AUTOMATIC KAPPA WEIGHTING FOR INSTRUMENTAL VARIABLE MODELS OF COMPLIER TREATMENT EFFECTS

RAHUL SINGH† AND LIYANG SUN‡

ABSTRACT. We propose debiased machine learning estimators for complier parameters, such as local average treatment effect, with high dimensional covariates. To do so, we characterize the doubly robust moment function for the entire class of complier parameters as the combination of Wald and κ weight formulations. We directly estimate the κ weights, rather than their components, in order to eliminate the numerically unstable step of inverting propensity scores of high dimensional covariates. We prove our estimator is balanced, consistent, asymptotically normal, and semiparametrically efficient, and use it to estimate the effect of 401(k) participation on the distribution of net financial assets.

Keywords: high dimensional, doubly robust, complier

JEL: C14, C36, C45

Date: November 1, 2021.

We thank Alberto Abadie, Anish Agarwal, Isaiah Andrews, Joshua Angrist, Abhijit Banerjee, Victor Chernozhukov, Avi Feller, Anna Mikusheva, Whitney Newey, Garima Sharma, and Suhas Vijaykumar for helpful comments. The paper also benefited from the comments of conference audiences at Econometric Society World Congress 2020, NABE Tech Economics Conference 2020, and NeurIPS CausalML Workshop 2019. Replication code is available at https://github.com/lsun20/rrr_kappa_replication.

†rahul.singh@mit.edu, MIT Department of Economics.

‡lsun20@berkeley.edu, UC Berkeley and CEMFI. Support from the Jerry A. Hausman Graduate Dissertation Fellowship and Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010, is gratefully acknowledged.
1. Introduction

We consider the problem of estimating low dimensional complier parameters using a binary instrumental variable $Z$, which is valid conditional on high dimensional covariates $X$. Angrist et al. [1996] prove that identification of the local average treatment effect (LATE) based on the instrumental variable does not require any functional form restrictions. Using $\kappa$ weighting, Abadie [2003] extends identification for an entire class of complier parameters that may be of interest to applied researchers in addition to LATE. We characterize the doubly robust moment function for this entire class of complier parameters by augmenting $\kappa$ weighting with the classic Wald formulation. The doubly robust moment function allows the possibility that the treatment effect for different individuals in the population may vary flexibly according to their covariates [Fröhlich, 2007]. The doubly robust moment function also allows for nonlinear models [Abadie, 2003], which are often appropriate when output $Y$ and treatment $D$ are binary. Finally, it allows for model selection of covariates and their transformations using machine learning.

We propose a new approach, automatic $\kappa$ weighting (Auto-$\kappa$), that uses machine learning to select covariates and their transformations that approximate heterogeneity and nonlinearity well. Unlike previous work on complier parameters, our approach also selects covariates and their transformations that are well balanced among the full population, the subpopulation that is assigned the instrument, and the subpopulation that is not assigned the instrument. We propose a machine learning update to $\kappa$ weighting to achieve this goal.

Approximating heterogeneity and nonlinearity with a regularized machine learning estimator can introduce bias into complier parameter estimation. Similarly, performing model selection with machine learning can introduce bias. However, by situating the estimation problem in a debiased machine learning (DML) framework, we correct for such bias and obtain valid inference [Belloni et al., 2012, 2014, Chernozhukov et al., 2018a, Farrell, 2015, Robins et al., 2013]. Our key insight is that the $\kappa$ weight is the Riesz representer to the Wald
formula, so a doubly robust estimating equation that combines both classic formulations not only guards against mis-specification but also debiases machine learning.

We consider a general class of complier parameters including local average treatment effect, average complier characteristics, and complier counterfactual outcome distributions. Average complier characteristics help to assess the external validity of any study that uses instrumental variable identification [Angrist and Evans, 1998, Angrist and Fernández-Val, 2013]—whose treatment effects are we estimating when we use a particular instrument? Counterfactual outcome distributions are particularly important in welfare analysis of schooling, subsidized training, union status, minimum wages, and transfer programs [Abadie, 2002, Abadie et al., 2002].

We make the following contributions.

**Computational.** It appears that we propose the first machine learning estimator of any kind for average complier characteristics. More generally, we present a doubly robust estimator for a broad class of complier parameters that combines two classic approaches: a Wald formulation and a $\kappa$ weight formulation. In our approach, we directly estimate the $\kappa$ weights rather than their components in order to eliminate the numerically unstable step of inverting propensity scores of high dimensional covariates. To our knowledge, we propose the first machine learning estimator of complier parameters that eliminates the need for ad hoc trimming or censoring of propensity scores. Trimming and censoring extreme propensity scores are common practices in economic research, yet these practices have little theoretical justification.

**Statistical.** We characterize the doubly robust moment function for the entire class of complier parameters studied by Abadie [2003]. To do so, we reinterpret the $\kappa$ weight as the Riesz representer to the Wald formula, which appears to be a new insight. We prove that our estimator, which uses this doubly robust moment function as an estimating equation, is consistent, asymptotically normal, and semiparametrically efficient. We also prove a finite sample balancing property that distinguishes our estimator from previous work.
As a consequence of our main results, we propose a hypothesis test, free of functional form restrictions, to evaluate whether two different instruments induce the same subpopulation of compliers and hence identify the same causal parameters. It appears that no semiparametric test previously exists for this important question about the external validity of instruments. For complier counterfactual outcome distributions over a fixed grid, we justify simultaneous confidence bands by Gaussian multiplier bootstrap.

**Empirical.** We show that Auto-κ outperforms alternative approaches to estimating complier parameters in mean square error, while the simultaneous confidence bands still achieve nominal coverage. In other words, the new approach provides confidence intervals that are smaller yet still valid. We also demonstrate that the new approach does not require ad hoc pre-processing such as trimming or censoring. This is our main contribution for empirical economic practice using a binary instrument and high dimensional covariates. We use Auto-κ to estimate the effect of 401(k) participation on the distribution of net financial assets, instrumenting with 401(k) eligibility.

Section 2 summarizes related work. Section 3 previews the main ideas before Section 4 demonstrates the practicality of Auto-κ through simulations and a real world program evaluation. Sections 5, 6, and 7 formalize the setting, estimation procedure, and theoretical guarantees, respectively. Section 8 concludes.

