Bradic, Jelena; Chernozhukov, Victor; Newey, Whitney K.; Zhu, Yinchu

Working Paper

Minimax semiparametric learning with approximate sparsity

cemmap working paper, No. CWP32/21

cemmap

centre for microdata methods and practice

Minimax semiparametric learning with approximate sparsity

Jelena Bradic
Victor Chernozhukov
Whitney K. Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP32/21

Economic and Social Research Council

Minimax Semiparametric Learning With Approximate Sparsity*

Jelena Bradic[†] *UCSD*

Victor Chernozhukov[‡] *MIT*

Whitney K. Newey[§] *MIT*

Yinchu Zhu[¶] *Brandeis University*

March 2021.

Abstract

This paper is about the ability and means to root-n consistently and efficiently estimate linear, mean-square continuous functionals of a high dimensional, approximately sparse regression. Such objects include a wide variety of interesting parameters such as the covariance between two regression residuals, a coefficient of a partially linear model, an average derivative, and the average treatment effect. We give lower bounds on the convergence rate of estimators of such objects and find that these bounds are substantially larger than in a low dimensional, semiparametric setting. We also give automatic debiased machine learners that are $1/\sqrt{n}$ consistent and asymptotically efficient under minimal conditions. These estimators use no cross-fitting or a special kind of cross-fitting to attain efficiency with faster than $n^{-1/4}$ convergence of the regression. This rate condition is substantially weaker than the product of convergence rates of two functions being faster than $1/\sqrt{n}$, as required for many other debiased machine learners.

Keywords: Approximate sparsity, Lasso, debiased machine learning, linear functional, Riesz representer.

*This research was supported by NSF grants 1712481 and 1757140. J. Robins provided helpful comments.

[†]Department of Mathematics and Halicioglu Data Science Institute, UCSD, La Jolla, CA 92093. U.S.A. E-mail: jbradic@ucsd.edu.

[‡]Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: vchern@mit.edu.

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

[¶]Brandeis University, Waltham, MA 02453, E-mail: yinchuzhu@brandeis.edu.

1 Introduction

This paper is about the ability and means to $1/\sqrt{n}$ consistently and efficiently estimate linear, mean-square continuous functionals of a high dimensional, approximately sparse regression. Such objects include a wide variety of interesting parameters such as the covariance between two regression residuals, a coefficient of a partially linear model, an average derivative, and the average treatment effect. We give lower bounds on the convergence rate of estimators of such objects and find that these bounds are substantially larger than in a low dimensional, semiparametric setting. We also give automatic debiased machine learners that are $1/\sqrt{n}$ consistent and asymptotically efficient under minimal conditions. These estimators use no cross-fitting or a special kind of cross-fitting to attain efficiency with faster than $n^{-1/4}$ convergence of the regression. This rate condition is substantially weaker than the product of convergence rates of two functions being faster than $n^{-1/2}$, as assumed for many other debiased machine learners

High dimensional regressions are potentially important in many applications. There may be many covariates of interest for the covariance between two regression residuals, a coefficient of a partially linear model, an average derivative, or an average treatment effect. This variety of important examples motivates interest in high dimensional regressions. Modern regression models and methods are flexible in ways that are useful with high dimensional regressors. In particular, approximately sparse models specify that the regression function can be well approximated by a linear combination of relatively few important regressors but the identity of the important regressors is unknown. This specification is different than nonparametric models where the identity of important functions in a series approximation is known. Approximate sparsity is also different than sparsity in allowing all coefficients to be nonzero rather than only a few, making approximate sparsity a more plausible assumption in many settings.

There has been a recent growth of the literature on inference of high dimensional models as cited later in the Introduction. A major development has been the use of debiasing methods based on ideas dating back to Neyman orthogonalization for inference based on regularized regression learners. The typical assumptions of these recent methods require rate double robustness, where the product of the regression rate and another rate is faster than $1/\sqrt{n}$. This condition is stronger than needed for classical semiparametric settings. This leads to a natural question: Are we sure that the debiasing approach is the ultimate solution, rather than a specific fix? We shed light on this question by deriving the necessary conditions for $1/\sqrt{n}$ estimation and showing that these conditions are also sufficient for achieving the $1/\sqrt{n}$ rate if we carefully design the debiasing formulation. These results also reveal a fundamental difference between the high dimensional models and the classical semiparametric models.

We find that necessary conditions for $1/\sqrt{n}$ consistent estimation under approximate sparsity are substantially stronger than in low dimensional semiparametric models where the identity of important regressors is known. These necessary conditions depend on the sparse approximation

rates $s^{-\xi_1}$ for the regression and $s^{-\xi_2}$ for another function α_0, where s is the number of important regressors and α_0 is specified in Section 2. The necessary condition is that $\max\{\xi_1, \xi_2\} > 1/2$. In the classic nonparametric setting where the identity of the important regressors is known the corresponding necessary condition for partially linear regression is $\xi_1 + \xi_2 \geq 1/2$; see Ritov and Bickel (1990), Robins et al. (2009), and Section 2. The difference between the approximately sparse condition and the classic nonparametric condition is illustrated in Figure 1. The red line is the boundary for the necessary condition $\xi_1 + \xi_2 \geq 1/2$ for attaining $1/\sqrt{n}$ consistency when the identity of the important regressors is known. The blue box is the boundary for the necessary condition $\max\{\xi_1, \xi_2\} > 1/2$ given here for attaining $1/\sqrt{n}$ consistency under approximate sparsity where the identity of the important regressors is unknown. The requirement for $1/\sqrt{n}$ consistency for the approximately sparse case is stronger, in that the approximation rates must be outside the square, than the requirement for when the identity of the important functions is known, where the rates must be on or above the triangle.

We also give a lower bound for convergence rates when $\tilde{\xi} = \max\{\xi_1, \xi_2\} \leq 1/2$. We find that the lower bound is $(\ln(p)/n)^{2\tilde{\xi}/(2\tilde{\xi}+1)}$ which is illustrated in Figure 2. Here the lower bound is also determined by the maximum of the sparse approximation rates. One can think of the difference of this lower bound with the minimal condition $\xi_1 + \xi_2 \geq 1/2$ for $1/\sqrt{n}$ consistency with known important regressors as a cost of not knowing the important regressors. For example when $\xi_1 = \xi_2 = 1/4$, where $1/\sqrt{n}$ may be possible with known important regressors, the rate lower bound with approximate sparsity is $(\ln(p)/n)^{1/3}$, a smaller power of $(\ln(p)/n)$. This cost is additional to the well known, slightly slower attainable convergence rates of regression learners when the identity of important regressors is unknown, e.g. see Bickel, Ritov, and Tsybakov (2009), Belloni et al. (2012), and Cai and Guo (2017). The attainable mean-square convergence rate for a regression is a power of sample size n when the identity of important regressors is known but is the same power of $\ln(p)/n$ when the identity of important regressors is unknown, where p is the number of potential regressors. If p grows no faster than a power of n this only reduces the rate of convergence from a power of n to the same power of $\ln(p)/n$. In contrast, for $\xi_1 = \xi_2 = 1/4$ the attainable rate of convergence for the parameters we consider drops from $1/\sqrt{n}$ to larger than $n^{-1/3}$ when the identity of the important regressors becomes unknown.

We give estimators of objects of interest that are $1/\sqrt{n}$ consistent and efficient when $\max\{\xi_1, \xi_2\} > 1/2$. These estimators are based on Lasso regression and Lasso minimum distance estimation of the other function α_0. We give an estimator of an average product that is $1/\sqrt{n}$ consistent under minimal conditions. This estimator uses special cross fitting, where the regressor second moments use the same observations as the sample average of regression products. We also give an automatic debiased machine learner (Chernozhukov, Newey, and Singh, 2018) of a linear functional of a regression, where the functional of interest depends only on regressors, that is $1/\sqrt{n}$ consistent when $\xi_1 > 1/2$, a minimal condition. This estimator uses no cross-fitting.

We obtain $1/\sqrt{n}$ consistency by combining special or no cross-fitting with the Lasso first order conditions to ensure that the size of a key second order vanishes faster than $1/\sqrt{n}$.

The condition $\max\{\xi_1, \xi_2\} > 1/2$ is substantially weaker than conditions previously imposed to obtain $1/\sqrt{n}$ consistency of estimators under approximate sparsity. For example $\xi_1 > 1/2$ corresponds to the rate of convergence of Lasso regression being faster than $n^{-1/4}$. Previous conditions require that the product of convergence rates for the learners of the regression and α_0 is faster than $n^{-1/2}$, that is $\xi_1/(2\xi_1 + 1) + \xi_2/(2\xi_2 + 1) > 1/2$ for Lasso. This is the rate double robustness condition of Belloni, Chernozhukov, and Hansen (2014) and Farrell (2015) that is stronger than $\xi_1 > 1/2$. Figure 1 illustrates the difference between these conditions. The hyperbola gives the lower boundary of the set where $\xi_1/(2\xi_1+1)+\xi_2/(2\xi_2+1) > 1/2$. The box is the lower boundary of the set where $\max\{\xi_1, \xi_2\} > 1/2$, as before. The exterior of the box is well inside the points above the hyperbola, especially along each of the axes near the box.

We find that there is a robustness and efficiency trade-off between the estimators of linear functionals with and without cross-fitting. The estimator without cross fitting attains $1/\sqrt{n}$ consistency under just $\xi_1 > 1/2$ but does require that Lasso estimates the conditional expectation of the outcome variable given the regressors. When the Lasso dictionary can approximate any function of the regressors this requirement is not restrictive but otherwise it is. When the conditional expectation cannot be approximated by the Lasso dictionary we show that the estimator without sample splitting is still $1/\sqrt{n}$ consistent under the strong condition that $\xi_2 > 1/2$ in addition to $\xi_1 > 1/2$. The estimator with cross fitting is $1/\sqrt{n}$ consistent under the weaker condition that $\xi_1/(2\xi_1+1)+\xi_2/(2\xi_2+1) > 1/2$ and in this sense is more robust to Lasso not estimating the conditional expectation.

Our results show that orthogonal moment functions with optimal regression learners provide efficient estimators under minimal conditions for approximate sparsity. When the important regressors are known orthogonal moment functions are not needed for efficient estimation but undersmoothing is required, meaning the bias of the regression learner vanishes fast than the variance, e.g. see Newey and Robins (2018). Thus we find that orthogonal moment functions with optimal regression learners seem particularly well suited to approximately sparse models.

The approximately sparse specification we consider is a special case of those of Belloni et al. (2012) and Belloni, Chernozhukov, and Hansen (2014). Some of the estimators we consider use orthogonal moment functions from Chernozhukov et al. (2016) but with special sample splitting. The estimator of α_0 we use is a Lasso minimum distance estimator like Chernozhukov, Newey, and Singh (2018) with special or no cross-fitting.

The debiased machine learning we consider is based on the zero derivative of the estimating equation with respect to each nonparametric component, as in Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), and Robins et al. (2013), and Chernozhukov et al. (2018). This kind of debiasing is different than bias correcting the regression learner, as in Zhang and Zhang

(2014), Belloni Chernozhukov, and Wang (2014), Belloni, Chernozhukov, and Kato (2015), Javanmard and Montanari (2014a,b; 2015), van de Geer et al. (2014), Ren et al. (2015), Bradic and Kolar (2017), and Zhu and Bradic (2018). The functionals we consider are different than those analyzed in Cai and Guo (2017). The continuity properties of functionals we consider provide additional structure that we exploit, namely a Riesz representer, an object that was not considered in Cai and Guo (2017) Targeted maximum likelihood (van der Laan and Rubin, 2006) based on machine learners has been considered by van der Laan and Rose (2011) and large sample theory given by Luedtke and van der Laan (2016), Toth and van der Laan (2016), and Zheng, Luo, and van der Laan (2016).

Mackey, Syrgkanis, and Zadik (2018) showed that weak sparsity conditions would suffice for $1/\sqrt{n}$ consistency of a certain estimator of a partially linear conditional mean when certain variables are independent and non Gaussian. The estimator given there will not be consistent for the objects and model we consider.

Recently Hirshberg and Wager (2020) showed in independent work that their minimax estimator of a regression functional is $1/\sqrt{n}$ consistent for a finite and high dimensional model under weak rate conditions similar to those considered here. The Lasso based estimator we give is relatively simple to compute, is doubly robust, has simple standard errors that are robust to Lasso estimating the conditional mean, and allows for an infinite dimensional model.

In summary, the contributions of this paper are a lower bound on the convergence rate of certain linear, mean-square continuous functionals of an approximately sparse regression and estimators that attain $1/\sqrt{n}$ consistency under minimal conditions with weaker rate requirements than previous estimators.

In Section 2 we describe the objects of interest here. Section 3 describes approximately sparse models. Section 4 gives lower bounds on convergence rates. Estimators that are efficient under minimal conditions are given in Sections 5 and 6.

2 Linear Functionals of a Regression

We consider parameters that depend linearly on a conditional expectation. To describe such an object, let W denote a data observation, and consider a subvector $(Y, X')'$ of W, where Y is a scalar outcome with finite second moment and X is a covariate vector. Denote the conditional expectation of Y given X as

$$\rho_0(x) = E[Y|X = x].$$

Let $m(w, \rho)$ denote a function of the function ρ (i.e. a functional of ρ), where ρ denotes a possible conditional expectation function. The objects of interest here have the form

$$\theta_0 = \mathrm{E}[m(W, \rho_0)]. \tag{2.1}$$

This parameter of interest is an expectation of some known formula $m(W, \rho)$ of a data observation w and a regression function ρ.

We focus on objects where there is $\alpha_0(X)$ with $E[\alpha_0(X)^2] < \infty$ such that

$$\mathrm{E}[m(W, \rho)] = \mathrm{E}[\alpha_0(X)\rho(X)] \text{ for all } \rho(X) \text{ with } E[\rho(X)^2] < \infty. \tag{2.2}$$

By the Riesz representation theorem, existence of such a $\alpha_0(X)$ is equivalent to $E[m(W, \rho)]$ being a mean-square continuous functional of ρ, i.e. there is $C > 0$ such that $E[m(W, \rho)] \leq C \|\rho\|_2$ for all $\rho \in \mathcal{B}$, where $\|\rho\|_2 = \sqrt{E[\rho(X)^2]}$. We will refer to this $\alpha_0(X)$ as the Riesz representer (Rr). Existence of the Rr is equivalent to the semiparametric variance bound for θ_0 being finite, as mentioned in Newey (1994) and shown in Hirshberg and Wager (2020).

There are many important examples of this type of object.

Example 1: (Average Product) A leading example for our results is the average product

$$\theta_0 = E[Z\rho_0(X)].$$

Here $m(W, \rho) = Z\rho(X)$. By iterated expectations for any $\rho(X)$ with $E[\rho(X)^2] < \infty$,

$$E[Z\rho(X)] = E[E[Z|X]\rho(X)] = E[\alpha_0(X)\rho(X)], \ \alpha_0(X) = E[Z|X].$$

The object θ_0 is the part of the covariance between two regression residuals that depends on an unknown function. That covariance is

$$E[\{Z - \alpha_0(X)\}\{Y - \rho_0(X)\}] = E[ZY] - E[Z\rho_0(X)] = E[ZY] - \theta_0,$$

where the first equality follows by orthogonality of $\alpha_0(X)$ and $Y - \rho_0(X)$. This covariance is useful in the analysis of covariance while controlling for nonparametric regression on X. We focus on θ_0 as the part of the covariance between two regression residuals that depends on the regression $\rho_0(X)$.

Example 2: (Weighted Average Derivative). Here $X = (D, Z)$ for a continuously distributed random variable D, $\rho_0(x) = \rho_0(d, x)$, $\omega(d)$ is a pdf, and

$$\theta_0 = E\left[\int \omega(u)\frac{\partial \rho_0(u, Z)}{\partial d}du\right] = E\left[\int S(u)\rho_0(u, Z)\omega(u)du\right],$$

where $S(u) = -\omega(u)^{-1}\partial\omega(u)/\partial u$ is the negative score for the pdf $\omega(u)$ and the second equality follows by integration by parts. Here $m(w, \rho) = \int S(u)\rho(u, z)\omega(u)du$. This θ_0 can be interpreted as an average treatment effect on Y of a continuous treatment D; see Chernozhukov, Newey,

and Singh (2021). Multiplying and dividing by the conditional pdf $f(d|z)$ of $D = d$ given $Z = z$ we find that for any $\rho(X)$ with $E[\rho(X)^2] < \infty$,

$$E[m(W,\rho)] = E[\int S(u)\rho(u,Z)\omega(u)du] = E[f(D|Z)^{-1}S(D)\omega(D)\rho(X)] = E[\alpha_0(X)\rho(X)],$$
$$\alpha_0(X) = f(D|Z)^{-1}S(D)\omega(D).$$

Example 3: (Average Treatment Effect). Here $X = (D, Z)$ and $\rho_0(x) = \rho_0(d, z)$, where $D \in \{0, 1\}$ is the treatment indicator and Z are covariates. The object of interest is

$$\theta_0 = E[\rho_0(1, Z) - \rho_0(0, Z)].$$

If potential outcomes are mean independent of treatment D conditional on covariates Z, then θ_0 is the average treatment effect (Rosenbaum and Rubin, 1983). Here $m(w, \rho) = \rho(1, z) - \rho(0, z)$. Let $\pi_0(Z) = \Pr(D = 1|Z)$ be the propensity score. Note that $E[\rho(1, Z)] = E[\pi(Z)^{-1}D\rho(d, Z)] = E[\pi_0(Z)^{-1}D\rho(X)]$ and similarly $E[\rho(0, Z)] = E[\{1 - \pi_0(Z)\}^{-1}(1 - D)\rho(X)]$ Then for any $\rho \in \mathcal{B}$

$$E[m(W,\rho)] = E[\rho(1,Z) - \rho(0,Z)] = E[\alpha_0(X)\rho(X)],\ \alpha_0(X) = \frac{D}{\pi_0(Z)} - \frac{1-D}{1-\pi_0(Z)}$$

Our results are based on a dictionary of random variables

$$(b_1(X),\ b_2(X),\ ...),\ E[b_j(X)^2] = 1,$$

where X can be infinite dimensional. An important example has $b_j(X) = X_j$ for $X = (X_1, X_2, ...)$. We will assume for most of our results that the conditional expectation $E[Y|X]$ can be well approximated by a linear combination of this dictionary. Let $\mathcal{B}$ denote the set of function with finite second moment that is the closure in mean square of the set of linear combinations of dictionary functions, i.e. the set of functions that can be approximated arbitrarily well in mean square by $\sum_{j=1}^{\infty} \tilde{\gamma}_j b_j(X)$ where only a finite number of $\tilde{\gamma}_j$ are nonzero. We will maintain through much of this paper that

$$\rho_0(X) = E[Y|X] \in \mathcal{B}. \tag{2.3}$$

In this way $E[Y|X]$ is assumed to be an infinite dimensional linear regression. If $\mathcal{B}$ contains all measurable functions with finite second moment then $\rho_0(X) \in \mathcal{B}$ does not impose any restrictions on ρ_0 but otherwise it does.

3 Approximate Sparsity

In this Section we describe the approximate sparsity condition that determines the achievable convergence rate for estimators of θ_0. We also clarify the distinction between approximately

sparse models and those where identity of the important regressors is known and explain the key conditions on which our results are based. For the dictionary of functions $(b_1(x), b_2(x), ...)$ discussed in the previous Section let $b(x) = (b_1(x), ..., b_p(x))'$ denote the $p \times 1$ vector of the first p components. For a scalar random variable $a(X)$ let $\|a\|_2 = \sqrt{E[a(X)^2]}$ denote the mean square norm. For a $p \times 1$ constant vector d let $\|d\|_0$ denote the number of nonzero elements of d. For any constants $C, \xi > 0$, and positive integer $t \in \mathbb{N}$ we define

$$\mathcal{M}_{C,\xi} := \left\{ v \in \mathbb{R}^p : \min_{\|a\|_0 \leq t} \|b(\cdot)'(v - a)\|_2 \leq Ct^{-\xi} \ \forall t \in \mathbb{N} \right\}.$$

In this definition t is the number of nonzero elements of a. This $\mathcal{M}_{C,\xi}$ is the set of $p \times 1$ coefficients v such that $b(X)'v$ can be approximated in mean square by $b(X)'a$, at a rate $t^{-\xi}$, where t is the number of nonzero components of a. The idea is that $b(X)'v$ is the true regression and a are the coefficients of a sparse approximation to $b(X)'v$ with approximation rate $t^{-\xi}$.

Approximately sparse specifications are different than more familiar nonparametric specifications in ways that are useful in high dimensional settings. Approximate sparsity allows for very many potential regressors (possibly many more than sample size) when relatively few important regressors give a good approximation but the identity of those few is not known. In contrast, series approximations are based on relatively few regressors, often many fewer than the sample size. Approximately sparse and series approximations are similar in that they both depend on a few regressors giving a good approximation. They differ in that series regression requires that the identity of the important regressors is known, while with approximate sparsity their identity need not be known. This difference is useful in high dimensional settings, where there are potentially very many regressors needed to approximate a function of many variables. Typically, economics and statistics provide little guidance about which regressors are important. With approximate sparsity, such information is not needed, since very many terms can be included among the potential regressors.

We can be precise about this key difference between approximate sparsity and series approximations by comparing $\mathcal{M}_{C,\xi}$ with a class of functions corresponding to series approximations. For $\mathcal{M}_{C,\xi}$ the nonzero components of a are allowed to be coefficients of any of the p dictionary functions, where p can be large, even larger than sample size. In the definition of $\mathcal{M}_{C,\xi}$ it is unknown which dictionary functions are used in the sparse approximate at rate $Ct^{-\xi}$. A series approximation from the semiparametric literature would require that the unknown function be well approximated by a linear combination of the first t functions. The set of unknown v allowed here would be

$$\mathcal{S}_{C,\xi} = \left\{ v \in \mathbb{R}^p : \min_{\{(a_1,...,a_t,0,...,0\}} \|b(\cdot)'(v - a)\|_2 \leq Ct^{-\xi} \ \forall t \in \mathbb{N} \right\}$$

For example, suppose that X is continuously distributed with compact support and that dictionary functions are products of all nonnegative powers of x that are weakly increasing in order

with j. If x has dimension d and v is such that $b(x)'v$ has bounded derivatives of order s then it is well known that there is C and an ordering of $b(x)$ such that the inequality in the definition of $\mathcal{S}_{C,\xi}$ is satisfied with $\xi = s/d$. This ξ is the well known rate for approximation of functions in a Holder class. Comparing $S_{C,\xi}$ with $\mathcal{M}_{C,\xi}$ we see that the approximately sparse class $\mathcal{M}_{C,\xi}$ extends the notion of a series approximation to allow the best approximating functions to be unknown. Approximate sparsity means there is an $t^{-\xi}$ approximation rate without specifying the order/direction/location of the elements of $b(x)$ that give this rate. Notice that if $\xi \geq \tilde{\xi}$, then $\mathcal{M}_{C,\xi} \subseteq \mathcal{M}_{C,\tilde{\xi}}$, similarly to $\mathcal{S}_{C,\xi}$ shrinking with ξ.

