0)\}^2]\}^{1/2} = O(K^{-\zeta_m})$.

The first condition will be satisfied with $d_K = 1$ in the examples under specific regularity conditions detailed below. The second condition gives a rate for the mean square error convergence of $m(z,\gamma_K) - m(z,\gamma_0)$ as K grows. In all of the examples this rate will be $\zeta_m = \zeta_\gamma$. In other examples, including those where $m(z,\gamma)$ and $v(z)$ depend on derivatives with respect to x, we will have d_K growing with K and $\zeta_m < \zeta_\gamma$.

For the statement of the results to follow it is convenient to work with the remainder term

$$\bar{\Delta}_n^* = \sqrt{n}K^{-\zeta_\gamma - \zeta_\alpha} + K^{-\zeta_\gamma} + K^{-\zeta_\alpha} + \sqrt{\frac{K}{n}}.$$

This remainder coincides with the fast remainder Δ_n^* when the spline order is high enough with $\kappa \geq \max\{s_\gamma, s_\alpha\} - 1$. The only cases where it would not be possible to choose such a κ are for the Haar basis where $\kappa = 0$.

4 The Plug-in Estimator

In this Section we derive bounds on the size of remainders for the plug-in estimator. Some bounds are given for general plug-in estimators, some for plug-ins that are series regression with Haar splines, and some for other splines. We begin with a result that applies to all plug-ins. We drop the CF designation because all the estimators from this point on will use cross-fitting.

The cross-fit form of the plug-in estimator allows us to partly characterize its properties under weak conditions on a general plug-in estimator that need not be a series regression. This characterization relies on independence of $\hat{\gamma}_\ell$ from the observations in I_ℓ to obtain relatively simple stochastic equicontinuity remainders. Also, this result accounts for the overlap across groups in observations used to form $\hat{\gamma}_\ell$. Let $\mathcal{A}_n$ denote an event that occurs with probability approaching one. For example, $\mathcal{A}_n$ could include the set of data points where $\hat{\Sigma}_\ell$ is nonsingular for each ℓ.

Lemma 1: *If Assumptions 1 and 2 are satisfied and there is Δ_n^m such that*

$$1(\mathcal{A}_n)\left\{\int [m(z,\hat{\gamma}_\ell) - m(z,\gamma_0)]^2 F_0(dz)\right\}^{1/2} = O_p(\Delta_n^m), (\ell = 1, ..., L),$$

then for $\bar{m}(\gamma) = \int m(z,\gamma)F_0(dz)$,

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n [m(z_i,\gamma_0) - \beta_0] + \sqrt{n}\sum_{\ell=1}^L \frac{\bar{n}_\ell}{n}[\bar{m}(\hat{\gamma}_\ell) - \beta_0] + O_p(\Delta_n^m).$$

If in addition there is Δ_n^ϕ such that for each $(\ell = 1, ..., L)$,

$$\sqrt{\hat{n}_\ell}[\bar{m}(\hat{\gamma}_\ell) - \beta_0] = \frac{1}{\sqrt{\hat{n}_\ell}}\sum_{i \notin I_\ell} \alpha_0(x_i)[y_i - \gamma_0(x_i)] + O_p(\Delta_n^\phi),$$

then for $\delta(z) = m(z,\beta_0) - \beta_0 + \alpha_0(x)[y - \gamma_0(x)]$

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(z_i) + O_p(\Delta_n^m + \Delta_n^\phi + n^{-1}).$$

This result gives a decomposition of remainder bounds into two kinds. The first Δ_n^m is a stochastic equicontinuity bound that has the simple mean-square form given here because of the sample splitting. The second Δ_n^ϕ is a bound that comes from the asymptotically linear expansion of the linear functional estimator $\bar{m}(\hat{\gamma}_\ell)$. For general b-splines we can apply Ichimura and Newey (2017) to obtain Δ_n^ϕ. For zero order splines we give here sharper remainder bounds.

For series estimators the stochastic equicontinuity remainder bound Δ_n^m will be

$$\Delta_n^m = \sqrt{(d_K + 1)\frac{K}{n}} + K^{-\zeta_m},$$

where d_K and ζ_m are as given in Assumption 6. As mentioned above, in the examples in this paper $d_K \leq C$ and $\zeta_m = \zeta_\gamma$. Here we can take $\Delta_n^m \leq C\bar{\Delta}_n^*$, so the stochastic equicontinuity remainder bound is the same size as $\bar{\Delta}_n^*$.

Our next result gives remainder bounds for the Haar basis.

THEOREM 2: *If Assumptions 1-6 are satisfied,* $\kappa = 0$, *and* $K[\ln(n)]^2/n \longrightarrow 0$ *then*

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + O_p(\bar{\Delta}_n^* + \Delta_n^m + K^{-\zeta_\gamma} \ln(n)).$$

If in addition d_K *is bounded as a function of* K *and* $\zeta_m = \zeta_\gamma$ *then* $\Delta_n^m \leq C\bar{\Delta}_n^*$.

Here we see that for a Haar basis the order of the remainder term for the plug-in estimator is a sum of the stochastic equicontinuity term Δ_n^m and $\bar{\Delta}_n^*$, with $K^{-\zeta_\gamma} \ln(n)$ being the size of the fast remainder up to $\ln(n)$. In the examples and other settings where d_K is bounded and $\zeta_m = \zeta_\gamma$ the Δ_n^m remainder will just be of order $\bar{\Delta}_n^*$. The following result states conditions for the examples.

COROLLARY 3: *Suppose that Assumptions 1-3 and 5 are satisfied,* $\kappa = 0$, $K[\ln(n)]^2/n \longrightarrow 0$, *and* $\gamma_0(x)$ *is Holder of order* s_γ. *If either i)* $\hat{\beta}$ *is the expected conditional covariance estimator,* $E[a_i|x_i = x]$ *is Holder of order* s_α, $E[a_i^2|x_i]$ *is bounded, or ii)* $\hat{\beta}$ *is the missing data mean estimator,* $\Pr(a_i = 1|x_i)$ *is bounded away from zero and is Holder of order* s_α, *or iii)* $\hat{\beta}$ *is the average derivative estimator,* $\omega(x)$ *and* $f_0(x)$ *are Holder of order* s_α, *and* $f_0(x)$ *is bounded away from zero on the set where* $\omega(x) > 0$, *then*

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + O_p(\bar{\Delta}_n^* + K^{-\zeta_\gamma} \ln(n)).$$

The remainder bound means that the plug-in estimator can attain root-n consistency under minimal conditions, when the dimension r is small enough. There will exist K such that $\bar{\Delta}_n^*$ goes to zero if an only if

$$1/2 < \zeta_\gamma + \zeta_\alpha = \frac{\min\{1, s_\gamma\} + \min\{1, s_\alpha\}}{r}. \tag{4.1}$$

This condition can be satisfied for $r < 4$ but not for $r \geq 4$. For $r = 1$ this condition will be satisfied if and only if

$$s_\gamma + s_\alpha > \frac{1}{2},$$

which is the minimal condition of Robins et al.(2009) for existence of a semiparametric efficient estimator for the expected conditional covariance and missing data parameters when $r = 1$. For $r = 2$ we note that

$$\min\{1, s_\gamma\} + \min\{1, s_\alpha\} \geq 1 \text{ if and only if } s_\gamma + s_\alpha \geq 1.$$

For $r=2$ equation (4.1) is $\min\{1,s_\gamma\}+\min\{1,s_\alpha\}>1$, which requires both $s_\alpha>0$ and $s_\gamma>0$ and so is slightly stronger than the Robins et al.(2009) condition $s_\gamma+s_\alpha>1$. For $r=3$ the situation is more complicated. Equation (4.1) is stronger than the corresponding condition $s_\gamma+s_\alpha>3/2$ of Robins et al.(2009), although it is the same for the set of (s_γ,s_α) where $s_\gamma\le 1$ and $s_\alpha\le 1$. Along the diagonal where $s_\alpha=s_\gamma$ the two conditions coincide as $s_\gamma>3/4$.

The limited nature of these results is associated with the Haar basis, which limits the degree to which smoothness of the underlying function results in a faster approximation rate. If Theorem 2 and Corollary 3 could be extended to other, higher order b-splines, this limitation could be avoided. For the present we are only able to do this for the doubly robust estimator of a partially linear projection, as discussed in the next Section.

There is a key result that allows us to obtain the remainder bound $\bar{\Delta}_n^*$ in Theorem 2. Let $\hat{h}_2=\sum_{i=1}^n p(x_i)[\gamma_0(x_i)-\gamma_K(x_i)]/n$, $\hat{\Sigma}=\sum_{i=1}^n p(x_i)p(x_i)^T/n$, and $\Sigma=E[p(x_i)p(x_i)^T$. We show in the Appendix that for the Haar basis

$$\lambda_{\max}(E[(\Sigma-\hat{\Sigma})^j\hat{h}_2\hat{h}_2^T(\Sigma-\hat{\Sigma})^j])\le\frac{K^{-2\zeta_\gamma}}{n}\left(\frac{CK}{n}\right)^j. \tag{4.2}$$

If b-spline bases other than Haar also satisfied this condition then we could obtain results analogous to Theorem 2 and Corollary 3 for these bases. We do not yet know if other bases satisfy this condition. The Haar basis is convenient in $p(x)^Tp(x)$ being piecewise constant. Cattaneo and Farrell (2013) exploited other special properties of the Haar basis to obtain sharp uniform nonparametric rates.

For b-splines of any order we can obtain remainder rates by combining Lemma 1 with Theorem 8 of Ichimura and Newey (2017).

Theorem 4: *If Assumptions 1-6 are satisfied then*

$$\sqrt{n}(\hat{\beta}-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(z_i)+O_p(\bar{\Delta}_n^*+\Delta_n^m+\bar{\Delta}_n),\bar{\Delta}_n=\left(\frac{K\ln K}{n}\right)^{1/2}K^{(1/2)-\zeta_\gamma}.$$

If in addition d_K is bounded as a function of K and $\zeta_m=\zeta_\gamma$ then $\Delta_n^m\le C\bar{\Delta}_n^*$.

Here we see that the remainder bound for splines with $\kappa>0$ has an additional term $\bar{\Delta}_n$. When ζ_γ is large enough, i.e. $\gamma_0(x)$ is smooth enough and the order of the spline is big enough, so that $\zeta_\gamma>1/2$, the additional $\bar{\Delta}_n$ will be no larger than $\bar{\Delta}_n^*$. Also, when $\zeta_\gamma>1/2$ the condition of Robins et al.(2009) for semiparametric efficient estimation is met for the expected conditional covariance and missing data examples for any ζ_α. Thus, when $\gamma_0(x)$ is smooth enough to meet the Robins et al.(2009) condition without imposing any smoothness on $\alpha_0(x)$ the plug-in estimator will have the remainder bound $\bar{\Delta}_n^*$.

More generally there will exist a K such that $\bar{\Delta}_n+\bar{\Delta}_n^*$ goes to zero if and only if

$$2\min\{\kappa+1,s_\gamma\}+\min\{\kappa+1,s_\alpha\}>r. \tag{4.3}$$

This condition is slightly stronger than that of Robins et al.(2009) which is $2s_\gamma + 2s_\alpha > r$. Also, the remainder may go to zero when when K is chosen to maximize the rate at which the mean square error of $\hat{\gamma}_0(x)$ goes to zero. Setting $K^{-2\zeta_r}$ proportional to K/n is such a choice of K. Here the remainder term goes to zero for $\min\{\kappa+1, s_\gamma\} > r/[2(1+r)]$ and $\min\{\kappa+1, s_\alpha\} > r/2$, a stronger condition for s_γ and the same condition for s_α as would hold if the remainder were $\bar{\Delta}_n^*$.