## 2. Related work

Our estimator can be viewed as a machine learning update to κ weighting, a popular estimator introduced by Abadie [2003]. In κ weighting, any complier parameter can be expressed as a weighted average of the corresponding population quantity. Abdulkadiroğlu et al. [2017] consider a κ weighting estimator of LATE. Likewise, Abadie et al. [2002], Angrist [2001], Angrist et al. [2016] propose κ weighting estimators of counterfactual outcome distributions. The weight involves a propensity score in the denominator. Theoretically, this literature has not yet justified the use of a regularized machine learning estimator for the κ weight in high
dimensional settings. By elucidating the relationship between $\kappa$ weighting and DML, we provide this justification. Moreover, we are able to estimate the $\kappa$ weight directly without estimating and inverting the propensity score.

Various works propose estimators of specific complier parameters by original DML with explicit propensity scores. Ogburn et al. [2015] and Chernozhukov et al. [2018a] present a DML estimator for LATE. Belloni et al. [2017] present a DML estimator for counterfactual outcome distributions. All of these estimators involve plugging in an estimated propensity score in the denominator of a balancing weight, which is numerically unstable, particularly in high dimensional settings.

This paper subsumes our earlier draft [Singh and Sun, 2019]. In independent work, Sun and Tan [2020] study a regularized calibration estimator of LATE with explicit propensity score estimation and cumbersome sequential fitting. Qiu et al. [2021] propose a regularized two stage regression with a non-convex objective that is not doubly robust. The procedure involves explicit propensity scores, constrained optimization with composite gradient descent, another constrained optimization to obtain a sparse precision matrix, and extensive censoring throughout.

Unlike previous work, we present a general characterization of the doubly robust moment function for the entire class of complier parameters studied by Abadie [2003], we estimate average complier characteristics with machine learning, and we avoid the numerically unstable step of inverting propensity scores. Our proposed procedure is also relatively simple (and we provide R code).

We directly estimate the $\kappa$ weight in the spirit of automatic debiased machine learning (Auto-DML) [Chernozhukov et al., 2018b,c]. Whereas Chernozhukov et al. [2018b,c] consider parameters of the full population identified by selection on observables, we consider parameters of the complier subpopulation identified by instrumental variables. See Chernozhukov et al. [2018b,c, 2021] for a comparison between Auto-DML and other approaches to semiparametric estimation that use machine learning, e.g. targeted maximum likelihood
[Van der Laan and Rose, 2011], efficient score [Ning and Liu, 2017], approximate residual balancing [Athey et al., 2018], and augmented minimax linear estimation [Hirshberg and Wager, forthcoming].

Our paper contributes to the growing literature on instrumental variables in machine learning. Both Hartford et al. [2017] and Singh et al. [2019] consider the problem of nonparametric instrumental variable regression, where the target parameter is the structural function $h$ that summarizes the counterfactual relationship $Y = h(D, X) + e$, where $e$ is confounding noise. Athey et al. [2019] further assume that 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 assume that the confounding noise $e$ is additively separable—a model proposed by Newey and Powell [2003]. In this setting, Hartford et al. [2017] introduce nonlinearity with neural networks, Singh et al. [2019] with RKHS methods, and Athey et al. [2019] with random forests.

In our setting, we do not assume additive separability of confounding noise—a model considered by Angrist et al. [1996]. Complier parameters are scalar summaries of the reduced form $E[Y|Z, X]$ and first stage $E[D|Z, X]$ (or appropriate generalizations thereof). We allow generic machine learning for nonlinear estimation and model selection of both the reduced form and first stage en route to estimation of complier parameters. Such estimators are called semiparametric [Newey, 1994].

Recently, Angrist and Frandsen [2019] outline three possible ways in which machine learning may be useful to update classic empirical strategies in labor economics: (i) selecting covariates and their transformations under selection on observables; (ii) selecting instruments and their transformations under IV identification; and (iii) selecting covariates and their transformations under IV identification. The authors provide strong empirical evidence in favor of (i), and pose (ii) and (iii) as areas for further research. In this work, we propose a new approach for (iii) that performs well with simulated and real economic data. Auto-$\kappa$
uses machine learning to select covariates (and their transformations) that not only approximate nonlinearity and heterogeneity but also confer balance, whereas existing methods such as $\kappa$ weighting and original DML do not.

3. Preview

We provide a concise preview of the framework, new estimation procedure, and theoretical guarantees developed in this paper. For clarity, we focus on the familiar example of local average treatment effect (LATE) in this initial discussion: $\theta_0 = \mathbb{E}[Y^{(1)} - Y^{(0)}|D^{(1)} > D^{(0)}]$. We study average complier characteristics and complier counterfactual outcome distributions in subsequent sections of the main text, and we study the entire class of complier parameters from Abadie [2003] in Appendix C.

3.1. Classic approaches: Wald formula and $\kappa$ weight. Under standard identification conditions of conditional independence, exclusion, overlap, and monotonicity [Angrist et al., 1996], LATE can be identified as

$$
\theta_0 = \frac{\mathbb{E}\{\mathbb{E}[Y|Z = 1, X] - \mathbb{E}[Y|Z = 0, X]\}}{\mathbb{E}\{\mathbb{E}[D|Z = 1, X] - \mathbb{E}[D|Z = 0, X]\}}
$$

following [Frölich, 2007, Theorem 1]. We call this expression the expanded Wald formula.

Therefore the direct Wald approach would involve estimating the reduced form $\mathbb{E}[Y|Z, X]$ and first stage $\mathbb{E}[D|Z, X]$, then plugging these estimates into the expanded Wald formula. Such an approach is called the “plug-in”, and it is valid only when both regressions are estimated with correctly specified and unregularized models. It is not a valid approach when either regression is incorrectly specified, leading to the name “forbidden regression” [Angrist and Pischke, 2008]. It is also invalid when the covariates are high dimensional and a regularized machine learning estimator is used to estimate either regression.
In seminal work, Abadie [2003] proposes an alternative formulation of a broad class of complier parameters including LATE. Abadie [2003] defines the \( \kappa \) weights

\[
\kappa^{(0)}(W) = (1 - D) \frac{(1 - Z) - (1 - \pi_0(X))}{\{1 - \pi_0(X)\}\pi_0(X)}, \quad \kappa^{(1)}(W) = D \frac{Z - \pi_0(X)}{\{1 - \pi_0(X)\}\pi_0(X)}
\]

where \( \pi_0(X) = \Pr(Z = 1|X) \) is the instrument propensity score and where we use the shorthand \( W = (Y, D, Z, X)' \).

