Automatic Debiased Machine Learning of Causal and Structural Effects[*]

Victor Chernozhukov *MIT*

Whitney K. Newey *MIT*

Rahul Singh *MIT*

January 29, 2021

Abstract

Many causal and structural effects depend on regressions. Examples include policy effects, average treatment effects, causal mediation, and structural parameters of economic models. The regressions may be high dimensional, making machine learning useful. Plugging machine learners into identifying equations can lead to poor inference due to bias from regularization and/or model selection. This paper gives automatic debiasing for linear and nonlinear functions of regressions. The debiasing is automatic in using Lasso and the function of interest without the full form of the bias correction. The debiasing can be applied to any regression learner, including neural nets, random forests, Lasso, boosting, and other high dimensional methods. In addition to providing the bias correction we give standard errors that are robust to misspecification, convergence rates for the bias correction, and primitive conditions for asymptotic inference for a variety of estimators of structural and causal effects. The automatic debiased machine learning is applied to estimating the average treatment effect on the treated for the NSW job training data and to estimating demand elasticities from Nielsen scanner data while allowing preferences to be correlated with prices and income.

Keywords: Debiased machine learning, causal parameters, structural parameters, regression effects, Lasso, Riesz representation.

[*] The present paper formed the basis of the Fisher-Schultz Lecture given by Victor Chernozhukov at the 2019 European Meeting of the Econometric Society in Manchester. This research was supported by NSF grant 1757140. Jamie Robins and Yinchu Zhu provided helpful comments.

1 Introduction

Many causal and structural parameters of economic interest depend on regressions, i.e. on conditional expectations or least squares projections. Examples include policy effects, average derivatives, regression decompositions, average treatment effects, causal mediation, and parameters of economic structural models. Often, regressions may be high dimensional, depending on many variables. There may be many covariates for policy effects, average derivatives, treatment effects, and many prices and covariates in the economic demand for some commodity. This paper is about estimating economic and causal parameters that depend on high dimensional regressions.

Machine learning is a collection of modern, adaptive statistical learning methods for estimating regression functions and other statistical objects. These methods exploit structured parsimony restrictions (such as approximate sparsity) on regressions, together with various forms of regularization and model selection, to enable high quality prediction in high dimensional settings. Key methods include neural nets (deep learning), random forests, and Lasso. The goal of this paper is to deploy these methods to infer causal and structural parameters that depend on regression functions, including treatment, policy, derivative, and decomposition effects as well as economic structural parameters.

Machine learning is different than other methods in ways that are useful in high dimensional settings. For example, Lasso has good properties with 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 (the regression is approximately sparse). In contrast, series regression is based on relatively few regressors, often much fewer than the sample size. Lasso and series regression 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 Lasso their identity need not be known. For Lasso, the important regressors just need to be included somewhere among the many potential regressors. 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 Lasso, such information is not needed, since very many terms can be included among the potential regressors. Other machine learning methods, such as random forests and neural nets, are also well suited to high dimensional regression.

Machine learners provide remarkably good predictions in a variety of settings but are inherently biased. The bias arises from using regularization and/or model selection to control the variance of the prediction. To obtain small mean squared prediction errors, machine learners regularize and/or select among models so that variance and squared bias are approximately equal. Although such equality is good for prediction, it is not good for inference. Confidence intervals based on estimators with approximately equal variance and squared bias will tend to have poor coverage. This inference problem can be even worse when machine learners are plugged into a formula for a causal or structural effect. These formulae often involve averaging over regressor values which reduces variance without affecting much the bias. Variance could also potentially also be a problem but machine learners control that for prediction purposes.

The squared bias of causal and structural estimators that plug-in regularized machine learners can shrink slower than the variance, leading to extremely poor confidence interval coverage and estimators that are not root-n consistent. Chernozhukov et al. (2017, 2018) give Lasso and random forest examples respectively and Chernozhukov et al. (2020) shows that Lasso plug-in estimators are not root-n consistent. Model selection inherent in machine learners also creates inference problems. Model selection creates bias from incorrect model choice under local alternatives, making the usual asymptotic confidence intervals invalid over local alternatives, as shown by Leeb and Potscher (2008a,b). Estimators of parameters of interest obtained by plugging in machine learners can inherit this problem, as pointed out by Belloni, Chernozhukov, and Kato (2015) and Chernozhukov, Hansen, and Spindler (2015) and shown in Chernozhukov et al. (2020).

To reduce regularization and model selection bias we use a Neyman orthogonal moment function where there is no first-order effect of the regression on the expected moment function. The orthogonal moment function is constructed by adding to an identifying moment the nonparametric influence function of the regression on the identifying moment function. This construction is model free, being based on the probability limit of the regression learner for a general distribution, as in Chernozhukov et al. (2016, 2020). As a result the orthogonality property and standard errors associated with this moment function are model free, including being robust to assumptions about the regression.

The orthogonal moment function depends on another unknown function $\bar{\alpha}$ in addition to the regression. We develop a Lasso minimum distance learner of $\bar{\alpha}$ that is automatic, in the sense that it depends only on the identifying moment function and not on the form of $\bar{\alpha}$. The structure of the identifying moment function is used to approximate $\bar{\alpha}$ as a linear combination of a dictionary (i.e. basis) of known functions. We plug the Lasso learner of $\bar{\alpha}$ and a regression learner into the orthogonal moment functions to construct an automatic debiased machine learner (Auto-DML) of parameters of interest. We give estimators for a wide variety of effects, including policy effects, average derivatives, regression decompositions, and bounds on average equivalent variation where debiased machine learners were not previously available. We allow for the identifying moment functions to be nonlinear in regressions. We also give novel estimators of average treatment effects and causal mediation.

We allow any regression learner, including neural nets, random forests, Lasso, and other high dimensional learners to be used in the orthogonal moment function. The primary requirement of the regression learner is that the product of mean-square convergence rates for the learner of $\bar{\alpha}$ and the regression learner is faster than $n^{-1/2}$. Under this condition and a few other regularity conditions we show root-n consistency and asymptotic normality of the learner (estimator) of the parameter of interest. We give convergence rates for the Lasso learner of $\bar{\alpha}$ and combine them with existing convergence rates for regressions to verify conditions for particular estimators. A learner of $\bar{\alpha}$ and large sample theory is given for parameters that depend nonlinearly on regressions as well as parameters that are linear in a regression.

The large sample theory in this paper takes the probability limit of the regression learner and $\bar{\alpha}$ to be fixed. It would be straightforward to extend the results to allow the regression limit and $\bar{\alpha}$ to change with sample size. Such a change would allow us to accommodate sparse specifications where number of nonzero coefficients in the true regression grows with the sample size but would complicate notation and detail. We choose to work with a fixed regression for simplicity while accommodating high dimensional regressions via approximate sparsity.

We give an application to estimating the treatment effect on the treated of job training from the National Supported Work Demonstration (NSW). For many large sets of covariates, we find similar estimates based on neural net, random forest, and Lasso regressions with the automatic bias correction for each. We also give an application to estimating price elasticities from scanner panel data while allowing endogeneity of prices. We estimate the elasticities from Auto-DML of an average derivative that includes many covariates that account for correlated random effects. We find price elasticities that are much smaller than cross-section elasticities, consistent with though larger than fixed effects elasticities found in Chernozhukov, Hausman, and Newey (2019). We also find that plug in estimates are similar to the cross-section elasticity estimates, so that debiasing is important in this application.

The estimators of objects of interest we give use cross-fitting, as in Chernozhukov et al. (2018), where orthogonal moment functions are averaged over groups of observations, the regression and $\bar{\alpha}$ learners use all observations not in the group, and each observation is included in the average over one group. Cross-fitting removes a source of bias and eliminates any need for Donsker conditions for the regression learner. Early work that used similar sample splitting ideas includes Bickel (1982), Schick (1986), and Klaasen (1987).

An Auto-DML for average equivalent variation with a Lasso learner of $\bar{\alpha}$ was given in Chernozhukov, Hausman, and Newey (2016). Convergence rates and asymptotic normality results based on the Dantzig selector were given in Chernozhukov and Newey (2018) and Chernozhukov, Newey, and Robins (2018). Chernozhukov, Newey, and Singh (2018) gave convergence rates for a Lasso learner of $\bar{\alpha}$ while allowing for any regression learner and nonlinear functions of regressions. This paper builds on and supersedes Chernozhukov, Newey, and Singh (2018). Chernozhukov, Newey, and Singh (2019) builds on and supersedes Chernozhukov, Newey, and Robins (2018) and is distinguished from this paper and previous work in giving and analyzing Auto-DML for local (nonparametric) effects. All of these papers make use of model free, estimator based orthogonal moment functions for regression learners given in Chernozuhkov et al. (2016), with the automatic debiasing in Chernozhukov et al. (2020) building on this paper. The combined use of cross-fitting and orthogonal moment functions for debiased machine learning is like Chernozhukov et al. (2018). The Auto-DML given here innovates by not requiring an explicit formula for the bias correction that is required in Chernozhukov et al. (2018) and earlier papers.

This work builds upon ideas in classical semi- and nonparametric learning theory with low-dimensional regressions using traditional smoothing methods (Van Der Vaart, 1991; Newey 1994; Bickel et al., 1993; Robins and Rotnitzky, 1995; Van der Vaart, 1998), that do not apply to the current high-dimensional setting. The orthogonal moment functions developed in Chernozhukov et al. (2016) and used here build on previous work on model free orthogonal moment functions. Hasminskii and Ibragimov (1979) and Bickel and Ritov (1988) suggest such estimators for functionals of a density. Newey (1994) develops such scores for densities and regressions from computation of the semiparametric efficiency bound for regular functionals. Doubly robust estimating equations for treatment effects as in Robins et al. (1995) and Robins and Rotnitzky (1995) constitute model based orthogonal moment functions and have motivated much subsequent work. Newey, Hsieh, and Robins (1998, 2004) extend model free orthogonal moment functions to any functional of a density or distribution in a low dimensional setting. Model free, orthogonal moments for any learner are given and their general properties derived in Chernozhukov et al. (2016, 2020). We use those model free, orthogonal moment functions for regressions.

This paper also builds upon and contributes to the literature on modern orthogonal/debiased estimation and inference, including Zhang and Zhang (2014), Belloni et al. (2011, 2014a,b), Robins et al. (2013), van der Laan and Rose (2011), Javanmard and Montanari (2014a,b, 2015), Van de Geer et al. (2014), Farrell (2015), Ning and Liu (2017), Chernozhukov et al. (2015), Neykov et al. (2018), Ren et al. (2015), Jankova and Van De Geer (2015, 2016a, 2016b), Bradic and Kolar (2017), Zhu and Bradic (2017a,b). This prior work is about regression coefficients, treatment effects, and semiparametric likelihood models. The objects of interest 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 the Riesz representer, an object that is 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 et al. (2016). We use moment function methods with automatic debiasing rather than the likelihood based method for a known form for orthogonal moments in van der Laan and Rose (2011).

Various papers have considered direct estimation of $\bar{\alpha}$ for treatment effects, where $\bar{\alpha}$ is a

Riesz representer that depends on inverse propensity scores. Our work is the first to present a framework for direct estimation of the Riesz representer of a broad class of linear and nonlinear functionals, in a high-dimensional setting, without requiring strong Donsker class assumptions. The earliest reference of which we know is Robins et al.(2007), which gives a linear estimator for $\bar{\alpha}$ for only the average treatment effect. Vermeulen and Vansteelandt (2015) base parametric propensity score and regression estimators on double robustness conditions for the average treatment effect. We differ in using a linear approximation to $\bar{\alpha}$, which is restrictive in a parametric setting but is general in high dimensional and/or nonparametric settings. Newey and Robins (2018) present and analyze estimators based on regression splines, while we present and analyze sparse methods for the high-dimensional setting. The Lasso minimum distance learner of $\bar{\alpha}$ given in Chernozhukov, Newey, and Singh (2018) and here is a direct estimator of the Riesz representer for a broad class of linear and nonlinear functionals that can be interpreted as being based on orthogonality of the moment functions. Chernozhukov et al. (2020) extends this learner of $\bar{\alpha}$ to functionals of high dimensional regression quantiles and other functions.

In independent work on treatment effects Avagyan, V. and S. Vansteelandt (2017) give a model assisted estimator based on regularized first order conditions and Tan (2020) developed a model assisted, multistep method of doubly robust estimation with Lasso type regression learners having standard errors that are robust to misspecification of the regression or propensity score. For treatment effects the estimator here is simpler in being single step, allows for any regression learner (e.g. neural nets), and is model free and has correct standard errors if either or both the regression and the propensity score are misspecified. Also Farrell et al. (2020) gave a neural nets and model based estimator of the average treatment effect and Wooldridge and Zhu (2020) give a Lasso based debiased machine learner for panel data with correlated random effects that depend on high dimensional regressions. Our results also allow for a neural net regression learner but are model free with specification robust standard error.

Chernozhukov, Newey, and Robins (2018) gave Auto-DML for linear functionals using the Dantzig selector. More recently Hirshberg and Wager (2018) developed augmented linear estimators for the linear functionals in the low dimensional case where the regression estimator is in a Donsker class using the same orthogonal moment functions considered here. Chernozhukov, Newey, and Singh (2018) gave Auto-DML for linear and nonlinear functionals, including average derivatives, of any regression leaner. More recently Hirshberg and Wager (2020) gave Lasso based debiasing for an average derivative of a Lasso learner of a single index high dimensional regression and Rothenhausler and Yu (2019) Lasso based debiasing for Lasso regression of an average derivative. Singh and Sun (2019) extend the present work to the instrumental variable setting and present estimators of the local average treatment effect, average complier characteristics, and complier counter factual distributions. Previous to the current paper Farbmacher et al. (2020) gave a DML for causal mediation. More recently we give here Auto-DML for causal mediation analysis as an example in Section 5

In summary, contributions of the paper include the construction of DML for a wide range of interesting policy effects and structural parameters where DML was not previously available. This construction is based on a Lasso minimum distance learner of $\bar{\alpha}$ that is developed here. Also, the debiasing and inference is model free and robust to misspecification, and carried out in a single step, unlike previous estimators of average treatment effects. For average treatment effects we construct DML for a variety of regression learners, such as neural nets, random forests, or high dimensional methods.

In Section 2 we describe objects of interest including examples and associated orthogonal moment functions. In Section 3 we give the Lasso learner of $\bar{\alpha}$, the Auto-DML estimator, and a consistent estimator of its asymptotic variance. Section 4 derives mean square convergence rates for the Lasso learner of $\bar{\alpha}$ and conditions for root-n consistency and asymptotic normality of Auto-DML including primitive conditions in examples. Section 5 gives Auto-DML for nonlinear functionals of multiple regressions and as an example develops Auto-DML for causal mediation analysis. Section 6 gives Auto-DML for regression decomposition and applies this to estimating the average treatment on the treated for the NSW experiment. Section 7 gives Auto-DML estimates of price elasticities that allow for correlated random effects in scanner panel data. Section 8 offers some conclusions and possible extensions.

2 Average Linear Effects and Orthogonal Moment Functions

For expositional purposes, we consider in this Section parameters that depend linearly on a single 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 X is a covariate vector. Denote the conditional expectation of Y given $X \in \mathcal{X}$ as

$$\gamma_0(x) = E[Y|X = x].$$

Let $m(w, \gamma)$ denote a function of the function γ (i.e. a functional of γ), where γ denotes a possible conditional expectation function $\gamma : \mathcal{X} \longrightarrow \mathbb{R}$, that depends on a data observation w and is linear in γ. We will consider effects of the form

$$\theta_0 = E[m(W, \gamma_0)].$$

The parameter of interest here is an expectation of some known formula $m(W, \gamma)$ of a data observation W and a regression γ.

We also give results in later Sections for important parameters having more general forms. In Section 5 we generalize to allow $m(W, \gamma)$ to be nonlinear in multiple regressions and an estimator of causal effects with mediation. In Section 6 we give estimators of regression decompositions and their properties. These important examples extend the framework of this Section to parameters that are nonlinear in multiple regressions

Several important examples of linear effects are:

Example 1: (Average Policy Effect). An average effect of a counter factual shift in the distribution of regressors from a known F_0 to another known F_1, when γ_0 does not vary with the distribution of X, is

$$\theta_0 = \int \gamma_0(x) d\mu(x); \;\; \mu(x) = F_1(x) - F_0(x).$$

Here $m(w, \gamma) = \int \gamma(x) d\mu(x)$ which does not depend on w.

Example 2: (Weighted Average Derivative). Here $X = (D, Z)$ for a continuously distributed random variable D, $\gamma_0(x) = \gamma_0(d, x)$, $\omega(d)$ is a pdf, and

$$\theta_0 = E\left[\int \omega(u) \frac{\partial \gamma_0(u, Z)}{\partial d} du\right] = E\left[\int S(u)\gamma_0(u, Z)\omega(u)du\right] = E[S(U)\gamma_0(U, Z)],$$

where $S(u) = -\omega(u)^{-1}\partial\omega(u)/\partial u$ is the negative score for the pdf $\omega(u)$, the second equality follows by integration by parts, and U is a random variable that is independent of Z with pdf $\omega(u)$. This U could be thought of as one simulation draw from the pdf $\omega(u)$. Here $m(w, \gamma) = S(u)\gamma(u, x)$ where W includes U.

This θ_0 can be interpreted as an average treatment effect on Y of a continuous treatment D in a model where $Y = Y(D)$ for a potential outcome stochastic process $Y(d)$ that is independent of D conditional on covariates Z. By conditional independence

$$E[\gamma_0(u, Z)] = \int E[Y(D)|D = u, Z = z]F_Z(dz) = E[Y(u)|Z = z]F_Z(dz) = E[Y(u)],$$

for $\omega(u) > 0$ assuming that the joint pdf of (D, X) is positive where $\omega(D) > 0$, as in Chamberlain (1984), Wooldridge (2001), and Blundell and Powell (2004). The $E[Y(u)]$ is the average outcome at $D = u$ and is sometimes referred to as the average structural function. Assuming that we can interchange the order of differentiation and integration,

$$\theta_0 = \int \omega(u)\frac{\partial E[\gamma_0(u, Z)]}{\partial u} du = \int \frac{\partial E[Y(u)]}{\partial u}\omega(u)du = \int E[\frac{\partial Y(u)}{\partial u}]\omega(u)du,$$

similarly to Imbens and Newey (2009) and Rothenhäusler and Yu (2019). Regarding $E[\partial Y(u)/\partial u]$ as the average treatment effect at u we see that θ_0 is an average treatment effect. Alternatively, θ_0 can be regarded as an average derivative of the average structural function. The averaging over a known pdf $\omega(u)$ helps fulfill regularity conditions for the Auto-DML developed here that can be used to estimate θ_0 for high dimensional covariates Z.

Example 3: (Average Treatment Effect). Here $X = (D, Z)$ and $\gamma_0(x) = \gamma_0(d, z)$, where $D \in \{0, 1\}$ is the treatment indicator and Z are covariates. The object of interest is

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

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 $m(w, \gamma) = \gamma(1, z) - \gamma(0, z)$.

Example 4: (Average Equivalent Variation Bound). An economic example is a bound on average equivalent variation for heterogenous demand. 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, $\omega(z)$ some weight function, and U be a random variable that is uniformly distributed over $(\check{p}_1, \bar{p}_1)$ and independent of (Y, X). The U can be thought of as one simulation draw from a uniform distribution on $(\check{p}_1, \bar{p}_1)$. The object of interest is

$$\theta_0 = E\left[\Lambda(U, Z)\gamma_0(U, Z)\right], \; \Lambda(u, z) = \omega(z)1(\check{p}_1 < u < \bar{p}_1)(\bar{p}_1 - \check{p}_1)\frac{z_1}{u}\exp(-\kappa[u - \check{p}_1]).$$

If 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 for all individuals, then θ_0 is an upper (lower) bound on the weighted average over consumers of 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 $m(w, \gamma) = \Lambda(u, z)\gamma(u, z)$, where W includes U.

We focus on $m(w, \gamma)$ where there exists a function $\alpha_0(X)$ with $E[\alpha_0(X)^2] < \infty$ and

$$\mathrm{E}[m(W, \gamma)] = \mathrm{E}[\alpha_0(X)\gamma(X)] \text{ for all } \gamma \text{ such that } \mathrm{E}[\gamma(X)^2] < \infty. \tag{2.1}$$

By the Riesz representation theorem, existence of such a $\alpha_0(X)$ is equivalent to $E[m(W, \gamma)]$ being a mean-square continuous functional of γ, i.e. $E[m(W, \gamma)] \leq C \|\gamma\|$ for all γ, where $\|\gamma\| = \sqrt{E[\gamma(X)^2]}$ and $C > 0$. 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 stated in Newey (1994) and shown in Hirshberg and Wager (2018) for conditional expectations and in Chernozhukov, Newey, and Singh (2019) more generally for least squares projections. Thus, in assuming existence of $\alpha_0(X)$ we are just assuming that θ_0 has a finite semiparametric variance bound.

Each of Examples 1-4 has such a Rr. Let $f(x)$ denote the pdf of X in Example 1, $f(d|z)$ the pdf of D conditional on Z in Example 2, $\pi_0(z) = \Pr(D = 1|Z = z)$ the propensity score in Example 3, and $f(p_1|z)$ the pdf of P_1 conditional on Z in Example 4. Table 1 summarizes the functional $m(w, \gamma)$ and the Rr in each of the examples:

Table 1: m and Rr for Examples 1-4		
Effect	$m(W,\gamma)$	Riesz Representer
Policy Effect	$\int \gamma(x)[f_1(x)-f_0(x)]dx$	$f(X)^{-1}[f_1(X)-f_0(X)]$
Weighted Average Derivative	$S(U)\gamma(U,Z)$	$f(D\|Z)^{-1}\omega(D)S(D)$
Average Treatment Effect	$\gamma(1,Z)-\gamma(0,Z)$	$\pi_0(Z)^{-1}D-(1-\pi_0(Z))^{-1}(1-D)$
Equivalent Variation Bound	$\Lambda(U,Z)\gamma(U,Z)$	$(\bar{p}_1-\check{p}_1)^{-1}f(P_1\|Z)^{-1}\Lambda(P_1,Z)$

Equation (2.1) follows in Example 1 by multiplying and dividing by $f(x)$ inside the integral, in Example 2 by integration by multiplying and dividing by $f(d|z)$, in Example 3 in a standard way for average treatment effects, and in Example 4 by multiplying and dividing by $f(p_1|z)$. For $E[\alpha_0(X)^2]<\infty$ to hold the denominator must not be too small relative to the numerator in each $\alpha_0(X)$, on average. For instance Example 3 must have $E[\{\pi_0(Z)(1-\pi_0(Z))\}^{-1}]<\infty$.