5 Partially Linear Projection

In this Section we consider a series estimator of partially linear projection coefficients. We give this example special attention because the DCDR estimator will have a remainder bound that is only $\bar{\Delta}_n^*$. The remainder bounds we find for other doubly robust estimators may be larger. What appears to make the partially linear projection special in this respect is that $\alpha_0(x)$ is a conditional expectation of an observed variable. In other cases where $\alpha_0(x)$ is not a conditional expectation we do not know if the remainder bound will be $\bar{\Delta}_n^*$ for bases other than Haar.

The parameter vector of interest in this Section is

$$\beta_0 = \left(E[\{a_i - E[a_i|w_i]\}a_i^T]\right)^{-1} E[\{a_i - E[a_i|w_i]\}y_i].$$

This vector β_0 can be thought of as the coefficients of a_i in a projection of y_i on the set of functions of the form $a_i^T\beta + \lambda(x_i)$ that have finite mean square. Note that this definition of β_0 places no substantive restrictions on the distribution of data, unlike the conditional expectation partially linear model where $E[y_i|a_i, w_i] = a_i^T\beta_0 + \xi_0(x_i)$.

The object β_0 is of interest in a treatment effects model where a_i is a binary treatment, y_i is the observed response, x_i are covariates, and outcomes with and without treatment are assumed to be mean independent of a_i conditional on w_i. Under an ignorability condition that the outcome is mean independent of treatment conditional on covariates, $E[y_i|a_i = 1, x_i] - E[y_i|a_i = 0, x_i]$ is the average treatment effect conditional on x_i. Also for $\pi_i = \Pr(a_i = 1|x_i)$,

$$\beta_0 = \frac{E[\pi_i(1-\pi_i)\{E[y_i|a_i = 1, x_i] - E[y_i|a_i = 0, x_i]\}]}{E[\pi_i(1-\pi_i)]}.$$

Here we have the known interpretation of β_0 as a weighted average of conditional average treatment effects, with weights $\pi_i(1-\pi_i)/E[\pi_i(1-\pi_i)]$.

It is straightforward to construct a DCDR estimator of β_0. Let $\gamma_0(x_i) = E[y_i|x_i]$ and $\alpha_0(x_i) = -E[a_i|x_i]$ as before, except that a_i may now be a vector. Also let I_ℓ denote the index set for the ℓ^{th} group, and $\hat{I}_\ell$ and $\tilde{I}_\ell$ the index sets for the observations used to obtain $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$ respectively. For any function $g(z)$ let

$$\bar{F}\{g(z)\} = \frac{1}{\bar{n}_\ell}\sum_{i\in I_\ell} g(z_i), \hat{F}\{g(z)\} = \frac{1}{\hat{n}_\ell}\sum_{i\in \hat{I}_\ell} g(z_i), \tilde{F}\{g(z)\} = \frac{1}{\tilde{n}_\ell}\sum_{i\in \tilde{I}_\ell} g(z_i).$$

These represent sample averages over each of the groups of observations. Let $\hat{\gamma}_\ell(x)$, $\hat{\alpha}_\ell(x)$, and $\tilde{\alpha}_\ell(x)$ be series estimators of $\gamma_0(x)$ and $\alpha_0(x)$ given by

$$\hat{\gamma}_\ell(x) = p(x)^T\hat{\delta}_\ell, \hat{\alpha}_\ell(x) = p(x)^T\hat{\delta}_{\ell\alpha}, \tilde{\alpha}_i(x) = p(x)^T\tilde{\delta}_{\ell\alpha},$$
$$\hat{\delta} = \hat{\Sigma}^{-1}\hat{h},\ \hat{\delta}_\alpha = \hat{\Sigma}^{-}\hat{h}_\alpha,\ \tilde{\delta}_\alpha = \tilde{\Sigma}^{-}\tilde{h}_\alpha,\ \hat{\Sigma} = \hat{F}\{p(x)p(x)^T\},\ \tilde{\Sigma} = \tilde{F}\{p(x)p(x)^T\},$$
$$\hat{h} = \hat{F}\{p(x)y\},\ \hat{h}_\alpha = \hat{F}\{p(x)a\},\ \tilde{h}_\alpha = \tilde{F}\{p(x)a\}.$$

The estimator we consider is

$$\tilde{\beta} = \left(\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i - \tilde{\alpha}_\ell(x_i)][a_i - \hat{\alpha}_\ell(x_i)]^T\right)^{-1}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i - \tilde{\alpha}_\ell(x_i)][y_i - \hat{\gamma}_\ell(x_i)]. \tag{5.1}$$

This estimator can be thought of as an instrumental variables estimator with left hand sides variable $y_i - \hat{g}_i(x_i)$, right hand side variables $a_i - \hat{\alpha}_i(x_i)$, and instruments $a_i - \tilde{\alpha}_i(x_i)$. Here the instrumental variables form is used to implement the cross-fitting and not to correct for endogeneity. This form means that every element of the matrix that is inverted and of the vector it is multiplying is a DCDR estimator of an expected conditional covariance like that described earlier.

THEOREM 5: *If Assumptions 1 - 3 and 5 are satisfied, $\lambda_0(x) = E[y_i - a_i^T\beta_0|x_i = x]$ is Holder of order s_γ and each component of $E[a_i|x_i = x]$ is Holder of order s_α, $H = E[Var(a_i|x_i)]$ exists and is nonsingular, and $\Omega = E[\{a_i - \alpha_0(x_i)\}\{a_i - \alpha_0(x_i)\}^T\varepsilon_i^2]$ exists then for $\varepsilon_i = y_i - a_i^T\beta_0 - \lambda_0(x_i)$ and $\psi(z_i) = H^{-1}(a_i - E[a_i|x_i])\varepsilon_i$,*

$$\sqrt{n}(\tilde{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i) + O_p(\bar{\Delta}_n^*).$$

The regularity conditions here are somewhat stronger than those of Donald and Newey (1994), who do not require any restrictions on the marginal distribution of x_i nor use any sample splitting. This strengthening is useful to achieve the fast remainder for partially linear projections rather than for the coefficients β_0 in the conditional mean model $E[y_i|a_i, x_i] = a_i^T\beta_0 + \lambda_0(x_i)$ of Donald in Newey (1994). The upper bound on the rate at which K can grow is slightly stricter than in Donald and Newey (1994) due to the presence of the $\ln(K)$ term in Assumption 5. Thus, under somewhat stronger conditions than those of Donald and Newey (1994) the DCDR estimator of a partially linear projection has a fast remainder just as in Donald and Newey (1994). Consequently, the estimator will be root-n consistent under minimal conditions.

When the Robins et al. (2009) minimal condition $(s_\gamma + s_\alpha)/r > 1/2$ holds, consider a spline with $\kappa > \max\{s_\gamma, s_\alpha\} - 1$, so that $\zeta_\gamma + \zeta_\alpha = (s_\gamma + s_\alpha)/r > 1/2$. Then there will exist a K

such that $\bar{\Delta}_n^* \longrightarrow 0$ and hence $\tilde{\beta}$ will be semiparametric efficient. Thus we see that the DCDR estimator $\tilde{\beta}$ of equation (5.1) will be semiparametric efficient under nearly minimal conditions and has a fast remainder term.

6 The Doubly Robust Estimator

In this Section we show that the DCDR estimator has improved properties relative to the plug-in estimator, in the sense that the remainder bounds are smaller for the DCDR robust estimator. We have not yet been able to obtain the fast remainder for the doubly robust estimator for general splines, for the same reasons as for plug-in estimators.

Before giving results for series estimators we give a result that applies to any doubly robust estimator of a linear functional. Let $\mathcal{A}_n$ denote an event that occurs with probability approaching one. For example, $\mathcal{A}_n$ could include the set of data points where $\hat{\Sigma}_\ell$ is nonsingular.

LEMMA 6: *If Assumptions 1 and 2 are satisfied, $\hat{\gamma}_\ell(x)$ and $\hat{\alpha}_\ell(x)$ do not use observations in I_ℓ, $Var(y_i|x_i)$ is bounded, and there are Δ_n^m, Δ_m^γ,* and Δ_m^α, *such that for each* $(\ell = 1, ..., L)$,

$$1(\mathcal{A}_n)\left\{\int [m(z,\hat{\gamma}_\ell) - m(z,\gamma_0)]^2 F_0(dz)\right\}^{1/2} = O_p(\Delta_n^m),$$

$$1(\mathcal{A}_n)\left\{\int \alpha_0(x)^2[\hat{\gamma}_\ell(x) - \gamma_0(x)]^2 F_0(dz)\right\}^{1/2} = O_p(\Delta_n^\gamma),$$

$$1(\mathcal{A}_n)\left\{\int [\tilde{\alpha}_\ell(x) - \alpha_0(x)]^2 F_0(dz)\right\}^{1/2} = O_p(\Delta_n^\alpha),$$

then

$$\sqrt{n}(\tilde{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(z_i) - \frac{1}{\sqrt{n}}\sum_{\ell=1}^L \sum_{i\in I_\ell}[\tilde{\alpha}_\ell(x_i) - \alpha_0(x_i)][\hat{\gamma}_\ell(x_i) - \gamma_0(x_i)] + O_p(\Delta_n^m + \Delta_n^\gamma + \Delta_n^\alpha).$$

This result does not require that $\hat{\gamma}_\ell(x)$ and $\hat{\alpha}_\ell(x)$ be computed from different samples. It only uses the sample splitting in averaging over different observations that are used to construct $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$. Also, it is known from Newey, Hsieh, and Robins (1998, 2004) and Chernozhukov et. al. (2016) that adding the adjustment term to the plug-in estimator makes the remainder second order. The conclusion of Lemma 6 gives an explicit form of that result. Under weak conditions that only involve mean-square convergence the doubly robust estimator has a remainder that is the sum of three stochastic equicontinuity remainders and the quadratic, split sample remainder involving the product of the estimation remainders for the two nonparametric estimators $\hat{\gamma}$ and $\tilde{\alpha}$.

For series estimators the DCDR estimator will have $\bar{\Delta}_n^*$ as its primary remainder for the Haar basis

THEOREM 7: *If Assumptions 1-6 are satisfied, $\kappa = 0$, and $K[\ln(n)]^2/n \longrightarrow 0$ then*

$$\sqrt{n}(\tilde{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + O_p(\bar{\Delta}_n^* + \Delta_n^m).$$

If in addition d_K is bounded as a function of K and $\zeta_m = \zeta_\gamma$ then $\Delta_n^m \leq C\bar{\Delta}_n^*$.

One improvement of the DCDR estimator over the plug-in estimator is that the remainder no longer contains the $K^{-\zeta_\gamma}\ln(n)$ term. The elimination of this term is the direct result of the DCDR estimator having a smaller remainder than the plug-in estimator.

For splines of order $\kappa > 0$ we can obtain a result for the DCDR estimator that improves on the plug-in remainder bound.

THEOREM 8: *If Assumptions 1-6 are satisfied then*

$$\sqrt{n}(\tilde{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + O_p(\bar{\Delta}_n^* + \Delta_n^m + \tilde{\Delta}_n), \tilde{\Delta}_n = \sqrt{\frac{K^3 \left[\ln(K)\right]^2 (1 + d_K)}{n^3}} K^{(1/2)-\zeta_\gamma}.$$

If in addition d_K is bounded as a function of K and $\zeta_m = \zeta_\gamma$ then $\Delta_n^m \leq C\bar{\Delta}_n^*$.