In the rest of the paper, we find a major difference between $\mathcal{M}_{C,\xi}$ and $\mathcal{S}_{C,\xi}$ for estimating the functional $\mathrm{E}[m(W,\rho_0)] = \mathrm{E}[\alpha_0(X)\rho_0(X)]$. Robins et al. (2009) showed that for Holder classes a necessary condition for existence of a $1/\sqrt{n}$ estimator is $\xi_1 + \xi_2 \geq 1/2$ (i.e. the region that is not below the triangle in Figure 1). It can be shown that this requirement is also necessary for existence of a $1/\sqrt{n}$ estimator of the average product for the class $S_{C,\xi}$. For brevity we omit this demonstration from the paper. We demonstrate here that for $(\rho(\cdot), \alpha(\cdot)) \in \mathcal{M}_{C,\xi_1} \times \mathcal{M}_{C,\xi_2}$, the $1/\sqrt{n}$ rate is only possible outside the box in Figure 1, i.e. where $\max\{\xi_1, \xi_2\} > 1/2$.

4 Lower Bound on Convergence Rate

In this Section we give a lower bound on the rate of convergence of learners of $\theta_0 = E[m(W,\rho_0)]$ for the expected product, partially linear regression, and the weighted average derivative. We work with data $\{W_i\}_{i=1}^n$ that is i.i.d. where the distribution of W_i can change with n. This setup is common for results on lower bounds for convergence rates where the lower bound is uniform across a set of data generating processes.

4.1 Average Product

We first consider the average product $\theta_0 = E[Z\rho_0(X)]$. We will derive the bound for the case where X is an infinite dimensional vector and $b_j(X) = X_j$, $(j = 1, 2, ...)$. Let $b(X) = (b_1(X), ..., b_p(X))'$ and for notational simplicity let $X_i = b(X_i)$. We make the following assumption:

Assumption 1: *For each n the data $(Y_i, Z_i, X_i{}')$ is jointly Gaussian with mean zero, $EX_iX_i' = I_p$, and*

$$E[Y_i|X_i] = X_i'\gamma,\ E[Z_i|X_i] = X_i'\pi.$$

The lower bound derived under this condition is a minimax lower bound in any model where Assumption 1 is satisfied as a special case. A lower bound on a convergence rate obtained for a

particular model is a lower bound over any class of models that include the particular model as a special case. In this model the average product for sample size n is

$$\theta = E[Z_i E[Y_i|X_i]] = E[E[Z_i|X_i]E[Y_i|X_i]] = \pi' E[X_i X_i']\gamma = \pi'\gamma,$$

where the second equality follows by iterated expectations. We will consider the supremum of the expected length of confidence intervals as the data generating process varies over a range of possible parameter values for models satisfying Assumption 1, so we do not regard θ as fixed at a value θ_0 or θ_n for any sample size.

Let

$$\Omega = E(Q_i Q_i'),\ Q_i = (Y_i - X_i'\gamma, Z_i - X_i'\pi).$$

For $\xi_1, \xi_2 > 0$ we define the parameter space

$$\Theta_{\xi_1,\xi_2} := \left\{\beta = (\gamma, \pi, \Omega) :\ \gamma \in \mathcal{M}_{C_0,\xi_1}, \pi \in \mathcal{M}_{C_0,\xi_2},\ \text{eigenvalues of } \Omega \text{ belong to } [M^{-1}, M].\right\}$$

where C_0, $M > 0$ are constants. For $\beta = (\gamma, \pi, \Omega)$, we consider the functional $\phi(\beta) = \gamma'\pi$.

Let $\mathcal{C}(\Theta)$ be the set of $1-\alpha$ confidence intervals for $\phi(\beta)$ that are valid uniformly over $\beta \in \Theta$. We are interested in the following

$$\mathcal{L}(\Theta, \tilde{\Theta}) = \inf_{CI \in \mathcal{C}(\tilde{\Theta})} \sup_{\beta \in \Theta} E_\beta |CI|$$

for $\Theta \subseteq \tilde{\Theta}$, where $|CI|$ denotes the length of a confidence interval. If $\mathcal{L}(\Theta, \tilde{\Theta})$ depends on $\tilde{\Theta}$ instead of Θ, then there is no adaptivity between Θ and $\tilde{\Theta}$. If $\Theta = \tilde{\Theta}$, then $\mathcal{L}(\Theta, \Theta)$ is the minimax rate over Θ. The primary goal is to study this object with $\Theta = \Theta_{\xi_1,\xi_2}$ and $\tilde{\Theta} = \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$, where $\xi_1 \geq \tilde{\xi}_1$ and $\xi_2 \geq \tilde{\xi}_2$. This means $\Theta_{\xi_1,\xi_2} \subseteq \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$. We will assume that we are in a high-dimensional setting where $p > n$ for large enough n by imposing the condition that there exists a constant $\kappa > 0$ such that $n \leq p^{1-\kappa} \ln p$ for large enough n.

Theorem 1: *If Assumption 1 is satisfied and there exists a constant $\kappa > 0$ such that $n \leq p^{1-\kappa} \ln p$ for large enough n then for any $\xi_1 \geq \tilde{\xi}_1 \geq 0$, $\xi_2 \geq \tilde{\xi}_2 \geq 0$, and $\tilde{\xi} = \max\{\tilde{\xi}_1, \tilde{\xi}_2\} \leq 1/2$,*

$$\mathcal{L}\left(\Theta_{\xi_1,\xi_2}, \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}\right) \geq D\left(\frac{\ln p}{n}\right)^{2\tilde{\xi}/(2\tilde{\xi}+1)},$$

where $D > 0$ is a constant depending only on $\tilde{\xi}_1$, $\tilde{\xi}_2$, κ, α, C_0, M.

Theorem 1 has two important implications. First, when $\tilde{\xi} = \max\{\tilde{\xi}_1, \tilde{\xi}_2\}$ is much smaller than 1/2, the rate $\mathcal{L}\left(\Theta_{\xi_1,\xi_2}, \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}\right)$ can be much slower than the parametric rate $n^{-1/2}$. Second, Theorem 1 implies that adaptivity to smoothness is not possible. Notice that in Theorem 1, the lower bound for $\mathcal{L}\left(\Theta_{\xi_1,\xi_2}, \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}\right)$ only depends on $\max\{\tilde{\xi}_1, \tilde{\xi}_2\}$ and has nothing to do with

(ξ_1, ξ_2). This means that any confidence interval that is valid over $\Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$ with $\max\{\tilde{\xi}_1, \tilde{\xi}_2\} \leq 1/2$ cannot have expected width $n^{-1/2}$ even at points in a smaller parameter space Θ_{ξ_1,ξ_2}, no matter how small Θ_{ξ_1,ξ_2} is. Hence, there does not exist a confidence interval that satisfies both of the following properties: (1) being valid over $\Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$ with $\max\{\tilde{\xi}_1, \tilde{\xi}_2\} < 1/2$ and (2) having expected width $O(n^{-1/2})$ on a smaller (potentially much smaller) space Θ_{ξ_1,ξ_2}.One implication is that it is not possible to distinguish between $\max\{\xi_1, \xi_2\} \leq 1/2$ and $\max\{\xi_1, \xi_2\} > 1/2$ from the data. Consequently, the condition of $\max\{\xi_1, \xi_2\} > 1/2$ that is necessary in order to obtain the $1/\sqrt{n}$ rate on Θ_{ξ_1,ξ_2} cannot be tested in the data.

4.2 Partial Linear Coefficient and Average Derivative

We now consider the coefficient of a partially linear model under the following condition. The data generating process is i.i.d. with (Y_i, Z_i, X_i') satisfying

Assumption 2: *(Y_i, Z_i, X_i') is jointly Gaussian with mean zero, $E[X_i X_i'] = I_p$,*

$$Y_i = Z_i\theta + X_i'\mu + \varepsilon_i, \;\; Z_i = X_i'\pi + u_i, \tag{4.1}$$

where $E[X_i\varepsilon_i] = E[X_i u_i] = 0$ and $E[Z_i\varepsilon_i] = 0$.

Let $\sigma_u^2 = Eu_i^2$ and $\sigma_\varepsilon^2 = E\varepsilon_i^2$. The distribution of the data is now parameterized by $\lambda = (\theta, \mu, \pi, \sigma_u^2, \sigma_\varepsilon^2)$. Let $C_1, C_2, \xi_1, \xi_2 > 0$, we define the following parameter space

$$\Lambda_{\xi_1,\xi_2} = \left\{\lambda = (\theta, \mu, \pi, \sigma_u^2, \sigma_\varepsilon^2) : \theta \;\in \mathbb{R}, \; \mu \in \mathcal{M}_{C_0,\xi_1}, \; \pi \in \mathcal{M}_{C_0,\xi_2}, \;\; \{\sigma_u^2, \sigma_\varepsilon^2\} \subset [M^{-1}, M]\right\},$$

where $M \geq 2$ is a constant.

We notice that the conditional covariance model can be written in the partial linear form. Assume that (Y_i, Z_i, X_i) has the distribution indexed by $\beta = (\gamma, \pi, \Omega)$ as in Assumption 1. Then by straight-forward algebra, we can see that the model in Assumption 2 can be written in terms of the model in Assumption 1 with $\lambda = f(\beta) = (\theta, \mu, \pi, \sigma_u^2, \sigma_\varepsilon^2)$, where $\theta = \Omega_{1,2}/\Omega_{2,2}$, $\mu = \gamma - \pi\Omega_{1,2}/\Omega_{2,2}$, $\sigma_u^2 = \Omega_{2,2}$ and $\sigma_\varepsilon^2 = \Omega_{1,1} - \Omega_{1,2}^2/\Omega_{2,2}$. It turns out that this relationship allows us to translate the lower bound in Theorem 1 to a lower bound for θ in Assumption 2.

Theorem 2: *If Assumption 2 is satisfied and there exists a constant $\kappa > 0$ such that $n \leq p^{1-\kappa}\ln p$ for large enough n then for any $\xi_1 \geq \tilde{\xi}_1 \geq 0$, $\xi_2 \geq \tilde{\xi}_2 \geq 0$, and $\tilde{\xi} = \max\{\tilde{\xi}_1, \tilde{\xi}_2\} \leq 1/2$,*

$$\mathcal{L}(\Lambda_{\xi_1,\xi_2}, \Lambda_{\tilde{\xi}_1,\tilde{\xi}_2}) \geq D(n^{-1}\ln p)^{2\tilde{\xi}/(2\tilde{\xi}+1)},$$

where $D > 0$ is a constant depending only on $\tilde{\xi}_1$, $\tilde{\xi}_2$, κ, α, C_0.

By Theorem 2 the condition $\tilde{\xi} = \max\{\tilde{\xi}_1, \tilde{\xi}_2\} > 1/2$ is also a necessary condition for attaining the $1/\sqrt{n}$ rate in partial linear models. The same adaptivity discussions apply. We would also like to point out that although $\mathcal{L}(\Lambda_{\xi_1,\xi_2}, \Lambda_{\tilde{\xi}_1,\tilde{\xi}_2})$ and $\mathcal{L}(\Theta_{\xi_1,\xi_2}, \Theta_{\tilde{\xi}_1,\tilde{\xi}_2})$ measure the expected length of confidence intervals, the rates are not due to the possibility of $|CI|$ taking extreme values with a small probability because the object of interest is bounded over the parameter set.

The average derivative is a harder problem than partial linear models and hence the lower bound in Theorem 2 applies to the problem of average derivative. To see this, consider a function of (Z_i, X_i). A special case is when the partial derivative with respect to Z_i is constant. In this special case, the average derivative problem becomes learning a coefficient in a partial linear model. By Theorem 2, even in this special problem, $\max\{\tilde{\xi}_1, \tilde{\xi}_2\} > 1/2$ is a necessary condition for attaching the parametric rate. Therefore, in general, one needs to impose $\max\{\tilde{\xi}_1, \tilde{\xi}_2\} > 1/2$ to obtain the $1/\sqrt{n}$ rate for the average derivative.

An implication of this Section is that when $\max\{\xi_1, \xi_2\} \leq 1/2$ an estimator of θ_0 can converge no faster than $\sqrt{\ln(p)/n}$. In the next two Sections we give estimators that attain $1/\sqrt{n}$ consistency when $\max\{\xi_1, \xi_2\} > 1/2$.

5 Efficient Estimation of the Average Product Via Special Cross-Fitting

The estimators we consider are based on an orthogonal moment function of the form

$$m(W, \rho) - \theta + \alpha(X)[Y - \rho(X)],\ \alpha \in \mathcal{B},\ \rho \in \mathcal{B},$$

as in Chernozhukov et al. (2016) and Chernozhukov, Newey, and Singh (2018). The $\hat{\rho}$ and $\hat{\alpha}$ we use to construct $\hat{\theta}$ have probability limits

$$\bar{\rho}(X) = proj(Y|\mathcal{B}) = proj(\rho_0(X)|\mathcal{B}),\ \bar{\alpha}(X) = proj(\alpha_0(X)|\mathcal{B})$$

respectively. The orthogonal moment functions are doubly robust for the parameter θ_0 in that

$$\theta_0 = E[m(W, \bar{\rho}) + \bar{\alpha}(X)\{Y - \bar{\rho}(X)\}] \text{ if either } E[Y|X] \in \mathcal{B} \text{ or } \alpha_0(X) \in \mathcal{B}.$$

Most of the results of this Section and the next will assume that $E[Y|X] \in \mathcal{B}$. The influence function for the estimators we consider will be

$$\bar{\psi}(w) = m(w, \bar{\rho}) - \bar{\theta} + \bar{\alpha}(x)[y - \bar{\rho}(x)].$$

This is the efficient influence function for the parameter $\bar{\theta} = E[m(W, \bar{\rho})] = E[\bar{\alpha}(X)\bar{\rho}(X)]$ in a model where the data generating process is unrestricted except for regularity conditions, as shown in Chernozhukov, Newey, and Singh (2019). We will give conditions for estimators of θ

to have the influence function $\bar{\psi}(w)$ and hence be asymptotically efficient for $\bar{\theta}$, mostly where $E[Y|X] \in \mathcal{B}$ and hence $\theta_0 = \bar{\theta}$.

In this Section we give an estimator of the average product of Example 1 that is efficient under the minimal condition $\max\{\xi_1, \xi_2\} > 1/2$ when $E[Y|X] \in \mathcal{B}$. This estimator uses special cross-fitting. To describe the estimator we partition the observations into two sets of about equal size I_1 and I_2 and let ℓ index the sets. For a $p \times 1$ vector d let $\|d\|_1 = \sum_{j=1}^p |d_j|$ and let $r > 0$ be a Lasso penalty to be specified later in this Section. The estimator will be constructed using

$$\hat{\rho}_\ell(x) = b(x)'\hat{\gamma}_\ell,\ \hat{\gamma}_\ell = \arg\min_\gamma \gamma'\hat{\Sigma}_\ell\gamma - 2\tilde{\mu}_\ell'\gamma + 2r\|\gamma\|_1,\ (\ell = 1, 2)$$
$$\hat{\alpha}_\ell(x) = b(x)'\hat{\pi}_\ell,\ \hat{\pi}_\ell = \arg\min_\pi \pi'\hat{\Sigma}_\ell\pi - 2\tilde{M}_\ell'\pi + 2r\|\pi\|_1,$$
$$\hat{\Sigma}_\ell = \frac{1}{n_\ell}\sum_{i\in I_\ell} b(X_i)b(X_i)',\ \tilde{\mu}_\ell = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} b(X_i)Y_i,\ \tilde{M}_\ell = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} b(X_i)Z_i,$$

Here $\hat{\rho}_\ell(x)$ and $\hat{\alpha}_\ell(x)$ are L_1 penalized high dimensional regression estimators that differ from conventional Lasso estimates in the dictionary second moment matrix $\hat{\Sigma}_\ell$ and the cross moment matrices $\tilde{\mu}_\ell$ and $\tilde{M}_\ell$ being constructed from different samples. The average product estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{2}\sum_{i\in I_\ell}\{\hat{\rho}_\ell(X_i)Z_i + \hat{\alpha}_\ell(X_i)Y_i - \hat{\alpha}_\ell(X_i)\hat{\rho}_\ell(X_i)\}. \tag{5.1}$$

This estimator has a familiar doubly robust form (e.g. see Robins et al. 2008), but differs from previous estimators in the way cross-ftting is done. Here the sum in equation (5.1) is over the same observations used to form $\hat{\Sigma}_\ell$ whereas previous estimators use different observations for these components. An asymptotic variance estimator for this $\hat{\theta}$ is

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{2}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}^2,\ \hat{\psi}_{i\ell} = \hat{\rho}_\ell(X_i)Z_i + \hat{\alpha}_\ell(X_i)Y_i - \hat{\alpha}_\ell(X_i)\hat{\rho}_\ell(X_i) - \hat{\theta}.$$

An important advantage of the cross-fitting here is realized in the treatment of a key quadratic remainder

$$T_1 = (\hat{\pi}_\ell - \pi)'\hat{\Sigma}_\ell(\hat{\gamma}_\ell - \gamma), \tag{5.2}$$

where γ and π are coefficients of the population least squares regression of Y and Z on $b(X)$ respectively. In previous work this type of remainder is often bounded by the product of convergence rates of $\hat{\rho}_\ell$ and $\hat{\alpha}_\ell$ similarly to Belloni, Chernozhukov, and Hansen (2014) and Farrell (2015). Under approximate sparsity that type of bound on $\hat{R}$ requires that (ξ_1, ξ_2) are located above the hyperbola in Figure 1, where $\xi_1/(2\xi_1+1) + \xi_2/(2\xi_2+1) > 1/2$, in order to guarantee $\sqrt{n}T_1 \xrightarrow{p} 0$. When $\xi_1 > 1/2$ we replace those previous remainder bounds with a smaller one by using the Lasso first order conditions for $\hat{\pi}_\ell$ and a bound on $\|\hat{\gamma}_\ell - \gamma\|_1$ to show that $\sqrt{n}T_1 \xrightarrow{p} 0$

under very weak conditions for $\hat{\pi}_\ell$. We also use the special cross fitting where $\hat{\Sigma}_\ell$ is from a sample independent of $\tilde{\mu}_\ell$ and $\tilde{M}_\ell$ to ensure that another remainder is small. These two features of $\hat{\theta}$ lead to efficiency of $\hat{\theta}$ under the minimal condition $\xi_1 > 1/2$.

Because the estimator $\hat{\theta}$ is the same when we swap the positions of $(\hat{\rho}_\ell(X_i), Y_i)$ and $(\hat{\alpha}_\ell(X_i), Z_i)$ it will also be efficient when $\xi_2 > 1/2$ and very weak conditions are satisfied for $\hat{\gamma}_\ell$. In this way it will follow that $\hat{\theta}$ is efficient under the minimal condition $\max\{\xi_{1.}, \xi_2\} > 1/2$ for $1/\sqrt{n}$ consistent estimation given in Section 4. We focus on the condition $\xi_1 > 1/2$ with the understanding that the estimator will also be efficient when $\xi_2 > 1/2$.

We consider the properties of $\hat{\theta}$ under two possible data generating processes with a corresponding average product parameter for each. For one of the processes the distribution of W can change with n so that

$$E[Y|X] = \rho_n(X) = b(X)'\gamma,\ \mu - \Sigma\gamma = 0,\ \mu = E[b(X)Y],\ \Sigma = E[b(X)b(X)'],$$

where we suppress an n subscript on the expectations for notational convenience. Here we impose that the conditional expectation $E[Y|X]$ is a linear combination of the $p \times 1$ vector $b(X)$ for each n. The true average product for this data generating process is

$$\theta_n = E[Z\rho_n(X)] = E[\alpha_n(X)\rho_n(X)],\ \alpha_n(X) = b(X)'\pi,\ M - \Sigma\pi = 0,\ M = E[b(X)Z].$$

Here $\alpha_n(X)$ is the linear projection of Z on the $p \times 1$ vector $b(X)$. We do not require that $\alpha_n(X)$ is a conditional expectation, unlike $\rho_n(X)$.

This data generating process is like that considered for the lower bound of Section 4 with $E[Y|X]$ being a linear combination of the $p \times 1$ dictionary $b(X)$. Section 4 allowed the average product to range over a set of possible values for each sample size, so the lower bound given there will apply to estimation of θ_n. We show that $\hat{\theta}$ is a $1/\sqrt{n}$ consistent estimator of θ_n under the minimal condition $\xi_1 > 1/2$. This result provides an upper bound for the convergence rate of estimators of θ_n.

The parameter θ_n only depends on finite dimensional regressions. In the second data generating process the parameter of interest depends on an infinite dimensional regression. For this second data generating process W_i will have the same distribution for each n and

$$E[Y|X] = \rho_0(X) \in \mathcal{B}.$$

Here $E[Y|X]$ may depend on an infinite number of the basis functions $(b_1(X), b_2(X), ...)$. The average product for this data generation process is

$$\theta_0 = E[Z\rho_0(X)] = E[\alpha_0(X)\rho_0(X)] = E[\bar{\alpha}(X)\rho_0(X)],\ \bar{\alpha}(X) = proj(Z|\mathcal{B})(X).$$

We do not require that $\bar{\alpha}(X)$ be a conditional expectation. Under $\xi_1 > 1/2$ and additional conditions we show that $\hat{\theta}$ is an asymptotically efficient estimator of θ_0 and $\hat{V}$ is a consistent estimator of asymptotic variance of $\hat{\theta}$.

