De-biased Machine Learning in Instrumental Variable Models for Treatment Effects

Rahul Singh∗
MIT Economics
rahul.singh@mit.edu

Liyang Sun∗
MIT Economics
lsun20@mit.edu

Abstract

Instrumental variable identification is a strategy in causal statistics for estimating the counterfactual effect of treatment \( D \) on output \( Y \) controlling for covariates \( X \) using observational data. Even in the presence of an unmeasured confounder of \((Y, D)\), the treatment effect on the subpopulation of compliers can nonetheless be identified if an instrumental variable \( Z \) is available. We introduce a de-biased machine learning (DML) approach to estimating complier parameters with high-dimensional data. Complier parameters include local average treatment effect, average complier characteristics, and complier counterfactual outcome distributions. In our approach, the de-biasing is itself performed by machine learning, a variant called automatic de-biased machine learning (Auto-DML). We prove our estimator is consistent, asymptotically normal, and semi-parametrically efficient. In experiments, our estimator outperforms state-of-the-art alternatives, and it does not require ad hoc trimming or censoring of a learned propensity score. We use it to estimate the effect of 401(k) participation on the distribution of net financial assets.

1 Introduction

Instrumental variable (IV) identification is a strategy in causal statistics for estimating the counterfactual effect of treatment \( D \) on output \( Y \) controlling for covariates \( X \) using observational data [50]. Even in the presence of an unmeasured confounder of \((Y, D)\), the treatment effect can nonetheless be identified if an instrumental variable \( Z \) is available. Intuitively, \( Z \) only influences \( Y \) via \( D \), identifying the counterfactual relationship of interest. This solution comes at a price; the analyst can no longer measure parameters of the entire population such as average treatment effect (ATE). Instead, the analyst can only measure parameters defined for the subpopulation of compliers such as local average treatment effect (LATE). A complier is an individual whose treatment status \( D \) is affected by variation in the instrument \( Z \). In public policy, the instruments take the form of changes in eligibility criteria for social programs. Compliers are thus of policy interest as they are exactly the subpopulation to be affected by eligibility changes. In digital platforms, the instruments take the form of randomized recommendations. Compliers are thus of business interest as they are exactly the subpopulation to be affected by different recommendations.

To fix ideas, we provide examples with continuous outcome \( Y \), binary treatment \( D \), and binary instrument \( Z \). Charter school admission by lottery \((Z)\) only influences student test scores \((Y)\) via actually attending the charter school \((D)\), identifying the counterfactual effect of the charter school on test scores even if there is selection bias in which students choose to accept an offer of admission [8, 10]. However, the analyst can only learn the treatment effect on the subpopulation of complier students: those who would attend the charter school if they won the lottery and who would not attend the charter school if they lost the lottery. As another example, the randomized recommendation to
enroll on a digital platform \((Z)\) only influences platform engagement \((Y)\) via actual enrollment \((D)\), identifying the counterfactual effect of enrollment on engagement even in the scenario of imperfect compliance \([51]\). However, the analyst can only learn the treatment effect on the subpopulation of complier consumers: those who would enroll if recommended and who would not enroll if not recommended.

In the present work, we introduce a de-biased machine learning (DML) approach to estimating complier parameters with high-dimensional data \([22, 33]\). Our instrumental variable identification assumption is in terms of potential outcomes, and it does not require any functional form restrictions \([9]\). As such, we allow the possibility of heterogeneous treatment effects, i.e. that the treatment effect varies with covariate \(X\) as in \([34]\). We also allow for nonlinear models, which are often appropriate when output \(Y\) is binary. Approximating heterogeneity and nonlinearity with a regularized, black-box machine learning (ML) algorithm can introduce bias into complier parameter estimation \([22]\). However, by situating the estimation problem in a DML framework, we correct for this bias. In our approach, the de-biasing is itself performed by machine learning, a variant called automatic de-biased machine learning (Auto-DML) \([28, 29]\). We present a general estimator, then specialize it to the tasks of learning LATE, average complier characteristics, and complier counterfactual outcome distributions. Counterfactual outcome distributions are particularly important in welfare analysis of schooling, subsidized training, union status, minimum wages, and transfer programs \([1, 3]\).

We make five contributions. First, we extend the theory of de-biased machine learning with automatic bias correction pioneered by \([28, 29]\). Whereas \([28, 29]\) consider parameters of the full population identified by selection on observables, we consider parameters of the complier subpopulation identified by instrumental variables. Second, we prove our estimator is consistent, asymptotically normal, and semi-parametrically efficient. Rather than considering any specific parameter, our results apply to a general class of parameters identified by instrumental variables. Third, we justify simultaneous confidence bands by Gaussian multiplier bootstrap \([23]\). Fourth, we re-interpret a widely-used algorithm for estimating complier parameters called \(\kappa\)-weighting as the Riesz representer in DML; it is in fact a component of the de-biasing term. Fifth, we show our approach outperforms alternative approaches to estimating complier parameters and does not require ad hoc pre-processing, suggesting Auto-DML may be an effective paradigm in high-dimensional causal inference.

For clarity of exposition, we analyze the setting with discrete or continuous outcome \(Y\), binary treatment \(D\), and binary instrument \(Z\). This setting has a simple, intuitive definition of the complier subpopulation. The same analysis goes through with discrete or continuous treatment \(D\) and discrete instrument \(Z\), albeit with more elaborate definitions. For the case of continuous instrument \(Z\), similar results are possible at slower rates by using our approach to extend results by \([30]\), who consider the unconfounded setting with continuous treatment \(D\). In this sense, the present work provides a framework for de-biased machine learning of causal parameters using instrumental variables.

Section 2 discusses related work. Section 3 defines complier parameters in terms of potential outcomes, and Section 4 characterizes their de-biased moment functions. Section 5 presents the Auto-DML algorithm for complier parameters and a bootstrap procedure for simultaneous confidence bands. Section 6 proves consistency, asymptotic normality, semi-parametric efficiency, and validity of the bootstrap. Section 7 compares the empirical performance of Auto-DML with other estimators. Section 8 concludes.

## 2 Related Work

Several approaches have been proposed to estimate specific complier parameters by DML. Both \([44]\) and \([22]\) present a DML estimator for LATE. The justification in \([44]\) is via inverse propensity weighting, while the justification in \([22]\) is by interpreting LATE as a ratio of ATEs. In \([17]\), the authors present a DML estimator for counterfactual outcome distributions with simultaneous confidence bands. All of these estimators involve plugging in an estimated propensity score in the denominator, which is numerically unstable. Unlike previous work, we present a general justification that covers a broad class of estimators, and we present an Auto-DML variant that eliminates the numerically unstable step of plugging in an estimated propensity score. As far as we know, ours is the first DML and Auto-DML estimator of complier characteristics. For a comparison between Auto-DML and other approaches to semi-parametric estimation that use machine learning–namely
We now formalize our causal assumption about the instrument $Z$. A potential outcome is a latent random variable expressing a counterfactual outcome given a hypothesis. With neural networks, [49] with RKHS methods, [13] with random forests, and [51] with black-box ML, we provide this justification. Moreover, by introducing the Auto-DML variant, we are able to learn the $\kappa$-weight directly without estimating its components or even knowing its functional form.

Finally, our paper contributes to the growing literature on instrumental variables in machine learning. Both [35] and [49] consider the problem of nonparametric instrumental variable regression, where the target parameter is the structural function $h$ that summarizes the counterfactual relationship: $Y = h(D, X) + e$, and $e$ is confounding noise. In [13] and [51], the authors further assume the function $h$ can be decomposed as $h(D, X) = \mu(X) + \tau(X)D$, where $\tau(X)$ is interpretable as a heterogeneous treatment effect. Importantly, these works [35, 49, 13, 51] assume that the confounding noise $e$ is additively separable—a model proposed by [42]. In this setting, [35] introduce nonlinearity with neural networks, [49] with RKHS methods, [13] with random forests, and [51] with black-box ML. In our setting, we do not assume additive separability of confounding noise—a model considered by [9]. Our target parameters are functionals of the underlying regression $\mathbb{E}[V|Z,X]$, where $V$ is a vector of relevant random variables defined below. Such parameters are called semi-parametric. (If the instrument $Z$ were continuous, then the target parameter would be non-parametric.) We allow black-box ML for nonlinear estimation of $\mathbb{E}[V|Z,X]$.

3 Problem Setting and Definitions

Let $W = (Y, D, Z, X)'$ concatenate the random variables. $Y \in \mathcal{Y} \subset \mathbb{R}$ is the continuous outcome, $D \in \{0, 1\}$ is the binary treatment, $Z \in \{0, 1\}$ is the binary instrumental variable, and $X \in \mathcal{X} \subset \mathbb{R}^{\text{dim}(X)}$ is the covariate. We observe $n$ i.i.d. observations $\{W_i\}_{i=1}^n$. Wherever possible, we suppress index $i$ to lighten notation.

Instrumental variable identification requires an assumption expressed in terms of potential outcomes. A potential outcome is a latent random variable expressing a counterfactual outcome given a hypothetical intervention. We recommend [39], [45], and [36] for a clear introduction to this framework for causal inference. Following the notation of [9], we denote by $Y^{(z,d)}$ the potential outcome under the intervention $Z = z$ and $D = d$. We denote by $D^{(z)}$ the potential treatment under the intervention $Z = z$. Compliers are the subpopulation for whom $D^{(1)} > D^{(0)}$.

We now formalize our causal assumption about the instrument $Z$, quoting [9]. This prior knowledge, described informally in the introduction, allows us to define and recover the counterfactual effect of treatment $D$ on outcome $Y$ for compliers.

Assumption 1 (Identification), Assume

1. Independence: $\{Y^{(z,d)}\}_{z,d \in \{0,1\}}, \{D^{(z)}\}_{d \in \{0,1\}} \perp Z|X$
2. Exclusion: $\mathbb{P}(Y^{(1,d)} = Y^{(0,d)}|X) = 1$ for $d \in \{0,1\}$
3. Overlap: $\pi_0(X) := \mathbb{P}(Z = 1|X) \in (0,1)$
4. Monotonicity: $\mathbb{P}(D^{(1)} \geq D^{(0)}|X) = 1$ and $\mathbb{P}(D^{(1)} > D^{(0)}|X) > 0$

The independence condition states that the instrument $Z$ is as good as randomly assigned conditional on covariates $X$. The exclusion condition imposes that the instrument $Z$ only affects the outcome $Y$ via the treatment $D$. We can therefore simplify notation: $Y^{(d)} = Y^{(1,d)} = Y^{(0,d)}$. The overlap condition ensures that there are no covariate values for which the instrument is deterministic. The monotonicity condition rules out the possibility of defiers: individuals who will always pursue an opposite treatment status from their assignment.
We extend DML to estimate complier parameters. Specifically, we demonstrate how Assumption 1, γ0, the regression of a random vector V = (V1, ..., Vj)' conditional on (Z, X) as
\[ \gamma_0(Z, X) = \begin{bmatrix} \gamma_{V1}^0(Z, X) \\ \vdots \\ \gamma_{Vj}^0(Z, X) \end{bmatrix} = \mathbb{E}[V | Z, X] \]
where \( \gamma_{Vj}^0(Z, X) = \mathbb{E}[V_j | Z, X] \). The random vector V is observable and depends on the complier parameter of interest; we specify its components for LATE, complier characteristics, and counterfactual outcome distributions in Theorem 1. We denote the classic Horvitz-Thompson weight with
\[ \pi_0(Z) = \frac{Z}{\pi_0(X)} - \frac{1 - Z}{1 - \pi_0(X)} = \frac{Z - \pi_0(X)}{\pi_0(X)[1 - \pi_0(X)]}, \quad \pi_0(X) = \mathbb{P}(Z = 1 | X) \]
Lastly, we denote by \( | \cdot | \) the \( \ell_q \) norm of a vector, and we denote by \( \| \cdot \| \) the \( L_2 \) norm of a random variable, i.e. \( \| V_j \| = \sqrt{\mathbb{E}[V_j^2]} \).

4 Learning Problem

DML is a method of moments framework for semi-parametric estimation with de-biasing and strong statistical guarantees [24]. We review the DML learning problem: in stage 1, learn the regression \( \gamma_0 \) and an additional nuisance parameter called the Riesz representer; in stage 2, estimate the parameter of interest \( \theta_0 \) using method of moments with a de-biased moment function and the stage 1 estimates. We extend DML to estimate complier parameters. Specifically, we demonstrate how Assumption 1, expressed in terms of potential outcomes, implies a moment function and a corresponding de-biased moment function for the complier parameters in Definition 1. Its de-biasing term is precisely the normalized \( \kappa \)-weight, a result we prove in Appendix A.3.

4.1 De-biased Learning

Consider a causal parameter \( \theta_0 \) implicitly defined by
\[ \mathbb{E}[m(W, \gamma_0, \theta)] = 0 \quad \text{iff} \quad \theta = \theta_0 \]
Here \( m \) is called the moment function, and it defines the causal parameter \( \theta_0 \). \( \gamma_0 \) is the regression, a nuisance parameter that must be estimated in order to estimate the parameter of interest \( \theta_0 \).

The plug-in approach involves estimating \( \hat{\gamma} \) in stage 1 by some black-box ML algorithm, and estimating \( \hat{\theta} \) in stage 2 by method of moments with moment function \( m \). The plug-in approach is badly biased when \( \hat{\gamma} \) involves regularization [22]. (The plug-in approach is not biased if \( \hat{\gamma} \) is estimated by OLS, an unbiased estimator. However, the OLS estimator will poorly approximate a nonlinear regression \( \gamma_0 \).