Here we see that the remainder bound for the DCDR estimator will generally be smaller than the remainder bound for the plug-in estimator because the term $K\ln(K)/n$ is raised to the $3/2$ power rather than the $1/2$ power. Here it turns out that there will exist a K such that all of the remainder terms go to zero if

$$4\zeta_\gamma + 3\zeta_\alpha \geq 2.$$

For example, if $s_\gamma = s_\alpha$ and $\kappa \geq \max\{s_\gamma, s_\alpha\} - 1$, this requires $s_\gamma > 2r/7$, which is only slightly stronger than the $s_\gamma > r/4$ condition of Robins et al.(2009) that is required for existence of a semiparametric efficient estimator. Also, existence of K such that the remainder will be of size no larger than $\bar{\Delta}_n^*$ requires

$$2\zeta_\gamma + \zeta_\alpha \geq 1.$$

For example, if $\zeta_\gamma = \zeta_\alpha$ this requires $\zeta_\gamma > 1/3$, which is weaker than the condition $\zeta_\gamma > 1/2$ for the remainder for the plug-in estimator. In these ways the DCDR estimator improves on the plug-in estimator.

7 Appendix

This Appendix gives the proofs of the results in the body of the paper. We begin with the proofs of Lemma 1 and Lemma 6 because they are not restricted to series estimators.

Proof of Lemma 1: Define $\hat{\Delta}_{i\ell} = m(z_i, \hat{\gamma}) - m(z_i, \gamma_0) - \bar{m}(\hat{\gamma}_\ell) + \beta_0$ for $i \in I_\ell$ and let $Z(I_\ell)^c$ denote the set of observations z_i for $i \notin I_\ell$. Note that $E[\hat{\Delta}_{i\ell}|Z(I_\ell)^c] = 0$ by construction for $i \in I_\ell$. Also by independence of the observations, $E[\hat{\Delta}_{i\ell}\hat{\Delta}_{j\ell}|Z(I_\ell)^c] = 0$ for $i, j \in I_\ell$. Furthermore, $E[\hat{\Delta}_{i\ell}^2|Z(I_\ell)^c] \leq \int [m(z, \hat{\gamma}_\ell) - m(z, \gamma_0)]^2 F_0(dz) = O_p((\Delta_n^m)^2)$ for $i \in I_\ell$. Then we have

$$E[\left(\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}\hat{\Delta}_{i\ell}\right)^2 |Z(I_\ell)^c] = \frac{1}{n}E[\left(\sum_{i\in I_\ell}\hat{\Delta}_{i\ell}\right)^2 |Z(I_\ell)^c] = \frac{\bar{n}_\ell}{n}E[\hat{\Delta}_{i\ell}^2|Z(I_\ell)^c] = O_p((\Delta_n^m)^2).$$

Therefore, by the Markov inequality we have $\sum_{i\in I_\ell}\hat{\Delta}_{i\ell}/\sqrt{n} = O_p(\Delta_n^m)$. The first conclusion then follows from

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sum_{\ell=1}^{L}\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}\hat{\Delta}_{i\ell} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(z_i, \gamma_0) - \beta_0] + \sqrt{n}\sum_{\ell=1}^{L}\frac{\bar{n}_\ell}{n}[\bar{m}(\hat{\gamma}_\ell) - \beta_0].$$

For the second conclusion note by the subsamples being as close to equal size as possible,

$$\frac{\bar{n}_\ell}{\hat{n}_\ell} = \frac{\bar{n}_\ell/n}{\hat{n}_\ell/n} = \frac{1/L}{(L-1)/L} + O(n^{-1}) = \frac{1}{(L-1)} + O(n^{-1}).$$

Then by

$$\begin{aligned}
\sqrt{n}\sum_{\ell=1}^{L}\frac{\bar{n}_\ell}{n}[\bar{m}(\hat{\gamma}_\ell) - \beta_0] &= \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\bar{n}_\ell\sqrt{\frac{1}{\hat{n}_\ell}}\sqrt{\hat{n}_\ell}[\bar{m}(\hat{\gamma}_\ell) - \beta_0] = \sum_{\ell=1}^{L}\frac{\bar{n}_\ell}{\hat{n}_\ell}\frac{1}{\sqrt{n}}\sum_{i\notin I_\ell}\phi(z_i) + O_p(\Delta_n^\phi) \\
&= \frac{1}{L-1}\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\notin I_\ell}\phi(z_i) + O_p(\Delta_n^\phi + n^{-1}) \\
&= \frac{1}{L-1}\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}(\sum_{i=1}^{n}\phi(z_i) - \sum_{i\in I_\ell}\phi(z_i)) + O_p(\Delta_n^\phi + n^{-1}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(z_i) + O_p(\Delta_n^\phi + n^{-1}).
\end{aligned}$$

The conclusion then follows by the triangle inequality. *Q.E.D.*

Proof of Lemma 6: By adding and subtracting terms it follows that for $\varepsilon_i = y_i - \gamma_0(x_i)$ and $\phi(z_i) = \alpha_0(x_i)[y_i - \gamma_0(x_i)]$

$$\begin{aligned}
\tilde{\alpha}_\ell(x_i)[y_i - \hat{\gamma}_\ell(x_i)] &= \phi(z_i) - \alpha_0(x_i)[\hat{\gamma}(x_i) - \gamma_0(x_i)] + [\tilde{\alpha}_\ell(x_i) - \alpha_0(x_i)]\varepsilon_i \\
&\quad - [\tilde{\alpha}_\ell(x_i) - \alpha_0(x_i)][\hat{\gamma}(x_i) - \gamma_0(x_i)].
\end{aligned}$$

The first conclusion of Lemma 1 with $m(z,\gamma)=\alpha_0(x)\gamma(x)$ gives

$$\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\alpha_0(x_i)[\hat{\gamma}_\ell(x_i)-\gamma_0(x_i)]=\sqrt{n}\sum_{\ell=1}^{L}\frac{\bar{n}_\ell}{n}\int\alpha(x)[\hat{\gamma}_\ell(x)-\gamma_0(x)]F_0(dx)+O_p(\Delta_n^\gamma).$$

Assumption 1 and the first conclusion of Lemma 1 also give

$$\begin{aligned}\sqrt{n}\sum_{\ell=1}^{L}\frac{\bar{n}_\ell}{n}\int\alpha(x)[\hat{\gamma}_\ell(x)-\gamma_0(x)]F_0(dx)&=\sqrt{n}\sum_{\ell=1}^{L}\frac{\bar{n}_\ell}{n}[\bar{m}(\hat{\gamma}_\ell)-\beta_0]\\&=\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[m(z_i,\hat{\gamma}_\ell)-m(z_i,\gamma_0)]+O_p(\Delta_n^m).\end{aligned}$$

In addition, if we take $\gamma=\alpha$ and $m(z,\alpha)=\alpha(x)\varepsilon$ then $\int m(z,\alpha)F_0(dz)=0$, so that by Lemma 1,

$$\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[\tilde{\alpha}_\ell(x_i)-\alpha_0(x_i)]\varepsilon_i=O_p(\Delta_n^\alpha).$$

Then collecting terms we have

$$\begin{aligned}\sqrt{n}(\tilde{\beta}-\beta_0)&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(z_i,\gamma_0)-\beta_0]\\&+\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(z_i,\hat{\gamma})-m(z_i,\gamma_0)+\tilde{\alpha}_\ell(x_i)[y_i-\hat{\gamma}_\ell(x_i)]\}\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i)+\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\alpha_0(x_i)[\hat{\gamma}_\ell(x_i)-\gamma_0(x_i)]+O_p(\Delta_n^m+\Delta_n^\gamma)\\&\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{-\alpha_0(x_i)[\hat{\gamma}(x_i)-\gamma_0(x_i)]+[\tilde{\alpha}_\ell(x_i)-\alpha_0(x_i)]\varepsilon_i\}\\&-\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[\tilde{\alpha}_\ell(x_i)-\alpha_0(x_i)][\hat{\gamma}(x_i)-\gamma_0(x_i)]\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i)-\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[\tilde{\alpha}_\ell(x_i)-\alpha_0(x_i)][\hat{\gamma}(x_i)-\gamma_0(x_i)]\\&+O_p(\Delta_n^m+\Delta_n^\gamma+\Delta_n^\alpha).Q.E.D.\end{aligned}$$

We now turn to proofs of the results involving series estimators. Let $\Sigma=E[p(x_i)p(x_i)^T]$. It follows from Assumption 3 that Σ is nonsingular, so we can replace $p(x)$ by $\Sigma^{-1/2}p(x)$ and so normalize $\Sigma=I$ without changing the assumptions. We impose this normalization throughout. Also, throughout the Appendix C will denote a generic constant not depending on n or K.

We next prove the key result in eq. (4.2) for a zero order spline. Let $r(x) = \gamma_0(x) - \gamma_K(x)$ and $\hat{h}_2 = \sum_{i=1}^{n} p(x_i)r(x_i)/n$ as in the body of the paper. Also let $\|A\|_{op}$ denote the operator norm of a symmetric matrix A, being the largest absolute value of eigenvalues.

Lemma A1: *If Assumptions 1-6 are satisfied,* $\kappa = 0$, $K[\ln(n)]^2/n \longrightarrow 0$, *then for* $\hat{U} = \sum_{j=0}^{J-1}(I - \hat{\Sigma})^j \hat{h}_2$, $\hat{W} = \hat{\Sigma}^{-1}(I - \hat{\Sigma})^J \hat{h}_2$, $J = int[\ln(n)]$ *and any constant* $\Delta > 0$,

$$\left\| E[\hat{U}\hat{U}^T] \right\|_{op} \le C\frac{K^{-2\zeta_\gamma}[\ln(n)]^2}{n}, \hat{W}^T\hat{W} = o_p(n^{-\Delta}).$$

Proof: Let $Q_i = p(x_i)p(x_i)^T$, $\Delta_i = I - Q_i$, and $h_i = p(x_i)r(x_i)$. Note that $E[\Delta_i] = 0$ and $E[h_i] = 0$. For each j let $L = 2j + 2$. Let $\hat{U}_j = (I - \hat{\Sigma})^j \hat{h}_2$. Then we have

$$E[\hat{U}_j\hat{U}_j^T] = \frac{1}{n^{2j+2}} \sum_{i_1,...,i_L=1}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i_{j+1}} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)].$$

Consider any $(i_1, ..., i_L)$ such that $i_{j+1} \neq i_{j+2}$. Let $i^* = i_{j+1}$ and let $Z_{i^*}^c$ denote the vector of observations other than z_{i^*}. Note that

$$E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i_{j+1}} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)] = E[E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i^*} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right) |Z_{i^*}^c]].$$

We proceed to show that

$$E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i^*} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right) |Z_{i^*}^c] = 0.$$

Note that conditional on $Z_{i^*}^c$ we can treat all terms where $i_\ell \neq i^*$ as constant. Also, because $i_{j+1} \neq i_{j+2}$ all terms where $i_\ell = i^*$ depend only on $p(x_{i^*})$. Therefore for the scalar $r(x) = \gamma_0(x) - \gamma_K(x)$ we have

$$E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i^*} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right) |Z_{i^*}^c] = E[A_1(p(x_{i^*}))p(x_{i^*})r(x_{i^*})A_2(p(x_{i^*}))] = E[A(p(x_{i^*}))r(x_{i^*})],$$

where $A_1(p)$ and $A_2(p)$ are $K \times K$ and $1 \times K$ matrices of functions of p and $A(p) = A_1(p)pA_2(p)$. Let X_k denote the interval where $p_k(x)$ is nonzero. Note that $p_k(x) = 1(x \in X_k)c_k$ for a constant c_k, and hence

$$A(p(x_{i^*})) = \sum_{k=1}^{K} A_k 1(x_{i^*} \in X_k), A_k = A((0, ..., 0, c_k, 0, ..., 0)^T).$$