This data generating process is different than that considered for the lower bound Section 4. Here $E[Y|X]$ is infinite dimensional rather than finite dimensional. We account for nonparametric infinite dimensional $E[Y|X]$ by restricting the bias in approximating $\rho_0(X)$ and $\bar{\alpha}(X)$ by linear combinations of $b(X)$, as further discussed following Assumption 9 below. With this and other conditions the $1/\sqrt{n}$ consistency of $\hat{\theta}$ serves as an upper convergence rate bound for estimating θ_0 over a wider range of data generating processes than considered in Section 4. The asymptotic efficiency of $\hat{\theta}$ and consistency of $\hat{V}$ are also of interest for asymptotic inference for θ_0.

To specify conditions for $\hat{\theta}$ we begin with an approximate sparsity condition for the regression of Y on $b(X)$. For a $p \times 1$ vector d let $\|d\|_2 = \sqrt{\sum_{j=1}^p d_j^2}$. For any nonincreasing function $f : \mathbb{N} \mapsto [0, \infty)$, we define

$$\mathcal{M}_f := \left\{ v \in \mathbb{R}^p : \min_{\|a\|_0 \leq t} \|v - a\|_2 \leq f(t) \; \forall t \in \mathbb{N} \right\}.$$

We impose the following approximate sparsity condition on the least squares coefficients γ that satisfy $\mu - \Sigma\gamma = 0$.

Assumption 3: There is $C > 0$ and $\xi_1 > 1/2$ such that for all n large enough $\gamma \in \mathcal{M}_f$ *for* $f(t) = Ct^{-\xi_1}$ and $t \leq C(\ln(p)/n)^{-2/(2\xi_1+1)}$.

When the maximum eigenvalue of $\Sigma = E[b(X)b(X)']$ is bounded, which we will assume, this condition strengthens the sparse approximation rate condition of Section 4 to apply to the vector of population least squares coefficients γ rather the mean-square projection $b(X)'\gamma$. Assumption 3 is equivalent to the sparse approximation rate of Section 4 when the smallest eigenvalue of Σ is bounded away from zero. Since $\Sigma = I$ for the model from which the lower bound is constructed, the lower bound continues to hold under this Assumption 3. Therefore, a result showing $1/\sqrt{n}$ consistency under Assumption 3 is sharp in that the assumed sparsity condition is no stronger than the minimal condition $\max\{\xi_1, \xi_2\} > 1/2$.

We will impose weak conditions on $\alpha_n(X)$ and $\bar{\alpha}(X)$. In particular we do not require a sparse approximation rate for $\alpha_n(X)$ or $\bar{\alpha}(X)$. In many settings, such as Example 3, conditions on $\bar{\alpha}(X)$ embody common support restrictions for treatment effects. Imposing weak conditions on $\bar{\alpha}(X)$ allows weak common support conditions. For some of our results we will only require that $E[\alpha_0(X)^2] < \infty$, which is a minimal condition for $1/\sqrt{n}$ consistent estimation.

We do require that $\bar{\alpha}(X)$ or $\alpha_n(X)$ has a sparse mean-square approximation but we do not require any rate. When the distribution W does not vary with n a sparse approximation always exists by the definition of $\bar{\alpha}$ as an element of $\mathcal{B}$. When the distribution of W varies with n we impose the following sparse approximation existence condition.

Assumption 4: *There is* $C > 0$, $\delta_n \longrightarrow 0$, and $\tilde{\pi}$ with $\|\tilde{\pi}\|_0 = O(\delta_n^2 n/\ln(p))$ such that $(\pi - \tilde{\pi})'\Sigma(\pi - \tilde{\pi}) = O(\delta_n^2)$.

This condition is weaker than the sparse approximation rate assumed for the bounds in Section 4. Consequently imposing it does not affect the sharpness of our convergence rates for $\hat{\theta}$. Also, this Assumption is automatically satisfied when the distribution of W does not vary with n, as demonstrated in the following result.

Lemma 3: *If the distribution of* W_i *does not vary with* n *then Assumption 4 is satisfied.*

This is a simple result that follows from the definition of $\mathcal{B}$ as the set of functions that can be approximated arbitrarily well by a finite linear combination of $(b_1(X), b_2(X), ...)$. The convergence $\delta_n \longrightarrow 0$ follows by that condition. The rate at which $\delta_n \longrightarrow 0$ is not restricted in any way.

The following condition imposes restrictions on the Lasso penalty r. Let $\varepsilon_n = \sqrt{\ln(p)/n}$.

Assumption 5: $\varepsilon_n = o(r)$, $r = o(\varepsilon_n \delta_n^{-1})$, *and* $r = o(n^c \varepsilon_n)$ *for every* $c > 0$.

Here we allow r to be slightly larger than ε_n which simplifies the statements of the results without affecting their sharpness. The penalty size r is restricted to go to zero faster than $\varepsilon_n \delta_n^{-1}$. One can guarantee that such a bound is satisfied in a wide variety of cases by choosing r to be proportional to ε_n times several compositions of $\ln(n)$ (*e.g.* $r \approx \ln(\ln(\ln(n)))\varepsilon_n$). This assumption could be avoided by specifying that $r \approx C\varepsilon_n$ for a large enough C and that all results hold with large probability. We impose Assumption 5 for simplicity.

The next condition imposes that the elements of the dictionary $b(X)$ are uniformly bounded. For a $p \times 1$ vector d let $\|d\|_\infty = \max_{j \leq p} |d_j|$, $\|d\|_1 = \sum_{j=1}^p |d_j|$, and $\varepsilon_n = \sqrt{\ln(p)/n}$.

Assumption 6: *There is* $C > 0$ *such that for all* n, $\|b(X)\|_\infty \leq C$, *and the largest eigenvalue of* $\Sigma = E[b(X)b(X)']$ *is bounded uniformly in* p *and* n.

This condition simplifies the analysis considerably. It could be relaxed to allow for sub Gaussian regressors as in the lower bounds of Section 4, although that seems to require a different asymptotic variance estimator where $\hat{\alpha}(X)$ is trimmed. We will maintain Assumption 6 for simplicity.

We will also make use of a sparse eigenvalue condition as in much of the Lasso literature. Let J denote a subvector of $(1, ..., p)$, γ_J be the vector consisting of $\gamma_{Jj} = \gamma_j$ for $j \in J$ and $\gamma_{Jj} = 0$ otherwise, and γ_{J^c} be the corresponding vector for J^c (e.g. so that $\gamma = \gamma_J + \gamma_{J^c}$).

Assumption 7: *There are c, $C > 0$ such that with probability approaching one for all $s = O(n/\ln(p))$,*

$$\min_{|J|\leq s} \min_{\|\gamma_{J^c}\|_1 \leq 3\|\gamma_J\|_1} \frac{\gamma'\hat{\Sigma}\gamma}{\gamma_J'\gamma_J} \geq c.$$

We first consider the case where the distribution of W_i can change with n and show that the estimator $\hat{\theta}$ is $1/\sqrt{n}$ consistent for $\theta_n = E[Z\rho_n(X)]$. The next condition specifies that $\rho_n(X)$ is a conditional expectation

Assumption 8: *$E[Y|X] = \rho_n(X) = b(X)'\gamma$ for γ satisfying $\mu = \Sigma\gamma$ and there is $C > 0$ such that $Var(Y|X) \leq C$.*

This condition was imposed in Section 4 so the lower bound from there applies here. In this case the parameter of interest is

$$\theta_n = E[ZE[Y|X]] = E[Zb(X)']\gamma = M'\gamma = \pi'\Sigma\gamma.$$

Let

$$\rho_n(x) = b(x)'\gamma,\ \alpha_n(x) = b(x)'\pi,\ \psi_n(w) = \rho_n(x)z + \alpha_n(x)y - \alpha_n(x)\rho_n(x) - \theta_n.$$

Theorem 4: *If Assumptions 3-8 are satisfied and there is $C > 0$ with $E[Z^2|X] \leq C$, $E[\rho_n(X)^2] \leq C$, $E[\alpha_n(X)^2] \leq C$ then*

$$\hat{\theta} = \theta_n + \frac{1}{n}\sum_{i=1}^{n} \psi_n(W_i) + o_p(n^{-1/2}) = \theta_n + O_p(n^{-1/2}).$$

For $\xi_2 > 1/2$ this conclusion also is satisfied with Z, π, and M interchanged with Y, γ, and μ in Assumptions 3–5 and 8.

Theorem 4 shows that $\hat{\theta}$ is a $1/\sqrt{n}$ consistent estimator of θ_n under the minimal approximate sparsity condition $\max\{\xi_1, \xi_2\} > 1/2$ and a few additional regularity conditions, thus providing a sharp upper bound on the convergence rate of estimators of θ_n.

Next we consider the second data generating process where the distribution of W_i does not vary with n and the parameter of interest is $\theta_0 = E[Z\rho_0(X)]$. We impose the following additional condition

Assumption 9: i) $E[Y|X] \in \mathcal{B}$; ii) $\sqrt{n}\,\|\rho_0 - b'\gamma\|_2\,[\|\bar{\alpha} - b'\pi\|_2 + \varepsilon_n^{-1} r\delta_n] \longrightarrow 0$;

The condition that $\sqrt{n}\,\|\rho_0 - b'\gamma\|_2\,\|\bar{\alpha} - b'\pi\|_2 \longrightarrow 0$ is a kind of "tail bias" condition that requires that the product of approximation errors from the least squares regression of $\rho_0(X)$ and $\bar{\alpha}(X)$ respectively on $b(X)$ goes to zero faster than $1/\sqrt{n}$. This condition is different than

analogous tail bias conditions in the classic semiparametric setting. The condition here concerns the bias from using all the potential regressors p and p is allowed to be much bigger than n. In semiparametric settings the analogous product bias condition only uses s elements of b with $s < n$. The condition here is much weaker than in the standard semiparametric setting. For example suppose that $\|\rho_0 - b'\gamma\|_2 = o(p^{-c})$ for some $c > 0$. Then by choosing $p = n^{1/2c}$ we would have $\sqrt{n}\,\|\rho_0 - b'\gamma\|_2 = \sqrt{n}o(n^{-1/2}) \longrightarrow 0$ and Assumption 8 ii) would be satisfied. Assumption 8 ii) is more general in allowing $\|\rho_0 - b'\gamma\|_2$ to shrink slower than $n^{-1/2}$ when $\|\bar{\alpha} - b'\pi\|_2$ and $\varepsilon_n^{-1} r\delta_n$ shrink fast enough.

We impose an additional condition for consistency of the asymptotic variance estimator $\hat{V}$.

Assumption 10: i) $\rho_0(X) = \sum_{j=1}^{\infty} \gamma_{j0} b_j(X)$ *with* $\sum_{j=1}^{\infty} |\gamma_{j0}| < \infty$; *ii) there is a constant* C *such that for all positive integers* $t \leq C(\ln(p)/n)^{-2/(2\xi_1+1)}$ *there is* J *such that* $\tilde{\gamma}_0 = (\gamma_{10}, ..., \gamma_{p0})'$ *satisfies* $\|\gamma - \tilde{\gamma}_{0J}\|_2 \leq Ct^{-\xi_1}$ *and* $\|\tilde{\gamma}_{0J^c}\|_1 \longrightarrow 0$.

Let

$$\psi_0(w) = z\rho_0(x) - \theta_0 + \bar{\alpha}(x)[y - \rho_0(x)].$$

Theorem 5: *If Assumptions 3,5–7, and 9 are satisfied and* $E[Z^2|X]$ *and* $Var(Y|X)$ *are bounded then*

$$\hat{\theta} = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} \psi_0(W_i) + o_p(n^{-1/2}),\ \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),\ V = E[\psi_0(W)^2].$$

If Assumption 10 is also satisfied then $\hat{V} \xrightarrow{p} V$. *These conclusions also hold under the same conditions with* Z *and* Y *interchanged.*

Theorem 5 shows that efficient estimation of the average product θ_0 is possible under the approximate sparsity condition $\max\{\xi_1, \xi_2\} > 1/2$ under the conditions given here. It is interesting to compare this result with the efficient estimation results for the average product when identity of important regressors is known. Newey and Robins (2018) gave such results when the elements of $b(x)$ are splines and $\rho_0(X)$ and $\alpha_0(X)$ are elements of Holder classes of functions with sparse approximation rates of ξ_1 and ξ_2 respectively implied by the Holder assumptions. There it was found that an estimator like $\hat{\theta}$ but with a different type of cross fitting was asymptotically efficient when $\xi_1 + \xi_2 > 1/2$. With approximate sparsity as allowed by Theorem 5 the condition for efficiency is $\max\{\xi_1, \xi_2\} > 1/2$. This difference in minimal conditions for efficiency is expected from the lower bound in Section 4.

It is also interesting to compare the respective estimators. The estimator here is based on Lasso which results in $\hat{\rho}_\ell$ and $\hat{\alpha}_\ell$ estimating ρ_0 and $\bar{\alpha}$ at optimal respective rates. In contrast the efficient estimator in Newey and Robins (2018) uses a spline estimator of ρ_0 and specifies that

the number of series terms grows faster than is optimal for estimating the function. The results here show that such undersmoothing is not needed for asymptotic efficiency in approximately sparse models.

6 Efficient Estimation of Functionals Without Cross-fitting

In this Section we give an estimator for general linear functionals of the type considered in Section 2. The estimator is debiased machine learning with a Lasso regression and estimator of $\bar{\alpha}(X)$ without cross-fitting. We show that the estimator is $1/\sqrt{n}$ consistent and asymptotically efficient under $\xi_1 > 1/2$, one of the minimal conditions of Section 4, when the functional of interest depends just on X. These results depend on $E[Y|X]$ being a linear combination of the dictionary $(b_1(X), b_2(X), ...)$. We also show that this estimator is doubly robust and that when $\xi_2 > 1/2$ in addition to $\xi_1 > 1/2$ it is $1/\sqrt{n}$ consistent when $E[Y|X]$ is not a linear combination of $(b_1(X), b_2(X), ...)$ and/or the functional of interest depends on W and not just X.

The estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\{m(W_i, \hat{\rho}) + \hat{\alpha}(X_i)[Y_i - \hat{\rho}(X_i)]\}, \tag{6.1}$$

$$\hat{\rho}(x) = b(x)'\hat{\gamma},\ \hat{\gamma} = \arg\min_{\gamma} \gamma'\hat{\Sigma}\gamma - 2\hat{\mu}'\gamma + 2r\left\|\gamma\right\|_1,$$

$$\hat{\alpha}(x) = b(x)'\hat{\pi},\ \hat{\pi} = \arg\min_{\pi} \pi'\hat{\Sigma}\pi - 2\hat{M}'\pi + 2r\left\|\pi\right\|_1,$$

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n} m(W_i, b),\ \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} b(X_i)Y_i,\ \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} b(X_i)b(X_i)'.$$

This $\hat{\rho}(X)$ is Lasso regression and the $\hat{\alpha}(X)$ is the Lasso minimum distance learner of $\bar{\alpha}(X)$ given in Chernozhukov, Newey, and Singh (2018). An asymptotic variance estimator for this $\hat{\theta}$ is

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i^2,\ \hat{\psi}_i = m(W_i, \hat{\rho}) + \hat{\alpha}(X_i)[Y_i - \hat{\rho}(X_i)] - \hat{\theta}.$$

The estimators $\hat{\theta}$ and $\hat{V}$ are like the automatic debiased machine learning estimator in Chernozhukov, Newey, and Singh (2018) with Lasso $\hat{\rho}$. They do not use cross-fitting in that $\hat{\theta}$ uses the same observations in the average as are used to construct $\hat{\rho}(X)$ and $\hat{\alpha}(X)$.

This absence of cross fitting makes this estimator computationally convenient because $\hat{\gamma}$ and $\hat{\alpha}$ only need to be computed once rather than the multiple times required for cross-fitting. It would be interesting to explore finite sample consequences of not using cross-fitting though that is beyond the scope of this version of the paper.

The absence of cross-fitting has the same important role here as does the special cross fitting in Section 5 in helping to control the size of the remainder T_1 in equation (5.2). Also dependence

of $m(X,\gamma)$ on only X will mean that $\hat{\pi}$ will depend only on X which helps control an additional remainder.

The first additional condition we impose in this Section is:

Assumption 11: *There is $\bar{\alpha}(X) \in \mathcal{B}$ such that $E[m(W,\rho)] = E[\alpha_0(X)\rho(X)]$ for all $\rho \in \mathcal{B}$ and there is $C > 0$ such that $\max_{j \leq p} |m(X, b_j)| \leq C$ for all p.*

Similarly to Section 5 we will first show that the estimator $\hat{\theta}$ is $1/\sqrt{n}$ consistent for $\theta_n = E[m(W,\rho_n)]$ in the first data generating process with $E[Y|X] = \rho_n(X)$ where the distribution of W can change with n. Here let $M = E[m(X,b)]$, π be a $p \times 1$ vector of coefficients satisfying $M - \Sigma\pi = 0$, $\theta_n = E[m(W,\rho_n)] = M'\gamma$, $\rho_n(x) = b(x)'\gamma$, $\alpha_n(x) = b(x)'\pi$, and

$$\psi_n(w) = m(w,\rho_n) - \theta_n + \alpha_n(x)[y - \rho_n(x)].$$

Theorem 6: *If $m(W,\rho)$* depends only on X, Assumptions *3-8 and 11 are satisfied and there is $C > 0$ with $E[m(X,\rho_n)^2] \leq C$* and *$E[\alpha_n(X)^2] \leq C$ then*

$$\hat{\theta} = \theta_n + \frac{1}{n}\sum_{i=1}^{n} \psi_n(W_i) + o_p(n^{-1/2}) = \theta_n + O_p(n^{-1/2}).$$

This result shows $1/\sqrt{n}$ consistency of $\hat{\theta}$ for the parameter θ_n for the first data generating process that changes with n. We also show asymptotic efficiency with the second data generating process where parameter of interest is $\theta_0 = E[m(W,\rho_0)] = E[\alpha_0(X)\rho_0(X)]$.

Assumption 12: *Either i) $|m(W,\rho)| \leq a(W)\sup_x |\rho(x)|$, $E[a(W)^2] < \infty$ and Assumption 10 is satisfied or ii) $E[m(W,\rho)^2] \leq CE[\rho(X)^2]$ for all $\rho \in B$ and $\bar{\alpha}(X)$ is bounded.*

Part i) of this condition does not restrict $\bar{\alpha}(X)$ while part ii) requires that $\bar{\alpha}(X)$ is bounded. Let

$$\psi_0(w) = m(w,\rho_0) - \theta_0 + \bar{\alpha}(x)[y - \rho_0(x)].$$

Theorem 7: *If $m(W,\rho)$ depends only on X and Assumptions 3, 5–7, 9, 11, and 12 are satisfied then*

$$\hat{\theta} = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} \psi_0(W_i) + o_p(n^{-1/2}),\ \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,V),\ V = E[\psi_0(W)^2].$$

If Assumption 10 is also satisfied then $\hat{V} \xrightarrow{p} V$.

The same conclusion was obtained by Chernozhukov, Newey, and Singh (2018) under stronger approximate sparsity conditions. Here the only approximate sparsity condition assumed is

$\xi_1 > 1/2$, i.e. that (ξ_1, ξ_2) is to the right of the box in Figure 1. Chernozhukov, Newey, and Singh (2018) require that (ξ_1, ξ_2) is above the hyperbola in Figure 1. When (ξ_1, ξ_2) is below the hyperbola and to the right of the box the estimator here will be asymptotically efficient whereas the one in Chernozhukov, Newey, and Singh (2018) is not known to be. Also Theorem 7 requires no approximately sparse approximation rate for $\bar{\alpha}(X)$ which is required in Chernozhukov, Newey, and Singh (2018). This feature of Theorem 7 will be useful when approximate sparsity for $\bar{\alpha}(X)$ is deemed a strong condition.

Theorem 7 does assume that $\rho_0(X) = E[Y|X]$, i.e. that the regression function is a linear combination of the dictionary $(b_1(X), b_2(X), ...)$, which is not assumed in Chernozhukov Newey and Singh (2018). This condition is not restrictive when $(b_1(X), b_2(X), ...)$ can approximate any function of X in mean square but otherwise is restrictive. We do not know whether it is possible to attain asymptotic efficiency without $\rho_0(X) = E[Y|X]$ when the only approximate sparsity condition is $\xi_1 > 1/2$ but this is a topic of ongoing research. Theorem 7 is also specific to Lasso regression while Chernozhukov, Newey, and Singh (2018) applies to any regression learner that converges at a power of the sample size.

Theorem 7 applies to many interesting objects of interest including the weighted average derivative of Example 2 and the average treatment effect of Example 3. To illustrate Theorem 7 we give results for these two examples. These and the rest of the following results are given for the 2nd data generating process where the object of interest is $\theta_0 = E[m(W, \rho_0)]$. For simplicity we only give results where Assumption 10 is satisfied.

Here is an asymptotic efficiency result for the weighted average derivative of Example 2:

Corollary 8 (Example 2): *If Assumptions 3, 5–7, 9, and 10 are satisfied, there is $C > 0$ such that $|S(u)| \leq C$ for all $u \in \mathbb{R}$ and $E[f_{D|Z}(D|Z)^{-1}\omega(D)] < \infty$ then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ and $\hat{V} \xrightarrow{p} V$.*

This result imposes weaker conditions on approximate sparsity and on $\bar{\alpha}(X)$ than imposed in Chernozhukov, Newey, and Singh (2020). It imposes a stronger condition in requiring that $\rho_0(X) = E[Y|X]$.

Here is an asymptotic efficiency result for the average treatment effect of Example 3.

Corollary 9 (Example 3): *If Assumptions 3, 5–7, 9, and 10 are satisfied and $E[\Pr(D = 1|Z)^{-1}\{1 - \Pr(D = 1|Z)\}^{-1}] < \infty$ then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ and $\hat{V} \xrightarrow{p} V$.*

This result is notable in imposing no restrictions on the propensity score $\Pr(D = 1|Z)$ other than the minimal condition for existence of a $1/\sqrt{n}$ consistent estimator and in having weaker approximate sparsity conditions than in Chernozhukov, Newey, and Singh (2020). It imposes a stronger condition in requiring that $\rho_0(X) = E[Y|X]$.

An important property of the estimator $\hat{\theta}$ is that it is doubly robust, meaning that it may be consistent for an object of interest even when $E[Y|X]$ is not in $\mathcal{B}$. To explain let $\bar{\rho}(X) = E[Y|X]$ and suppose that the object of interest is

$$\bar{\theta} = E[m(X, \bar{\rho})].$$

Let $\bar{\alpha}(X) = proj(\alpha_0|\mathcal{B})(X)$. Then we have

THEOREM 10: *If Assumptions 3, 5–7, 9, 11, and 12 are satisfied then* $\hat{\theta} \xrightarrow{p} \theta_0 = E[\alpha_0(X)\rho_0(X)]$ and $\bar{\theta} = \theta_0$ if either $E[Y|X] \in \mathcal{B}$ or $\alpha_0(X) \in \mathcal{B}$.