The \( \kappa \) weights have the property that

\[
\theta_0 = \frac{\mathbb{E}[\kappa^{(1)}(W) \cdot Y - \kappa^{(0)}(W) \cdot Y]}{\mathbb{E}[1 - \frac{D(1-Z)}{1-\pi_0(X)} - \frac{(1-D)Z}{\pi_0(X)}]}.
\]

In words, the mean of the product of \( Y \) and \( \kappa^{(d)} \) gives, up to a scaling, the average potential outcome \( Y^{(d)} \) of compliers when treatment is \( D = d \).

Therefore the \( \kappa \) weight approach would involve estimating the propensity score \( \hat{\pi} \) and plugging this estimate into the \( \kappa \) weight formula. Intuitively, the \( \kappa \) weight approach is like a multistage inverse propensity weighting (IPW). Impressively, it remains agnostic about the functional form of the reduced form \( \mathbb{E}[Y|Z, X] \) and first stage \( \mathbb{E}[D|Z, X] \). It is valid only when \( \hat{\pi} \) is estimated with a correctly specified and unregularized model. It is invalid if \( \hat{\pi} \) is incorrectly specified or if covariates are high dimensional and a regularized machine learning estimator is used to estimate \( \hat{\pi} \). Moreover, the inversion of \( \hat{\pi} \) can lead to numerical instability in high dimensional settings.

3.2. Doubly robust moment. Next, we introduce the moment function and doubly robust moment function formulations of LATE. For the special case of LATE, these formulations were first derived by Tan [2006] with the goal of addressing mis-specification of the regressions and the propensity score. Consider the expanded Wald formulation. Rearranging and using the notation \( V = (Y, D)' \), \( \gamma_0(Z, X) = \mathbb{E}[V|Z, X] \), and \( \begin{bmatrix} 1 & -\theta \end{bmatrix} \) as the row vector \( (1, -\theta) \), we

---

\( \text{Abadie [2003] also introduces } \kappa(W) \text{ that identifies } \mathbb{E}[g(Y, D, X)|D^{(1)} > D^{(0)}]. \text{ The parameters of interest in the main text only involve } \kappa^{(0)} \text{ and } \kappa^{(1)}, \text{ so we reserve a discussion of } \kappa(W) \text{ for Appendix C.} \)
arrive at the moment function formulation of LATE

\[
\mathbb{E}\left\{\left[1 - \theta\right] \left[\gamma_0(1, X) - \gamma_0(0, X)\right]\right\} = 0 \text{ if and only if } \theta = \theta_0.
\]

Denote the the balancing weight as

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

Tan [2006] shows that for LATE, the doubly robust moment function is

\[
\mathbb{E}\left\{\left[1 - \theta\right] \left[\gamma_0(1, X) - \gamma_0(0, X)\right] + \alpha_0(Z, X) \left[1 - \theta\right] \left[V - \gamma_0(Z, X)\right]\right\} = 0 \text{ if and only if } \theta = \theta_0.
\]

The doubly robust formulation remains valid if either the vector-valued regression \(\gamma_0\) or propensity score \(\pi_0\) is incorrectly specified.

3.3. A new synthesis with machine learning: Auto-\(\kappa\). Our key observation is the connection between the \(\kappa\) weight and the balancing weight \(\alpha_0\):

\[
\kappa^{(0)}(W) = (D - 1)\alpha_0(Z, X), \quad \kappa^{(1)}(W) = D\alpha_0(Z, X).
\]

This simple observation allows us to (i) characterize the doubly robust moment function for a broad class of complier parameters, generalizing Tan [2006] to the full class defined by Abadie [2003]; (ii) propose a new class of estimators with regularized machine learning for flexible estimation and model selection of the regression \(\hat{\gamma}\) in a way that approximates nonlinearity and heterogeneity; (iii) also use regularized machine learning for flexible estimation and model selection of the balancing weight \(\hat{\alpha}\) in a way that guarantees balance; (iv) replace the numerically unstable step of estimating and inverting \(\hat{\pi}\) with the numerically stable step of estimating \(\hat{\alpha}\) directly. As such, it allows us to combine the classic Wald and \(\kappa\) weight formulations while also updating them to incorporate machine learning.

We propose an Auto-\(\kappa\) estimation procedure for the complier parameter \(\hat{\theta}\) as follows. Partition the sample into, say, 5 folds \(\{I_\ell\}_{\ell=1:5}\).
(1) For each fold $\ell$, estimate $\hat{\gamma}_{-\ell}$ and $\hat{\alpha}_{-\ell}$ from observations not in $I_\ell$.

(2) Estimate $\hat{\theta}$ as the solution to

$$\frac{1}{n} \sum_{\ell=1}^{5} \sum_{i \in I_\ell} \left[ 1 - \theta \right] [\hat{\gamma}_{-\ell}(1, X_i) - \hat{\gamma}_{-\ell}(0, X_i)] + \hat{\alpha}_{-\ell}(Z_i, X_i) \left[ 1 - \theta \right] [V_i - \hat{\gamma}_{-\ell}(Z_i, X_i)] = 0.$$ 

We prove that our Auto-$\kappa$ estimator is $\sqrt{n}$ consistent, asymptotically normal, and semiparametrically efficient. Moreover, it has favorable finite sample properties that distinguish it from asymptotically equivalent estimators. We build on this main result to propose (i) a hypothesis test to evaluate whether two different instruments induce the same subpopulation of compliers and (ii) simultaneous confidence bands for complier counterfactual outcome distributions over a fixed grid.

4. Simulations and Program Evaluation

4.1. Simulations. We compare the performance of our proposed Auto-$\kappa$ estimator with $\kappa$ weighting [Abadie, 2003] and the original DML with explicit propensity scores [Chernozhukov et al., 2018a] in simulations. We focus on counterfactual distributions as our choice of complier parameter $\theta_0$ over the grid $U$ specified on the horizontal axis of Figure 1.