Therefore by orthogonality of each $p_k(x_i)$ with $r(x_i)$ in the population,

$$E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i^*} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right) |Z_{i^*}^c] = \sum_{k=1}^{K} A_k E[1(x_{i^*} \in X_k)r(x_{i^*})] = \sum_{k=1}^{K} A_k c_k^{-1} E[p_k(x_{i^*})r(x_{i^*})] = 0.$$

Therefore by iterated expectations, if $i_{j+1} \neq i_{j+2}$ we have

$$E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i_{j+1}} h_{i_{j+2}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)] = 0.$$

It then follows that for $\Psi = E[h_{i_{j+1}} h_{i_{j+1}}^T] = E[r(x_i)^2 p(x_i) p(x_i)^T]$ and $\tilde{\Delta}_{i_{j+1}} = h_{i_{j+1}} h_{i_{j+1}}^T - \Psi$,

$$\begin{aligned}
E[\hat{U}_j \hat{U}_j^T] &= \frac{1}{n^{2j+2}} \sum_{i_1,..,i_{j+1},i_{j+3},...,i_L=1}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) h_{i_{j+1}} h_{i_{j+1}}^T \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)] = T_1^j + T_2^j, \\
T_1^j &= \frac{1}{n^{2j+1}} \sum_{i_1,..,i_j,i_{j+3},...,i_L=1}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \Psi \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)], \\
T_2^j &= \frac{1}{n^{2j+2}} \sum_{i_1,..,i_{j+1},i_{j+3},...,i_L=1}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)].
\end{aligned}$$

Consider first T_2^j. Note that Δ_i and $\tilde{\Delta}_i$ are diagonal matrices, so that $E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)]$ is a diagonal matrix, with k^{th} diagonal element given by $E[\left(\Pi_{\ell=1}^{j}\Delta_{k,i_\ell}\right) \tilde{\Delta}_{k,i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{k,i_\ell}\right)]$ where

$$\Delta_{k,i} = p_k(x_i)^2 - E[p_k(x_i)^2], \tilde{\Delta}_{k,i_{j+1}} = r(x_i)^2 p_k(x_i)^2 - E[r(x_i)^2 p_k(x_i)^2].$$

The largest absolute value of the eigenvalues of a diagonal matrix is the maximum of the absolute values of the diagonal elements, so it suffices to show that the conclusion holds for these diagonal elements. We will consider the k^{th} diagonal element but for notational convenience drop the k subscript in what follows.

Note that $p_k(x_i)^2 \leq BK$ for some B that does not vary with k or j. Also, for any random variable Y_i and $\mu = E[Y_i]$, note that by Jensen's inequality, $|\mu|^s \leq E[|Y_i|^s]$ for $s \geq 1$. Then for any positive s,

$$E[|Y_i - \mu|^s] \leq E[(|Y_i| + |\mu|)^s] \leq E[2^{s-1}(|Y_i|^s + |\mu|^s)] \leq 2^{s-1}(E[|Y_i|^s] + |\mu|^s) \leq 2^s E[|Y_i|^s]$$

Then for any positive integer s, by the triangle inequality and the definitions of Δ_i,

$$|E[\Delta_i^s]| \leq 2^s E[p_k(x_i)^{2s}] \leq 2^s (BK)^{s-1} E[p_k(x_i)^2] \leq (4BK)^{s-1} \leq (CK)^{s-1}. \tag{7.1}$$

Also, by $r(x_i)^2 \leq DK^{-2\zeta_\gamma}$ we have

$$\begin{aligned}
\left|E[(\Delta_i)^s \tilde{\Delta}_i]\right| &\leq E[|\Delta_i|^s (r(x_i)^2 p_k(x_i)^2 + E[r(x_i)^2 p_k(x_i)^2])] \qquad (7.2)\\
&\leq E[(p_k(x_i)^2 + E[p_k(x_i)^2])^{s+1}] DK^{-2\zeta_\gamma} \\
&\leq 2^{s+1} E[p_k(x_i)^{2s+2}] DK^{-2\zeta_\gamma} \leq 2^{s+1}(BK)^s DK^{-2\zeta_\gamma} \\
&\leq (4(D+1)BK)^s K^{-2\zeta_\gamma} \leq (CK)^s K^{-2\zeta_\gamma}.
\end{aligned}$$

Next consider

$$T_2^j = \frac{1}{n^{2j+2}} \sum_{i_1,..,i_{j+1},i_{j+3}....,i_{2j+2}=1}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)].$$

The only terms in this sum that are nonzero are those where every index i_ℓ is equal to at least one other index $i_{\ell'}$, i.e. where each index is "matched" with at least one other. Let $\tilde{\imath} = (i_1, .., i_{j+1}, i_{j+3}...., i_{2j+2})^T$ denote the $2j+1$ dimensional vector of indices where each i_ℓ is an integer in $[1, n]$. Let Υ_d denote a set of all such $\tilde{\imath}$ with specified indices that are equal to each other, but those matched indices are not equal to any other indices. For example, one Υ_d is the set of $\tilde{\imath}$ with $i_1 = i_{j+1} = i_{j+3} = \cdots = i_{2J+2}$ and another is the set of $\tilde{\imath}$ with $i_1 = i_2, i_3 = \cdots = i_{2J+2}, i_2 \neq i_3$. For each d each group of index coordinates that are equal to each other can be thought of as a group of matching indices that we index by g_d. Let m_{g_d} denote the number of indices in group g_d and G_d denote the total number of groups. Note that the total number of indices is $2j+1 = \sum_{g_d=1}^{G_d} m_{g_d}$. Also, by eqs. (7.1) and (7.2) for each $\tilde{\imath} \in \Upsilon_d$ we have

$$|E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)]| \leq K^{-2\zeta_\gamma} \prod_{g_d=1}^{G_d} (CK)^{m_{g_d}-1} = K^{-2\zeta_\gamma} (CK)^{2j+1-G_d}.$$

Also, the number of indices in Υ_d is less than or equal to n^{G_d} since each match can be regarded as a single index. Therefore,

$$\begin{aligned}\left|\frac{1}{n^{2j+2}} \sum_{\tilde{\imath}\in\Upsilon_d}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)]\right| &\leq \frac{1}{n^{2j+2}} \sum_{\tilde{\imath}\in\Upsilon_d}^{n} \left|E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)]\right| \\ &\leq \left(\frac{1}{n^{2j+2}}\right) n^{G_d} K^{-2\zeta_\gamma} (CK)^{2j+1-G_d} \\ &= \frac{1}{n} K^{-2\zeta_\gamma} \left(\frac{CK}{n}\right)^{2j+1-G_d}.\end{aligned}$$

By hypothesis $K/n \longrightarrow 0$ so that for large enough n we have $CK/n < 1$. For such n we have $(CK/n)^{2j+1-G_d}$ decreasing in G_d. Also, the largest G_d is j, because each group must contain at least two elements. Therefore, for large enough n we have

$$\left|\frac{1}{n^{2j+2}} \sum_{\tilde{\imath}\in\Upsilon_d}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)]\right| \leq \frac{1}{n} K^{-2\zeta_\gamma} \left(\frac{CK}{n}\right)^{j+1}.$$

Note that the bound on the right does not depend on d. Let D denote the total number of possible Υ_d. Then since $E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)] = 0$ if $\tilde{\imath} \notin \cup_{d=1}^{D}\Upsilon_d$ we have

$$\left|T_2^j\right| \leq \sum_{d=1}^{D} \left|\frac{1}{n^{2j+2}} \sum_{\tilde{\imath}\in\Upsilon_d}^{n} E[\left(\Pi_{\ell=1}^{j}\Delta_{i_\ell}\right) \tilde{\Delta}_{i_{j+1}} \left(\Pi_{\ell=j+3}^{L}\Delta_{i_\ell}\right)]\right| \leq \frac{D}{n} K^{-2\zeta_\gamma} \left(\frac{CK}{n}\right)^{j+1}.$$

Note that there are exactly j^{2j+1} ways of forming $2j+1$ indices into j groups. Ignoring the fact that we can exclude ways where any group has only one index we have the bound $D \leq j^{2j+1}$. Plugging in this bound into the above inequality and maximizing over diagonal elements gives

$$\left\|T_2^j\right\|_{op} \leq \frac{j^{2j+1}K^{-2\zeta_\gamma}}{n}\left(\frac{CK}{n}\right)^{j+1}.$$

Arguing similarly for T_1^j gives

$$\left\|T_1^j\right\|_{op} \leq \frac{j^{2j}K^{-2\zeta_\gamma}}{n}\left(\frac{CK}{n}\right)^{j},$$

where we take $0^0 = 1$.

Next note that by $K\ln(n)^2/n \longrightarrow 0$ we have $CK/n \leq 1/[2\ln(n)^2]$ for large enough n. Also, $j/\ln(n) \leq 1$ for all $j < J$. Then for n large enough

$$\sum_{j=0}^{J-1} j^{2j}\left(\frac{CK}{n}\right)^j \leq \sum_{j=0}^{J-1} j^{2j}\left(\frac{1}{2\ln(n)^2}\right)^j \leq \sum_{j=0}^{J-1}\left(\frac{j}{\ln(n)}\right)^{2j}\left(\frac{1}{2}\right)^j \leq \sum_{j=0}^{J-1}\left(\frac{1}{2}\right)^j \leq \sum_{j=0}^{\infty}\left(\frac{1}{2}\right)^j = \frac{1}{1-\varepsilon_n} \leq 2.$$

Similarly it follows that for large enough n,

$$\sum_{j=0}^{J-1} j^{2j+1}\left(\frac{CK}{n}\right)^{j+1} \leq \frac{1}{2\ln(n)}\sum_{j=0}^{J-1}\left(\frac{j}{\ln(n)}\right)^{2j+1}\left(\frac{1}{2}\right)^j \leq \frac{1}{\ln(n)}.$$

Then we have for large enough n,

$$\begin{aligned}\left\|\sum_{j=0}^{J-1} E[\hat{U}_j\hat{U}_j^T]\right\|_{op} &\leq \left\|\sum_{j=0}^{J-1}(T_1^j + T_2^j)\right\|_{op} \leq \sum_{j=0}^{J-1}\left(\left\|T_1^j\right\|_{op} + \left\|T_2^j\right\|_{op}\right)\\ &\leq \frac{K^{-2\zeta_\gamma}}{n}\left(2+\frac{1}{\ln(n)}\right) \leq \frac{CK^{-2\zeta_\gamma}}{n}.\end{aligned}$$

Also by the Cauchy Schwartz inequality, $\hat{U}\hat{U}^T = \left(\sum_{j=0}^{J-1}\hat{U}_j\right)\left(\sum_{j=0}^{J-1}\hat{U}_j\right)^T \leq J^2\sum_{j=0}^{J-1}\hat{U}_j\hat{U}_j^T$. Therefore, for large enough n,

$$\left\|E[\hat{U}\hat{U}^T]\right\|_{op} \leq J^2\left\|\sum_{j=0}^{J-1}E[\hat{U}_j\hat{U}_j^T]\right\|_{op} \leq \frac{C\ln(n)^2K^{-2\zeta_\gamma}}{n},$$

giving the first conclusion.