The DML approach uses a more sophisticated moment function [40].
\[ \mathbb{E}[\psi(W, \gamma_0, \alpha_0, \theta)] = 0 \quad \text{iff} \quad \theta = \theta_0 \]
\[ \psi(W, \gamma, \alpha, \theta) = m(W, \gamma, \alpha, \theta) + \phi(W, \gamma, \alpha, \theta) \]
\( \phi \) is called the de-biasing term. We derive \( \phi \) such that \( \psi \) is doubly robust. In particular, we derive \( \phi \) such that
\[ \mathbb{E}[\psi(W, \gamma, \alpha_0, \theta_0)] = 0 \quad \forall \gamma \text{ s.t. } \mathbb{E}[\gamma_j(Z, X)]^2 < \infty \]
\[ \mathbb{E}[\psi(W, \gamma_0, \alpha_0, \theta_0)] = 0 \quad \forall \alpha \text{ s.t. } \mathbb{E}[\alpha(Z, X)]^2 < \infty \]
so stage 2 estimation of \( \hat{\theta} \) by method of moments with moment function \( \psi \) is asymptotically invariant to estimation error of either \( \hat{\gamma} \) or \( \hat{\gamma}_0 \). In this sense, introducing the additional term \( \phi \) serves to de-bias the original moment function \( m \). The learning problem for the causal parameter \( \theta_0 \) is now Neyman orthogonal to the learning problem for the nuisance parameters \( (\gamma_0, \alpha_0) \).

Importantly, the doubly robust moment function \( \psi \) introduces an additional nuisance parameter \( \alpha_0 \), a component of the Riesz representer, which must be estimated in stage 1. Whereas DML involves estimating \( \hat{\alpha} \) by estimating its components and knowing its functional form, we estimate \( \hat{\alpha} \) directly by Auto-DML.

### 4.2 Example: LATE

To fix ideas, we present the example of LATE.

\[
\theta_0 = \mathbb{E}[Y^{(1)} - Y^{(0)} | D^{(1)} > D^{(0)}]
\]

\[
= \frac{\mathbb{E}[Y | Z = 1] - \mathbb{E}[Y | Z = 0]}{\mathbb{E}[D | Z = 1] - \mathbb{E}[D | Z = 0]}
\]

\[
= \frac{\mathbb{E} \{ \mathbb{E}[Y | Z = 1, X] - \mathbb{E}[Y | Z = 0, X] \}}{\mathbb{E} \{ \mathbb{E}[D | Z = 1, X] - \mathbb{E}[D | Z = 0, X] \}}
\]

where the first expression is from Definition 1, the second expression is the Wald formula appealing to Assumption 1 and the classic result of [37, Theorem 1], and the third expression we call the expanded Wald formula appealing to the law of iterated expectations (LIE). Rearranging and using the notation

\[
\gamma_0(Z, X) = \mathbb{E}[V | Z, X], \quad \mathbf{V} = \begin{bmatrix} Y \\ D \end{bmatrix}
\]

we arrive at the moment function \( m \) formulation of LATE.

\[
\mathbb{E} \{ [1 - \theta] (\gamma_0(1, X) - \gamma_0(0, X)) \} = 0 \text{ iff } \theta = \theta_0
\]

The plug-in approach involves estimating \( \hat{\gamma} \) in stage 1 by some black-box ML algorithm, and estimating \( \hat{\theta} \) in stage 2 by method of moments with this moment function. Equivalently, the plug-in approach involves estimating \( \mathbb{E}[Y | Z, X] \) and \( \mathbb{E}[D | Z, X] \) and plugging these estimates into the expanded Wald formula.

It is possible to directly derive the de-biasing term \( \phi \) and hence the doubly robust moment function \( \psi \) for LATE by standard arguments. Using the notation

\[
\alpha_0(Z, X) = \frac{Z}{\pi_0(X)} - \frac{1 - Z}{1 - \pi_0(X)}, \quad \pi_0(X) = \mathbb{P}(Z = 1 | X)
\]

the doubly robust moment function \( \psi \) formulation of LATE is

\[
\mathbb{E} \{ [1 - \theta] (\gamma_0(1, X) - \gamma_0(0, X)) + \alpha_0(Z, X) [1 - \theta] (V - \gamma_0(Z, X)) \} = 0 \text{ iff } \theta = \theta_0
\]

It turns out that a single, unifying argument—given in Theorem 1 below—can derive the doubly robust moment function for not only LATE but also average complier characteristics and complier counterfactual outcome distributions. In Appendix A.2, we compare this de-biased LATE characterization with the so-called forbidden regression discussed in [11, Section 4.6.1]. The de-biased framework allows for consistent estimation even for forbidden regressions, i.e. nonlinear first-stage and reduced-form specifications.

### 4.3 Complier Parameters

As our first result, we derive the doubly robust moment functions for the complier parameters in Definition 1. We show that these moment functions share a common structure.

**Theorem 1** (Doubly robust moment functions). Under Assumption 1, the doubly robust moment functions for LATE, average complier characteristics, and complier counterfactual outcome distributions are of the form

\[
\psi(w, \gamma, \alpha, \theta) = m(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta)
\]

\[
m(w, \gamma, \theta) = A(\theta)[\gamma(1, x) - \gamma(0, x)]
\]

\[
\phi(w, \gamma, \alpha, \theta) = \alpha(z, x) A(\theta)[v - \gamma(z, x)]
\]

where
1. For LATE, \( V = (Y, D)' \) and \( A(\theta) = [1 - \theta] \)
2. For complier characteristics, \( V = (Df(X)', D)' \) and \( A(\theta) = [1 - \theta] \)
3. For complier counterfactual distributions, \( V^y = ((D - 1)1_{Y \leq y}, D1_{Y \leq y}, D)' \) and \( A(\theta^y) = \begin{bmatrix} 1 & 0 & -\delta y \\ 0 & 1 & -\delta y \end{bmatrix} \)

Formally, \( \alpha_0(z, x)A(\theta_0) \) is the Riesz representer to the continuous linear functional \( \gamma \mapsto \mathbb{E}[A(\theta_0)(\gamma(1, x) - \gamma(0, x))] \), i.e.
\[
\mathbb{E}[A(\theta_0)(\gamma(1, X) - \gamma(0, X))] = \mathbb{E}[\alpha_0(Z, X)A(\theta_0)\gamma(Z, X)], \quad \forall \gamma \text{ s.t. } \mathbb{E}[\gamma_j(Z, X)]^2 < \infty
\]
Indeed, this fact directly follows from the classic Horvitz-Thompson derivation that \( \alpha_0(z, x) \) is the Riesz representer to the continuous linear functional \( \gamma \mapsto \mathbb{E}[\gamma(1, X) - \gamma(0, X)] \), i.e.
\[
\mathbb{E}[\gamma(1, X) - \gamma(0, X)] = \mathbb{E}[\alpha_0(Z, X)\gamma(Z, X)], \quad \forall \gamma \text{ s.t. } \mathbb{E}[\gamma_j(Z, X)]^2 < \infty
\]
In Appendix A.3, we review the classic Horvitz-Thompson derivation. We also prove a more general version of Theorem 1 for the entire class of complier parameters, and we demonstrate that the \( \kappa \)-weight is a reparametrization of the Riesz representer \( \alpha_0(z, x)A(\theta_0) \).

5 Algorithm

In [24], the authors show it is data-efficient and theoretically elegant to use sample splitting in DML [19, 48]. The Auto-DML algorithm of [28, 29] is as follows.

**Algorithm 1** (De-biased machine learning). Partition the sample into subsets \( \{I_\ell\}_{\ell=1:L} \).
1. For each \( \ell \), estimate \( \hat{\gamma}_{-\ell} \) and \( \hat{\alpha}_{-\ell} \) from observations not in \( I_\ell \)
2. Estimate \( \hat{\theta} \) as the solution to
\[
\frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta) \bigg|_{\theta = \hat{\theta}} = 0
\]

Our theoretical guarantees apply to Dantzig selector or Lasso estimators of \( \hat{\alpha}_{-\ell} \), originally presented in [28] and [29], respectively. In what follows, we restrict attention to Lasso.

Consider the projection of \( \alpha_0(z, x) \) onto \( p \)-dimensional dictionary \( b(z, x) \). A high-dimensional dictionary \( b \) allows for flexible approximation of \( \alpha_0 \) [18]. With \( \ell_1 \)-regularization, the objective becomes
\[
\rho_L = \underset{\rho}{\text{argmin}} \mathbb{E}[\alpha_0(Z, X) - \rho' b(Z, X)]^2 + 2\lambda_L |\rho|_1
\]
where \( \lambda_L = \sqrt{\frac{\ln p}{n}} \) is a theoretical regularization level. Next, extend the Riesz representer result component-wise.
\[
\mathbb{E}[b(1, X) - b(0, X)] = \mathbb{E}[\alpha_0(Z, X)b(Z, X)], \quad \forall b \text{ s.t. } \mathbb{E}[b_j(Z, X)]^2 < \infty
\]
Expanding the square, ignoring terms without \( \rho \), and using this Riesz representer result
\[
\rho_L = \underset{\rho}{\text{argmin}} -2\rho' \mathbb{E}[b(1, X) - b(0, X)] + \rho' \mathbb{E}[b(Z, X)b(Z, X)] |\rho| + 2\lambda_L |\rho|_1
\]
The empirical analogue to the above expression yields an estimator of \( \hat{\rho} \). In this paper, we consider \( \hat{\alpha}(z, x) = \hat{\rho}' b(z, x) \) as in [29].

**Algorithm 2** (Regularized Riesz representer). For observations in \( I_{-\ell} = \{1, \ldots, n\} \setminus I_\ell \)
1. Calculate \( p \times p \) matrix \( \hat{G}_{-\ell} = \frac{1}{n-n_\ell} \sum_{i \in I_{-\ell}} b(Z_i, X_i) b(Z_i, X_i)' \)
2. Calculate \( p \times 1 \) vector \( \hat{M}_{-\ell} = \frac{1}{n-n_\ell} \sum_{i \in I_{-\ell}} b(1, X_i) - b(0, X_i) \)
3. Set \( \hat{\alpha}_{-\ell}(z, x) = b(z, x) \hat{\alpha}_{-\ell} \) where \( \hat{\alpha}_{-\ell} = \arg\min_{\rho} \rho^T \hat{G}_{-\ell} - 2 \rho^T \hat{M}_{-\ell} + 2 \lambda_n |\rho|_1 \)

In Appendix A.4, we discuss how Algorithm 2 automatically attenuates the influence of outliers, which is a central issue in applied statistical research. Specifically, we provide a finite sample balancing property in Proposition 3. In Appendix A.5, we provide and justify an iterative tuning procedure for data-driven regularization parameter \( \lambda_n \).

Likewise, we can project \( \gamma_0(z, x) \) onto \( p \)-dimensional dictionary \( b(z, x) \) using the functional \( b \mapsto \mathbb{E}[b(Z, X) \mathbb{V}] \). Our theoretical results are agnostic about the choice of estimator \( \hat{\gamma} \); it may be this estimator or any other black-box ML algorithm satisfying the rate condition specified in Assumption 7.

Suppose we wish to form a simultaneous confidence band for the components of \( \hat{\theta} \), particularly relevant for the estimation of counterfactual outcome distributions based on a grid \( \mathcal{U} \subseteq \mathcal{Y} \). The following algorithm allows us to do so from some estimator \( \hat{C} \) for the asymptotic variance \( C \) of \( \theta \).

Let \( \hat{S} = \text{diag}(\hat{C}) \) and \( S = \text{diag}(C) \) collect the diagonal elements of these matrices.

**Algorithm 3** (Simultaneous confidence band). Given \( C \),

1. Calculate \( \hat{\Sigma} = \hat{S}^{-1/2} \hat{C} \hat{S}^{-1/2} \)
2. Sample \( Q \overset{i.i.d.}{\sim} \mathcal{N}(0, \hat{\Sigma}) \) and compute the value \( c \) as the \( (1 - \alpha) \)-quantile of sampled \( |Q|_\infty \)
3. Form the confidence band \( [t_j, u_j] = \left[ \hat{\theta}_j - c \sqrt{\frac{C_{jj}}{n}}, \hat{\theta}_j + c \sqrt{\frac{C_{jj}}{n}} \right] \) where \( \hat{C}_{jj} \) is the diagonal entry of \( \hat{C} \) corresponding to \( j \)-th element \( \hat{\theta}_j \) of \( \theta \).

6 Consistency and Asymptotic Normality

We adapt the assumptions of [29] to our setting. First, we place weak assumptions on the dictionary \( b \), propensity score \( \pi_0 \), conditional variance \( \text{var}(\mathbb{V}|Z, X) \), and Jacobian \( J \).

**Assumption 2** (Bounded dictionary). There exists \( C > 0 \) s.t. \( \max_j |b_j(Z, X)| \leq C \) almost surely.

Alternatively, it is possible to allow the bound on the dictionary to be a sequence \( B_n^b \) that increases in \( n \); the core analysis remains the same with additional notation.

**Assumption 3** (Regularity). Assume

1. \( \pi_0(x) \in (\bar{c}, 1 - \bar{c}) \) for some \( \bar{c} > 0 \)
2. \( \text{var}(\mathbb{V}|Z, X) \) is bounded
3. Jacobian \( J = \mathbb{E} \left[ \frac{\partial \psi(W, \gamma_0, \alpha, \theta)}{\partial y} \bigg|_{\theta = \theta_0} \right] \) is nonsingular

6.1 Stage 1

Next we analyze our stage 1 estimators \( (\hat{\gamma}, \hat{\alpha}) \) of the nuisance parameters \( (\gamma_0, \alpha_0) \). We articulate assumptions required for convergence of \( \hat{\alpha} \) under two regimes: the regime in which \( \alpha_0 \) is dense and the regime in which \( \alpha_0 \) is sparse.