This result shows that the probability limit $\bar{\theta}$ of $\hat{\theta}$ will be equal to the parameter of interest θ_0 if either $E[Y|X]$ is in $\mathcal{B}$ or if $\alpha_0(X)$ is in $\mathcal{B}$. For instance in the average treatment effect of Example 3 the probability limit of $\hat{\theta}$ is the average treatment effect if either the projection of Y on $\mathcal{B}$ is $E[Y|X]$ or if $\alpha_0(X) = D/\Pr(D = 1|Z) - (1 - D)/[1 - \Pr(D = 1|Z)]$ is a linear combination of the dictionary $(b_1(X), b_2(X), ...)$.

If population least square coefficients π satisfying $M - \Sigma\pi = 0$ have a $\xi_2 > 1/2$ sparse approximation rate, in addition to $\xi_1 > 1/2$ and the other conditions previously imposed, then $\hat{\theta}$ will be an asymptotically efficient estimator for θ_0 and $\hat{V}$ a consistent estimator of its asymptotic variance without $\rho_0(X) = E[Y|X]$ and without $m(W, \rho)$ depending only on X.

THEOREM 11: *If Assumptions 3, 5–7, 11, and 12 ii) are satisfied, Assumption 3 is satisfied with* π *substituted for* γ *and* $\xi_2 > 1/2$, *and* $\sqrt{n}\,\|\rho_n - \bar{\rho}\|\,\|\alpha_n - \bar{\alpha}\| \longrightarrow 0$ *then for* $\bar{\theta} = E[m(W, \bar{\rho})]$,

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

Thus we see that if there is a sparse approximation rate $\xi_2 > 1/2$ for π and other conditions previously discussed are satisfied then $\hat{\theta}$ will be consistent and asymptotically normal estimator of $\bar{\theta}$ and $\hat{V}$ will be a consistent estimator of its asymptotic variance. This result does not require that $E[Y|X] \in \mathcal{B}$ nor does it require that $m(W, \rho)$ depends only on X.

In summary, Theorem 7 showed that the automated debiased machine learner without sample splitting $\hat{\theta}$ is asymptotically efficient under a minimal condition $\xi_1 > 1/2$. Theorem 10 showed that this $\hat{\theta}$ is doubly robust, providing a consistent estimator of an object that may be equal to a parameter of interest when $\rho_0(X) \neq E[Y|X]$. Theorem 11 rounds out these properties by showing that $\hat{\theta}$ is consistent and asymptotically normal for the parameter $\bar{\theta}$ and the variance estimator $\hat{V}$ is consistent when π is also approximately sparse with $\xi_2 > 1/2$ when $\rho_0(X) \neq E[Y|X]$ and/or when $m(W, \rho)$ does not depend on just X. Thus we find that $\hat{\theta}$ and $\hat{V}$ have several attractive properties. They are simpler to compute than the cross-fit version. The $\hat{\theta}$ is asymptotically efficient under a minimal sparsity condition when $E[Y|X] \in \mathcal{B}$, is doubly robust, and inference

based on $\hat{\theta}$ and $\hat{V}$ is asymptotically correct even when $\rho_0(X) \neq E[Y|X]$ if π has $\xi_2 > 1/2$. That is, $\hat{\theta}$ is asymptotically efficient under a minimal condition, is doubly robust, and $\hat{\theta}$ and $\hat{V}$ provide specification robust large sample inference when $\xi_2 > 1/2$.

References

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica* 80, 2369-2429.

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

Belloni, A., V. Chernozhukov, and K. Kato (2015): "Uniform Post Selection Inference for Least Absolute Deviation Regression and Other Z-Estimation Problems," *Biometrika*, 102: 77–94. ArXiv, 2013.

Belloni, A., V. Chernozhukov, L. Wang (2014): "Pivotal Estimation via Square-Root Lasso in Nonparametric Regression," *Annals of Statistics* 42, 757–788.

Bickel, P.J., Y.Ritov, and A.Tsybakov (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics* 37, 1705–1732.

Bradic, J. and M. Kolar (2017): "Uniform Inference for High-Dimensional Quantile Regression: Linear Functionals and Regression Rank Scores," *arXiv preprint arXiv:1702.06209*.

Cai, T.T. and Z. Guo (2017): "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *Annals of Statistics* 45, 615-646.

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

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

Chernozhukov, V., W.K. Newey, and R. Singh (2018): "Learning L_2 Continuous Regression Functionals Via Regularized Riesz Representers," arXiv.

Chernozhukov, V., W.K. Newey, and Singh (2019): "De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," arXiv.

Chernozhukov, V., W.K. Newey, and Singh (2021): "Automatic Debiased Machine Learning of Causal and Structural Effects," https://arxiv.org/abs/1809.05224.

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

Hirshberg, D.A. and S. Wager (2020): "Augmented Minimax Linear Estimation," https://arxiv.org/pdf/1712.00038.pdf.

Hoeffding, W. (1963): "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association* 58, 13-30.

Javanmard, A. and A. Montanari (2014a): "Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory," *IEEE Transactions on Information Theory* 60, 6522–6554.

Javanmard, A. and A. Montanari (2014b): "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research* 15: 2869–2909.

Javanmard, A. and A. Montanari (2015): "De-Biasing the Lasso: Optimal Sample Size for Gaussian Designs," arXiv preprint arXiv:1508.02757.

Luedtke, A. R. and M. J. van der Laan (2016): "Optimal Individualized Treatments in Resource-limited Settings," *The International Journal of Biostatistics* 12, 283-303.

Mackey, L. V. Syrgkanis, and I. Zadik (2018): "Orthogonal Machine Learning: Power and Limitations," *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80.

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

Newey, W.K. and J.M. Robins (2018): "Cross Fitting and Fast Remainder Rates for Semiparametric Estimation," arxiv.

Ren, Z., T. Sun, C.H. Zhang, and H. Zhou (2015): "Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Models," *Annals of Statistics* 43, 991–1026.

Ritov, Y. and P. J. Bickel (1990), "Achieving Information Bounds in Non and Semiparametric Models," Annals of Statistics 18, 925-938.

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

Robins, J, E. Tchetgen Tchetgen, L. Li, and A. van der Vaart (2009): "Semiparametric Minimax Rates," *Electronic Journal of Statistics"* 3, 1305-1321.

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

Rosenbaum, P.R. and D.B. Rubin (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, 41–55.

Toth, B. and M. J. van der Laan (2016), "TMLE for Marginal Structural Models Based On An Instrument," U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 350.

Van De Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42:

1166–1202.

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

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

Van der Vaart, A.W. and J.A. Wellner (1996): *Weak Convergence and Empirical Processes With Applications to Statistics*," Springer.

Zhang, C. and S. Zhang (2014): "Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models," *Journal of the Royal Statistical Society, Series B* 76, 217–242.

Zheng, W., Z. Luo, and M. J. van der Laan (2016), "Marginal Structural Models with Counterfactual Effect Modifiers," U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 348.

Zhu, Y. and J. Bradic (2018): "Linear Hypothesis Testing in Dense High-Dimensional Linear Models," *Journal of the American Statistical Association* 113, 1583-1600.

Zhu, Y. and J. Bradic (2018): "Breaking the Curse of Dimensionality in Regression," arXiv:1708.00430.

Figure 1: Minimal conditions for the root-n rate

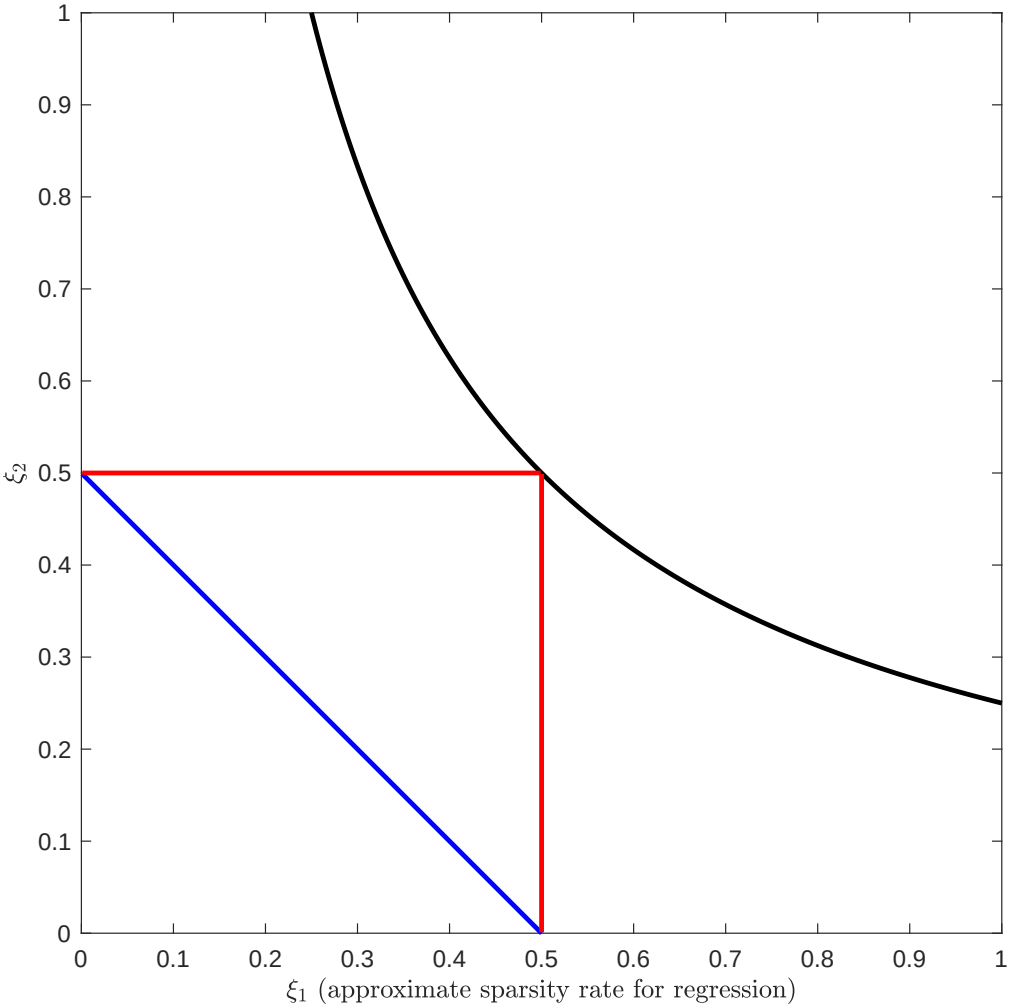
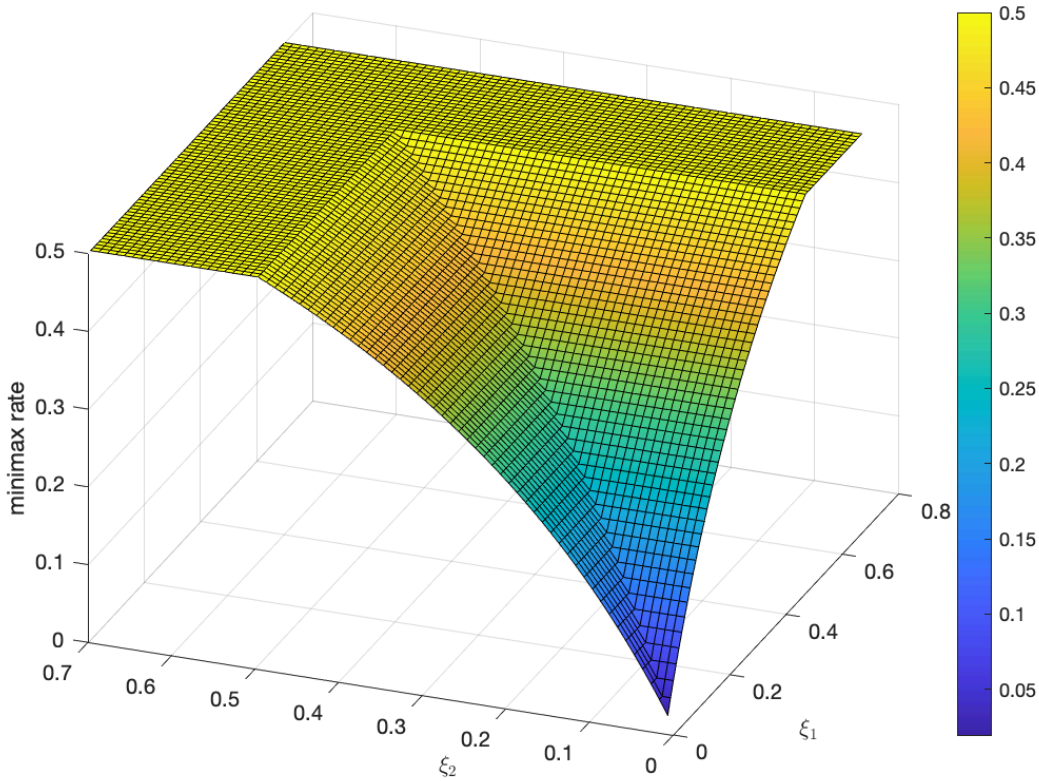


Figure 2: Minimal conditions for the root-n rate



The above figure displays $2\tilde{\xi}/(2\tilde{\xi}+1)$, where $\tilde{\xi} = \max\{\xi_1, \xi_2\}$.

A Online Appendix: Proofs of Theorems

Proof of Theorem 1: Notice that for $\beta = (\gamma, \pi, \Omega)$, the distribution of $W_i = (Y_i, Z_i, X_i')' \in \mathbb{R}^{p+2}$ is P_β, which is $N(0, \Sigma_\beta)$, where

$$\Sigma_\beta := \begin{pmatrix} \Omega_{11} + \|\gamma\|_2^2 & \Omega_{12} + \gamma'\pi & \gamma' \\ \Omega_{12} + \gamma'\pi & \Omega_{22} + \|\pi\|_2^2 & \pi' \\ \gamma & \pi & I_p \end{pmatrix} \qquad \text{with} \qquad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{pmatrix}. \tag{A.1}$$

We define $\beta_* = (0, 0, I_2)$. Clearly, $\Sigma_{\beta_*} = I_{p+2}$ and $\beta_* \in \Theta_{f_1, f_2}$.

Let $k = \left\lfloor c_1 \sqrt{n/\ln p} \right\rfloor$, where $c_1 > 0$ is a constant to be determined below. Define $\mathcal{Q}_k = \{v \in \{0,1\}^p : \|v\|_0 = k\}$. Let $N = |\mathcal{Q}_k|$. Clearly, $N = \binom{p}{k}$. We list elements in $\mathcal{Q}_k$, i.e., $Q_k = \{\delta_1, ..., \delta_N\}$. For $1 \le j \le N$, define $\gamma_j = c_n \delta_j$ and $\pi_j = c_n \delta_j$, where $c_n = c_0 \sqrt{n^{-1} \ln p}$ and c_0 is a constant chosen as follows. Now we choose any constants $c_0, c_1 > 0$ that satisfy

$$c_0 \le \sqrt{\kappa/12}, \ c_0 c_1 \le \min\{M_1, 2C_0\} \ \text{ and } c_0^2 c_1 \le \frac{\sqrt{n/\ln p}}{2}(1 - M_2^{-1}). \tag{A.2}$$

One can easily verify that (A.2) guarantees

$$c_0^2 k n^{-1} \ln p \le 1/2 \qquad \text{and} \qquad k^2/p^{1-6c_0^2} = o(1). \tag{A.3}$$

Now we define

$$\beta_j = (\gamma_j, \pi_j, \bar{\Omega}) \qquad \text{with} \qquad \bar{\Omega} = \begin{pmatrix} 1 - c_n^2 k & -c_n^2 k \\ -c_n^2 k & 1 - c_n^2 k \end{pmatrix}.$$

We have

$$\Sigma_{\beta_j} := \begin{pmatrix} 1 & 0 & c_n \delta_j' \\ 0 & 1 & c_n \delta_j' \\ c_n \delta_j & c_n \delta_j & I_p \end{pmatrix}. \tag{A.4}$$

Let ψ be an arbitrary random variable satisfying $P(|\psi| \le 1) = 0$ to be chosen later. Notice that

$$\begin{aligned} \left| N^{-1} \sum_{j=1}^N E_{\beta_j} \psi - E_{\beta_*} \psi \right| &= \left| E_{\beta_*} N^{-1} \sum_{j=1}^N \psi \left(\frac{dP_{\beta_j}}{dP_{\beta_*}} - 1 \right) \right| \le E_{\beta_*} \left| N^{-1} \sum_{j=1}^N \left(\frac{dP_{\beta_j}}{dP_{\beta_*}} - 1 \right) \right| \\ &\le \sqrt{E_{\beta_*} \left(N^{-1} \sum_{j=1}^N \left(\frac{dP_{\beta_j}}{dP_{\beta_*}} - 1 \right) \right)^2} = \sqrt{N^{-2} \sum_{j_2=1}^N \sum_{j_1=1}^N E_{\beta_*} \frac{dP_{\beta_{j_1}}}{dP_{\beta_*}} \frac{dP_{\beta_{j_2}}}{dP_{\beta_*}} - 1}. \end{aligned} \tag{A.5}$$

The rest of the proof proceeds in three steps. We first verify that $\beta_j \in \Theta_{\tilde{\xi}_1, \tilde{\xi}_2}$ and then conduct computations to bound (A.5). Then we derive the desired result.

Step 1: show $\beta_j \in \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$ for $1 \le j \le N$.

By (A.2), we have that

$$\|\gamma_j\|_1 \vee \|\pi_j\|_1 = c_n k = c_0 k\sqrt{n^{-1}\ln p} \le c_0 c_1 \le M_1.$$

To verify $\gamma_j \in \mathcal{M}_{C_0,\tilde{\xi}_1}$, we need to show that $c_n\sqrt{k-t} \le C_0 t^{-\tilde{\xi}_1}$ $\forall 1 \le t \le k$. Since $\tilde{\xi}_1 \le 1/2$, we only need to show $c_n\sqrt{k-t} \le C_0 t^{-1/2}$ $\forall 1 \le t \le k$. Equivalently, this is to show that

$$C_0^2 t^{-1} + c_n^2 t \ge c_n^2 k \qquad \forall 1 \le t \le k.$$

Notice that $C_0^2 t^{-1} + c_n^2 t \ge 2C_0 c_n$. It suffices to show $2C_0 c_n \ge c_n^2 k$. This holds by $c_n = c_0\sqrt{n^{-1}\ln p}$, $k \le c_1\sqrt{n/\ln p}$ and $c_0 c_1 \le 2C_0$ (due to (A.2)).

Similarly, the analogous argument can verify that $\pi_j \in \mathcal{M}_{C_0,\tilde{\xi}_2}$.

Notice that the eigenvalues of $\bar{\Omega}$ are 1 and $1-2c_n^2 k$. Since $c_n = c_0\sqrt{n^{-1}\ln p}$, $k = c_1\sqrt{n/\ln p}$ and $c_0^2 c_1 \le (1-M_2^{-1})\sqrt{n/\ln p}/2$ (due to (A.2)), we have that $1-2c_n^2 k \ge M_2^{-1}$. Thus, eigenvalues of $\bar{\Omega}$ are between M_2^{-1} and M_2. Therefore, $\beta_j \in \Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$.

Step 2: computing likelihood.

By Lemma 3 in Cai and Guo (2017), we have that

$$E_{\beta_*}\frac{dP_{\beta_{j_1}}}{dP_{\beta_*}}\frac{dP_{\beta_{j_2}}}{dP_{\beta_*}} = \left[\det\left(I_{p+2} - (\Sigma_{\beta_*}^{-1}\Sigma_{\beta_{j_1}} - I_{p+2})(\Sigma_{\beta_*}^{-1}\Sigma_{\beta_{j_2}} - I_{p+2})\right)\right]^{-n/2}. \tag{A.6}$$

By (A.4) and $\Sigma_{\beta_*} = I_{p+2}$, we have that

$$\Sigma_{\beta_*}^{-1}\Sigma_{\beta_j} - I_{p+2} = \begin{pmatrix} 0 & 0 & c_n\delta_j' \\ 0 & 0 & c_n\delta_j' \\ c_n\delta_j & c_n\delta_j & 0 \end{pmatrix}.$$

Then we have

$$\begin{aligned}
&I_{p+2} - (\Sigma_{\beta_*}^{-1}\Sigma_{\beta_{j_1}} - I_{p+2})(\Sigma_{\beta_*}^{-1}\Sigma_{\beta_{j_2}} - I_{p+2}) \\
&= I_{p+2} - \begin{pmatrix} 0 & 0 & c_n\delta_{j_1}' \\ 0 & 0 & c_n\delta_{j_1}' \\ c_n\delta_{j_1} & c_n\delta_{j_1} & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 & c_n\delta_{j_2}' \\ 0 & 0 & c_n\delta_{j_2}' \\ c_n\delta_{j_2} & c_n\delta_{j_2} & 0 \end{pmatrix} \\
&= I_{p+2} - \begin{pmatrix} c_n^2\delta_{j_1}'\delta_{j_2} & c_n^2\delta_{j_1}'\delta_{j_2} & 0 \\ c_n^2\delta_{j_1}'\delta_{j_2} & c_n^2\delta_{j_1}'\delta_{j_2} & 0 \\ 0 & 0 & 2c_n^2\delta_{j_1}\delta_{j_2}' \end{pmatrix} = \begin{pmatrix} 1-c_n^2\delta_{j_1}'\delta_{j_2} & -c_n^2\delta_{j_1}'\delta_{j_2} & 0 \\ -c_n^2\delta_{j_1}'\delta_{j_2} & 1-c_n^2\delta_{j_1}'\delta_{j_2} & 0 \\ 0 & 0 & I_p - 2c_n^2\delta_{j_1}\delta_{j_2}' \end{pmatrix}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\det\left[I_{p+2} - (\Sigma_{\beta_*}^{-1}\Sigma_{\beta_{j_1}} - I_{p+2})(\Sigma_{\beta_*}^{-1}\Sigma_{\beta_{j_2}} - I_{p+2})\right] \\
&= \det\left(I_p - 2c_n^2\delta_{j_1}\delta_{j_2}'\right) \times \det\begin{pmatrix} 1 - c_n^2\delta_{j_1}'\delta_{j_2} & -c_n^2\delta_{j_1}'\delta_{j_2} \\ -c_n^2\delta_{j_1}'\delta_{j_2} & 1 - c_n^2\delta_{j_1}'\delta_{j_2} \end{pmatrix} \\
&\overset{\text{(i)}}{=} \left(1 - 2c_n^2\delta_{j_1}'\delta_{j_2}\right) \times \det\begin{pmatrix} 1 - c_n^2\delta_{j_1}'\delta_{j_2} & -c_n^2\delta_{j_1}'\delta_{j_2} \\ -c_n^2\delta_{j_1}'\delta_{j_2} & 1 - c_n^2\delta_{j_1}'\delta_{j_2} \end{pmatrix} \\
&= \left(1 - 2c_n^2\delta_{j_1}'\delta_{j_2}\right) \times \left\{\left(1 - c_n^2\delta_{j_1}'\delta_{j_2}\right)^2 - \left(-c_n^2\delta_{j_1}'\delta_{j_2}\right)^2\right\} = \left(1 - 2c_n^2\delta_{j_1}'\delta_{j_2}\right)^2,
\end{aligned}
$$

where (i) follows by Sylvester's determinant identity. By (A.6), we have that

$$
E_{\beta_*}\frac{dP_{\beta_{j_1}}}{dP_{\beta_*}}\frac{dP_{\beta_{j_2}}}{dP_{\beta_*}} = \left(1 - 2c_n^2\delta_{j_1}'\delta_{j_2}\right)^{-n} \overset{\text{(i)}}{\leq} \exp\left(6nc_n^2\delta_{j_1}'\delta_{j_2}\right), \tag{A.7}
$$

where (i) follows by $c_n^2\delta_{j_1}'\delta_{j_2} \leq c_n^2 k \leq 1/2$ and the fact that $(1-x)^{-n} < \exp(3xn)$ for any $x \in [0, 1/2]$. (To see this, define $f(x) = -3x - \ln(1-x)$. Notice that $f(\cdot)$ is convex on $[0, 1/2]$ by checking $f''(\cdot)$. Also notice that $f(0) < 0$ and $f(1/2) < 0$. Hence, $f(x) < 0$ on $[0, 1/2]$. This means $-\ln(1-x) \leq 3x$. Multiplying both sides by n and taking exponential, we obtain $(1-x)^{-n} \leq \exp(3xn)$.)