Equation (2.1) implies that the effect of interest can be represented in three different ways, as

$$\theta_0 = E[m(W,\gamma_0)] = E[\alpha_0(X)\gamma_0(X)] = E[\alpha_0(X)Y],$$

where the last equality follows by iterated expectations. Any of these three expressions could be used to estimate θ_0. We could estimate θ_0 from the first expression using a learner (estimator) of γ_0. We could also estimate θ_0 from the last expression using a learner of $\alpha_0(X)$. In addition we could use learners of both γ_0 and α_0 to estimate θ_0 from the middle expression. We focus here on using a learner of γ_0, though α_0 will be important for the bias correction to follow.

We rely on a regression learner (estimator) $\hat{\gamma}$ of γ_0 to estimate θ_0. The $\hat{\gamma}$ can be any of a variety of machine learners including neural nets, random forests, Lasso, and other high dimensional methods. All we require is that $\hat{\gamma}$ converge in mean square at a sufficiently fast rate, as specified in Section 4.

Whatever the choice of $\hat{\gamma}$, estimating θ_0 by plugging $\hat{\gamma}$ into $m(W,\gamma)$ and averaging over observations on W can lead to large biases when $\hat{\gamma}$ involves regularization and/or model selection, as discussed in the Introduction. For that reason we use an orthogonal moment function for θ_0, where the regression learner $\hat{\gamma}$ has no first-order effect on the moments. We follow Chernozhukov et al. (2016, 2020) in basing the orthogonal moment function on the probability limit (plim) $\gamma(F)$ of $\hat{\gamma}$ when one observation W has CDF F, where F is unrestricted except for regularity conditions. Here $\gamma(F)$ can be thought of as the plim of $\hat{\gamma}$ under general misspecification, where $\gamma(F)$ need not be the conditional expectation $E_F[Y|X]$.

The plim $\gamma(F)$ of $\hat{\gamma}$ depends on the learner. For example Lasso, the Dantzig selector, boosting, and other high dimensional methods are based on a sequence of potential regressors $X=(X_1,X_2,...)$. These learners have the form

$$\hat{\gamma}(x)=\sum_{j=1}^{\infty}\hat{\beta}_j x_j,\ \hat{\beta}_{j'}\neq 0 \text{ for a finite number of } j',$$

where $x=(x_1,x_2,...)$ denotes a possible realization of X. Because each $\hat{\gamma}(X)$ is a linear combination of $X=(X_1,X_2,...)$ the plim $\gamma(F)$ of $\hat{\gamma}$ will also be a linear combination of X, or at least will be approximated by such a linear combination. Define Γ to be the mean square closure of the set of finite linear combinations of X, i.e. Γ is the set of $\gamma(X)$ such that $E[\gamma(X)^2]<\infty$ and for every $\varepsilon>0$ there exists $(\beta_j^\varepsilon)_{j=1}^\infty$ such that $\beta_{j'}^\varepsilon\neq 0$ for a finite number of j' and $E[\{\gamma(X)-\sum_{j=1}^\infty \beta_j^\varepsilon X_j\}^2]<\varepsilon$. It will be the case that $\gamma(F)\in\Gamma$. Because Lasso and other high dimensional methods are being used for least squares prediction of Y it will also be the case that

$$\gamma(F)=\arg\min_{\gamma\in\Gamma} E_F[\{Y-\gamma(X)\}^2], \tag{2.2}$$

This $\gamma(F)$ minimizes population least squares criteria over the (mean square closure of) linear combinations of X, i.e. it is the best linear predictor of Y by linear combinations of X. Here $\gamma(F)$ is the infinite dimensional linear regression that is nonpararmetrically estimated by Lasso and other high dimensional methods.

Neural nets and random forests may have a different $\gamma(F)$. A neural net or random forest is often a nonparametric regression estimator for a finite (but high) dimensional X. In that case

$$\gamma(F)=E_F[Y|X],$$

which satisfies equation (2.2) when Γ is the set of all (measurable) functions of X with finite second moment. The plim of Lasso and other high dimensional methods will also be this $\gamma(F)$ if $X=(X_1,X_2,...)$ can approximate any function of a fixed set of regressors, but otherwise will not. A third type of learner $\hat{\gamma}$ is one that imposes additivity restrictions on $\hat{\gamma}$, such as $\hat{\gamma}(X)=\hat{\gamma}_1(X_1)+\hat{\gamma}_2(X_2)$, allowing for nonparametric learners $\hat{\gamma}_1(X_1)$ and $\hat{\gamma}_2(X_2)$. In that case $\gamma(F)$ will be satisfy equation (2.2) where Γ is the mean square closure of functions that are additive in X_1 and X_2.

We use the orthogonal moment function from Chernozhukov et al. (2016, 2020) for a regression learner $\hat{\gamma}$ having plim $\gamma(F)$ satisfying equation (2.2) for any linear, closed Γ. The orthogonal moment function is constructed by adding to the identifying moment function $m(w,\gamma)-\theta$ the nonparametric influence function of of $E[m(W,\gamma(F))]$. As shown in Newey (1994) the nonparametric influence function of $E[m(W,\gamma(F))]$ is

$$\bar{\alpha}(X)[Y-\bar{\gamma}(X)],$$

where $\bar{\gamma}(X)$ is the solution to equation (2.2) for $F=F_0$ and $\bar{\alpha}\in\Gamma$ satisfies $E[m(W,\gamma)]=E[\bar{\alpha}(X)\gamma(X)]$ for all $\gamma\in\Gamma$. As in Chernozhukov, Newey, and Singh (2019),

$$\bar{\alpha}=\arg\min_{\alpha\in\Gamma} E[\{\alpha_0(X)-\alpha(X)\}^2]. \tag{2.3}$$

Evaluating the nonparametric influence function at possible values γ and α of $\bar{\gamma}$ and $\bar{\alpha}$ and adding it to the the identifying moment function gives the orthogonal moment function

$$\psi(w,\theta,\gamma,\alpha)=m(w,\gamma)-\theta+\alpha(x)[y-\gamma(x)]. \tag{2.4}$$

The moment function $\psi(w, \theta, \gamma, \alpha)$ depends on a possible value α of the unknown function $\bar{\alpha}$ as well as a possible value γ of the plim $\bar{\gamma}$ of the regression learner. A learner $\hat{\alpha}$ of $\bar{\alpha}$ is needed to use this orthogonal moment function to estimate θ_0. In Section 3 we will describe how to construct $\hat{\alpha}$. In Chernozhukov et al. (2016, 2020) $\psi(w, \theta, \gamma, \alpha)$ is shown to be orthogonal without being specific about the form of $\hat{\alpha}$. For exposition we repeat that demonstration here. Consider any $\gamma, \alpha \in \Gamma$, representing possible realizations of learners $\hat{\gamma}$ and $\hat{\alpha}$ that are in Γ. The well known necessary and sufficient conditions for equation (2.2) with $F = F_0$ are that $E[\alpha(X)\{Y - \bar{\gamma}(X)\}] = 0$ for all $\alpha \in \Gamma$. Therefore

$$\begin{aligned} E[\psi(W, \theta, \gamma, \alpha) - \psi(W, \theta, \bar{\gamma}, \bar{\alpha})] &= E[m(W, \gamma)] - E[m(W, \bar{\gamma})] + E[\alpha(X)\{Y - \gamma(X)\}] \\ &= E[\alpha_0(X)\{\gamma(X) - \bar{\gamma}(X)\}] + E[\alpha(X)\{Y - \gamma(X)\}] \\ &= E[\bar{\alpha}(X)\{\gamma(X) - \bar{\gamma}(X)\}] + E[\alpha(X)\{\bar{\gamma}(X) - \gamma(X)\}] \\ &= -E[\{\alpha(X) - \bar{\alpha}(X)\}\{\gamma(X) - \bar{\gamma}(X)\}], \end{aligned} \tag{2.5}$$

where the second equality follows by equation (2.1) and the third equality by the necessary and sufficient condition for equation (2.3) that $E[\{\alpha_0(X) - \bar{\alpha}(X)\}\gamma(X)] = 0$ for all $\gamma \in \Gamma$. Here we see that $\psi(w, \theta, \gamma, \bar{\alpha})$ "partials out" γ in the sense that

$$E[m(W, \gamma) + \bar{\alpha}(X)\{Y - \gamma(X)\}] = E[m(W, \bar{\gamma})]$$

does not depend on γ. Also equation (2.5) gives an explicit formula showing that the effect of γ and α on $E[\psi(W, \theta, \gamma, \alpha)]$ is second order and hence $\psi(W, \theta, \gamma, \alpha)$ is orthogonal.

The orthogonality property of $\psi(W, \theta, \gamma, \alpha)$ only depends on γ, $\alpha \in \Gamma$ and $\bar{\gamma}$ satisfying equation (2.2). In particular orthogonality does not depend on either $\bar{\gamma}$ being $E[Y|X]$ or on $\bar{\alpha} = \alpha_0$. In this sense orthogonality of $\psi(W, \theta, \gamma, \alpha)$ is model free, i.e. nonparametric. Consequently the estimator of θ will be asymptotically normal and standard errors consistent even if either $\bar{\gamma} \neq \gamma_0$ or $\bar{\alpha} \neq \alpha_0$ or both. This robustness of the standard errors results from the orthogonality of the moments only depending on the $\bar{\gamma}$ limit of the regression estimator, so that the sample average of the estimated orthogonal moment function will be asymptotically equivalent to the sample average at the truth, without an model assumptions.

The orthogonal moment function could also be viewed as the efficient influence function of $E[m(W, \bar{\gamma})]$ which clarifies that the Auto-DML is an efficient semiparametric estimator of $E[m(W, \bar{\gamma})]$. Viewing $\psi(w, \theta, \gamma, \alpha)$ in this way is not useful for debiasing because the results of Chernozhukov et. al. (2016, 2020) already imply model free orthogonality.

The moment function $\psi(w, \theta, \gamma, \alpha)$ is doubly robust for estimation of the true parameter θ_0. Evaluating at $\theta_0, \bar{\gamma}, \bar{\alpha}$ and taking the expectation gives

$$\begin{aligned} E[\psi(W, \theta_0, \bar{\gamma}, \bar{\alpha})] &= E[m(W, \bar{\gamma})] - \theta_0 + E[\bar{\alpha}(X)\{Y - \bar{\gamma}(X)\}] \\ &= E[\alpha_0(X)\{\bar{\gamma}(X) - \gamma_0(X)\}] + E[\bar{\alpha}(X)\{\gamma_0(X) - \bar{\gamma}(X)\}] \\ &= -E[\{\bar{\alpha}(X) - \alpha_0(X)\}\{\bar{\gamma}(X) - \gamma_0(X)\}], \end{aligned} \tag{2.6}$$

which is zero for $\bar{\gamma} = \gamma_0$ or $\bar{\alpha} = \alpha_0$. Thus $E[\psi(W, \theta_0, \bar{\gamma}, \bar{\alpha})] = 0$, so that the orthogonal moment condition identifies θ_0, when either $\bar{\gamma}(X) = E[Y|X]$ or $\alpha_0(X) \in \Gamma$. These conditions both hold when the regression learner is nonparametric so that Γ is the set of all functions of X with finite second moment. For high dimensional regressions where Γ is the closed linear span of $X = (X_1, X_2, ...)$ the plim of the learner $\hat{\gamma}$ may not be $E[Y|X]$ but the orthogonal moment function still identifies θ_0 when $\alpha_0(X) \in \Gamma$. That is, θ_0 is identified when $\alpha_0(X)$ can be approximated arbitrarily well in mean square by a linear combination of X. This robustness condition can be interpreted in each of Examples 1-4:

Example 1: For high dimensional $\hat{\gamma}$, where Γ is the mean square closure of linear combinations of X, $E[\psi(W, \theta_0, \bar{\gamma}, \bar{\alpha})] = 0$ even when $\bar{\gamma}(X) \neq E[Y|X]$ if $\alpha_0(X) = [f_1(X) - f_0(X)]/f(X) \in \Gamma$.

Example 2: For high dimensional $\hat{\gamma}$, where Γ is the mean square closure of linear combinations of X, $E[\psi(W, \theta_0, \bar{\gamma}, \bar{\alpha})] = 0$ even when $\bar{\gamma}(X) \neq E[Y|X]$ if $\alpha_0(X) = f(D|Z)^{-1}\omega(D)S(D) \in \Gamma$.

Example 3: For the average treatment effect where Γ is nonparametric, so that $\bar{\gamma}(X) = E[Y|X]$ and $\bar{\alpha}(X) = \alpha_0(X)$, the orthogonal moment function in equation (2.4) corresponds to the seminal doubly robust moment function of Robins, Rotnitzky, and Zhao (1994). When $\hat{\gamma}$ is high dimensional, with say $X = (DZ, (1-D)\tilde{Z})$ for sequences $Z = (Z_1, Z_2, ...)$ and $\tilde{Z} = (\tilde{Z}_1, \tilde{Z}_2, ...)$, with each $\tilde{Z}_j$ a function of Z, the orthogonal moment function is

$$\psi(W, \theta, \bar{\gamma}, \bar{\alpha}) = \bar{\gamma}(1, Z) - \bar{\gamma}(1, 0) - \theta + \bar{\alpha}(X)[Y - \bar{\gamma}(X)].$$

This orthogonal moment function is different than those previously considered in $\bar{\alpha}(X)$ being the projection of $\alpha_0(X)$ on Γ rather than $\alpha_0(X)$. Here $E[\psi(W, \theta_0, \bar{\gamma}, \bar{\alpha})] = 0$ if linear combinations of Z and $\tilde{Z}$ can approximate abitrarily well $\pi_0(Z)^{-1}$ and $[1 - \pi_0(Z)]^{-1}$ respectively, even when $\bar{\gamma}(X) \neq E[Y|X]$.

For brevity we omit further discussion of Example 4 from the paper and refer the interested reader to Chernozhukov, Hausman, and Newey (2019).

3 Estimation

To estimate (learn) θ_0 we use cross-fitting where the orthogonal moment function $\psi(w, \gamma, \alpha, \theta)$ is averaged over observations different than used to estimate $\bar{\gamma}$ and $\bar{\alpha}$. We assume that the data W_i, $(i = 1, ..., n)$ are 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. In practice $L = 5$ (5-fold) or $L = 10$ (10-fold) cross-fitting is often used. 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 $\hat{\theta}$ and an associated asymptotic variance estimator $\hat{V}$ have explicit forms

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}, \tag{3.1}$$

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

Any regression learner $\hat{\gamma}_\ell$ can be used here as long as its mean-square convergence rate is a power of $1/n$, as assumed in Section 4. Such a convergence rate is available for neural nets (Chen and White, 1999, and Farrell, Liang, and Misra, 2018), random forests (Syrgkanis and Zampetakis, 2020), boosting (Luo and Spindler, 2016), Lasso (Bickel, Ritov, and Tsybakov, 2009), and other high dimensional methods. As a result any of these regression learners can be used to construct an Auto-DML $\hat{\theta}$ from equation (3.1), in conjunction with a learner $\hat{\alpha}_\ell$ of $\bar{\alpha}$.

To describe $\hat{\alpha}_\ell$ let $b(x) = (b_1(x), ..., b_p(x))$ be a $p \times 1$ dictionary of functions of x, where p can be large, with each $b_j(x)$ standardized to have mean 0 and standard deviation 1, to be further discussed later in this Section. For convenience we ignore dependence of $b(x)$ on the data in the notation. The learner $\hat{\alpha}_\ell$ given here is

$$\hat{\alpha}_\ell(x) = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} m(W_i,1) + b(x)'\hat{\rho}_\ell,\ \hat{\rho}_\ell = \arg\min_\rho\{-2\hat{M}_\ell'\rho + \rho'\hat{G}_\ell\rho + 2r\sum_{j=1}^{J}|\rho_j|\}, \tag{3.2}$$

$$\hat{M}_\ell = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} m(W_i,b),\ \hat{G}_\ell = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} b(X_i)b(X_i)',$$

where n_ℓ is the number of observations not in I_ℓ and $r > 0$ is a positive scalar. This $\hat{\alpha}_\ell$ is used in equation (3.1) to construct $\hat{\theta}$ and $\hat{V}$.

To explain and motivate $\hat{\alpha}_\ell$ it is notationally convenient to drop the ℓ subscript, with the understanding that $\hat{\alpha}_\ell$ is computed using only observations not in I_ℓ for each ℓ, as in equation (3.2). It is also notationally convenient to drop the 0 mean normalization of $b(x)$ and consider $\hat{\alpha}$ having the form

$$\hat{\alpha}(x) = b(x)'\hat{\rho}, \tag{3.3}$$

where $\hat{\rho}$ is a vector of estimated coefficients.

The $\hat{\alpha}$ depends on the choice of dictionary $b(x)$ and penalty degree r. For the dictionary we require that each $b_j(x)$ belongs to the set Γ of possible plims of $\hat{\gamma}(x)$ discussed in Section 2 and that linear combinations of the dictionary "span" Γ.

Assumption 1: $b(x) = (b_1(x), ..., b_p(x))'$ where i) $b_j \in \Gamma$ for all j and ii) for any $\alpha \in \Gamma$ and $\varepsilon > 0$ there is p and $\rho \in \mathbb{R}^p$ such that $E[\{\alpha(X) - b(X)'\rho\}^2] < \varepsilon$.

One key feature of this condition is that each $b_j \in \Gamma$. This feature allows us to use $m(w, \gamma)$ to construct $\hat{\alpha}$ and will guarantee that $\hat{\alpha} \in \Gamma$, as required for the orthogonality shown in equation (2.5). Another key feature is that linear combinations of $b(x)$ can approximate anything that belongs to Γ. This feature will lead to $\hat{\alpha}$ estimating $\bar{\alpha}$. The link imposed by Assumption 1, between the regression learner $\hat{\gamma}$ and the dictionary $b(x)$ used to construct $\hat{\alpha}$, is important for the orthogonality property of $\psi(w, \gamma, \alpha, \theta)$ and hence for $\hat{\theta}$ to be asymptotically normal and $\hat{V}$ to be a consistent estimator of the asymptotic variance under general misspecification.

Assumption 1 requires that linear combinations of $b(x)$ must be able to approximate any γ in the set of possible plims of $\hat{\gamma}$ and that each b_j must be a possible plim of $\hat{\gamma}$. For Lasso and other high dimensional regression learners where $X = (X_1, X_2, ...)$ Assumption 1 will be satisfied for

$$b(x) = (x_1, ..., x_p)'. \tag{3.4}$$

Evidently each element $b_j(X) = X_j$ is an element of Γ and the spanning condition is satisfied because any linear combination of X with a finite number of nonzero coefficients will also be a linear combination of $b(x)$ for p large.

For neural nets, random forests, and other learners that nonparametrically estimate $E[Y|X]$, Assumption 1 will require that a linear combination of $b(X)$ can approximate any function of X for large enough p. Such a $b(x)$ can be formed from low order multivariate powers of components of x, with a full set of approximating functions included as p grows. In applications one may use a variety of nonlinear functions including powers of transformations of X.

The learner $\hat{\alpha}$ also depends on the choice of penalty degree r. An important, useful feature of Lasso is that $r = A\sqrt{\ln(p)/n}$ for a constant A gives the fastest possible mean square convergence rate for Lasso, that optimally trades off bias and variance. In Appendix A, we describe cross-validation and theoretical methods for choosing the choosing r based on data that have proven stable across several different applications. We also provide R code, available upon request, for the construction of $\hat{\alpha}(x)$ and $\hat{\theta}$.

We can motivate $\hat{\rho}$ in $\hat{\alpha}(x) = b(x)'\hat{\rho}$ as being based on the Riesz representation in equation (2.1) and $\bar{\alpha}$ satisfying equation (2.3), which imply that for $m(w, b) = (m(w, b_1), ..., m(w, b_p))'$,

$$M := E[m(W, b)] = E[\alpha_0(X)b(X)] = E[\bar{\alpha}(X)b(X)], \tag{3.5}$$

where the last equality is satisfied by $b_j \in \Gamma$, which implies $E[b_j(X)\{\alpha_0(X) - \bar{\alpha}(X)\}] = 0$ for each j. We see that the cross moments M between the true, unknown $\bar{\alpha}(x)$ and the dictionary $b(x)$ are equal to the expectation of the known vector of functions $m(w, b)$. Also, the second moment matrix $G = E[b(X)b(X)']$ of the dictionary is an expectation of a known function of the data. Estimating M and G enables learning coefficients ρ of the least squares regression of $\bar{\alpha}(X)$ on $b(X)$, satisfying $M = G\rho$. We learn ρ using a Lasso minimum distance objective function to allow for large p. Let

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

be unbiased estimators of M and G. The coefficient estimator is given by

$$\hat{\rho} = \arg\min_{\rho}\{-2\hat{M}'\rho + \rho'\hat{G}\rho + 2r\left\|\rho\right\|_1\}, \ \left\|\rho\right\|_1 = \sum_{j=1}^{p}|\rho_j|. \tag{3.6}$$

The estimator $\hat{\rho}$ can be interpreted as a minimum distance version of Lasso. Here $\hat{M}$ is analogous to $\sum_{i=1}^{n} Y_i b(X_i)/n$ in Lasso. The objective function in equation (3.6) can be thought of as the Lasso objective with $\sum_{i=1}^{n} Y_i b(X_i)/n$ replaced by $\hat{M}$ and $\sum_{i=1}^{n} Y_i^2/n$ dropped. In this way the objective function is a penalized approximation to the least squares regression of $\alpha_0(x)$ on $b(x)$, where $2r\left\|\rho\right\|_1$ is the penalty. We refer to this as minimum distance Lasso because $\hat{M}$ does not have the product form of Lasso regression.

The learner $\hat{\alpha}(x)$ of $\bar{\alpha}(x)$ is automatic in being based on $\hat{M}$ and $\hat{G}$, neither of which requires knowledge of the form of $\bar{\alpha}$. In particular, $\hat{\alpha}(x) = b(x)'\hat{\rho}$ does not depend on plugging in nonparametric estimates of components of $\bar{\alpha}(x)$. Instead, $b(x)'\hat{\rho}$ is linear in the dictionary $b(x)$ and uses the known function $m(w, \gamma)$ in the construction of $\hat{M}$ to obtain the learner $\hat{\rho}$. This automatic nature of $\hat{\alpha}(x)$ is especially useful for Lasso and other high dimensional regression learners where $b(x)$ can be taken to be the first p elements of $x = (x_1, x_2, ...)$, and where $\bar{\alpha}(x)$ is a least squares projection of $\alpha_0(X)$ on Γ, as in Section 2. The projection $\bar{\alpha}(x)$ will generally not have a simple form that can be learned by plugging in nonparametric learners to an explicit formula.

