Learning L_2-Continuous Regression Functionals via Regularized Riesz Representers[*]

Victor Chernozhukov[†] *MIT*

Whitney K. Newey[‡] *MIT*

Rahul Singh[§] *MIT*

December 31, 2018

Abstract

Many objects of interest can be expressed as an L_2 continuous functional of a regression, including average treatment effects, economic average consumer surplus, expected conditional covariances, and discrete choice parameters that depend on expectations. Debiased machine learning (DML) of these objects requires learning a Riesz representer (RR). We provide here Lasso and Dantzig regularized learners of the RR and corresponding debiased learners of affine and nonlinear functionals. We give convergence rates for the regularized RR and conditions for root-n consistency and asymptotic normality of the functionial learners. We allow for a wide variety of regression learners that can converge at relatively slow rates. We give DML estimators and results for nonlinear functionals in addition to affine functionals.

Keywords: Regression functionals, Riesz representers, Lasso, Dantzig, debiased machine learning.

[*] This research was supported by NSF grant 1757140. J. Robins and Y. Zhu provided helpful comments.

[†] 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.

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

1 Introduction

Many statistical objects of interest can be expressed as an L_2 (mean square) continuous functional of a conditional expectation (regression). Examples of affine regression functionals include average treatment effects, policy effects, economic average consumer surplus, and the expected conditional covariance of two random variables. Nonlinear functionals include discrete choice models that depend on regressions. Often the regression may be high dimensional, depending on many variables. There may be many covariates for a treatment effect when treatment was assigned in a complicated way. There are often many prices and covariates in the economic demand for some commodity. This variety of important examples motivates the learning of L_2 continuous regression functionals.

Plugging a machine learner into a functional of interest can be badly biased; e.g. see Chernozhukov et al. (2018). We use debiased/double machine learning (DML, Chernozhukov et al. 2018), based on estimating equations that have zero derivative with respect to each nonparametric component. Such debiased estimating equations are sometimes referred to as Neyman orthogonal. They can be constructed by adding the influence function of a functional of the regression learner limit. We also debias using sample splitting (Bickel, 1982, Schick, 1986), where we average over data observations different than those used by the nonparametric learners. The resulting estimators of regression functionals have second order remainders which leads to root-n consistency under regularity conditions we give.

The influence function of an L_2 continuous functional of a regression limit is the product of the regression residual with the Riesz representer (RR) of the functional derivative, as shown in Newey (1994). Therefore, DML of regression functionals requires a machine learner of the RR. We provide here ℓ_1 regularized RR learners: Lasso and Dantzig selector. These automatically learn the RR from the empirical analog of equations that implicitly characterize it, without needing to know its form. We derive convergence rates for these regularized RR's and give conditions sufficient for root-n consistency and asymptotic normality of the DML estimator. DML also requires a regression learner for its construction. We allow for a variety of regression learners, requiring only a sufficiently fast L_2 convergence rate for the regression. We give a consistent estimator of the asymptotic variance. Results are given for nonlinear functionals as well as for affine ones. We impose only L_2 convergence conditions on the RR and regression learners, so that our results apply to many possible machine learners.

Debiasing via DML 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). 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), Neykov et al. (2015), Ren et al. (2015), Jankova and van de Geer (2015, 2016a,b), Bradic and Kolar (2017); Zhu and Bradic (2018). These two debiasing approaches bear some resemblance when the functional of interest is a coefficient of a partially linear model (as discussed in Chernozhukov et al., 2018), but are quite different for other functionals. The differences between these methods seem analogous to the difference between nonparametric estimation and root-n consistent functional estimation in the semiparametric literature (see Bickel, Klassen, Ritov, and Wellner, 1993 and Van der Vaart, 1991). Inference for a nonparametric regression requires bias correcting or undersmoothing the regression estimator while root-n consistent functional estimation can be based on learners that are not debiased or undersmoothed (see Newey 1994 for series regression). Similarly, DML based inference does not require the use of debiased learners. As we show, any regression learner having a fast enough convergence rate will suffice when combined with the RR learners given here.

The functionals we consider are different than those analyzed in Cai and Guo (2017). We consider nonlinear functionals as well as linear functionals where the linear combination coefficients are estimated, neither of which is allowed for in Cai and Guo (2017). Also the L_2 continuity of the linear functionals provides additional structure that we exploit, involving the RR, which is not exploited 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 et al. (2016). Here we provide DML learners via regularized RR, which are relatively simple to implement and analyze and directly target functionals of interest.

L_2 continuity does place us squarely in a semiparametric setting where root-n consistent efficient semiparametric estimation of the object of interest is possible under sufficient regularity conditions; see Jankova and Van De Geer (2016a). Our results apply to different objects than considered by Ning and Liu (2017), who considered machine learning of the efficient score for a parameter of an explicit semiparametric form for the distribution of the data. Unlike Ning and Liu (2017), we do not work with an explicit semiparametric form for the distribution of the data. Instead we focus on learning functionals of a nonparametric regression. Our estimators can be thought of as being based on DML of a functional of interest rather than the efficient score for a parameter of interest in an explicit form of a semiparametric model. There are many interesting examples, including those we have given, where learning via DML is more convenient and natural than embedding the functional of interest in a large, explicit semiparametric form.

We build on previous work on debiased estimating equations constructed by adding an influence function. Hasminskii and Ibragimov (1979) and Bickel and Ritov (1988) suggested such estimators for functionals of a density. Doubly robust estimating equations as in Robins, Rotnitzky, and Zhao (1995) and Robins and Rotnitzky (1995) have this structure. Newey, Hsieh, and Robins (1998, 2004) and Robins et al. (2008) further developed theory. For an affine functional the doubly robust learner we consider is given in Chernozhukov et al. (2016). We make use of simple and general regularity conditions in Chernozhukov et al. (2016) that only require L_2 convergence of nonparametric learners.

The RR learners we consider are linear in a dictionary of functions. Such RR learners were previously used in Newey (1994) for asymptotic variance estimation and in Robins et al. (2007) for estimation of the inverse of the propensity score with missing data. Recently Newey and Robins (2017) considered such RR learning in efficient semiparametric estimation of linear regression functionals with low dimensional regressors. Hirshberg and Wager (2018) gave different RR estimators when the regression is restricted to a Donsker class. None of these works are about machine learning.

The Athey, Imbens, and Wager (2018) learner of the average treatment effect is based on a specific regression learner and on approximate balancing weights when the regression is linear and sparse. Our estimator allows for a wide variety of regression learners and does not restrict the regression to be sparse or linear. We do this via regularized RR learning that can also be interpreted as learning of balancing weights or inverse propensity scores, as further discussed in Section 4.

Zhu and Bradic (2017) showed that it is possible to attain root-n consistency for the coefficients of a partially linear model when the regression function is dense. Our results apply to a wide class of affine and nonlinear functionals and similarly allow the regression learner to converge at relatively slow rates.

Chernozhukov, Newey, and Robins (2018) have previously given the Dantzig learner of the RR. We innovate here by allowing the functional to depend on data other than the regressors, by giving a Lasso learner of the RR, by deriving convergence rates for both Lasso and Dantzig as learners of the true RR rather than a sparse approximation to it, by allowing for a general regression learner rather than just Dantzig, and by providing learners for nonlinear functionals. These results are innovative relative to other previous work in the ways described in the previous paragraphs.

In Section 2 we describe the objects we are interested in, their DML estimators, give a Lasso learner of the RR, and an estimator of the asymptotic variance for DML. Section 3 derives L_2 convergence rates of Lasso and Dantzig RR learners. Section 4 gives conditions for root-n consistency and asymptotic normality of DML and consistency of the asymptotic variance, in general and for the examples. Section 5 shows how to construct Lasso and Dantzig RR learners for nonlinear functionals and gives large sample inference results for the DML estimator and its asymptotic variance estimator.

2 Learning Affine Functionals

For expositional purposes we first consider objects of interest that are L_2 continuous affine functionals of a conditional expectation. To describe such an object let W denote a data observation and consider a subvector $(Y, X')'$ where Y is a scalar outcome with finite second moment and the covariate vector X that takes values $x \in \mathcal{X}$, a Borel subset of $\mathbb{R}^d$. Denote the conditional expectation of Y given X as

$$\gamma_0(x) = \mathrm{E}[Y \mid X = x].$$

Let $m(w, \gamma)$ denote an affine functional of a possible conditional expectation function $\gamma : X \longrightarrow \mathbb{R}$ that depends on the data observation W. The object of interest is

$$\theta_0 = \mathrm{E}[m(W, \gamma_0)]. \tag{2.1}$$

We focus on functionals where $E[m(W, \gamma) - m(W, 0)]$ is a mean square continuous linear functional of γ. This continuity property is equivalent to the semiparametric variance bound for θ_0 being finite, as discussed in Newey (1994). In this case the Riesz representation theorem implies existence of $\alpha_0(x)$ with $E[\alpha_0(X)^2]$ finite and

$$E[m(W, \gamma) - m(W, 0)] = E[\alpha_0(X)\gamma(X)] \tag{2.2}$$

for all $\gamma(x)$ with $E[\gamma(X)^2]$ finite. We refer to $\alpha_0(x)$ as the RR.

There are many important examples of this type of object. One is the average treatment effect. Here $X = (D, Z)$ and $\gamma_0(x) = \gamma_0(d, z)$, where $D \in \{0, 1\}$ is the indicator of the receipt of the treatment and Z are covariates. The object of interest is

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)].$$

When the treatment effect is mean independent of the treatment D conditional on covariates Z then θ_0 is the average treatment effect, Rosenbaum and Rubin (1983). Here $m(w, \gamma) = \gamma(1, z) - \gamma(0, z)$ and the RR is $\alpha_0(x) = d/\pi_0(z) - (1 - d)/[1 - \pi_0(z)]$ where $\pi_0(z)$ is the propensity score $\pi_0(z) = \Pr(D = 1|Z = z)$. Thus $E[m(W, \gamma)]$ is mean square continuous when $E[1/\pi_0(Z)] < \infty$ and $E[1/\{1 - \pi_0(Z)\}] < \infty$.

Another interesting example is the average effect of changing the conditioning variables according to the map $x \longrightarrow t(x)$. The object of interest is

$$\theta_0 = E[\gamma_0(t(X)) - \gamma_0(X)] = \int \gamma_0(x) dF_t(dx) - E[Y],$$

where F_t denotes the CDF of $t(X)$. The object θ_0 is the average policy effect of a counterfactual change of covariate values similar to Stock (1989). Here $m(w, \gamma) = \gamma(t(x)) - y$ and the RR is

$\alpha_0(x) = f_t(x)/f_0(x)$ where $f_0(x)$ is the pdf of X and $f_t(x)$ is the pdf of $t(X)$. $E[m(W,\gamma)]$ is mean square continuous if $E[\alpha_0(X)^2] = \int f_0(x)^{-1} f_t(x)^2 dx < \infty$.

A third object of interest is a bound on average consumer surplus for economic demand functions. Here Y is the share of income spent on a commodity and $X = (P_1, Z)$, where P_1 is the price of the commodity and Z includes income Z_1, prices of other goods, and other observable variables affecting utility. Let $\check{p}_1 < \bar{p}_1$ be lower and upper prices over which the price of the commodity can change, κ a bound on the income effect, and $\omega(z)$ some weight function. The object of interest is

$$\theta_0 = E[\omega(Z)\int_{\check{p}_1}^{\bar{p}_1} (\frac{Z_1}{u})\gamma_0(u,Z)\exp(-\kappa[u-\check{p}_1])du],$$

where Z_1 is income and u is a variable of integration. When individual heterogeneity in consumer preferences is independent of X and κ is a lower (upper) bound on the derivative of consumption with respect to income across all individuals then θ_0 is an upper (lower) bound on the weighted average over consumers of exact consumer surplus (equivalent variation) for a change in the price of the first good from $\check{p}_1$ to $\bar{p}_1$; see Hausman and Newey (2016). Here $m(w,\gamma) = \omega(z)\int_{\check{p}_1}^{\bar{p}_1}(z_1/u)\gamma_0(u,z)\exp(-\kappa[u-\check{p}_1])du$ and the RR is

$$\alpha_0(x) = f_0(p_1|z)^{-1}\omega(z)1(\check{p}_1 < p_1 < \bar{p}_1)(z_1/p_1)\exp(-\kappa[p_1-\check{p}_1]),$$

where $f_0(p_1|z)$ is the conditional pdf of P_1 given Z.

A fourth example is the average conditional covariance between Y and some other variable, say W_1. In this case the object of interest is

$$\theta_0 = E[Cov(Y, W_1|X)] = E[W_1\{Y - \gamma_0(X)\}].$$

This object is useful in the analysis of covariance while controlling for regressors X and is an important component in the coefficient β_0 of W_1 for a partially linear regression of Y on W_1 and unknown functions of x. This object differs from the previous three examples in $m(w,\gamma)$ depending on w other than the regressors x. Here $m(w,\gamma) = w_1\{y - \gamma(x)\}$ and the RR is $\alpha_0(x) = -E[W_1|X=x]$.