For the second conclusion note that for any $\Delta > 0$,

$$\ln\{n^\Delta[\ln(n)]^{-2\ln(n)+2}\} = \ln(n)[\Delta - 2\ln(\ln(n))] + 2\ln(\ln(n)) \longrightarrow -\infty.$$

It follows that $[\ln(n)]^{-2\ln(n)+2} = o(n^{-\Delta})$ for any Δ. Also, by $K/n = o([1/\ln(n)]^2)$ we have $K\ln(K)/n = o(1/\ln(n))$, so that

$$\left(\frac{K\ln(K)}{n}\right)^{2J} = o([\ln(n)]^{-2int(\ln(n))}) = o([\ln(n)]^{-2(\ln(n))+2}) = o(n^{-\Delta}),$$

for any $\Delta > 0$. Then we have

$$\hat{1}\hat{W}^T\hat{W} \leq 4\hat{h}_2^T(I-\hat{\Sigma})^{2J}\hat{h}_2 \leq 4\hat{h}_2^T\hat{h}_2 \left\|I-\hat{\Sigma}\right\|_{op}^{2J} = O_p(\frac{K^{1-2\zeta_\gamma}}{n}\left[\frac{K\ln(K)}{n}\right]^{2J}) = o_p(n^{-\Delta}),$$

for any $\Delta > 0$ by Rudelson's (1999) law of large numbers for random matrices, giving the second conclusion. *Q.E.D.*

In the Appendix we focus on one subset $\bar{I} = I_\ell$ of observations and let $\hat{I}$ and $\tilde{I}$ denote the observations used to compute $\hat{\delta}$ and $\tilde{\delta}_\alpha$ respectively. Let $\bar{n}$, $\hat{n}$, $\tilde{n}$ denote the number of elements of $\bar{I}$, $\hat{I}$, and $\tilde{I}$ respectively and

$$\bar{F}\{g(z)\} = \frac{1}{\bar{n}}\sum_{i\in\bar{I}} g(z_i), \hat{F}\{g(z)\} = \frac{1}{\hat{n}}\sum_{i\in\hat{I}} g(z_i), \tilde{F}\{g(z)\} = \frac{1}{\tilde{n}}\sum_{i\in\tilde{I}} g(z_i),$$

denote averages over the respective subsets of observations.

Next we make a few definitions we will use throughout. Let ζ_γ, ζ_α, δ, γ_K, δ_α, and α_K be as defined following Assumption 4. Also, let

$$\begin{aligned}
&\varepsilon_i = y_i - \gamma_0(x_i), r_i = \gamma_0(x_i) - \gamma_K(x_i), \eta_i = v(z_i) - p(x_i)\alpha_0(x_i), r_i^\alpha = \alpha_0(x_i) - \alpha_K(x_i),\\
&\hat{h}_1 = \hat{F}\{p(x)\varepsilon\}, \hat{h}_2 = \hat{F}\{p(x)r\}, \tilde{h}_1^\alpha = \tilde{F}\{\eta\}, \tilde{h}_2^\alpha = \tilde{F}\{p(x)r^\alpha\}, \hat{\Sigma} = \hat{F}\{p(x)p(x)^T\}, \tilde{\Sigma} = \tilde{F}\{p(x)p(x)^T\},\\
&\hat{\Delta}_1 = \hat{\Sigma}^{-}\hat{h}_1, \hat{\Delta}_2 = \hat{\Sigma}^{-}\hat{h}_2, \tilde{\Delta}_1^\alpha = \tilde{\Sigma}^{-}\tilde{h}_1^\alpha, \tilde{\Delta}_2^\alpha = \tilde{\Sigma}^{-}\tilde{h}_2^\alpha, \bar{\Sigma} = \bar{F}\{p(x)p(x)^T\},
\end{aligned}$$

One piece of algebra we will use throughout is that, when $\hat{\Sigma}$ and $\tilde{\Sigma}$ are nonsingular, by adding and subtracting $\hat{\Sigma}^{-1}\hat{F}\{p(x)\gamma_0(x)\}$ and $\tilde{\Sigma}^{-1}\tilde{F}\{p(x)\alpha_0(x_i)\}$ respectively we have

$$\hat{\delta} - \delta = \hat{\Delta}_1 + \hat{\Delta}_2, \tilde{\delta}_\alpha - \delta_\alpha = \tilde{\Delta}_1^\alpha + \tilde{\Delta}_2^\alpha. \tag{7.3}$$

Some properties of these objects will be useful in the proofs to follow. We collect these properties in the following result. Let $\hat{1}$ and $\tilde{1}$ denote the indicator function that the smallest eigenvalue of $\hat{\Sigma}$ or $\tilde{\Sigma}$ is larger than $1/2$ respectively. As in Belloni et al.(2015) $\Pr(\hat{1}=1) \longrightarrow 1$ and $\Pr(\bar{1}=1) \longrightarrow 1$. Also, let $\hat{Z}^c$, $\tilde{Z}^c$, $\bar{Z}^c$ denote all the other observations other than those indexed by $\hat{I}$, $\tilde{I}$, or $\bar{I}$ respectively and $X = (x_1, ..., x_n)$.

LEMMA A2: *If Assumptions 1-6 are is satisfied then*

$$\text{i) } \hat{1}\left\|\hat{\Delta}_1\right\| = O_p\left(\frac{K}{n}\right);\ \text{ii) } \hat{1}\left\|\hat{\Delta}_2\right\| = o_p\left(K^{-2\zeta_\gamma}\frac{K}{n}\right);$$

$$\text{iii) } \hat{1}\left\|\hat{\Delta}_1^\alpha\right\| = O_p\left(\frac{(1+d_K)K}{n}\right),\ \text{iv) } \hat{1}\left\|\hat{\Delta}_2^\alpha\right\| = O_p\left(\frac{K}{n}\right),$$

$$\text{v) } \hat{1}\left\|\hat{\delta}-\delta\right\|^2 = O_p\left(\frac{K}{n}\right);\ \text{vi) } \tilde{1}\left\|\tilde{\delta}_\alpha-\delta_\alpha\right\|^2 = O_p\left(\frac{d_K K}{n}\right),$$

$$\text{vii) } \hat{1}E[\hat{\Delta}_1\hat{\Delta}_1^T|X,\hat{Z}^c] \le \frac{C}{n}I,\ \text{viii) } \hat{1}\int[\hat{\gamma}(x)-\gamma_0(x)]^2F_0(dx) = O_p(\frac{K}{n}+K^{-2\zeta_\gamma}),$$

$$\text{ix) } \tilde{1}\int[\tilde{\alpha}(x)-\alpha_0(x)]^2F_0(dx) = O_p\left(\frac{(d_K+1)K}{n}+K^{-2\zeta_\alpha}\right).$$

Proof: Note that for $\varepsilon_i = y_i - \gamma_0(x_i)$, $E[\varepsilon_i^2|x_i] = Var(y_i|x_i) \le C$. Note that $\hat{1}\hat{\Sigma}^{-2} \le 4I$ in the positive semi-definite semi-order so that

$$E[\hat{1}\left\|\hat{\Delta}_1\right\|^2] \le 4E[\hat{h}_1^T\hat{h}_1] = \frac{4}{\hat{n}^2}\sum_{i,j\in\hat{I}}E[p(x_i)^Tp(x_j)\varepsilon_i\varepsilon_j] = \frac{4}{\hat{n}}E[\|p(x_i)\|^2\varepsilon_i^2] \le \frac{4C}{\hat{n}}E[\|p(x_i)\|^2] = O(\frac{K}{n}).$$

The first conclusion then follows by the Markov inequality. Next, we have $\sup_x|\gamma_K(x)-\gamma_0(x_i)| = O(K^{-\zeta})$ and hence for

$$E[\hat{1}\left\|\hat{\Delta}_2\right\|^2] \le 4E[\hat{h}_2^T\hat{h}_2] = \frac{4}{\hat{n}^2}\sum_{i,j\in\hat{I}}E[p(x_i)^Tp(x_j)r_ir_j] = \frac{4}{\hat{n}}E[\|p(x_i)\|^2]O(K^{-2\zeta_\gamma}) = O\left(K^{-2\zeta_\gamma}\frac{K}{n}\right),$$

so the second equality also follows by the Markov inequality. Next, note that

$$E[\eta_i^T\eta_i] \le 2E[v(z_i)^Tv(z_i)] + 2E[\alpha_0(x_i)^2\|p(x_i)\|^2] = O(K(d_K+1)).$$

Then we have

$$E[\tilde{1}\left\|\tilde{\Delta}_1^\alpha\right\|^2] \le 4E[\tilde{h}_1^{\alpha T}\tilde{h}_1^\alpha] = \frac{4}{\tilde{n}^2}\sum_{i,j\in\tilde{I}}E[\eta_i^T\eta_j] = \frac{4}{\hat{n}}E[\eta_i^T\eta_i] = O(\frac{K(d_K+1)}{n}),$$

so the third conclusion follows from the Markov inequality. The fourth conclusion follows exactly like the second conclusion. the fifth and sixth conclusions follow by eq. (7.3) and the triangle inequality.

Next, note that by independence of the observations

$$E[\hat{1}\hat{\Delta}_1\hat{\Delta}_1^T|X,\hat{Z}^c] = \hat{1}\hat{\Sigma}^{-1}E[\hat{h}_1\hat{h}_1^T|X]\hat{\Sigma}^{-1} = \hat{1}\hat{\Sigma}^{-1}\left\{\frac{1}{\hat{n}^2}\sum_{i,j\in\hat{I}}p(x_i)p(x_j)^TE[\varepsilon_i\varepsilon_j|X]\right\}\hat{\Sigma}^{-1}$$
$$= \hat{1}\hat{\Sigma}^{-1}\left\{\frac{1}{\hat{n}^2}\sum_{i\in\hat{I}}p(x_i)p(x_i)^TE[\varepsilon_i^2|x_i]\right\}\hat{\Sigma}^{-1} \le \hat{1}\frac{C}{\hat{n}}\hat{\Sigma}^{-1} \le \frac{2C}{n}I,$$

giving the seventh conclusion.

Next, note that $\int p(x)[\gamma_K(x) - \gamma_0(x)]F_0(dx) = 0$, so that

$$\hat{1}\int[\hat{\gamma}(x) - \gamma_0(x)]^2 F_0(dx) = \hat{1}\int[\hat{\gamma}(x) - \gamma_K(x) + \gamma_K(x) - \gamma_0(x)]^2 F_0(dx)$$
$$= \hat{1}\left\|\hat{\delta} - \delta\right\|^2 + \hat{1}K^{-2\zeta_\gamma} = O_p\left(\frac{K}{n} + K^{-2\zeta_\gamma}\right),$$

giving the eighth conclusion. The last conclusion follows similarly. *Q.E.D.*

Next, we give an important intermediate result:

LEMMA A3: *If Assumptions 1-6 are satisfied then*

$$\hat{1}\int[m(z,\hat{\gamma}) - m(z,\gamma_0)]^2 F_0(dz) = O_p\left(\frac{d_K K}{n} + K^{-2\zeta_m}\right).$$

Proof: By linearity of $m(z,\gamma) - m(z,0)$, we have $m(z,\hat{\gamma}) - m(z,\gamma_K) = v(z)^T(\hat{\delta} - \delta)$. Then by Lemma A2,

$$\hat{1}\int[m(z,\hat{\gamma}) - m(z,\gamma_0)]^2 F_0(dz) \leq 2\hat{1}\int[m(z,\hat{\gamma}) - m(z,\gamma_K)]^2 F_0(dz) + 2\hat{1}\int[m(z,\gamma_K) - m(z,\gamma_0)]^2 F_0(dz)$$
$$\leq 2\hat{1}(\hat{\delta} - \delta)^T E[v(z_i)v(z_i)^T](\hat{\delta} - \delta) + O(K^{-2\zeta_m})$$
$$\leq 2d_K\hat{1}\left\|\hat{\delta} - \delta\right\|^2 + O(K^{-2\zeta_m}) = O_p\left(\frac{d_K K}{n} + K^{-2\zeta_m}\right). \ Q.E.D.$$

The proof of the results for the doubly robust estimators will make use of a few Lemmas, that we now state.