The learner $\hat{\alpha}(x) = b(x)'\hat{\rho}$ also avoids inverting a learner of a conditional probability or pdf. The finite sample properties of methods that rely on inverses of learners can be poor; see Singh and Sun (2019) for recent examples. Instead, $\hat{\alpha}$ approximates and learns $\bar{\alpha}$ by a linear combination of functions. In this way the $\hat{\alpha}$ given here avoids potential instability from inverting a high dimensional estimator. The inverse of a conditional probability or density is present in $\alpha_0(x)$ in all of the examples in this paper. We anticipate that this feature is present quite generally for causal and structural models involving shifts in regressors, because the Rr equation (2.1) involves an expectation with respect to the data distribution rather than the shifted distribution. Thus absence of an inverse of a machine learner in $\hat{\alpha}$ may prove to be widely useful. In some economic structural models the linearity of $\hat{\alpha}$ in $b(x)$ may not be quite as appealing, because inverse densities can have semiparametric form and so mitigate the problem of inverting a high dimensional learner. An example is the dynamic discrete choice learner of Chernozhukov et al. (2016, 2020).

This learner $\hat{\alpha}(x)$ can be thought of as being based on orthogonality of the moment function with respect to γ. Let τ denote a scalar and $b_j(x)$ an element of $b(x)$. Then by equation (3.5)

$$\frac{\partial}{\partial \tau} E[\psi(W, \theta, \gamma + \tau b_j, \bar{\alpha})] = E[m(W, b_j) - \bar{\alpha}(X) b_j(X)] = 0, \ (j = 1, ..., p).$$

Replacing the expectation by a sample average and $\bar{\alpha}(X)$ by $b(X)'\rho$ gives

$$\frac{1}{n} \sum_{i=1}^{n} \{m(W_i, b_j) - [b(X_i)'\rho] b_j(X_i)\} = \hat{M} - \hat{G}\rho.$$

This sample average is a scaled version of the derivative of objective function in equation (3.6) without the penalty term. The first-order conditions for equation (3.6) will set $\hat{\rho}$ so that this object is close to zero, subject to the penalty, i.e. will solve penalized versions of a moment equation. Thus, the Lasso minimum distance learner can be thought of as a way of using orthogonality of $\psi(W, \theta, \gamma, \alpha)$ with respect to γ to learn $\bar{\alpha}$ while penalizing to facilitate high dimensional estimation. In Section 6 we use an extension of this approach to construct an Auto-DML when $m(W, \gamma)$ is nonlinear in γ.

To illustrate $\hat{\alpha}$ we consider the choice of dictionary and the form of $\hat{\alpha}$ for Examples 1-3.

Example 1: If the regression learner $\hat{\gamma}$ is nonparametric the dictionary $b(X)$ should also be nonparametric while if $\hat{\gamma}$ is a high dimensional regression the dictionary should be chosen as in equation (3.4). Here $m(w, b) = \int b(x)[f_1(x) - f_0(x)]dx$ does not depend on the data observation w and the first order conditions for $\hat{\rho}$ imply that for each j,

$$\left| \int b_j(x)[f_1(x) - f_0(x)]dx - \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} b_j(X_i) \hat{\alpha}_\ell(X_i) \right| \leq r.$$

Here $\hat{\alpha}_\ell(X_i)$ acts to approximately re-weight so that the integral of $b_j(x)$ over the policy shift is approximately equal to the sample average of the re-weighted function $b_j(X_i)\hat{\alpha}_\ell(X_i)$.

Example 2: The dictionary $b(X)$ should be chosen as in Example 1. Also by $m(w, b) = S(u)\gamma(u, z)$ the first order conditions for $\hat{\rho}$ imply that for each j,

$$\left| \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} \{S(U_i) b_j(U_i, Z_i) - b_j(X_i) \hat{\alpha}_\ell(X_i)\} \right| \leq r.$$

Here $\hat{\alpha}_\ell(X_i)$ acts approximately as a re-weighting scheme, making the $b_j(U_i, Z_i)$ be approximately equal to the sample average of the re-weighted function $b_j(X_i)\hat{\alpha}_\ell(X_i)$.

Example 3: The dictionary should be chosen similarly to Example 1. For instance suppose that $X = (DZ, (1-D)Z)$, where $Z = (Z_1, Z_2, ...)$ is a sequence or possible covariates. Then the dictionary

$$b(x) = (dq(z)', (1-d)q(z)')', \ q(z) = (z_1, ..., z_{p/2})', \tag{3.7}$$

would satisfy Assumption 1. The estimator $\hat{\alpha}_\ell$ has an interesting form for this dictionary. Note that $m(w, b) = b(1, z) - b(0, z) = (q(z)', 0')' - (0', q\,(z)')' = (q(z)', -q(z)')$. Then

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

Let $\hat{\rho}^1_\ell$ be the estimated coefficients of $dq(z)$ and $\hat{\rho}^0_\ell$ be the estimated coefficients of $(1-d)q(z)$. Then the learner of $\bar{\alpha}(X_i)$ is

$$\hat{\alpha}_\ell(X_i) = D_i \hat{\omega}^1_{\ell i} - (1 - D_i)\hat{\omega}^0_{\ell i}, \ \hat{\omega}^1_{\ell i} = q(Z_i)'\hat{\rho}^1_\ell, \ \hat{\omega}^0_{\ell i} = -q(Z_i)'\hat{\rho}^0_\ell,$$

where $\hat{\omega}^1_{\ell i}$ and $\hat{\omega}^0_{\ell i}$ might be thought of as "weights." These weights sum to one if $q(z)$ includes a constant but may be negative. The first order conditions for $\hat{\alpha}$ are that for each j,

$$\left| \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} q_j(Z_i)[1 - D_i \hat{\omega}^1_{\ell i}] \right| \le r, \ \left| \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} q_j(Z_i)[1 - (1 - D_i)\omega^0_{\ell i}] \right| \le r. \tag{3.8}$$

Here $\hat{\rho}_\ell$ sets the weights $\hat{\omega}^1_{\ell i}$ and $\hat{\omega}^0_{\ell i}$ to approximately "balance" the overall sample average with the treated and untreated averages for each element of the dictionary $q(z)$. The constraints of equation (3.8) are like the balancing conditions of Zubizarreta (2015) and Athey, Imbens, and Wager (2018). The source of these constraints is regularized least squares approximation of $\bar{\alpha}(x) = proj(\pi_0(z)^{-1}d - [1 - \pi_0(z)]^{-1}(1-d)|Z)$ by a linear combination of the dictionary $b(x)$. The approach of this paper shows that this type of balancing is sufficient to debias any regression learner under regularity conditions in Section 4.

4 Large Sample Inference

In this Section, we give mean square convergence rates for the Lasso minimum distance learner of $\hat{\alpha}$ and root-n consistency and asymptotic normality results for the learner $\hat{\theta}$ of the object of interest and its asymptotic variance estimator $\hat{V}$. Let ε_n denote a sequence that converges to zero no faster than $\sqrt{\ln(p)/n}$ and for a random variable $a(W)$ let $\|a\| = \sqrt{E[a(W)^2]}$

Assumption 2: *There exists $C > 1, \xi > 0$ such that for each positive integer $s \le C\varepsilon_n^{-2/(2\xi+1)}$ there is $\bar{\rho}$ with s nonzero elements such that*

$$\|\bar{\alpha} - b'\bar{\rho}\| \le C(s)^{-\xi}.$$

Here $\|\bar{\alpha} - b'\bar{\rho}\|$ is the mean square approximation error from using the linear combination $b'\bar{\rho}$ to approximate $\bar{\alpha}$. This approximate sparsity condition specifies that there is a sparse $\bar{\rho}$, having only s nonzero elements, so that the approximation error is bounded by $C(s)^{-\xi}$. Note that it is not required that $\bar{\alpha}$ be equal to linear combination of s terms, i.e. it is not required that α_0 be strictly sparse. Assumption 2 does allow unknown identity for the elements of $b(x)$ that give the approximation rate $s^{-\xi}$. In this way this condition allows for high dimensional x where statistics and economics to not provide much guidance for which elements of $b(x)$ are important.

The ε_n in this condition represents a convergence rate for $\hat{M}$ and $\hat{G}$ that will be no faster than $\sqrt{\ln(p)/n}$ under the conditions given here. When s is chosen to be approximately $C\varepsilon_n^{-2/(2\xi+1)}$, which is the largest s allowed by this Assumption 2, s will grow no faster than $(\sqrt{n/\ln(p)})^{2/(2\xi+1)} \leq n^{1/(2\xi+1)}$, which grows slower than n. Because $p \geq s$ is implicitly required by this condition, Assumption 2 puts a quite a weak restriction on p. Note though that an important feature of Assumption 2 is that the sparse approximation is based on functions included in the $p \times 1$ dictionary $b(x)$. Thus larger values of p give more flexibility and will help Assumption 2 to be satisfied.

Our results will require a convergence rate for $\hat{\alpha}$ that is faster than some power of n and Assumption 2 is a natural condition that leads to such a rate. Sufficient conditions for such a rate are well known from the approximation literature when $\bar{\alpha}(x)$ belongs to a Besov or Holder class of function and linear combinations of $b(x)$ can approximate any function of x.

We will also make use of a sparse eigenvalue condition as considered in much of the Lasso literature. Let ρ denote a $p \times 1$ vector, ρ_J a $J \times 1$ subvector of ρ, and ρ_{J^c} the vector consisting of components of ρ that are not in ρ_J. Also for a matrix A let $\|A\|_1 = \sum_{i,j} |a_{ij}|$.

Assumption 3: *$G = E[b(X)b(X)']$ has largest eigenvalue bounded uniformly in n and there is $C, c > 0$ such that for all $s \approx C\varepsilon_n^{-2}$ with probability approaching one*

$$\min_{J \leq s} \min_{\|\rho_{J^c}\|_1 \leq 3\|\rho_J\|_1} \frac{\rho'\hat{G}\rho}{\rho_J'\rho_J} \geq c$$

This is a sparse eigenvalue condition that is familiar from the Lasso literature, including Bickel, Ritov, Tsybakov (2009), Belloni and Chernozhukov (2013), and Rudelson and Zhou (2013).

We will work with a dictionary $b(X)$ with elements that are uniformly bounded.

Assumption 4: *There is $C > 0$ such that with probability one $\sup_j |b_j(X)| \leq C$.*

This condition implies a convergence rate of $\sqrt{\ln(p)/n}$ for $\left\|\hat{G} - G\right\|_\infty$, where $\|A\|_\infty = \max_{i,j} |a_{ij}|$ for a matrix $A = [a_{ij}]$.

Lasso mean square convergence rates are often stated in terms of finite sample bounds. Because the focus of this paper is root-n consistency for $\hat{\theta}$ and for that we only need convergence at certain powers of n we can simplify the statement of convergence rates without affecting the conditions for $\hat{\theta}$ by allowing the Lasso regularization value r to shrink slightly slower than ε_n. This does lead to approximate sparseness conditions that are strict inequalities on the size of ξ but Bradic et al. have shown that strict inequalities are necessary for root-n consistent estimation, meaning that there is no loss of generality in these conditions. We also limit the growth of p to be slower than some power of n.

Assumption 5: $\varepsilon_n = o(r)$, $r = o(n^c \varepsilon_n)$ for all $c > 0$, and there $C > 0$ such that $p \leq Cn^C$.

We also hypothesize a convergence rate for $\hat{M}$.

Assumption 6: $\left\| \hat{M} - M \right\|_\infty = O_p(\varepsilon_n)$ for $\varepsilon_n \longrightarrow 0$.

We use this condition to accommodate $\hat{M}$ that can depend on the regression learner $\hat{\gamma}$ as needed for Section 5.

Theorem 1: *If Assumptions 1 - 6 are satisfied then for all* $c > 0$,

$$\|\hat{\alpha} - \bar{\alpha}\| = o_p(n^c \varepsilon_n^{2\xi/(2\xi+1)}).$$

This theorem is based on extending Lemmas of Bradic et al. (2021) to allow ε_n to shrink slower than $\sqrt{\ln(p)/n}$. The extension will be used in Section 5 to obtain convergence rates when $\hat{M}$ depends on a nonparametric estimator.

The sparse eigenvalue condition of Assumption 3 seems strong in some settings. It is possible to drop Assumption 3 and Assumption 2 if the following condition is satisfied:

Assumption 7: $\bar{\alpha}(X) = \sum_{j=1}^{\infty} \rho_{j0} b_j(X)$, $\sum_{j=1}^{\infty} |\rho_{j0}| < \infty$, *and for* $C > 0$ *and* $\bar{s} = C\sqrt{n}$ *the* $b_j(x)$ *corresponding to the largest* s *values of* $|\rho_{j0}|$ *are included in* $b(x)$.

This condition allows us to drop Assumption 2 because absolute summability of the coefficients ρ_{0j} implies a sparse approximation rate of $\xi = 1/2$. It also allows $\hat{G}$ to converge at a rate slower ε_n in order to accommodate nonparametric estimation in $\hat{G}$.

Theorem 2: *If Assumptions 1 and 5-7 are satisfied and* $\left\| \hat{G} - G \right\|_\infty = O_p(\varepsilon_n)$ *then for all* $c > 0$,

$$\|\hat{\alpha} - \bar{\alpha}\| = o_p(n^c \sqrt{\varepsilon_n}).$$

This result extends Chatterjee and Javarov (2015) to allow ε_n to shrink slower than $\sqrt{\ln(p)/n}$. When $\varepsilon_n = \sqrt{\ln(p)/n}$ in Assumption 6 this result gives a mean square convergence rate for $\hat{\alpha}$ that is faster than $n^{-1/4+c}$ for all $c > 0$, without a sparse eigenvalue condition.

We now use these results to obtain root-n consistency and asymptotic normality for the Auto-DML $\hat{\theta}$ and consistency of its asymptotic variance estimator $\hat{V}$. We impose some additional regularity conditions.

Assumption 8: *There is $C > 0$ such that with probability one* $\max_{j\leq p} |m(W, b_j))| \leq C$.

Under this condition Assumption 6 will be satisfied with $\varepsilon_n = \sqrt{\ln(p)/n}$. This condition will be satisfied under by Assumption 4 in each of Examples 1-3 under conditions of Corollaries 4-6 to follow.

Assumption 9: *$E[\{Y - \bar{\gamma}(X)\}^2|X]$ and $\bar{\alpha}(X)$ are bounded.*

We impose this condition for simplicity; it could be weakened. We also impose the following condition.

Assumption 10: *$E[m(W, \gamma_0)^2] < \infty$ and $\int [m(w, \hat{\gamma}) - m(w, \bar{\gamma})]^2 F_W(dw) \xrightarrow{p} 0$.*

This condition will be implied by existence of $C > 0$ with $|E[m(W, \gamma)^2]| \leq C \|\gamma\|^2$ for all γ, which will be satisfied in the examples we consider under regularity conditions to be specified.

Assumption 11: *With probability approaching one $\hat{\gamma}_\ell \in \Gamma$ and there is $d_\gamma > 0$ such that $\|\hat{\gamma} - \bar{\gamma}\| = O_p(n^{-d_\gamma})$ and either Assumptions 2 and 3 are satisfied with*

$$\frac{\xi}{2\xi + 1} + d_\gamma > \frac{1}{2}, \tag{4.1}$$

or Assumption 7 is satisfied and $d_\gamma > 1/4$.

This assumption allows $\hat{\gamma}$ to be any learner that converges in mean square at a rate that is some power of n. By Theorem 1 the mean square convergence rate for $\hat{\alpha}$ is as close as desired to $n^{-\xi/(2\xi+1)}$, so Assumption 11 is the rate double robustness condition of Belloni, Chernozhukov, and Hansen (2014) and Farrell (2015) that the product of convergence rates for $\hat{\alpha}$ and $\hat{\gamma}$ must go to zero faster than $1/\sqrt{n}$. Under Assumptions 2 and 3 a full trade-off in rates between $\hat{\alpha}$ and $\hat{\gamma}$ is permitted, with Assumption 11 being satisfied for any ξ if d_γ is large enough and for any d_γ if ξ is large enough. Under Assumption 7 this trade-off is not present, with $d_\gamma > 1/4$ being required by Assumption 11.

The following gives the large sample inference results for $\hat{\theta}$ and $\hat{V}$. Define

$$\bar{\theta} = E[m(W, \bar{\gamma})],\ \psi(w) = m(w, \bar{\gamma}) - \bar{\theta} + \bar{\alpha}(x)[y - \bar{\gamma}(x)],\ V = E[\psi(W)^2].$$

Here $\bar{\theta}$ will be the object estimated by $\hat{\theta}$ when neither of the double robustness conditions $\bar{\gamma}(X) = E[Y|X]$ or $\bar{\alpha}(X) \in \Gamma$ are satisfied.

Theorem 3: *If Assumptions 1-5, and 8-11 are satisfied then* $\sqrt{n}(\hat{\theta} - \bar{\theta}) \xrightarrow{d} N(0, V)$. *If in addition Assumption 7 is satisfied then* $\hat{V} \xrightarrow{p} V$.

It is possible to construct a consistent estimator of V without Assumption 7 by using a trimmed version of $\hat{\alpha}_\ell(x)$ but we omit that demonstration to avoid further complicating $\hat{V}$. The conclusion of Theorem 3 implies that asymptotic test statistics and confidence intervals can be formed in the usual manner from $\hat{\theta}$ and $\hat{V}$. Theorem 3 is proven by using the convergence rate results of Theorem 1 and Theorem 2 to show that the hypotheses of Lemma 15 of Chernozhukuv et al. (2020) are satisfied.

Most of the conditions of Theorem 3 are quite general, with only Assumptions 8 and 10 pertaining to a particular $m(w, \gamma)$. It is straightforward to specify conditions under which Assumptions 8 and 10 are satisfied for Examples 1-3.

Corollary 4 (Example 1): *If Assumptions 1-5, 9, and 11 are satisfied and there is* $C > 0$ *such that* $|[f_1(x) - f_0(x)]/f(x)| \leq C$ *then* $\sqrt{n}(\hat{\theta} - \bar{\theta}) \xrightarrow{d} N(0, V)$. *If in addition Assumption 7 is satisfied then* $\hat{V} \xrightarrow{p} V$.

The specific regularity condition for the policy effect in Corollary 4 is that the Rr $\alpha_0(X) = [f_1(X) - f_0(X)]/f(x)$ be bounded.

Corollary 5 (Example 2): *If Assumptions 1-5, 9, and 11 are satisfied and there is* $C > 0$ *such that* $|S(u)| \leq C$, $f(D|Z)^{-1} |\omega(D)S(D)| \leq C$ *then* $\sqrt{n}(\hat{\theta} - \bar{\theta}) \xrightarrow{d} N(0, V)$. *If in addition Assumption 7 is satisfied then* $\hat{V} \xrightarrow{p} V$.

The regularity conditions for the weighted average derivative in Corollary 5 are that the score $S(u)$ be bounded and the Rr $\alpha_0(X) = f(D|Z)^{-1}\omega(D)S(D)$ also be bounded.

Corollary 6 (Example 3): *If Assumptions 1, 4-5, 9, and 11 are satisfied and there is* $C > 0$ *with* $\pi_0(Z) \in [C, 1 - C]$ *then* $\sqrt{n}(\hat{\theta} - \bar{\theta}) \xrightarrow{d} N(0, V)$. *If in addition Assumption 7 is satisfied then* $\hat{V} \xrightarrow{p} V$.

The additional condition in Corollary 6 is that the propensity score be bounded away from 0 and 1, an overlap condition that is common in asymptotic theory for estimators of the average treatment effect. Together Corollaries 4-6 demonstrate how simple primitive conditions involving $m(w, \gamma)$ can be specified so that the Auto-DML $\hat{\theta}$ of an object of interest will be asymptotically normal and the asymptotic variance estimator $\hat{V}$ consistent.

5 Nonlinear Effects of Multiple Regressions

Some important effects of interest are expectations of nonlinear functions of multiple regressions. Causal mediation analysis is an important example that we consider in this Section. The regression decomposition in Section 6 is also an important example. In this Section we give Auto-DML for such effects. Such effects have the form $\theta_0 = E[m(W, \gamma_0)]$ where $m(w, \gamma)$ is nonlinear in a possible value γ of multiple regressions $(\gamma_1(X_1), ..., \gamma_K(X_K))'$ with regressors X_k specific to each regression $\gamma_k(X_k)$. The corresponding orthogonal moment functions are like those discussed in Section 3 except that the bias correction is a sum of K terms with the k^{th} term being the bias correction for the learner of γ_k, as in Newey (1994, p. 1357). The estimated bias corrections are like those of Section 4 with the k^{th} term being the product of a Lasso learner $\hat{\alpha}_{k\ell}(X_k)$ and the residual $Y_k - \hat{\gamma}_{k\ell}(X_k)$. Each $\hat{\alpha}_{k\ell}(X_k)$ differs from Section 3 in the corresponding $\hat{M}_{k\ell}$ being a derivative evaluated at a preliminary estimator of $\bar{\gamma}$. Because the construction of $\hat{\theta}$ is so closely related to that in Section 3 we proceed immediately with its description here and fill in details concerning the orthogonal moment function below.

The Auto-DML of a nonlinear effect is similar to equation (3.1) in being

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i, \hat{\gamma}_\ell) + \sum_{k=1}^{K}\hat{\alpha}_{k\ell}(X_{ki})[Y_{ki} - \hat{\gamma}_{k\ell}(X_{ki})]\}, \tag{5.1}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}^2,\ \hat{\psi}_{i\ell} = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \sum_{k=1}^{K}\hat{\alpha}_{k\ell}(X_{ki})[Y_{ki} - \hat{\gamma}_{k\ell}(X_{ki})],$$

where each $\hat{\alpha}_{k\ell}(X_{ki})$ is obtained as follows: For each k let $b_k(x_k) = (b_{k1}(x_k), ..., b_{kp}(x_k))'$ be a $p\times 1$ dictionary vector specific to the k^{th} regression $\gamma_k(x_k)$ and let $\hat{\gamma}_{\ell,\ell^i}$ be the vector of regressions computed from all observations not in either I_ℓ or $I_{\ell'}$. Also let τ denote a scalar, and e_k the k^{th} column of the K dimensional identity matrix. Then

$$\hat{\alpha}_{k\ell}(X_{ki}) = b_k(X_{ki})'\hat{\rho}_{k\ell},\ \hat{\rho}_{k\ell} = \arg\min_{\rho}\{-2\hat{M}_{k\ell}'\rho + \rho'\hat{G}_{k\ell}\rho + 2r_k\left\|\rho\right\|_1\},\ \ \left\|\rho\right\|_1 = \sum_{j=1}^{p}|\rho_j|, \tag{5.2}$$

$$\hat{M}_{k\ell} = (\hat{M}_{k\ell 1}, ..., \hat{M}_{k\ell p})',\ \hat{G}_{k\ell} = \left(\frac{1}{n - n_\ell}\right)\sum_{i\notin I_\ell} b_k(X_{ki})b_k(X_{ki})',$$

$$\hat{M}_{k\ell j} = \frac{d}{d\tau}\left(\frac{1}{n - n_\ell}\right)\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}} m(W_i, \hat{\gamma}_{\ell,\ell'} + \tau e_k b_{kj})\Bigg|_{\tau=0},\ (j = 1, ..., p).$$

where b_{kj} denotes the j^{th} element of the dictionary $b_k(x_k)$ as a function of x_k. Thus the $\hat{\alpha}_{k\ell}(X_i)$ in equation (5.1) is a Lasso minimum distance estimator like that of Section 3 that is specific to $\hat{\gamma}_k$ and uses the $\hat{M}_{k\ell}$ from equation (5.2) rather than the one in equation (3.2).