DML of θ_0 can be carried out using the doubly robust moment function

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

given in Chernozhukov et al. (2016). This function has the doubly robust property that

$$0 = E[\psi(W,\theta_0,\gamma_0,\alpha)] = E[\psi(W,\theta_0,\gamma,\alpha_0)],$$

for all γ and α. Consequently, $\psi(w,\theta,\gamma,\alpha)$ is debiased in that any functional derivative of $E[\psi(W,\theta_0,\gamma_0,\alpha)]$ with respect to α and of $E[\psi(W,\theta_0,\gamma,\alpha_0)]$ with respect to γ is zero. Therefore a DML learner $\hat{\theta}$ can be constructed from machine learning estimators $\hat{\gamma}$ and $\hat{\alpha}$ by plugging these into the moment function $\psi(w,\theta,\gamma,\alpha)$ in place of γ and α and solving for $\hat{\theta}$ from setting the sample moment of $\psi(w,\theta,\hat{\gamma},\hat{\alpha})$ to zero.

To help avoid potentially severe finite sample bias and to avoid regularity conditions based on $\hat{\gamma}$ and $\hat{\alpha}$ being in a Donsker class, which machine learning estimators are usually not, we also use sample splitting. We construct $\hat{\gamma}$ and $\hat{\alpha}$ from observations that are not being averaged over. Let the data be W_i, $(i = 1, ..., n)$, assumed to be i.i.d.. Let I_ℓ, $(\ell = 1, ..., L)$ be a partition of the observation index set $\{1, ..., n\}$ into L distinct subsets of about equal size. Let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ be estimators constructed from the observations that are *not* in I_ℓ. We construct the estimator $\hat{\theta}$ by setting the sample average of $\psi(W_i,\theta,\hat{\gamma}_\ell,\hat{\alpha}_\ell)$ to zero and solving for θ. This estimator has the explicit form

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}. \tag{2.3}$$

A variety of regression learners $\hat{\gamma}_\ell$ of the nonparametric regression $E[Y|X]$ could be used here, as discussed in the Introduction. We also need an estimator $\hat{\alpha}_\ell$ to construct $\hat{\theta}$. We give here Lasso and Dantzig learners $\hat{\alpha}_\ell$. These learners make use of a $p \times 1$ dictionary of functions $b(x)$ where p can be much bigger than n. The learners take the form

$$\hat{\alpha}(x) = b(x)'\hat{\rho}, \tag{2.4}$$

where $\hat{\rho}$ is a vector of estimated coefficients. For notational convenience we drop the ℓ subscript, with the understanding that the description which follows should be applied only to the observations not in I_ℓ for each ℓ. The learners for α_0 are based on the fact that the Riesz representation implies that for $m(w,b) = (m(w,b_1), ..., m(w,b_p))'$,

$$M = E[m(W,b) - m(W,0)] = E[\alpha_0(X)b(X)].$$

Here we see that the cross moments M between the true, unknown RR $\alpha_0(x)$ and the dictionary $b(x)$ are equal to the expectation of a known vector of functions $m(w,b)-m(w,0)$. Consequently an unbiased estimator of $M = E[\alpha_0(X)b(X)]$ can be constructed as

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n}\{m(W_i,b) - m(W_i,0)\}.$$

Likewise an unbiased estimator of $G = E[b(X)b(X)']$ can be constructed as

$$\hat{G} = \frac{1}{n}\sum_{i=1}^{n}\{b(X_i)b(X_i)'\}.$$

The estimator $\hat{M}$ is analogous to $\sum_{i=1}^{n} Y_i b(X_i)/n$ in Lasso and Dantzig regression. Just as $\sum_{i=1}^{n} Y_i b(X_i)/n$ is an unbiased estimator of $E[\gamma_0(X)b(X)]$ so is $\hat{M}$ an unbiased estimator of M.

Minimum distance versions of Lasso and Dantzig can be constructed by replacing $\sum_{i=1}^{n} Y_i b(X_i)/n$ in the Lasso objective function and Dantzig constraints by $\hat{M}$. Doing this for Lasso, while dropping $\sum_{i=1}^{n} Y_i^2/n$ term in the Lasso objective, gives an estimator

$$\hat{\rho}_L = \arg\min_{\rho}\{-2\hat{M}'\rho + \rho'\hat{G}\rho + 2r_L|\rho|_1\}. \tag{2.5}$$

The objective function here is a ℓ_1 penalized approximation to the least squares regression of $\alpha_0(x)$ on $b(x)$, where $2r_L$ is the penalty. Making the analogous replacement in the constraints of the Dantzig selector gives a Dantzig estimator

$$\hat{\rho}_D = \arg\min_{\rho} |\rho|_1 \, s.t. |\hat{M} - \hat{G}\rho|_\infty \leq \lambda_D, \tag{2.6}$$

where $\lambda_D > 0$ is the slackness size. These two minimization problems can be thought of as minimum distance versions of Lasso and Dantzig respectively.

Either of these $\hat{\rho}$ may be used in equation (2.4) to form an estimator of the RR. This estimator of the RR may then be substituted in equation (2.3) along with a machine learning regression estimator to construct an estimator of the object of interest. We derive the properties of $\hat{\theta}$ under weak conditions that only require a relatively slow L_2 convergence rate for $\hat{\gamma}$. Our results on Lasso and Dantzig minimum distance can be applied to show that these produce fast enough convergence rates without assuming sparseness of the ℓ_1 regularized approximation to the true regression.

It is interesting to note that the estimator $b(x)'\hat{\rho}$ of the RR does not require any knowledge of the form of $\alpha_0(x)$. In particular it does not depend on plugging in nonparametric estimates of components of $\alpha_0(x)$. Instead it is a linear in $b(x)$ estimator that uses $\hat{M}$ as an estimator of M in an ℓ_1 regularized least squares approximation of the least squares projection of $\alpha_0(x)$ on $b(x)$.

In the next Section we will derive convergence rates for the Lasso and Dantzig estimators of the RR and in Section 4 formulate sufficient conditions for root-n consistency and asymptotic normality of $\hat{\theta}$ from equation (2.3). For asymptotic inference we also need a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\theta}-\theta_0)$. We can construct such a variance estimator by plugging in $\hat{\gamma}$ and $\hat{\alpha}$ into the influence function formula. Let

$$\hat{\psi}_i = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)], \; i \in I_\ell, \; (\ell = 1, ..., L).$$

An estimator of the asymptotic variance is then the sample variance $\hat{V}$ of $\hat{\psi}_i$ given by

$$\hat{V} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\hat{\psi}_i - \bar{\psi})^2, \; \bar{\psi} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i. \tag{2.7}$$

To summarize, based on an estimated RR we have given a doubly robust machine learning estimator of a linear functional of a nonparametric regression. We have given Lasso and Dantzig estimators of the RR that are linear in approximating functions. We have also given an estimator of the asymptotic variance of $\sqrt{n}(\hat{\theta}-\theta_0)$.

3 Properties of Lasso and Dantzig Minimum Distance

In this Section we derive L_2 convergence rates for Lasso and Dantzig minimum distance estimators. We apply these result to obtain rates for regularized estimators of RRs. We begin with some conditions. We make a standard assumption concerning the dictionary $b(x)$ of approximating functions:

Assumption 1: *There is B_n^b such that with probability one,*

$$max_{1\leq j\leq p}|b_j(X)|\leq B_n^b.$$

As usual this condition implies that

$$|\hat{G}-G|_\infty=O_p(\varepsilon_n^G),\ \varepsilon_n^G=(B_n^b)^2\sqrt{\frac{\ln(p)}{n}}.$$

The rates of convergence of the RR learner will depend on the ε_n^G. Leading cases have B_n^b not depending on n so that $\varepsilon_n^G=\sqrt{\ln(p)/n}$. The RR rates will also depend on the convergence rate for $|\hat{M}-M|_\infty$. Here we impose a general condition in order to cover nonlinear functionals and additional cases.

Assumption 2: *There is ε_n^M such that*

$$|\hat{M}-M|_\infty\leq O_p(\varepsilon_n^M),$$

This condition has the flexibility to be applied to various cases, including nonlinear functionals as described in Section 5. In what follows we will give the form of ε_n^M in specific settings. When $\hat{M}$ is a sample average of functions that are bounded uniformly in n then $\varepsilon_n^M=\sqrt{\ln(p)/n}$.

We also explicitly treat the bias in approximating $\alpha_0(x)$ by a linear combination of the dictionary $b(x)$. We consider two types of bias conditions. The first type does not rely on any sparsity conditions.

Assumption 3: *There is ρ_n such that $\|\alpha_0-b'\rho_n\|^2=O(\max\{\varepsilon_n^G,\varepsilon_n^M\})$.*

Sparsity plays no role in this condition. Assumption 3 is clearly satisfied with $\|\alpha_0-b'\rho_n\|^2=0$ in the no bias case where $\alpha_0(x)$ equals a linear combination of $b(x)$. When there is ρ_n such that

$\|\alpha_0 - b'\rho_n\|^2$ shrinks faster than some power of p then this condition will be satisfied when p grows faster than a high enough power of n. These conditions are sufficient to obtain a convergence rate for the Lasso and Dantzig RR's. Let $B_n = |\rho_n|_1$ for ρ_n from Assumption 3.

Theorem 1: *If Assumptions 1 - 3 are satisfied then for any r_L such that $\varepsilon_n^M + \varepsilon_n^G(1+B_n) = o(r_L)$,*

$$\|\alpha_0 - \hat{\alpha}_L\|^2 = O_p((1+B_n)r_L), \;\; |\hat{\rho}_L|_1 = O_p(1+B_n).$$

Also for λ_D such that $\varepsilon_n^M + \varepsilon_n^G(1+B_n) = o(\lambda_D)$,

$$\|\alpha_0 - \hat{\alpha}_D\|^2 = O_p((1+B_n)\lambda_D), \;\; |\hat{\rho}_D|_1 = O_p(1+B_n).$$

The Lasso penalty degree r_L and the Dantzig slackness degree λ_D help determine the convergence rate of the Lasso and Dantzig RR. When $\varepsilon_n^M \leq \varepsilon_n^G$ that rate will be arbitarily close to $(B_n^g)^2(1+B_n)^2\sqrt{\ln(p)/n}$. This rate will be fast enough for root-n consistency of the functional learners when the regression converges fast enough, as discussed in Section 4. A leading case of this result is when B_n^g and B_n are bounded and $\varepsilon_n^M = \sqrt{\ln(\rho)/n}$. The rate for this cawse will be r_L as shown in the following result:

Corollary 2: *If Assumptions 1 and 2 are satisfied with $\varepsilon_n = \varepsilon_n^M = \varepsilon_n^G = \sqrt{\ln(p)/n}$ and there is a $C > 0$ and ρ_n such that $|\rho_n|_1 \leq C$ and $\|\alpha_0 - b'\rho_n\|^2 \leq C\varepsilon_n$ then for any r_L with $\varepsilon_n = o(r_L)$ we have $\|\alpha_0 - \hat{\alpha}_L\|^2 = O_p(r_L)$ and $\|\alpha_0 - \hat{\alpha}_D\|^2 = O_p(r_L)$.*

Here $\hat{\alpha}_L$ and $\hat{\alpha}_D$ converge at an L_2 rate close to $n^{-1/4}$.

Faster convergence rates can be obtained under sparsity conditions. One useful condition is a sparse approximation rate as in the following hypothesis. Let $\varepsilon_n = \max\{\varepsilon_n^G, \varepsilon_n^M\}$.

Assumption 4: *There exists $C > 0$ and $\bar{\rho}$ with $\bar{s}$ nonzero elements such that*

$$\|\alpha_0 - b'\bar{\rho}\|^2 \leq C\bar{s}\varepsilon_n^2$$

Intuitively $\|\alpha_0 - b'\bar{\rho}\|^2$ will be the squared bias from using the linear combination $b'\bar{\rho}$ to approximate α_0. The term $\bar{s}\varepsilon_n^2$ is a variance like term. Assumption 4 specifies $\bar{s}$ so that squared bias is no larger than the variance term. Since the squared bias will generally decrease with $\bar{s}$ for some choice of $\bar{\rho}$ and $\bar{s}\varepsilon_n^2$ increases linearly with $\bar{s}$, such an $\bar{s}$ will generally exist. Specifying $\bar{s}$ to be as small as possible while maintaining Assumption 4 leads to the fastest convergence rates in our results, which essentially set variance equal to squared bias. For example suppose that $\alpha_0(x)$ is sparse, being a linear combination of $\bar{s}$ members of the dictionary $b(x)$. Then by choosing $\bar{\rho}$ equal the coefficients of that linear combination we have $\alpha_0(X) = b(X)'\bar{\rho}$, so that $\|\alpha_0 - b'\bar{\rho}\|^2 = 0$ and Assumption 4 is satisfied.