Now we combine (A.5) and (A.7), obtaining

$$
\begin{aligned}
\left(N^{-1}\sum_{j=1}^{N} E_{\beta_j}\psi - E_{\beta_*}\psi\right)^2 &\leq N^{-2}\sum_{j_2=1}^{N}\sum_{j_1=1}^{N}\exp\left(6nc_n^2\delta_{j_1}'\delta_{j_2}\right) - 1 \\
&\overset{\text{(i)}}{<} \exp\left(\exp\left(6nc_n^2 + \ln(k^2/p)\right)\right) - 1 = \exp\left(\frac{k^2}{p^{1-6c_0^2}}\right) - 1 \overset{\text{(iii)}}{=} o(1),
\end{aligned} \tag{A.8}
$$

where (i) follows by Lemma A.1, (ii) follows by $c_n = c_0\sqrt{n^{-1}\ln p}$ and (iii) follows by (A.3).

Step 3: derive the desired result.

Let $CI_n = [l_n, u_n]$ be an arbitrary confidence interval for $\gamma'\pi$ with nominal coverage probability $1-\alpha$ on $\Theta_{\tilde{\xi}_1,\tilde{\xi}_2}$. In other words,

$$
\inf_{\beta\in\Theta_{\tilde{\xi}_1,\tilde{\xi}_2}} P_\beta\left(l_n \leq \phi(\beta) \leq u_n\right) \geq 1 - \alpha. \tag{A.9}
$$

We now choose the random variable

$$
\psi = \mathbf{1}\left\{c_n^2 k \notin CI\right\}.
$$

By (A.8), we have

$$
N^{-1}\sum_{j=1}^{N} E_{\beta_j}\psi \leq E_{\beta_*}\psi + o(1). \tag{A.10}
$$

Notice that $\phi(\beta_j) = \gamma_j'\pi_j = c_n^2 k$ and $\phi(\beta_*) = 0$. We observe that

$$\begin{aligned}
P_{\beta_*}\left(u_n - l_n \geq c_n^2 k\right) &\geq P_{\beta_*}\left(0 \in [l_n, u_n] \text{ and } c_n^2 k \in [l_n, u_n]\right) \\
&\overset{\text{(i)}}{\geq} P_{\beta_*}\left(0 \in [l_n, u_n]\right) - P_{\beta_*}\left(c_n^2 k \notin [l_n, u_n]\right) \\
&= P_{\beta_*}\left(\phi(\beta_*) \in [l_n, u_n]\right) - E_{\beta_*}\psi \\
&\overset{\text{(ii)}}{\geq} 1 - \alpha - E_{\beta_*}\psi \\
&\overset{\text{(iii)}}{\geq} 1 - \alpha - N^{-1}\sum_{j=1}^{N} E_{\beta_j}\psi + o(1) \\
&= 1 - \alpha - N^{-1}\sum_{j=1}^{N} P_{\beta_j}(c_n^2 k \notin CI) + o(1) \\
&= 1 - \alpha - N^{-1}\sum_{j=1}^{N} P_{\beta_j}(\phi(\beta_j) \notin CI) + o(1) \\
&\overset{\text{(iv)}}{\geq} 1 - \alpha - \alpha + o(1),
\end{aligned} \tag{A.11}$$

where (i) follows by $P(A \bigcap B) \geq P(A) - P(B^c)$ for any events A, B, (ii) follows by (A.9) and $\beta_* \in \Theta_{\tilde{\xi}_1, \tilde{\xi}_2}$, (iii) follows by (A.10) and (iv) follows by (A.9) and $\beta_j \in \Theta_{\tilde{\xi}_1, \tilde{\xi}_2}$.

Now we observe that

$$E_{\beta_*}|CI| = E_{\beta_*}(l_n - u_n) \geq c_n^2 k P_{\beta_*}\left(u_n - l_n \geq c_n^2 k\right) \geq c_n^2 k(1 - 2\alpha + o(1)).$$

Notice that $c_n^2 k \gtrsim \sqrt{n^{-1}\ln p}$. Since CI is an arbitrary confidence interval, the proof is complete. *Q.E.D.*

Lemma A.1 *Let $k \in \mathbb{N}$ and define $\mathcal{Q}_k = \{v \in \{0,1\}^p : \|v\|_0 = k\}$. Let v and u be two independent vectors that have a uniform distribution on $\mathcal{Q}_k$. Then for any $D \geq 0$,*

$$E\exp\left(Du'v\right) < \exp\left(\exp\left(D + \ln(k^2/p)\right)\right).$$

Proof: Let $N = |\mathcal{Q}_k|$. We list elements in $\mathcal{Q}_k$, i.e., $Q_k = \{x_1, ..., x_N\}$. Then

$$E\exp\left(Du'v\right) = N^{-2}\sum_{j_2=1}^{N}\sum_{j_1=1}^{N}\exp\left(Dx_{j_1}'x_{j_2}\right) \overset{\text{(i)}}{=} N^{-1}\sum_{j=1}^{N}\exp\left(Dx_1'x_j\right) = E\exp(Dx_1'v),$$

(i) follows by the observation that $\sum_{j_1=1}^{N}\exp(Dx_{j_1}'x_{j_2})$ does not depend on j_2. Without loss of generality, we take $x_1 = (1, ..., 1, 0, ..., 0)'$, i.e., the vector whose first k entries are nonzero.

Let $\mathcal{C}_{n,k}$ be the population that consists of n elements with $n-k$ elements being 0 and the remaining k being 1. Let $\{\xi_i\}_{i=1}^k$ be a random sample without replacement from the population

of $\mathcal{C}_{n,k}$. We observe that $x_1'v$ has the same distribution as $\sum_{i=1}^k \xi_i$. Then

$$E\exp(Dx_1'v) = E\exp\left(D\sum_{i=1}^k \xi_i\right).$$

Let $\{\zeta_i\}_{i=1}^k$ be a random sample with replacement from $\mathcal{C}_{n,k}$. In other words, $\{\zeta_i\}_{i=1}^k$ is i.i.d Bernoulli with $E(\zeta_i) = k/p$. Since $x \mapsto \exp(Dx)$ is a convex function, we can use Theorem 4 of Hoeffding (1963) and obtain that

$$\begin{aligned} E\exp\left(D\sum_{i=1}^k \xi_i\right) \le E\exp\left(D\sum_{i=1}^k \zeta_i\right) &= [E\exp(D\zeta_1)]^k \\ &\overset{\text{(i)}}{=} \left(1 - \frac{k}{p} + \frac{k}{p}\exp(D)\right)^k \\ &\overset{\text{(ii)}}{\le} \exp\left(\frac{k^2}{p}[\exp(D) - 1]\right) \\ &< \exp\left(\frac{k^2}{p}\exp(D)\right) = \exp\left(\exp\left(D + \ln(k^2/p)\right)\right), \end{aligned}$$

where (i) follows by the moment generating function of Bernoulli distributions, (ii) follows by the elementary inequality $1 + x \le \exp(x)$ for $x \ge 0$. The proof is complete. *Q.E.D.*

Proof of Theorem 2: Let $\beta_* = (\gamma, \pi, \Omega)$ with $\gamma = \pi = 0$ and $\Omega = I_2$. Let

$$\Theta^{(1)} = \left\{\beta = (\gamma, \pi, \Omega):\ \gamma = \pi = c_n\delta,\ \delta \in \{0,1\}^p,\ \|\delta\|_0 = k,\ \Omega = \begin{pmatrix} 1 - c_n^2 k & -c_n^2 k \\ -c_n^2 k & 1 - c_n^2 k \end{pmatrix}\right\},$$

where $c_n = c_0\sqrt{n^{-1}\ln p}$ and $k = \left\lfloor c_1\sqrt{n/\ln p}\right\rfloor$ with c_0, c_1 satisfying (A.2) in the proof of Theorem 1. Notice that $c_n^2 k \le c_1 c_0^2\sqrt{n^{-1}\ln p} = o(1)$ is smaller than $1/4$ for large n.

In (A.8) from the proof of Theorem 1, we have already proved that for any random variable ψ satisfying $|\psi| \le 1$, we have that

$$\inf_{\beta\in\Theta^{(1)}} E_\beta\psi \le E_{\beta_*}\psi + o(1). \tag{A.12}$$

Recall that for any $\beta = (\gamma, \pi, \Omega)$, we can formulate it in the corresponding partial linear form with $\lambda = f(\beta) = (\theta, \mu, \pi, \sigma_u^2, \sigma_\varepsilon^2)$, where $\theta = \Omega_{1,2}/\Omega_{2,2}$, $\mu = \gamma - \pi\Omega_{1,2}/\Omega_{2,2}$, $\sigma_u^2 = \Omega_{2,2}$ and $\sigma_\varepsilon^2 = \Omega_{1,1} - \Omega_{1,2}^2/\Omega_{2,2}$. We denote $\lambda_* = f(\beta_*)$. We use the notation $f_1(\beta) = \theta = \Omega_{1,2}/\Omega_{2,2}$. Notice that P_β and P_λ with $\lambda = f(\beta)$ are the same probability measure. For this reason, we use P_β and P_λ exchangeably. Now let $CI_*(\cdot) = [u_*(\cdot), l_*(\cdot)]$ be an arbitrary confidence interval for θ that has uniform coverage $1 - \alpha$ over $\Lambda_{\tilde{\xi}_1,\tilde{\xi}_2}$.

Recall that $c_n^2 k \le 1/4$ for large n. Therefore, for $\beta \in \Theta^{(1)}$ and for large n, we have $\theta = f_1(\beta) = -c_n^2 k/(1 - c_n^2 k) \le -2c_n^2 k$, $3/4 \le \sigma_u^2 \le 1$, $1/2 \le \sigma_\varepsilon^2 \le 1$ and $\|\rho\|_1 \vee \|\pi\|_1 \le 2M_1$.

Moreover, for $\beta \in \Theta^{(1)}$ and for large n, $\mu = \gamma - \pi\Omega_{1,2}/\Omega_{2,2} = (1-\theta)\gamma \in \mathcal{M}_{C_0,\tilde{\xi}_1}$ since $\theta = o(1)$ and we can always shrink c_1 to $c_1/2$ if needed. Therefore, for large n, $f(\Theta^{(1)}) \subset \Lambda_{\tilde{\xi}_1,\tilde{\xi}_2}$. Therefore, CI_* has uniform $1-\alpha$ coverage over $\Theta^{(1)}$ for large n.

Now we consider

$$\psi_* = \mathbf{1}\left\{CI_* \backslash A \neq \emptyset\right\},$$

where $A = [-c_n^2 k, c_n^2 k]$. By (A.12), we have

$$\inf_{\beta \in \Theta^{(1)}} E_\beta \psi_* \leq E_{\beta_*}\psi_* + o(1).$$

Notice that for $\beta \in \Theta^{(1)}$, we have

$$\begin{aligned} E_\beta \psi_* &= E_\lambda \psi_* \qquad (\text{with } \lambda = f(\beta)) \\ &= P_\lambda\left(CI_* \backslash A \neq \emptyset\right) \\ &\geq P_\lambda\left(f_1(\beta) \in CI_* \text{ and } f_1(\beta) \notin A\right) \\ &\overset{\text{(i)}}{=} P_\lambda\left(f_1(\beta) \in CI_*\right) = P_\lambda(\theta \in CI_*) \overset{\text{(ii)}}{\geq} 1-\alpha, \end{aligned}$$

where (i) follows by the fact that $f_1(\beta) \leq -2c_n^2 k$ and thus $f_1(\beta) \notin A$ and (ii) follows by the fact that CI_* is a confidence interval for θ. Hence, the above displays imply

$$P_{\lambda_*}\left(CI_* \backslash A \neq \emptyset\right) = E_{\lambda_*}\psi_* = E_{\beta_*}\psi_* \geq 1-\alpha-o(1). \tag{A.13}$$

On the other hand, we notice that

$$\begin{aligned} &P_{\lambda_*}\left(CI_* \backslash A \neq \emptyset\right) \\ &= P_{\lambda_*}\left(CI_* \backslash A \neq \emptyset \text{ and } 0 \in CI_*\right) + P_{\lambda_*}\left(CI_* \backslash A \neq \emptyset \text{ and } 0 \notin CI_*\right) \\ &\overset{\text{(i)}}{\leq} P_{\lambda_*}\left(CI_* \backslash A \neq \emptyset \text{ and } 0 \in CI_*\right) + \alpha \overset{\text{(ii)}}{\leq} P_{\lambda_*}\left(|CI_*| \geq c_n^2 k\right) + \alpha, \end{aligned} \tag{A.14}$$

where (i) follows by the fact that $\theta = 0$ at $\lambda_* = f(\beta_*)$ and that CI_* is a confidence interval for θ and (ii) follows by the fact that $\{CI_* \backslash A \neq \emptyset\} \bigcap \{0 \in CI_*\} \subset \{|CI_*| \geq c_n^2 k\}$. (To see this last step, let $CI_* = [l_*, u_*]$. Notice that $0 \in CI_*$ means $l_* \leq 0 \leq u_*$. Also notice that $CI_* \backslash A \neq \emptyset$ means the event $\{l_* < -c_n^2 k\} \bigcup \{u_* > c_n^2 k\}$. If $l_* < -c_n^2 k$, then $0 \leq u_*$ would imply $u_* - l_* \geq c_n^2 k$; if $u_* > c_n^2 k$, then $l_* \leq 0$ implies $u_* - l_* \geq c_n^2 k$. Thus, in both cases, we have $|CI_*| = u_* - l_* \geq c_n^2 k$.)

Now we combine (A.13) and (A.14), obtaining $P_{\lambda_*}\left(|CI_*| \geq c_n^2 k\right) \geq 1 - 2\alpha - o(1)$. *Q.E.D.*

For the proof of results in Sections 5 and 6 let $\varepsilon_n = \sqrt{\ln(p)/n}$, $s_0 \geq C\varepsilon_n^{-2/(2\xi_2+1)}$, and π be coefficients of the least squares projection of $\alpha_0(X)$ on $b(X)$, satisfying

$$M - \Sigma\pi = E[b(X)\{\alpha_0(X) - b(X)'\pi\}] = 0.$$

By Assumption 3 we can define $J_0 \subset \{1, ..., p\}$ as indices of a sparse approximation with $|J_0| = s_0$, where $|A|$ denotes the number of elements of a matrix, and coefficients $\tilde{\pi}_j$ for $j \in J_0$ such that for $\tilde{\pi} = (\tilde{\pi}_1, ..., \tilde{\pi}_J)'$, with $\tilde{\pi}_j = 0$ for $j \notin J_0$,

$$\|\pi - \tilde{\pi}\|_2 \leq C s_0^{-\xi_2} \leq C \varepsilon_n^{-2\xi_2/(2\xi_2+1)}$$

Also define π_* as

$$\pi_* \in \arg\min_v (\pi - v)\prime\Sigma(\pi - v) + 2\varepsilon_n \sum_{j \in J_0^c} |v_j|. \tag{A.15}$$

Lemma A1: $\|\Sigma(\pi_* - \pi)\|_\infty \leq C\varepsilon_n$.

Proof: Let $e_j \in \mathbb{R}^p$ denote the j-th column of I_p. The first-order condition for π^* imply that for $j \in J_0$, we have $e_j'\Sigma(\pi_* - \pi) = 0$; for $j \in J_0^c$, we have that $e_j'\Sigma(\pi_* - \pi) + \varepsilon_n z_j = 0$, where $z_j = \text{sign}(\pi_{*,j})$ if $\pi_{*,j} \neq 0$ and $z_j \in [-1, 1]$ if $\pi_{*,j} = 0$. Therefore, for any j, we have that $|e_j'\Sigma(\pi_* - \pi)| \leq \varepsilon_n$. Hence, $\|\Sigma(\pi_* - \pi)\|_\infty \leq \varepsilon_n$. *Q.E.D.*

Lemma A2: $(\pi - \pi_*)'\Sigma(\pi - \pi_*) \leq C\varepsilon_n^{4\xi_2/(2\xi_2+1)}$.

Proof: By the definition of π_*, we have that by the largest eigenvalue of Σ bounded,

$$\begin{aligned}(\pi - \pi_*)'\Sigma(\pi - \pi_*) + \varepsilon_n \sum_{j \in J_0^c} |\pi_{*,j}| &\leq (\pi - \tilde{\pi})'\Sigma(\pi - \tilde{\pi}) + \varepsilon_n \sum_{j \in J_0^c} |\tilde{\pi}_j| = (\pi - \tilde{\pi})\prime\Sigma(\pi - \tilde{\pi}) \\ &\leq C \|\pi - \tilde{\pi}\|_2^2 \leq C\varepsilon_n^{-4\xi_2/(2\xi_2+1)}. \ Q.E.D.\end{aligned}$$

Let J be the vector of indices of nonzero elements of π_*.

Lemma A3: $|J| \leq C\varepsilon_n^{-2/(2\xi_2+1)}$.

Proof: For all $j \in J \backslash J_0$ the first order conditions to equation (A.15) imply $|e_j'\Sigma(\pi_* - \pi)| = \varepsilon_n$. Therefore, It follows that

$$\sum_{j \in J \backslash J_0} \left(e_j'\Sigma(\pi_* - \pi)\right)^2 = \frac{1}{4}\varepsilon_n^2 |J \backslash J_0|.$$

In addition,

$$\begin{aligned}\sum_{j \in J \backslash J_0} \left(e_j'\Sigma(\pi_* - \pi)\right)^2 &\leq \sum_{j=1}^p \left(e_j'\Sigma(\pi_* - \pi)\right)^2 = (\pi_* - \pi)'\Sigma \left(\sum_{j=1}^p e_j e_j'\right) \Sigma(\pi_* - \pi) \\ &= (\pi_* - \pi)'\Sigma^2(\pi_* - \pi) \leq \lambda_{\max}(\Sigma)\{(\pi - \pi_*)'\Sigma(\pi - \pi_*)\} \leq C\varepsilon_n^{4\xi_2/(2\xi_2+1)},\end{aligned}$$

where the last inequality follows by Lemma A2 and $\lambda_{\max}(\Sigma) \leq C$. Combining the above two displays, we obtain

$$\frac{1}{4}\varepsilon_n^2 |J \backslash J_0| \leq C\varepsilon_n^{4\xi_2/(2\xi_2+1)}.$$

Dividing through by ε_n^2 gives $|J\backslash J_0| \leq C\varepsilon_n^{-2/(2\xi_2+1)}$. Thus by $s_0 \leq C\varepsilon_n^{-2/(2\xi_2+1)}$,

$$|J| = |J_0| + |J\backslash J_0| = s_0 + |J\backslash J_0| \leq s_0 + C\varepsilon_n^{-2/(2\xi_2+1)} \leq C\varepsilon_n^{-2/(2\xi_2+1)}. \; Q.E.D.$$

Lemma A4: *If* $\xi_2 > 1/2$ *then* $\|\pi_* - \pi\|_1 \lesssim \varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)}$.

Proof: By Lemma B1 and $\xi_2 > 1/2$, we have that

$$\left\|\pi_{J_0^c}\right\|_1 \leq s_0^{1/2-\xi_2} \leq C\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)}.$$

Let $J_1 = J \cup J_0$ note that $J \subset J_1$ and $J_0 \subset J_1$ imply $J_1^c \subset J^c$ and $J_1^c \subset J_0^c$, so that

$$\left\|(\pi_*)_{J_1^c} - \pi_{J_1^c}\right\|_1 = \left\|\pi_{J_1^c}\right\|_1 \leq \left\|\pi_{J_0^c}\right\|_1.$$

Also, by Lemma A3,

$$|J_1| \leq |J| + |J_0| \leq C\varepsilon_n^{-2/(2\xi_2+1)} + s_0 \leq C\varepsilon_n^{-2/(2\xi_2+1)}$$

Therefore we have

$$\begin{aligned}\|\pi_* - \pi\|_1 &= \|(\pi_*)_{J_1} - \pi_{J_1}\|_1 + \left\|(\pi_*)_{J_1^c} - \pi_{J_1^c}\right\|_1 \leq \|(\pi_*)_{J_1} - \pi_{J_1}\|_1 + \left\|\pi_{J_0^c}\right\|_1 \\ &\leq \sqrt{|J_1|}\,\|(\pi_*)_{J_1} - \pi_{J_1}\|_2 + C\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)} \\ &\leq C\varepsilon_n^{-1/(2\xi_2+1)}\|\pi_* - \pi\|_2 + C\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)} \leq C\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)}. \; Q.E.D.\end{aligned}$$

Lemma A5: $\|\hat{\Sigma}\pi_* - \Sigma\pi_*\|_\infty = O_p(\varepsilon_n)$, $\|\hat{\Sigma}\pi - \Sigma\pi\|_\infty = O_p(\varepsilon_n)$.