The $\hat{M}_{k\ell j}$ given here generalizes equation (3.2) to allow for nonlinearity of $m(w, \gamma)$ in γ. The derivative with respect to the scalar τ in $\hat{M}_{k\ell j}$ is generally simple to compute analytically using the chain rule of calculus, as we will illustrate for causal mediation analysis. When $m(w,\gamma)$ is linear in a single γ this derivative just evaluates $m(W_i,\gamma)$ at $\gamma = b_j$, giving the $\hat{M}_{\ell j}$ of equation (3.2). As with linear $m(w,\gamma)$ the $\hat{M}_{k\ell j}$ and the rest of the $\hat{\theta}$ depends just on $m(w,\gamma)$ and the first step. Thus the $\hat{\theta}$ in equation (5.1) is automatic, in the same way as the estimator of equation (3.1), in only requiring $m(w,\gamma)$ and the regression residuals $Y$for its construction.

The $\hat{M}_{k\ell j}$ given here does depend on a cross-fit regression learner $\hat{\gamma}_{\ell,\ell'}$ in order to allow for the nonlinearity of $m(w,\gamma)$ in γ. The cross-fitting will make the sample average used in the construction of $\hat{M}_{k\ell j}$ independent of the regression learner $\hat{\gamma}_{\ell,\ell'}$ used in its construction. This independence helps $\hat{M}_{k\ell j}$ be uniformly consistent over $j = 1,...,p$ for large p with only mean square convergence convergence rates for $\hat{\gamma}_{\ell,\ell'}$. This feature of the theory helps $\hat{\theta}$ to be root-n consistent and asymptotically normal for a wide variety of regression learners $\hat{\gamma}_{\ell,\ell'}$.

The dictionary $b_k(x_k)$ used in the construction of $\hat{\alpha}_{k\ell}(x_k)$ should be chosen analogously to the $b(x)$ in Section 3. Each b_{kj} should be an element of the set Γ_k of possible plim's of $\hat{\gamma}_k$. Also linear combinations of $b_k(x_k)$ should be able to approximate any element of Γ_k arbitrarily well in mean square. That is, Assumption 1 should be satisfied with Γ_k and $b_k(x)$ replacing Γ and $b(x)$ respectively. In particular if $\hat{\gamma}_k$ is a high dimensional regression then $b(x) = (x_{k1},...,x_{kp})'$ will do. If $\hat{\gamma}_k$ is a nonparametric estimator then $b_k(x_k)$ should be chosen so that linear combinations can approximate any function of x_k.

An important difference between the Lasso minimum distance learner in Section 3 and each $\hat{\alpha}_{k\ell}(x_k)$ here is that the penalty size r_k must be chosen to be larger $\sqrt{\ln(p)/n}$ when $m(w,\gamma)$ depends nonlinearly on γ. The reason for larger r_k is that $\hat{M}_{k\ell}$ depends on the machine learner $\hat{\gamma}_{\ell,\ell'}$ and so will converge at a slower rate, leading to a requirement that r_k converge to zero slightly slower than the mean square convergence rate of $\hat{\gamma}_{\ell,\ell'}$. A choice of r_k proportional to $n^{-1/4}$ will generally suffice for this purpose, since $\hat{\gamma}_{\ell,\ell'}$ will be required to converge faster than $n^{-1/4}$.

This estimator will not be doubly robust due to the nonlinearity of $m(w,\gamma)$ in γ; see Chernozhukov et al. (2016). 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 $\hat{\alpha}_{k\ell}(x_k)$ will consistently estimate $\bar{\alpha}_k(x_k)$ such that $\sum_{k=1}^{K}\bar{\alpha}_k(x)[y_k - \bar{\gamma}_k(x_k)]$ is the influence function for $E[m(W,\gamma(F))]$ at $\gamma(F) = \bar{\gamma}$ where $\gamma(F)$ =plim$(\hat{\gamma})$.

Example 5: (Causal Mediation Analysis) Causal mediation analysis provides an interesting example of a nonlinear function of multiple regressions. This effect allows for intermediate variables, referred to mediators, that lie between treatment and outcome. In this example there is an outcome variable Y, a treatment indicator $D \in \{0,1\}$, and covariates Z similar to the average treatment effect in Example 3. In addition there is a mediation variable that we will denote by Q, where we assume that $Q \in \{1, ..., K-1)$ for an integer $K \geq 3$. Let

$\gamma_{K0}(D, Q, Z) = E[Y|D, Q, Z]$, $\gamma_{k0}(D, Z) = \Pr(Q = k|D, Z) = E[1(Q = k)|D, Z]$, $(k = 1, ..., K-1)$.

The causal mediation effect of Imai, Keele, and Tingley (2010, Theorem 1) is

$$\theta_0(d, d') = E[\sum_{k=1}^{K-1} \gamma_{K0}(d, k, Z)\gamma_{k0}(d', Z)].$$

This effect, or parameter, has the form $\theta_0(d, d') = E[m(W, \gamma)]$ for $W = (Y, D, Q, Z)$ and

$$m(W, \gamma) = \sum_{k=1}^{K-1} \gamma_K(d, k, Z)\gamma_k(d', Z).$$

In this example we have $X_k = (D, Z)$, $(k = 1, ..., K-1)$ and $X_K = (D, Q, Z)$. To construct the Auto-DML $\hat{\theta}$ we need to choose the dictionaries $b_k(X_k)$ for each k. We choose

$$b_K(X_K) = (b_{K1}(D, Q, Z), ..., b_{Kp}(D, Q, Z))'$$

to be a nonparametric dictionary if $\hat{\gamma}_K$ is a nonparametric estimator such as a neural net or random forest or choose $b_K(D, Q, Z)$ to be the leading p regressors used in a high dimension regression learner $\hat{\gamma}_K$. For $k \leq K-1$ we choose the same dictionary $b_k(X_k) = b_1(D, Z)$ with

$$b_1(D, Z) = (b_{11}(D, Z), ..., b_{1p}(D, Z))',$$

for each $k \leq K-1$. We specify $b_1(D, Z)$ to be a nonparametric dictionary if each $\hat{\gamma}_k$ is a nonparametric estimator such as a neural net or random forest or choose $b_1(D, Z)$ to be the leading p regressors used in a high dimension regression learner for each $\hat{\gamma}_k$.

It is straightforward to compute each $\hat{M}_{k\ell j}$. Note that for $k \leq K-1$,

$$\begin{aligned}
\left.\frac{d}{d\tau} m(W, \gamma + \tau e_k b_{kj})\right|_{\tau=0} &= \left.\frac{d}{d\tau}\gamma_K(d, k, Z)\{\gamma_k(d', Z) + \tau b_{1j}(d', Z)\}\right|_{\tau=0} = \gamma_K(d, k, Z) b_{1j}(d', Z), \\
\left.\frac{d}{d\tau} m(W, \gamma + \tau e_K b_{Kj})\right|_{\tau=0} &= \left.\frac{d}{d\tau}\{\sum_{k=1}^{K-1}\{\gamma_K(d, k, Z) + \tau b_{Kj}(d, k, Z)\}\gamma_k(d', Z)]\right|_{\tau=0} \\
&= \sum_{k=1}^{K-1} b_{Kj}(d, k, Z)\gamma_k(d', Z).
\end{aligned}$$

Then we have

$$\hat{M}_{k\ell j} = \frac{1}{n - n_\ell}\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} \hat{\gamma}_{K\ell,\ell'}(d, k, Z_i) b_{1j}(d', Z_i), \ (k = 1, ..., K-1),$$

$$\hat{M}_{K\ell j} = \frac{1}{n - n_\ell}\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}}\sum_{k=1}^{K-1} b_{Kj}(d, k, Z_i)\hat{\gamma}_{k\ell,\ell'}(d', Z_i), \ (j = 1, ..., p).$$

We can then compute $\hat{\alpha}_{k\ell}(x)$ as in equation (5.2) and $\hat{\theta}$ for $Y_{ki} = 1(Q_i = k)$, $(k = 1, ..., K-1)$ and $Y_{Ki} = Y_i$ as in equation (5.1).

The orthogonal moment function corresponding to this estimator is

$$\psi(W, \gamma, \alpha, \theta) = \sum_{k=1}^{K-1} \gamma_K(d, k, Z)\gamma_k(d', Z) - \theta + \alpha_K(D, Q, Z)[Y - \gamma_K(D, Q, Z)]$$

$$+ \sum_{k=1}^{K-1} \alpha_k(D, Z)[1(Q = k) - \gamma_k(D, Z)], \;\; \gamma_K, \alpha_K \in \Gamma_K, \;\; \gamma_k, \alpha_k \in \Gamma_1, \;\; (k \leq K-1).$$

where Γ_K is the set of possible plims of $\hat{\gamma}_K$ and Γ_1 is the set of plims of $\hat{\gamma}_k$ for $k \leq K-1$. This moment function differs from the multiply robust moment function of Tchetgen Tchetgen and Shipster (2012) in imposing the constraint that each γ_k and α_k are contained in the set Γ_k of possible plim's of $\hat{\gamma}_k$. For example, when $\hat{\gamma}_K$ is a high dimensional regression estimator γ_K and α_K must be elements of the mean square span of $(X_1, X_2, ...)$ similarly to Section 2. It has the multiple robustness feature that for $\bar{\theta} = E[m(W, \bar{\gamma})]$ and any $\alpha = (\alpha_1, ..., \alpha_K) \in \Pi_{k=1}^K \Gamma_k$,

$$E[\psi(W, \bar{\gamma}, \alpha, \bar{\theta})] = 0,$$

shown in Chernozhukov et al. (2020) to be a general feature of orthogonal moment functions constructed from the influence function of $E[m(W, \gamma(F))]$. It also has other multiple robustness features. For α_{k0}, $(k = 1, ..., K)$ given in the proof of Corollary 9 in the Appendix, when $\alpha_{k0} \in \Gamma_1$, $(k \leq K-1)$ and $\alpha_{K0} \in \Gamma_K$,

$$E[\psi(W, \gamma_{10}, ..., \gamma_{K-1,0}, \gamma_K, \alpha_0, \theta_0)] = 0, \;\; E[\psi(W, \gamma_1, ..., \gamma_{K-1}, \gamma_{K0}, \alpha_0, \theta_0)] = 0,$$

for any $\gamma_K \in \Gamma_K$ and $\gamma_k \in \Gamma_1$, $(k \leq K-1)$.

We now return to the general learner $\hat{\theta}$ and give regularity conditions for asymptotic normality and consistent estimation of the asymptotic variance of $\hat{\theta}$. For $\tilde{\gamma} = (\tilde{\gamma}_1, ..., \tilde{\gamma}_K)' \in \Pi_{k=1}^K \Gamma_k$ and $\gamma_k \in \Gamma_k$ let

$$D_k(W, \gamma_k, \tilde{\gamma}) := \left. \frac{\partial m(W, \tilde{\gamma} + e_k \tau \gamma_k)}{\partial \tau} \right|_{\tau=0}$$

be the Gateaux derivative of $m(W, \gamma)$ with respect to γ_k when it exists. Comparing this definition with equation (5.2) we see that each $\hat{M}_{kj\ell}$ is an average of values of this Gateaux derivative. We impose the following condition on these derivatives.

Assumption 12: *There are $C, \varepsilon > 0$, $a_{kj}(w)$, and $A_k(w, \gamma)$ such that for all γ with $\|\gamma - \bar{\gamma}\| \leq \varepsilon$, $D_k(W, b_{kj}, \gamma)$ exists and for $k = 1, ..., K$*

$$D_k(W, b_{kj}, \gamma) = a_{kj}(W) A_k(W, \gamma), \;\; \max_{j \leq p} |E[a_{kj}(W)\{A_k(W, \gamma) - A_k(W, \bar{\gamma})\}]| \leq C \|\gamma - \bar{\gamma}\|,$$

$$\max_{j \leq p} |a_{kj}(W)| \leq C, \;\; E[A_k(W, \gamma)^2] \leq C.$$

This condition and the use of the cross-fit $\hat{\gamma}_{\ell,\ell'}$ in $\hat{M}_{k\ell}$ lead to a convergence rate for $\hat{M}_{k\ell}$. Let $M_{kj} = E[D_k(W, b_{kj}, \bar{\gamma})]$ and $M_k = (M_{k1}, ..., M_{kp})$, $(j = 1, ..., p; k = 1, ..., K)$.

Lemma 7: *If there is* $0 < d_\gamma < 1/2$ such that $\|\hat{\gamma}_{k\ell,\ell'} - \bar{\gamma}_{k\ell,\ell'}\| = O_p(n^{-d_\gamma})$, $(k = 1, ..., K; \ell, \ell' = 1, ...L)$, *and Assumption 12 is satisfied then*

$$\left\| \hat{M}_{k\ell} - M_k \right\|_\infty = O_p(n^{-d_\gamma}).$$

This result can be utilized to obtain mean square convergence rates for $\hat{\alpha}_k$ from Theorems 1 and 2. As for linear functionals the limit $\bar{\alpha}_k$ of the estimators $\hat{\alpha}_k$ are important for the properties of $\hat{\theta}$. Here the $\bar{\alpha}_k$ are associated with the Gateaux derivatives $D_k(W, \gamma_k, \bar{\gamma})$, $(k = 1, ..., K)$. The following condition specifies each $\bar{\alpha}_k$ and specifies the size of the remainder in a linearization using the Gateaux derivatives.

Assumption 13: *i) For* $(k = 1, ..., K)$ *there is* $\bar{\alpha}_k \in \Gamma_k$ *such that for all* $\gamma_k \in \Gamma_k$, $E[D_k(W, \gamma_k, \bar{\gamma})] = E[\bar{\alpha}_k(X_k)\gamma_k(X_k)]$; *ii)* $\bar{\alpha}_k(X_k)$ *and* $E[\{Y_k - \bar{\gamma}_k(X_k)\}^2 | X_k]$ *are bounded; iii) there are* ε, $C > 0$ *such that for all* $\gamma \in \Pi_{k=1}^K \Gamma_k$ *with* $\|\gamma - \bar{\gamma}\| < \varepsilon$,

$$|E[m(W, \gamma) - m(W, \bar{\gamma}) - \sum_{k=1}^{K} D_k(W, \gamma_k - \bar{\gamma}, \bar{\gamma})]| \leq C\|\gamma - \bar{\gamma}\|^2.$$

Here each $\bar{\alpha}_k$ is specified as the Riesz representer for the linear functional $E[D_k(W, \gamma_k, \bar{\gamma})]$ on $\gamma \in \Gamma_k$ as in Newey (1994, equation 4.4). Here the linearization $E[D_k(W, \gamma_k, \bar{\gamma})]$ has the role that was fulfilled by the linear functional $E[m(W, \gamma)]$ earlier. Indeed when $m(W, \gamma)$ is linear then $m(W, \gamma)$ will be its Gateaux derivative.

From Lemma 7 we see that the convergence rate for each $\hat{M}_{k\ell}$ is the convergence rate n^{-d_γ} of $\hat{\gamma}$ rather than $\sqrt{\ln(p)/n}$. As a result rate conditions for root-n consistency are different in the nonlinear $m(W, \gamma)$ case than in the linear one. The following condition imposes the rate conditions for a nonlinear functional.

Assumption 14: *There is* $1/4 < d_\gamma < 1/2$ *such that* $\|\hat{\gamma}_k - \bar{\gamma}_k\| = O_p(n^{-d_\gamma})$, $(k = 1, ..., K)$ *and for* $\bar{\alpha} = \bar{\alpha}_k$ *and* $b(x) = b_k(x_k)$, *either i) Assumptions 2 and 3 are satisfied and* $d_\gamma(1 + 4\xi)/(1 + 2\xi) > 1/2$ *or ii) Assumption 7 is satisfied and* $d_\gamma > 1/3$.

The requirement $d_\gamma > 1/4$ given here is familiar for estimators that depend nonlinearly on unknown functions, e.g. Newey (1994).. Condition i) allows d_γ to be any rate greater than $1/4$ if ξ is large enough. Condition ii), which drops the sparse eigenvalue assumption but requires absolute summability of the coefficients of each $\bar{\alpha}_k$, requires $d_\gamma > 1/3$.

The following gives the large sample inference results for $\hat{\theta}$ and $\hat{V}$. Define

$$\bar{\theta} = E[m(W, \bar{\gamma})],\ \psi(w) = m(w, \bar{\gamma}) - \bar{\theta} + \sum_{k=1}^{K} \bar{\alpha}_k(x_k)[y_k - \bar{\gamma}_k(x)],\ V = E[\psi(W)^2].$$

Here $\bar{\theta}$ will be the object estimated by $\hat{\theta}$ for $\bar{\gamma}$ =plim$(\hat{\gamma})$.

Theorem 8: *If for* $\Gamma = \Gamma_k$*,* $b(x) = b_k(x_k)$*,* $r = r_k$ *for* $(k = 1, ..., K)$ *and* $\varepsilon_n = n^{-d_\gamma}$ Assumptions 1, 4, 5, 10, and 12-14 are satisfied *then* $\sqrt{n}(\hat{\theta} - \bar{\theta}) \xrightarrow{d} N(0, V)$*. If in addition Assumption 7 is satisfied for* $\bar{\alpha} = \bar{\alpha}_k$ *and* $b = b_k$ *for each* $(k = 1, ..., K)$ *then* $\hat{V} \xrightarrow{p} V$*.*

Example 6: It is straightforward to specify regularity conditions for causal mediation that are sufficient for the conditions of Theorem 8 to hold.

Assumption 15: $\bar{\gamma}_k(X_k)$ *is bounded* $(k = 1, ..., K)$*, there is* $C > 0$ *such that* $\Pr(D = d, Q = q|Z) > C$ *for all* $d \in \{0, 1\}$*,* $q \in \{1, ..., K - 1\}$*, and* $E[\{Y - \bar{\gamma}_K(D, Q, Z)\}^2|D, Q, Z] \leq C$*.*

This condition is used to guarantee that $\bar{\alpha}_k(X_k)$ is bounded for each k. For brevity the form of $\bar{\alpha}_k(X_k)$ and $\psi(w)$ is given in the Appendix

Corollary 9: *If for* $\Gamma = \Gamma_k$*,* $b(x) = b_k(x_k)$*,* $r = r_k$ *for* $(k = 1, ..., K)$ *and* $\varepsilon_n = n^{-d_\gamma}$ *Assumptions 1, 4, 5, 14, and 15 are satisfied and there is* $C > 0$ *such that* $|\hat{\gamma}_k(x_k)| \leq C$ *for all* x_k *then* $\sqrt{n}(\hat{\theta} - \bar{\theta}) \xrightarrow{d} N(0, V)$*. If in addition Assumption 7 is satisfied for* $\bar{\alpha} = \bar{\alpha}_k$ *and* $b = b_k$ *for each* $(k = 1, ..., K)$ *then* $\hat{V} \xrightarrow{p} V$*.*

The conditions of this result are simple relative to the general regularity conditions in Assumptions 12 and 13. This simplicity is facilitated by $m(W, \gamma)$ being quadratic in γ. The condition that $|\hat{\gamma}_k(x_k)| \leq C$ is not strong for $k = 1, ..., K - 1$ because $Y_{ki} \in \{0, 1\}$. For $k = K$ this restriction could be imposed by truncating $\hat{\gamma}_k(x)$ for some C larger than a known bound on $\gamma_K(X_k)$ without affecting Assumption 14. In this way Corollary 9 provides a quite simple set of conditions for Auto-DML of causal mediation effects.

6 Regression Decomposition and the Average Treatment Effect on the Treated

In this Section we consider regression decompositions and the average treatment effect on the treated (ATET). We also give an application of estimating the ATET using Auto-DML.

Example 6: (Regression Decomposition and ATET): The effect of some dummy variable $D \in \{0, 1\}$ on an outcome variable Y is often of interest. Regression analysis can be used to decompose the unconditional effect into an effect conditional on covariates and an effect from a shift in the covariate distribution when D shifts. One way to carry out such a decomposition takes the form

$$E[Y|D=1]-E[Y|D=0]=\Delta_{response}+\Delta_{composition},$$
$$\Delta_{response}=E[Y|D=1]-\frac{E[D\gamma_0(0,Z)]}{\Pr(D=1)},\ \Delta_{composition}=\frac{E[D\gamma_0(0,Z)]}{\Pr(D=1)}-E[Y|D=0],$$

where $\gamma_0(D,Z)=E[Y|D,Z]$. We will focus here on the response effect

$$\theta_0=\Delta_{response}=\frac{E[D\gamma_0(1,Z)]-E[D\gamma_0(0,Z)]}{\Pr(D=1)}=\frac{E[D\{\gamma_0(1,Z)-\gamma_0(0,Z)\}]}{\Pr(D=1)}.$$

This θ_0 is the average effect of changing D on the outcome Y conditional on Z, averaged over the subpopulation with $D=1$. One could also consider a corresponding effect on the subpopulation with $D=0$. That could also be estimated using Auto-DML similarly to θ_0 but for brevity we omit this discussion.

This θ_0 is also the ATET when D is a treatment indicator and potential outcomes are mean independent of treatment conditional on covariates Z. Thus the estimator $\hat{\theta}$ and the asymptotic variance estimator $\hat{V}$ we give could be applied for inference for the ATET. We do so in the application given later in this Section.

The key regression functional of interest for θ_0 is

$$\begin{aligned}E[D\gamma_0(0,Z)]&=E[\pi_0(Z)\gamma_0(0,Z)]=E[\pi_0(Z)\frac{1-D}{1-\pi_0(Z)}\gamma_0(0,Z)] \qquad (6.1)\\&=E[\alpha_0(X)\gamma_0(X)],\ \alpha_0(X)=\frac{(1-D)\pi_0(Z)}{1-\pi_0(Z)}.\end{aligned}$$

Here $\alpha_0(X)$ is the Rr of a linear effect as in Section 2 with $m(w,\gamma)=d\gamma(0,z)$. The condition $E[\alpha_0(X)^2]<\infty$ for a finite semiparametric variance bound is $E[1/\{1-\pi_0(Z)\}]<\infty$.