We consider a simple simulation design where $Y$ is a continuous outcome, $D$ is a binary treatment, $Z$ is a binary instrumental variable, and $X$ is a continuous covariate. Appendix B includes more details on the simulation design. Each simulation consists of $n = 1000$ observations. We conduct 1000 such simulations and implement each estimator as follows.

(1) $\kappa$ weight. We estimate the propensity score $\hat{\pi}$ by logistic regression, which we then use in the weights $\hat{\kappa}^{(0)}(W), \hat{\kappa}^{(1)}(W)$ and subsequently the estimator $\hat{\theta}$.

(2) DML. We use 5 folds. We estimate the propensity score $\hat{\pi}$ by $\ell_1$ regularized logistic regression, using a high dimensional dictionary of basis functions $b(X)$ consisting of fourth order polynomials of $X$. We estimate $\hat{\gamma}$ by Lasso as described in Section 6, using a high dimensional dictionary of basis functions $b(Z, X)$ consisting of fourth order polynomials of $X$ and interactions between $Z$ and the polynomials.
(3) **Auto-κ.** The key difference from DML is that instead of estimating the propensity score, we directly estimate the balancing weight $\hat{\alpha}$ by Lasso as described in Section 6, using a high dimensional dictionary of basis functions $b(Z, X)$ consisting of fourth order polynomials of $X$ and interactions between $Z$ and the polynomials. Subsequently, we estimate $\hat{\theta}$ and construct simultaneous confidence bands by steps outlined in Section 6.

Since the true propensity scores $\pi_0(X)$ are highly nonlinear, we expect $\kappa$ weighting and DML to encounter issues of numerical instability. Furthermore, $\kappa$ weighting might not be as efficient as the DML and Auto-$\kappa$ estimators, which have the semiparametrically efficient asymptotic variance.

For each value in the grid $\mathcal{U}$, Table 1 presents the bias and the root mean square error (RMSE) of each estimator across simulation draws. The last column averages the performance across grid points. Figure 1 visualizes the median as well as the 10% and 90% quantiles across simulation draws. Auto-$\kappa$ outperforms DML by a large margin due to numerical stability. Even though Auto-$\kappa$ uses regularized machine learning to estimate $(\hat{\gamma}, \hat{\alpha})$, regularization bias does not percolate into bias for estimating the counterfactual distribution due to the doubly robust moment function. In terms of efficiency, Auto-$\kappa$ substantially outperforms $\kappa$ weighting. Lastly, the simultaneous confidence bands based on the Auto-$\kappa$ estimator have coverage probability of 98.4% for the CDF of $Y^{(0)}$ and 93.6% for the CDF of $Y^{(1)}$, which are quite close to the nominal level of 95%.

Numerical instability from inverting $\hat{\pi}$ is a known issue. In practice, researchers might try trimming and censoring. Trimming means excluding observations for which $\hat{\pi}$ is extreme [Belloni et al., 2017]. Censoring means imposing bounds on $\hat{\pi}$ for such observations.

**Remark 4.1** (Trimming and censoring). *Auto-$\kappa$ without trimming or censoring outperforms $\kappa$ weighting and DML even with trimming and censoring in this simulation design. Compare Figure 1 (no pre-processing) in the main text with Figures A.1 (trimming) and A.3 (censoring) in Appendix B to see this phenomenon.*
This property is convenient, since ad hoc trimming and censoring have limited theoretical justification [Crump et al., 2009]. In Appendix B, we consider an alternative simulation design where the true propensity scores are more uniformly distributed and all three estimators perform well.

4.2. Effect of 401(k) on net financial assets. After validating the favorable bias, RMSE, and coverage properties of our Auto-$\kappa$ estimator, we apply it to real world data to estimate the counterfactual distributions of employee net financial assets with and without 401(k) participation, using 401(k) eligibility as the instrument. We follow the identification strategy of Poterba and Venti [1994], Poterba et al. [1995]. The authors assume that when 401(k) was introduced, workers ignored whether different jobs 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 Section 5 are thus satisfied. Since ineligibility implies no participation, the monotonicity condition of Section 5 is satisfied by construction.

We use data from the 1991 US Survey of Income and Program Participation [Abadie, 2003, Belloni et al., 2017, Chernozhukov and Hansen, 2004, 2005, Ogburn et al., 2015]. 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. We follow Chernozhukov et al. [2018a] and control for covariates $X$ that include age, income, years of education, family size, marital status, two earner status, benefit pension status, IRA participation, and home ownership.

The data include $n = 9915$ observations. We follow Chernozhukov et al. [2018a] in the choice of the dictionary $b(Z, X)$, which includes polynomials of continuous covariates, interactions among all covariates, and interactions between covariates and instrument. Applying
the dictionary on the covariates gives a high dimensional setting with \( p = 277 \). For the choice of grid, we take \( \mathcal{U} \) to be the 19 values \( y \) corresponding to the 5\textsuperscript{th} to 95\textsuperscript{th} percentiles of \( \mathcal{Y} \) in increments of 5 percentiles. To implement DML, Chernozhukov et al. [2018a] drop observations in the group with \( Z = 0 \) with estimated propensity scores \( \hat{\pi} \) that exceed the maximum and minimum propensity scores in the group with \( Z = 1 \)—a form of trimming. We implement Auto-\( \kappa \) to estimate the effect of 401(k) participation on the distribution of net financial assets, controlling for high dimensional covariates. Importantly, we do not conduct trimming for Auto-\( \kappa \), and nonetheless our results are numerically stable.