Another important example is the approximately sparse case where there is $(\tilde{b}_1(x), \tilde{b}_2(x), ...)$ and $C > 0$ with $|\tilde{b}_j(X)| \leq C$ and

$$\alpha_0(x) = \sum_{j=1}^{\infty} \tilde{b}_j(x)\tilde{\rho}_j, \;\; |\tilde{\rho}_j| \leq Cj^{-d}. \tag{3.1}$$

Assume that for each p the vector $(\tilde{b}_1(x), ..., \tilde{b}_{\bar{s}}(x))$ is a subvector of $b(x)$ when $\bar{s}$ is small enough. Choose $\bar{\rho}_k = \tilde{\rho}_j$ if $b_k(x) = \tilde{b}_j(x)$ for some $j \leq \bar{s}$ and otherwise let $\bar{\rho}_k = 0$. Then for some $\bar{C} > 0$,

$$b(X)'\bar{\rho} = \sum_{j=1}^{\bar{s}} \tilde{b}_j(X)\tilde{\rho}_j, \;\; |\alpha_0(X) - b(X)'\bar{\rho}| = \left| \sum_{j=s+1}^{\infty} \tilde{b}_j(X)\tilde{\rho}_j \right| \leq C^2 \sum_{j=\bar{s}+1}^{\infty} j^{-d} \leq \bar{C}(\bar{s})^{-d},$$

so that

$$\|\alpha_0 - b'\bar{\rho}\|^2 \leq \bar{C}^2 \left(\bar{s}\right)^{-2d}.$$

In this case the smallest $\bar{s}$ so that Assumption 4 is satisfied will satisfy $\bar{s} \leq \tilde{C}(\varepsilon_n)^{-2d/(1+2d)}$ for some $\tilde{C} > 0$, so that

$$\bar{s}\varepsilon_n^2 \leq \tilde{C}(\varepsilon_n)^{4d/(1+2d)}.$$

For $\varepsilon_n = \sqrt{\ln(p)/n}$ we will have

$$\bar{s}\varepsilon_n^2 \leq \tilde{C}\left(\frac{\ln p}{n}\right)^{2d/(1+2d)}. \tag{3.2}$$

Here the variance like term is bounded above by a power of $\ln(p)/n$.

To obtain faster rates we also impose sparse eigenvalue conditions. Let $\mathcal{J} = \{1, ..., p\}$, $\mathcal{J}_\rho$ be the subset of $\mathcal{J}$ with $\rho_j \neq 0$, and $\mathcal{J}_\rho^c$ be the complement of $\mathcal{J}_\rho$ in $\mathcal{J}$.

Assumption 5: *G is nonsingular and has largest eigenvalue uniformly bounded in n. Also there is $k > 3$ such such that*

$$\inf_{\{\delta:\delta\neq 0, \sum_{j\in\mathcal{J}_{\rho_L}^c} |\delta_j| \leq k\sum_{j\in\mathcal{J}_{\rho_L}} |\delta_j|\}} \frac{\delta' G\delta}{\sum_{j\in\mathcal{J}_{\rho_L}} \delta_j^2} > 0, \;\; s_D = \sup_{\delta\neq 0, |\rho_D+\delta|_1 \leq |\rho_D|_1} \frac{|\delta|_1^2}{\delta' G\delta} < \infty.$$

The first condition is a population version of a restricted eigenvalue condition of Bickel, Ritov, and Tsybakov (2009). The other condition specifies that the effective dimension s_D is finite. The effective dimension is the reciprocal of the identifiability factors that were introduced in Chernozhukov et al. (2013) as a generalization of the restricted eigenvalue. Let $\bar{B}_n = |\bar{\rho}|_1$ for $\bar{\rho}$ in Assumption 4.

Theorem 3: *If Assumptions 1, 2, 4, and 5 are satisfied and $\varepsilon_n^M + \varepsilon_n^G(1 + \bar{B}_n) = o(r_L)$ then*

$$\|\alpha_0 - \hat{\alpha}_L\|^2 = O_p(\bar{s}r_L^2),$$

and

$$\|\alpha_0 - \hat{\alpha}_D\|^2 = O_p(s_D r_L^2 + \bar{s}\varepsilon_n^2).$$

For example consider again the approximately sparse case with $\varepsilon_n^M = \varepsilon_n^G = \sqrt{\ln(p)/n}$ and $\bar{B}_n$ bounded. Then for the $\bar{s}$ given in equation

These rate results are useful in specifying conditions for root-n consistency and asymptotic normality of $\hat{\theta}$ and consistency of the asymptotic variance estimator, to which we now turn.

4 Large Sample Inference For Affine Functionals

In this Section we give conditions for root-n consistency and asymptotic normality of the estimator $\hat{\theta}$. We also show that the asymptotic variance estimator is consistent. These results allow us to carry out large sample inference about the object of interest in the usual way. We also apply the general results to each of the examples. Recall that the estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}. \tag{4.1}$$

where $\hat{\alpha}_\ell(x) = b(x)'\hat{\rho}_\ell$.

We impose the following conditions.

Assumption 6: *$Var(Y|X)$ is bounded, $\alpha_0(x)$ is bounded, $E[m(W,\gamma_0)^2] < \infty$, and $E[\{m(W,\gamma) - m(W,\gamma_0)\}^2]$ is continuous at γ_0 in $\|\gamma - \gamma_0\|$.*

Boundedness of $Var(Y|X)$ is standard in the regression literature. It may be possible to weaken the second and third conditions but it is beyond the scope of the paper to do so. All of these conditions are imposed to make sure that only L_2 rates are needed for $\hat{\gamma}$ and for $\hat{\alpha}$. This helps the results apply to machine learning estimators where only L_2 convergence rates are available.

Assumption 7: *There are B_n^m and $A(W)$ such that $A(W)$ is sub-Gaussian and*

$$max_{1\le j\le p}|m(W,b_j)| \le B_n^m A(W).$$

This is a primitive condition that leads to a convergence rate for $\hat{M}$.

Lemma 4: *If Assumption 7 is satisfied then*

$$|\hat{M} - M|_\infty = O_p(B_n^m\sqrt{\frac{\ln(p)}{n}}).$$

Note that for $m(w,b_j)=yb_j(x)$ the minimization problems in equations (2.5) and (2.6) are those for the Lasso and Dantzig regression respectively. Thus the convergence rates of Theorems 1 and 3 apply to obtain population L_2 rates for Lasso and Dantzig learners for γ_0.

Our results for $\hat{\theta}$ will rely on a convergence rate for $\hat{\gamma}$. In order to allow these results to apply to as wide a variety of machine learning estimators $\hat{\gamma}$ as possible we just hypothesize such a rate.

Assumption 8: $\|\hat{\gamma}-\gamma_0\| = O_p(n^{-d_\gamma})$, $0<d_\gamma<1/2$.

The results of Section 3 imply such a rate for Lasso or Dantzig selector. The next condition imposes rates that will be sufficient for root-n consistency of $\hat{\theta}$. Let

$$\varepsilon_n^\alpha = [B_n^m + \left(B_n^b\right)^2(1+B_n)]\sqrt{\frac{\ln(p)}{n}}$$

For simplicity we give results just for Lasso; analogous results for the Dantzig selecter will follow similarly.

Assumption 9: $\varepsilon_n^\alpha = o(r_L)$ *and either i) Assumption 3 is satisfied and* $n(1+B_n)r_L(\varepsilon_n^\gamma)^2 \longrightarrow 0$; *or ii) Assumptions 4 and 5 are satisfied and* $n\bar{s}r_L^2(\varepsilon_n^\gamma)^2 \longrightarrow 0$.

This condition will be sufficient for $\sqrt{n}\,\|\hat{\alpha}_L-\alpha_0\|\,\|\hat{\gamma}-\gamma_0\| \xrightarrow{p} 0$ which leads to asymptotic normality of $\hat{\theta}$. For example, consider an approximately sparse α_0 as in equation (3.1), where $B_n^m+\left(B_n^b\right)^2(1+B_n)\le C$ for a positive constant C. Then by Theorem 3 and equation (3.2) Assumption 9 will be statisfied, with r_L going to zero slightly slower than $\sqrt{\ln(p)/n}$, when

$$\frac{d}{1+2d}+d_\gamma>\frac{1}{2}. \tag{4.2}$$

This condition allows for a tradeoff between d, which determines how well a sparse approximation to $\alpha_0(x)$ works, and the convergence rate d_γ for $\hat{\gamma}$. In particular, $\hat{\gamma}$ may converge at a rate that is any small power of n as long as d is large enough.

We also impose a rate condition that is useful for consistency of $\hat{V}$.

Assumption 10: $(1+B_n)B_n^b n^{-d_\gamma}\longrightarrow 0$.

When B_n and B_n^b are bounded this condition is automatically satisfied. The following gives the large sample inference results for $\hat{\theta}$ and $\hat{V}$.

Theorem 5: *If Assumptions 1, and 7-9 are satisfied then for* $\psi_0(w)=m(w,\gamma_0)-\theta_0+\alpha_0(x)[y-\gamma_0(x)]$,

$$\sqrt{n}(\hat{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_0(W_i)+o_p(1).$$

If in addition Assmption 10 is satisfied then $\hat{V} \xrightarrow{p} V = E[\psi_0(W)^2]$.

This result allows γ_0 to be "dense" and estimated at relatively slow rates if $\hat{\alpha}$ converges at a sufficiently fast L_2 rate, as illustrated in equation (4.2). Conversely, we can allow $\hat{\alpha}$ to converge at the relatively slow $n^{-1/4}$ rate of Theorem 3 if $\hat{\gamma}$ converges fast enough. In this way Theorem 5 also allows for α_0 to be "dense" and estimated at slow rates if $\hat{\gamma}$ converges fast enough. We now give more specific regularity conditions for the examples.

4.1 Average Treatment Effect

For the average treatment effect we consider a dictionary of the form $b(x) = [dq(z)', (1-d)q(z)']'$ where $q(z)$ is a $(p/2) \times 1$ dictionary of functions of the covariates z. Note that $m(w,b) = [q(z)', -q(z)']'$ so that

$$\hat{M}_\ell = \begin{pmatrix} \bar{q}_\ell \\ -\bar{q}_\ell \end{pmatrix}, \bar{q}_\ell = \frac{1}{n-n_\ell}\sum_{i \notin I_\ell} q(Z_i).$$

Let $\hat{\rho}^d_\ell$ be the estimated coefficients of $dq(z)$ and $\hat{\rho}^{1-d}_\ell$ the estimated coefficients of $(1-d)q(z)$. Then the RR learner is given by

$$\hat{\alpha}_\ell(X_i) = D_i\hat{\omega}^d_{\ell i} + (1-D_i)\hat{\omega}^{1-d}_{\ell i},\ \hat{\omega}^d_{\ell i} = q(Z_i)'\hat{\rho}^d_\ell,\ \hat{\omega}^{1-d}_{\ell i} = q(Z_i)'\hat{\rho}^{1-d}_\ell,$$

where $\hat{\omega}^d_{\ell i}$ and $\hat{\omega}^{1-d}_{\ell i}$ might be thought of as "weights." These weights sum to one if $q(z)$ includes a constant but need not be nonnegative. The first order conditions for Lasso and the constraints for Dantzig are that for each j,

$$\left|\frac{1}{n-n_\ell}\sum_{i \notin I_\ell} q_j(Z_i)[1 - D_i\hat{\omega}^d_{\ell i}]\right| \le r, \left|\frac{1}{n-n_\ell}\sum_{i \notin I_\ell} q_j(Z_i)[1 + (1-D_i)\omega^{1-d}_{\ell i}]\right| \le r, \tag{4.3}$$

where $r = r_L$ for Lasso and $r = \lambda_D$ for Dantzig. Here we see that RR learner sets the weights $\hat{\omega}^d_{\ell i}$ and $\hat{\omega}^{1-d}_{\ell i}$ to approximately "balance" the overall sample average with the treated and untreated averages for each element of the dictionary $q(z)$. The resulting learner of the ATE is

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in I_\ell}\{\hat{\gamma}_\ell(1, Z_i) - \hat{\gamma}_\ell(0, Z_i) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}. \tag{4.4}$$

The conditions we give are sufficient for this estimator to be root-n consistent when $\hat{\gamma}_\ell$ has a sufficiently fast L_2 convergence rate. The constraints of equation (4.3) are similar to those of Zubizarreta (2015) and Athey, Imbens, and Wager (2017) though the source of these constraints is ℓ_1 regularized best L_2 approximation of the RR $\alpha(x) = \pi_0(z)^{-1}d - [1-\pi_0(z)]^{-1}(1-d)$ by a linear combination of the dictionary $b(x)$. We show here that this type of balancing is sufficient to debias any regression learner under sufficient regularity conditions.

Theorem 6: *If i) there is $C > 0$ with $C < \pi_0(z) = \Pr(D = 1|z) < 1 - C$, $Var(Y|X)$ is bounded; ii) there is B_n^q with $max_{j \le p/2} \sup_z |q_j(Z)| \le B_n^q$ and Assumptions 8 and 9 are satisfied for*

$$\varepsilon_n^\alpha = [B_n^q + (B_n^q)^2 (1 + B_n)]\sqrt{\frac{\ln(p)}{n}}$$

then for $\alpha_0(x) = \pi_0(z)^{-1}d - [1 - \pi_0(z)]^{-1}(1 - d)$ and $\psi_0(w) = \gamma_0(1, z) - \gamma_0(0, z) - \theta_0 + \alpha_0(x)[y - \gamma_0(x)]$,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_0(W_i) + o_p(1).$$

If in addition Assumption 10 is satisfied then $\hat{V} \xrightarrow{p} V = E[\psi_0(W)^2]$.

In comparison with Athey, Imbens, and Wager (2018) this result depends on relatively fast estimation of the RR, or equivalently the dictionary balancing weights, while allowing for relatively slow estimation of the regression.This result can be applied to any regression estimator $\hat{\gamma}$ and we do not require that γ_0 be sparse. The DML form allows us to trade-off rates at which the conditional mean γ_0 and the inverse propensity score are estimated while maintaining root-n consistency, as in equation (4.2) when α_0 is approximately sparse.