The effect $\theta_0=\Delta_{response}=ATET$ is a special of the nonlinear effect in Section 5 where $\gamma=(\gamma_1,\gamma_2)$, $Y_1=Y$, $X_1=(D,Z)$, $Y_2=D$, $X_2=1$, and

$$m(w,\gamma)=\frac{y_2}{\gamma_2}[y_1-\gamma_1(0,z)].$$

The orthogonal moment function for this object is

$$\psi(w,\gamma,\gamma_2,\alpha,\theta)=\frac{1}{\gamma_2}\{d[y-\gamma(0,z)-\theta]-\alpha(x)[y-\gamma(x)]\},$$

where for notational convenience we let $y_1=y$, $y_2=d$, and $\gamma_1=\gamma$. Similarly to Section 2 this moment function is doubly robust in that

$$E[\psi(W,\bar{\gamma},\gamma_{20},\bar{\alpha},\theta_0)]=0$$

if either $\bar{\gamma}(X) = E[Y|X]$ or $\alpha_0(X) \in \Gamma$.

An Auto-DML is given by

$$\hat{\theta} = \frac{1}{n_D}\{\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{D_i[Y_i - \hat{\gamma}_\ell(0, Z_i)] - \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}\}, \tag{6.2}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}^2,\ \hat{\psi}_{i\ell} = (\frac{n}{n_D})\{D_i[Y_i - \hat{\gamma}_\ell(0, Z_i) - \hat{\theta}] - \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\},$$

where n_D is the number of treated observations and $\hat{\alpha}_\ell(x)$ is the Lasso learner of the Rr for $m(w, \gamma) = d\gamma(0, z)$. Similarly to the ATE in Example 3 we specify the dictionary to be $b(x) = [dq(z)', (1-d)q(z)']'$, where $q(z) = (z_1, ..., z_{p/2})'$ when $\hat{\gamma}_\ell$ is high dimensional and $q(z)$ is a vector of approximating functions when $\hat{\gamma}_\ell$ is nonparametric. Then $m(w, b_j) = d \cdot b_j(0, z) = d \cdot 1(j > p/2)q_{j-p/2}(z)$, so that

$$\hat{M}_\ell = \frac{1}{n - n_\ell}\sum_{i\notin I_\ell} m(W_i, b) = \begin{pmatrix} 0 \\ \bar{q}_\ell \end{pmatrix},\ \bar{q}_\ell = \frac{1}{n - n_\ell}\sum_{i\notin I_\ell} D_i q(Z_i).$$

Then by block diagonality of $\hat{G}_\ell$ and the first block of $\hat{M}_\ell$ being zero

$$\hat{\alpha}_\ell(x) = (1-d)q(z)'\hat{\rho}_{\ell 2},\ \hat{\rho}_{\ell 2} = \arg\min_{\rho}\{-2\bar{q}_\ell'\rho_2 + \rho_2'\hat{G}_2\rho_2 + 2r\left\|\rho_2\right\|_1\},$$

$$\hat{G}_{\ell 2} = \frac{1}{n - n_\ell}\sum_{i\notin I_\ell}(1 - D_i)q(Z_i)q(Z_i)'.$$

The first order conditions for the Lasso coefficients $\hat{\rho}_{\ell 2}$ are

$$\left|\frac{1}{n - n_\ell}\sum_{i\notin I_\ell} q_j(Z_i)[D_i - (1 - D_i)\hat{\omega}_{\ell i}]\right| \le r,\ \hat{\omega}_{\ell i} = q(Z_i)'\hat{\rho}_{\ell 2},\ (j = 1, ..., p/2). \tag{6.3}$$

The $\hat{\alpha}_\ell$ learner sets the “weights” $\hat{\omega}_{\ell i}$ to approximately “balance” the treated and untreated averages for each element of $q(z)$.

Corollary 10: *If i) there is $C > 0$ with $\pi_0(Z) < 1 - C$; ii) Assumptions 1, 5, and 11 are satisfied; iii)* $\max_{j\le p/1}|q_j(z)| \le C$; *and iv)* $Var(Y|X) \le C$ *and* $\bar{\alpha}(X)$ is bounded; *then for* $\bar{\theta} = E[D\{Y - \bar{\gamma}(0, Z)\}]/\Pr(D = 1)$ *and* $\psi(W) = \Pr(D = 1)^{-1}\{D[Y - \bar{\gamma}(0, Z) - \bar{\theta}] - \bar{\alpha}(X)[Y - \bar{\gamma}(X)]\}$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),\ \hat{V} \xrightarrow{p} V,\ V = E[\psi(W)^2].$$

As an empirical application, we use the Auto-DML of the ATT to estimate the effect of job training in the National Supported Work Demonstration (NSW), a job training program for disadvantaged workers that operated in the mid-1970s. We follow the empirical strategy of

LaLonde (1986) and Dehijia and Wahba (1999), who compare the difference-in-means estimator applied to an experimental data set with various econometric estimators applied to "quasi-experimental" data sets. The experimental data set consists of the treatment and control groups from a field experiment. A quasi-experimental data set consists of the treatment group from a field experiment and a comparison group from an unrelated national survey.

We use sample selection and variable construction as in Dehijia and Wahba (1999) and Farrell (2015). The outcome Y is earnings in 1978. The treatment D is an indicator of participation in job training. We consider three specifications of covariates Z. We impose common support of the propensity score for the treated and untreated groups based on covariates Z as in Farrell (2015). In estimation, we consider the fully-interacted dictionary $b(D, Z) = (1, D, Z, DZ)$ for all three specifications of Z.

The covariate specifications are as follows.

1. Demographics and earnings, with quadratic terms of continuous variables. In particular, the covariates are: age, education, black indicator, Hispanic indicator, married indicator, 1974 earnings, 1975 earnings, age squared, education squared, 1974 earnings squared, and 1975 earnings squared. This specification is moderately flexible. It is one that an analyst may reasonably implement without knowing the experimental benchmark ex ante. Here $dim(Z) = 11$ and $p = dim(b(D, Z)) = 24$.
2. Demographics and earnings, with quadratic terms of continuous variables and constructed indicators. In particular, the covariates are: those in specification 1; unemployed in 1974 indicator, unemployed in 1975 indicator, and no degree indicator. This specification includes some domain knowledge about which signals employers may respond to while making hiring decisions. Note that it does not include conveniently hand-crafted basis functions to get closer to the experimental benchmark. Here $dim(Z) = 14$ and $p = dim(b(D, Z)) = 30$.
3. A high dimensional specification where the covariates are: those in specification 2; all possible first order interactions, and all polynomials up to order five of the continuous variables (age, education, 1974 earnings, 1975 earnings). This specification was introduced by Farrell (2015). Here $dim(Z) = 171$ and $p = dim(b(D, X)) = 344$.

We estimate the Rr with Lasso minimum distance, and the regression with Lasso minimum distance, random forests (RF), or neural networks (NN). For Lasso minimum distance, we use the tuning procedure described in Appendix A. We use the same settings of random forest and neural networks as Chernozhukov et al. (2018). We use $L = 5$ folds in cross-fitting. Standard errors are calculated by delta method as described in Appendix **??**.

Tables 1, 2, and 3 summarize results for the NSW, PSID, and CPS data sets, respectively. For comparison, LaLonde (1986) reports 1794 (633) by difference-in-means applied to the NSW data, which is the experimental benchmark. Farrell (2015) reports 1737 (869) by group Lasso

spec.	treated	untreated	Lasso ATT	Lasso SE	RF ATT	RF SE	NN ATT	NN SE
1	185	254	1693.31	645.62	1673.83	661.21	1760.50	643.40
2	185	250	1670.65	646.30	1863.39	664.87	1716.11	648.39
3	185	179	1460.95	867.55	2407.61	780.79	1705.24	814.90

Table 1: ATT using NSW treatment and NSW control, by Auto-DML

spec.	treated	untreated	Lasso ATT	Lasso SE	RF ATT	RF SE	NN ATT	NN SE
1	185	844	877.38	831.85	1705.86	848.49	-1001.85	855.25
2	185	1620	2051.33	846.53	1178.24	856.15	1104.49	863.59
3	185	2336	1754.76	843.58	1573.81	836.15	2390.93	939.77

Table 2: ATT using NSW treatment and PSID comparison, by Auto-DML

applied to the PSID data using specification 3. Our results are broadly consistent. To emphasize the importance of debiasing, we report analogous tables using the biased, plug-in approach in Appendix C.

7 Panel Average Derivative and Demand Elasticities

In this Section, we apply Auto-DML to estimating demand elasticities while allowing for individual preferences that are correlated with prices and total expenditure. Specifically, we estimate own-price elasticity in a panel data model with correlated random slopes. We apply this approach to Nielsen scanner data.

A panel data model requires double indexing. Let Y_{it}, $(t = 1, ..., T_i, i = 1, ..., n)$, denote the share of total expenditure on some good for household i in time period t. Let X_{it} be a vector of log prices, log expenditure, and covariates. Let $\tilde{X}_i = (X'_{i1}, ..., X'_{i,T_i})'$ collect observations over all time periods for individual i into one vector. We allow for an unbalanced panel where different households may have different numbers of observations T_i as in Wooldridge (2019).

spec.	treated	untreated	Lasso ATT	Lasso SE	RF ATT	RF SE	NN ATT	NN SE
1	185	7687	419.57	576.08	1553.72	617.10	1386.00	587.70
2	185	13326	1050.58	568.55	1607.67	592.33	1664.22	583.42
3	185	13449	1335.90	602.38	1750.20	630.49	1676.90	681.86

Table 3: ATT using NSW treatment and CPS comparison, by Auto-DML

Consider the demand model of Chernozhukov, Hausman, and Newey (2019) given by

$$E[Y_{it}|\tilde{X}_i, B_{it}] = b_1(X_{it})' B_{it}. \tag{7.1}$$

The K-dimensional dictionary $b_1(X_{it})$ is a vector of functions of X_{it} that includes a constant and, for example, powers of log price and log expenditure. B_{it} represents household specific preferences that may vary over time and that may be correlated with regressors from each time period. We assume the conditional mean of B_{it} is time stationary with

$$E[B_{it}|\tilde{X}_i] = [I_K \otimes \tilde{H}_i]' \pi_0, \quad \pi_0 = (\pi_{10}', ..., \pi_{K,0}')', \tag{7.2}$$

where I_K is a K-dimensional identity matrix. $\tilde{H}_i$ is a vector of functions of $\tilde{X}_i$ with length that does not depend on T_i. This panel model is like that of Chamberlain (1982, 1992), Chernozhukov et al. (2013b), Graham and Powell (2012), and Wooldridge (2019), as further discussed in Chernozhukov, Hausman, and Newey (2019).

We will consider identifying and estimating transformations of $\beta_0 = E[B_{it}]$. β_0 is interpretable as the average marginal effect of changing $b_1(X_{it})$. The transformations we consider will be interpretable as average income, own-price, and cross-price elasticities. By law of iterated expectations, our model implies

$$\beta_0 = E[B_{it}] = [I_K \otimes E[\tilde{H}_i]]' \pi_0. \tag{7.3}$$

Combining (7.1), (7.2), and (7.3), we summarize the correlated random effects model as follows.

$$\begin{aligned} \gamma_0(\tilde{X}_i) &= E[Y_{it}|\tilde{X}_i] \\ &= b_1(X_{it})'\{\beta_0 + E[B_{it}|\tilde{X}_i] - \beta_0\} \\ &= b_1(X_{it})'\{\beta_0 + [I_K \otimes \tilde{H}_i]'\pi_0 - [I_K \otimes E[\tilde{H}_i]]'\pi_0\} \\ &= b_1(X_{it})'\beta_0 + [b_1(X_{it}) \otimes (\tilde{H}_i - E[\tilde{H}_i])]'\pi_0. \end{aligned} \tag{7.4}$$

In summary, the choice of K-dimensional dictionary $b_1(X_{it})$ in the demand model (7.1) induces a p-dimensional dictionary $b_{it} = b(\tilde{X}_i) = (b_1(X_{it})', [b_1(X_{it}) \otimes (\tilde{H}_i - E[\tilde{H}_i])]')'$ in the correlated random effects model (7.4). In practice, we replace $E[\tilde{H}_i]$ with $\frac{1}{n}\sum_{i=1}^n \tilde{H}_i$ and set $\tilde{H}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} b_1(X_{it})$.

Example 10: Demand elasticities. Denote $X_{it} = (D_{it}, Z_{it})$ where D_{it} is log own price. By the derivation in Chernozhukov et al. (2019) for budget share regressions, an average own-price elasticity is

$$\theta_0^* = \frac{\theta_0}{E[Y]} - 1, \quad \theta_0 = E\left[\frac{\partial \gamma_0(\tilde{X}_i)}{\partial d}\right].$$

Own-price elasticity θ_0^* is a smooth transformation of a linear effect θ_0, which in this case is average derivative. Auto-DML of own-price elasticity is then given by

$$\hat{\theta}^* = \frac{\hat{\theta}}{\frac{1}{n\sum_{i=1}^n T_i}\sum_{i=1}^n \sum_{t=1}^{T_i} Y_{it}} - 1$$

where $\hat{\theta}$ is the Auto-DML of average derivative from Example 4. Income elasticity and cross-price elasticity have a similar structure; see Appendix **??** for details.

For completeness, we present $\hat{M}_\ell$ for average derivative using the panel data dictionary b_{it}.

$$\hat{M}_\ell = \frac{1}{\sum_{i=1}^n T_i - \sum_{i\in I_\ell} T_i}\sum_{i\notin I_\ell}\sum_{t=1}^{T_i}\frac{\partial b_{it}}{\partial d} = \frac{1}{\sum_{i=1}^n T_i - \sum_{i\in I_\ell} T_i}\sum_{i\notin I_\ell}\sum_{t=1}^{T_i}\begin{pmatrix}\frac{\partial b_1(X_{it})}{\partial d} \\ 0\end{pmatrix}.$$

Recall Theorem 9 provides consistency and asymptotic normality guarantees for Auto-DML $\hat{\theta}$. A more sophisticated estimator $\hat{V}$ of the asymptotic variance of $\hat{\theta}$ is required that accounts for clustering of observations by household. See the Appendix **??** for details. Importantly, the cluster structure is also preserved in cross-fitting. Clustering methods for DML were previously used by Chiang et al. (2019) and Chernozhukov, Hausman, and Newey (2019). The consistency of own-price elasticity $\hat{\theta}^*$ follows from the continuous mapping theorem, and the asymptotic normality of $\hat{\theta}^*$ follows from delta method.

As an empirical application, we apply Auto-DML to estimate own-price elasticity of milk and soda with Nielsen scanner data. The empirical work here is the researchers' own analyses based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

The data we use are a subset of the Nielsen Homescan Panel as in Burda, Harding, and Hausman (2008, 2012). The data include 1483 households from the Houston-area zip codes for the years 2004-2006. The number of monthly observations for each household ranges from 12 to 36, with some households being added and taken away throughout the three years covered. 609 households are included the entire time. Expenditures include all purchases of the household in each month. The original data had time stamps for purchases. If a household purchased a good more than once in a month, the "monthly price" is the average price that the household paid (i.e. total amount spent on good/total quantity purchased). We include observations with zero expenditure share as justified in Chernozhukov, Hausman, and Newey (2019). For those observations, $Y_{it} = 0$ and own price is imputed in the ways described in Chernozhukov, Hausman, and Newey (2019).

We consider 15 groups of goods: bread, butter, cereal, chips, coffee, cookies, eggs, ice cream, milk, orange juice, salad, soda, soup, water, and yogurt. As in Burda, Harding, and Hausman

good	elasticity	SE
milk	-0.64	0.010
soda	-0.57	0.013

Table 4: Average own-price elasticity, by Auto-DML

(2008, 2012), we choose these groups because they make up a relatively large proportion of total food expenditure. We consider budget share regressions for two of these goods: milk and soda. Y_{it} is share of expenditure spent on milk (soda) by household i in month t. We take as $b_1(X_{it})$ the concatenation of the following variables: fourth order polynomial of log expenditure; fourth order polynomial of log price for milk (soda); up to fourth order interactions thereof; and log price of other goods. For $\tilde{H}_i$, we use the time averages of $b_1(X_{it})$. Note that $K = dim(b_1(X_{it})) = 42$ and $p = 1521$.

We estimate own-price elasticity according to the procedure outlined previously in this Section. We estimate both the Rr and the regression with Lasso minimum distance. For Lasso minimum distance, we use the tuning procedure described in Appendix A. We use $L = 5$ folds in cross-fitting. We calculate clustered standard errors by delta method, as described in Appendix **??**.

Table 4 summarizes results for the milk and soda own-price elasticities using Auto-DML. For comparison, the cross sectional estimates for milk and soda elasticities are -1.42 (0.020) and -0.86 (0.020), respectively (Table 1 of Chernozhukov, Hausman, and Newey 2019). Our results show that allowing for correlated random coefficients lowers these elasticity estimates by large magnitudes. These results confirm the finding in Table 5 of Chernozhukov, Hausman, and Newey (2019), that panel elasticity estimates allowing for correlation of preferences with prices and total expenditure are much smaller than cross-section estimates. Our own-price elasticity estimates are not as small as their slope fixed effect estimates, which for milk are between -0.65 (0.013) and -0.51 (0.052) and for soda are between -0.56 (0.017) and -0.36 (0.056) depending on choice of regularization parameter. We find that correlated random slope estimates of elasticities are close to, though not quite as small as, all fixed effects estimates and are much smaller than cross-section estimates.

For further comparison, we report results from the plug-in approach in Table 5. The plug-in elasticity estimates are much closer to the cross-section estimates than the Auto-DML estimates. The results of this table confirm the importance of debiasing in this application. In Appendix C, we report additional results from unregularized, high-dimensional linear regressions.

good	elasticity	SE
milk	-1.43	0.012
soda	-1.00	0.000

Table 5: Average own-price elasticity, by plug-in

8 Conclusions

In this paper we have given an automatic method of debiasing a machine learner of a parameter of interest that depends on a high dimensional and/or nonparametric regression. Using a Lasso or Dantzig minimum distance method we have orthogonal using only the form of the object of interest, without knowing the exact form of the bias correction. We have allowed the regression learner to be anything that converges in mean square at a fast enough rate. We have shown root-n consistency and asymptotic normality and given a consistent asymptotic variance estimator for a wide variety of causal and structural estimators, including GMM depending on first step regressions. We have applied these methods to estimate the average treatment effect on the treated in a job training experiment and have found similar results for Lasso, neural nets, and random forests regressions. We also have also estimated a correlated random slopes specification for consumer demand from scanner data and found estimates that are similar to fixed slope effect elasticities.

A Computing Auto-DML

A.1 Tuning

A.1.1 Procedure

The estimating equation (3.6) takes as given the value of regularization parameter r_L. For practical use, we provide an iterative tuning procedure to empirically determine r_L. Due to its iterative nature, the tuning procedure is most clearly stated as a replacement for equation (3.6).

Recall that the inputs to equation (3.6) are observations in I_ℓ^c, i.e. excluding fold ℓ. The analyst must also specify the p-dimensional dictionary b. For notational convenience, we assume b includes the intercept in its first component: $b_1(x) = 1$. In this tuning procedure, the analyst must further specify a low-dimensional sub-dictionary b^{low} of b. As in equation (3.6), the output of the tuning procedure is $\hat{\rho}_\ell$, an estimator of the Rr coefficient trained only on observations in I_ℓ^c.

The tuning procedure is as follows. For observations in I_ℓ^c

1. Initialize $\hat{\rho}_\ell$ using b^{low}

$$\begin{aligned} \hat{G}_\ell^{\text{low}} &= \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} b^{\text{low}}(X_i)b^{\text{low}}(X_i)' \\ \hat{M}_\ell^{\text{low}} &= \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} m(W_i, b^{\text{low}}) \\ \hat{\rho}_\ell &= \begin{bmatrix} \left(\hat{G}_\ell^{\text{low}}\right)^{-1}\hat{M}_\ell^{\text{low}} \\ 0 \end{bmatrix} \end{aligned}$$

2. Calculate moments

$$\begin{aligned} \hat{G}_\ell &= \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} b(X_i)b(X_i)' \\ \hat{M}_\ell &= \frac{1}{n-n_\ell}\sum_{i\notin I_\ell} m(W_i, b) \end{aligned}$$

3. While $\hat{\rho}_\ell$ has not converged
   (a) Update normalization

$$\hat{D}_\ell = \left[diag\left(\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}[b(X_i)b(X_i)'\hat{\rho}_\ell - m(W_i, b)]^2\right)\right]^{\frac{1}{2}}$$

(b) Update $(r_L, \hat{\rho}_\ell)$

$$r_L = \frac{c_1}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2p}\right)$$

$$\hat{\rho}_\ell = \arg\min_\rho \rho'\hat{G}_\ell\rho - 2\rho'\hat{M}_\ell + 2r_L c_3|\hat{D}_{\ell,1}\cdot\rho_1| + 2r_L\sum_{j=2}^{p}|\hat{D}_{\ell,j}\cdot\rho_j|$$

where ρ_j is the j-th coordinate of ρ and $\hat{D}_{\ell,j}$ is the j-th diagonal entry of $\hat{D}_\ell$.

In step 1, b^{low} is sufficiently low-dimensional that $\hat{G}_\ell^{\text{low}}$ is invertible. In practice, we take $dim(b^{low}) = dim(b)/40$.

In step 3, (c_1, c_2, c_3) are hyper-parameters taken as $(1, 0.1, 0.1)$ in practice. We implement the optimization via generalized coordinate descent with soft-thresholding. See Appendix A.2 for a detailed derivation of this soft-thresholding routine. We use the same techniques as Chernozhukov, Newey, and Singh (2018) to improve numerical stability in high dimensional settings. We use $\hat{D}_\ell + 0.2I$ instead of $\hat{D}_\ell$, and we cap the maximum number of iterations at 10. We also use warm start: in a given iteration, the optimization to determine $\hat{\rho}_\ell$ is initialized as the value of $\hat{\rho}_\ell$ in the previous iteration.

A.1.2 Justification

The iterative tuning procedure is analogous to Algorithm A.1 and therefore justified by an argument analogous to Theorem 1 of Belloni et al. (2012).

The analogy is as follows. The normalization $\hat{D}_\ell$ is the square root of the empirical second moment of the dictionary times the regression residual, just as $\hat{\Gamma}_\ell$ in Belloni et al. (2012). The formula for the regularization parameter is the same, after accounting for the fact that the objective in the present work uses r_L whereas the objective in eq. 2.4 of Belloni et al. (2012) uses $\frac{\lambda}{n}$.

A.2 Optimization

A.2.1 Procedure

The tuning procedure, an elaboration of estimating equation (3.6), involves the minimization of a generalized Lasso objective. We generalize the coordinate descent approach for Lasso (Fu 1998, Daubechies et al. 2004, Friedman et al. 2007, Friedman et al. 2010) to the minimum distance Lasso objective used in the present work. Specifically, we use the following coordinate-wise soft-thresholding update.

To lighten notation, we abstract from sample splitting, estimation of the moments and normalization, and special treatment of the intercept. We also scale the objective by 1/2:

$$\hat{\rho} = \arg\min_\rho \frac{1}{2}\rho'G\rho - \rho'M + r_L\|D\rho\|_1$$

We denote the j-th element of a generic vector V by V_j. We denote the (j,k)-entry of the matrix G by G_{jk}.