Proof: By $(\pi - \pi_*)'\Sigma(\pi - \pi_*) \longrightarrow 0$ and $\pi'\Sigma\pi \leq E[\alpha_0(X)^2]$ it follows that $E[(b(X)'\pi_*)^2] = \pi_*'\Sigma\pi_* \leq C$. The first conclusion then follows by uniform boundedness of the elements of $b(X)$ and Lemma B2 with $X_{i,j} = b_j(X_i)$ and $X_{i0} = b(X)'\pi_*$. The second conclusion follows similarly *Q.E.D.*

Next let

$$\hat{\pi} = \arg\min_{\pi}\{-2\hat{M}'\pi + \pi'\hat{\Sigma}\pi + 2r\,\|\pi\|_1\},$$

for $\hat{M}$ to be specified later in this appendix.

Lemma A6: *If* $\left\|\hat{M} - M\right\|_\infty = O_p(\varepsilon_n)$ *and* $\varepsilon_n = o(r)$ *then for* $\Delta = \hat{\pi} - \pi^*$ *and any* $\hat{J}$ *such that* $(\pi^*)_{\hat{J}^c} = 0$, *with probability approaching one,*

$$\Delta'\hat{\Sigma}\Delta \leq 3r\|\Delta\|_1,\; \|\Delta_{\hat{J}^c}\|_1 \leq 3\|\Delta_{\hat{J}}\|_1.$$

Proof: By the definition of the estimator, we have

$$\hat{\pi}'\hat{\Sigma}\hat{\pi} - 2\hat{M}'\hat{\pi} + 2r\|\hat{\pi}\|_1 \leq \pi_*'\hat{\Sigma}\pi_* - 2\hat{M}'\pi_* + 4r\|\pi_*\|_1.$$

Plugging $\hat{\pi} = \pi_* + \Delta$ into the above equation and rearranging the terms gives

$$\Delta'\hat{\Sigma}\Delta + 2r\|\pi_* + \Delta\|_1 \leq 2r\|\pi_*\|_1 + 2(\hat{M} - \hat{\Sigma}\pi_*)'\Delta. \tag{A.16}$$

Note that $\|\hat{M} - M\|_\infty = O_p(\varepsilon_n)$ and by Lemma A5 $\|\hat{\Sigma}\pi_* - \Sigma\pi_*\|_\infty = O_p(\varepsilon_n)$. Then by Lemma 1, $M = \Sigma\pi$, and the triangle inequality,

$$\begin{aligned}\|\hat{M} - \hat{\Sigma}\pi_*\|_\infty &\leq \|\hat{\Sigma}\pi_* - \Sigma\pi_*\|_\infty + \|\hat{M} - M\|_\infty + \|M - \Sigma\pi_*\|_\infty \\ &\leq O_p(\varepsilon_n) + \|M - \Sigma\pi\|_\infty + \|\Sigma(\pi_* - \pi)\|_\infty = O_p(\varepsilon_n),\end{aligned}$$

Therefore, by the Holder inequality we have $\left|(\hat{M} - \hat{\Sigma}\pi_*)'\Delta\right| \leq \|\hat{M} - \hat{\Sigma}\pi_*\|_\infty\|\Delta\|_1$, so that

$$\Delta'\hat{\Sigma}\Delta + 2r\|\pi_* + \Delta\|_1 \leq 2r\|\pi_*\|_1 + 2\varepsilon_n\|\Delta\|_1.$$

By $\varepsilon_n = o(r)$ it follows that with probability approaching one $2\varepsilon_n \leq r$ and

$$\Delta'\hat{\Sigma}\Delta + 2r\|\pi_* + \Delta\|_1 \leq 2r\|\pi_*\|_1 + r\|\Delta\|_1.$$

The triangle inequality implies $\|\pi_*\|_1 = \|\pi_* + \Delta - \Delta\|_1 \leq \|\pi_* + \Delta\|_1 + \|\Delta\|_1$ so subtracting $2r\|\pi_* + \Delta\|_1$ from both sides gives the first conclusion.

Next, since $\Delta'\hat{\Sigma}\Delta \geq 0$ it also follows from equation (A.16) that $2r\|\pi_* + \Delta\|_1 \leq 2r\|\pi_*\|_1 + r\|\Delta\|_1$ with probability approaching one, so dividing through by r gives

$$2\|\pi_* + \Delta\|_1 \leq 2\|\pi_*\|_1 + \|\Delta\|_1.$$

It follows by $(\pi_*)_{\hat{J}^c} = 0$ that $\|\pi_* + \Delta\|_1 = \|(\pi_*)_{\hat{J}} + \Delta_{\hat{J}}\|_1 + \|\Delta_{\hat{J}^c}\|_1$ and $\|\pi_*\|_1 = \|(\pi_*)_{\hat{J}}\|_1$. Substituting in the previous display then gives

$$\begin{aligned}2\|(\pi_*)_{\hat{J}} + \Delta_{\hat{J}}\| + 2\|\Delta_{\hat{J}^c}\|_1 &\leq 2\|(\pi_*)_{\hat{J}}\|_1 + \|\Delta\|_1 = 2\|(\pi_*)_{\hat{J}}\|_1 + \|\Delta_{\hat{J}}\|_1 + \|\Delta_{\hat{J}^c}\|_1 \\ &\leq 2\left(\|(\pi_*)_{\hat{J}} + \Delta_{\hat{J}}\|_1 + \|\Delta_{\hat{J}}\|_1\right) + \|\Delta_{\hat{J}}\|_1 + \|\Delta_{\hat{J}^c}\|_1 \\ &= 2\|(\pi_*)_J + \Delta_J\|_1 + 3\|\Delta_J\|_1 + \|\Delta_{J^c}\|_1.\end{aligned}$$

Subtracting $2\|(\pi_*)_J + \Delta_J\|_1 + \|\Delta_{J^c}\|_1$ from both sides gives the second conclusion. *Q.E.D.*

Lemma A7: *If* $\left\|\hat{M} - M\right\|_\infty = O_p(\varepsilon_n)$ *and* $\varepsilon_n = o(r)$ *then* $\Delta'\hat{\Sigma}\Delta = O_p((r/\varepsilon_n)^2\varepsilon_n^{4\xi_2/(2\xi_2+1)})$, $\|\Delta\|_1 = O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)})$, $\|\Delta\|_2 = O_p((r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)})$.

Proof: For $\hat{J} = J$ it follows from the sparse eigenvalue condition and Lemma A6 that with high probability

$$\begin{aligned}\|\Delta_J\|_2^2 &\leq C\Delta'\hat{\Sigma}\Delta \leq Cr\,\|\Delta\|_1 \leq Cr\,\|\Delta\|_1 = Cr(\|\Delta_J\|_1 + \|\Delta_J^c\|_1) \leq Cr\,\|\Delta_J\|_1 \\ &\leq Cr\sqrt{|J|}\,\|\Delta_J\|_2 \leq Cr\varepsilon_n^{-1/(2\xi_2+1)}\,\|\Delta_J\|_2 = C(r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)}\,\|\Delta_J\|_2\,.\end{aligned}$$

Dividing through by $\|\Delta_J\|_2$ then gives

$$\|\Delta_J\|_2 \leq C(r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)}.$$

Plugging this back in the final expression in the previous inequality gives the first conclusion.

For the second conclusion note that by Lemma A6,

$$\begin{aligned}\|\Delta\|_1 &= \|\Delta_{J^c}\|_1 + \|\Delta_J\|_1 \leq 4\,\|\Delta_J\|_1 \leq 4\sqrt{|J|}\,\|\Delta_J\|_2 \\ &\leq C\varepsilon_n^{-1/(2\xi_2+1)}(r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)} = C(r/\varepsilon_n)\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)}.\end{aligned}$$

For the third conclusion let N denote the indices corresponding to the largest $|J|$ entries in Δ_{J^c}, so that $N \subset J^c$, $|N| = |J|$ and $|\Delta_j| \geq |\Delta_k|$ for any $j \in J^c \cap N$ and $k \in J^c \backslash N$. By Lemma A6 for $\hat{J} = J \cup N$ it follows exactly as in second previous display that

$$\|\Delta_{\hat{J}}\|_2 \leq C(r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)}.$$

By Lemma 6.9 of van de Geer and Buhlmann (2011) and Lemma A6,

$$\|\Delta_{\hat{J}^c}\|_2 \leq (|J|)^{-1/2}\|\Delta_{\hat{J}^c}\|_1 \leq (|J|)^{-1/2}3\|\Delta_{\hat{J}}\|_1 \leq 3(|J|)^{-1/2}\sqrt{|J|}\|\Delta_J\|_2 \leq C(r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)}.$$

Therefore, by the triangle inequality,

$$\|\Delta\|_2 \leq \|\Delta_{\hat{J}}\|_2 + \|\Delta_{\hat{J}^c}\|_2 \leq C(r/\varepsilon_n)\varepsilon_n^{2\xi_2/(2\xi_2+1)},$$

giving the third conclusion. *Q.E.D.*

Lemma A8: *If* $\left\|\hat{M} - M\right\|_\infty = O_p(\varepsilon_n)$ *and* $\varepsilon_n = o(r)$ *then* $\left\|\hat{\Sigma}(\hat{\pi} - \pi)\right\|_\infty = O_p(r)$.

Proof: The Lasso first order conditions imply $\left\|\hat{\Sigma}\hat{\pi} - \hat{M}\right\|_\infty = O(r)$. By Lemma A5 $\left\|\hat{\Sigma}\pi - \Sigma\pi\right\|_\infty = O_p(\varepsilon_n)$. Then by the triangle inequality,

$$\begin{aligned}\left\|\hat{\Sigma}(\hat{\pi} - \pi)\right\|_\infty &\leq \left\|\hat{\Sigma}\hat{\pi} - \hat{M}\right\|_\infty + \left\|\hat{M} - M\right\|_\infty + \|M - \Sigma\pi\|_\infty + \left\|\left(\Sigma - \hat{\Sigma}\right)'\pi\right\|_\infty \\ &= O_p(r) + O_p(\varepsilon_n) + 0 + O_p(\varepsilon_n) = O_p(r).\ Q.E.D.\end{aligned}$$

Lemma A9: *If Assumptions 4-7 are satisfied then* $\Delta'\hat{\Sigma}\Delta = O_p((r\varepsilon_n^{-1}\delta_n)^2) = o_p(1)$.

Proof: Consider $\tilde{\pi}$ from Assumption 4 or Lemma 3 and let π_* be as defined in equation (A.15) for $J_0 = \{j : \tilde{\pi}_j \neq 0\}$. By the definition of π_*, we have

$$(\pi - \pi_*)'\Sigma(\pi - \pi_*) + \varepsilon_n \sum_{j \in J_0^c} |\pi_{*,j}| \leq (\pi - \tilde{\pi})'\Sigma(\pi - \tilde{\pi}) + \varepsilon_n \sum_{j \in J_0^c} |\tilde{\pi}_j| \leq (\pi - \tilde{\pi})'\Sigma(\pi - \tilde{\pi})$$
$$= O(\delta_n^2).$$

Let J be the vector of indices of nonzero elements of π_*. For all $j \in J \backslash J_0$ the first order conditions to equation (A.15) imply $|e_j'\Sigma(\pi_* - \pi)| = \varepsilon_n$. Therefore, it follows that

$$\sum_{j \in J \backslash J_0} \left(e_j'\Sigma(\pi_* - \pi)\right)^2 = \varepsilon_n^2 |J \backslash J_0|.$$

In addition

$$\sum_{j \in J \backslash J_0} \left(e_j'\Sigma(\pi_* - \pi)\right)^2 \leq \sum_{j=1}^{p} \left(e_j'\Sigma(\pi_* - \pi)\right)^2 = (\pi_* - \pi)'\Sigma \left(\sum_{j=1}^{p} e_j e_j'\right) \Sigma(\pi_* - \pi)$$
$$= (\pi_* - \pi)'\Sigma^2(\pi_* - \pi) \leq \lambda_{\max}(\Sigma)\{(\pi - \pi_*)'\Sigma(\pi - \pi_*)\} \leq C\delta_n^2.$$

Then from the previous equation $\varepsilon_n^2 |J \backslash J_0| \leq C\delta_n^2$ implying $|J \backslash J_0| \leq C\delta_n^2 \varepsilon_n^{-2}$. By Lemma 3 we also have $|J_0| \leq C\delta_n^2 \varepsilon_n^{-2}$, so by summing

$$|J| \leq C\delta_n^2 \varepsilon_n^{-2}.$$

Then by Lemma A6, we have

$$\Delta'\hat{\Sigma}\Delta \leq 3r\|\Delta\|_1 = 3r(\|\Delta_{J^c}\|_1 + \|\Delta_J\|_1) \leq 12r\|\Delta_J\|_1 \leq Cr\sqrt{|J|}\,\|\Delta_J\|_2 \leq Cr\varepsilon_n^{-1}\delta_n \|\Delta_J\|_2 .$$

By Assumption 6, $\|\Delta_J\|_2^2 \leq C\Delta'\hat{\Sigma}\Delta$ with probability approaching one. Then dividing through by $\|\Delta_J\|_2$ gives $\|\Delta_J\|_2 \leq Cr\varepsilon_n^{-1}\delta_n$ with probability approaching one. Plugging this inequality in the previous equation gives

$$\Delta'\hat{\Sigma}\Delta \leq C(r\varepsilon_n^{-1}\delta_n)^2$$

with probability approaching one, implying the conclusion. *Q.E.D.*

Proof of Lemma 3: By $\bar{\alpha}(X) \in \mathcal{B}$ there exists a sequence $\bar{\alpha}_k(x)$ consisting of finite dimensional linear combinations of $(b_1(x), b_2(x), ...)$ such that $\|\bar{\alpha}_k - \bar{\alpha}\|_2 \longrightarrow 0$. Since each $\bar{\alpha}_k(x)$ is a finite dimensional linear combination there exists p_k such that for $b^k(x) = (b_1(x), ..., b_{p_k}(x))'$ we have $\bar{\alpha}_k(x) = b^k(x)'\gamma_k$ for some γ_k. Let $\alpha_k(X)$ be the least square projection of $\bar{\alpha}(X)$ on $b^k(X)$. Then by $\|\bar{\alpha} - \alpha_k\|_2 \leq \|\bar{\alpha} - \bar{\alpha}_k\|_2$ it follows that

$$\|\bar{\alpha} - \alpha_k\|_2 \longrightarrow 0.$$

Let $\delta_k = \|\bar{\alpha} - \alpha_k\|_2$. By $n/\ln(p) \longrightarrow \infty$ and $p \longrightarrow \infty$ we can choose k_n so that $p_{k_n} \leq p$ and

$$p_{k_n} \leq \delta_{k_n}^2 \frac{n}{\ln(p)}.$$

Let $\delta_n = \delta_{k_n}$ and $\tilde{\pi} = (\pi'_{k_n}, 0')'$ where π'_{k_n} are the coefficients for α_{k_n}. Then $\|\tilde{\pi}\|_0 \leq p_{k_n} \leq \delta_n^2 n/\ln(p)$ and $\|\alpha_n - \bar{\alpha}\|^2 \leq \|\alpha_{k_n} - \bar{\alpha}\|^2$ so that

$$\begin{aligned}(\pi - \tilde{\pi})'\Sigma(\pi - \tilde{\pi}) &= \|\alpha_n - \alpha_{k_n}\|^2 = \|\alpha_n - \bar{\alpha} - (\alpha_{k_n} - \bar{\alpha})\|^2 \\ &\leq 2\|\alpha_n - \bar{\alpha}\|^2 + 2\|\alpha_{k_n} - \bar{\alpha}\|^2 \leq 4\delta_{k_n}^2 = 4\delta_n^2. \ Q.E.D.\end{aligned}$$

Proof of Theorem 4: Let $\theta_n = \gamma'\Sigma\pi = \mu'\pi$, $\rho_n(x) = b(x)'\gamma$, and $\alpha_n(x) = b(x)'\pi$. Let

$$\hat{U} = \hat{\mu} - \hat{\Sigma}\gamma, \ \hat{R} = \hat{M} - \hat{\Sigma}\pi.$$

Then by adding and subtracting terms, for $m(w, \rho_n) = z\rho_n(x)$,

$$\begin{aligned}\hat{\theta} - \theta_n &= \hat{M}'\hat{\gamma} + \hat{\pi}'(\hat{\mu} - \hat{\Sigma}\hat{\gamma}) - \mu'\pi = (\hat{\mu} - \mu)'\pi + \hat{\mu}'(\hat{\pi} - \pi) + \hat{M}'\hat{\gamma} - \hat{\pi}'\hat{\Sigma}\hat{\gamma} \qquad (A.17) \\ &= (\hat{\mu} - \mu)'\pi + (\hat{U} + \hat{\Sigma}\gamma)'(\hat{\pi} - \pi) + \hat{M}'\hat{\gamma} - \hat{\pi}'\hat{\Sigma}\hat{\gamma} \\ &= (\hat{\mu} - \mu)'\pi + \hat{U}'(\hat{\pi} - \pi) + \gamma'\hat{\Sigma}(\hat{\pi} - \pi) + \hat{M}'\hat{\gamma} - \pi'\hat{\Sigma}\hat{\gamma} - (\hat{\pi} - \pi)\prime\hat{\Sigma}\hat{\gamma} \\ &= (\hat{\mu} - \mu)'\pi + \hat{U}'(\hat{\pi} - \pi) + (\gamma - \hat{\gamma})'\hat{\Sigma}(\hat{\pi} - \pi) + \hat{R}'\hat{\gamma} \\ &= \hat{M}'\gamma + \pi'(\hat{\mu} - \hat{\Sigma}\gamma) - \theta_n + (\gamma - \hat{\gamma})'\hat{\Sigma}(\hat{\pi} - \pi) + \hat{U}'(\hat{\pi} - \pi) + \hat{R}'(\hat{\gamma} - \gamma) \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi_n(W_i) + T_1 + T_2 + T_3, \ \psi_n(w) = m(w, \rho_n) - \theta_n + \alpha_n(x)[y - \rho_n(x)], \\ T_1 &= (\gamma - \hat{\gamma})'\hat{\Sigma}(\hat{\pi} - \pi), \ T_2 = \hat{R}'(\hat{\gamma} - \gamma), \ T_3 = \hat{U}'(\hat{\pi} - \pi).\end{aligned}$$

It follows by Lemma B2 that $\left\|\hat{M} - M\right\|_\infty = O_p(\varepsilon_n)$ and $\|\hat{\mu} - \mu\|_\infty = O_p(\varepsilon_n)$. By Lemmas A4 and A7 applied to $\hat{\gamma}$, γ_*, γ in place of $\hat{\pi}$, π_*, and π and the triangle inequality, $\|\hat{\gamma} - \gamma\|_1 = O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi-1)/(2\xi+1)})$. Then by Lemma A8, the Holder inequality, and $r/\varepsilon_n = o(n^c)$ for any $c > 0$,

$$\begin{aligned}|T_1| &\leq \left\|\hat{\Sigma}(\hat{\pi} - \pi)\right\|_\infty \|\hat{\gamma} - \gamma\|_1 = O_p(r)O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi-1)/(2\xi+1)}) \\ &= O_p((r/\varepsilon_n)^2\varepsilon_n^{4\xi/(2\xi+1)}) = o_p(n^{-1/2}).\end{aligned}$$

Also,

$$\begin{aligned}\left\|\hat{R}\right\|_\infty &\leq \left\|\hat{M} - M\right\|_\infty + \left\|(\Sigma - \hat{\Sigma})\pi\right\|_\infty + \|M - \Sigma\pi\|_\infty \qquad (A.18) \\ &= O_p(\varepsilon_n) + O_p(\varepsilon_n) + 0 = O_p(\varepsilon_n).\end{aligned}$$

Therefore it follows that

$$|T_2| \leq \left\|\hat{R}\right\|_\infty \|\hat{\gamma} - \gamma\|_1 = O_p(\varepsilon_n) O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi-1)/(2\xi+1)})$$
$$= O_p((r/\varepsilon_n)\varepsilon_n^{4\xi/(2\xi+1)}) = o_p(n^{-1/2}).$$

Next, note that for $\varepsilon_i = Y_i - \rho_n(X_i)$,

$$T_3 = \frac{1}{n}\sum_i \varepsilon_i \{b(X_i)'(\hat{\pi} - \pi)\}.$$

Then for $\tilde{X} = (X_1, ..., X_n)$ and $\tilde{W}$ the observations not in I_ℓ it follows by $\rho_n(X_i) = E[Y_i|X_i]$ that $E[\varepsilon_i|\tilde{X}, \tilde{W}] = 0$. Then by $\hat{\pi}$ depending only on $\tilde{X}$ and $\tilde{W}$ and $Var(Y|X) \leq C$,

$$E[T_3|\tilde{X}, \tilde{W}] = \frac{1}{n}\sum_i E[\varepsilon_i|\tilde{X}, \tilde{W}]\{b(X_i)'(\hat{\pi} - \pi)\} = 0,$$
$$Var(T_3|\tilde{X}, \tilde{W}) = \frac{1}{n^2}\sum_i Var(Y_i|X_i)\{b(X_i)'(\hat{\pi} - \pi)\}^2 \leq C\frac{1}{n^2}\sum_i \{b(X_i)'(\hat{\pi} - \pi)\}^2$$
$$= C(\hat{\pi} - \pi)'\hat{\Sigma}(\hat{\pi} - \pi)/n \leq C\Delta'\hat{\Sigma}\Delta/n + C(\pi - \pi_*)'\hat{\Sigma}(\hat{\pi} - \pi_*)/n$$
$$= o_p(1) + O_p(E[(\pi - \pi_*)'\hat{\Sigma}(\pi - \pi_*)]/n)$$
$$= o_p(n^{-1}) + O_p((\pi - \pi_*)'\Sigma(\pi - \pi_*)/n) = o_p(n^{-1}),$$

where the second inequality follows by the triangle and Cauchy-Schwartz inequalities, the fourth equality by the conclusion of Lemma A7 and the Markov inequality, and the last equality by Lemma A2. Then by conditional Markov inequality it follows that

$$|T_3| = o_p(n^{-1/2}).$$

The triangle inequality then gives $T_1 + T_2 + T_3 = o_p(n^{-1/2})$, giving the first conclusion.