4.2 Average Policy Effect

For the average policy effect let $b(x)$ be a dictionary satisfying Assumption 3. Note that $m(w, b) = b(t(x)) - y$, so that

$$\hat{M}_\ell = \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} b(t(X_i)).$$

For $\hat{\rho}_\ell$ equal to the Lasso or Dantzig coefficients, the learner of the RR is given by $\hat{\alpha}_\ell(x) = b(x)'\hat{\rho}_\ell$. The first order conditions for Lasso and the Dantzig constraints are that for each j

$$\left| \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} [b_j(t(X_i)) - b_j(X_i)\hat{\alpha}_\ell(X_i)] \right| \le r.$$

Here $\hat{\alpha}_\ell(X_i)$ acts approximately as a reweighting scheme in making the sample average of the dictionary after transformation $b(t(X_i))$ be approximately equal to the sample average of the reweighted dictionary $b(X_i)\hat{\alpha}_\ell(X_i)$. The resulting learner of the average policy effect is

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{\hat{\gamma}_\ell(t(X_i)) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)] - Y_i\}. \tag{4.5}$$

Theorem 7: *If i) there is $C > 0$ with $1/C \le \alpha_0(x) = f_t(x)/f_0(x) \le C$, $Var(Y|X)$ is bounded; ii) Assumptions 1, 8, and 9 are satisfied for*

$$\varepsilon_n^\alpha = [B_n^b + \left(B_n^b\right)^2 (1 + B_n)]\sqrt{\frac{\ln(p)}{n}};$$

then for $\psi_0(w) = \gamma_0(t(x)) - y - \theta_0 + \alpha_0(x)[y - \gamma_0(x)]$,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_0(W_i) + o_p(1).$$

If in addition Assumption 10 is satisfied then $\hat{V} \xrightarrow{p} V = E[\psi_0(W)^2]$.

The third example, estimation of a bound for average equivalent variation, is treated in detail in Chernozhukov, Hausman, and Newey (2018). We consider here the fourth example.

4.3 Expected Conditional Covariance

For the expected conditional covariance let $b(x)$ be a dictionary satisfying Assumption 3. Note that $m(w, b) - m(w, 0) = -w_1 b(x)$ so that

$$\hat{M}_\ell = \frac{-1}{n - n_\ell} \sum_{i \notin I_\ell} b(X_i) W_{1i}.$$

Here the Lasso or Dantzig are those obtained from Lasso or Dantzig regression where the dependent variable is $-W_{1i}$. The resulting learner of the expected conditional covariance is

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{W_{1i} + \hat{\alpha}_\ell(X_i)\}[Y_i - \hat{\gamma}_\ell(X_i)]\}. \tag{4.6}$$

This estimator

Theorem 8: *If i)* $E[W_1^2|X]$ *and* $E[Y^2|X]$ *are bounded,* $E[W_1^2 Y^2] < \infty$*; ii)* W_1 *is sub-Gaussian and Assumptions 1, 8, and 9 are satisfied for*

$$\varepsilon_n^\alpha = [B_n^b + \left(B_n^b\right)^2 (1 + B_n)] \sqrt{\frac{\ln(p)}{n}};$$

then for $\psi_0(w) = [w + \alpha_0(x)][y - \gamma_0(x)] - \theta_0$,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_0(W_i) + o_p(1).$$

If in addition Assumption 10 is satisfied then $\hat{V} \xrightarrow{p} V = E[\psi_0(W)^2]$.

This result gives root-n consistency and asymptotic normality of the expected conditional covariance estimator when the regression estimator converges fast enough in L_2 and when $E[W_1|X]$ is estimated by Lasso or Dantzig. This asymmetric treatment may be useful in settings where one wants to allow one of the conditional expectation functions to be estimated at a slower rate.

For further bias reduction estimation of $E[Y|X]$ and $E[W_1|X]$ from different samples may be warranted, as in Newey and Robins (2018). It is beyond the scope of this paper to analyze such estimators.

5 Nonlinear Functionals

Debiased machine learning estimators of $\theta_0 = E[m(W, \gamma_0)]$ for nonlinear $m(w, \gamma)$ can also be constructed. The estimator is similar to the linear functional case except that the RR is that of a linearization and a different $\hat{M}$ is needed. In this Section we show how to construct $\hat{M}$ that can be used to machine learn the RR and give conditions that are sufficient for valid large sample inference for nonlinear functionals.

As before a RR is important in the construction of the estimator. Here the RR is that for a linearization of the functional. Suppose that $m(w, \gamma)$ has a Gateaux derivative $D(w, \zeta, \gamma)$ where ζ represents a deviation from γ and $D(w, \zeta, \gamma)$ is linear in ζ. That is suppose that

$$\left.\frac{d}{d\tau} m(w, \gamma + \tau\zeta)\right|_{\tau=0} = D(w, \zeta, \gamma),$$

where τ is a scalar. We will assume that $E[D(W, \gamma, \gamma_0)]$ is a linear mean square continuous functional of γ so that there is a RR $\alpha_0(x)$ satisfying

$$E[D(W, \gamma, \gamma_0)] = E[\alpha_0(X)\gamma(X)],$$

for all $\gamma(x)$ with finite second moment. This Riesz representation is analogous to equation (2.2) with the functional $m(w, \gamma) - m(w, 0)$ replaced by the first order approximation $D(w, \gamma, \gamma_0)$. The Riesz representation implies that for $D(w, b, \gamma_0) = (D(w, b_1, \gamma_0), ..., D(w, b_p, \gamma_0))'$,

$$M = E[D(W, b, \gamma_0)] = E[\alpha_0(X)b(X)].$$

A learner $\hat{\theta}$ can be constructed from an estimator $\hat{\alpha}_\ell(x)$ of the RR $\alpha_0(x)$ and a learner $\hat{\gamma}_\ell(x)$ of $E[Y|X = x]$ exactly as in equation (2.3). This estimator may not be doubly robust due to the nonlinearity of $m(w, \gamma)$ in γ. Nevertheless it will have zero first order bias and so be root-n consistent and asymptotically normal under sufficient regularity conditions. It has zero first order bias because $\alpha_0(x)[y - \gamma_0(x)]$ is the influence function for $E[m(W, \gamma)]$, as shown in Newey (1994), and because a sample average plus an average of an estimate of that influence function has zero order bias; see Chernozhukov et al. (2016).

An estimator $\hat{\alpha}_\ell(x)$ is needed to construct $\hat{\theta}$. We continue to consider estimators $\hat{\alpha}_\ell(x)$ described in Section 2, but based on a different $\hat{M}_\ell$, where it is now convenient to include an ℓ subscript. For a machine learning estimator $\hat{\gamma}_{\ell,\ell'}$ of $E[Y|X]$ obtained from observations not in either I_ℓ or $I_{\ell'}$ the estimator $\hat{M}_\ell$ is given by

$$\hat{M}_\ell = (\hat{M}_{\ell 1}, ..., \hat{M}_{\ell p})',$$
$$\hat{M}_{\ell j} = \frac{d}{d\tau}\left(\frac{1}{n - n_\ell}\right)\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} m(W_i, \hat{\gamma}_{\ell,\ell'} + \tau b_j) = \left(\frac{1}{n - n_\ell}\right)\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} D(W_i, b_j, \hat{\gamma}_{\ell,\ell'}).$$

This estimator uses further sample splitting where $\hat{M}$ is constructed by averaging over observations that are not used in $\hat{\gamma}_{\ell,\ell'}$. For convenience we have used the same partitioning of the observations as before. This additional sample splitting helps us allow for p to still be large in this setting where we are plugging in a nonparametric estimator into many sample moments.

Next we obtain a convergence rate for $\hat{M}$.

Assumption 11: *There is $\varepsilon > 0$, B_n^D, B_n^Δ and sub-Gaussian $A(W)$ such that for all γ with $\|\gamma - \gamma_0\| \leq \varepsilon$,* i)

$$\max_j |D(W, b_j, \gamma)| \leq B_n^D A(W),$$

ii) $\max_j |E[D(W, b_j, \gamma) - D(W, b_j, \gamma_0)]| \leq B_n^\Delta \|\gamma - \gamma_0\|$.

Lemma 9: *If Assumptions 7 and 8 are satisfied then*

$$|\hat{M} - M|_\infty = O_p(\varepsilon_n^M), \ \varepsilon_n^M = (B_n^D \sqrt{\frac{\ln(p)}{n}} + B_n^\Delta \varepsilon_n^\gamma).$$

To allow for nonlinearity of $m(w, \gamma)$ in γ we impose the following condition

Assumption 12: *There are ε, $C > 0$ such that for all γ with $\|\gamma - \gamma_0\| \leq \varepsilon$,*

$$|E[m(W, \gamma) - m(W, \gamma_0) - D(W, \gamma - \gamma_0, \gamma_0)]| \leq C\|\gamma - \gamma_0\|^2.$$

This condition implies that $E[m(W, \gamma)]$ is Frechet differentiable in $\|\gamma - \gamma_0\|$ with derivative $E[D(W, \gamma - \gamma_0, \gamma_0)]$. It is a specific condition that corresponds to $E[m(W, \gamma)]$ being an L_2 differentiable function.

Let

$$\varepsilon_n^\alpha = [B_n^D + (B_n^b)^2(1 + B_n)]\sqrt{\frac{\ln(p)}{n}} + B_n^\Delta \varepsilon_n^\gamma$$

Modify Assumption 9.

Theorem 10: *If Assumptions 1, 6, 8-9, and 11-12 are satisfied with $\varepsilon_n^\gamma = o(n^{-1/4})$ and $E[m(W, \gamma_0)^2] < \infty$, then for $\psi_0(w) = m(w, \gamma_0) - \theta_0 + \alpha_0(x)[y - \gamma_0(x)]$,*

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_0(W_i) + o_p(1).$$

If in addition Assumption 10 is satisfied then $\hat{V} \xrightarrow{p} V = E[\psi_0(W)^2]$.

6 Appendix: Proofs of Results

In this Appendix we give the proofs of the results of the paper, partly based on useful Lemmas that are stated and proved in this Appendix. The first Lemma states a well known necessary condition for minimizing the Lasso objective function.

Lemma A0: *For any* $p \times 1$ vector $\hat{M}$, $p \times p$ *positive semi-definite* $\hat{G}$, *and* $r > 0$, if $\rho^* = \arg\min_\rho\{-2\hat{M}'\rho + \rho'\hat{G}\rho + 2r|\rho|_1\}$ *then*

$$|\hat{M} - \hat{G}\rho^*|_\infty \leq r.$$

Proof: Because the objective function is convex in ρ, a necessary condition for minimization is that 0 belongs to the sub-differential of the objective, i.e.

$$0 \in -2\hat{M} + 2\hat{G}\rho^* + 2r([-1,1] \times ... \times [-1,1])'.$$

Therefore for each j we have

$$0 \leq -2\hat{M}_j + 2e_j'\hat{G}\rho^* + 2r, \; 0 \geq -2\hat{M}_j + 2e_j'\hat{G}\rho^* - 2r,$$

where e_j is the j^{th} unit vector. Dividing through by 2 and adding $\hat{M}_j - e_j'\hat{G}\rho^*$ both sides of each inequality gives

$$-r \leq \hat{M}_j - e_j'\hat{G}\rho^* \leq r,$$

that is,

$$|\hat{M}_j - e_j'\hat{G}\rho^*| \leq r.$$

The conclusion follows because this inequality holds for each j. *Q.E.D.*

The following result gives the rate of convergence of $|\hat{G} - G|_\infty$. Let $\|A(W)\|_{\Psi_2}$ be the sub-Gaussian norm of a random variable $A(W)$ as in Vershynin (2018).

Lemma A1: *If Assumption 3 is satisfied then*

$$|\hat{G} - G|_\infty = O_p(\varepsilon_n^G), \; \varepsilon_n^G = (B_n^b)^2\sqrt{\frac{\ln(p)}{n}}.$$

Proof: Define

$$T_{ijk} = b_j(X_i)b_k(X_i) - E[b_j(X_i)b_k(X_i)], \; U_{jk} = \frac{1}{n}\sum_{i=1}^{n} T_{ijk}.$$

For any constant C,

$$\Pr(|\hat{G}-G|_\infty \geq C\varepsilon_n^G) \leq \sum_{j,k=1}^{p} \mathbb{P}(|U_{jk}| > C\varepsilon_n^G) \leq p^2 \max_{j,k} \mathbb{P}(|U_{jk}| > C\varepsilon_n^G)$$

Note that $E[T_{ijk}] = 0$ and

$$|T_{ijk}| \leq |b_j(X_i)| \cdot |b_k(X_i)| + E[|b_j(X_i)| \cdot |b_k(X_i)|] \leq 2(B_n^b)^2.$$

Define $K = \|T_{ijk}\|_{\Psi_2} \leq 2(B_n^b)^2$. By Hoeffding's inequality there is a constant c such that

$$\begin{aligned} p^2 \max_{j,k} \mathbb{P}(|U_{jk}| > C\varepsilon_n^G) &\leq 2p^2 \exp\left(-\frac{cn(C\varepsilon_n^G)^2}{K^2}\right) \leq 2p^2 \exp\left(-\frac{cn(C\varepsilon_n^G)^2}{4(B_n^b)^4}\right) \\ &\leq 2\exp\left(\ln(p)[2-\frac{cC^2}{4}]\right) \longrightarrow 0. \end{aligned}$$

For any $C > \sqrt{8/c}$. Thus for large enough C, $\Pr(|\hat{G}-G|_\infty \geq C\varepsilon_n^G) \longrightarrow 0$, implying the conclusion. *Q.E.D.*