For $j = 1 : p$

1. Calculate loadings that do not depend on ρ_j

$$\begin{aligned} z_j &= G_{jj} \\ \pi_j &= M_j - \sum_{k \neq j} \rho_k G_{jk} \end{aligned}$$

2. Update coordinate ρ_j

$$\begin{aligned} \rho_j &= \frac{\pi_j + D_j r_L}{z_j}, \quad && \text{if } \pi_j < -D_j r_L \\ &= 0, && \text{if } \pi_j \in [-D_j r_L, D_j r_L] \\ &= \frac{\pi_j - D_j r_L}{z_j}, && \text{if } \pi_j > D_j r_L \end{aligned}$$

A.2.2 Justification

In this Section, we derive the coordinate-wise soft-thresholding update and argue that the procedure converges to the minimizer.

Observe that

$$\frac{\partial}{\partial \rho_j} \left[\frac{1}{2} \rho' G \rho - \rho' M \right] = -\pi_j + \rho_j z_j$$

and the loadings (z_j, π_j) do not depend on ρ_j.

The subgradient of the penalty term is

$$\begin{aligned} \frac{\partial}{\partial \rho_j} r_L \|D\rho\|_1 &= -D_j r_L && \text{if } \rho_j < 0 \\ &= [-D_j r_L, D_j r_L] && \text{if } \rho_j = 0 \\ &= D_j r_L && \text{if } \rho_j > 0 \end{aligned}$$

In summary, the subgradient of the objective is

$$\begin{aligned} \frac{\partial}{\partial \rho_j} \left[\frac{1}{2} \rho' G \rho - \rho' M + r_L \|D\rho\|_1 \right] &= -\pi_j + \rho_j z_j - D_j r_L && \text{if } \rho_j < 0 \\ &= [-\pi_j - D_j r_L, -\pi_j + D_j r_L] && \text{if } \rho_j = 0 \\ &= -\pi_j + \rho_j z_j + D_j r_L && \text{if } \rho_j > 0 \end{aligned}$$

Rearranging yields the component-wise update.

algorithm	MSE	R^2
Lasso	0.0060	0.17
generalized Lasso	0.0060	0.17
theoretical r_L	0.0014	0.48
normalization $\hat{D}$	0.0016	0.56
iteration: cold start	0.0014	0.50
iteration: warm start	0.0014	0.50
max iteration	0.0014	0.50
$\hat{D} + 0.2I$	0.0014	0.46

Table 6: 100 simulations

In our minimum distance Lasso procedure, the objective is of the form of eq. 21 of Friedman et al. (2007).

$$Q(\theta) = g(\theta) + \sum_k h^k(\theta^k)$$
$$g(\theta) = \frac{1}{2}\theta' G\theta - M'\theta$$
$$h^k(\theta^k) = |\theta^k|$$

where g is differentiable and convex and $\{h^k\}$ are convex. Therefore coordinate descent converges to the minimizer of the objective (Tseng, 2001).

A.3 Minimum Distance Lasso Using Simulated Data

We first validate the minimum distance Lasso estimator for $\hat{\rho}$ on a design in which the truth is known. We compare our implementation to the Lasso implementation `LassoShooting.fit` in the `hdm` package at each point of departure: minimum distance Lasso formulation, theoretical r_L, normalization $\hat{D}$, iteration, and stabilization. Altogether, this exercise confirms the validity of each technique introduced in the tuning procedure.

In this design, the ground truth is $\rho_0 = (1, 1, 1, 0, 0, ...)$ where $dim(\rho_0) = 101$. The data generating process is

$$Y = X'\rho_0 + \epsilon$$

where $X = (1, X_1, ..., X_{100})'$, $X_j \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\epsilon \sim \mathcal{N}(0, 1)$. Recall that the regression coefficient ρ_0 can be recovered by using the functional $m(w, \gamma) = y\gamma(x)$ in the minimum distance Lasso formulation.

In Table 6, we report MSE defined as $|\hat{\rho} - \rho_0|_2^2$ of various implementations. Table 6 is cumulative in the sense that each row implements one additional technique relative to the preceding row. Before using theoretical r_L, we use $r_L = 0.5$. We use the estimator reported in the final row in the empirical examples of Sections 6 and 7; it is precisely the estimator defined in the tuning procedure.

B 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. We first give a series of Lemmas like those in Bradic et al. (2021) except that ε_n is allowed to be larger than $\sqrt{\ln(p)/n}$ in order to allow $m(w,\gamma)$ being nonlinear in γ. These Lemmas are used to prove Theorem 1. Let ε_n be as given in Assumptions 2 and 6 and $s_0 \geq C\varepsilon_n^{-2/(2\xi+1)}$. By Assumption 2 we can define J_0 as indices of a sparse approximation with $|J_0| = s_0$ and coefficients $\tilde{\rho}_j$ for $j \in J_0$ such that for $\tilde{\alpha}(x) = \sum_{j\in J_0} \tilde{\rho}_j b_j(X)$,

$$E[\{\alpha_0(X) - \tilde{\alpha}(X)\}^2] \leq Cs_0^{-2\xi}.$$

Define ρ to be the coefficients of a linear projection of $\alpha_0(X)$ on $b(X)$ so that $\breve{\alpha}(X) = b(X)'\rho$ satisfies

$$E[b(X)\{\alpha_0(X) - \breve{\alpha}(X)\}] = 0.$$

Also define ρ_* as

$$\rho_* \in \arg\min_{v} \, (\rho - v)'G(\rho - v) + \varepsilon_n \sum_{j\in J_0^c} |v_j|. \tag{B.1}$$

Lemma A1: $\|G(\rho_* - \rho)\|_\infty \leq \varepsilon_n$.

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

Lemma A2: $(\rho - \rho_*)'G(\rho - \rho_*) \leq C\varepsilon_n^{4\xi/(2\xi+1)}$ *and* $\|\bar{\alpha} - b'\rho^*\| = O(\varepsilon_n^{2\xi/(2\xi+1)})$.

Proof: Define $\tilde{\rho} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_p)'$ as

$$\tilde{\rho}_j = \begin{cases} \tilde{\rho}_j & \text{if } j \in J_0 \\ 0 & \text{otherwise.} \end{cases}$$

By the definition of ρ_*, we have that

$$(\rho - \rho_*)'G(\rho - \rho_*) + \varepsilon_n \sum_{j\in J_0^c} |\rho_{*,j}| \leq (\rho - \tilde{\rho})'G(\rho - \tilde{\rho}) + \varepsilon_n \sum_{j\in J_0^c} |\tilde{\rho}_j| = (\rho - \tilde{\rho})'G(\rho - \tilde{\rho}). \tag{B.2}$$

Note that $\tilde{\alpha}(x) = b(x)'\tilde{\rho}$, so by the defintion of $\breve{\alpha}(x) = b(x)'\rho$ we have

$$\begin{aligned}(\rho - \tilde{\rho})'G(\rho - \tilde{\rho}) &= E[\{b(X)'(\rho - \tilde{\rho})\}^2] = \|\breve{\alpha} - \tilde{\alpha}\|_2^2 = \|\alpha_0 - \tilde{\alpha} - (\alpha_0 - \breve{\alpha})\|_2^2 \leq 2(\|\alpha_0 - \tilde{\alpha}\|_2^2 + \|\alpha_0 - \breve{\alpha}\|_2^2) \\ &\leq 4\,\|\alpha_0 - \breve{\alpha}\|^2 \leq 4Cs_0^{-2\xi} \leq C\varepsilon_n^{4\xi/(2\xi+1)}.\end{aligned}$$

where the last inequality follows by by $s_0 \geq C\varepsilon_n^{-2/(2\xi+1)}$. The result then follows eq. (B.2) by and $\varepsilon_n \sum_{j\in J_0^c} |\rho_{*,j}| \geq 0$. *Q.E.D.*

Define J to be the vector of indices of nonzero elements of ρ_* and $|A|$ be be the number non zero elements of any finite set A.

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

Proof: For all $j \in J\backslash J_0$ the first order conditions to equation (B.1) imply $|e_j'G(\rho_* - \rho)| = \varepsilon_n/2$. Therefore, It follows that

$$\sum_{j\in J\backslash J_0} (e_j'G(\rho_* - \rho))^2 = \frac{1}{4}\varepsilon_n^2 |J\backslash J_0|.$$

In addition,

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

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

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

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

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

Lemma A4: $\|\hat{G}\rho_* - G\rho_*\|_\infty = O_p(\sqrt{\ln(p)/n})$.

Proof: By $(\rho - \rho_*)'G(\rho - \rho_*) \longrightarrow 0$ and $\rho' G\rho \leq E[\alpha_0(X)^2]$ it follows that $E[\{b(X)'\rho_*\}^2] = \rho_*'G\rho_* \leq C$. The conclusion then follows by Assumption 4 and Lemma B2 of Bradic et al. (2021). *Q.E.D.*

Lemma A5: *For* $\Delta = \hat{\rho} - \rho^*$ *and any* $\hat{J}$ *such that* $(\rho^*)_{\hat{J}^c} = 0$, *with probability one then with probability approaching one,*

$$\Delta'\hat{G}\Delta \leq 3r\|\Delta\|_1,\ \|\Delta_{\hat{J}_2}\|_1 \leq 3\|\Delta_{\hat{J}}\|_1.$$

Proof: By the definition of the estimator $\hat{\rho}$, we have

$$\hat{\rho}'\hat{G}\hat{\rho} - 2\hat{M}'\hat{\rho} + 2r\|\hat{\rho}\|_1 \le \rho_*'\hat{G}\rho_* - 2\hat{M}'\rho_* + 2r\|\rho_*\|_1.$$

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

$$\Delta'\hat{G}\Delta + 2r\|\rho_* + \Delta\|_1 \le 2r\|\rho_*\|_1 + 2(\hat{M} - \hat{G}\rho_*)'\Delta. \tag{B.3}$$

By the definition of ρ and $M = E[b(X)\bar{\alpha}(X)]$ we have $G\rho - M = 0$. Then by Assumption 6, Lemma A1, Lemma A4, and the triangle inequality

$$\begin{aligned}\|\hat{G}\rho_* - \hat{M}\|_\infty &\le \|\hat{G}\rho_* - G\rho_*\|_\infty + \|M - \hat{M}\|_\infty + \|G\rho_* - M\|_\infty \\ &\le O_p(\varepsilon_n) + \|G\rho - M\|_\infty + \|G(\rho_* - \rho)\|_\infty = O_p(\varepsilon_n).\end{aligned}$$

Therefore, by the Holder inequality we have $\left|(\tilde{\mu} - \hat{G}\rho_*)'\Delta\right| \le \|\tilde{\mu} - \hat{G}\rho_*\|_\infty\|\Delta\|_1 = O_p(\varepsilon_n)\|\Delta\|_1$, so that by $\varepsilon_n = o(r)$,

$$\Delta'\hat{G}\Delta + 2r\|\rho_* + \Delta\|_1 \le 2r\|\rho_*\|_1 + O_p(\varepsilon_n)\|\Delta\|_1 \le 2r\|\rho_*\|_1 + r\|\Delta\|_1,$$

with probability appraoching one. Then the triangle inequality $\|\rho_*\|_1 = \|\rho_* + \Delta - \Delta\|_1 \le \|\rho_* + \Delta\|_1 + \|\Delta\|_1$ and subtracting $2r\|\rho_* + \Delta\|_1$ from both sides gives the first conclusion.

Next, since $\Delta'\hat{G}\Delta \ge 0$ it also follows from equation (B.3) that $2r\|\rho_* + \Delta\|_1 \le 2r\|\rho_*\|_1 + r\|\Delta\|_1$, so dividing through by r gives

$$2\|\rho_* + \Delta\|_1 \le 2\|\rho_*\|_1 + \|\Delta\|_1.$$

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

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

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

Lemma A6: $\|\Delta\|_2 = O_p((r/\varepsilon_n)\varepsilon_n^{2\xi/(2\xi+1)})$.

Proof: For $\hat{J} = J$ it follows from Assumption 5 and Lemma A5 that with probability approaching one,

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

Dividing through by $\|\Delta_J\|_2$ then gives with probability approaching one,

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

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 A5 for $\hat{J} = J \cup N$ it follows exactly as in second previous display that

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

By Lemma 6.9 of van de Geer and Buhlmann (2011) and Lemma A5,

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

Therefore, by the triangle inequality with probability approaching one,

$$\|\Delta\|_2 \leq \|\Delta_{\hat{J}}\|_2 + \|\Delta_{\hat{J}^c}\|_2 \leq C(r/\varepsilon_n)\varepsilon_n^{2\xi/(2\xi+1)}. \; Q.E.D.$$

Proof of Theorem 1: By Lemma A6,

$$\Delta' G \Delta \leq \lambda_{\max}(G)\,\|\Delta\|_2^2 = O_p((r/\varepsilon_n)^2\varepsilon_n^{4\xi/(2\xi+1)}).$$

Then by Lemma A2, the triangle inequality, and Assumption 5, for any $c > 0$,

$$\begin{aligned}\|\bar{\alpha} - \hat{\alpha}\|^2 &\leq 2\,\|\bar{\alpha} - b'\rho^*\|^2 + 2\,\|b'(\rho^* - \hat{\rho})\|^2 = O(\varepsilon_n^{4\xi/(2\xi+1)}) + \Delta' G \Delta \\ &= O_p((r/\varepsilon_n)^2\varepsilon_n^{4\xi/(2\xi+1)}) = o_p(n^{2c}\varepsilon_n^{4\xi/(2\xi+1)}).\end{aligned}$$

Taking square roots of both sides gives the conclusion. *Q.E.D.*

Next we give a series of Lemmas that are used to prove Theorem 2.

Lemma A7: *If Assumption 7 is satisfied then Assumption 2 is satisfied with* $\xi = 1/2$.

Proof: Let J_s denote the indices of the s largest coefficients in absolute value and $j_s \in J_s$ be such that $|\rho_{0j_s}| \leq |\rho_{0j}|$ for all $j \in J_s$. Then

$$s\,|\rho_{0j_s}| \leq \sum_{j \in J_s} |\rho_{j0}| \leq \sum_{j=1}^{\infty} |\rho_{j0}| = C. \tag{B.4}$$

By Assumption 7 $J_s \subset \{1, ..., p\}$. Define

$$\alpha_p(X) := \sum_{j=1}^{p} \rho_{0j} b_j(X), \;\; \alpha_s(X) := \sum_{j \in J_s} \rho_{0j} b_j(X).$$

Let $\rho^p = (\rho_{01}, ..., \rho_{0p})$ and ρ^s be the vector with $\rho_j^s = \rho_{0j}$ if $j \in J_s$ and $\rho_j^s = 0$ otherwise. Then by $|\rho_{0j}| \leq |\rho_{0j_s}|$ for all $j \notin J_s$,

$$\begin{aligned}\|\alpha_p - \alpha_s\|^2 &= (\rho^p - \rho^s)'G(\rho^p - \rho^s) \leq C\,\|\rho^p - \rho^s\|^2 = C\sum_{j\notin J_s}\rho_{0j}^2 \leq C\,|\rho_{0j_s}|\sum_{j\notin J_s}|\rho_{0j}| \\ &\leq C\,|\rho_{0j_s}|\sum_{j=1}^{\infty}|\rho_{0j}| \leq C\,|\rho_{0j_s}| \leq C/s.\end{aligned}$$

It then follows by Assumption 7 and the triangle and Cauch-Scwartz inequalities that

$$\|\bar{\alpha} - \alpha_s\|^2 \leq 2\,\|\bar{\alpha} - \alpha_p\|^2 + 2\,\|\alpha_p - \alpha_s\|^2 \leq C/s. \ Q.E.D.$$

Define $\rho_* \in \arg\min_\rho \left\{\|\bar{\alpha} - b'\rho\|^2 + \varepsilon_n\,|\rho|_1\right\}$.

Lemma A8: *If Assumption 7 is satisfied then*

$$\|\bar{\alpha} - b'\rho_*\|^2 \leq C\varepsilon_n, \ \ |\rho_*|_1 \leq C.$$

Proof: Note that by $\xi = 1/\ \not{2}$ as in Lemma A7 we have $s = \varepsilon_n^{-2/(2\xi-1)} = \varepsilon_n^{-1}$. By Lemma A7 and the definition of ρ_*,

$$\|\bar{\alpha} - b'\rho_*\|^2 + \varepsilon_n\,|\rho_*|_1 \leq \|\bar{\alpha} - b'\rho_s\|^2 + \varepsilon_n\,|\rho_s|_1 \leq C\varepsilon_n.$$

The conclusion follows from the terms on the left-hand side both being positive. *Q.E.D.*

Lemma A9: *If* $\varepsilon_n = o(r)$ *then* $\|\hat{\rho}\|_1 = O_p(1)$.

Proof: For $\Delta = \hat{\rho} - \rho_*$ equation (B.3) can be written as

$$\Delta\prime\hat{G}\Delta + 2r\|\hat{\rho}\|_1 \leq 2r\|\rho_*\|_1 + 2(\hat{M} - \hat{G}\rho_*)'\Delta. \tag{B.5}$$

By Lemma A8 $\|\bar{\alpha} - b'\rho_*\|^2 \longrightarrow 0$ so that $E[(b(X)'\rho^*)^2] \leq C$. Then by Assumption 7, Lemma A8, and the Holder inequality it follows that

$$\left\|(\hat{G} - G)\rho_*\right\|_\infty \leq \left\|\hat{G} - G\right\|_\infty |\rho_*|_1 = O_p(\varepsilon_n)O_p(1) = O_p(\varepsilon_n).$$

Note that the first order conditions for the minimization of

$$\begin{aligned}\|\bar{\alpha} - b'\rho\|^2 + \varepsilon_n\,|\rho|_1 &= C + \rho'G\rho - 2E[\bar{\alpha}(X)b(X)]'\rho + \varepsilon_n\,|\rho|_1 \\ &= C + \rho'G\rho - 2M'\rho + \varepsilon_n\,|\rho|_1\end{aligned}$$

imply that $\|G\rho_* - M\|_\infty = O(\varepsilon_n)$, similarly to Lemma A2. Then by the triangle inequality,

$$\left\|\hat{M} - \hat{G}\rho_*\right\|_\infty \leq \left\|\hat{M} - M\right\|_\infty + \left\|(\hat{G} - G)\rho_*\right\|_\infty + \|M - G\rho_*\|_\infty = O_p(\varepsilon_n).$$

Then by the $\Delta\prime\hat{G}\Delta \geq 0$, the Holder and triangle inequalities, and dividing equation (B.5) by $2r$ we have

$$\|\hat{\rho}\|_1 \leq \|\rho_*\|_1 + \left\|\hat{M} - \hat{G}\rho_*\right\|_\infty \|\Delta\|_1 / r \leq C + O_p(\varepsilon_n/r)(\|\hat{\rho}\|_1 + \|\rho_*\|_1) = C + o_p(1)\|\hat{\rho}\|_1.$$

Then noting that $o_p(1)\|\hat{\rho}\|_1 \leq (1/2)\|\hat{\rho}\|_1$ with probability approaching one we have

$$\|\hat{\rho}\|_1 \leq C. \ Q.E.D.$$

Proof of Theorem 2: It follows by Lemma A9 that $\left\|(G - \hat{G})\hat{\rho}\right\|_\infty \leq \left\|G - \hat{G}\right\|_\infty \|\hat{\rho}\|_1 = O_p(\varepsilon_n)O_p(1) = O_p(\varepsilon_n)$. Also, the first order conditions for Lasso imply $\left\|-\hat{G}\hat{\rho} + \hat{M}\right\|_\infty \leq r$. Also $\left\|\hat{M} - M\right\|_\infty = O_p(\varepsilon_n)$ and $\|-G\rho_* + M\|_\infty \leq \varepsilon_n$ by the first order conditions for ρ_*. Then by the triangle inequality

$$\|G(\hat{\rho} - \rho_*)\|_\infty \leq \left\|(G - \hat{G})\hat{\rho}\right\|_\infty + \left\|-\hat{G}\hat{\rho} + \hat{M}\right\|_\infty + \left\|\hat{M} - M\right\|_\infty + \|-G\rho^* + M\|_\infty = O_p(r).$$

Then by Lemma A8

$$(\hat{\rho} - \rho_*)'G(\hat{\rho} - \rho_*) \leq \|\hat{\rho} - \rho_*\|_1 \|G(\hat{\rho} - \rho_*)\|_\infty \leq (\|\hat{\rho}\|_1 + \|\rho_*\|_1)O_p(r) = O_p(r).$$

Then we have

$$\|\bar{\alpha} - \hat{\alpha}\|^2 \leq 2\|\bar{\alpha} - b'\rho_*\|^2 + 2\|b'(\hat{\rho} - \rho_*)\|^2 = O_p(\varepsilon_n) + 2(\hat{\rho} - \rho_*)'G(\hat{\rho} - \rho_*) = O_p(r) = o_p(n^{2c}\varepsilon_n),$$

for any $c > 0$. Taking square roots of both sides of the inequality gives the conclusion. *Q.E.D.*

Lemma A10: *If Assumption 4 is satisfied then* $\left\|\hat{G} - G\right\|_\infty = O_p(\sqrt{\ln(p)/n})$.

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,

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

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

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

Define $K = 2C_b^2/\sqrt{\ln 2} \geq \|T_{ijk}\|_{\Psi_2}$. By Hoeffding's inequality (Vershynin, 2018) there is a constant c such that

$$\begin{aligned} p^2 \max_{j,k} \Pr(|U_{jk}| > C\varepsilon_n^*) &\leq 2p^2 \exp\left(-\frac{c(nC\varepsilon_n^*)^2}{nK^2}\right) \\ &= 2p^2 \exp\left(-\frac{cC^2 \ln(p)}{K^2}\right) \\ &\leq 2 \exp\left(\ln(p)[2 - \frac{cC^2}{K^2}]\right) \longrightarrow 0 \end{aligned}$$

for any $C > K\sqrt{2/c}$. Thus for large enough C, $\Pr(|\hat{G} - G|_\infty \geq C\sqrt{\ln(p)/n}) \longrightarrow 0$, implying the conclusion. *Q.E.D.*

Proof of Theorem 3: The proof proceeds verifying Assumptions 1-3 of Chernozhukov et al. (2020, LR). Assumption 1 i) of LR is implied by Assumption 10. Let $\phi(w, \gamma, \alpha) = \alpha(x)[y - \gamma(x)]$. Note that by Assumption 9,

$$\begin{aligned} \int \{\phi(w, \hat{\alpha}_\ell, \bar{\gamma}) - \phi(w, \bar{\alpha}, \bar{\gamma})\}^2 F(dw) &= \int \{\hat{\alpha}_\ell(x) - \bar{\alpha}(x)\}^2 [y - \bar{\gamma}(x)]^2 F(dw) \\ &\leq C \left\|\hat{\alpha}_\ell - \bar{\alpha}\right\|^2 \xrightarrow{p} 0, \\ \int \{\phi(w, \bar{\alpha}, \hat{\gamma}_\ell) - \phi(w, \bar{\alpha}, \bar{\gamma})\}^2 F(dw) &= \int \bar{\alpha}(x)^2 [\hat{\gamma}_\ell(x) - \bar{\gamma}(x)]^2 F(dx) \\ &\leq C \left\|\hat{\gamma}_\ell - \bar{\gamma}\right\|^2 \xrightarrow{p} 0, \end{aligned}$$

giving Assumptions 1 ii) and 1 iii) of LR.