**Assumption 4** (Dense Riesz representer). Assume there exist some \( \rho_n \in \mathbb{R}^p \) and \( \alpha_0 \)

\[
|\rho_n|_1 \leq C \quad \text{and} \quad \|\alpha_0 - b^T \rho_n\|^2 = O\left(\sqrt{\frac{\ln p}{n}}\right)
\]

Assumption 4 is a statement about the quality of approximation of \( \alpha_0 \) by dictionary \( b \). It is satisfied if, for example, \( \alpha_0 \) is a linear combination of \( b \). It is possible to allow the bound on \( |\rho_n|_1 \) to be a sequence \( B_n \) that increases in \( n \); the core analysis remains the same with additional notation.

**Assumption 5** (Sparse Riesz representer). Assume

1. there exist \( C > 1, \xi > 0 \) s.t. for all \( \bar{s} \leq C \left( \frac{\ln p}{n} \right)^{-\frac{\xi}{1 + \xi}} \), there exists some \( \rho \in \mathbb{R}^p \) with \( |\rho|_1 \leq C \) and \( \bar{s} \) nonzero elements s.t. \( \|\alpha_0 - b^T \rho\|^2 \leq C(\bar{s})^{-\xi} \)
2. $G = \mathbb{E}[b(Z, X)b(Z, X)']$ is nonsingular with largest eigenvalue uniformly bounded in $n$

3. there exists $k > 3$ s.t. for $\rho \in \{\rho_L, \rho\}$

$$RE(k) = \inf_{\delta' G \delta \geq 0, \sum_{j \in \mathcal{J}_p} |\delta_j| \leq k} \frac{\delta' G \delta}{\sum_{j \in \mathcal{J}_p} \delta_j^2} > 0$$

where $\mathcal{J}_p = \text{support}(\rho)$

4. $\ln p = O(\ln n)$

Assumption 5 is a statement about the quality of approximation of $\alpha_0$ by a subset of dictionary $b$. It is satisfied if, for example, $\alpha_0$ is sparse or approximately sparse [29]. $RE$ is the population version of the restricted eigenvalue condition of [20]. It is possible to allow the bound on $|\hat{\rho}|_1$ to be a sequence $B_n$ that increases in $n$; the core analysis remains the same with additional notation.

We quote stage 1 convergence guarantees for the estimator $\hat{\alpha}$ in Algorithm 2 from [29]. We obtain a slow rate for dense $\alpha_0$ and a fast rate for sparse $\alpha_0$. In both cases, we require the data-driven regularization parameter $\lambda_n$ to approach 0 slightly slower than $\sqrt{\ln p / n}$.

**Assumption 6** (Regularization). $\lambda_n = a_n \sqrt{\ln p / n}$ for some $a_n \to \infty$

For example, one could set $a_n = \ln(\ln(n))$ [21]. In Appendix A.5, we provide and justify an iterative tuning procedure to determine data-driven regularization parameter $\lambda_n$.

**Theorem 2** (Dense Riesz representer rate). Under Assumptions 1, 2, 4, and 6,

$$\|\hat{\alpha} - \alpha_0\|^2 = O_p \left(a_n \sqrt{\ln p / n}\right), \quad |\hat{\rho}|_1 = O_p(1)$$

**Theorem 3** (Sparse Riesz representer rate). Under Assumptions 1, 2, 5, and 6,

$$\|\hat{\alpha} - \alpha_0\|^2 = O_p \left(a_n^2 \left(\frac{\ln p / n}{\sqrt{\ln p / n}}\right)^{-\frac{1}{2+2\xi}}\right), \quad |\hat{\rho}|_1 = O_p(1)$$

Whereas Theorem 2 does not require an explicit sparsity condition, Theorem 3 does. When $\xi > \frac{1}{2}$, the rate in Theorem 3 is faster than the rate in Theorem 2 for $a_n$ growing slowly enough. Interpreting the rate in Theorem 3, $n^{-\frac{1}{2+2\xi}}$ is the well-known rate of convergence if the identity of the nonzero components of $\hat{\rho}$ were known. The fact that their identity is unknown introduces a cost of $(\ln p)^{-\frac{1}{2+2\xi}}$. The cost $a_n^2$ can be made arbitrarily small.

We place a rate assumption on black-box estimator $\hat{\gamma}$. It is a weak condition that allows $\hat{\gamma}$ to converge at a rate slower than $n^{-\frac{1}{2}}$. Importantly, it allows the analyst a broad variety of choices of ML algorithms to estimate $\gamma_0$. In our empirical application in Section 7.2, we choose Lasso and neural networks.

**Assumption 7** (Regression rate). $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-d_\gamma})$ where

1. in the dense Riesz representer regime, $d_\gamma \in \left(\frac{1}{2}, \frac{1}{2}\right)$

2. in the sparse Riesz representer regime, $d_\gamma \in \left(\frac{1}{2} - \frac{\xi}{1+2\xi}, \frac{1}{2}\right)$

These regime-specific lower bounds on $d_\gamma$ are sufficient conditions for the DML product condition.

**Corollary 1.** Under Assumptions 1, 2, either 4 or 5, 6, and 7, $\|\hat{\alpha} - \alpha_0\| \cdot \|\hat{\gamma} - \gamma_0\| = o_p(n^{-\frac{1}{2}})$

The product condition in Corollary 1 formalizes the trade-off in estimation error permitted in learning the stage 1 nuisance parameters $(\gamma_0, \alpha_0)$. In particular, faster convergence of $\hat{\alpha}$ permits slower convergence of $\hat{\gamma}$. Prior information about $\alpha_0$ used to estimate $\hat{\alpha}$, encoded by sparsity or perhaps by additional moment restrictions, can be helpful in this way. We will appeal to this product condition while proving learning guarantees for stage 2 causal parameter $\theta_0$. (Faster convergence of $\hat{\alpha}$ does not imply faster convergence of $\hat{\theta}$, which already occurs at the parametric rate. Nor does it imply efficiency gains in the asymptotic variance of $\hat{\theta}$, which is already the semi-parametric lower bound.)
6.2 Stage 2

We now present the main theorem of this paper. We prove our Auto-DML estimator for complier parameters is consistent, asymptotically normal, and semi-parametrically efficient, appealing to the theory in [24] to generalize the main result in [29].

**Assumption 8.** $\theta_0, \tilde{\theta} \in \Theta$, a compact parameter space

**Theorem 4** (Auto-DML consistency and asymptotic normality). Suppose Assumptions 1, 2, 3, either 4 or 5, 6, 7, and 8 hold. Then $\tilde{\theta} \overset{p}{\to} \theta_0$, $\sqrt{n}(\hat{\theta} - \theta_0) \overset{d}{\to} N(0, \Sigma)$, and $\hat{C} \overset{p}{\to} C$

\[
\begin{align*}
\mathbf{J} &= \mathbb{E} \left[ \frac{\partial \psi_0(\mathbf{W})}{\partial \theta} \right], \quad \mathbf{J} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \frac{\partial \psi_0(\mathbf{W})}{\partial \theta}, \quad \mathbf{\Omega} = \mathbb{E}[\psi_0(\mathbf{W})\psi_0(\mathbf{W})'], \quad \mathbf{\Omega} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \psi_0(\mathbf{W})\psi_0(\mathbf{W})' \\
\mathbf{C} &= \mathbf{J}^{-1}\mathbf{\Omega}^{-1}, \quad \hat{\mathbf{C}} = \hat{\mathbf{J}}^{-1}\hat{\mathbf{\Omega}}^{-1}, \quad \psi_0(\mathbf{w}) = \psi(\mathbf{w}, \gamma_0, \alpha_0, \theta_0), \quad \hat{\psi}_i(\theta) = \psi(\mathbf{W}_i, \gamma_{-i}, \alpha_{-i}, \theta)
\end{align*}
\]

It follows that $\hat{\theta}$ is semi-parametrically efficient [24]. See [5] for a discussion of semi-parametric efficiency.

6.3 Simultaneous Confidence Band

Finally, we prove the validity of the bootstrap procedure presented in Algorithm 3 for simultaneous inference on the counterfactual distributions $\theta_0$. Appealing to the theory in [23], we assume the following sufficient conditions.

**Assumption 9** (Grid size). There are positive constants $(c', C')$ such that $(\log((dn)))^{7}/n \leq C' n^{-c'}$ where $d = \text{dim}(\mathcal{U})$ is the dimension of the grid on which the counterfactual distributions $\theta_0$ are evaluated

**Assumption 10** (Tail bounds). $c'$ and $C'$ also satisfy

1. $\mathbb{P} \left( \sqrt{\log d} \Delta_1 > C' n^{-c'} \right) < C' n^{-c'}$

2. $\mathbb{P} \left( ( \log((dn)))^{2} \Delta_2 > C' n^{-c'} \right) < C' n^{-c'}$

where

\[
\begin{align*}
\Delta_1 &= \left\| \frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \left( \hat{S}^{-1/2} \hat{J}^{-1} \hat{\psi}_i(\theta_0) - S^{-1/2} J^{-1} \psi_0(\mathbf{W}_i) \right) \right\|_{\infty} \\
\Delta_2 &= \left\| \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \left( \hat{S}^{-1/2} \hat{J}^{-1} \hat{\psi}_i(\theta_0) - S^{-1/2} J^{-1} \psi_0(\mathbf{W}_i) \right) \right\|_{\infty}^2
\end{align*}
\]

and the square is taken element-wise.

See [16] for examples of primitive conditions under which Assumption 10 holds.

**Theorem 5** (Simultaneous confidence band). Under the assumptions of Theorem 4 as well as Assumptions 9 and 10, the confidence band in Algorithm 3 jointly covers the true counterfactual distributions $\theta_0$ at all grid points $y \in \mathcal{U}$ with probability approaching the nominal level, i.e.

$\lim_{n \to \infty} \mathbb{P}((\theta_0)_y \in [l_j, u_j] \forall j) = 1 - \alpha$.

7 Experiments

We compare the performance of our Auto-DML estimator with original DML [22] and κ-weighting [2] in simulations. We focus on counterfactual distributions as our choice of complier parameter $\theta_0$. We then apply our Auto-DML estimator to real-world data to estimate the counterfactual distributions of employee net financial assets with and without 401(k) participation.
7.1 Simulation

We consider a simulation design detailed in Appendix A.7. Each simulation consists of \( n = 1000 \) observations, and we use a dictionary \( b \) with dimension \( p = 10 \). We apply Auto-DML, DML, and \( \kappa \)-weighting to estimate a counterfactual outcome distribution at each value in the grid \( \mathcal{U} \) specified on the horizontal axis of Figure 1.

For each algorithm, we implement 500 simulations and visualize the median as well as the 10\% and 90\% quantiles for each value in the grid \( \mathcal{U} \). Figure 1 summarizes results: Auto-DML outperforms DML by a large margin due to numerical stability. Auto-DML modestly outperforms \( \kappa \)-weighting, perhaps because the former uses regularized ML to estimate nuisance parameters while the latter does not.

\[ \beta_0^\text{truth} = \Pr(Y^{(0)} \leq y \mid D^{(1)} > D^{(0)}) \]

\[ \delta_0^\text{truth} = \Pr(Y^{(1)} \leq y \mid D^{(1)} > D^{(0)}) \]

Figure 1: Counterfactual distribution simulation

In Appendix A.7.4, we consider whether DML and \( \kappa \)-weighting can be improved by addressing the numerical instability that results from inverting \( \hat{\pi} \). Specifically, we try trimming and censoring. Trimming means excluding observations for which \( \hat{\pi} \) is extreme [17]. Censoring means imposing bounds on \( \hat{\pi} \) for such observations. We find that trimming improves DML. There is no noticeable improvement for \( \kappa \)-weighting with either trimming or censoring. Auto-DML without trimming or censoring outperforms DML and \( \kappa \)-weighting even with trimming or censoring. This property is convenient, since ad hoc trimming and censoring have limited theoretical justification [31]. In Appendix A.4, we formalize properties of Auto-DML that explain why it does not require trimming or censoring.

7.2 Effect of 401(k) on Assets

Next, we use Auto-DML to investigate the effect of 401(k) participation on the distribution of net financial assets, using 401(k) eligibility as the instrument. We follow the identification strategy of [46] and [47]. The authors assume that when 401(k) was introduced, workers ignored whether a given job offered 401(k) and instead made employment decisions based on income and other observable job characteristics; after conditioning on income and job characteristics, 401(k) eligibility was as good as randomly assigned at the time. The independence and exclusion conditions of Assumption 1 are thus satisfied. Since ineligibility implies no participation, the monotonicity condition of Assumption 1 is satisfied by construction.

We use data from the 1991 US Survey of Income and Program Participation, studied in [2, 26, 27, 44, 17]. We use sample selection and variable construction as in [22]. The outcome \( Y \) is net financial assets (NFA) defined as the sum of IRA balances, 401(k) balances, checking accounts, US saving bonds, other interest-earning accounts, stocks, mutual funds, and other interest-earning assets minus non-mortgage debt. The treatment \( D \) is participation in a 401(k) plan. The instrument \( Z \) is eligibility to enroll in a 401(k) plan. The covariates \( X \) are age, income, years of education, family size, marital status, two-earner status, benefit pension status, IRA participation, and home-ownership.