For the second conclusion note that

$$\psi_n(W) = \rho_n(X)Z + \alpha_n(X)Y - \alpha_n(X)\rho_n(X) - \theta_n$$
$$= T_4 + T_5 - \theta_n,\ T_4 = \rho_n(X)Z,\ T_5 = \alpha_n(X)[Y - \rho_n(X)].$$

Note that

$$E[T_4^2] = E[\rho_n(X)^2 E[Z^2|X]] \leq CE[\rho_n(X)^2] \leq C,$$
$$E[T_5^2] = E[\alpha_n(X)^2\{Y - \rho_n(X)\}^2] = E[\alpha_n(X)^2 Var(Y|X)] \leq CE[\alpha_n(X)^2] \leq C,$$
$$|\theta_n| = E[\alpha_n(X)\rho_n(X)] \leq \sqrt{E[\alpha_n(X)^2]}\sqrt{E[\rho_n(X)^2]} \leq C.$$

It then follows that $E[\psi_n(W)^2] \leq C$, so the second conclusion follows by $E[\psi_n(W)] = 0$. *Q.E.D.*

Before proving Theorem 5 we give two additional Lemmas.

Lemma A10: If $\sqrt{n}\,\|\rho_0-\rho_n\|\,\|\bar{\alpha}-\alpha_n\| \longrightarrow 0$ then $\sqrt{n}\,|E[\alpha_n(X)\rho_n(X)]-E[\bar{\alpha}(X)\rho_0(X)]| \longrightarrow 0$.

Proof: Note that $\rho_n(X)$ and $\alpha_n(X)$ are least squares projections of $\rho_0(X)$ and $\bar{\alpha}(X)$ on $b(X)$ respctively so that $E[\alpha_n(X)\rho_n(X)]=E[\alpha_n(X)\rho_0(X)]=E[\bar{\alpha}(X)\rho_n(X)]$. Then

$$\begin{aligned}
&|E[\bar{\alpha}(X)\rho_0(X)]-E[\alpha_n(X)\rho_n(X)]| \\
&=|E[\bar{\alpha}(X)\rho_0(X)]-E[\alpha_n(X)\rho_0(X)]-E[\bar{\alpha}(X)\rho_n(X)]+E[\alpha_n(X)\rho_n(X)]| \\
&=|E[\{\bar{\alpha}(X)-\alpha_n(X)\}\{\rho_0(X)-\rho_n(X)\}]| \leq \|\bar{\alpha}-\alpha_n\|_2\,\|\rho_0-\rho_n\|_2 = o(1/\sqrt{n}).
\end{aligned}$$

where the last inequality is Cauchy-Schwartz and the final equality follows by hypothesis. *Q.E.D.*

Lemma A11: If Assumptions 6 and 10 are satisfied then $\sup_x|\rho_n(x)-\rho_0(x)| \longrightarrow 0$ and $\sup_x|\hat{\rho}(x)-\rho_0(x)|=o_p(1)$.

Proof: Let $\tilde{\gamma}_0$ be the $p\times 1$ vector consisting of the first p elements of $\gamma_0=(\gamma_{01},\gamma_{02},...)$. For J as in Assumption 10 it follows by Lemma B1 and the first inequality in the proof of Lemma A4 that

$$\begin{aligned}
\|\gamma-\tilde{\gamma}_0\|_1 &\leq \|\gamma-\gamma_{0J}\|_1+\|\gamma_{0J^c}\|_1 \leq \|\gamma_J-\gamma_{0J}\|_1+\|\gamma_{J^c}\|_1+o(1) \\
&\leq \sqrt{|J|}\,\|\gamma_J-\gamma_{0J}\|_2+o(1) \leq \sqrt{|J|}\,\|\gamma-\gamma_{0J}\|_2+o(1) \\
&\leq C\,|J|^{-\xi_1+1/2}+o(1)=o(1).
\end{aligned}$$

By Assumptions 6 and 10 it then follows that

$$\sup_x|\rho_n(x)-\rho_0(x)| \leq C\sum_{j=1}^{p}|\gamma_j-\gamma_{j0}|+C\sum_{j=p+1}^{\infty}|\rho_{0j}| \leq C\,\|\gamma-\tilde{\gamma}_0\|_1+o(1) \longrightarrow 0,$$

giving the first conclusion.

For the second conclusion note that it was shown previously that $\left\|\hat{M}-M\right\|_\infty = O_p(\varepsilon_n)$. Then by Lemmas A4 and A7, the triangle inequality, and $\xi_1>1/2$

$$\|\hat{\gamma}-\tilde{\gamma}_0\|_1 \leq \|\hat{\gamma}-\gamma_*\|_1+\|\gamma_*-\gamma\|_1+\|\gamma-\tilde{\gamma}_0\|_1=o_p(1).$$

The second conclusion then follows by

$$\sup_x|\hat{\rho}_n(x)-\rho_0(x)| \leq C\sum_{j=1}^{p}|\hat{\gamma}_j-\gamma_{j0}|+C\sum_{j=p+1}^{\infty}|\rho_{0j}| \leq C\,\|\hat{\gamma}-\tilde{\gamma}_0\|_1+o(1)=o_p(1).\ Q.E.D.$$

Proof of Theorem 5: It follows exactly as in the Theorem 4 that $T_1=o_p(n^{-1/2})$ and $T_2=o_p(n^{-1/2})$. For T_3 let $\varepsilon_i=Y_i-\rho_0(X_i)$ and $d_{in}=\rho_0(X_i)-\rho_n(X_i)$. Then

$$T_3=T_{31}+T_{32},\ T_{31}=\frac{1}{n}\sum_i\varepsilon_i\{b(X_i)'(\hat{\pi}-\pi)\},\ T_{32}=\frac{1}{n}\sum_i d_{in}\{b(X_i)'(\hat{\pi}-\pi)\}.$$

It follows as in the proof of Theorem 4 that $T_{31} = o_p(n^{-1/2})$. Also by the Cauchy-Schwartz and Markov inequalities and Assumption 8,

$$\begin{aligned}|T_{32}| &\leq \sqrt{\frac{1}{n}\sum_i d_{in}^2}\sqrt{(\hat{\pi}-\pi)'\hat{\Sigma}(\hat{\pi}-\pi)} = O_p(\sqrt{E[\{d_{in}^2\}]})O_p(r\varepsilon_n^{-1}\delta_n)\\ &= O_p(\|\rho_0-\rho_n\|_2\, r\varepsilon_n^{-1}\delta_n) = o_p(n^{-1/2}).\end{aligned}$$

Then it follows as in the proof of Theorem 4 that

$$\sqrt{n}(\hat{\theta}-\theta_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_n(W_i) + o_p(1).$$

Next note that $|\theta_n - \theta_0| = o(n^{-1/2})$ by Lemma A10. Also, note that

$$\begin{aligned}\psi_n(w)-\psi(w) &= z[\rho_n(x)-\rho_0(x)] + y[\alpha_n(x)-\bar{\alpha}(x)] - \alpha_n(x)\rho_n(x) + \bar{\alpha}(x)\rho_0(x) - \theta_n + \theta_0\\ &= T_{4n}(w) + T_{5n}(w) + T_{6n}(x) - \theta_n + \theta_0\\ T_{4n}(w) &= [z-\bar{\alpha}(x)][\rho_n(x)-\rho_0(x)],\ T_{5n}(w) = [y-\rho_0(x)][\alpha_n(x)-\bar{\alpha}(x)],\\ T_{6n}(x) &= -[\rho_n(x)-\rho_0(x)][\alpha_n(x)-\bar{\alpha}(x)].\end{aligned}$$

Note that for $T_4 = \sum_{i=1}^n T_{4n}(X_i)/\sqrt{n}$ it follows by $Var(Z|X) \leq C$ that

$$E[T_4^2] = E[\{Z-\bar{\alpha}(X)\}^2\{\rho_n(X)-\rho_0(X)\}^2] \leq CE[\{\rho_n(X)-\rho_0(X)\}^2] \longrightarrow 0.$$

It then follows by the Markov inequality that $T_4 \xrightarrow{p} 0$ and similarly $T_5 \xrightarrow{p} 0$. Also by the triangle and Cauchy-Schwartz inequalities $T_6 = \sum_{i=1}^n T_{6n}(X_i)/\sqrt{n}$ satisfies

$$E[|T_6|] \leq \sqrt{n}E[T_{6n}(X)] \leq \sqrt{n}\,\|\rho_n-\rho_0\|_2\,\|\alpha_n-\bar{\alpha}\|_2 \longrightarrow 0.$$

It then follows from Theorem 4 that

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= \sqrt{n}(\hat{\theta}-\theta_n) + \sqrt{n}(\theta_n-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_n(W_i) + o_p(1) + o(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(W_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^n[\psi_n(W_i)-\psi(W_i)] + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(W_i) + o_p(1).\end{aligned}$$

The first conclusion of Theorem 5 conclusion then follows by the central limit theorem.

Recall that for $\ell \in \{1,2\}$ and $i \in I_\ell$, $\hat{\psi}_{i\ell} = \hat{\rho}_\ell(X_i)Z_i + \hat{\alpha}_\ell(X_i)(Y_i - \hat{\rho}_\ell(X_i)) - \hat{\theta}$. Let $\psi_i = \rho_0(X_i)Z_i + \bar{\alpha}(X_i)(Y_i - \rho_0(X_i)) - \theta_0$. By the law of large numbers, $|I_\ell|^{-1}\sum_{i\in I_\ell}\psi_i^2 = V + o_p(1)$. Thus, it suffices to show that $|I_\ell|^{-1}\sum_{i\in I_\ell}(\hat{\psi}_{i\ell}^2 - \psi_i^2) = o_p(1)$. Defining $\delta_{i\ell} = \hat{\psi}_{i\ell} - \psi_i$, we need to show that

$$|I_\ell|^{-1}\sum_{i\in I_\ell}\delta_{i\ell}^2 + 2|I_\ell|^{-1}\sum_{i\in I_\ell}\psi_i\delta_{i\ell} = o_p(1). \tag{A.19}$$

Notice that

$$\left| |I_\ell|^{-1} \sum_{i \in I_\ell} \psi_i \delta_{i\ell} \right| \leq \sqrt{|I_\ell|^{-1} \sum_{i \in I_\ell} \psi_i^2} \times \sqrt{|I_\ell|^{-1} \sum_{i \in I_\ell} \delta_{i\ell}^2} \overset{\text{(i)}}{=} \sqrt{O_p(1)} \times \sqrt{|I_\ell|^{-1} \sum_{i \in I_\ell} \delta_{i\ell}^2},$$

where (i) follows by $E(|I_\ell|^{-1} \sum_{i \in I_\ell} \psi_i^2) = V = O(1)$. Therefore, to show (A.19), we only need to verify

$$|I_\ell|^{-1} \sum_{i \in I_\ell} \delta_{i\ell}^2 = o_p(1). \tag{A.20}$$

Let $\varepsilon_i = Y_i - \rho_0(X_i)$. Notice that $E(\varepsilon_i \mid X_i) = 0$. We observe that

$$\delta_{i\ell} = (\hat{\rho}_\ell(X_i) - \rho_0(X_i))Z_i + (\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))\varepsilon_i - \hat{\alpha}_\ell(X_i)(\hat{\rho}_\ell(X_i) - \rho_0(X_i)) + (\theta_0 - \hat{\theta}). \tag{A.21}$$

We now bound these terms. For the first term, we notice that $\hat{\rho}_\ell$ depends on $(\hat{\Sigma}_\ell, \tilde{\mu}_\ell)$ and $\tilde{\mu}_\ell$ is computed using $\{W_i\}_{i \notin I_\ell}$. Therefore, for $i \in I_\ell$, $E(Z_i^2 \mid \hat{\rho}_\ell(X_i), X_i) = E(Z_i^2 \mid X_i)$, which is assumed to be bounded. It follows that

$$\begin{aligned} &E\left[|I_\ell|^{-1} \sum_{i \in I_\ell} (\hat{\rho}_\ell(X_i) - \rho_0(X_i))^2 Z_i^2 \mid \hat{\rho}_\ell, \{X_i\}_{i \in I_\ell} \right] \\ &\qquad = O_p(1) \times |I_\ell|^{-1} \sum_{i \in I_\ell} (\hat{\rho}_\ell(X_i) - \rho_0(X_i))^2 \overset{\text{(i)}}{=} O_p(1) \times o_p(1) = o_p(1), \end{aligned} \tag{A.22}$$

where (i) follows by Lemma A11. Similarly, we notice that $\hat{\alpha}_\ell$ depends on $(\hat{\Sigma}_\ell, \tilde{M}_\ell)$ and $\tilde{M}_\ell$ is computed using $\{W_i\}_{i \notin I_\ell}$. Therefore, for $i \in I_\ell$, $E(\varepsilon_i^2 \mid \hat{\alpha}_\ell(X_i), X_i) = E(\varepsilon_i^2 \mid X_i)$, which is assumed to be bounded. It follows that

$$E\left[|I_\ell|^{-1} \sum_{i \in I_\ell} (\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))^2 \varepsilon_i^2 \mid \hat{\alpha}_\ell, \{X_i\}_{i \in I_\ell} \right] = O_p(1) \times |I_\ell|^{-1} \sum_{i \in I_\ell} (\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))^2.$$

Since $\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i) = \hat{\alpha}_\ell(X_i) - \alpha_n(X_i) + \alpha_n(X_i) - \bar{\alpha}(X_i) = b(X_i)'(\hat{\pi}_\ell - \pi) + \alpha_n(X_i) - \bar{\alpha}(X_i)$, we have that

$$\begin{aligned} |I_\ell|^{-1} \sum_{i \in I_\ell} (\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))^2 &\leq 2|I_\ell|^{-1} \sum_{i \in I_\ell} \left(b(X_i)'(\hat{\pi}_\ell - \pi)\right)^2 + 2|I_\ell|^{-1} \sum_{i \in I_\ell} \left(\alpha_n(X_i) - \bar{\alpha}(X_i)\right)^2 \\ &= 2(\hat{\pi}_\ell - \pi)'\hat{\Sigma}_\ell(\hat{\pi}_\ell - \pi) + 2|I_\ell|^{-1} \sum_{i \in I_\ell} \left(\alpha_n(X_i) - \bar{\alpha}(X_i)\right)^2 \\ &\overset{\text{(i)}}{=} 2(\hat{\pi}_\ell - \pi)'\hat{\Sigma}_\ell(\hat{\pi}_\ell - \pi) + o_p(1) \\ &\overset{\text{(ii)}}{\leq} 2\left(\|\hat{\Sigma}_\ell^{1/2}(\hat{\pi}_\ell - \pi_*)\|_2 + \|\hat{\Sigma}_\ell^{1/2}(\pi_* - \pi)\|_2 \right)^2 + o_p(1) \\ &\leq 4\|\hat{\Sigma}_\ell^{1/2}(\hat{\pi}_\ell - \pi_*)\|_2^2 + 4\|\hat{\Sigma}_\ell^{1/2}(\pi_* - \pi)\|_2^2 + o_p(1) \\ &= 4(\hat{\pi}_\ell - \pi_*)'\hat{\Sigma}_\ell(\hat{\pi}_\ell - \pi_*) + 4(\pi_* - \pi)'\hat{\Sigma}_\ell(\pi_* - \pi) + o_p(1) \\ &\overset{\text{(iii)}}{=} o_p(1), \end{aligned} \tag{A.23}$$

where (i) follows by $E(\alpha_n(X_i)-\bar{\alpha}(X_i))^2 = o(1)$, (ii) follows by $\sqrt{(\hat{\pi}_\ell - \pi)'\hat{\Sigma}_\ell(\hat{\pi}_\ell - \pi)} = \|\hat{\Sigma}_\ell^{1/2}(\hat{\pi}_\ell - \pi)\|_2 \leq \|\hat{\Sigma}_\ell^{1/2}(\hat{\pi}_\ell - \pi_*)\|_2 + \|\hat{\Sigma}_\ell^{1/2}(\pi_* - \pi)\|_2$ and (iii) follows by Lemmas A2 and A9. The above two displays imply that

$$|I_\ell|^{-1}\sum_{i\in I_\ell}(\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))^2\varepsilon_i^2 = o_p(1). \tag{A.24}$$

By (A.23), we have that

$$\begin{aligned}|I_\ell|^{-1}\sum_{i\in I_\ell}\hat{\alpha}_\ell(X_i)^2 &= |I_\ell|^{-1}\sum_{i\in I_\ell}(\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i) + \bar{\alpha}(X_i))^2 \\ &\leq 2|I_\ell|^{-1}\sum_{i\in I_\ell}(\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))^2 + 2|I_\ell|^{-1}\sum_{i\in I_\ell}\bar{\alpha}(X_i)^2 = O_p(1).\end{aligned}$$

Therefore, Lemma A11 implies that

$$|I_\ell|^{-1}\sum_{i\in I_\ell}\hat{\alpha}_\ell(X_i)^2(\hat{\rho}_\ell(X_i) - \rho_0(X_i))^2 \leq \sup_x|\hat{\rho}(x) - \rho(x)|^2|I_\ell|^{-1}\sum_{i\in I_\ell}\hat{\alpha}_\ell(X_i)^2 = o_p(1). \tag{A.25}$$

Since $\hat{\theta} - \theta_0 = o_p(1)$, we now combine (A.21) with (A.22), (A.24) and (A.25), obtaining

$$\begin{aligned}|I_\ell|^{-1}\sum_{i\in I_\ell}\delta_{i\ell}^2 &\leq 4|I_\ell|^{-1}\sum_{i\in I_\ell}(\hat{\rho}_\ell(X_i) - \rho_0(X_i))^2 Z_i^2 + 4|I_\ell|^{-1}\sum_{i\in I_\ell}(\hat{\alpha}_\ell(X_i) - \bar{\alpha}(X_i))^2\varepsilon_i^2 \\ &\quad + 4|I_\ell|^{-1}\sum_{i\in I_\ell}\hat{\alpha}_\ell(X_i)^2(\hat{\rho}_\ell(X_i) - \rho_0(X_i))^2 + 4(\hat{\theta} - \theta_0)^2 = o_p(1).\end{aligned}$$

We have proved (A.20)..*Q.E.D.*

Proof of Theorem 6: By Assumption 8 and Lemma B2 it follows that $\hat{M} = \sum_{i=1}^n m(X_i, b)/n$ satisfies $\left\|\hat{M} - M\right\|_\infty = O_p(\varepsilon_n)$. It then follows exactly as in the proof of Theorem 3 that the remainder decomposition given there is satisfied and that $T_1 = o_p(n^{-1/2})$ and $T_2 = o_p(n^{-1/2})$. Here $\hat{\pi}$ depends only on $\tilde{X} = (X_1, ..., X_n)$, so that

$$\begin{aligned}E[T_3|\tilde{X}] &= \frac{1}{n}\sum_i E[\varepsilon_i|\tilde{X}]\{b(X_i)'(\hat{\pi} - \pi)\} = 0, \\ Var(T_3|\tilde{X}) &= \frac{1}{n^2}\sum_i Var(Y_i|X_i)\{b(X_i)'(\hat{\pi} - \pi)\}^2 \leq C(\hat{\pi} - \pi)'\hat{\Sigma}(\hat{\pi} - \pi)/n = o_p(n^{-1}),\end{aligned}$$

where the last inequality follows as in the proof of Theorem 3. The conditional Markov inequality then gives $T_3 = o_p(n^{-1})$, so the first equality in the conclusion follows as in the proof of Theorem 3. Also note that $E[m(W, \rho_n)^2] \leq C$ and $E[\{\alpha_n(X)[Y - \rho_n(X)]\}^2] = E[\alpha_n(X)^2 Var(Y|X)] \leq CE[\alpha_n(X)^2] \leq C$, so the second equality follows by Chebyshev's inequality. *Q.E.D.*

Proof of Theorem 7: It follows exactly as in the proof of Theorem 5 that

$$\sqrt{n}(\hat{\theta} - \theta_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_n(W_i) + o_p(1).$$

Next note that $|\theta_n - \theta_0| = o(n^{-1/2})$ by Lemma A10. Also, note that

$$\begin{aligned}
\psi_n(w) - \psi_0(w) &= m(x, \rho_n - \rho_0) + \alpha_n(x)[y - \rho_n(x)] - \bar{\alpha}(x)[y - \rho_0(x)] - \theta_n + \theta_0 \\
&= T_{4n}(w) + T_{5n}(w) + T_{6n}(w) + T_{7n}(w) - \theta_n + \theta_0 \\
T_{4n}(w) &= m(x, \rho_n - \rho_0) - \theta_n + \theta_0,\ T_{5n}(w) = [\alpha_n(x) - \bar{\alpha}(x)][y - \rho_0(x)], \\
T_{6n}(w) &= -[\rho_n(x) - \rho_0(x)][\alpha_n(x) - \bar{\alpha}(x)],\ T_{7n}(w) = \bar{\alpha}(x)[\rho_n(x) - \rho_0(x)].
\end{aligned}$$

Let $T_j = \sum_{i=1}^n T_{jn}(W_i)/\sqrt{n}$, $(j = 4, 5, 6, 7)$. It follows as in the proof of Theorem 5 that $T_5 = o_p(1)$ and $T_6 = o_p(1)$. Also we have $E[T_{4n}(W)] = 0$. By Lemma A11, $E[m(W, \rho_n - \rho_0)^2] \leq CE[a(W)^2] \sup_x |\rho_n(x) - \rho_0(x)|^2 \longrightarrow 0$ if Assumption 12 i) is satisfied and $E[m(W, \rho_n - \rho_0)^2] \leq C \|\rho_n - \rho_0\|^2 \longrightarrow 0$ if Assumption 12 ii) is satisfied. Therefore

$$E[T_{4n}(W)^2] \leq 2E[m(W, \rho_n - \rho_0)^2] + 2|\theta_n - \theta_0|^2 \longrightarrow 0,$$

so that $T_4 = o_p(1)$.