In what follows let $\varepsilon_n = \max\{\varepsilon_n^G, \varepsilon_n^M\}$,

$$\rho_L = \arg\min_\rho\{\|\alpha_0 - b'\rho\|^2 + 2\varepsilon_n|\rho|_1\},\ \rho_D = \arg\min_\rho |\rho|_1 \text{ subject to } |M - G\rho|_\infty \leq \varepsilon_n$$

Lemma A2: *If Assumption 3 is satisfied then*

$$\begin{aligned} \|\alpha_0 - b'\rho_L\|^2 &\leq C(1+B_n)\varepsilon_n, \quad |\rho_L|_1 \leq C(1+B_n), \\ \|\alpha_0 - b'\rho_D\|^2 &\leq C(1+B_n)\varepsilon_n, \quad |\rho_D|_1 \leq C(1+B_n). \end{aligned}$$

Proof: The first conclusion follows immediately from

$$\|\alpha_0 - b'\rho_L\|^2 + 2\varepsilon_n|\rho_L|_1 \leq \|\alpha_0 - b'\rho_n\|^2 + 2\varepsilon_n|\rho_n|_1 \leq \varepsilon_n(C + 2B_n) \leq C(1+B_n)\varepsilon_n.$$

Also, the first order conditions for ρ_L imply that $|M - G\rho_L|_\infty \leq \varepsilon_n$, so that ρ_L is feasible for the Dantzig minimization problem, and hence

$$|\rho_D|_1 \leq |\rho_L|_1 \leq C(1+B_n).$$

Also by the triangle inequality

$$\begin{aligned} \|b'(\rho_L - \rho_D)\|^2 &= (\rho_L - \rho_D)'G(\rho_L - \rho_D) \leq |\rho_L - \rho_D|_1 |G(\rho_D - \rho_L)|_\infty \\ &\leq 2(|\rho_L|_1 + |\rho_D|_1)|M - G\rho_D - (M - G\rho_L)|_\infty \leq C(1+B_n)\varepsilon_n. \end{aligned}$$

The second conclusion then follows from

$$\|\alpha_0 - b'\rho_D\|^2 \leq 2\|\alpha_0 - b'\rho_L\|^2 + 2\|b'(\rho_L - \rho_D)\|^2 \leq C\varepsilon_n(1 + B_n).\ Q.E.D.$$

LEMMA A3: *If Assumptions 1-3 are satisfied and $\varepsilon_n^M + \varepsilon_n^G(1 + B_n) = o(r_L)$ then $|\hat{\rho}_L|_1 = O_p(1 + B_n)$.*

Proof: The first order conditions for ρ_L imply

$$|M - G\rho_L|_\infty \leq \varepsilon_n.$$

Then by the triangle and Holder inequalities, Lemma A2, and $\varepsilon_n \leq \varepsilon_n^M + \varepsilon_n^G(1 + B_n)$,

$$\begin{aligned} |\hat{M} - \hat{G}\rho_L|_\infty &\leq |M - G\rho_L|_\infty + |\hat{M} - M|_\infty + |(G - \hat{G})\rho_L|_\infty \\ &= O_p(\varepsilon_n + \varepsilon_n^M) + |G - \hat{G}|_\infty |\rho_L|_1 \\ &= O_p(\varepsilon_n^M + \varepsilon_n^G(1 + B_n)) = o_p(r_L). \end{aligned}$$

By the definition of $\hat{\rho}_L$,

$$-2\hat{M}'\hat{\rho}_L + \hat{\rho}_L'\hat{G}\hat{\rho}_L + 2r_L|\hat{\rho}_L|_1 \leq -2\hat{M}'\rho_L + \rho_L'\hat{G}\rho_L + 2r_L|\rho_L|_1.$$

Subtracting the first two terms on the left-hand side of this inequality from both sides gives

$$\begin{aligned} 2r_L|\hat{\rho}_L|_1 &\leq 2\hat{M}'(\hat{\rho}_L - \rho_L) - [\hat{\rho}_L'\hat{G}\hat{\rho}_L - \rho_L'\hat{G}\rho_L] + 2r_L|\rho_L|_1 \\ &= 2\hat{M}'(\hat{\rho}_L - \rho_L) - [(\hat{\rho}_L - \rho_L)'\hat{G}(\hat{\rho}_L - \rho_L) + 2\rho_L'\hat{G}(\hat{\rho}_L - \rho_L)] + 2r_L|\rho_L|_1 \\ &\leq 2(\hat{M} - \hat{G}\rho_L)'(\hat{\rho}_L - \rho_L) + 2r_L|\rho_L|_1 \\ &\leq 2\left|\hat{M} - \hat{G}\rho_L\right|_\infty |\hat{\rho}_L - \rho_L|_1 + 2r_L|\rho_L|_1. \end{aligned}$$

Dividing through both sides of this inequality by $2r_L$ gives

$$|\hat{\rho}_L|_1 \leq o_p(1)|\hat{\rho}_L - \rho_L|_1 + |\rho_L|_1 \leq |\rho_L|_1 + o_p(1)(|\hat{\rho}_L|_1 + |\rho_L|_1). \tag{6.1}$$

It follows that with probability approaching one (w.p.a.1),

$$|\hat{\rho}_L|_1 \leq |\rho_L|_1 + \frac{1}{2}(|\hat{\rho}_L|_1 + |\rho_L|_1).$$

Subtracting $|\hat{\rho}_L|_1/2$ from both sides and multiplying through by 2 gives w.p.a.1,

$$|\hat{\rho}_L|_1 \leq 3|\rho_L|_1 \leq C(1 + B_n).Q.E.D.$$

Proof of Theorem 1: The population and sample Lasso first order conditions give

$$|M - G\rho_L|_\infty \le \varepsilon_n,\ |\hat{M} - \hat{G}\hat{\rho}_L|_\infty \le r_L.$$

Then by Lemma A3 and the triangle and Holder inequalities,

$$\begin{aligned} |G(\hat{\rho}_L - \rho_L)|_\infty &\le |(G - \hat{G})\hat{\rho}_L|_\infty + |\hat{G}\hat{\rho}_L - \hat{M}|_\infty + |\hat{M} - M|_\infty + |M - G\rho_L|_\infty \\ &\le |G - \hat{G}|_\infty |\hat{\rho}_L|_1 + |\hat{G}\hat{\rho}_L - \hat{M}|_\infty + |\hat{M} - M|_\infty + |M - G\rho_L|_\infty \\ &= O_p(\varepsilon_n^G(1 + B_n) + r_L + \varepsilon_n^M + \varepsilon_n) = O_p(r_L). \end{aligned} \tag{6.2}$$

Similarly, the Dantzig constraints imply

$$|M - G\rho_D|_\infty \le \varepsilon_n,\ |\hat{M} - \hat{G}\hat{\rho}_D|_\infty \le \lambda_D,$$

Also $\hat{\rho}_L$ satisfies the Dantzig constraints so by Lemma A3,

$$|\hat{\rho}_D|_1 \le |\hat{\rho}_L|_1 = O_p(1 + B_n).$$

Then as in equation (6.2),

$$\begin{aligned} |G(\hat{\rho}_D - \rho_D)|_\infty &\le |(G - \hat{G})\hat{\rho}_D|_\infty + |\hat{G}\hat{\rho}_D - \hat{M}|_\infty + |\hat{M} - M|_\infty + |M - G\rho_D|_\infty \\ &= O_p(\varepsilon_n^G(1 + B_n) + \lambda_D + \varepsilon_n^M + \varepsilon_n) = O_p(\lambda_D). \end{aligned}$$

By Lemmas A2 and A3,

$$\begin{aligned} \|\hat{\alpha}_L - \alpha_0\|^2 &\le 2\,\|\hat{\alpha}_L - b'\rho_L\|^2 + 2\,\|b'\rho_L - \alpha_0\|^2 \\ &\le 2(\hat{\rho}_L - \rho_L)'G(\hat{\rho} - \rho_L) + C(1 + B_n)\varepsilon_n \\ &\le 2\,|\hat{\rho}_L - \rho_L|_1\,|G(\hat{\rho} - \rho_L)|_\infty + O((1 + B_n)r_L) \\ &\le O_p(1 + B_n)O_p(r_L) + O((1 + B_n)r_L) = O_p(r_L), \end{aligned}$$

giving the first conclusion. The second conclusion follows similarly, with

$$\|\hat{\alpha}_D - \alpha_0\|^2 \le 2(\hat{\rho}_D - \rho_D)'G(\hat{\rho}_D - \rho_D) + C(1 + B_n)\varepsilon_n = O_p(r_L).\ Q.E.D.$$

We next give a result bounding the approximation error $\|\alpha_0 - b'\rho_L\|^2$ where $\alpha_L = b(x)'\rho_L$ is the population Lasso approximation to $\alpha_0(x)$.

Lemma A4: *If Assumptions 4 and 5 are satisfied then there is $C > 0$ such that for all ρ,*

$$\begin{aligned} &\|\alpha_0 - b'\rho_L\|^2 \le C[\|\alpha_0 - b'\rho\|^2 + \varepsilon_n^2\mathcal{M}(\rho)], \\ &\|\alpha_0 - b'\rho_D\|^2 \le C[\|\alpha_0 - b'\rho\|^2 + \varepsilon_n^2\{\mathcal{M}(\rho) + \mathcal{M}(\rho_L)\}]. \end{aligned}$$

Proof: For any ρ let $\alpha_\rho(x) = b(x)'\rho$, $\delta = \rho_L - \rho$, $\mathcal{J} = \{1, ..., p\}$, $\mathcal{J}_\rho$ be the subset of $\mathcal{J}$ with $\rho_j \neq 0$, and $\mathcal{J}_\rho^c$ be the complement of $\mathcal{J}_\rho$ in $\mathcal{J}$. Then

$$\|\alpha_0 - \alpha_L\|^2 + 2\varepsilon_n |\rho_L|_1 \leq \|\alpha_0 - \alpha_\rho\|^2 + 2\varepsilon_n |\rho|_1 .$$

Adding $-2r_0 |\rho_L|_1 + \varepsilon_n |\delta|_1$ to both sides gives

$$\begin{aligned}\|\alpha_0 - \alpha_L\|^2 + \varepsilon_n |\delta|_1 &\leq \|\alpha_0 - \alpha_\rho\|^2 + 2\varepsilon_n |\rho|_1 - 2\varepsilon_n |\rho_L|_1 + \varepsilon_n |\delta|_1 \leq \|\alpha_0 - \alpha_\rho\|^2 + 2\varepsilon_n(|\rho|_1 - |\rho_L|_1 + |\delta|_1) \\ &= \|\alpha_0 - \alpha_\rho\|^2 + 2\varepsilon_n \sum_{j=1}^{p} (|\rho_j| - |\rho_{Lj}| + |\rho_{Lj} - \rho_j|) \\ &= \|\alpha_0 - \alpha_\rho\|^2 + 2\varepsilon_n \sum_{j \in \mathcal{J}_\rho} (|\rho_j| - |\rho_{Lj}| + |\rho_{Lj} - \rho_j|) \leq \|\alpha_0 - \alpha_\rho\|^2 + 4\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| .\end{aligned}$$

Subtracting $\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j|$ from both sides gives

$$\|\alpha_0 - \alpha_L\|^2 + \varepsilon_n \sum_{j \in \mathcal{J}_\rho^c} |\delta_j| \leq \|\alpha_0 - \alpha_\rho\|^2 + 3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| . \tag{6.3}$$

Choose any $\xi = 3/(k-3)$. If $3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| \leq \xi \|\alpha_0 - \alpha_\rho\|^2$ then

$$\|\alpha_0 - \alpha_L\|^2 \leq \|\alpha_0 - \alpha_L\|^2 + \varepsilon_n \sum_{j \in \mathcal{J}_\rho^c} |\delta_j| \leq \|\alpha_0 - \alpha_\rho\|^2 + 3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| \leq (1 + \xi) \|\alpha_0 - \alpha_\rho\|^2 .$$

Now suppose that $3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| > \xi \|\alpha_0 - \alpha_\rho\|^2$. Then

$$\begin{aligned}\varepsilon_n \sum_{j \in \mathcal{J}_\rho^c} |\delta_j| &\leq \|\alpha_0 - \alpha_L\|^2 + \varepsilon_n \sum_{j \in \mathcal{J}_\rho^c} |\delta_j| \leq \|\alpha_0 - \alpha_\rho\|^2 + 3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| \\ &\leq (1 + 1/\xi) 3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| = k\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| .\end{aligned}$$

Then dividing through by ε_n it follows by Assumption 5 that there is $\bar{C}$ not depending on ρ such that

$$\sum_{j \in \mathcal{J}_\rho} \delta_j^2 \leq \bar{C} \delta' G \delta = \bar{C} \|\alpha_L - \alpha_\rho\|^2 .$$

Also by the Cauchy-Schwartz and triangle inequalities

$$\sum_{j \in \mathcal{J}_\rho} |\delta_j| \leq \sqrt{\mathcal{M}(\rho)} \sqrt{\sum_{j \in \mathcal{J}_\rho} \delta_j^2} \leq \sqrt{\mathcal{M}(\rho)} \sqrt{\bar{C}} \|\alpha_L - \alpha_\rho\| \leq \sqrt{\bar{C}} \sqrt{\mathcal{M}(\rho)} (\|\alpha_0 - \alpha_L\| + \|\alpha_0 - \alpha_\rho\|),$$

so that

$$\begin{aligned}\|\alpha_0 - \alpha_L\|^2 &\leq \|\alpha_0 - \alpha_L\|^2 + \varepsilon_n \sum_{j \in \mathcal{J}_\rho^c} |\delta_j| \leq \|\alpha_0 - \alpha_\rho\|^2 + 3\varepsilon_n \sum_{j \in \mathcal{J}_\rho} |\delta_j| \\ &\leq \|\alpha_0 - \alpha_\rho\|^2 + 3\varepsilon_n \sqrt{\bar{C}} \sqrt{\mathcal{M}(\rho)} (\|\alpha_0 - \alpha_L\| + \|\alpha_0 - \alpha_\rho\|).\end{aligned}$$