We use data from the 1991 US Survey of Income and Program Participation, studied in [2, 26, 27, 44, 17]. We use sample selection and variable construction as in [22]. The outcome \( Y \) is net financial assets (NFA) defined as the sum of IRA balances, 401(k) balances, checking accounts, US saving bonds, other interest-earning accounts, stocks, mutual funds, and other interest-earning assets minus non-mortgage debt. The treatment \( D \) is participation in a 401(k) plan. The instrument \( Z \) is eligibility to enroll in a 401(k) plan. The covariates \( X \) are age, income, years of education, family size, marital status, two-earner status, benefit pension status, IRA participation, and home-ownership.

The data include \( n = 9875 \) observations after we numerically impose the overlap condition of Assumption 1, following [32] and [22]. In Appendix A.8, we show that our results remain unchanged without this pre-processing step. We follow [17] in the choice of grid points \( \mathcal{U} \) and [22] in the choice
of dictionary \( b \). We take \( \mathcal{U} \) as the 5\(^{th}\) through 95\(^{th}\) percentiles of \( Y \), a total of 91 different values of \( y \). We consider a high-dimensional dictionary with \( p = 277 \). See Appendix A.8 for further details on the dictionary and Auto-DML implementation.

Figure 2 visualizes point estimates and simultaneous 95% confidence bands. We find that 401(k) participation significantly shifts out the distribution of NFA, consistent with results reported in [17]. For compliers, the distribution of potential NFA under participation first order stochastic dominates the distribution of potential NFA under non-participation. Moreover, the Auto-DML algorithm is robust in the high dimensional setting, yielding similar results using Lasso or a neural network to estimate \( \hat{\gamma} \). Our counterfactual distribution estimates are non-decreasing. Generically, this property may not hold in the finite sample [38]; in such case, monotone rearrangement is possible [25].

![Estimated CDF of \( Y(1) \) and \( Y(0) \)](image)

(a) \( \hat{\gamma} \) estimated by Lasso

(b) \( \hat{\gamma} \) estimated by neural network

Figure 2: Effect of 401(k) on net financial assets for compliers

8 Conclusion

We extend de-biased machine learning with automatic bias correction to the task of learning causal parameters from confounded, high-dimensional data. The procedure is easily implemented and semi-parametrically efficient. As a contribution to the instrumental variable literature, we reinterpret the \( \kappa \)-weight as the Riesz representer in the problem of learning complier parameters and we allow for high-dimensional covariates. As a contribution to the de-biased machine learning literature, we generalize the theory of Auto-DML and provide a framework for estimating causal parameters identified by instrumental variables. In simulations, Auto-DML outperforms DML and \( \kappa \)-weighting and eliminates the ad hoc step of trimming or censoring, suggesting Auto-DML may be an effective paradigm in high-dimensional causal inference.

Acknowledgements

We are grateful to Alberto Abadie, Anish Agarwal, Isaiah Andrews, Joshua Angrist, Abhijit Banerjee, Victor Chernozhukov, Joseph Kibe, Anna Mikusheva, Whitney Newey, Garima Sharma, and Suhas Vijaykumar for helpful comments.
References


A Appendix

Contents

A.1 Notation Glossary ................................................. 15
A.2 Forbidden Regression and Parametric Estimation ................. 16
  A.2.1 2SLS ...................................................... 17
  A.2.2 Forbidden Regression ........................................ 17
  A.2.3 Auto-DML ................................................. 17
A.3 Identification ..................................................... 18
A.4 Balancing Property ............................................... 20
A.5 Tuning .......................................................... 21
A.6 Consistency and Asymptotic Normality ......................... 22
  A.6.1 Lemmas ..................................................... 22
  A.6.2 Stage 1 ..................................................... 23
  A.6.3 Stage 2 ..................................................... 23
  A.6.4 Simultaneous Confidence Band ................................ 28
A.7 Simulation ........................................................ 29
  A.7.1 Design ...................................................... 29
  A.7.2 Algorithms .................................................. 29
  A.7.3 Sensitivity .................................................. 30
  A.7.4 Trimming and Censoring .................................... 31
A.8 Effect of 401(k) on Assets ....................................... 32
  A.8.1 Dictionaries ................................................ 32
  A.8.2 Algorithm .................................................. 32
  A.8.3 Results ..................................................... 32

A.1 Notation Glossary

Let $W = (Y, D, Z, X)'$ concatenate the random variables. $Y \in \mathcal{Y} \subset \mathbb{R}$ is the continuous outcome, $D \in \{0, 1\}$ is the binary treatment, $Z \in \{0, 1\}$ is the binary instrumental variable, and $X \in \mathcal{X} \subset \mathbb{R}^{\text{dim}(X)}$ is the covariate. We observe $n$ i.i.d. observations $\{W_i\}_{i=1}^n$. Where possible, we suppress index $i$ to lighten notation.

Following the notation of [9], we denote by $Y(z,d)$ the potential outcome under the intervention $Z = z$ and $D = d$. Due to Assumption 1, we can simplify notation: $Y^{(d)} = Y^{(1,d)} = Y^{(0,d)}$. We denote by $D(z)$ the potential treatment under the intervention $Z = z$. Compliers are the subpopulation for whom $D(1) > D(0)$.

Using the notation of [29], we denote the regression of random vector $V = (V_1, \ldots, V_J)'$ conditional on $(Z, X)$ as

$$
\gamma_0(Z, X) = \left[ \begin{array}{c} \gamma^V_0(Z, X) \\ \vdots \\ \gamma^V_J(Z, X) \end{array} \right] = \mathbb{E}[V | Z, X]
$$

where $\gamma^V_0(Z, X) = \mathbb{E}[V_j | Z, X]$. The random vector $V$ is observable and depends on the complier parameter of interest; we specify its components for LATE, complier characteristics, and counterfactual outcome distributions in Theorem 1.
We denote the propensity score \( \pi_0(x) = \mathbb{P}(Z = 1|X = x) \). We denote the classic Horvitz-Thompson weight with
\[
\alpha_0(z, x) = \frac{z}{\pi_0(x)} - \frac{1 - z}{1 - \pi_0(x)} = \frac{z - \pi_0(x)}{\pi_0(x)(1 - \pi_0(x))}
\]
We denote by \(|\cdot|_q\) the \( \ell_q \) norm of a vector. We denote by \( \|\cdot\| \) the \( \ell_2 \) norm of a random variable \( V_j \), i.e. \( \|V_j\| = \sqrt{\mathbb{E}[V_j^2]} \). For random vector \( V = (V_1, ..., V_J)' \), we slightly abuse notation by writing
\[
\|V\| = \begin{bmatrix} \|V_1\| \\ \|V_2\| \\ \vdots \\ \|V_J\| \end{bmatrix}, \quad \mathbb{E}[V]^2 = \|V\|^2 = \begin{bmatrix} \|V_1\|^2 \\ \|V_2\|^2 \\ \vdots \\ \|V_J\|^2 \end{bmatrix}
\]
Likewise, we write the element-wise absolute value as
\[
|V| = \begin{bmatrix} |V_1| \\ |V_2| \\ \vdots \\ |V_J| \end{bmatrix}
\]
Consider a causal parameter \( \theta_0 \in \Theta \), where \( \Theta \) is some compact parameter space. It is implicitly defined by moment function \( m \).
\[
\mathbb{E}[m(W, \gamma_0, \theta)] = 0 \text{ iff } \theta = \theta_0
\]
We denote the doubly robust moment function for \( \theta_0 \) by
\[
\mathbb{E}[\psi(W, \gamma_0, \alpha_0, \theta)] = 0 \text{ iff } \theta = \theta_0
\]
\[
\psi(w, \gamma, \alpha, \theta) = m(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta)
\]
\( \phi \) is called the de-biasing term.

In sample splitting, we partition the sample into \( L \) folds \( \{I_\ell\}_{\ell=1:L} \), each with \( n_\ell = n/L \) observations. We denote by \( (\hat{\gamma}_\ell, \hat{\alpha}_\ell) \) the estimates from observations not in \( I_\ell \). We denote by \( b(z, x) \) a \( p \)-dimensional dictionary of basis functions.

The population regularized Riesz representer (RRR) parameter \( \rho_L \) is the solution to
\[
\rho_L = \arg\min_{\rho} \mathbb{E}[\alpha_0(Z, X) - \rho' b(Z, X)]^2 + 2\lambda_L |\rho|_1
\]
where \( \lambda_L = \sqrt{n \ln p} \) is the theoretical regularization parameter.

The sample RRR parameter \( \hat{\rho}_{-\ell} \) estimated from \( I_{-\ell} \) is the solution to
\[
\hat{\rho}_{-\ell} = \arg\min_{\rho} \hat{\rho}' \hat{G}_{-\ell} \rho - 2\rho' \hat{M}_{-\ell} + 2\lambda_n |\rho|_1
\]
where
\[
\hat{G}_{-\ell} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b(Z_i, X_i) b(Z_i, X_i)', \quad \hat{M}_{-\ell} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b(1, X_i) - b(0, X_i)
\]
and \( \lambda_n \) is a data-driven regularization parameter. We denote \( \hat{\alpha}_{-\ell}(z, x) = b(z, x)' \hat{\rho}_{-\ell} \).

In estimating a simultaneous confidence band, we denote the \( d \)-dimensional grid \( \mathcal{U} \subset \mathcal{Y} \). \( \hat{C} \) is the estimator of the asymptotic variance \( C \) of \( \theta \). Let \( S = \text{diag}(\hat{C}) \) and \( S = \text{diag}(C) \) collect the diagonal elements of these matrices.

The remaining symbols are concisely defined in the assumptions and theorems of Section 6

A.2 Forbidden Regression and Parametric Estimation

We compare our semi-parametric approach with standard parametric approaches to estimation using instrumental variables. Specifically, we relate Auto-DML to the so-called forbidden regression discussed in [11, Section 4.6.1]. We focus our discussion on LATE as the target parameter \( \theta_0 \). We show that specifications which may be inconsistent in the parametric framework are indeed consistent in the semi-parametric framework. By considering an Auto-DML approach to LATE, we may “taste” from the forbidden regressions.
A.2.1 2SLS

To simplify the exposition, we consider a model without covariates $X$. The linear (parametric) specification of [11, Section 4.6.1] is

$$D = \pi_0 Z + \epsilon_1, \quad \mathbb{E}[Z\epsilon_1] = 0, \quad \mathbb{E}[\epsilon_1] = 0$$

$$Y = \theta_0 D + \epsilon_2, \quad \mathbb{E}[Z\epsilon_2] = 0, \quad \mathbb{E}[\epsilon_2] = 0$$

Economists refer to the former equation as the first-stage and the latter as the second-stage. Substituting the former into the latter yields an equation called the reduced-form. Note that the first-stage and second-stage of 2SLS are different than stage 1 and stage 2 of DML; we maintain the semantic distinction throughout.

$\theta_0$ is LATE. By standard projection geometry

$$\bar{D} := P_Z D = \pi_0 Z$$

$$\bar{Y} := P_Z Y = P_Z(\theta_0 D) = \theta_0 \bar{D}$$

where $P_Z$ denotes the projection onto instrument $Z$. It follows that $\theta_0$ can be obtained by projecting $\bar{Y}$ on $\bar{D}$. Since $Y = \bar{Y} + Y^\perp$ where $\bar{Y}, \bar{D} \in \text{span}\{Z\}$ and $Y^\perp \in \text{span}\{Z\}^\perp$, $\theta_0$ can also be obtained by projecting $Y$ on $D$. This procedure is the widely-used two-stage least squares algorithm (2SLS).

2SLS is robust to functional form mis-specification in two senses. First, if the model is saturated—i.e. if $(D, Z, X)$ have discrete support—then the linear specification is w.l.o.g. Second, even if the first-stage relationship is actually nonlinear, projection geometry ensures consistent estimation of a well-defined best linear approximation of LATE. (In the absence of covariates $X$ or under the assumption of homogeneous treatment effects, the estimand of 2SLS is LATE. In the presence of covariates and heterogeneous treatment effects, the estimand of 2SLS is not LATE but rather a variance-reweighting of the expanded Wald formula [11, Theorem 4.5.1]. By contrast, even in the presence of heterogeneous treatment effects, the estimand of Auto-DML is LATE.)

A.2.2 Forbidden Regression

A forbidden regression is a variant of 2SLS in which the analyst attempts to guess the functional form of the nonlinear first-stage relationship. For example, for binary treatment $D$ one may consider the probit specification

$$D = \Phi(\hat{\pi}_0 Z) + \hat{\epsilon}_1, \quad \mathbb{E}[\hat{\epsilon}_1 | Z] = 0$$

$$Y = \theta_0 D + \hat{\epsilon}_2, \quad \mathbb{E}[\hat{\epsilon}_2 | Z] = 0$$

Note that the moment restrictions are stronger than in 2SLS: conditional moment restrictions implying correct specification of the conditional means, rather than unconditional moment restrictions implying orthogonal noise. Substituting the former into the latter yields a reduced-form.

Under correct specification, $\theta_0$ may be consistently estimated using the first-stage and reduced-form, similar to 2SLS. However, this approach is not robust to functional form mis-specification. Under mis-specification, projection geometry no longer applies and the resulting estimator for $\theta_0$ is inconsistent.

The concept of a forbidden regression highlights the drawbacks of using a parametric approach to LATE, i.e. a finite-dimensional nuisance parameter ($\pi_0$ or $\hat{\pi}_0$) and finite-dimensional causal parameter $\theta_0$. Any nonlinear parametric approach is highly sensitive to mis-specification. Indeed, to guarantee consistency under mis-specification, the analyst must revert back to the linear specification and projection geometry of 2SLS. Forfeiting data fit entails forfeiting statistical precision. Another important drawback is that 2SLS consistently estimates an approximation to LATE in a highly restricted class.