LEMMA A4: *If Assumptions 1-6 are satisfied then the hypotheses of Lemma 6 are satisfied with*

$$\Delta_n^m = \sqrt{\frac{d_K K}{n}} + K^{-\zeta_m}, \Delta_n^\gamma = \sqrt{\frac{K}{n}} + K^{-\zeta_\gamma}, \Delta_n^\alpha = \sqrt{\frac{d_K K}{n}} + K^{-\zeta_\alpha}.$$

Proof: The first conclusion follows by Lemma A3 and the second and third by parts viii) and ix) of Lemma A2. *Q.E.D.*

LEMMA A5: *If Assumptions 1-6 are satisfied and* $\hat{\gamma}_\ell$ and $\tilde{\alpha}_\ell$ are computed from distinct samples then for $\bar{\Sigma} = \bar{F}\{p(x)p(x)^T\}$

$$\sqrt{n}\bar{F}\{[\tilde{\alpha}_\ell(x) - \alpha_0(x)][\hat{\gamma}_\ell(x) - \gamma_0(x)]\} = \sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha + O_p(\bar{\Delta}_n^* + \Delta_n^m).$$

Proof: Let $\bar{h}_2 = \bar{F}\{p(x)[\gamma_K(x) - \gamma_0(x)]\}$ and $\bar{h}_2^\alpha = \bar{F}\{p(x)[\alpha_K(x) - \alpha_0(x)]\}$. Note that

$$\begin{aligned}
&\bar{F}\{[\tilde{\alpha}_\ell(x) - \alpha_0(x)][\hat{\gamma}_\ell(x) - \gamma_0(x)]\} \\
&= \bar{F}\{[p(x)^T(\tilde{\delta}_\alpha - \delta_\alpha) + \alpha_K(x) - \alpha_0(x)][p(x)^T(\hat{\delta} - \delta) + \gamma_K(x) - \gamma_0(x)]\} \\
&= (\hat{\delta} - \delta)^T\bar{\Sigma}(\tilde{\delta}_\alpha - \delta_\alpha) + (\hat{\delta} - \delta)^T\bar{h}_2^\alpha + (\tilde{\delta}_\alpha - \delta_\alpha)^T\bar{h}_2 + \bar{F}\{[\alpha_K(x) - \alpha_0(x)][\gamma_K(x) - \gamma_0(x)]\}.
\end{aligned}$$

By the Markov inequality

$$\sqrt{n}\bar{F}\{[\alpha_K(x) - \alpha_0(x)][\gamma_K(x) - \gamma_0(x)]\} = O_p(\sqrt{n}K^{-\zeta_\gamma - \zeta_\alpha}). \tag{7.4}$$

Note that

$$E[\tilde{h}_2^\alpha(\tilde{h}_2^\alpha)^T] = \frac{1}{\bar{n}}E[p(x_i)p(x_i)^T(r_i^\alpha)^2] \le C\frac{1}{n}I.$$

Therefore by Lemma A2 we have

$$E[\{\hat{1}(\hat{\delta} - \delta)^T\bar{h}_2^\alpha\}^2|\bar{Z}^c] = \hat{1}(\hat{\delta} - \delta)^T E[\tilde{h}_2^\alpha(\tilde{h}_2^\alpha)^T](\hat{\delta} - \delta) \le C\hat{1}\frac{1}{n}\left\|\hat{\delta} - \delta\right\|^2 = O_p(\frac{K}{n^2}).$$

Then by the Markov inequality it follows that

$$\sqrt{n}(\hat{\delta} - \delta)^T\bar{h}_2^\alpha = O_p(\sqrt{\frac{K}{n}}). \tag{7.5}$$

It follows similarly that

$$\sqrt{n}(\tilde{\delta}_\alpha - \delta_\alpha)^T\bar{h}_2 = O_p(\sqrt{\frac{d_K K}{n}}). \tag{7.6}$$

Next, note that

$$(\hat{\delta} - \delta)^T\bar{\Sigma}(\tilde{\delta}_\alpha - \delta_\alpha) = \hat{\Delta}_1^T\bar{\Sigma}(\tilde{\delta}_\alpha - \delta_\alpha) + \hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_2^\alpha + \hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha.$$

Let $\bar{1}$ be the event that $\lambda_{\max}(\bar{\Sigma}) \le 2$. Then by conclusion vii) of Lemma A2, and $\bar{1}$, $\hat{1}$, and $\tilde{1}$ all functions of X we have

$$\begin{aligned}
E[\bar{1}\hat{1}\tilde{1}\{\hat{\Delta}_1^T\bar{\Sigma}(\tilde{\delta}_\alpha - \delta_\alpha)\}^2|X, \hat{Z}^c] &= \bar{1}\hat{1}\tilde{1}(\tilde{\delta}_\alpha - \delta_\alpha)^T\bar{\Sigma}E[\hat{\Delta}_1\hat{\Delta}_1^T|X, \hat{Z}^c]\bar{\Sigma}(\tilde{\delta}_\alpha - \delta_\alpha) \\
&\le C\frac{1}{n}\bar{1}\tilde{1}(\tilde{\delta}_\alpha - \delta_\alpha)^T\bar{\Sigma}^2(\tilde{\delta}_\alpha - \delta_\alpha) \le 4C\frac{1}{n}\tilde{1}\left\|\tilde{\delta}_\alpha - \delta_\alpha\right\|^2 \\
&= O_p(\frac{d_K K}{n^2}).
\end{aligned}$$

Therefore we have

$$\sqrt{n}\hat{\Delta}_1^T\bar{\Sigma}(\tilde{\delta}_\alpha - \delta_\alpha) = O_p(\sqrt{\frac{d_K K}{n}}). \tag{7.7}$$

Finally, note that by the Cauchy-Schwartz inequality

$$\hat{1}\hat{\Delta}_2^T\hat{\Delta}_2 \le \hat{1}2\hat{h}_2^T\hat{\Sigma}^{-1}\hat{h}_2 \le 2\hat{F}\{[\gamma_K(x) - \gamma_0(x)]^2\} = O_p(K^{-2\zeta_\gamma}).$$

It follows similarly that $\hat{1}(\hat{\Delta}_2^{\alpha})^T(\hat{\Delta}_2^{\alpha}) = O_p(K^{-2\zeta_\alpha})$ so that

$$\sqrt{n}\bar{1}\hat{1}\tilde{1}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_2^{\alpha} \leq 2\sqrt{n}\sqrt{\hat{1}\hat{\Delta}_2^T\hat{\Delta}_2}\sqrt{\tilde{1}(\tilde{\Delta}_2^{\alpha})^T(\tilde{\Delta}_2^{\alpha})} = O_p(\sqrt{n}K^{-\zeta_\gamma-\zeta_\alpha}). \tag{7.8}$$

The conclusion then follows by eqs. (7.4), (7.5), (7.6), (7.7), (7.8), and the triangle inequality. Q.E.D.

Proof of Theorem 2: It follows by Lemma A2 that the first hypothesis of Lemma 1 is satisfied with $\Delta_n^m = \sqrt{d_K/n} + K^{-\zeta_m}$. Let

$$\bar{m}(\gamma) = \int [m(z,\gamma) - m(z,\gamma)]F_0(dz) = E[\alpha_0(x_i)\gamma(x_i)],$$

where the first equality is a definition and the second follows by Assumption 1. Then the first conclusion of Lemma 1 holds.

Next let $n = \hat{n}_\ell$ and $\hat{\gamma} = \hat{\gamma}_\ell$ for some ℓ and $\phi(z) = \alpha_0(x)[y - \gamma_0(x)]$. Then it follows as in Ichimura and Newey (2017), pp. 29 that

$$\hat{1}\sqrt{n}[\bar{m}(\hat{\gamma}) - \beta_0 - \frac{1}{n}\sum_{i=1}^{n}\phi(z_i)] = \hat{1}\left(\hat{R}_1 + \hat{R}_2 + \hat{R}_3\right), \hat{R}_1 = \sqrt{n}E[\alpha_0(x_i)\{\gamma_K(x_i) - \gamma_0(x_i)\}], \tag{7.9}$$

$$\hat{R}_2 = \sqrt{n}v^T\hat{\Sigma}^{-1}\hat{h}_2, \hat{R}_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\alpha_K(x_i) - \alpha_0(x_i)][y_i - \gamma_0(x_i)].$$

By $\gamma_K(x_i) - \gamma_0(x_i)$ orthogonal to $p(x_i)$ in the population and the Cauchy-Schwartz inequality,

$$\left|\hat{R}_1\right| = \sqrt{n}\left|E[\{\alpha_0(x_i) - \alpha_K(x_i)\}\{\gamma_0(x_i) - \gamma_K(x_i)\}]\right| \leq \sqrt{n}\{E[\{\alpha_0(x_i) - \alpha_K(x_i)\}^2]E[\{\gamma_0(x_i) - \gamma_K(x_i)\}^2]\}^{1/2}$$
$$= O(\sqrt{n}K^{-\zeta_\gamma-\zeta_\alpha}) = O(\bar{\Delta}_n^*).$$

Also,

$$E[\hat{R}_3^2] = E[\{\alpha_K(x_i) - \alpha(x_i)\}^2\varepsilon_i^2] \leq CE[\{\alpha_K(x_i) - \alpha(x_i)\}^2] = O(K^{-2\zeta_\alpha}),$$

so by the Markov inequality,

$$\hat{R}_3 = O_p(K^{-\zeta_\alpha}) = O_p(\bar{\Delta}_n^*).$$

Next, note that $\hat{R}_2 = \hat{R}_{21} + \hat{R}_{22}$ where $\hat{R}_{21} = v^T\hat{h}_2$ and $\hat{R}_{22} = \sqrt{n}v^T(\hat{\Sigma}^{-1} - I)\hat{h}_2$. As noted following Assumption 4, $\sup_x|\gamma_K(x) - \gamma_0(x)| = O(K^{-\zeta_\gamma})$, so that

$$E[\hat{R}_{21}^2] = v^TE[p(x_i)p(x_i)^Tr_i^2]v \leq O(K^{-2\zeta_\gamma})v^Tv \leq O(K^{-2\zeta_\gamma})E[\alpha_0(x_i)^2] = O(K^{-2\zeta_\gamma}).$$

Then by the Markov inequality

$$\hat{R}_{21} = O_p(K^{-\zeta_\gamma}) = O_p(\bar{\Delta}_n^*).$$

Finally, note that $(\hat{\Sigma}^{-1}-I)\hat{h}_2 = \hat{U}+\hat{W}$ for $\hat{U}$ and $\hat{W}$ defined in the statement of Lemma A1, so that for any $\Delta > 0$ we have

$$\begin{aligned}\hat{1}\hat{R}_{22}^2 &= \hat{1}n \cdot v^T(\hat{U}+\hat{W})(\hat{U}+\hat{W})v \leq 2\hat{1}n \cdot v^T(\hat{U}\hat{U}^T+\hat{W}\hat{W})v \\ &\leq CK^{-2\zeta_\gamma}[\ln(n)]^2 + O_p(n^{-\Delta+1}),\end{aligned}$$

for any C. It then follows by eq. (7.9) and the triangle inequality that

$$\sqrt{n}[\bar{m}(\hat{\gamma})-\beta_0] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi(z_i) + O_p(\bar{\Delta}_n^* + K^{-\zeta_\gamma}\ln(n)).$$

The first conclusion then follows from the second conclusion of Lemma 1. The second conclusion follows by $\Delta_n^m = C\sqrt{K/n} + K^{-\zeta_\gamma} = O(\bar{\Delta}_n^*)$ when d_K is bounded and $\zeta_m = \zeta_\gamma$. $Q.E.D.$

Proof of Corollary 3: To prove this result it suffices to show that Assumptions 4 and 6 are satisfied in each of the examples with d_K bounded and $\zeta_m = \zeta_\gamma$.