To verify Assumption 2 of LR, note that by Assumption 8 it follows similarly to Lemma A10 that Assumption 6 is satisfied for

$$\varepsilon_n = \sqrt{\ln(p)/n}.$$

Consider first the first case of Assumption 11 where Assumptions 2 and 3 are satisfied. By Theorem 1, for any $c > 0$ we have

$$\|\hat{\alpha}_\ell - \bar{\alpha}\| = o_p(n^c [\ln(n)/n]^{\xi/(2\xi+1)}).$$

Choose $c = [d_\gamma + \xi/(2\xi + 1) - 1/2]/2 > 0$. Then by Assumption 11,

$$\sqrt{n} \|\hat{\alpha}_\ell - \bar{\alpha}\| \|\hat{\gamma}_\ell - \bar{\gamma}\| = o_p(n^c [\ln(n)]^{\xi/(2\xi+1)} n^{1/2 - \xi/(2\xi+1) - d_\gamma}) = o_p(n^{-c} [\ln(n)]^{\xi/(2\xi+1)}) = o_p(1).$$

Consider now the second case of Assumption 11 where Assumption 7 is satisfied Then for $c = (1/4 + d_\gamma - 1/2)/2$, the conclusion of Theorem 2 gives

$$\sqrt{n} \|\hat{\alpha}_\ell - \bar{\alpha}\| \|\hat{\gamma}_\ell - \bar{\gamma}\| = o_p(n^c [\ln(n)]^{1/4} n^{-(1/4) - d_\gamma + 1/2}) = o_p(n^{-c} [\ln(n)]^{1/4}) = o_p(1).$$

Then by the Cauchy-Schwartz and conditional Markov inequalities we have

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}\{\hat{\alpha}_\ell(X_i)-\bar{\alpha}(X_i)\}\{\hat{\gamma}_\ell(X_i)-\bar{\gamma}(X_i)\}\right| \\
&\leq \sqrt{n}\sqrt{\frac{1}{n}\sum_{i\in I_\ell}\frac{\{\hat{\alpha}_\ell(X_i)-\bar{\alpha}(X_i)\}^2}{n}}\sqrt{\frac{1}{n}\sum_{i\in I_\ell}\frac{\{\hat{\gamma}_\ell(X_i)-\bar{\gamma}(X_i)\}^2}{n}} \\
&\leq \sqrt{n}\left\|\hat{\alpha}_\ell-\bar{\alpha}\right\|\left\|\hat{\gamma}_\ell-\bar{\gamma}\right\| = o_p(1),
\end{aligned}$$

so that Assumption 2 of LR is satisfied.

To verify Assumption 3 of LR, note that by Assumption 1 $\hat{\alpha}_\ell(x)=b(x)'\hat{\rho}_\ell\in\Gamma$, so that

$$\int\phi(w,\bar{\gamma},\hat{\alpha}_\ell)F_0(dw)=\int\hat{\alpha}_\ell(x)[y-\bar{\gamma}(x)]F(dw)=0$$

and $E[m(W,\gamma)-\bar{\theta}+\bar{\alpha}(X)\{Y-\gamma(X)\}]$ is affine in γ, giving Assumption 3 of LR. It then follow by Lemma 15 of LR that

$$\sqrt{n}(\hat{\theta}-\bar{\theta})=\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\psi(W_i,\hat{\gamma}_\ell,\hat{\alpha}_\ell,\bar{\theta})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i,\bar{\gamma},\bar{\alpha},\bar{\theta})+o_p(1).$$

The first conclusion then follows by the central limit theorem.

The second conclusion To show the second conclusion, let $\psi_i=\psi_0(W_i)$. Then for $i\in I_\ell$,

$$(\hat{\psi}_{i\ell}-\psi_i)^2\leq C\left(\sum_{j=1}^{3}R_{ij}+R\right),\ R_{i1}=[m(W_i,\hat{\gamma}_\ell)-m(W_i,\gamma_0)]^2,\ R_{i2}=\hat{\alpha}_\ell(X_i)^2\{\hat{\gamma}_\ell(X_i)-\bar{\gamma}(X_i)\}^2,$$
$$R_{i3}=\{\hat{\alpha}_\ell(X_i)-\bar{\alpha}(X_i)\}^2\{Y_i-\bar{\gamma}(X_i)\}^2,\ R=(\hat{\theta}-\bar{\theta})^2.$$

The first conclusion implies $R\xrightarrow{p}0$. Let $\mathcal{W}_{-\ell}$ denote the observations not in I_ℓ. By Assumption 10,

$$E[R_{i1}|\mathcal{W}_{-\ell}]=\int[m(w,\hat{\gamma}_\ell)-m(w,\bar{\gamma})]^2F_W(dw)=o_p(1).$$

By Assumption 4 and Lemma A9, uniformly in x

$$|\hat{\alpha}_\ell(x)|\leq\sum_{j=1}^{p}|b_j(x)|\,|\hat{\rho}_{\ell j}|\leq C\left\|\hat{\rho}_\ell\right\|_1=O_p(1).$$

Then by Assumption 11,

$$\begin{aligned}
E[R_{i2}|\mathcal{W}_{-\ell}]&\leq C\left\|\hat{\rho}_\ell\right\|_1\int\{\hat{\gamma}_\ell(x)-\bar{\gamma}(x)\}^2F_W(dw)=C\left\|\hat{\rho}_\ell\right\|_1\left\|\hat{\gamma}_\ell-\bar{\gamma}\right\|^2 \\
&\leq O_p(1)o_p(1)=o_p(1).
\end{aligned}$$

Also by Assumption 9 and iterated expectations

$$E[R_{i3}|\mathcal{W}_{-\ell}] \leq \int \{\hat{\alpha}_\ell(x) - \bar{\alpha}(x)\}^2 E[(Y - \bar{\gamma}(x))^2 | X = x] F_X(dx)$$
$$\leq C \int \{\hat{\alpha}_\ell(x) - \bar{\alpha}(x)\}^2 F_X(dx) = C \left\|\hat{\alpha}_\ell - \bar{\alpha}\right\|^2 = o_p(1).$$

Then by the triangle inequality,

$$E[\frac{1}{n}\sum_{i\in I_\ell}\sum_{j=1}^{3} R_{ij}|\mathcal{W}_{-\ell}] \leq E[R_{i1}|\mathcal{W}_{-\ell}] + E[R_{i3}|\mathcal{W}_{-\ell}] + E[R_{i3}|\mathcal{W}_{-\ell}] = o_p(1).$$

It then follows by the conditional Markov inequality that $\sum_{i\in I_\ell}\sum_{j=1}^{3} R_{ij}/n = o_p(1)$. The triangle inequality and adding up over ℓ then gives $(\hat{\psi}_{i\ell} - \psi_i)^2$

$$\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}(\hat{\psi}_{i\ell} - \psi_i)^2 = o_p(1).$$

Note also that by Assumptions 9 and 10,

$$E[\psi_i^2] \leq C(1 + E[m(W,\bar{\gamma})^2] + E[\bar{\alpha}(X)^2\{Y - \bar{\gamma}(X)\}^2]) < \infty.$$

Then

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}^2 = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}(\hat{\psi}_{i\ell} - \psi_i + \psi_i)^2 = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}(\hat{\psi}_{i\ell} - \psi_i)^2 + 2\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}(\hat{\psi}_{i\ell} - \psi_i)\psi_i + \frac{1}{n}\sum_{i=1}^{n}\psi_i^2;$$

Furthermore by the Cauchy-Schwartz and Markov inequalities we have

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

Then $\hat{V} \xrightarrow{p} V$ follows by the triangle inequality and the law of large numbers. *Q.E.D.*

Proof of Corollary 4: Note that by Assumption 4,

$$|m(W,b_j)| \leq \int |b_j(x)| \left[f_1(x) + f_2(x)\right]dx \leq C$$

so that Assumption 8 is satisfied. Also, $|\alpha_0(X)| = |[f_1(x) - f_0(x)]/f(x)| \leq C$ by hypothesis, so by the Cauchy-Schwartz inequality,

$$E[m(W,\gamma)^2] = |E[\gamma(X)\alpha_0(X)]|^2 \leq C\left|E[|\gamma(X)|]\right|^2 \leq CE[\gamma(X)^2],$$

implying Assumption 10. The conclusion then follows by Theorem 3. $Q, E.D.$

Proof of Corollary 5: Integration by parts and Assumption 4 give

$$|m(W,b_j)| = \left| -\int \frac{\partial w(D)}{\partial d} b(D,Z)dD \right| \leq C \int \left| \frac{\partial w(D)}{\partial d} \right| |b(D,Z)| \leq C \int_{\mathcal{D}} dD \leq C,$$

so Assumption 8 is satisfied. Also by hypothesis $|\alpha_0(X)| = |\partial w(D)/\partial d| / f(D|Z) \leq C$, so that

$$\begin{aligned} E[m(W,\gamma)^2] &= E[\left[-\int \frac{\partial w(D)}{\partial d} \gamma(D,Z)dD \right]^2] = E[\{E[\alpha_0(X)\gamma(X)|Z]\}^2] \\ &\leq E[E[\alpha_0(X)^2\gamma(X)^2|Z]] = E[\alpha_0(X)^2\gamma(X)^2] \leq CE[\gamma(X)^2], \end{aligned}$$

so Assumption 10 is satisfied. The conclusion then follows by Theorem 3. $Q, E.D.$

Proof of Corollary 6: By Assumption 4 and $m(w,\gamma) = \gamma(1,z) - \gamma(0,z)$ so by the triangle inequality

$$|m(W,b_j)| = |b_j(1,Z) - b_j(0,Z)| \leq C,$$

and Assumption 8 is satisfied. Also by hypothesis $|\alpha_0(X)| = |\partial w(D)/\partial d| / f(D|Z) \leq C$, so that

$$\begin{aligned} E[m(W,\gamma)^2] &\leq CE[\gamma(1,Z)^2] + CE[\gamma(0,Z)^2] = CE[\frac{D}{\pi_0(Z)}\gamma(1,Z)^2] + CE[\frac{1-D}{1-\pi_0(Z)}\gamma(0,Z)^2] \\ &= CE[\left\{ \frac{D}{\pi_0(Z)} + \frac{1-D}{1-\pi_0(Z)} \right\} \gamma(X)^2] \leq CE[\gamma(X)^2], \end{aligned}$$

so Assumption 10 is satisfied. The conclusion then follows by Theorem 3. $Q, E.D.$

Proof of Lemma 7: Define

$$\bar{M}_k(\gamma) = (\bar{M}_{k1}(\gamma), ..., \bar{M}_{kp}(\gamma))', \ \bar{M}_{kj}(\gamma) = \int D_k(W, b_{kj}, \gamma)F(dW).$$

For notational convenience we henceforth suppress the k superscript. Let $\mathcal{A}_{\ell,\ell'}$ be the event that $\|\hat{\gamma}_{\ell,\ell'} - \bar{\gamma}\| \leq \varepsilon$ and note that $\Pr(\mathcal{A}_{\ell,\ell'}) \longrightarrow 1$ for each ℓ and ℓ'. When $\mathcal{A}_{\ell,\ell'}$ occurs,

$$\int A(W, \hat{\gamma}_{\ell,\ell'})^2 F(dW) \leq C,$$

by Assumption 11. Define

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

Note that for any constant C' and the event $\mathcal{A} = \{\max_j |U_{\ell' j}(\tilde{\gamma}_{\ell,\ell'})| \geq C'\varepsilon_n^*\}$ where $\varepsilon_n^* = \sqrt{\ln(p)/n}$

$$\begin{aligned} \Pr(\mathcal{A}) &= \Pr(\mathcal{A}|\Gamma_{\ell,\ell'})\Pr(\Gamma_{\ell,\ell'}) + \Pr(\mathcal{A}|\Gamma_{\ell,\ell'}^c)1 - \Pr(\Gamma_{\ell,\ell'}) \qquad \text{(B.6)} \\ &\leq \Pr(\max_j |U_{\ell' j}(\tilde{\gamma}_{\ell,\ell'})| \geq C'\varepsilon_n^* |\Gamma_{\ell,\ell'}) + 1 - \Pr(\Gamma_{\ell,\ell'}). \end{aligned}$$

By Lemma B2 of Bradic et al. (2021) there is C' large enough that for any $\delta > 0$ with probability approaching one,

$$\Pr(\max_j |U_{\ell' j}(\tilde{\gamma}_{\ell,\ell'})| \geq C' \varepsilon_n^* | \Gamma_{\ell,\ell'}) < \delta/2.$$

Also $1 - \Pr(\Gamma_{\ell,\ell'}) \longrightarrow 0$, so that $\Pr(\mathcal{A}) < \delta$ for all n large enough. Therefore

$$\|U_{\ell'}(\tilde{\gamma}_{\ell,\ell'})\|_\infty = \max_j |U_{\ell' j}(\tilde{\gamma}_{\ell,\ell'})| = O_p(\varepsilon_n^*).$$

Next, for each ℓ it follows that $n - n_\ell = \sum_{\ell' \neq \ell} n_{\ell'}$ and

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

Also by Assumption and $\Pr(\Gamma_{\ell,\ell'}) \longrightarrow 1$ for each ℓ and ℓ', ξ

$$\left| \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_\ell} \bar{M}(\tilde{\gamma}_{\ell,\ell'}) - M \right|_\infty = \left| \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_{\ell'}} [\bar{M}(\tilde{\gamma}_{\ell,\ell'}) - M] \right|_\infty \leq C \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_{\ell'}} \|\hat{\gamma}_{\ell,\ell'} - \gamma_0\| = O_p(n^{-d_\gamma}).$$

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

Proof of Theorem 8: The proof proceeds verifying Assumptions 1-3 of Chernozhukov et al. (2020, LR) similarly to the proof of Theorem 3. By Assumption 14, if Assumptions 2 and 3 are satisfied it follows that Assumption 6 is satisfied with $\varepsilon_n = n^{-d\gamma}$, so by Theorem 1,

$$\|\hat{\alpha}_k - \bar{\alpha}\| = o_p(n^c n^{-d_\gamma 2\xi/(2\xi+1)}).$$

Then for $c = [d_\gamma(2\xi/(2\xi+1) + d_\gamma - 1/2]/2 = [d_\gamma(4\xi+1)/(2\xi+1) - 1/2]/2 > 0$ we have $\|\hat{\alpha}_k - \bar{\alpha}_k\|^2 = o_p(1)$ and

$$\sqrt{n}\,\|\hat{\alpha}_k - \bar{\alpha}_k\|\,\|\hat{\gamma}_k - \bar{\gamma}_k\| = \sqrt{n} o_p(n^c n^{-d_\gamma 2\xi/(2\xi+1)}) O_p(n^{-d\gamma}) = o_p(n^c n^{-2c}) = o_p(1),$$

for each k. Similarly, by Assumption 14 if Assumption 7 is satisfied (rather than Assumptions 2 and 3) then by Theorem 2 for any $c > 0$,

$$\|\hat{\alpha}_k - \bar{\alpha}_k\| = o_p(n^c n^{-d_\gamma/2}).$$

Then for $c = [d_\gamma/2 + d_\gamma - 1/2]/2 > 0 = [3d_\gamma/2 - 1/2]/2$ we have $\|\hat{\alpha}_k - \bar{\alpha}\| = o_p(1)$ and

$$\sqrt{n}\,\|\hat{\alpha}_k - \bar{\alpha}_k\|\,\|\hat{\gamma}_k - \bar{\gamma}_k\| = \sqrt{n} o_p(n^c n^{-d_\gamma/2}) O_p(n^{-d\gamma}) = o_p(n^c n^{-2c}) = o_p(1),$$

for each k.

Next, Assumption 1 i) of LR is implied by Assumption 10. Let $\phi_k(w, \gamma_k, \alpha_k) = \alpha_k(x_k)[y_k - \gamma_k(x_k)]$ and

$$\phi(w, \gamma, \alpha) = \sum_{k=1}^{K} \phi_k(w, \gamma_k, \alpha_k)$$

Note that by $E[\{Y_k - \bar{\gamma}(X_k)\}^2|X_k]$ and $\bar{\alpha}_k(X_k)$ bounded,

$$\begin{aligned}\int \{\phi_k(w,\hat{\alpha}_{k\ell},\bar{\gamma}_k) - \phi_k(w,\bar{\alpha}_k,\bar{\gamma}_k)\}^2 F(dw) &= \int \{\hat{\alpha}_{k\ell}(x_k) - \bar{\alpha}_k(x_k)\}^2 [y_k - \bar{\gamma}_k(x_k)]^2 F(dw) \\ &\leq C \left\|\hat{\alpha}_{k\ell} - \bar{\alpha}_k\right\|^2 \xrightarrow{p} 0, \\ \int \{\phi_k(w,\bar{\alpha}_k,\hat{\gamma}_{k\ell}) - \phi_k(w,\bar{\alpha}_k,\bar{\gamma}_k)\}^2 F(dw) &= \int \bar{\alpha}_k(x_k)^2 [\hat{\gamma}_{k\ell}(x_k) - \bar{\gamma}(x_k)]^2 F(dx_k) \\ &\leq C \left\|\hat{\gamma}_{k\ell} - \bar{\gamma}_k\right\|^2 \xrightarrow{p} 0,\end{aligned}$$

so that Assumptions 1 ii) and 1 iii) of LR are satisfied by the triangle inequality.

By the Cauchy-Schwartz and conditional Markov inequalities we have

$$\begin{aligned}&\left|\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}\{\hat{\alpha}_{k\ell}(X_{ki}) - \bar{\alpha}_k(X_{ki})\}\{\hat{\gamma}_{k\ell}(X_{ki}) - \bar{\gamma}_k(X_{ki})\}\right| \\ &\leq \sqrt{n}\sqrt{\frac{1}{n}\sum_{i\in I_\ell}\frac{\{\hat{\alpha}_{k\ell}(X_i) - \bar{\alpha}_k(X_i)\}^2}{n}}\sqrt{\frac{1}{n}\sum_{i\in I_\ell}\frac{\{\hat{\gamma}_{k\ell}(X_{ki}) - \bar{\gamma}_k(X_{ki})\}^2}{n}} \\ &= O_p(\sqrt{n}\left\|\hat{\alpha}_{k\ell} - \bar{\alpha}_k\right\| \left\|\hat{\gamma}_{k\ell} - \bar{\gamma}_k\right\|) = o_p(1).\end{aligned}$$

Then by the triangle inequality Assumption 2 of LR is satisfied.

To verify Assumption 3 of LR, note that by Assumption 1 $\hat{\alpha}_\ell(x) = b(x)'\hat{\rho}_\ell \in \Gamma$, so that

$$\int \phi_k(w,\bar{\gamma}_k,\hat{\alpha}_{k\ell})F_0(dw) = \int \hat{\alpha}_{k\ell}(x_k)[y_k - \bar{\gamma}_k(x_k)]F(dw) = 0.$$

Also note that for each k,

$$\begin{aligned}E[\phi_k(W,\gamma_k,\bar{\alpha}_k)] &= E[\bar{\alpha}_k(X_k)\{Y_k - \gamma_k(X_k)\}] = E[\bar{\alpha}_k(X_k)\{\bar{\gamma}_k(X_k) - \gamma_k(X_k)\}] \\ &= E[D_k(W,\bar{\gamma}_k,\bar{\gamma})] - E[D_k(W,\gamma_k,\bar{\gamma})] = -E[D_k(W,\gamma_k - \bar{\gamma}_k,\bar{\gamma})].\end{aligned}$$

Then by Assumption 13 for all γ with $\|\gamma - \bar{\gamma}\| < \varepsilon$,

$$\begin{aligned}\left|E[\psi(W,\gamma,\bar{\alpha},\bar{\theta})]\right| &= \left|E[m(W,\gamma) - m(W,\bar{\gamma}) + \sum_{k=1}^{K}\phi_k(W,\gamma_k,\bar{\alpha}_k)]\right| \\ &= \left|E[m(W,\gamma) - m(W,\bar{\gamma}) - \sum_{k=1}^{K}D_k(W,\gamma_k - \bar{\gamma}_k,\bar{\gamma})]\right| \leq C\left\|\gamma - \bar{\gamma}\right\|^2,\end{aligned}$$

giving Assumption 3 of LR.

It then follow by Lemma 15 of LR that

$$\sqrt{n}(\hat{\theta} - \bar{\theta}) = \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\psi(W_i,\hat{\gamma}_\ell,\hat{\alpha}_\ell,\bar{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i,\bar{\gamma},\bar{\alpha},\bar{\theta}) + o_p(1).$$

The first conclusion then follows by the central limit theorem and $\psi(W,\bar{\gamma},\bar{\alpha},\bar{\theta})=\psi(W)$.