A.2.3 Auto-DML

By contrast, the Auto-DML approach is semi-parametric, with infinite-dimensional nuisance parameters ($\gamma_0, \alpha_0$) and finite-dimensional causal parameter $\theta_0$. In Theorem 1, we show that the Auto-DML estimator for LATE accommodates black-box machine learning of

$$\gamma_0(Z, X) := \mathbb{E}[V | Z, X] = \mathbb{E} \left[ \begin{array}{c} Y \\ D \end{array} \right] | Z, X$$
These conditional expectations correspond to the reduced-form and first-stage estimating equations of 2SLS, respectively. Equivalently, they correspond to the numerator and denominator of the expanded Wald formula in Section 4.2. In particular, the Auto-DML estimator accommodates the forbidden regression: an analyst may specify probit first-stage, linear second-stage, and the resulting reduced-form.

By using a semi-parametric approach to LATE, any nonlinear approach to estimating the first-stage and the reduced-form is permitted. Mis-specification is no longer a concern; rather, the approximation of $\gamma_0$ must be of sufficiently high quality as stated in Hypothesis 7. By Theorem 4, such an estimator is consistent, asymptotically normal, and semi-parametrically efficient. Auto-DML consistently estimates an approximation to LATE in a much broader class, namely the class in which $E[\gamma_0(Z, X)]^2 < \infty$.

A.3 Identification

We review the derivation of the classic Horvitz-Thompson weight, relate Auto-DML to $\kappa$-weighting, and prove a general identification result. We then specialize this result to LATE, complier characteristics, and counterfactual outcome distributions.

Proposition 1. $\alpha_0(z, x)$ is the Riesz representer to the continuous linear functional $\gamma \mapsto E[\gamma(1, X) - \gamma(0, X)]$, i.e.

$$E[\gamma(1, X) - \gamma(0, X)] = E[\alpha_0(Z, X)\gamma(Z, X)], \quad \forall \gamma \text{ s.t. } E[\gamma(Z, X)]^2 < \infty$$

Proof. Observe that

$$E \left[ \frac{Z}{\pi_0(X)} \gamma(Z, X) \bigg| X = x \right] = E \left[ \frac{Z}{\pi_0(X)} \gamma(Z, X) \bigg| Z = 1, X \right] \pi_0(X) = \gamma(1, X)$$

and likewise

$$E \left[ \frac{1 - Z}{1 - \pi_0(X)} \gamma(Z, X) \bigg| X = x \right] = \gamma(0, X)$$

In summary, we can write

$$E[\gamma(1, X) - \gamma(0, X)] = \int \{ \gamma(1, x) - \gamma(0, x) \} dP(x)$$

$$= \int \left\{ E \left[ \frac{Z}{\pi_0(X)} \gamma(Z, X) \bigg| X = x \right] - E \left[ \frac{1 - Z}{1 - \pi_0(X)} \gamma(Z, X) \bigg| X = x \right] \right\} dP(x)$$

$$= E \left[ \frac{Z}{\pi_0(X)} \gamma(Z, X) \bigg| X = x \right] - E \left[ \frac{1 - Z}{1 - \pi_0(X)} \gamma(Z, X) \bigg| X = x \right]$$

Definition 2. Define

$$\kappa^{(0)}(w) = (1 - d) \frac{(1 - z) - (1 - \pi_0(x))}{[1 - \pi_0(x)]\pi_0(x)}$$

$$\kappa^{(1)}(w) = d \frac{z - \pi_0(x)}{[1 - \pi_0(x)]\pi_0(x)}$$

$$\kappa(w) = [1 - \pi_0(x)]\kappa^{(0)}(w) + \pi_0(x)\kappa^{(1)}(w)$$

These are the $\kappa$-weights introduced in [2].
Proposition 2. The $\kappa$-weights can be rewritten as
\[
\kappa^{(0)}(w) = \alpha_0(z,x)[d-1] \\
\kappa^{(1)}(w) = \alpha_0(z,x)d \\
\kappa(w) = \alpha_0(z,x)[d-1 + \pi_0(x)]
\]

Proof. $\alpha_0(z,x) = \frac{z}{\pi_0(x)} - \frac{1-z}{1-\pi_0(x)} = \frac{z-\pi_0(x)}{\pi_0(x)[1-\pi_0(x)]}$

Theorem 6. Suppose Assumption 1 holds. Let $g(y,d,x,\theta)$ be a measurable, real-valued function s.t. $E[g(Y, D, X, \theta)]^2 < \infty$ for all $\theta \in \Theta$.

1. If $\theta_0$ is defined by the moment condition $E[g(Y^{(0)}, X, \theta_0)]|D^{(1)} > D^{(0)}] = 0$, let $v(w, \theta) = |d - 1| g(y, x, \theta)$
2. If $\theta_0$ is defined by the moment condition $E[g(Y^{(1)}, X, \theta_0)]|D^{(1)} > D^{(0)}] = 0$, let $v(w, \theta) = d \cdot g(y, x, \theta)$
3. If $\theta_0$ is defined by the moment condition $E[g(Y, D, X, \theta_0)]|D^{(1)} > D^{(0)}] = 0$, let $v(w, \theta) = [d - 1 + \pi_0(x)] g(y, d, x, \theta)$

Then the doubly robust moment function for $\theta_0$ is of the form
\[
\psi(w, \gamma, \alpha, \theta) = \tilde{m}(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta) \\
\tilde{m}(w, \gamma, \theta) = \gamma(1, x, \theta) - \gamma(0, x, \theta) \\
\phi(w, \gamma, \alpha, \theta) = \alpha(z, x)[v(w, \theta) - \gamma(z, x)]
\]

where $\gamma_0(z, x, \theta) := E[v(W, \theta)|z, x]$

Proof. Consider the first case. Under Assumption 1, we can appeal to [2, Theorem 3.1].

\[
0 = E[g(Y^{(0)}, X, \theta_0)]|D^{(1)} > D^{(0)}] = \frac{E[k^{(0)}(W) g(Y^{(0)}, X, \theta_0)]}{P(D^{(1)} > D^{(0)})}
\]

Hence
\[
0 = E[k^{(0)}(W) g(Y^{(0)}, X, \theta_0)] \\
= E[\alpha_0(Z, X)(D - 1) g(Y^{(0)}, X, \theta_0)] \\
= E[\alpha_0(Z, X) v(W, \theta)] \\
= E[\alpha_0(Z, X) \gamma_0(Z, X, \theta_0)] \\
= E[\gamma_0(1, X, \theta_0) - \gamma_0(0, X, \theta_0)]
\]

appealing to Assumption 1, Proposition 2, and the fact that $\alpha_0$ is the Riesz representer for $\gamma \mapsto E[\gamma(1, X, \theta_0) - \gamma(0, X, \theta_0)]$. Likewise for the second and third cases.

Proof of Theorem 1. Suppose we can decompose $v(w, \theta) = h(w, \theta) + a(\theta)$ for some function $a(\cdot)$ that does not depend on data. Then we can replace $v(w, \theta)$ with $h(w, \theta)$ without changing $\hat{m}$ and $\phi$. This is because $\gamma^v(z, x, \theta) = \gamma^h(z, x, \theta) + a(\theta)$ and hence $v(w, \theta) - \gamma^v(z, x) = h(w, \theta) - \gamma^h(z, x)$. Whenever we use this reasoning, we write $v(w, \theta) \propto h(w, \theta)$.

1. For LATE we can write $\theta_0 = \delta_0 - \beta_0$, where $\delta_0$ is defined by the moment condition $E[Y^{(1)} - \delta_0]|D^{(1)} > D^{(0)}] = 0$ and $\beta_0$ is defined by the moment condition $E[Y^{(0)} - \beta_0]|D^{(1)} > D^{(0)}] = 0$. Applying Case 2 of Theorem 6 to $\delta_0$, we have $v(w, \delta) = d \cdot (y - \delta)$. Applying Case 1 of Theorem 6 to $\beta_0$, we have $v(w, \beta) = (d - 1) \cdot y - d \cdot \beta$. Writing $\delta = \delta - \beta$, the moment function for $\theta_0$ can thus be derived with $v(w, \theta) = v(w, \delta) - v(w, \beta) = y - d \cdot \theta$. Note that this expression decomposes into $V = (Y, D)'$ and $A(\theta) = [1 \quad -\theta]$ in Theorem 1.
2. For average complier characteristics, \( \theta_0 \) is defined by the moment condition \( \mathbb{E}[f(X) - \theta_0]^{(1)} > D^{(0)} = 0 \). Applying Case 2 of Theorem 6 setting \( g(Z^{(1)}, X, \theta_0) = f(X) - \theta_0 \), we have \( v(w, \theta) = d \cdot (f(x) - \theta_0) \). This expression decomposes into \( V = \{Df(X)^{(1)}, D^{(0)}\} \) and \( A(\theta) = \{1 - \theta\} \) in Theorem 1. 

3. For complier distribution of \( Y^{(0)} \), \( \beta^{(0)}_0 \) is defined by the moment condition \( \mathbb{E}[1_{Y^{(0)} \leq \bar{y}} - \beta^{(0)}_0]^{(1)} > D^{(0)} = 0 \). Applying Case 1 of Theorem 6 to \( \beta^{(0)}_0 \), we have \( v(w, \beta^{(0)}_0) = (d - 1) \cdot (1_{y \leq \bar{y}} - \beta^{(0)}_0) \propto (d - 1) \cdot 1_{y \leq \bar{y}} - d \cdot \beta^{(0)}_0 \). For complier distribution of \( Y^{(1)} \), \( \delta^{(0)}_0 \) is defined by the moment condition \( \mathbb{E}[1_{Y^{(1)} \leq \bar{y}} - \delta^{(0)}_0]^{(1)} > D^{(0)} = 0 \). Applying Case 2 of Theorem 6 to \( \delta_0 \), we have \( v(w, \delta^{(0)}_0) = d \cdot (1_{y \leq \bar{y}} - \delta^{(0)}_0) \). Concatenating \( v(w, \beta^{(0)}_0) \) and \( v(w, \delta^{(0)}_0) \), we arrive at the decomposition in Theorem 1.

\[ \square \]

A.4 Balancing Property

Next, we discuss how Algorithm 2 automatically attenuates the influence of outliers, which is a central issue in applied statistical research. We demonstrate that the finite sample balancing property of Auto-DML, shown by [29] for parameters of the full population, does in fact generalize to parameters of the complier subpopulation analyzed in the present work.

Recall that in Algorithm 1, the asymptotic influence \( \psi(W_i, \gamma_0, \alpha_0, \theta_0) \) of observation \( W_i \) in fold \( \ell \) is estimated by empirical influence \( \psi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \hat{\theta}) \), where the Riesz representer estimator \( \hat{\alpha}_\ell \) is calculated according to Algorithm 2.

In original DML, the propensity score \( \hat{\pi}_{\ell} \) is explicitly estimated to serve as a component of the Horvitz-Thompson weight

\[
\hat{\alpha}^\text{original}_{\ell}(Z, X_i) = \frac{Z_i}{\hat{\pi}_{\ell}(X_i)} - \frac{1 - Z_i}{1 - \hat{\pi}_{\ell}(X_i)}
\]

In the finite sample, \( \hat{\pi}_{\ell}(X_i) \) can be close to 0 or 1, causing the empirical influence in original DML \( \psi(W_i, \hat{\gamma}_\ell, \hat{\alpha}^\text{original}_{\ell}, \hat{\theta}) \) to diverge. This scenario would arise if, for example, there exists an imbalanced stratum \( x \): there is an outlier \( W_i \) with \( Z_i = 1 \) and \( x_i = x \) but no other observations \( W_j \) with \( Z_j = 0 \) and \( x_j \) close to \( x \).

This issue may cause an analyst to worry about the choice of covariates \( X \), or to introduce ad hoc trimming or censoring of propensity scores. It is a general concern in propensity score-based methods, including matching, \( \kappa \)-weighting, and original DML. Auto-DML automatically addresses outliers and finite-sample imbalance in three ways.

First, Auto-DML considers \( \alpha_0 \) rather than \( \pi_0 \) to be a nuisance parameter. \( \hat{\alpha}_\ell \) enters \( \psi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \hat{\theta}) \) additively, whereas \( \hat{\pi}_\ell \) enters \( \psi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \hat{\theta}) \) inversely.

Second, Auto-DML confines a finite sample guarantee of balance on average. Consider the choice of dictionary \( b \) and corresponding partition of parameter \( \rho \) to be

\[
b(z, X) = \begin{bmatrix} zq(x) \\ (1 - z)q(x) \end{bmatrix}, \quad \rho = \begin{bmatrix} \rho^{(z=1)} \\ \rho^{(z=0)} \end{bmatrix}
\]

**Proposition 3** (Balance).

\[
\left| \frac{1}{n - n_\ell} \sum_{i \in L_\ell} q(X_i) - \frac{1}{n - n_\ell} \sum_{i \in L_\ell} q(X_i) Z_i \cdot \hat{\omega}^{(z=1)}_{\ell, i} \right| \leq \lambda_n, \quad \hat{\omega}^{(z=1)}_{\ell, i} = q(X_i) \rho^{(z=1)}_{\ell}
\]

\[
\left| \frac{1}{n - n_\ell} \sum_{i \in L_\ell} q(X_i) - \frac{1}{n - n_\ell} \sum_{i \in L_\ell} q(X_i)(1 - Z_i) \cdot \hat{\omega}^{(z=0)}_{\ell, i} \right| \leq \lambda_n, \quad \hat{\omega}^{(z=0)}_{\ell, i} = q(X_i) \rho^{(z=0)}_{\ell}
\]

**Proof.** Immediate from the Lasso first order condition \( \square \)
Proposition 3 shows that the weights \( \{ \omega_{E,i}^{(z=1)}, \omega_{E,i}^{(z=0)} \} \) serve to approximately balance the overall sample average with the sample average of the group having \( Z = 1 \) and the sample average of the group having \( Z = 0 \), across each element of dictionary \( q \). The result is similar to the balancing conditions of [53] and [12]. Auto-DML automatically calculates these weights.