Figure 3 visualizes point estimates and simultaneous 95\% confidence bands based on the Auto-\( \kappa \) estimator. We estimate \( \hat{\alpha} \) by Lasso as described in Section 6. We estimate \( \hat{\gamma} \) by either (i) Lasso as described in Section 6 or (ii) neural network with a single hidden layer of eight neurons as in Chernozhukov et al. [2018a]. Subsequently, we estimate \( \hat{\theta} \) and construct simultaneous confidence bands by steps outlined in Section 6. We find that 401(k) participation significantly shifts out the distribution of NFA, consistent with results reported in Belloni et al. [2017]. The impact is heterogenous: 401(k) participation has a small impact on NFA at low quantiles while appearing to have a larger impact at higher quantiles. However, the impact is not statistically significant for the highest quantiles. From the figure, it is clear that an analyst would reject the null hypothesis that 401(k) participation has no effect and also reject the null hypothesis of a constant treatment effect. Auto-\( \kappa \) gives robust performance with high dimensional covariates, yielding similar results using Lasso or a neural network to estimate \( \hat{\gamma} \).

5. Framework

5.1. Complier parameters. Suppose we are interested in the effect of a binary treatment \( D \in \{0, 1\} \) on a continuous outcome \( Y \in \mathcal{Y} \subset \mathbb{R} \). There is a binary instrumental variable \( Z \in \{0, 1\} \) available, as well as a potentially high dimensional covariate \( X \in \mathcal{X} \subset \mathbb{R}^{\text{dim}(X)} \).
We observe \( n \) i.i.d. observations \( \{W_i\}_{i=1}^{n} \) where \( W = (Y, D, Z, X)' \) concatenate the random variables. Wherever possible, we suppress index \( i \) to lighten notation.

Following the notation of Angrist et al. [1996], 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 \( D(1) > D(0) \). In this paper, we focus on the classic instrumental variable model for treatment effects summarized by the following assumption.

**Assumption 1** (Instrumental variable identification). Assume

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

Independence states that the instrument \( Z \) is as good as randomly assigned conditional on covariates \( X \). Exclusion imposes that the instrument \( Z \) only affects the outcome \( Y \) via the treatment \( D \). We can therefore simplify notation: \( Y^{(d)} = Y^{(1,d)} = Y^{(0,d)} \). Overlap ensures that there are no covariate values for which the instrument is deterministic. Monotonicity rules out the possibility of defiers: individuals who will always pursue an opposite treatment status from their instrument assignment.

For this instrumental variable model, we focus on the following complier parameters in the main text.

**Definition 5.1** (Complier parameters). Consider the following parameters of the complier subpopulation:

1. LATE is \( \theta_0 = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid D^{(1)} > D^{(0)}] \).
2. Average complier characteristics are \( \theta_0 = \mathbb{E}[f(X) \mid D^{(1)} > D^{(0)}] \) for any finite-dimensional measurable real function \( f(\cdot) \) of covariate \( X \) such that \( \mathbb{E}[f_j(X)] < \infty \).
(3) Complier counterfactual outcome distributions are $\theta_0 = \{\theta^y_0\}_{y \in \mathcal{U}}$ where

$$\theta^y_0 = \begin{bmatrix} \beta^y_0 \\ \delta^y_0 \end{bmatrix} = \begin{bmatrix} \mathbb{P}(Y(0) \leq y | D(1) > D(0)) \\ \mathbb{P}(Y(1) \leq y | D(1) > D(0)) \end{bmatrix}$$

and $\mathcal{U} \subset \mathcal{Y}$ is a fixed grid of finite dimension.

LATE is an extremely popular estimand in empirical economics. Average complier characteristics help to assess the external validity of any study that uses instrumental variable identification since LATE is defined for the subpopulation of compliers only. Counterfactual outcome distributions are particularly important in welfare analysis of social policies as we have seen in Section 4.

**Remark 5.1 (General class of complier parameters).** In Appendix C, we consider a more general class of complier parameters. In particular, we consider complier parameters implicitly defined by any of the following expressions:

1. $\mathbb{E}[g(Y(0), X, \theta) | D(1) > D(0)] = 0$ if and only if $\theta = \theta_0$;
2. $\mathbb{E}[g(Y(1), X, \theta) | D(1) > D(0)] = 0$ if and only if $\theta = \theta_0$;
3. $\mathbb{E}[g(Y, D, X, \theta) | D(1) > D(0)] = 0$ if and only if $\theta = \theta_0$.

This is the full class studied by Abadie [2003].

### 5.2. Doubly robust moment function.

In Section 3, we previewed the doubly robust moment function for LATE, and alluded to its robustness to mis-specification. We now give a more careful and more general exposition that includes a broad class of complier parameters. In particular, we argue that a broad class of complier parameters share a common structure.

We first show that any complier parameter $\theta_0$ in Definition 5.1 admits an expanded Wald formulation. In other words, for some appropriate definition of the vector $V = (V_1, ..., V_J)'$ and accordingly of the vector valued regression $\gamma_0(Z, X) = \mathbb{E}[V | Z, X]$,

$$\mathbb{E} \{ A(\theta)[\gamma_0(1, X) - \gamma_0(0, X)] \} = 0 \text{ if and only if } \theta = \theta_0,$$
where $A(\theta)$ is a matrix whose initial columns are the identity matrix and whose final column is $-\theta$.

Next, we show that this moment function can be made doubly robust by including an additional term. The doubly robust moment is instead

$$
\mathbb{E} \{ A(\theta)[\gamma_0(1, X) - \gamma_0(0, X)] + \alpha_0(Z, X)A(\theta)[V - \gamma_0(Z, X)] \} = 0 \text{ if and only if } \theta = \theta_0.
$$

Recall from Section 3 that $\alpha_0$ is the balancing weight defined as

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

and also that the $\kappa$ weights are

$$
\kappa^{(0)}(W) = (D - 1)\alpha_0(Z, X), \quad \kappa^{(1)}(W) = D\alpha_0(Z, X).
$$

Therefore the doubly robust moment formulation is, in general, a combination of the expanded Wald formulation and the $\kappa$ weight formulation. We now state the formal result, which nests Tan [2006] as a special case for LATE.

**Theorem 5.2 (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

$$
\psi(W, \gamma, \alpha, \theta) = A(\theta)[\gamma(1, X) - \gamma(0, X)] + \alpha(Z, X)A(\theta)[V - \gamma(Z, X)]
$$

where

1. For LATE, we set $V = (Y, D)'$ and $A(\theta) = \begin{pmatrix} 1 & -\theta \end{pmatrix}$ [Tan, 2006].
2. For complier characteristics, we set $V = (Df(X)', D)'$ and $A(\theta) = \begin{pmatrix} 1 & -\theta \end{pmatrix}$.
3. For complier counterfactual distributions, we set $V^y = ((D - 1)1_{Y \leq y}, D1_{Y \leq y}, D)'$ and $A(\theta^y) = \begin{pmatrix} 1 & 0 & -\beta^y \\ 0 & 1 & -\delta^y \end{pmatrix}$. 