Next, by Lemma A11, $E[\bar{\alpha}(X)^2\{\rho_n(X) - \rho_0(X)\}^2] \leq E[\bar{\alpha}(X)^2] \sup_x |\rho_n(x) - \rho_0(x)|^2 \longrightarrow 0$ if Assumption 12 i) is satisfied and $E[\bar{\alpha}(X)^2\{\rho_n(X) - \rho_0(X)\}^2] \leq C \|\rho_n - \rho_0\|^2 \longrightarrow 0$ if Assumption 12 ii) is satisfied. Therefore

$$E[T_{7n}(W)^2] \leq E[\bar{\alpha}(X)^2\{\rho_n(X) - \rho_0(X)\}^2] \longrightarrow 0.$$

Define $T_{8n}(W) = T_{7n}(W) - (\theta_n - \theta_0)$. Note that $E[T_{8n}(W)] = 0$ and

$$E[T_{8n}(W)^2] \leq 2E[T_{7n}(W)^2] + 2|\theta_n - \theta_0|^2 \longrightarrow 0.$$

Therefore $\sum_{i=1}^n T_{8n}(W_i)/\sqrt{n} = o_p(1)$. Then by $\sqrt{n}(\theta_n - \theta_0) \longrightarrow 0$,

$$T_7 = \frac{1}{\sqrt{n}} \sum_{i=1}^n T_{8n}(W_i) + \sqrt{n}(\theta_n - \theta_0) = o_p(1) + o(1) = o_p(1).$$

Recall $\hat{\psi}_i = m(X_i, \hat{\rho}) + \hat{\alpha}(X_i)(Y_i - \hat{\rho}(X_i)) - \hat{\theta}$. Let $\psi_i = m(X_i, \rho_0) + \bar{\alpha}(X_i)(Y_i - \rho_0(X_i)) - \theta_0$. Define $\delta_i = \hat{\psi}_i - \psi_i$. Following essentially the same argument as in the proof of Theorem 5, we only need to verify

$$n^{-1} \sum_{i=1}^n \delta_i^2 = o_p(1).$$

Notice that

$$\delta_i = m(X_i, \hat{\rho} - \rho_0) + (\hat{\alpha}(X_i) - \bar{\alpha}(X_i))\varepsilon_i - \hat{\alpha}(X_i)(\hat{\rho}(X_i) - \rho_0(X_i)) + (\theta_0 - \hat{\theta}),$$

where $\varepsilon_i = Y_i - \rho_0(X_i)$.

Following almost the same arguments as in the proof of Theorem 5, we have

$$n^{-1}\sum_{i=1}^{n}(\hat{\alpha}(X_i)-\bar{\alpha}(X_i))^2\varepsilon_i^2 = o_p(1)$$

and

$$n^{-1}\sum_{i=1}^{n}\hat{\alpha}(X_i)^2(\hat{\rho}(X_i)-\rho_0(X_i))^2 = o_p(1).$$

It suffices to show

$$n^{-1}\sum_{i=1}^{n}m(X_i,\hat{\rho}-\rho_0)^2 = o_p(1). \tag{A.26}$$

To do so, we observe that

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}m(X_i,\hat{\rho}-\rho_0)^2 &= n^{-1}\sum_{i=1}^{n}\left(m(X_i,\hat{\rho}-\rho_n)+m(X_i,\rho_n-\rho_0)\right)^2 \\ &\le 2n^{-1}\sum_{i=1}^{n}m(X_i,\hat{\rho}-\rho_n)^2 + 2n^{-1}\sum_{i=1}^{n}m(X_i,\rho_n-\rho_0)^2. \end{aligned}$$

By Assumption 9 and $\sup_x|\rho_n(x)-\rho_0(x)| = o(1)$, we have that

$$Em(X_i,\rho_n-\rho_0)^2 \le \left[Ea(X_i)^2\right]\times\sup_x|\rho_n(x)-\rho_0(x)|^2 = o(1).$$

Therefore,

$$n^{-1}\sum_{i=1}^{n}m(X_i,\hat{\rho}-\rho_0)^2 \le 2n^{-1}\sum_{i=1}^{n}m(X_i,\hat{\rho}-\rho_n)^2 + o_p(1).$$

By Assumption 9, we notice that

$$|m(X_i,\hat{\rho}-\rho_n)| = |m(X_i,b)'(\hat{\gamma}-\gamma)| \le \max_j|m(X_i,b_j)|\cdot\|\hat{\gamma}-\gamma\|_1 \le C\|\hat{\gamma}-\gamma\|_1.$$

The above two displays imply

$$n^{-1}\sum_{i=1}^{n}m(X_i,\hat{\rho}-\rho_0)^2 \le 2C^2\|\hat{\gamma}-\gamma\|_1^2 + o_p(1).$$

We define γ_* similar to the definition of π_*. Recall that $\xi_1 > 1/2$. Then by the same argument as Lemmas A4 and A7 with $(\pi,\pi_*,\hat{\pi})$ replaced by $(\gamma,\gamma_*,\hat{\gamma})$, we have $\|\gamma_*-\gamma\|_1 = O_p(\varepsilon_n^{(2\xi_1-1)/(2\xi_1+1)})$ and $\|\hat{\gamma}-\gamma_*\|_1 = O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi_1-1)/(2\xi_1+1)})$. By $\xi_1 > 1/2$ and Assumption 5, we have $\|\hat{\gamma}-\gamma\|_1 = o_p(1)$. By the above display, we have proved (A.26). *Q.E.D.*

Proof of Corollary 8: We proceed to verify that Assumptions 11 and 12 i) are satisfied so that the result follows by Theorem 7. In Example 2

$$m(X,\rho) = S(U)\rho(U,Z).$$

By hypothesis and iterated expectations

$$E[\bar{\alpha}(X)^2] = E[E[f_{D|Z}(D|Z)^{-2}S(D)^2\omega(D)^2|Z]] = E[\int f_{D|Z}(U|Z)^{-1}S(U)^2\omega(U)dU]$$
$$\leq CE[\int f_{D|Z}(U|Z)^{-1}\omega(U)du] < \infty.$$

Note that for $\rho \in \mathcal{B}$ by multiplying and dividing by $f(D|Z)$

$$E[m(X,\rho)] = E[\int S(u)\rho(u,Z)\omega(u)du] = E[E[f(D|Z)^{-1}S(D)\omega(D)\rho(D,Z)|Z]]$$
$$= E[\bar{\alpha}(X)\rho(X)] = E[\bar{\alpha}(X)\rho(X)],\ \bar{\alpha}(X) = proj(\bar{\alpha}|\mathcal{B})(X),$$

giving the first condition of Assumption 11. Also $|m(X,b_j)| = |S(U)|\,|b_j(U,Z)| \leq C$ by $S(u)$ and $b_j(x)$ bounded, giving the second condition of Assumption 11.

To show that Assumption 12 i) is satisfied note that $f(U|Z)^{-1}\omega(U)$ is finite with probability one for the joint distribution of (U,Z), so that the support of (U,Z) is a subset of the support of (D,Z). Then by $S(u)$ bounded over the real line,

$$|m(W,\rho)| = |S(U)\rho(U,Z)| \leq C\sup_{x\in X}|\rho(x)|\,.\ Q.E.D.$$

Proof of Corollary 9: We proceed to verify that Assumptions 11 and 12 i) are satisfied so that the result follows by Theorem 7. In Example 3,

$$m(X,\rho) = \rho(1,Z) - \rho(0,Z).$$

Define $\alpha_0(X) := \pi_0(1,Z)^{-1}D - \pi_0(0,Z)^{-1}(1-D)$. By hypothesis and iterated expectations

$$E[\bar{\alpha}(X)^2] \leq 2E[\pi_0(1,Z)^{-2}D^2] + 2E[\pi_0(0,Z)^{-2}(1-D)^2]$$
$$= 2E[\pi_0(1,Z)^{-2}D] + 2E[\pi_0(0,Z)^{-2}(1-D)] = 2E[\pi_0(1,Z)^{-1}] + 2E[\pi_0(0,Z)^{-1}]$$
$$\leq 4E[\pi_0(1,Z)^{-1}\pi_0(0,Z)^{-1}] < \infty.$$

Note that for $\rho \in \mathcal{B}$ by $E[\pi_0(d,Z)^{-1}1(D=d)|Z] = 1$,

$$E[m(X,\rho)] = E[\rho(1,Z)] - E[\rho(0,Z)] = E[\frac{1(D=1)}{\pi_0(1,Z)}\rho(1,Z)] - E[\frac{1(D=0)}{\pi_0(0,Z)}\rho(0,Z)]$$
$$= E[\frac{1(D=1)}{\pi_0(1,Z)}\rho(D,Z)] - E[\frac{1(D=0)}{\pi_0(0,Z)}\rho(D,Z)] = E[\alpha_0(X)\rho(X)]$$
$$= E[\bar{\alpha}(X)\rho(X)],\ \alpha_0(X) = proj(\bar{\alpha}|\mathcal{B})(X).$$

giving the first condition of Assumption 11. Also $|m(X,b_j)| \leq |b_j(1,Z)| + |b_j(0,Z)| \leq C$ and $b_j(x)$ bounded, giving the second condition of Assumption 11.

To show that Assumption 12 i) is satisfied note that $\pi_0(1,Z) > 0$ and $\pi_0(0,Z) > 0$ with probability one, so that the support of $(1,Z)$ and $(0,Z)$ are subsets of the support of (D,Z). Then $|m(X,\rho)| \leq |\rho(1,Z)| + |\rho(0,Z)| \leq 2\sup_{x\in X}|\rho(x)|$. *Q.E.D.*

Proof of Theorem 10: In the proof of Theorems 5 and 7 the only place that $E[Y|X] \in \mathcal{B}$ is used is in showing that $T_3 = o_p(n^{-1/2})$. The expansion in equation (A.17) continues to hold and $T_1 = o_p(n^{-1/2})$ and $T_2 = o_p(n^{-1/2})$ are satisfied as in the proof of Theorem 4. Note that

$$E[\{Y - \rho_n(X)\}^2] \leq 2E[Y^2] + 2E[\rho_n(X)^2] \leq C.$$

Also it follows as in the proof of Theorem 4 that $(\hat{\pi} - \pi)'\hat{\Sigma}(\hat{\pi} - \pi) = o_p(1)$. Then by the Cauchy-Schwartz and Markov inequalities and $\varepsilon_i = Y_i - \rho_n(X_i)$,

$$|T_3| \leq \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} \left\{(\hat{\pi} - \pi)'\hat{\Sigma}(\hat{\pi} - \pi)\right\}^{1/2} = O_p(1)o_p(1) = o_p(1).$$

It then follows as in the proof of Theorem 6 and 7 that

$$\hat{\theta} = \theta_n + \frac{1}{n}\sum_{i=1}^n \psi_n(W_i) + o_p(1) = \bar{\theta} + \frac{1}{n}\sum_{i=1}^n \psi_0(W_i) + o_p(1),$$

so $\hat{\theta} = \bar{\theta} + o_p(1)$ by the law of large numbers, for $\bar{\theta} = E[m(W,\bar{\rho})] = E[\bar{\alpha}(X)\bar{\rho}(X)]$, giving the first conclusion

Because $\bar{\rho}(X)$ and $\bar{\alpha}(X)$ are projections of $\rho_0(X)$ and $\alpha_0(X)$ on $\mathcal{B}$ respectively, we have

$$\theta_0 = E[\alpha_0(X)\bar{\rho}(X)] = E[\bar{\alpha}(X)\rho_0(X)],$$

so the second conclusion follows from $\bar{\theta} = E[\bar{\alpha}(X)\bar{\rho}(X)]$. *Q.E.D.*

Proof of Theorem 11: In the proof of the previous results the only place that $\rho_0(X) = E[Y|X]$ and $m(W,\rho)$ depending only on X were required was in the proof that $T_3 = o_p(n^{-1/2})$. Therefore the conclusion will follow from the proof that $T_3 = o_p(n^{-1/2})$ under the hypotheses of this Theorem.

Note that $T_3 = \hat{U}'(\hat{\pi} - \pi)$ for $\hat{U} = \hat{\mu} - \hat{\Sigma}\gamma$ as in the proof of Theorem 4. It follows similarly to equation (2.2) that $\left\|\hat{U}\right\|_\infty = O_p(\varepsilon_n)$. Also it follows similarly to $\|\hat{\gamma} - \gamma\|_1 = O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi_1-1)/(2\xi_1+1)})$, as shown in the proof of Theorem 4, that

$$\|\hat{\pi} - \pi\|_1 = O_p((r/\varepsilon_n)\varepsilon_n^{(2\xi_2-1)/(2\xi_2+1)}).$$

Therefore

$$|T_3| = \left|\hat{U}'(\hat{\pi} - \pi)\right| \leq \left\|\hat{U}\right\|_\infty \|\hat{\pi} - \pi\|_1 = O_p((r/\varepsilon_n)\varepsilon_n^{4\xi_2/(2\xi_2+1)}) = o_p(n^{-1/2}).\ Q.E.D.$$

A.1 Appendix B: Some General Lemmas

LEMMA B1: *For any $a \in \mathbb{R}^p$ such that $\|a - b_s\|_2 \leq Cs^{-r}$ for any $s \geq 0$, where $C, r > 0$ are constants and $b_s = \arg\min_{\|v\|_0 \leq s} \|a - v\|_2$. If $r > 1/2$ and $s \geq 2$ then $\|a - b_s\|_1 \leq Ds^{1/2-r}$, where $D > 0$ is a constant depending only on C and r.*

Proof: Without loss of generality, we assume that $|a_1| \geq |a_2| \geq \cdots \geq |a_p| \geq 0$. Then clearly, $a - b_k = (0, 0, \ldots, 0, a_{k+1}, a_{k+2}, \ldots, a_p)\prime$ for $k \geq 0$. By assumption, we have that for any $k \geq 0$,

$$\sum_{j=k+1}^{p} a_j^2 \leq C^2 k^{-2r}. \tag{A.27}$$

Let $g \in \mathbb{N}$ be defined as $2^g < p/s \leq 2^{g+1}$. With a slight abuse of notation, we extend a to be a $2^{g+1}s$-dimensional vector with $a_j = 0$ for $j > p$. Then we have that

$$\sum_{j=s+1}^{p} a_j = \sum_{m=0}^{g} \sum_{j=2^m s+1}^{2^{m+1}s} a_j \leq \sum_{m=0}^{g} \sqrt{2^m s \sum_{j=2^m s+1}^{2^{m+1}s} a_j^2} \leq \sum_{m=0}^{g} \sqrt{2^m s \sum_{j=2^m s+1}^{p} a_j^2}$$
$$\overset{\text{(i)}}{\leq} \sum_{m=0}^{g} \sqrt{2^m s C^2 (2^m s)^{-2r}} = Cs^{1/2-r} \sum_{m=0}^{g} \left(2^{1/2-r}\right)^m < Cs^{1/2-r} \sum_{m=0}^{\infty} \left(2^{1/2-r}\right)^m \overset{\text{(ii)}}{=} C \frac{1}{1 - 2^{1/2-r}} s^{1/2-r},$$

where (i) follows by () applied to $k = 2^m s$ and (ii) follows by the fact that $2^{1/2-r} < 1$ (since $r > 1/2$). Q.E.D

Let $X_i = (X_{i,1}, ..., X_{i,p})' \in \mathbb{R}^p$ and $X_{i,0}$ be a scalar, with all random variables allowed to depend on n.

LEMMA B2: *If there is C such that $\max_{1 \leq j \leq p} |X_{i,j}| \leq C$ and $E[X_{0,i}^2] \leq C$ then for $D_i = X_i X_{0,i} - E[X_i X_{0,i}]$ and $\bar{D} = \sum_{i=1}^n D_i/n$,*

$$\left\|\bar{D}\right\|_\infty = O_p(\sqrt{\ln(p)/n}).$$

Proof: We prove this result using symmetrization. Note that

$$E[D_{i,j}^2] \leq E[X_{i,j}^2 X_{i,0}^2] \leq CE[X_{i,0}^2] \leq C.$$

Let $\varepsilon_1, ..., \varepsilon_n$ be i.i.d Rademacher random variables independent of X_i for for all observations, i.e., $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$. Define the symmetrized quantity $W_{*,j} = \sum_{i=1}^n D_{i,j}\varepsilon_i$ Since ε_i is sub-Gaussian (due to $\varepsilon_i \in \{-1, 1\}$), there exists a constant $\kappa > 0$ such that for any $t \in \mathbb{R}$,

$E\exp(t\varepsilon_i) \leq \exp(\kappa t^2)$. By Hoeffding's lemma, we can simply take $\kappa = 1/2$. Since $\{\varepsilon_i\}_{i=1}^n$ is independent of X we have

$$E\left[\exp(tW_{*,j})|X\right] = E\left[\prod_{i=1}^n \exp[tD_{i,j}\varepsilon_i] \mid X\right] = \prod_{i=1}^n E\left[\exp[tD_{i,j}\varepsilon_i] \mid X\right]$$
$$\leq \prod_{i=1}^n \exp\left(t^2 D_{i,j}^2/2\right) = \exp\left(t^2 \sum_{i=1}^n D_{i,j}^2/2\right).$$

Similarly, apply the same argument to $-W_{*,j}$ to obtain $E\left[\exp(-tW_{*,j}) \mid X\right] \leq \exp\left(t^2 \sum_{i=1}^n D_{i,j}^2/2\right)$. Since $\exp(t|W_{*,j}|) \leq \exp(tW_{*,j}) + \exp(-tW_{*,j})$, we have

$$E\left[\exp(t|W_{*,j}|) \mid X\right] \leq 2\exp\left(t^2 \sum_{i=1}^n D_{i,j}^2/2\right).$$

Next let $z > 0$ be a non-random quantity to be chosen later and $\|W_*\|_\infty = \max_{1\leq j\leq p} |W_{*,j}|$. By Lemma 2.3.7 of van der Vaart and Wellner (1996) we have

$$(1-\beta_n(z))\, P\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^n D_{i,j}\right| > z\right) \leq 2P\left(\|W_*\|_\infty > z/4\right), \tag{A.28}$$

where

$$\beta_n(z) = 1 - 4z^{-2} n \max_{1\leq j\leq p} E[D_{i,j}^2] \geq 1 - Cz^{-2}n.$$

Note that $|E[D_{i,j}]|^2 \leq E[D_{i,j}^2] \leq C$, so that

$$D_{i,j}^2 \leq C(X_{i0}^2 + 1).$$

For any $M_n \longrightarrow \infty$ let $\mathcal{A} = \left\{n^{-1}\sum_{i=1}^n X_{i,0}^2 > M_n\right\}$. Since $E[X_{i,0}^2]$ is bounded uniformly in n the Markov inequality, for any $M_n \longrightarrow \infty$ we have $P(\mathcal{A}) = o(1)$. On the event $\mathcal{A}^c$ we have $D_{ij}^2 \leq C(M_n + 1) := \tilde{M}_n$ for all j, so that

$$E\left[\exp\left(t\|W_*\|_\infty\right) \mid X\right] = E\left[\exp\left(t \max_{1\leq j\leq p} |W_{*,j}|\right) \mid X\right] \leq \sum_{j=1}^p E\left(\exp(t|W_{*,j}|) \mid X\right)$$
$$\leq 2\sum_{j=1}^p \exp\left(t^2 \sum_{i=1}^n D_{i,j}^2/2\right) \leq 2p\exp\left(\tilde{M}_n t^2 n\right),$$

Let $t > 0$ be a non-random quantity to be chosen. By the Markov inequality we have

$$P\left(\|W_*\|_\infty > z/4 \mid X\right)\mathbf{1}_{\mathcal{A}^c} \leq P\left(\exp(t\|W_*\|_\infty) > \exp(tz/4) \mid X\right)\mathbf{1}_{\mathcal{A}^c}$$
$$\leq \exp(-tz/4) E\left[\exp(t\|W_*\|_\infty) \mid X\right] \cdot \mathbf{1}_{\mathcal{A}^c}$$
$$\leq \exp(-tz/4)\cdot 2p\exp\left(\tilde{M}_n t^2 n\right) = \exp\left(-\frac{1}{4}tz + \ln(2p) + \tilde{M}_n t^2 n\right).$$

Now choose $t = z/[8n\tilde{M}_n]$ to obtain

$$P\left(\|W_*\|_\infty > z/4 \mid X\right)\mathbf{1}_{\mathcal{A}^c} \leq \exp\left(-z^2/(nK_n) + \ln(2p)\right),$$

where $K_n = 64\tilde{M}_n$. Therefore we have

$$\begin{aligned} P\left(\|W_*\|_\infty > z/4 \mid X\right) &= P\left(\|W_*\|_\infty > z/4 \mid X\right)\left(\mathbf{1}_{\mathcal{A}^c} + \mathbf{1}_{\mathcal{A}}\right) \\ &\leq \exp\left(-z^2/(nK_n) + \ln(2p)\right) + \mathbf{1}_{\mathcal{A}}. \end{aligned}$$

Then by iterated expectations and $E[\mathbf{1}_{\mathcal{A}}] = \Pr(\mathcal{A}) \longrightarrow 0$,

$$P\left(\|W_*\|_\infty > z/4\right) \leq \exp\left(-z^2/(nK_n) + \ln(2p)\right) + o(1).$$

Finally choose $z_n = 2\sqrt{K_n n \ln(2p)}$ and note that

$$\beta_n(z) := 1 - Cz^{-2}n = 1 - \frac{C}{4K_n \ln(2p)} \longrightarrow 1.$$

Define $B_n := \sqrt{n/\ln(p)}\left\|\bar{D}\right\|_\infty$. Then by equation (A.28) we have for large enough n

$$\begin{aligned} \Pr(B_n \geq 2\sqrt{K_n \frac{\ln(2p)}{\ln(p)}}) &= \Pr(\left\|n\bar{D}\right\|_\infty > z_n) \leq CP\left(\|W_*\|_\infty > \frac{z_n}{4}\right) \\ &\leq \exp\left(-z_n^2/(nK_n) + \ln(2p)\right) + o(1) = \exp\left(-4\ln(2p) + \ln(2p)\right) + o(1) \longrightarrow 0 \end{aligned}$$

Note that this equation will hold for any K_n going to infinity and that $\ln(2p)/\ln(p) \longrightarrow 1$, so that $\Pr(B_n \geq k_n) \longrightarrow 0$ for all $k_n \longrightarrow \infty$, i.e. B_n is bounded in probability. *Q.E.D.*