Note that

$$3\varepsilon_n\sqrt{\bar{C}}\sqrt{\mathcal{M}(\rho)}\left\|\alpha_0-\alpha_\rho\right\| \leq \frac{9}{4}\varepsilon_n^2\bar{C}\mathcal{M}(\rho)+\left\|\alpha_0-\alpha_\rho\right\|^2,$$

$$3\varepsilon_n\sqrt{\bar{C}}\sqrt{\mathcal{M}(\rho)}\left\|\alpha_0-\alpha_L\right\| = 6\varepsilon_n\sqrt{\bar{C}}\sqrt{\mathcal{M}(\rho)}(\frac{1}{2}\left\|\alpha_0-\alpha_L\right\|) \leq 9\varepsilon_n\bar{C}\mathcal{M}(\rho)+\frac{1}{4}\left\|\alpha_0-\alpha_L\right\|^2.$$

Substituting these two inequalities in the previous one, subtracting $\left\|\alpha_0-\alpha_L\right\|^2$ from both sides, collecting terms, and multiplying through by $4/3$ gives

$$\left\|\alpha_0-\alpha_L\right\|^2 \leq \frac{4}{3}\{2\left\|\alpha_0-\alpha_\rho\right\|^2+\bar{C}(\frac{9}{4}+9)\varepsilon_n^2\mathcal{M}(\rho)\} = \frac{8}{3}\left\|\alpha_0-\alpha_\rho\right\|^2+15\bar{C}\varepsilon_n^2\mathcal{M}(\rho).$$

The conclusion for Lasso then follows for $C=\max\{1+\xi, 8/3, 15\bar{C}\}$.

For Dantzig, note that for $\delta=\rho_D-\rho_L$ we have

$$\begin{aligned}\left\|\alpha_0-\alpha_L\right\|^2 &= \left\|\alpha_0-\alpha_D+\alpha_D-\alpha_L\right\|^2\\ &= \left\|\alpha_0-\alpha_D\right\|^2+2E[\{\alpha_0(X)-b(X)'\rho_D\}b(X)'\delta]+\delta'G\delta\\ &= \left\|\alpha_0-\alpha_D\right\|^2+2\delta'(M-G\rho_D)+\delta'G\delta.\end{aligned}$$

Solving gives

$$\begin{aligned}\left\|\alpha_0-\alpha_D\right\|^2 &= \left\|\alpha_0-\alpha_L\right\|^2-2\delta'(M-G\rho_D)-\delta'G\delta \leq \left\|\alpha_0-\alpha_L\right\|^2+2\left|\delta\right|_1\left|M-G\rho_D\right|_\infty-\delta'G\delta.\\ &\leq \left\|\alpha_0-\alpha_L\right\|^2+2\left|\delta\right|_1\varepsilon_n-\delta'G\delta.\end{aligned}$$

By feasibility of ρ_L for the Dantzig problem, $\left|\rho_L+\delta\right|_1=\left|\rho_D\right|_1\leq\left|\rho_L\right|_1$. Therefore,

$$\sum_{j\in\mathcal{J}_{\rho_L}^c}\left|\delta_j\right|+\sum_{j\in\mathcal{J}_{\rho_L}}\left|\rho_{Lj}+\delta_j\right| = \left|\rho_L+\delta\right|_1 \leq \left|\rho_L\right|_1 = \sum_{j\in\mathcal{J}_{\rho_L}}\left|\rho_{Lj}\right|.$$

Subtracting gives and the triangle inequality gives

$$\sum_{j\in\mathcal{J}_{\rho_L}^c}\left|\delta_j\right| \leq \sum_{j\in\mathcal{J}_{\rho_L}}\left|\rho_{Lj}\right|-\sum_{j\in\mathcal{J}_{\rho_L}}\left|\rho_{Lj}+\delta_j\right| \leq \sum_{j\in\mathcal{J}_{\rho_L}}\left|\delta_j\right| \leq k\sum_{j\in\mathcal{J}_{\rho_L}}\left|\delta_j\right|.$$

Then by Assumption 5 there is a constant $\bar{C}$ such that $\delta'G\delta\geq\bar{C}\sum_{j\in\mathcal{J}_{\rho_L}}\delta_j^2$. It then follows by the Cauchy-Schwartz inequality that

$$\begin{aligned}2\left|\delta\right|_1\varepsilon_n-\delta'G\delta &= 2(\sum_{j\in\mathcal{J}_{\rho_L}^c}\left|\delta_j\right|+\sum_{j\in\mathcal{J}_{\rho_L}}\left|\delta_j\right|)\varepsilon_n-\delta'G\delta \leq 4(\sum_{j\in\mathcal{J}_{\rho_L}}\left|\delta_j\right|)\varepsilon_n-\delta'G\delta\\ &\leq 4\sqrt{\mathcal{M}(\rho_L)}\sqrt{\sum_{j\in\mathcal{J}_{\rho_L}}\delta_j^2}\varepsilon_n-\bar{C}\sum_{j\in\mathcal{J}_{\rho_L}}\delta_j^2 = 2\sqrt{4\mathcal{M}(\rho_L)\varepsilon_n^2/\bar{C}}\sqrt{\bar{C}\sum_{j\in\mathcal{J}_{\rho_L}}\delta_j^2}-\bar{C}\sum_{j\in\mathcal{J}_{\rho_L}}\delta_j^2\\ &\leq 4\mathcal{M}(\rho_L)\varepsilon_n^2/\bar{C}.\end{aligned}$$

Therefore the second conclusion follows by the first conclusion. *Q.E.D.*

Lemma A5: *If Assumption 5 is satisfied then there is $C > 0$ such that for all ρ,*

$$\mathcal{M}(\rho_L) \leq C[\varepsilon_n^{-2} \|\alpha_0 - b'\rho\|^2 + \mathcal{M}(\rho)].$$

Proof of Lemma A5: Let $e_L(x) = \alpha_0(x) - \alpha_L(x)$. Note that for $\bar{\lambda} = \lambda_{\max}(G)$ we have $G^{-1} \geq 1/\bar{\lambda}$. Also, for $\rho_{Lj} \neq 0$ the first order conditions for ρ_{Lj} imply $E[b_j(X)e_L(X)] = \varepsilon_n sgn(\rho_{Lj})$ for $\rho_{Lj} \neq 0$. Then as usual for the population least squares regression of $e_L(X)$ on $b(X)$,

$$\begin{aligned}\|\alpha_0 - \alpha_L\|^2 &= E[e_L(X)^2] \geq E[e_L(X)b(X)']G^{-1}E[b(X)e_L(X)] \geq \frac{1}{\bar{\lambda}}E[e_L(X)b(X)']E[b(X)e_L(X)] \\ &\geq \frac{1}{\bar{\lambda}} \sum_{j \in \mathcal{J}_{\rho_L}} \{E[e_L(X)b_j(X)]\}^2 = \frac{1}{\bar{\lambda}}\mathcal{M}(\rho_L)\varepsilon_n^2\end{aligned}$$

The first conclusion of Lemma A4 and dividing both sides by $\varepsilon_n^2/\bar{\lambda}$ gives the conclusion. *Q.E.D.*

Lemma A6: *If Assumptions 1, 2, and 5 are satisfied and $\varepsilon_n^M + \varepsilon_n^G(1 + \bar{B}_n) = o(r_L)$ then with probability approaching one $\sum_{j \in \mathcal{J}_{\rho_L}^c} |\hat{\rho}_{Lj} - \rho_{Lj}| \leq 3 \sum_{j \in \mathcal{J}_{\rho_L}} |\hat{\rho}_{Lj} - \rho_{Lj}|$.*

Proof: It follows as in equation (6.1) of the proof of Lemma A3 that

$$|\hat{\rho}_L|_1 \leq |\rho_L|_1 + o_p(1)|\hat{\rho}_L - \rho_L|_1.$$

Therefore with probability approaching one,

$$|\hat{\rho}_L|_1 \leq |\rho_L|_1 + \frac{1}{2}|\hat{\rho}_L - \rho_L|_1.$$

Note that $|\rho_{Lj}| + |\hat{\rho}_{Lj} - \rho_{Lj}| - |\hat{\rho}_{Lj}| = 0$ when $\rho_{Lj} = 0$ and that $|\rho_{Lj}| - |\hat{\rho}_{Lj}| \leq |\hat{\rho}_{Lj} - \rho_{Lj}|$ by the triangle inequality. Then adding $|\hat{\rho}_L - \rho_L|_1/2$ to and subtracting $|\hat{\rho}_L|_1$ from both sides gives

$$\begin{aligned}\frac{1}{2}|\hat{\rho}_L - \rho_L|_1 &\leq |\rho_L|_1 + |\hat{\rho}_L - \rho_L|_1 - |\hat{\rho}_L|_1 = \sum_{j=1}^{p}(|\rho_{Lj}| + |\hat{\rho}_{Lj} - \rho_{Lj}| - |\hat{\rho}_{Lj}|) \\ &= \sum_{j \in A_L}(|\rho_{Lj}| + |\hat{\rho}_{Lj} - \rho_{Lj}| - |\hat{\rho}_{Lj}|) \leq 2\sum_{j \in A_L}|\hat{\rho}_{Lj} - \rho_{Lj}|.\end{aligned}$$

Note that $|\hat{\rho}_L - \rho_L|_1 = \sum_{j \in A_L^c} |\hat{\rho}_{Lj} - \rho_{Lj}| + \sum_{j \in A_L} |\hat{\rho}_{Lj} - \rho_{Lj}|$, so multiplying both sides by 2 and subtracting $\sum_{j \in A_L} |\hat{\rho}_{Lj} - \rho_{Lj}|$ from both sides gives the result. *Q.E.D.*

Proof of Theorem 3: Choose $\bar{\rho}$ so that $\|\alpha_0 - b'\bar{\rho}\|^2 \leq \bar{s}\varepsilon_n^2$. the triangle inequality and Lemma A4,

$$\begin{aligned}\|\hat{\alpha}_L - \alpha_0\|^2 &\leq 2\|\hat{\alpha}_L - \alpha_L\|^2 + 2\|\alpha_L - \alpha_0\|^2 \\ &\leq 2(\hat{\rho}_L - \rho_L)'G(\hat{\rho}_L - \rho_L) + \bar{s}\varepsilon_n^2.\end{aligned} \tag{6.4}$$

It follows as in the proof of Theorem 1 that $|G(\hat{\rho}_L - \rho_L)|_\infty = O_p(r_L)$. Also by Lemma A5, $\mathcal{M}(\rho_L) \leq C\bar{s}$. Then by Lemma A6 and Assumption 5, with probability approaching one

$$\begin{aligned} |\hat{\delta}|_1^2 &= (\sum_{j\in\mathcal{J}^c_{\rho_L}} |\hat{\delta}_j| + \sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|)^2 \leq (4\sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|)^2 \leq C\mathcal{M}(\rho_L)\sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|^2 \leq C\bar{s}\hat{\delta}'G\hat{\delta} \\ &\leq C\bar{s}|G\hat{\delta}|_\infty|\hat{\delta}|_1 = O_p(\bar{s}r_L)|\hat{\delta}|_1. \end{aligned}$$

Dividing through by $|\hat{\delta}|_1$ then gives $|\hat{\delta}|_1 = O_p(\bar{s}r_L)$. It follows that

$$\hat{\delta}'G\hat{\delta} \leq |G\hat{\delta}|_\infty|\hat{\delta}|_1 = O_p(\bar{s}r_L^2).$$

The conclusion for Lasso then follows from eq. (6.4).