See Appendix C for the proof. Theorem 5.2 provides a single argument to characterize the doubly robust moment functions of many complier parameters. In particular, the doubly robust moment function of average complier characteristics is new and useful for our proposed hypothesis test below.

**Remark 5.2** (General class of complier parameters). In Theorem C.1 in Appendix C, we prove a more abstract result for any complier parameter implicitly defined by the expressions in Remark 5.1, i.e. the full class studied by Abadie [2003]. The generality of this characterization appears to be new. It answers the open question posed by Sloczyński and Wooldridge [2018] of how to characterize the doubly robust moment function for any complier parameter.

Doubly robust moment functions confer many benefits for estimation. First, as the name suggests, they guard against mis-specification. Indeed, one can verify

\[
\mathbb{E} \left\{ A(\theta_0)[\gamma(1, X) - \gamma(0, X)] + \alpha_0(Z, X)A(\theta_0)[V - \gamma(Z, X)] \right\} = 0, \quad \forall \gamma \text{ s.t. } \mathbb{E}[\gamma(Z, X)^2] < \infty \\
\mathbb{E} \left\{ A(\theta_0)[\gamma_0(1, X) - \gamma_0(0, X)] + \alpha(Z, X)A(\theta_0)[V - \gamma_0(Z, X)] \right\} = 0, \quad \forall \alpha \text{ s.t. } \mathbb{E}[\alpha(Z, X)^2] < \infty.
\]

We see that the doubly robust moment function remains valid as an estimating equation if either \( \hat{\gamma} \) or \( \hat{\alpha} \) is incorrectly specified.

The debiased machine learning paradigm [Chernozhukov et al., 2016, 2018a] emphasizes another benefit of doubly robust moment functions. When covariates are high dimensional, a flexible, regularized machine learning estimator \( \hat{\gamma} \) that can perform model selection is desirable. When using a regularized (and therefore biased) machine learning estimator \( \hat{\gamma} \), the introduction of the additional object \( \hat{\alpha} \) serves to debias \( \hat{\theta} \), the estimator for the ultimate causal parameter of interest. We prove that \( \sqrt{n} \) consistency of \( \hat{\theta} \) is possible despite slower than \( \sqrt{n} \) rates on \( \|\hat{\gamma} - \gamma_0\| \) and \( \|\hat{\alpha} - \alpha_0\| \) as long as \( \|\hat{\gamma} - \gamma_0\| \cdot \|\hat{\alpha} - \alpha_0\| = o_p(n^{-\frac{1}{2}}) \) (where we denote \( \|f\| = \sqrt{\mathbb{E}[f(W)^2]} \)). In Section 7, we extend debiased machine learning theory to a broad class of complier parameters in this way.
Since the double robustness property and product rate condition are ultimately in terms of $\hat{\alpha}$ rather than $\hat{\pi}$, we propose estimating $\hat{\alpha}$ directly, building on recent insights of Chernozhukov et al. [2018b,c]. Whereas Chernozhukov et al. [2018b,c] propose estimators for parameters of the full population identified by selection on observables, we propose estimators for parameters of the complier subpopulation identified by instrumental variables. We formalize the finite sample balancing property conferred by this technique in Section 6.

6. Estimation

6.1. Automatic debiased machine learning. Debiased machine learning [Chernozhukov et al., 2016, 2018a] is a meta estimation procedure that combines doubly robust moment functions [Newey, 1994, Robins and Rotnitzky, 1995] with sample splitting [Bickel, 1982, Schick, 1986]. Given the doubly robust moment function of some causal parameter of interest as well as machine learning estimators ($\hat{\gamma}, \hat{\alpha}$) for its nonparametric components, DML generates an estimator of the causal parameter in the following way.

**Estimator 6.1 (Debiased 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

$$\frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta) \bigg|_{\theta = \hat{\theta}} = 0.$$  

In Theorem 5.2, we characterize the doubly robust moment function $\psi$ for complier parameters. What remains is an account of how to estimate the vector valued regression $\hat{\gamma}$ and the balancing weight $\hat{\alpha}$. Our theoretical results are agnostic about the choice of the vector valued regression $\hat{\gamma}$ as long as it satisfies the rate condition in Assumption 6 below. It could be, for example, a neural network or a Lasso regression.

For the balancing weight estimator $\hat{\alpha}$, we propose the following regularized estimator, adapting the so-called “regularized Riesz representer” of Chernozhukov et al. [2018b,c]. In the following discussion, we denote by $|\cdot|_q$ the $\ell_q$ norm of a vector, and we denote by $\|\cdot\|$ the
norm $\|V_j\| = \sqrt{\mathbb{E}[V_j^2]}$. Consider the projection of the balancing weight $\alpha_0(Z, X)$ onto the $p$-dimensional dictionary of basis functions $b(Z, X)$. A high dimensional dictionary $b$ allows for flexible approximation of $\alpha_0$ [Belloni et al., 2014]. With $\ell_1$-regularization, the objective becomes

$$\rho_L = \arg\min_{\rho} \|\alpha_0(Z, X) - \rho'b(Z, X)\|^2 + 2\lambda_L|\rho|_1$$

where $\lambda_L = \sqrt{\frac{\ln p}{n}}$ is a theoretical regularization level.

The main theoretical property of the balancing weight $\alpha_0$ is that it represents the functional $\gamma \mapsto \mathbb{E}[\gamma(1, X) - \gamma(0, X)]$ in the sense that

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

Applying this result to the dictionary of basis functions,

$$\mathbb{E}[b(1, X) - b(0, X)] = \mathbb{E}[\alpha_0(Z, X)b(Z, X)], \quad \forall b \text{ s.t. } \|b_j(Z, X)\| < \infty.$$  

Expanding the square, ignoring terms without $\rho$, and using this logic,

$$\rho_L = \arg\min_{\rho} -2\rho'\mathbb{E}[b(1, X) - b(0, X)] + \rho'\mathbb{E}[b(Z, X)b(Z, X)']\rho + 2\lambda_L|\rho|_1.$$  

The empirical analogue to the above expression yields an estimator of $\hat{\rho}$. The estimator is therefore $\hat{\alpha}(Z, X) = \hat{\rho}'b(Z, X)$, which we now formalize.