Third, Auto-DML is a Lasso-type estimator that delivers a sparse estimate \( \hat{\rho}_{-\ell} \). As such, it automatically determines which elements of dictionary \( q \) to use in the calculation of influence \( \psi(W, \gamma_{-\ell}, \alpha_{-\ell}, \theta) \). It discerns that, in the finite sample, some elements are best ignored, and those elements may be different for the group with \( Z = 1 \) and the group with \( Z = 0 \). If there exists an imbalanced stratum \( X = x \), it may not necessarily correspond to an imbalanced stratum \( q(x) = q' \). If \( q' \) is an imbalanced stratum, then Auto-DML can zero out specific components of \( q(x) \) where \( q' \) is imbalanced; geometrically, it can collapse those dimensions in the space of basis functions. In this sense, Auto-DML learns which transformations of covariates to match on.

### A.5 Tuning

Algorithm 2 takes as given the value of regularization parameter \( \lambda_n \). For practical use, we provide an iterative tuning procedure to empirically determine \( \lambda_n \). This is precisely the tuning procedure of [29], adapted from [28]. Due to its iterative nature, the tuning procedure is most clearly stated as a replacement for Algorithm 2.

Recall that the inputs to Algorithm 2 are observations in \( I_{-\ell} \), i.e. excluding fold \( \ell \). The analyst must also specify the \( p \)-dimensional dictionary \( b \). For notational convenience, we assume \( b \) includes the intercept in its first component: \( b_1(z, x) = 1 \). In this tuning procedure, the analyst must further specify a low-dimensional sub-dictionary \( b_{\text{low}} \) of \( b \). As in Algorithm 2, the output of the tuning procedure is \( \hat{\alpha}_{-\ell} \), an estimator of the Riesz representer trained only on observations in \( I_{-\ell} \).

The tuning procedure is as follows.

**Algorithm 4 (RRR with tuning).** For observations in \( I_{-\ell} \)

1. Initialize \( \hat{\rho}_{-\ell} \) using \( b_{\text{low}} \)

\[
\hat{G}_{-\ell}^{\text{low}} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b_{\text{low}}(Z_i, X_i) b_{\text{low}}(Z_i, X_i)'
\]

\[
\hat{M}_{-\ell}^{\text{low}} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b_{\text{low}}(1, X_i) - b_{\text{low}}(0, X_i)
\]

\[
\hat{\rho}_{-\ell} = \begin{bmatrix} \left( \hat{G}_{-\ell}^{\text{low}} \right)^{-1} \hat{M}_{-\ell}^{\text{low}} \\ 0 \end{bmatrix}
\]

2. Calculate moments

\[
\hat{G}_{-\ell} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b(Z_i, X_i) b(Z_i, X_i)'
\]

\[
\hat{M}_{-\ell} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b(1, X_i) - b(0, X_i)
\]

3. While \( \hat{\rho}_{-\ell} \) has not converged

   (a) Update normalization

   \[
   \hat{D}_{-\ell} = \sqrt{\frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} [b(Z_i, X_i) b(Z_i, X_i)'/\hat{\rho}_{-\ell} - (b(1, X_i) - b(0, X_i))]^2}
   \]

   (b) Update \((\lambda_n, \hat{\rho}_{-\ell})\)

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

\[
\hat{\rho}_{-\ell} = \arg\min_\rho \rho' \hat{G}_{-\ell} \rho - 2 \rho' \hat{M}_{-\ell} + 2 \lambda_n c_3 |\hat{D}_{-\ell,11} \cdot \rho_1| + 2 \lambda_n \sum_{j=2}^p |\hat{D}_{-\ell,1j} \cdot \rho_j| 
\]
where \( \rho_j \) is the \( j \)-th coordinate of \( \rho \) and \( \hat{D}_{-\ell,jj} \) is the \( j \)-th diagonal entry of \( \hat{D}_{-\ell} \).

4. Set \( \hat{\alpha}_{-\ell}(z, x) = b(z, x)/\hat{\rho}_{-\ell} \)

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

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 [29] for a detailed derivation of this soft-thresholding routine. In the optimization, we initialize at the previous value of \( \hat{\rho}_{-\ell} \). For numerical stability, we use \( \hat{D}_{-\ell} + 0.2I \) instead of \( \hat{D}_{-\ell} \), and we cap the maximum number of iterations at 10.

We justify Algorithm 4 in the same manner as [28, Section 5.1]. Specifically, we appeal to [15, Theorem 8] for the homoscedastic case and [14, Theorem 1] for the heteroscedastic case.

A.6 Consistency and Asymptotic Normality

A.6.1 Lemmas

Definition 3.

\[
G = \mathbb{E}[b(Z, X)b(Z, X)'] \quad M = \mathbb{E}[m(W, b, \theta_0)]
\]

Proposition 4. Under Assumption 2, \( |\hat{G} - G|_\infty = O_p\left(\sqrt{\frac{\ln n}{n}}\right)\)

Proof. [29, Lemma C1]

Proposition 5. Under Assumptions 1 and 2, \( |\hat{M} - M|_\infty = O_p\left(\sqrt{\frac{\ln n}{n}}\right)\)

Proof. [29, Lemma 4]

Denote \( \tilde{m}(w, \gamma) := \gamma(1, x) - \gamma(0, x) \).

Proposition 6. Under Assumptions 1 and 3

1. \( \mathbb{E}[\tilde{m}(W, \gamma_0)^2] < \infty \)
2. \( \mathbb{E}[\tilde{m}(W, \gamma) - \tilde{m}(W, \gamma_0)]^2 \) is continuous at \( \gamma_0 \) w.r.t. \( \|\gamma - \gamma_0\| \)
3. \( \max_j |\tilde{m}(W, b_j) - \tilde{m}(W, 0)| \leq C \)

Proof. [29, Theorem 6]

Proposition 7. Under Assumption 3

1. \( \mathbb{E}[\gamma_0(z, X)]^2 \leq C\mathbb{E}[\gamma_0(Z, X)]^2 \) for \( z \in \{0, 1\} \)
2. \( \mathbb{E}[\gamma(z, X) - \gamma_0(z, X)]^2 \leq C\|\gamma - \gamma_0\|^2 \) for \( d \in \{0, 1\} \)

Proof. [29, Theorem 6]

Proposition 8. Consider the estimator \( \hat{\theta} = \text{argmin}_{\theta \in \Theta} \hat{Q}(\theta) \), where \( \hat{Q} : \Theta \to \mathbb{R} \) estimates \( Q_0 : \Theta \to \mathbb{R} \). If

1. \( \Theta \) is compact
2. \( Q_0 \) is continuous in \( \theta \in \Theta \)
3. \( Q_0 \) is uniquely maximized at \( \theta_0 \)
Proof. [41, Theorem 2.1]

A.6.2 Stage 1

Proof of Theorem 2. Proposition 4, Proposition 5, and [29, Theorem 1]

Proof of Theorem 3. Proposition 4, Proposition 5, and [29, Theorem 3]. The argument that $|\hat{\varrho}| = O_p(1)$ is analogous to [29, Lemmas 2 and 3].

A.6.3 Stage 2

Proposition 9. Under Assumptions 1, 2, 3, either 4 or 5, 6, and 7,

$$\frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \psi(W_{i}, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \hat{\theta}) \xrightarrow{P} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{0}(W_{i})$$

Proof. [29, Theorem 5]

Proposition 10. Under Assumptions 1, 2, 3, either 4 or 5, 6, and 7, for each $\ell$

1. $\mathbb{E}[m(W, \hat{\gamma}_{-\ell}, \theta_{0}) - m(W, \gamma_{0}, \theta_{0})]^{2} \xrightarrow{P} 0$
2. $\mathbb{E}[\phi(W, \hat{\gamma}_{-\ell}, \alpha_{0}, \theta_{0}) - \phi(W, \gamma_{0}, \alpha_{0}, \theta_{0})]^{2} \xrightarrow{P} 0$
3. $\mathbb{E}[\phi(W, \gamma_{0}, \hat{\alpha}_{-\ell}, \theta_{0}) - \phi(W, \gamma_{0}, \alpha_{0}, \theta_{0})]^{2} \xrightarrow{P} 0$

Proof. First note that

$$\phi(W, \hat{\gamma}_{-\ell}, \alpha_{0}, \theta_{0}) - \phi(W, \gamma_{0}, \alpha_{0}, \theta_{0}) = \alpha_{0}(z, x)A(\theta_{0})[\gamma_{0}(z, x) - \hat{\gamma}_{-\ell}(z, x)]$$

$$\phi(W, \gamma_{0}, \hat{\alpha}_{-\ell}, \theta_{0}) - \phi(W, \gamma_{0}, \alpha_{0}, \theta_{0}) = [\hat{\alpha}_{-\ell}(z, x) - \alpha_{0}(z, x)]A(\theta_{0})[v - \gamma_{0}(z, x)]$$

1. Assumption 7 then Proposition 6
2. By Assumption 3 and Assumption 7
   $$\|\alpha_{0}A(\theta_{0})[\gamma_{0} - \hat{\gamma}_{-\ell}]\| \leq CA(\theta_{0})\|\gamma_{0} - \hat{\gamma}_{-\ell}\| \xrightarrow{P} 0$$
3. By Assumption 3, Theorem 2 or Theorem 3, and LIE w.r.t. $W_{-\ell} := \{W_{i}\}_{i \notin I_{\ell}}$
   $$\|[\hat{\alpha}_{-\ell} - \alpha_{0}]A(\theta_{0})[v - \gamma_{0}(z, x)]\| \leq \|\hat{\alpha}_{-\ell} - \alpha_{0}\|A(\theta_{0})C \cdot \Gamma \xrightarrow{P} 0$$

Proposition 11. Under Assumptions 1, 2, 3, either 4 or 5, 6, and 7,

$$\frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} [\phi(W_{i}, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_{0}) - \phi(W_{i}, \hat{\gamma}_{-\ell}, \alpha_{0}, \theta_{0})$$

$$- \phi(W_{i}, \gamma_{0}, \hat{\alpha}_{-\ell}, \theta_{0}) + \phi(W_{i}, \gamma_{0}, \alpha_{0}, \theta_{0})] \xrightarrow{P} 0$$

Proof. Note that

$$\phi(w, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_{0}) - \phi(w, \hat{\gamma}_{-\ell}, \alpha_{0}, \theta_{0}) - \phi(w, \gamma_{0}, \hat{\alpha}_{-\ell}, \theta_{0}) + \phi(w, \gamma_{0}, \alpha_{0}, \theta_{0})$$

$$= -[\hat{\alpha}_{-\ell}(z, x) - \alpha_{0}(z, x)]A(\theta_{0})[\hat{\gamma}_{-\ell}(z, x) - \gamma_{0}(z, x)]$$
Because convergence in first mean implies convergence in probability, it suffices to analyze

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \left[ \hat{\alpha}_{-\ell}(Z_{i}, X_{i}) - \alpha_{0}(Z_{i}, X_{i}) \right] A(\theta_{0}) \left[ \hat{\gamma}_{-\ell}(Z_{i}, X_{i}) - \gamma_{0}(Z_{i}, X_{i}) \right] \right]
\]

\[
\leq \sum_{\ell=1}^{L} \mathbb{E} \left[ \sqrt{n} \left[ \frac{1}{n} \sum_{i \in I_{\ell}} \left[ \hat{\alpha}_{-\ell}(Z_{i}, X_{i}) - \alpha_{0}(Z_{i}, X_{i}) \right] A(\theta_{0}) \left[ \hat{\gamma}_{-\ell}(Z_{i}, X_{i}) - \gamma_{0}(Z_{i}, X_{i}) \right] \right] \right]
\]

\[
= \sum_{\ell=1}^{L} \mathbb{E} \left[ \sqrt{n} \left[ \frac{1}{n} \sum_{i \in I_{\ell}} \left[ \hat{\alpha}_{-\ell}(Z_{i}, X_{i}) - \alpha_{0}(Z_{i}, X_{i}) \right] A(\theta_{0}) \left[ \hat{\gamma}_{-\ell}(Z_{i}, X_{i}) - \gamma_{0}(Z_{i}, X_{i}) \right] \right] \left| W_{-\ell} \right| \right]
\]

By element-wise Hölder and Proposition 1 we have convergence for each summand as

\[
\mathbb{E} \left[ \sqrt{n} \left[ \frac{1}{n} \sum_{i \in I_{\ell}} \left[ \hat{\alpha}_{-\ell}(Z_{i}, X_{i}) - \alpha_{0}(Z_{i}, X_{i}) \right] A(\theta_{0}) \left[ \hat{\gamma}_{-\ell}(Z_{i}, X_{i}) - \gamma_{0}(Z_{i}, X_{i}) \right] \right] \left| W_{-\ell} \right| \right]
\]

\[
\leq \mathbb{E} \left[ \frac{1}{n} \left[ \hat{\alpha}_{-\ell}(Z_{i}, X_{i}) - \alpha_{0}(Z_{i}, X_{i}) \right] A(\theta_{0}) \left[ \hat{\gamma}_{-\ell}(Z_{i}, X_{i}) - \gamma_{0}(Z_{i}, X_{i}) \right] \right] \left| W_{-\ell} \right|
\]

\[
\leq \sqrt{n} \left\| \hat{\alpha}_{-\ell} - \alpha_{0} \right\| A(\theta_{0}) \left\| \hat{\gamma}_{-\ell} - \gamma_{0} \right\|
\]

\[\xrightarrow{p} 0\]

**Proposition 12.** Under Assumption 1, for each \( \ell \)

1. \( \sqrt{n} \mathbb{E}[\psi(W, \hat{\gamma}_{-\ell}, \alpha_{0}, \theta_{0})] \xrightarrow{a.s.} 0 \)

2. \( \sqrt{n} \mathbb{E}[\phi(W, \gamma_{0}, \hat{\alpha}_{-\ell}, \theta_{0})] \xrightarrow{a.s.} 0 \)

**Proof.** Note that

\[
\mathbb{E}[\psi(W, \hat{\gamma}_{-\ell}, \alpha_{0}, \theta_{0})] = \mathbb{E}[A(\theta_{0})[\hat{\gamma}_{-\ell}(1, X) - \hat{\gamma}_{-\ell}(0, X)] + \alpha_{0}(Z, X)A(\theta_{0})[V - \hat{\gamma}_{-\ell}(Z, X)]]
\]

\[
\mathbb{E}[\phi(W, \gamma_{0}, \hat{\alpha}_{-\ell}, \theta_{0})] = \mathbb{E}[\hat{\alpha}_{-\ell}(Z, X)A(\theta_{0})[V - \gamma_{0}(Z, X)]]
\]

1. The LHS is exactly 0 by Theorem 1 as well as LIE w.r.t. \( W_{-\ell} \)

2. The LHS is exactly 0 by Theorem 1 as well as LIE w.r.t. \( W_{-\ell} \)

**Proposition 13.** Under Assumptions 1, 2, 3, either 4 or 5, 6, and 7,

1. \( J \) exists

2. \( \exists \) neighborhood \( \mathcal{N} \) of \( \theta_{0} \) w.r.t. \( \left\| \cdot \right\|_{2} \) s.t.