For the Dantzig selector, the triangle and Holder inequalities give

$$\begin{aligned} |\hat{M} - \hat{G}\rho_D|_\infty &\leq |\hat{M} - M|_\infty + |M - G\rho_D|_\infty + |(G - \hat{G})\rho_D|_\infty \\ &\leq |\hat{M} - M|_\infty + \varepsilon_n + |G - \hat{G}|_\infty|\rho_D|_1 \\ &= O_p(\varepsilon_n^M + \varepsilon_n^G(1 + \bar{B}_n)) = o_p(\lambda_D). \end{aligned}$$

It follows that with probability approaching one $|\hat{M} - \hat{G}\rho_D|_\infty \leq \lambda_D$, so that ρ_D is feasible for the sample Dantzig minimization problem. Also, as in the proof of Theorem 1

$$\begin{aligned} |G(\hat{\rho}_D - \rho_D)|_\infty &\leq |(G - \hat{G})\hat{\rho}_D|_\infty + |\hat{G}\hat{\rho}_D - \hat{M}|_\infty + |\hat{M} - M|_\infty + |M - G\rho_D|_\infty \\ &= O_p(\varepsilon_n^G(1 + \bar{B}_n) + \lambda_D + \varepsilon_n^M + \varepsilon_n) = o_p(\lambda_D). \end{aligned}$$

Feasibility of ρ_D with probability approaching one implies $|\rho_D + \hat{\delta}|_1 = |\hat{\rho}_D|_1 \leq |\rho_D|_1$, so by Assumption 5,

$$|\hat{\delta}|_1^2 \leq s_D\hat{\delta}'G\hat{\delta} \leq s_D|G\hat{\delta}|_\infty|\hat{\delta}|_1 = O_p(s_D\lambda_D)|\hat{\delta}|_1.$$

Dividing through by $|\hat{\delta}|_1$ then gives $|\hat{\delta}|_1 = O_p(s_D\lambda_D)$, so that

$$\hat{\delta}'G\hat{\delta} \leq |G\hat{\delta}|_\infty|\hat{\delta}|_1 = O_p(s_D\lambda_D^2).$$

The conclusion then follows by Lemmas A4 and A5 and the triangle inequality. *Q.E.D.*

Proof of Lemma 4: Define

$$T_{ij} = m(W_i, b_j) - E[m(W_i, b_j)],\ U_j = \frac{1}{n}\sum_{i=1}^n T_{ij}.$$

For any constant C,

$$\Pr(|\hat{M} - M|_\infty \geq C\varepsilon_n^M) \leq \sum_{j=1}^p \mathbb{P}(|U_j| > C\varepsilon_n^M) \leq p \cdot \max_j \mathbb{P}(|U_j| > C\varepsilon_n^M).$$

Note that $E[T_{ij}] = 0$ and

$$|T_{ij}| \leq |m(W_i, b_j)| + E[|m(W_i, b_j)|] \leq B_n^M \{A(W_i) + E[|A(W_i)|]\}.$$

Define $C_A = \|A(W_i)\|_{\Psi_2} + E[|A(W_i)|]$ and let $K = \|T_{ij}\|_{\Psi_2} \leq C_A B_n^M$. By Hoeffding's inequality there is a constant c such that

$$\begin{aligned} p \cdot \max_j \mathbb{P}(|U_j| > C\varepsilon_n^M) &\leq 2p \exp\left(-\frac{cn(C\varepsilon_n^M)^2}{K^2}\right) \leq 2p \exp\left(-\frac{cn(C\varepsilon_n^M)^2}{C_A^2 (B_n^M)^2}\right) \\ &\leq 2 \exp\left(\ln(p)[1 - \frac{cC^2}{C_A^2}]\right) \longrightarrow 0, \end{aligned}$$

for any $C > C_A/\sqrt{c}$. Thus for large enough C, $\Pr(|\hat{M} - M|_\infty \geq C\varepsilon_n^M) \longrightarrow 0$, implying the conclusion. *Q.E.D.*

Proof of Theorem 5: By Assumption 8 $\|\hat{\alpha}_L - \alpha_0\| \xrightarrow{p} 0$ so Assumption 6 implies

$$\int [m(W, \hat{\gamma}) - m(W, \gamma_0)]^2 F_0(dW) \xrightarrow{p} 0.$$

Let $\varepsilon_n^\gamma = n^{-d_\gamma}$. It also follows by Assumption 9, $\sqrt{n}\varepsilon_n^\gamma \longrightarrow \infty$, and Theorems 1 and 3 that $\|\hat{\alpha}_L - \alpha_0\| \xrightarrow{p} 0$. In addition by Assumption 8 and Theorems 1 and 3, $\sqrt{n}\|\hat{\alpha}_L - \alpha_0\| \|\hat{\gamma} - \gamma_0\| \xrightarrow{p} 0$. Then first conclusion then follows by Theorem 13 of Chernozhukov et al. (2018b)

To prove the second conclusion let $\psi_i = \psi_0(W_i)$ and $\varepsilon_i = Y_i - \gamma_0(X_i)$. Then for $i \leq\in I_\ell$,

$$\begin{aligned} &(\hat{\psi}_i - \psi_i)^2 \leq 3(R_{i1} + R_{i2} + R_{i3}) \\ &\qquad R_{i1} = [m(W_i, \hat{\gamma}_\ell) - m(W_i, \gamma_0)]^2, R_{i2} = \hat{\alpha}_\ell(X_i)^2 \{\hat{\gamma}(X_i) - \gamma_0(X_i)\}^2, \\ &\qquad R_{i3} = \{\hat{\alpha}_\ell(X_i) - \alpha_0(X_i)\}^2 \{Y_i - \gamma_0(X_i)\}^2. \end{aligned}$$

Let $Z_{-\ell}$ denote the observations not in I_ℓ. Then it follows as previously in this proof that

$$E[R_{i1}|Z_{-\ell}] = \int [m(W, \hat{\gamma}_\ell) - m(W, \gamma_0)]^2 F_0(dW) \xrightarrow{p} 0.$$

Also note that

$$\max_{i \in I_\ell} |\hat{\alpha}_\ell(X_i)| \leq |\hat{\rho}_{L\ell}|_1 \max_{i \in I_\ell} |b(X_i)|_\infty \leq O_p(1 + B_n) B_n^b = O_p((1 + B_n) B_n^b).$$

Therefore, the expectation conditional on the subvector of the data where $i \notin I_\ell$. Then for $i \in I_\ell$, by Assumption 5

$$\begin{aligned} E[R_{i2}|Z_{-\ell}] &\leq O_p((1 + B_n)^2 (B_n^b)^2) \int [\hat{\gamma}_\ell(X) - \gamma_0(X)]^2 F_0(dX) = O_p((1 + B_n)^2 (B_n^b)^2 (\varepsilon_n^\gamma)^2) \xrightarrow{p} 0, \\ E[R_{i3}|Z_{-\ell}] &= E[\{\hat{\alpha}_\ell(X_i) - \alpha_0(X_i)\}^2 \{Y_i - \gamma_0(X_i)\}^2 | X_i, Z_{-\ell}] = E[\{\hat{\alpha}_\ell(X_i) - \alpha_0(X_i)\}^2 Var(Y_i|X_i) | Z_{-\ell}] \\ &\leq C\|\hat{\alpha}_\ell - \alpha_0\|^2 \xrightarrow{p} 0. \end{aligned}$$

It then follows that

$$E[\frac{1}{n}\sum_{i\in I_\ell}(\hat{\psi}_i-\psi_i)^2|Z_{-\ell}]\leq 3\sum_{j=1}^{3}\frac{n_\ell}{n}E[R_{ij}|Z_{-\ell}]\xrightarrow{p}0.$$

It then follows by the triangle and conditional Markov inequalities and summing over ℓ that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i-\psi_i)^2\xrightarrow{p}0.$$

Then $\hat{V}\xrightarrow{p}0$ follows by the law of large numbers in the usual way. *Q.E.D.*

Proof of Theorem 6: Note that

$$m(w,b_j)=1(j\leq p/2)q_j(z)-1(j>p/2)q_{j-p/2}(z).$$

Therefore

$$\max_{1\leq j\leq p}|m(W,b_j)|\leq\max_{1\leq j\leq p/2}|q_j(Z)|\leq B_n^q$$

It then follows by hypothesis ii) of the statement of Theorem 6 that Assumption 1 is satisfied with $B_n^b=B_n^q$, Assumption 7 is satisfied with $A(W)=1$ and $B_n^m=B_n^q$. Then by Lemma 4 and by assumption it follows that Assumptions 8 and 9 are satisfied.

Next, it also follows by hypothesis i) and the form of $\alpha_0(x)$ that $Var(Y|X)$ and $\alpha_0(x)$ are bounded. In addition, by iterated expectations,

$$\begin{aligned}E[\gamma_0(1,Z)^2]&=E[\frac{D}{\pi_0(Z)}\gamma_0(1,Z)^2]=E[\frac{D}{\pi_0(Z)}\gamma_0(X)^2]\leq CE[\gamma_0(X)^2]<\infty,\\E[\{\gamma(1,Z)-\gamma_0(1,Z)\}^2]&=E[\frac{D}{\pi_0(Z)}\{\gamma(1,Z)-\gamma_0(1,Z)\}^2]=E[\frac{D}{\pi_0(Z)}\{\gamma(X)-\gamma_0(X)\}^2]\\&\leq C\|\gamma-\gamma_0\|^2.\end{aligned}$$

Combining these inequalities with the analogous inequalities for $\gamma(0,z)$ it follows that Assumption 6 is satisfied. The conclusions then follows by Theorem 5. *Q.E.D.*

Proof of Theorem 7: Note that

$$m(w,b_j)-m(w,0)=b_j(t(x)).$$

By $\alpha_0(x)$ bounded, the distribution of $t(X)$ is absolutely continuous with respect to the distribution of X so that by Assumption 1,

$$\max_{1\leq j\leq p}|m(W,b_j)|\leq\max_{1\leq j\leq p}|b_j(t(X))|\leq B_n^b$$

It then follows by hypothesis ii) of the statement of Theorem 7 that Assumption 7 is satisfied with $A(W) = 1$ and $B_n^m = B_n^b$. Then by Lemma 4 and by assumption it follows that Assumptions 8 and 9 are satisfied.

Next, it also follows by hypothesis i) that $Var(Y|X)$ and $\alpha_0(x)$ are bounded. In addition, by iterated expectations,

$$\begin{aligned} E[m(W,\gamma_0)^2] &\leq CE[\gamma_0(t(X))^2] + C = C\int \frac{f_t(x)}{f_0(x)}\gamma_0(x)^2 f_0(x)dx + C \\ &\leq CE[\gamma_0(X)^2] + C < \infty, \end{aligned}$$

$$E[\{m(W,\gamma) - m(W,\gamma_0)\}^2] = E[\{\gamma(t(X)) - \gamma_0(t(X)\}^2] = \int \frac{f_t(x)}{f_0(x)}\{\gamma(x) - \gamma_0(x)\}^2 f_0(x)dx \leq C\|\gamma - \gamma_0\|^2.$$

Thus we see that Assumption 6 is satisfied. The conclusion then follows by Theorem 5. *Q.E.D.*

Proof of Theorem 8: We have $m(w,\gamma) = w_1[y - \gamma(x)]$ so that

$$m(w, b_j) - m(w, 0) = -w_1 b_j(x).$$

Therefore by Assumption 3,

$$\max_{1\leq j\leq p}|m(W, b_j) - m(W, 0)| \leq |W_1| \max_{1\leq j\leq p}|b_j(X)| \leq B_n^b|W_1|$$

It then follows by hypothesis ii) of the statement of Theorem 8 that Assumption 7 is satisfied with $A(W) = |W_1|$ and $B_n^m = B_n^b$. Then by Lemma 4 and by assumption it follows that Assumptions 8 and 9 are satisfied.

Next, it also follows by hypothesis i) that $Var(Y|X)$ and $\alpha_0(x) = -E[W_1|x]$ are bounded. In addition, by hypothesis i),

$$\begin{aligned} E[m(W,\gamma_0)^2] &\leq CE[W_1^2\gamma_0(X)^2] + C < \infty, \\ E[\{m(W,\gamma) - m(W,\gamma_0)\}^2] &= E[E[W_1^2|X]\{\gamma(X) - \gamma_0(X)\}^2] \leq C\|\gamma - \gamma_0\|^2. \end{aligned}$$

Thus we see that Assumption 5 is satisfied. The conclusion then follows by Theorem 5. *Q.E.D.*

Proof of Lemma 9: Define

$$\hat{M}_\ell = (\hat{M}_{\ell 1}, ..., \hat{M}_{\ell p})', \ \hat{M}_{\ell j} = \left(\frac{1}{n - n_\ell}\right)\sum_{\tilde{\ell}\neq\ell}\sum_{i\in I_{\tilde{\ell}}} D(W_i, b_j, \hat{\gamma}_{\ell,\tilde{\ell}}),$$

$$\bar{M}(\gamma) = (\bar{M}_1(\gamma), ..., \bar{M}_p(\gamma))', \ \bar{M}_j(\gamma) = \int D(W, b_j, \gamma)F_0(dW).$$

Note that $M = \bar{M}(\gamma_0)$. Let $\Gamma_{\ell,\tilde{\ell}}$ be the event that $\|\hat{\gamma}_{\ell,\tilde{\ell}} - \gamma_0\| < \varepsilon$ and note that $\Pr(\Gamma_{\ell,\tilde{\ell}}) \longrightarrow 1$ for each ℓ and $\tilde{\ell}$. When $\Gamma_{\ell,\tilde{\ell}}$ occurs,

$$\max_j |D(W_i, b_j, \hat{\gamma}_{\ell,\tilde{\ell}})| \leq B_n^D A(W_i)$$

by Assumption 11. Define

$$T_{ij}(\gamma)=D(W_i,b_j,\gamma)-\bar{M}_j(\gamma),\ (i\in I_{\tilde{\ell}}),\ U_{\tilde{\ell}j}(\gamma)=\frac{1}{n_{\tilde{\ell}}}\sum_{i\in I_{\tilde{\ell}}}T_{ij}(\gamma).$$

Note that for any constant C and the event $\mathcal{A}=\{\max_j|U_{\tilde{\ell}j}(\hat{\gamma}_{\ell,\tilde{\ell}})|\geq C\varepsilon_n^D\}$

$$\begin{aligned}\Pr(\mathcal{A})&=\Pr(\mathcal{A}|\Gamma_{\ell,\tilde{\ell}})\Pr(\Gamma_{\ell,\tilde{\ell}})+\Pr(\mathcal{A}|\Gamma_{\ell,\tilde{\ell}}^c)\left[1-\Pr(\Gamma_{\ell,\tilde{\ell}})\right]\\&\leq\Pr(\max_j|U_{\tilde{\ell}j}(\hat{\gamma}_{\ell,\tilde{\ell}})|\geq C\varepsilon_n^D|\Gamma_{\ell,\tilde{\ell}})+1-\Pr(\Gamma_{\ell,\tilde{\ell}}).\end{aligned}$$