**Estimator 6.2** (Regularized balancing weight). Based on the observations in $I_{-\ell}$,

1. Calculate $p \times p$ matrix $\hat{G}_{-\ell} = \frac{1}{n-\ell} \sum_{i \in I_{-\ell}} b(Z_i, X_i)b(Z_i, X_i)'$,

2. Calculate $p \times 1$ vector $\hat{M}_{-\ell} = \frac{1}{n-\ell} \sum_{i \in I_{-\ell}} b(1, X_i) - b(0, X_i)$,

3. Set $\hat{\alpha}_{-\ell}(Z, X) = b(Z, X)'\hat{\rho}_{-\ell}$ where $\hat{\rho}_{-\ell} = \arg\min_{\rho} \rho'\hat{G}_{-\ell}\rho - 2\rho'\hat{M}_{-\ell} + 2\lambda_n|\rho|_1$.

In Appendix D, we provide and justify an iterative tuning procedure for data-driven regularization parameter $\lambda_n$. We refer to our proposed estimator, which combines the doubly robust moment function from Theorem 5.2 with the meta procedure in Estimator 6.1 and
the regularized balancing weights in Estimator 6.2, as automatic $\kappa$ weighting (Auto-$\kappa$) for complier parameters.

Likewise, we can project the regression $\gamma_0(Z, X)$ onto the $p$-dimensional dictionary of basis functions $b(Z, X)$ using the functional $b \mapsto E[b(Z, X)V']$. Our theoretical results are agnostic about the choice of estimator $\hat{\gamma}$; it may be this estimator or any other machine learning estimator satisfying the rate condition specified in Assumption 6 below.

6.2. Extensions: Hypothesis test and simultaneous confidence band. Suppose we wish to test the null hypothesis that two different instruments $Z_1$ and $Z_2$ induce the same complier subpopulation and hence identify the same causal parameters. Denote by $\hat{\theta}_1$ and $\hat{\theta}_2$ the Auto-$\kappa$ estimators for complier characteristics using the different instruments $Z_1$ and $Z_2$, respectively. The following procedure allows us to test this hypothesis from some estimator $\hat{C}$ for the asymptotic variance $C$ of $\hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2')'$.

**Estimator 6.3** (Hypothesis test for difference of complier subpopulations). *Given $\hat{C}$,*

1. Calculate the statistic $T = [\hat{\theta}_1 - \hat{\theta}_2]' \left\{ R \hat{C} R' \right\}^{-1} [\hat{\theta}_1 - \hat{\theta}_2]$ where $R = \begin{bmatrix} I & -I \end{bmatrix}$.
2. Compute the value $c_a$ as the $(1 - a)$-quantile of $\chi^2(1)$.
3. Reject the null hypothesis if $T > c_a$.

Suppose we wish to form a simultaneous confidence band for the components of $\hat{\theta}$, which may be the complier counterfactual outcome distribution based on a finite grid $\mathcal{U} \subseteq \mathcal{Y}$. The following procedure allows us to do so from some estimator $\hat{C}$ for the asymptotic variance $C$ of $\hat{\theta}$. Let $\hat{S} = diag(\hat{C})$ and $S = diag(C)$ collect the diagonal elements of these matrices.

**Estimator 6.4** (Simultaneous confidence band). *Given $\hat{C}$,*

1. Calculate $\hat{\Sigma} = \hat{S}^{-1/2}\hat{C}\hat{S}^{-1/2}$.
2. Sample $Q \overset{i.i.d.}{\sim} N(0, \hat{\Sigma})$ and compute the value $\hat{c}_a$ as the $(1 - a)$-quantile of sampled $|Q|_{\infty}$.
3. Form the confidence band $[l_j, u_j] = \left[ \hat{\theta}_j - \hat{c}_a \sqrt{\hat{C}_{jj} / n}, \hat{\theta}_j + \hat{c}_a \sqrt{\hat{C}_{jj} / n} \right]$ where $\hat{C}_{jj}$ is the diagonal entry of $\hat{C}$ corresponding to $j$-th element $\hat{\theta}_j$ of $\theta$. 
6.3. Finite sample balance. Next, we discuss how Auto-κ automatically attenuates the influence of outliers and guarantees balance. Recall that in Estimator 6.1, the asymptotic influence \( \psi(W_i, \gamma_0, \alpha_0, \theta_0) \) of observation \( W_i \) in fold \( \ell \) is estimated by empirical influence \( \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \hat{\theta}) \).

In original DML, the propensity score \( \hat{\pi}_{-\ell} \) is explicitly estimated to serve as a component of the balancing weight \( \hat{\alpha}_{-\ell}(Z_i, X_i) = Z_i \hat{\pi}_{-\ell}(X_i) - \frac{1-Z_i}{1-\hat{\pi}_{-\ell}(X_i)} \). In the finite sample, \( \hat{\pi}_{-\ell}(X_i) \) can be close to 0 or 1, causing the empirical influence in original DML \( \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \hat{\theta}) \) 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 selection 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-κ automatically addresses outliers and finite sample imbalance in three ways. First, Auto-κ considers estimation of \( \alpha_0 \) rather than \( \pi_0 \). \( \hat{\alpha}_{-\ell} \), estimated by Estimator 6.2, 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-κ confers a finite sample guarantee of balance on average. Consider the choice of dictionary of basis functions \( b \) 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 6.1** (Balance). Auto-κ with regularization \( \lambda_n \) yields, over any realization of data and for all \( n \)

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

\[
\left| \frac{1}{n-n_\ell} \sum_{i \in I-\ell} q(X_i) - \frac{1}{n-n_\ell} \sum_{i \in I-\ell} q(X_i)(1-Z_i) \cdot \hat{\omega}_{-\ell,i}^{(z=0)} \right|_\infty \leq \lambda_n;
\]
where \( \hat{\omega}_{-\ell,i}^{(z=1)} = q(X_i)'\hat{\rho}_{-\ell}^{(z=1)} \) and \( \hat{\omega}_{-\ell,i}^{(z=0)} = q(X_i)'\hat{\rho}_{-\ell}^{(z=0)} \).