The second conclusion follows by the triangle inequality as in the proof of Theorem 3 with R_{i2} replaced by $\hat{\alpha}_{k\ell}(X_{ki})^2\{\hat{\gamma}_{k\ell}(X_{ki})-\bar{\gamma}(X_{ki})\}^2$ and R_{i3} by $\{\hat{\alpha}_{\ell}(X_i)-\bar{\alpha}(X_i)\}^2\{Y_i-\bar{\gamma}(X_i)\}^2$. *Q.E.D.*

Proof of Corollary 9: The proof proceeds by showing that the conditions of Theorem 8 are satisfied. By $\bar{\gamma}_k$ bounded for each k and the triangle inequality, $E[m(W,\bar{\gamma})^2]<\infty$. Also, by the triangle inequality,

$$\begin{aligned}
[m(W,\hat{\gamma})-m(W,\bar{\gamma})]^2 &\leq C\sum_{k=1}^{K-1}\left|\hat{\gamma}_K(d,k,Z)\hat{\gamma}_k(d',Z)-\hat{\gamma}_K(d,k,Z)\bar{\gamma}_k(d',Z)\right|^2 \\
&\leq C\sum_{k=1}^{K-1}\left|\hat{\gamma}_K(d,k,Z)-\bar{\gamma}(d,k,Z)|^2\hat{\gamma}_k(d',Z)^2-|\bar{\gamma}_K(d,k,Z)|^2|\hat{\gamma}_k(d',Z)-\bar{\gamma}_k(d',Z)\right|^2 \\
&\leq C\sum_{k=1}^{K-1}(\left|\hat{\gamma}_K(d,k,Z)-\bar{\gamma}(d,k,Z)|^2+|\hat{\gamma}_k(d',Z)-\bar{\gamma}_k(d',Z)\right|^2).
\end{aligned}$$

Therefore we have

$$\int[m(w,\hat{\gamma})-m(w,\bar{\gamma})]^2F(dw)\leq\sum_{k=1}^{K-1}(\int|\hat{\gamma}_K(d,k,z)-\bar{\gamma}(d,k,z)|^2F_Z(dz)+\int|\hat{\gamma}_k(d',z)-\bar{\gamma}_k(d',z)|^2F_Z(dz)).$$

By $\pi(d,k|Z)=\Pr(D=d,Q=k|Z)\geq C$ for each d and k we have for any $\gamma_K(d,k,Z)$

$$\begin{aligned}
\int|\gamma_K(d,k,z)-\bar{\gamma}(d,k,z)|^2F_Z(dz) &= E[|\gamma_K(d,k,Z)-\bar{\gamma}(d,k,Z)|^2] \\
&= E[\frac{1(D=d,Q=k)}{\pi(D,Q|Z)}|\gamma_K(d,k,Z)-\bar{\gamma}(d,k,Z)|^2] \\
&= E[\frac{1(D=d,Q=k)}{\pi(D,Q|Z)}|\gamma_K(D,Q,Z)-\bar{\gamma}(D,Q,Z)|^2] \\
&\leq CE[|\gamma_K(D,Q,Z)-\bar{\gamma}(D,Q,Z)|^2]=C\left\|\gamma_K-\bar{\gamma}_K\right\|^2.
\end{aligned}$$

Applying this calculation to $\gamma_K=\hat{\gamma}_K$ gives

$$\int|\hat{\gamma}_K(d,k,z)-\bar{\gamma}(d,k,z)|^2F_Z(dz)\leq C\left\|\hat{\gamma}_K-\bar{\gamma}_K\right\|^2.$$

Also it follows by $\pi(d,k|Z)\geq C$ for each k that $\pi(d|Z)=\Pr(D=d|Z)\geq C$. Then similarly to the previous inequality we have)

$$|\gamma_k(d',Z)-\bar{\gamma}_k(d',Z)|\leq C\left\|\gamma_k-\bar{\gamma}_k\right\|^2,\ k=1,...,K-1.$$

Then collecting terms we have

$$\int[m(w,\hat{\gamma})-m(w,\bar{\gamma})]^2F(dw)\leq C\sum_{k=1}^{K}\left\|\hat{\gamma}_k-\bar{\gamma}_k\right\|^2\leq C\left\|\hat{\gamma}-\bar{\gamma}\right\|^2,$$

for $\|\gamma\| = \sum_{k=1}^{K} \|\gamma_k\|$. Thus Assumption 10 is satisfied.

Next, by the Gateaux derivative formula in the body of the paper for $(k = 1, ..., K-1)$ we have

$$D_k(W, b_{kj}, \gamma) = a_{kj}(W) A_k(W, \gamma),\ a_{kj}(W) = b_{1j}(d', Z),\ A_k(W, \gamma) = \gamma_K(d, k, Z).$$

It follows similarly to the verification of Assumption 10 and by Assumption 4 that

$$\max_{j \le p} |a_{kj}(W)| \le C, \text{ and } E[A_k(W, \gamma)^2] \le C \|\gamma\|^2,\ (k = 1, ..., K-1).$$

Also, we have

$$D_K(W, b_{Kj}, \gamma) = \sum_{k=1}^{K-1} b_{Kj}(d, k, Z) \gamma_{k1}(d', Z),$$

which also has the form like that Assumption 12 where the conclusion of Lemma 7 will also be satisfied. The second part of Assumption 12 follows by a similar argument, so that Assumption 12 is satisfied.

Turning now to Assumption 13, note that for $(k = 1, ..., K-1)$,

$$E[D_k(W, \gamma_k, \bar{\gamma})] = E[\bar{\gamma}_K(d, k, Z) \gamma_k(d', Z)] = E[\bar{\alpha}_k(X_k) \gamma_k(X_k)],\ \bar{\alpha}_k(X_k) = \frac{\bar{\gamma}_K(d, k, Z) 1(D = d')}{\pi(D|Z)},$$

$$E[D_K(W, \gamma_K, \bar{\gamma})] = E[\sum_{k=1}^{K-1} \gamma_K(d, k, Z) \bar{\gamma}_k(d', Z)] = E[\bar{\alpha}_K(X_K) \gamma_K(X_K)],$$

$$\bar{\alpha}_K(X_K) = \sum_{k=1}^{K-1} \frac{1(D = d, Q = k) \bar{\gamma}_k(d, Z)}{\pi(D, Q|Z)}.$$

Each of $\bar{\alpha}_k(X_k)$ is bounded by $\pi(D, Q|Z) \ge C$ and $\bar{\gamma}_k(X_k)$ bounded for each k.

To verify Assumption 13 iii) note that by algebra we have

$$m(W, \gamma) - m(W, \bar{\gamma}) - \sum_{k=1}^{K} D_k(W, \gamma_k - \bar{\gamma}_k, \bar{\gamma}) = \sum_{k=1}^{K-1} \{\gamma_K(d, k, Z) - \gamma_K(d, k, Z)\}\{\gamma_k(d', Z) - \bar{\gamma}_K(d, Z)\}.$$

Therefore by the Cauchy Scwhartz and triangle inequalities,

$$\begin{aligned}
&\left| E[m(W, \gamma) - m(W, \bar{\gamma}) - \sum_{k=1}^{K} D_k(W, \gamma_k - \bar{\gamma}_k, \bar{\gamma})] \right| \\
&= \left| E[\sum_{k=1}^{K-1} \{\gamma_K(d, k, Z) - \gamma_K(d, k, Z)\}\{\gamma_k(d', Z) - \bar{\gamma}_k(d', Z)\}] \right| \\
&\le \sum_{k=1}^{K-1} \left\{ E[|\gamma_K(d, k, Z) - \bar{\gamma}_K(d, k, Z)|^2] + E[|\gamma_k(d', Z) - \bar{\gamma}_k(d', Z)|^2] \right\} \le C \|\gamma - \bar{\gamma}\|^2,
\end{aligned}$$

where the last inequality follows similarly to previous results. The conclusion now follows by Theorem 8. *Q.E.D.*

Proof of Corollary 10: Note first that for any $\gamma(X)$ it follows as in the proof of Corollary 6 that by $\Pr(D=1|Z) < 1-C$,

$$E[D\gamma(0,Z)^2] \le E[\gamma(0,Z)^2] = E[\frac{1-D}{1-\pi_0(Z)}\gamma(0,Z)^2] = E[\frac{1-D}{1-\pi_0(Z)}\gamma(D,Z)^2] \le CE[\gamma(X)^2].$$

Also note that

$$E[D\gamma(0,Z)] = E[\pi_0(Z)\gamma(0,Z)] = E\left[\pi_0(Z)\frac{1-D}{1-\pi_0(X)}\gamma(0,Z)\right] = E[\alpha_0(X)\gamma(X)].$$

The remainder of the proof follows analogously to the proof of Theorem 3. *Q.E.D.*

B.1 Panel Average Derivative and Demand Elasticities

Since own-price elasticity θ_0^* is a deterministic mapping of $\tilde{\theta}_0 := (\theta_0, \mathbb{E}[Y_{it}])'$, we obtain the asymptotic variance V^* of θ_0^* from the asymptotic variance $\tilde{V}$ of $\tilde{\theta}_0$ using delta method. Specifically,

$$V^* = H\tilde{V}H'$$

where

$$H = \frac{\partial\theta_0^*}{\partial\tilde{\theta}_0} = \begin{bmatrix} \frac{1}{\mathbb{E}[Y_{it}]} & \frac{-\theta_0}{[\mathbb{E}[Y_{it}]]^2} \end{bmatrix}$$

and

$$\tilde{V} = \begin{bmatrix} \mathbb{E}[\psi_0(W_{it})]^2 & \mathbb{E}[\psi_0(W_{it})Y_{it}] \\ \cdot & \mathbb{E}[Y_{it}]^2 - \{\mathbb{E}[Y_{it}]\}^2 \end{bmatrix}.$$

We estimate the asymptotic variance V^* using the empirical analogue $\hat{V}^*$, where $\psi_0(W_{it})$ is replaced by

$$\hat{\psi}_{it} = \frac{\partial\hat{\gamma}_\ell(\tilde{X}_i)}{\partial d} - \hat{\theta} + \hat{\alpha}_\ell^*(\tilde{X}_i)[Y_{it} - \hat{\gamma}_\ell(\tilde{X}_i)], \quad i \in I_\ell.$$

The covariance estimator recognizes that household i's observations form a cluster T_i. For example, the estimator for the first component of $\tilde{V}$ is

$$\frac{1}{\sum_{i=1}^n T_i}\sum_{i=1}^n\sum_{t\in T_i}\sum_{s\in T_i}\hat{\psi}_{it}\hat{\psi}_{is}.$$

More generally, we may consider estimating not only own price elasticity but also income elasticity and cross price elasticity. The same arguments go through with light modification.

Concatenate the derivatives as

$$\theta_0 = \begin{bmatrix} (\theta_0)^{\text{income}} \\ (\theta_0)^{\text{own}} \\ (\theta_0)^{\text{cross}} \end{bmatrix} = \begin{bmatrix} E\left[\frac{\partial \gamma_0(\tilde{X}_i)}{\text{log income}}\right] \\ E\left[\frac{\partial \gamma_0(\tilde{X}_i)}{\text{log own price}}\right] \\ E\left[\frac{\partial \gamma_0(\tilde{X}_i)}{\text{log cross price}}\right] \end{bmatrix}$$

where the first and second components are scalars and the third component is a vector.

The elasticities are a smooth transform thereof. By arguments in Chernozhukov, Hausman, and Newey (2019)

$$\theta_0^* = \begin{bmatrix} (\theta_0^*)^{\text{income}} \\ (\theta_0^*)^{\text{own}} \\ (\theta_0^*)^{\text{cross}} \end{bmatrix} = \begin{bmatrix} \frac{(\theta_0)^{\text{income}}}{\mathbb{E}[Y_{it}]} - 1 \\ \frac{(\theta_0)^{\text{own}}}{\mathbb{E}[Y_{it}]} - 1 \\ \frac{(\theta_0)^{\text{cross}}}{\mathbb{E}[Y_{it}]} \end{bmatrix}.$$

Likewise the delta method argument goes through. Elasticites θ_0 are a deterministic mapping of $\tilde{\theta}_0 = ((\theta_0^*)', \mathbb{E}[Y_{it}])'$. We obtain the asymptotic variance V^* of θ_0^* from the asymptotic variance $\tilde{V}$ of $\tilde{\theta}_0$ using delta method. Specifically,

$$V^* = H\tilde{V}H'$$

where

$$H = \frac{\partial \theta_0^*}{\partial \tilde{\theta}_0} = \begin{bmatrix} \frac{1}{\mathbb{E}[Y_{it}]} \cdot I & \frac{-\theta_0}{[\mathbb{E}[Y_{it}]]^2} \end{bmatrix}$$

and $\tilde{V}$ is as before, where the influence function ψ_0 is vector-valued, corresponding to the vector θ_0.

As an aside, when using OLS, the empirical influence function used in estimating off-diagonal terms is

$$\psi_0(W_{it}) = (\mathbb{E}[b_{it}b_{it}'^{-1}b_{it}\epsilon_{it}$$

$$\hat{\psi}_{it} = \left(\frac{1}{\sum_{i=1}^n T_i} \sum_{i=1}^n \sum_{t \in T_i} \sum_{s \in T_i} b_{it}b_{it}'\right)^{-1} b_{it}\epsilon_{it}$$

where ϵ_{it} is the OLS residual for observation W_{it}. As before, we use a variance estimator that recognizes clustering.

C Additional Empirical Results

C.1 Regression Decomposition and ATT

C.2 Panel Average Derivative and Demand Elasticities

We present elasticity estimates from OLS with a simpler specification than the specification used in the main text. We take as $b_1(X_{it})$ the concatenation of the following variables: log

spec.	treated	untreated	Lasso ATT	Lasso SE	RF ATT	RF SE	NN ATT	NN SE
1	185	254	1181.65	577.82	1627.93	584.52	1651.97	577.34
2	185	250	1191.35	580.46	1685.15	585.05	1603.91	582.80
3	185	179	1103.52	590.74	1935.65	575.73	1746.51	590.21

Table 7: ATT using NSW treatment and NSW control, by plug-in

spec.	treated	untreated	Lasso ATT	Lasso SE	RF ATT	RF SE	NN ATT	NN SE
1	185	844	89.91	593.32	1214.55	617.46	-1795.95	579.00
2	185	1620	492.77	594.27	926.93	624.56	-0.35	595.95
3	185	2336	715.61	605.92	1001.66	609.32	1573.30	639.90

Table 8: ATT using NSW treatment and PSID comparison, by plug-in

spec.	treated	untreated	Lasso ATT	Lasso SE	RF ATT	RF SE	NN ATT	NN SE
1	185	7687	195.91	595.90	1350.67	635.41	1345.07	605.18
2	185	13326	851.08	599.47	1434.67	621.56	1580.90	611.67
3	185	13449	979.20	605.02	1253.08	627.57	1350.68	668.13

Table 9: ATT using NSW treatment and CPS comparison, by plug-in

elasticity	point	SE
income	0.42	0.05
own-price	-0.68	0.05
bread	-0.03	0.02
butter	0.00	0.02
cereal	0.00	0.02
chips	0.02	0.03
coffee	0.00	0.02
cookies	0.00	0.02
eggs	-0.03	0.03
ice cream	-0.03	0.03
orange juice	-0.01	0.05
salad	0.02	0.02
soda	-0.02	0.02
soup	-0.03	0.02
water	-0.01	0.02
yogurt	0.01	0.04

Table 10: Milk elasticities, by OLS

expenditure, and log price of each good. For $\tilde{H}_i$, we use the time averages of $b_1(X_{it})$. Note that $K = dim(b_1(X_{it})) = 16$ and $p = dim(b_{it}) = 288$. We calculate clustered standard errors derived by delta method as explained in Appendix **??**. Tables 10 and 11 summarize results.

Making the analogous replacement in the constraints of the Dantzig selector (Candes and Tao, 2007) gives a Dantzig estimator

$$\hat{\rho}_D = \arg\min_{\rho} |\rho|_1 \, s.t. |\hat{M} - \hat{G}\rho|_\infty \leq \lambda_D, \tag{C.1}$$

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 $\hat{\rho}_L$ or $\hat{\rho}_D$ may be used in equation (3.3) to form an estimator $\hat{\alpha}(x) = b(x)'\hat{\rho}_L$ or $\hat{\alpha}(x) = b(x)'\hat{\rho}_D$. This $\hat{\alpha}(x)$ may then be substituted in equation (3.1), along with a machine learner $\hat{\gamma}$ of the regression, to construct Auto-DML $\hat{\theta}$.

elasticity	point	SE
income	0.64	0.02
own-price	-0.65	0.04
bread	-0.01	0.03
butter	-0.04	0.02
cereal	0.01	0.03
chips	0.02	0.03
coffee	0.01	0.02
cookies	-0.03	0.02
eggs	-0.03	0.03
ice cream	0.01	0.03
milk	0.02	0.04
orange juice	-0.05	0.05
salad	-0.03	0.02
soup	0.01	0.03
water	0.01	0.02
yogurt	0.05	0.04

Table 11: Soda elasticities from regression, by OLS

References

Ahn, H. and C.F. Manski (1993): “Distribution Theory for the Analysis of Binary Choice under Uncertainty with Nonparametric Estimation of Expectations,” *Journal of Econometrics* 56, 291–321.

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

Avagyan, V. and S. Vansteelandt (2017): "Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation," https://arxiv.org/abs/1708.03787.

Belloni, A., D. Chen, and V. Chernozhukov (2012): “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain,” *Econometrica* 80, 2369–429.

Belloni, A. and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli* 19, 521–547.

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

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

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

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.

Blundell, R.W. and J.L. Powell (2004): "Endogeneity in Binary Response Models," *Review of Economic Studies* 71, 655-679.

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

Bradic, J., S. Wager, and Y. Zhu (2019): "Sparsity Double Robust Inference of Average Treatment Effects," https://arxiv.org/pdf/1905.00744.pdf.

Bradic, J., V. Chernozhukov, W. Newey, and Y. Zhu (2021): "Minimax Semiparametric Learning with Approximate Sparsity," arXiv.

Burda, M., M. Harding, J.A. Hausman (2008): “A Bayesian Mixed Logit Probit Model for Multinomial Choice,” *Journal of Econometrics* 147, 232–46.

Burda, M., M. Harding, J.A. Hausman (2012): “A Poisson Mixture Model of Discrete Choice,” *Journal of Econometrics* 166, 184–203.

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.

Chamberlain, G. (1982): “Multivariate Regression Models for Panel Data,” *Journal of Econometrics* 18, 5–46.

Chamberlain, G. (1982): “Efficiency Bounds for Semiparametric Regression,” *Econometrica* 60, 567–96.

Chamberlain, G. (1984): "Panel Data," *Handbook of Econometrics Vol 2*, Z. Griliches and M. Intriligator, eds., 1247-1318.

Chatterjee, S. and J. Jafarov (2015): “Prediction Error of Cross-Validated Lasso,” arXiv:1502.06291.

Chen, X. and H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory* 45, 682-691.

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

Chernozhukov, V., I. Fernandez-Val, J. Hahn, W. Newey (2013b): “Average and Quantile Effects in Nonseparable Panel Models,” *Econometrica* 81, 535–80.

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., J. C. Escanciano, H. Ichimura, W.K. Newey, and J. Robins (2016): “Locally Robust Semiparametric Estimation,” https://arxiv.org/abs/1608.00033v1.

Chernozhukov, V., J.A. Hausman, and W.K. Newey (2016): “Demand Analysis with Many Prices," Heterogeneity in Supply and Demand workshop, Boston College, December.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017): “Central Limit Theorems and Bootstrap in High Dimensions," *The Annals of Probability* 45: 2309–2352.

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. and W.K. Newey (2018): "Double Machine Learning for Linear Functionals of Projections," Workshop on the Interface of Machine Learning and Statistical Inference," Banff International Research Station, January.

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

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

Chernozhukov, V., W.K. Newey, and R. Singh (2019): "Double/De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," https://arxiv.org/abs/1802.08667v3

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W.K. Newey, and J. Robins (2020): “Locally Robust Semiparametric Estimation,” https://arxiv.org/abs/1608.00033v4.

Chiang, H.D., K. Kato, Y. Ma, Y. Sasaki (2019): “Multiway Cluster Robust Double/Debiased Machine Learning," arXiv:1909.03489.

Daubechies, I., M Defrise, and C. De Mol (2004): “An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint,” *Communications on Pure and Applied Mathematics* 57, 1413–57.

Dehejia, R.H. and S. Wahba (1999): “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association* 94 (448): 1053–62.

Farbmacher, M., M. Huber, L. Laffers, H. Langen, M. Spindler (2020): "Causal Mediation Analysis with Double Machine Learning," https://arxiv.org/abs/2002.12710.

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

Farrell, M., T. Liang, S. Misra (2018): "Deep Neural Networks for Estimation and Inference," https://arxiv.org/pdf/1809.09953.pdf.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007): “Pathwise Coordinate Optimization,” *The Annals of Applied Statistics* 1, 302–32.

Friedman, J., T. Hastie, and R. Tibshirani (2010): “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software* 33, 1-22.

Fu, W.J. (1998): “Penalized Regressions: The Bridge versus the Lasso,” *Journal of Computational and Graphical Statistics* 7, 397–416.

Graham, B. and J.L. Powell (2012): “Identification and Estimation of Average Partial Effects in “Irregular” Correlated Random Coefficient Panel Data Models,” *Econometrica* 80, 2105–52.

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 (2017): “Balancing Out Regression Error: Efficient Treatment Effect Estimation without Smooth Propensities,” arXiv:1712.00038v1.

Hirshberg, D.A. and S. Wager (2020): “Debiased Inference of Average Partial Effects in Single-Index Models,” *Journal of Business and Economic Statistics* 38, 19-24.

Hirshberg, D.A. and S. Wager (2019): “Augmented minimax linear estimation,” arXiv:1712.00038v5.

Imai, K, L. Keele, and D. Tingley (2010): "A General Approach to Causal Mediation Analysis," *Psychological Methods* 15, 309 –334.

Imbens, G.W. and W.K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica* 77, 1481-1512.

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:1601.00815.

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

LaLonde, R.J. (1986): “Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review* 76, 604–20.

Leeb, H., and B.M. Pötscher (2008a): “Recent Developments in Model Selection and Related Areas," *Econometric Theory* 24, 319–22.

Leeb H., and B.M. Pötscher (2008b): “Sparse Estimators and the Oracle Property, or the Return of Hodges’ Estimator," *Journal of Econometrics* 142, 201–211.

Luo, Ye and M. Spindler (2016): "High-Dimenstional L2 Boosting: Rate of Convergence," https://arxiv.org/pdf/1602.08927.pdf.

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:1801.09138.

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: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., 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..

Rothenhäusler, D. and B. Yu (2019): “Incremental Causal Effects,” arXiv:1907.13258.

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.

Rudelson, M. and S. Zhou (2013): "Reconstruction From Anisotropic Random Measurements," *IEEE Transactions on Informating Theory* 59, 3434–3447.

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

Singh, R. and L. Sun (2019): “De-biased Machine Learning for Compliers,” arXiv:1909.05244.

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

Syrgkanis, V., and M. Zampetakis (2020): "Estimation and Inference with Trees and Forests in High Dimensions," https://arxiv.org/abs/2007.03210.

Tchetgen Tchetgen, E.J. and I. Shipster (2012): "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness and Sensitivity Analysis," *The Annals of Statistics* 40, 1816-1845.

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.

Tseng, P. (2001): “Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization,” *Journal of Optimization Theory and Applications* 109, 475–94.

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. (1991): “On Differentiable Functionals,” *Annals of Statistics*, 19: 178–204.

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

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

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

Wooldridge, J.M. (2010): *Econometric Analysis of Cross-Section and Panel Data*, Cambridge, MIT Press.

Wooldridge, J.M. (2019): “Correlated Random Effects Models with Unbalanced Panels,” *Journal of Econometrics* 211, 137–50.

Wooldridge, J.M. and Y. Zhu (2020): "Inference in Approximately Sparse Correlated Random Effects Probit Models With Panel Data," *Journal of Business and Economic Statistics* 38, 1-18.

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

Zhu, Y. and J. Bradic (2017b): “Breaking the Curse of Dimensionality in Regression,” arXiv: 1708.00430.

Zubizarreta, J.R. (2015): “Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data,” *Journal of the American Statistical Association* 90 (429): 122–129.

Imai, K, L. Keele, and D. Tingley (2010): "A General Approach to Causal Mediation Analysis," *Psychological Methods* 15, 309–334.