For the conditional covariance $\alpha_0(x) = -E[a_i|x_i = x]$. This being Holder of order s_α is a hypothesis. Also, $v(z) = a \cdot p(x)$, so that

$$E[v(z_i)v(z_i)^T] = E[a_i^2 p(x_i)p(x_i)^T] \leq CE[p(x_i)p(x_i)^T] \leq CI$$

by $E[a_i^2|x_i]$ bounded. Also $\zeta_m = \zeta_\gamma$ by

$$E[\{m(z_i,\gamma_K) - m(z_i,\gamma_0)\}^2] = E[a_i^2\{\gamma_K(x_i) - \gamma_0(x_i)\}^2] \leq CK^{-2\zeta_\gamma}.$$

For the missing data mean $\alpha_0(x) = a/\pi_0(w)$ is Holder of order s_α by $\pi_0(w_i)$ being bounded away from zero and Holder of order s_α. Furthermore $v(z) = q(w)$, so that by Assumption 3,

$$E[v(z_i)v(z_i)^T] = E[q(w_i)q(w_i)^T] \leq CI,$$

and by a_i bounded and $\pi_0(w_i)$ bounded away from zero,

$$\begin{aligned}E[\{m(z_i,\gamma_K) - m(z_i,\gamma_0)\}^2] &= E[\{q(w_i)^T\delta - E[y_i|a_i = 1, w_i]\}^2] \\ &= E[\frac{a_i}{\pi_0(w_i)}\{\gamma_K(x_i) - \gamma_0(x_i)\}^2] \\ &\leq CE[\{\gamma_K(x_i) - \gamma_0(x_i)\}^2] \leq CK^{-2\zeta_\gamma}.\end{aligned}$$

For the average derivative example $\alpha_0(x) = \omega(x)/f_0(x)$ which is Holder of order s_α by each of $\omega(x)$ and $f_0(x)$ being Holder of order s_α and by $f_0(x)$ bounded away from zero where $\omega(x)$ is non zero. Furthermore $v(z) = \int \omega(x)p(x)dx$, so that by Cauchy-Schwartz,

$$\begin{aligned}E[v(z_i)v(z_i)^T] &= \int \omega(x)p(x)dx \int \omega(x)p(x)^T dx \\ &= E[\alpha_0(x_i)p(x_i)]E[\alpha_0(x_i)p(x_i)^T] \\ &\leq E[\alpha_0(x_i)^2]E[p(x_i)p(x_i)^T] \leq CI.\end{aligned}$$

Furthermore,

$$\begin{aligned}E[\{m(z_i,\gamma_K)-m(z_i,\gamma_0)\}^2] &= \{\int \omega(x)[\gamma_K(x)-\gamma_0(x)]dx\}^2 \\ &= E[\alpha_0(x_i)\{\gamma_K(x_i)-\gamma_0(x_i)\}]^2 \\ &\leq E[\alpha_0(x_i)^2]E[\{\gamma_K(x_i)-\gamma_0(x_i)\}^2] = O(K^{-2\zeta_\gamma}).\ Q.E.D.\end{aligned}$$

Proof of Theorem 4: The conclusion follows from Lemma 1 and Theorem 8 of Ichimura and Newey (2017) similarly to the proof of Theorem 2 above, with the conclusion of Theorem 8 of Ichimura and Newey (2017) replacing the argument following eq. (7.9) in the proof of Theorem 2. *Q.E.D.*

Proof of Theorem 5: Let $\hat{\lambda}_\ell(x)$ denote the series regression of $u_i = y - a_i^T\beta_0$ on $p(x_i)$ in the $\hat{I}_\ell$ sample. By a standard formula for instrumental variables estimation and series estimation,

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= \hat{H}^{-1}\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i-\hat{\alpha}_\ell(x_i)]\left\{y_i-\hat{\gamma}_\ell(x_i)-[a_i-\hat{\alpha}_\ell(x_i)]^T\beta_0\right\} \qquad (7.10)\\ &= \hat{H}^{-1}\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i-\hat{\alpha}_\ell(x_i)]\left[u_i-\hat{\lambda}_\ell(x_i)\right]\end{aligned}$$

Assume for the moment that a_i is a scalar and let $y_i = u_i$. Then $\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i-\hat{\alpha}_\ell(x_i)]\left[u_i-\hat{\lambda}_\ell(x_i)\right]/n$ is the doubly robust estimator with $m(z,\gamma)=a[y-\gamma(x)]$, i.e. for the expected conditional covariance. It then follows as in the proof of Corollary 3 that $\max\{\Delta_n^m,\Delta_n^\gamma,\Delta_n^\alpha\}\leq C\bar{\Delta}_n^*$. Then by Lemmas 6 and A5, for $\varphi(z)=[a_i-\alpha_0(x_i)]\varepsilon_i$,

$$\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i-\hat{\alpha}_\ell(x_i)]\left[u_i-\hat{\lambda}_\ell(x_i)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(z_i)+O_p(\bar{\Delta}_n^*)+\sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha.$$

Note that here $\alpha_0(x_i) = -E[a_i|x_i]$ so that

$$\tilde{h}_1^\alpha = \tilde{F}\{v(z)-\alpha_0(x)p(x)\} = -\tilde{F}\{[a-\alpha_0(x)]p(x)\}.$$

Then we have

$$E[\tilde{1}\tilde{\Delta}_1^\alpha\tilde{\Delta}_1^{\alpha T}|X,\tilde{Z}^c] = \tilde{1}\frac{1}{\tilde{n}}\tilde{\Sigma}^{-1}\tilde{F}\{p(x)p(x)^T Var(a_i|x_i=x)\}\tilde{\Sigma}^{-1}\leq\frac{C}{n}I.$$

Therefore it follows by Lemma A2 that

$$E[\hat{1}\tilde{1}(\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha)^2|X,\tilde{Z}^c] = \hat{1}\hat{\Delta}_2^T E[\tilde{1}\tilde{\Delta}_1^\alpha\tilde{\Delta}_1^{\alpha T}|X,\tilde{Z}^c]\hat{\Delta}_2 \leq \hat{1}\frac{C}{n}\hat{\Delta}_2^T\hat{\Delta}_2 = o_p(K/n^2).$$

Then by the Markov inequality

$$\sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha = o_p\left(\sqrt{\frac{K}{n}}\right) = O_p(\bar{\Delta}_n^*).$$

Consequently we have

$$\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i - \hat{\alpha}_\ell(x_i)]\left[u_i - \hat{\lambda}_\ell(x_i)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(z_i) + O_p(\bar{\Delta}_n^*). \tag{7.11}$$

Next, note that

$$\begin{aligned}\bar{F}\{[a-\tilde{\alpha}_\ell(x)][a-\hat{\alpha}_\ell(x)]\} &= \bar{F}\{[a-\alpha_0(x)+\alpha_0(x)-\tilde{\alpha}_\ell(x)][a-\alpha_0(x)+\alpha_0(x)-\hat{\alpha}_\ell(x)]\} \\ &= \bar{F}\{[a-\alpha_0(x)]^2 + [a-\alpha_0(x)][\alpha_0(x)-\tilde{\alpha}_\ell(x)] \\ &+ [a-\alpha_0(x)][\alpha_0(x)-\hat{\alpha}_\ell(x)] + [\alpha_0(x)-\tilde{\alpha}_\ell(x)][\alpha_0(x)-\hat{\alpha}_\ell(x)]\}\end{aligned}$$

Note that by Lemma A3 and $E[a_i^2|x_i]$ bounded,

$$\begin{aligned}\tilde{1}E[(\bar{F}\{[a-\alpha_0(x)][\alpha_0(x)-\tilde{\alpha}_\ell(x)]\})^2|\tilde{Z}] &= \frac{\tilde{1}}{\bar{n}}\int[a-\alpha_0(x)]^2[\alpha_0(x)-\tilde{\alpha}(x)]^2F_0(dz) \\ &\le C\frac{\tilde{1}}{n}\int E[a_i^2|x_i=x][\alpha_0(x)-\tilde{\alpha}(x)]^2F_0(dz) \\ &\le C\frac{\tilde{1}}{n}\int[\alpha_0(x)-\tilde{\alpha}(x)]^2F_0(dz) = O_p\left(\frac{1}{n}\left(\frac{K}{n}+K^{-2\zeta_\gamma}\right)\right).\end{aligned}$$

so that by the Markov inequality it follows that

$$\bar{F}\{[a-\alpha_0(x)][\alpha_0(x)-\tilde{\alpha}_\ell(x)]\} = O_p(\bar{\Delta}_n^*). \tag{7.12}$$

It follows similarly that

$$\bar{F}\{[a-\alpha_0(x)][\alpha_0(x)-\hat{\alpha}_\ell(x)]\} = O_p(\bar{\Delta}_n^*). \tag{7.13}$$

Also, by the Cauchy-Schwartz inequality

$$\hat{1}\tilde{1}\left|\bar{F}\{[\alpha_0(x)-\tilde{\alpha}_\ell(x)][\alpha_0(x)-\hat{\alpha}_\ell(x)]\}\right| \le (\tilde{1}\bar{F}\{[\alpha_0(x)-\tilde{\alpha}_\ell(x)]^2\})^{1/2}(\hat{1}\bar{F}\{[\alpha_0(x)-\hat{\alpha}_\ell(x)]^2\})^{1/2}.$$

Also,

$$E[\tilde{1}\bar{F}\{[\alpha_0(x)-\tilde{\alpha}_\ell(x)]^2\}|\tilde{Z}] = \tilde{1}\int[\tilde{\alpha}_\ell(x)-\alpha_0(x)]^2F_0(dx) = O_p\left(\frac{K}{n}+K^{-2\zeta_\gamma}\right),$$

so that $\tilde{1}\bar{F}\{[\alpha_0(x)-\tilde{\alpha}_\ell(x)]^2\} = O_p(K/n+K^{-2\zeta_\gamma})$. It follows similarly that $\hat{1}\bar{F}\{[\alpha_0(x)-\hat{\alpha}_\ell(x)]^2\} = O_p(K/n+K^{-2\zeta_\gamma})$, so that

$$= \bar{F}\{[\alpha_0(x)-\tilde{\alpha}_\ell(x)][\alpha_0(x)-\hat{\alpha}_\ell(x)]\} = O_p(\bar{\Delta}_n^*) \tag{7.14}$$

Also, note that by $E[\|a_i\|^4] < \infty$,

$$\bar{F}\{[a-\alpha_0(x)]^2\} = E[\{a-\alpha_0(x)\}^2] + O_p\left(\frac{1}{\sqrt{n}}\right) = E[\{a_i-\alpha_0(x_i)\}^2] + O_p(\bar{\Delta}_n^*).$$