Also

$$\Pr(\max_j|U_{\tilde{\ell}j}(\hat{\gamma}_{\ell,\tilde{\ell}})|\geq C\varepsilon_n^D|\Gamma_{\ell,\tilde{\ell}})\leq p\cdot\max_j\Pr(|U_{\tilde{\ell}j}(\hat{\gamma}_{\ell,\tilde{\ell}})|>C\varepsilon_n^D|\Gamma_{\ell,\tilde{\ell}}).$$

Note that $E[T_{ij}(\hat{\gamma}_{\ell,\tilde{\ell}})|\hat{\gamma}_{\ell,\tilde{\ell}}]=0$ for $i\in I_{\tilde{\ell}}$. Also, conditional on the event $\Gamma_{\ell,\tilde{\ell}}$,

$$|T_{ij}(\hat{\gamma}_{\ell,\tilde{\ell}})|\leq B_n^D\{A(W_i)+E[|A(W_i)|]\},\ i\in I_{\tilde{\ell}}.$$

Define $C_A=\|A(W_i)\|_{\Psi_2}+E[|A(W_i)|]$ and let $K(\hat{\gamma}_{\ell,\tilde{\ell}})=\|T_{ij}(\hat{\gamma}_{\ell,\tilde{\ell}})\|_{\Psi_2}\leq CB_n^D$, $i\in I_{\tilde{\ell}}$. By Hoeffding's inequality and the independence of $(W_i)_{i\in I_{\tilde{\ell}}}$ and $\hat{\gamma}_{\ell,\tilde{\ell}}$ there is a constant c such that

$$\begin{aligned}p\cdot\max_j\Pr(|U_{\tilde{\ell}j}(\hat{\gamma}_{\ell,\tilde{\ell}})|>C\varepsilon_n^D|\Gamma_{\ell,\tilde{\ell}})&=p\cdot\max_j E[\Pr(|U_{\tilde{\ell}j}(\hat{\gamma}_{\ell,\tilde{\ell}})|>C\varepsilon_n^D|\hat{\gamma}_{\ell,\tilde{\ell}})|\Gamma_{\ell,\tilde{\ell}}]\\&\leq 2pE[\exp\left(-\frac{cn(C\varepsilon_n^D)^2}{K(\hat{\gamma}_{\ell,\tilde{\ell}})^2}\right)|\Gamma_{\ell,\tilde{\ell}}]\leq 2p\exp\left(-\frac{cn(C\varepsilon_n^D)^2}{C_A^2(B_n^D)^2}\right)\\&\leq 2\exp\left(\ln(p)[1-\frac{cC^2}{C_A^2}]\right)\longrightarrow 0,\end{aligned}$$

for any $C>C_A/\sqrt{c}$. Let $U_{\tilde{\ell}}(\gamma)=(U_{\tilde{\ell}1}(\gamma),...,U_{\tilde{\ell}p}(\gamma))'$. It then follows from above that for large C, $\Pr(|U_{\tilde{\ell}}(\hat{\gamma}_{\ell,\tilde{\ell}})|_\infty\geq C\varepsilon_n^D)\longrightarrow 0$. Therefore $|U_{\tilde{\ell}}(\hat{\gamma}_{\ell,\tilde{\ell}})|_\infty=O_p(\varepsilon_n^D)$.

Next, for each ℓ,

$$\left|\hat{M}_\ell-\sum_{\tilde{\ell}\neq\ell}\frac{n_{\tilde{\ell}}}{n-n_\ell}\bar{M}(\hat{\gamma}_{\ell,\tilde{\ell}})\right|_\infty=\left|\sum_{\tilde{\ell}\neq\ell}\frac{n_{\tilde{\ell}}}{n-n_\ell}U_{\tilde{\ell}}(\hat{\gamma}_{\ell,\tilde{\ell}})\right|_\infty\leq\sum_{\tilde{\ell}\neq\ell}\frac{n_{\tilde{\ell}}}{n-n_\ell}|U_{\tilde{\ell}}(\hat{\gamma}_{\ell,\tilde{\ell}})|_\infty=O_p(\varepsilon_n^D).$$

Also by Assumption 9 ii) and the fact that $\Pr(\Gamma_{\ell,\tilde{\ell}})\longrightarrow 1$ for each ℓ and $\tilde{\ell}$

$$\begin{aligned}\left|\sum_{\tilde{\ell}\neq\ell}\frac{n_{\tilde{\ell}}}{n-n_\ell}\bar{M}(\hat{\gamma}_{\ell,\tilde{\ell}})-M\right|_\infty&=\left|\sum_{\tilde{\ell}\neq\ell}\frac{n_{\tilde{\ell}}}{n-n_\ell}[\bar{M}(\hat{\gamma}_{\ell,\tilde{\ell}})-M]\right|_\infty\leq B_n^\Delta\sum_{\tilde{\ell}\neq\ell}\frac{n_{\tilde{\ell}}}{n-n_\ell}\|\hat{\gamma}_{\ell,\tilde{\ell}}-\gamma_0\|\\&=O_p(B_n^\Delta\varepsilon_n^\gamma).\end{aligned}$$

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

Proof of Theorem 10: We prove the first conclusion by verifying the conditions of Lemma 14 of Chernozhukov et al. (2016). Let λ in Chernozhukov et al. (2018b) be α here and $\phi(w, \gamma, \lambda)$ in Chernozhukov et al. (2018b) be $\lambda(x)[y - \gamma(x)]$. By Assumption 6, $\varepsilon_n^\gamma \longrightarrow 0$, and $\varepsilon_n^\alpha \longrightarrow 0$ it follows that

$$\int [\phi(W, \hat{\gamma}, \lambda_0) - \phi(W, \gamma_0, \lambda_0)]^2 F_0(dW) = \int \lambda_0(X)^2 [\hat{\gamma}(X) - \gamma_0(X)]^2 F_0(dW) \leq C\|\hat{\gamma} - \gamma_0\|^2 \xrightarrow{p} 0.$$

$$\begin{aligned}\int [\phi(W, \gamma_0, \hat{\lambda}) - \phi(W, \gamma_0, \lambda_0)]^2 F_0(dW) &= \int [\hat{\lambda}(X) - \lambda_0(X)]^2 [Y - \gamma_0(X)]^2 F_0(dW) \\ &= \int [\hat{\lambda}(X) - \lambda_0(X)]^2 Var(Y|X) F_0(dX) \leq C\|\hat{\lambda} - \lambda_0\|^2 \xrightarrow{p} 0.\end{aligned}$$

Also by Assumption 6, $\int [m(W, \hat{\gamma}) - m(W, \gamma_0)]^2 F_0(dW) \xrightarrow{p} 0$, so all the conditions of Assumption 4 of Chernozhukov et al. (2018b) are satisfied.

Also by Assumptions 8 and 9 for ε_n^α given in the statement Theorem 10 and by Theorem 1 and 3 and the Cauchy-Schwartz inequality,

$$\begin{aligned}&\sqrt{n} \int |\phi(W, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi(W, \gamma_0, \hat{\lambda}_\ell) - \phi(W, \hat{\gamma}_\ell, \lambda_0) + \phi(W, \gamma_0, \lambda_0)| F_0(dW) \\ &= \sqrt{n} \int |\hat{\alpha}_\ell(X) - \alpha_0(X)||\hat{\gamma}_\ell(X) - \gamma_0(X)| F_0(dW) \leq \sqrt{n}\|\hat{\alpha}_\ell - \alpha_0\|\|\hat{\gamma}_\ell - \gamma_0\| \\ &= O_p(\sqrt{n}\Delta_n^\alpha \varepsilon_n^\gamma) \xrightarrow{p} 0.\end{aligned}$$

Therefore Assumption 5 of Chernozhukov et al. (2016) is satisfied.

Also, we have by Assumption 10

$$\begin{aligned}&\sqrt{n}\left|\int [m(W, \hat{\gamma}_\ell) - m(W, \gamma_0) + \alpha_0(X)\{Y - \hat{\gamma}_\ell(X)\}] F_0(dW)\right| \\ &= \sqrt{n}\left|\int [m(W, \hat{\gamma}_\ell) - m(W, \gamma_0) + \alpha_0(X)\{\gamma_0(X) - \hat{\gamma}_\ell(X)\}] F_0(dW)\right| \\ &= \sqrt{n}\left|\int [m(W, \hat{\gamma}_\ell) - m(W, \gamma_0) - D(W, \hat{\gamma}_\ell - \gamma_0, \gamma_0)] F_0(dW)\right| \\ &\leq C\sqrt{n}\|\hat{\gamma}_\ell - \gamma_0\|^2 = C\sqrt{n} o_p(1/\sqrt{n}) \xrightarrow{p} 0.\end{aligned}$$

Also,

$$\sqrt{n}\left|\int \hat{\alpha}(X)\{Y - \gamma_0(X)\}] F_0(dW)\right| = 0.$$

Therefore Assumption 6 of Chernozhukov et al. (2018b) is satisfied, so the first conclusion follows by Lemma 14 of Chernozhukov et al. (2018b). The second conclusion follows exactly as in the proof of Theorem 5. *Q.E.D.*

7 References

Athey, S., G. Imbens, and S. Wager (2018): “Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions,” *Journal of the Royal Statistical Society, Series B* 80, 597–623.

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. (1982): “On Adaptive Estimation,” *Annals of Statistics* 10, 647–671.

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

Bickel, P.J., C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

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.

Candes, E. and T. Tao (2007): “The Dantzig Selector: Statistical Estimation when p is much Larger than n,” *Annals of Statistics* 35, 2313–2351.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): “Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors,” *Annals of Statistics* 41, 2786–2819.

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

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey 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 J. Robins (2018): "Double/De-Biased Machine Learning Using Regularized Riesz Representers," arXiv.

Chernozhukov, V., J.A. Hausman, and W.K. Newey (2018): "Demand Analysis with Many Prices," forthcoming.

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

Hasminskii, R.Z. and I.A. Ibragimov (1979): “On the Nonparametric Estimation of Functionals,” in P. Mandl and M. Huskova (eds.), *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics, 21-25 August 1978*, Amsterdam: North-Holland, pp. 41-51.

Hausman, J.A. and W.K. Newey (2016): “Individual Heterogeneity and Average Welfare,” *Econometrica* 84, 1225–1248.

Hirshberg, D.A. and S. Wager (2018): "Augmented Minimax Linear Estimation," arXiv.

Jankova, J. and S. Van De Geer (2015): “Confidence Intervals for High-Dimensional Inverse Covariance Estimation,” *Electronic Journal of Statistics* 90, 1205–1229.

Jankova, J. and S. Van De Geer (2016a): “Semi-Parametric Efficiency Bounds and Efficient Estimation for High-Dimensional Models,” arXiv preprint arXiv:1601.00815.

Jankova, J. and S. Van De Geer (2016b): “Confidence Regions for High-Dimensional Generalized Linear Models under Sparsity,” arXiv preprint arXiv:1610.01353.

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.

Jing, B.Y., Q.M. Shao, and Q. Wang (2003): “Self-Normalized Cramér-Type Large Deviations for Independent Random Variables,” *Annals of Probability* 31, 2167–2215.

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.

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

Newey, W.K., F. Hsieh, and J.M. Robins (1998): “Undersmoothing and Bias Corrected Functional Estimation,” MIT Dept. of Economics working paper 98-17.

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

Newey, W.K. and J.M. Robins (2017): “Cross Fitting and Fast Remainder Rates for Semiparametric Estimation,” arxiv.

Neykov, M., Y. Ning, J.S. Liu, and H. Liu (2015): “A Unified Theory of Confidence Regions and Testing for High Dimensional Estimating Equations,” arXiv preprint arXiv:1510.08986.

Ning, Y. and H. Liu (2017): “A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models,” *Annals of Statistics* 45, 158-195.

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.

Robins, J.M. and A. Rotnitzky (1995): “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” *Journal of the American Statistical Association* 90 (429): 122–129.

Robins, J.M., A. Rotnitzky, and L.P. Zhao (1995): “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data,” *Journal of the American Statistical Association* 90, 106–121.

Robins, J.M., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007): “Comment: Performance of Double-Robust Estimators When ‘Inverse Probability’ Weights Are Highly Variable,” *Statistical Science* 22, 544–559.

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

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

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.

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

Stock, J.H. (1989): “Nonparametric Policy Analysis,” *Journal of the American Statistical Association* 84, 567–575.

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.

Tsybakov, A.B. (2009): *Introduction to Nonparametric Estimation.* New York: Springer.

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 Observa-*

tional and Experimental Data, Springer.

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

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

Van der Vaart, A.W. and J.A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York: Springer.

Vershynin, R. (2018): *High-Dimensional Probability*, New York: Cambridge University Press.

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 (2017): "Linear Hypothesis Testing in Dense High-Dimensional Linear Models," *Journal of the American Statistical Association* 112.

Zhu, Y. and J. Bradic (2018): "Breaking the Curse of Dimensionality in Regression," *Journal of Machine Learning Research*, forthcoming.

Zubizarreta, J.R. (2015): "Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data," *Journal of the American Statistical Association* 110, 910-922.