   (a) \( \left\| \hat{\gamma}_{-\ell} - \gamma_{0} \right\| \xrightarrow{p} 0 \)

   (b) \( \left\| \hat{\alpha}_{-\ell} - \alpha_{0} \right\| \xrightarrow{p} 0 \)

   (c) for \( \left\| \gamma - \gamma_{0} \right\| \) and \( \left\| \alpha - \alpha_{0} \right\| \) small enough, \( \psi(W_{i}, \gamma, \alpha, \theta) \) is diff. in \( \theta \) w.p.a. 1

   (d) \( \exists \xi > 0 \) and \( d(W) \) s.t. \( \mathbb{E}[d(W)] < \infty \) and for \( \left\| \gamma - \gamma_{0} \right\| \) small enough,

   \[
   \left\| \frac{\partial \psi(w, \gamma, \alpha, \theta)}{\partial \theta} - \frac{\partial \psi(w, \gamma, \alpha, \theta_{0})}{\partial \theta} \right\|_{2} \xrightarrow{p} \xi
   \]

3. \( \mathbb{E} \left| \frac{\partial \psi_{j}(W, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_{0})}{\partial \theta_{k}} - \frac{\partial \psi_{j}(W, \gamma_{0}, \alpha_{0}, \theta_{0})}{\partial \theta_{k}} \right| \xrightarrow{p} 0, \forall \ell, j, k \)

24
Proof. Note that
\[ \frac{\partial \psi(w, \gamma, \alpha, \theta)}{\partial \theta} = \frac{\partial A(\theta)}{\partial \theta} [\gamma(1, x) - \gamma(0, x)] + \alpha(0, x) \frac{\partial A(\theta)}{\partial \theta} [v - \gamma(z, x)] \]
where \( \frac{\partial A(\theta)}{\partial \theta} \) is a tensor consisting of 1s and 0s.

1. It suffices to show finite second moment. By triangle inequality and Assumption 3
\[ \left\| \frac{\partial A(\theta_0)}{\partial \theta} [\gamma_0(1, x) - \gamma_0(0, x)] + \alpha_0(z, x) \frac{\partial A(\theta)}{\partial \theta} [v - \gamma_0(z, x)] \right\| 
\leq \frac{\partial A(\theta_0)}{\partial \theta} \{ \| \gamma_0(1, x) - \gamma_0(0, x) \| + C \} \]
To bound the RHS, appeal to Proposition 7.
\[ \| \gamma_0(1, x) - \gamma_0(0, x) \| \leq \| \gamma_0(1, x) \| + \| \gamma_0(0, x) \| \leq C \| \gamma_0 \| < \infty \]

2. (a) Assumption 7
(b) Theorem 2 or Theorem 3
(c) \( \frac{\partial A(\theta)}{\partial \theta} \) is a tensor consisting of 1s and 0s
(d) the LHS is exactly \( \tilde{\theta} \)

3. It suffices to analyze the difference
\[ \xi = \hat{\gamma}_{-\ell}(1, x) - \hat{\gamma}_{-\ell}(0, x) + \hat{\alpha}_{-\ell}(z, x)[v - \hat{\gamma}_{-\ell}(z, x)] 
- \{ \gamma_0(1, x) - \gamma_0(0, x) + \alpha_0(z, x)[v - \gamma_0(z, x)] \} 
= \hat{\gamma}_{-\ell}(1, x) - \gamma_0(1, x) 
- \hat{\gamma}_{-\ell}(0, x) + \gamma_0(0, x) 
+ \hat{\alpha}_{-\ell}(z, x)[v - \hat{\gamma}_{-\ell}(z, x)] - \alpha_0(z, x)[v - \gamma_0(z, x)] 
+ \alpha_0(z, x)[v - \gamma_0(z, x)] - \alpha_0(z, x)[v - \gamma_0(z, x)] 
= \hat{\gamma}_{-\ell}(1, x) - \gamma_0(1, x) 
- \hat{\gamma}_{-\ell}(0, x) + \gamma_0(0, x) 
+ \hat{\alpha}_{-\ell}(z, x) - \alpha_0(z, x)[v - \gamma_0(z, x)] 
+ \alpha_0(z, x)[\gamma_0(z, x) - \hat{\gamma}_{-\ell}(z, x)] \]
where we use the decomposition
\[ \hat{\alpha}_{-\ell}(z, x)[v - \hat{\gamma}_{-\ell}(z, x)] - \alpha_0(z, x)[v - \hat{\gamma}_{-\ell}(z, x)] 
= [\hat{\alpha}_{-\ell}(z, x) - \alpha_0(z, x)] [v - \gamma_0(z, x) + \gamma_0(z, x) - \hat{\gamma}_{-\ell}(z, x)] \]
Hence
\[ E[|\xi|] 
\leq E[|\hat{\gamma}_{-\ell}(1, X) - \gamma_0(1, X)|] 
+ E[|\hat{\gamma}_{-\ell}(0, X) - \gamma_0(0, X)|] 
+ E[|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)| |v - \gamma_0(Z, X)|] 
+ E[|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)| |\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)|] 
+ E[|\alpha_0(Z, X)| |\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)|] \]
Consider the first two terms. By Jensen, Proposition 7, and Assumption 7
\[ E[|\hat{\gamma}_{-\ell}(1, X) - \gamma_0(1, X)|] \leq \| \hat{\gamma}_{-\ell}(1, x) - \gamma_0(1, x) \| \leq C \| \gamma_0 \| \overset{p}{\rightarrow} 0 \]
\[ E[|\hat{\gamma}_{-\ell}(0, X) - \gamma_0(0, X)|] \leq \| \hat{\gamma}_{-\ell}(0, x) - \gamma_0(0, x) \| \leq C \| \gamma_0 \| \overset{p}{\rightarrow} 0 \]
Consider the third term. By Hölder, Theorem 2 or Theorem 3, Assumption 3, and LIE w.r.t. \( W_{-\ell} \)
\[ E[|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)| |v - \gamma_0(Z, X)|] \leq \| \hat{\alpha}_{-\ell} - \alpha_0 \| \| v - \gamma_0(z, x) \| \leq C \| \hat{\alpha}_{-\ell} - \alpha_0 \| \overset{p}{\rightarrow} 0 \]
Proof. We verify the conditions of Proposition 8 with
\[ \mathbb{E}[\|\alpha_{-\ell}(Z, X) - \alpha_0(Z, X)\| \gamma_0(Z, X) - \gamma_{-\ell}(Z, X)] \leq \|\alpha_{-\ell} - \alpha_0\| \gamma_0 - \gamma_{-\ell} \overset{p}{\rightarrow} 0 \]
Consider the fifth term. By Assumption 3, Jensen, and Assumption 7
\[ \mathbb{E}[\alpha_0(Z, X)[\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)] \leq C\mathbb{E}[\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)] \leq C\|\gamma_0 - \hat{\gamma}_{-\ell}\| \overset{p}{\rightarrow} 0 \]

Proposition 14. Suppose Assumptions 1, 2, 3, either 4 or 5, 6, 7, and 8 hold. Then \( \hat{\theta} \overset{p}{\rightarrow} \theta_0 \)

Proof. We verify the conditions of Proposition 8 with
\[ Q_0(\theta) = \mathbb{E}[\psi_0(\theta)[\mathbb{E}]\psi_0(\theta)] \]
\[ \hat{Q}(\theta) = \left[ \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i(\theta) \right]' \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i(\theta) \]
\[ \psi_0(\theta) = \psi(W, \gamma_0, \alpha_0, \theta) \]
\[ \hat{\psi}_i(\theta) = \psi(W, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta) \]

1. Assumption 8
2. Theorem 1
3. Theorem 1
4. Define
\[ \eta_0(w) = \gamma_0(1, x) - \gamma_0(0, x) + \alpha_0(z, x)[v - \gamma_0(z, x)] \]
\[ \hat{\eta}_{-\ell}(w) = \hat{\gamma}_{-\ell}(1, x) - \hat{\gamma}_{-\ell}(0, x) + \hat{\alpha}_{-\ell}(z, x)[v - \hat{\gamma}_{-\ell}(z, x)] \]

It follows that
\[ \psi_0(\theta) = A(\theta)\eta_0(W), \quad \mathbb{E}\psi_0(\theta) = A(\theta)\mathbb{E}\eta_0(W) \]
\[ \hat{\psi}_i(\theta) = A(\theta)\hat{\eta}_{-\ell}(W_i), \quad \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i(\theta) = A(\theta) \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\eta}_{-\ell}(W_i) \]

It suffices to show \( \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\eta}_{-\ell}(W_i) \overset{p}{\rightarrow} \mathbb{E}\eta_0(W) \) since by continuous mapping theorem this implies that \( \forall \theta \in \Theta, \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i(\theta) \overset{p}{\rightarrow} \mathbb{E}\psi_0(\theta) \) and hence \( \hat{Q}(\theta) \overset{p}{\rightarrow} Q_0(\theta) \) uniformly.

We therefore turn to arguing the sufficient condition. Write
\[ \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\eta}_{-\ell}(W_i) - \mathbb{E}\eta_0(W) \]
\[ = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} [\hat{\eta}_{-\ell}(W_i) - \eta_0(W_i)] + \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \eta_0(W_i) - \mathbb{E}\eta_0(W) \]
Consider the first and second term. Denote \( \xi_i = \hat{\eta}_{-\ell}(W_i) - \eta_0(W_i) \) as in Proposition 13.3. Since convergence in mean implies convergence in probability, it suffices to analyze

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} |\xi_i| \right] \leq \frac{1}{n} \sum_{\ell=1}^{L} \mathbb{E} \left[ \frac{1}{n} \sum_{i \in I_\ell} |\xi_i| \right] = \sum_{\ell=1}^{L} \mathbb{E} \mathbb{E} \left[ \frac{1}{n} \sum_{i \in I_\ell} |\xi_i| W_{-\ell} \right] = \sum_{\ell=1}^{L} \frac{n}{n} \mathbb{E} ||\xi_i|| |W_{-\ell}| \\
\leq \sum_{\ell=1}^{L} \mathbb{E} \mathbb{E} ||\xi_i|| |W_{-\ell}| \\
\rightarrow 0
\]

by triangle inequality, LIE, and the proof of Proposition 13.3.

Consider the third and fourth term.

\[
\frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \eta_0(W_i) - \mathbb{E} \eta_0(W) = \frac{1}{n} \sum_{i=1}^{n} \eta_0(W_i) - \mathbb{E} \eta_0(W) \overset{P}{\rightarrow} 0
\]

by WLLN and the fact that \( \mathbb{E}||\eta_0(W)||^2 < \infty \), guaranteed by the argument in Proposition 15.3 below.

\[\Box\]

**Proposition 15.** Under Assumptions 1, 2, 3, either 4 or 5, 6, 7, and 8,

1. \( \hat{\theta} \overset{P}{\rightarrow} \theta_0 \)
2. \( J'J \) is nonsingular
3. \( \mathbb{E}||\psi_0(W)||^2 < \infty \)
4. \( \mathbb{E}||\phi(W, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_0) - \phi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0) - \phi(W, \gamma_0, \hat{\alpha}_{-\ell}, \theta_0) + \phi(W, \gamma_0, \alpha_0, \theta_0)||^2 \overset{P}{\rightarrow} 0 \)

**Proof.** As before

\[
\phi(W, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_0) - \phi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0) - \phi(W, \gamma_0, \hat{\alpha}_{-\ell}, \theta_0) + \phi(W, \gamma_0, \alpha_0, \theta_0) \\
= -[\hat{\alpha}_{-\ell}(z, x) - \alpha_0(z, x)] A(\theta_0) [\hat{\gamma}_{-\ell}(z, x) - \gamma_0(z, x)]
\]

1. Proposition 14
2. Assumption 3
3. By triangle inequality, Assumption 3, and Proposition 7

\[
||\gamma_0(1, x) - \gamma_0(0, x) + \alpha_0(z, x) [v - \gamma_0(z, x)]|| \leq ||\gamma_0(1, x) - \gamma_0(0, x)|| + CC'
\]

To bound the RHS, appeal to Proposition 7.

\[
||\gamma_0(1, x) - \gamma_0(0, x)|| \leq ||\gamma_0(1, x)|| + ||\gamma_0(0, x)|| \leq C ||\gamma_0|| < \infty
\]

4. It suffices to analyze

\[
|| -[\hat{\alpha}_{-\ell} - \alpha_0] A(\theta_0) [\hat{\gamma}_{-\ell} - \gamma_0] || \leq ||\hat{\alpha}_{-\ell} A(\theta_0) [\hat{\gamma}_{-\ell} - \gamma_0] || + ||\alpha_0 A(\theta_0) [\hat{\gamma}_{-\ell} - \gamma_0] ||
\]

Consider the first term. By Hölder, Assumption 2, and either Theorem 2 or Theorem 3

\[
||\hat{\alpha}_{-\ell}(z, x)|| = ||\hat{\rho}_v b(z, x)|| \leq ||\hat{\rho}_v||_1 ||b(z, x)||_{\infty} = O_p(1)
\]

27
where $C$ is the diagonal entry of $J$.