It then follows by eqs. (7.12), (7.13), (7.14) and the triangle inequality that

$$\bar{F}\{[a-\tilde{\alpha}_\ell(x)][a-\hat{\alpha}_\ell(x)]\} = E[\{a_i-\alpha_0(x_i)\}^2] + O_p(\bar{\Delta}_n^*).$$

Applying this argument to each element of $\hat{H} = \sum_{\ell=1}^{L}\sum_{i\in I_\ell}[a_i-\tilde{\alpha}_\ell(x_i)][a_i-\hat{\alpha}_\ell(x_i)]^T/n$ and each group of observations I_ℓ and summing up gives $\hat{H} = H + O_p(\bar{\Delta}_n^*)$. It then follows by a standard argument and nonsingularity of H that

$$\hat{H}^{-1} = H^{-1} + O_p(\bar{\Delta}_n^*). \tag{7.15}$$

Finally, it follows from eqs. (7.10), (7.11), (7.15) and from $\sum_{i=1}^{n}\varphi(z_i)/\sqrt{n} = O_p(1)$ that

$$\sqrt{n}(\hat{\beta}-\beta_0) = [H^{-1}+O_p(\bar{\Delta}_n^*)][\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(z_i)+O_p(\bar{\Delta}_n^*)] = H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(z_i)+O_p(\bar{\Delta}_n^*).\ Q.E.D.$$

Proof of Theorem 7: By Lemmas 6 and A5 it suffices to show that $\bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha = O_p(\Delta_n^m)$. Note that

$$\hat{1}\hat{\Delta}_2 = \hat{1}\hat{h}_2 + \hat{1}\hat{U} + \hat{1}\hat{W}.$$

By $E[\hat{h}_2\hat{h}_2^T] \le Cn^{-1}K^{-2\zeta_\gamma}I$ and Lemma A2 iii) we have

$$\begin{aligned} E[\left(\bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{h}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha\right)^2|\hat{Z}^c] &= n\bar{1}\tilde{1}\left(\tilde{\Delta}_1^\alpha\right)^T\bar{\Sigma}E[\hat{1}\hat{h}_2\hat{h}_2^T]\bar{\Sigma}\tilde{\Delta}_1^\alpha \le n\bar{1}\tilde{1}\left(\tilde{\Delta}_1^\alpha\right)^T\bar{\Sigma}E[\hat{h}_2\hat{h}_2^T]\bar{\Sigma}\tilde{\Delta}_1^\alpha \\ &\le CK^{-2\zeta_\gamma}\bar{1}\tilde{1}\left(\tilde{\Delta}_1^\alpha\right)^T\bar{\Sigma}^2\tilde{\Delta}_1^\alpha = O_p\left(\frac{(1+d_K)K^{1-2\zeta_\gamma}}{n}\right) = O_p((\Delta_n^m)^2). \end{aligned}$$

Also, by the first conclusion of Lemma A1 and by Lemma A2 iii),

$$\begin{aligned} E[\left(\bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{U}^T\bar{\Sigma}\tilde{\Delta}_1^\alpha\right)^2|\hat{Z}^c] &= n\bar{1}\tilde{1}\left(\tilde{\Delta}_1^\alpha\right)^T\bar{\Sigma}E[\hat{1}\hat{U}\hat{U}^T]\bar{\Sigma}\tilde{\Delta}_1^\alpha \le n\bar{1}\tilde{1}\left(\tilde{\Delta}_1^\alpha\right)^T\bar{\Sigma}E[\hat{U}\hat{U}^T]\bar{\Sigma}\tilde{\Delta}_1^\alpha \\ &\le CK^{-2\zeta_\gamma}\ln(n)^2\bar{1}\tilde{1}\left(\tilde{\Delta}_1^\alpha\right)^T\bar{\Sigma}^2\tilde{\Delta}_1^\alpha = O_p\left(\frac{(1+d_K)K^{1-2\zeta_\gamma}[\ln(n)]^2}{n}\right) = O_p((\Delta_n^m)^2). \end{aligned}$$

Also by the second conclusion of Lemma A1 and Lemma A2 iii), for $\Delta > 0$ large enough,

$$\bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha = O_p(n^{(1/2)-\Delta}\sqrt{(1+d_K)/n}) = O_p(\Delta_n^m).$$

The conclusion then follows by the Markov and triangle inequalities. *Q.E.D.*

Proof of Theorem 8: By Lemmas 6 and A5 it suffices to show that $\bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha = O_p(\Delta_n^m + \tilde{\Delta}_n)$. Note that

$$\bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{\Delta}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha = T_1 + T_2 + T_3, T_1 = \bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{h}_2^T\bar{\Sigma}\tilde{\Delta}_1^\alpha,$$
$$T_2 = \bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{\Delta}_2^T(I-\hat{\Sigma})\bar{\Sigma}\tilde{h}_1^\alpha,$$
$$T_3 = \bar{1}\hat{1}\tilde{1}\sqrt{n}\hat{\Delta}_2^T(I-\hat{\Sigma})\bar{\Sigma}(I-\tilde{\Sigma})\tilde{\Delta}_1^\alpha.$$

By Lemma A2 iii),

$$E[T_1^2|\hat{Z}^c] \leq \bar{1}\tilde{1}n(\tilde{\Delta}_1^\alpha)^T\bar{\Sigma}E[\hat{h}_2\hat{h}_2^T]\bar{\Sigma}\tilde{\Delta}_1^\alpha \leq CK^{-2\zeta_\gamma}\tilde{1}(\tilde{\Delta}_1^\alpha)^T\tilde{\Delta}_1^\alpha = O_p(K^{-2\zeta_\gamma}(\Delta_n^m)^2),$$

so by the Markov inequality, $T_1 = O_p(\Delta_n^m)$. By Lemma A2 ii),

$$E[T_2^2|\tilde{Z}^c] \leq \hat{1}\sqrt{n}\hat{\Delta}_2^T(I-\hat{\Sigma})E[\tilde{h}_1^\alpha\left(\tilde{h}_1^\alpha\right)^T](I-\hat{\Sigma})\hat{\Delta}_2 \leq Cd_K\hat{\Delta}_2^T(I-\hat{\Sigma})^2\hat{\Delta}_2$$
$$= O_p((1+d_K)\frac{K^{1-2\zeta_\gamma}}{n}\frac{K\ln(K)}{n}).$$

Note that by the Markov inequality and $K\ln(K)/n \longrightarrow 0$ it follows that $T_2 = O_p(\bar{\Delta}_n^* + \Delta_n^m)$. Finally, by the Caucy-Schwartz inequality and Lemma A2,

$$T_3 = O_p(\sqrt{\frac{K^3\ln(K)(1+d_K)}{n^3}}K^{(1/2)-\zeta_\gamma}) = O_p(\tilde{\Delta}_n).$$

The conclusion then follows by the triangle inequality. *Q.E.D.*

Acknowledgements

We appreciate the hospitality of the Cowles Foundation where much of the work for this paper was accomplished. We also appreciate the comments of M. Cattaneo, X. Chen, M. Jansson and seminar participants at UCL.

REFERENCES

Athey, S., G. Imbens, and S. Wager (2017): "Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing," *Journal of the Royal Statistical Society, Series B,* forthcoming.

Ayyagari, R. (2010): Applications of Influence Functions to Semiparametric Regression Models, Ph.D. Thesis, Harvard School of Public Health, Harvard University.

Belloni, A., V. Chernozhukov, D. Chetverikov, K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics* 186, 345–366.

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

Bickel, P. and Y. Ritov (1988): "Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates," *Sankhya: The Indian Journal of Statistics*, *Series A* 50, 381–393.

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.

Cattaneo, M.D., and M. Farrell (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics* 174, 127-143.

Cattaneo, M.D., and M. Jansson (2017): "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica*, forthcoming.

Cattaneo, M.D., M. Jansson, and X. Ma (2017): "Two-step Estimation and Inference with Possibly Many Included Covariates," working paper, Michigan.

Chernozhukov, V., J.C. Escanciano, H. Ichimura, W.K. Newey, J.M. Robins (2016): "Locally Robust Semiparametric Estimation," arXiv 1608.00033.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W.K. Newey, J.M. Robins (2017): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, forthcoming.

Donald, S.G. and W.K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis* 50, 30-40.

Firpo, S. and C. Rothe (2016): "Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions," working paper.

Gine, E. and R. Nickl (2008): "A Simple Adaptive Estimator of the Integrated Square of a Density," *Bernoulli* 14, 47–61.

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

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

Hirschberg, D.A and S. Wager (2017): "Balancing Out Regression Error: Efficient Treatment Effect Estimation without Smooth Propensities," arXiv:1712.00038.

Ichimura, H. and W.K. Newey (2017): "The Influence Function of Semiparametric Estimators," CEMMAP working paper CWP06/17.

Imbens G., J. Angrist, A. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics* 14, 57-67.

KANDASAMY, K., A. KRISHNAMURTHY, B. POCZOS, L. WASSERMAN, J. ROBINS (2015) "Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations," *Advances in Neural Information Processing Systems* 28 (NIPS 2015).

LAURENT, B. (1996): "Efficient Estimation of Integral Functionals of a Density," *Annals of Statistics* 24, 659-681.

MUKHERJEE, R., W.K. NEWEY, J.M. ROBINS (2017): "Semiparametric Efficient Empirical Higher Order Influence Function Estimators," arXiv:1705.07577.

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

NEWEY, W.K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

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.

POWELL, J.L., J.H. STOCK, AND T.M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

ROBINS, J.M. AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statististical Association* 90, 122–129.

ROBINS, J.M., A. ROTNITZKY, AND M. VAN DER LAAN (2000): "Comment on 'On Profile Likelihood' by S. A. Murphy and A. W. van der Vaart," *Journal of the American Statistical Association* 95, 431-435.

ROBINS, J., M. SUED, Q. LEI-GOMEZ, AND A. ROTNITZKY (2007): "Comment: Performance of Double-Robust Estimators When Inverse Probability' Weights Are Highly Variable," *Statistical Science* 22, 544–559.

ROBINS, J.M., E.T. TCHETGEN, L. LI, A. VAN DER VAART (2009): "Semiparametric Minimax Rates," *Electronic Journal of Statistics* 3, 1305–1321.

ROBINS, J.M., L. LI, E. TCHETGEN, AND A. VAN DER VAART (2008) "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," in *IMS Collections Vol. 2, Probability and Statistics: Essays in Honor of David A. Freedman,* D. Nolan and T. Speed (eds.), Beachwood, Ohio: Institute of Mathematical Statistics, 335-421.

ROBINS, J.M, P. ZHANG, R. AYYAGARI, R. LOGAN, E. TCHETGEN, L. LI, T. LUMLEY, A. VAN DER VAART A, HEI HEALTH REVIEW COMMITTEE (2013): "New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research," Research Report Health Eff Instm 175:3-129.

Robins, J.M., L. Li, R. Mukherjee, E. Tchetgen, A. van der Vaart (2017): "Minimax Estimation of a Functional on a Structured High Dimensional Model," *Annals of Statistics,* forthcoming.

Rotnitzky, A. and J.M. Robins (1995): "Semi-parametric Estimation of Models for Means and Covariances in the Presence of Missing Data," *Scandinavian Journal of Statistics* 22, 323–333.

Rudelson, M. (1999): "Random Vectors in the Isotropic Position," *Journal of Functional Analysis* 164, 60-72.

Scharfstein D.O., A. Rotnitzky, and J.M. Robins (1999): Rejoinder to "Adjusting For Nonignorable Drop-out Using Semiparametric Non-response Models," *Journal of the American Statistical Association* 94, 1135-1146.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1482.