Next, we prove the validity of Gaussian multiplier bootstrap for approximating this $c_0$, appealing to [23, Theorem 5.1]. Gaussian multiplier bootstrap is operationally equivalent to approximating coverage of confidence bands in the form of the rectangle

$$[(l_0), (u_0)] = \left[ \hat{\theta}_j - c_0 \sqrt{\frac{C_{jj}}{n}}, \hat{\theta}_j + c_0 \sqrt{\frac{C_{jj}}{n}} \right]$$

where $C_{jj}$ is the diagonal entry of $C$ corresponding to $j$-th element $\hat{\theta}_j$ of $\theta$.

The argument is as follows. Denote $[l_0, u_0] = \times_{j=1:2d}[(l_0)_j, (u_0)_j]$ and $d = \text{dim}(U)$. Then the simultaneous coverage probability is

$$\mathbb{P}(\theta_0 \in [l_0, u_0]) = \mathbb{P}(\sqrt{n}(\theta - \theta_0) \in S^{1/2}[-c_0, c_0]^{2d})$$

$$= \mathbb{P}(N(0, C) \in S^{1/2}[-c_0, c_0]^{2d}) + o(1)$$

$$= \mathbb{P}(S^{-1/2}N(0, C) \in [-c_0, c_0]^{2d}) + o(1)$$

$$= \mathbb{P}(|N(0, \Sigma)|_\infty \leq c_0) + o(1)$$

$$= 1 - \alpha + o(1)$$

Next, we prove the validity of Gaussian multiplier bootstrap for approximating this $c_0$, appealing to [23, Theorem 5.1]. Gaussian multiplier bootstrap is operationally equivalent to approximating $c_0$ with $c$ calculated in Algorithm 3. We match symbols with [23] to formalize the equivalence, designating notation from [23] with tildes.

$$\tilde{T} = \left| S^{-1/2} \sqrt{n}(\tilde{\theta} - \theta_0) \right|_\infty$$

$$\tilde{T}_0 = \left| S^{-1/2} \sqrt{n} \sum_{i=1}^n J^{-1} \hat{\psi}(W_i) \right|$$

and

$$\tilde{W} = \left| S^{-1/2} \sqrt{n} \sum_{i=1}^n \sum_{\ell=1}^L \tilde{J}^{-1} \hat{\psi}(\tilde{\theta}) e_i \right|_{\infty}$$

$$\tilde{W}_0 = \left| S^{-1/2} \sqrt{n} \sum_{i=1}^n J^{-1} \hat{\psi}(W_i) e_i \right|_{\infty}$$

where $e_i \sim N(0, 1)$.

By linearity of $\hat{\psi}(\theta)$, we can write

$$\frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \hat{\psi}(\theta) - \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \hat{\psi}(\theta_0) = \tilde{J}(\tilde{\theta} - \theta_0)$$
We estimate $\theta_0$ is exactly identified, the first term on the LHS is $\frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{\psi}_i(\hat{\theta}) = 0$. Premultiplying the above equation by $\hat{S}^{-1/2} \hat{J}^{-1}$, we observe that $\hat{S}^{-1/2} \sqrt{n}(\hat{\theta} - \theta_0)$ can be represented as an approximate mean: it is equal to a scaled average of $\psi_0(W_i)$ plus a remainder term $r_n$

$$\hat{S}^{-1/2} \sqrt{n}(\hat{\theta} - \theta_0) = - \frac{1}{\sqrt{n}} \hat{S}^{-1/2} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{J}^{-1} \hat{\psi}_i(\theta_0) = - \frac{1}{\sqrt{n}} \hat{S}^{-1/2} \sum_{i=1}^{n} J^{-1} \psi_0(W_i) + r_n$$

where

$$r_n := \frac{1}{\sqrt{n}} \hat{S}^{-1/2} \sum_{i=1}^{n} J^{-1} \psi_0(W_i) - \frac{1}{\sqrt{n}} \hat{S}^{-1/2} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{J}^{-1} \hat{\psi}_i(\theta_0)$$

After matching symbols, we verify [23, Condition M], which is a sufficient condition for [23, Theorem 5.1]. Assumption 10 verifies [23, Conditions M.i. M.ii]. As argued in [23, Comment 2.2], a sufficient condition for the first clause of [23, Condition M.iii] is that $\hat{S}^{-1/2} \hat{J}^{-1} \psi_0(W_i)$ is bounded. Assumption 9 verifies the second clause of [23, Condition M.iii].

Finally, we prove that $\hat{S}^{-1/2} \hat{J}^{-1} \psi_0(W_i)$ is indeed bounded. Observe that

$$\|J^{-1}\|_{op} = \lambda_{\text{max}}(J^{-1}) = \frac{1}{\lambda_{\text{min}}(J)} < \infty$$

where $\| \cdot \|_{op}$ is operator norm and the inequality appeals to Hypothesis 3.3. Recall that

$$\psi_0(W_i) = A(\theta_0)[\gamma_0(1, X_i) - \gamma_0(0, X_i)] + \alpha_0(Z_i, X_i) A(\theta_0)[V_i - \gamma_0(Z_i, X_i)]$$

By Hypothesis 3.1, $|\alpha_0(Z_i, X_i)| < \infty$. For complier counterfactual distributions, $V_i^y \in \{0, 1\} \times \{0, 1\}$ and hence $\gamma_0^y(Z_i, X_i) \in [-1, 0] \times \{0, 1\}$ for all $y \in \mathcal{U}$. $\square$

A.7 Simulation

A.7.1 Design

We implement 500 simulations. One simulation consists of a sample of $n = 1000$ observations. A given observation is generated from the following IV model:

$$X \sim i.i.d. Unif[0, 1], \quad Z | X \sim i.i.d. Bern(\pi_0(x)), \quad D | Z = z, X \sim i.i.d. Bern(\gamma_0(z, x)), \quad Y | Z = z, X \sim i.i.d. \mathcal{N}(\gamma_0(Y(z, x), 1)$$

where $Y$ is the continuous outcome, $D$ is the binary treatment, $Z$ is the binary instrumental variable, and $X$ is the covariate. $Y$ and $D$ are drawn independently given $Z$ and $X$. In particular,

$$\pi_0(x) = 0.05 \cdot 1_{x < 0.5} + 0.95 \cdot 1_{x > 0.5}, \quad \gamma_0^Y(z, x) = 2zx^2, \quad \gamma_0^D(z, x) = zx$$

From observations of $W = (Y, D, Z, X)'$, we estimate complier counterfactual outcome distributions $\hat{\theta} = (\hat{\theta}', \hat{\delta}')'$ at a few grid points. For $\beta_0 = \{\beta_0^y\} = \{P(Y(0) \leq y|D(1) > D(0))\}$, we set $y \in \{-3, -2, \ldots, 3, 4\}$. For $\delta_0 = \{\delta_0^y\} = \{P(Y(1) \leq y|D(1) > D(0))\}$, we set $y \in \{-2, -1, \ldots, 4, 5\}$. The true parameter values are

$$\beta_0^y = \frac{\int_0^1 [\Phi(y - 2x^2)(x - 1) + \Phi(y)] dx}{\int_0^1 x dx}$$

$$\delta_0^y = \frac{\int_0^1 [\Phi(y - 2x^2)x] dx}{\int_0^1 x dx}$$

A.7.2 Algorithms

We estimate $\hat{\delta}$ and $\hat{\alpha}$ with a dictionary $b(Z, X)$ consisting of fourth-order polynomials of $X$ and interactions between $Z$ and the polynomials. We estimate $\hat{\gamma}$ with a dictionary $b(X)$ consisting of fourth-order polynomials of $X$. We set $L = 5$ and $n_k = n/L$ for sample splitting.
Auto-DML. We estimate $\hat{\alpha}_{-\ell}$ by Lasso according to Algorithm 4. We estimate $\hat{\gamma}_V$ by Lasso according to Algorithm 4 and by neural network using the settings in [22]. Subsequently, we estimate $\theta$ by Algorithm 1.

DML-plugin. The only difference from Auto-DML is how we estimate $\hat{\alpha}_{-\ell}$. For DML-plugin, we first estimate $\hat{\pi}_{-\ell}$ by $\ell_1$-regularized logistic regression, and then form $\hat{\alpha}_{-\ell}$ by plugging $\hat{\pi}_{-\ell}$ into its formula.

kappa-weight. We estimate $\hat{\pi}$ by logistic regression, then plug $\hat{\pi}$ into Definition 2 to obtain $\hat{\kappa}^{(0)}(w), \hat{\kappa}^{(1)}(w), \text{or } \hat{\kappa}(w)$. The $\kappa$-weighted estimator is then the sample analog of formula given in [2]. For example, for $\theta_0 = \mathbb{E}[g(Y^{(0)}, X) | D^{(1)} > D^{(0)}]$, the $\kappa$-weighted estimator is

$$\hat{\theta}^\kappa = \frac{1}{n} \sum_{i=1}^{n} \kappa^{(0)}(W_i) \cdot g(Y_i, X_i)$$

MATLAB code for replicating the simulations is available at https://github.com/lsun20/rrr_kappa_replication.

A.7.3 Sensitivity

To investigate the sensitivity of Auto-DML to hyperparameters, we revisit our counterfactual distribution simulation exercise. We first consider the sensitivity of Auto-DML to the number of folds $L$ used in sample splitting. In Figure 3, we report simulation results for $L \in \{2, 5, 10\}$. Across $L$, we estimate $\hat{\gamma}$ by Lasso and tune the regularization parameter $\lambda$ according to Algorithm 4. We find that Auto-DML is insensitive to $L$.

Next, we consider the sensitivity of Auto-DML to the regularization parameter $\lambda$. We consider $\lambda \in \{0.5\lambda_n, \lambda_n, 2\lambda_n\}$ where $\lambda_n$ is the value obtained by Algorithm 4. Across $\lambda$, we estimate $\hat{\gamma}$ by Lasso and we fix $L = 5$. The simulation results in Figure 4 demonstrate that Auto-DML is insensitive to the value of $\lambda$. 

![Figure 3: Sensitivity to $L$](image)
A.7.4 Trimming and Censoring

We assess the effectiveness of trimming and censoring in the simulation considered in Section 7.1. We impose trimming according to \[ \hat{\pi} \notin [10^{-12}, 1 - 10^{-12}] \]. We impose censoring by setting \( \hat{\pi} < 10^{-12} \) to be \( 10^{-12} \) and \( \hat{\pi} > 10^{-12} \) to be \( 1 - 10^{-12} \). Trimming and censoring improve stability considerably for DML, though sometimes at the cost of introducing bias.

Figure 4: Sensitivity to \( \lambda \)

\[ \beta_0^y = \mathbb{P}(Y(0) \leq y \mid D(1) > D(0)) \]

\[ \delta_0^y = \mathbb{P}(Y(1) \leq y \mid D(1) > D(0)) \]

Figure 5: Trimming

\[ \beta_0^y = \mathbb{P}(Y(0) \leq y \mid D(1) > D(0)) \]

\[ \delta_0^y = \mathbb{P}(Y(1) \leq y \mid D(1) > D(0)) \]

Figure 6: Censoring
A.8 Effect of 401(k) on Assets

A.8.1 Dictionaries

We follow [22] in constructing the dictionary $b$. The raw covariates $X$ are age, income, years of education, family size, marital status, benefit pension status, IRA participation, and home-ownership. The dictionary includes polynomials of continuous covariates, interactions among all covariates, and interactions between covariates and treatment status.

A.8.2 Algorithm

We implement the same pre-processing step as [22] to facilitate comparison, which is interpretable as imposing the overlap condition: in the group with $Z = 0$, we drop observations with estimated propensity scores $\hat{\pi}$ that exceed the maximum and minimum propensity scores in the group with $Z = 1$. Note that we do not implement this pre-processing step in the trimming and censoring simulation exercises of Appendix A.7.4.

We then implement Auto-DML exactly as described in Appendix A.7.2. R code for replicating the empirical application is available at https://github.com/lsun20/rrr_kappa_replication.

A.8.3 Results

Finally, we implement Auto-DML without the pre-processing step. The full sample has $n = 9915$ observations. Figure 7 shows that our empirical findings remain the same.

(a) $\hat{\gamma}$ estimated by Lasso  
(b) $\hat{\gamma}$ estimated by neural network

Figure 7: Effect of 401(k) on net financial assets for compliers without pre-processing