See Appendix D for the proof. Proposition 6.1 shows that the weights \( \{\hat{\omega}_{\ell,i}^{(z=1)}, \hat{\omega}_{\ell,i}^{(z=0)}\} \) serve to approximately balance the overall sample average with the sample average of the group that is assigned the instrument \((Z = 1)\) and the sample average of the group that is not assigned the instrument \((Z = 0)\), across each basis function of the dictionary \(q\). The result is similar to the balancing conditions of Zubizarreta [2015] and Athey et al. [2018]. Auto-\(\kappa\) automatically calculates these weights. Note that this property does not hold for \(\kappa\) weight and original DML estimators, due to the inverse propensity score.

Third, Auto-\(\kappa\) is a Lasso type estimator that delivers a sparse estimate \(\hat{\rho}_{-\ell}\). As such, it automatically determines which basis functions in the dictionary \(q\) to use in the calculation of influence \(\psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \hat{\theta})\). It discerns that, in the finite sample, some basis functions are best ignored, and those basis functions may be different for the group assigned the instrument and the group not assigned the instrument. 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-\(\kappa\) can zero out specific components of \(q(x)\) where \(q'\) is imbalanced. In this sense, Auto-\(\kappa\) learns which transformations of covariates to match on—a type of model selection useful for causal inference. This property is a key reason why our estimator performs well in selecting covariates and their transformations under IV identification, which is one of the three ways that machine learning may be useful to labor economics outlined by Angrist and Frandsen [2019].

6.4. Finite sample orthogonality. By construction, the doubly robust moment function \(\psi\) ensures asymptotic invariance of \(\hat{\theta}\) to estimation error of either \(\hat{\gamma}\) or \(\hat{\alpha}\). As we will see, any consistent estimators \((\hat{\alpha}, \hat{\gamma})\) that satisfy the DML product condition \(\|\hat{\alpha} - \alpha_0\| \cdot \|\hat{\gamma} - \gamma_0\| = o_p(n^{-\frac{1}{2}})\) will deliver consistent and asymptotically normal estimators \(\hat{\theta}\) via Estimator 6.1; many approaches are asymptotically equivalent. We present a finite sample property of Auto-\(\kappa\), as a theoretical explanation of its numerical stability relative to previous approaches.
For this discussion, we use Estimator 6.2 to estimate both the balancing weight $\hat{\alpha}(z, x) = \hat{\rho}' b(z, x)$ and the regression $\hat{\gamma}(z, x) = \hat{\beta}' b(z, x)$. For simplicity, consider LATE so that $\theta_0$ is scalar. As a consequence, the notation simplifies: $\psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta) = \psi(W_i, \hat{\beta}_{-\ell}, \hat{\rho}_{-\ell}, \theta)$.

We have the following result.

**Proposition 6.2 (Orthogonality).** Auto-$\kappa$ with regularization $\lambda_n$ yields, over any realization of data and for all $n$

$$\left| \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} \frac{\partial}{\partial \beta} \psi(W_i, \hat{\beta}_{-\ell}, \hat{\rho}_{-\ell}, \theta_0) \right|_\infty \leq \{1 + |\theta_0|\} \lambda_n;$$

$$\left| \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} \frac{\partial}{\partial \rho} \psi(W_i, \hat{\beta}_{-\ell}, \hat{\rho}_{-\ell}, \theta_0) \right|_\infty \leq \{1 + |\theta_0|\} \lambda_n.$$

See Appendix D for the proof. This property differs from Neyman orthogonality in two important respects. First, it is a finite sample property rather than a population property. As such, it can distinguish Auto-$\kappa$ from asymptotically equivalent estimators. Second, the property holds approximately, with tolerance that vanishes at the rate of $\lambda_n$. Note that this property does not hold for $\kappa$ weight and original DML estimators, due to the inverse propensity score.

Proposition 6.2 connects our work with the literature on bias reduced doubly robust estimation, which seeks to improve the stability of balancing weights. Vermeulen and Vansteelandt [2015] propose parametric estimators of $(\alpha_0, \gamma_0)$ based on an exact version of the finite sample orthogonality property. The authors prove that their estimators locally minimize the squared first order asymptotic bias; the same is true about Auto-$\kappa$, up to tolerance $\lambda_n$. Unlike Vermeulen and Vansteelandt [2015], we consider instrumental variable identification, high dimensional covariates, and machine learning estimators.
7. Consistency and asymptotic normality

7.1. Regression and balancing weight. In this section, we present asymptotic statistical guarantees for \( \text{Auto-}\kappa \). We generalize the main results of Chernozhukov et al. [2018c] to the instrumental variable setting. To begin, we analyze our estimators \((\hat{\gamma}, \hat{\alpha})\) of the nonparametric functions \((\gamma_0, \alpha_0)\). First, we place a weak assumption on the dictionary of basis functions \(b\).

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

Next, 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 3** (Dense balancing weight). Assume there exist some \( \rho_n \in \mathbb{R}^p \) and \( C < \infty \) such that \( |\rho_n|_1 \leq C \) and \( \|\alpha_0 - b'\rho_n\|^2 = O\left(\sqrt{\frac{\ln p}{n}}\right) \).\(^3\)

Assumption 3 is a statement about the quality of approximation of the balancing weight \(\alpha_0\) by dictionary \(b\). It is satisfied if, for example, \(\alpha_0\) is a linear combination of \(b\).

**Assumption 4** (Sparse balancing weight). Assume

1. There exist \( C > 1, \xi > 0 \) such that for all \( \bar{s} \leq C\left(\frac{\ln p}{n}\right)^{-\frac{1}{3+2\xi}} \), there exists some \( \bar{\rho} \in \mathbb{R}^p \) with \( |\bar{\rho}|_1 \leq C \) and \( \bar{s} \) nonzero elements such that \( \|\alpha_0 - b'\bar{\rho}\|^2 \leq C(\bar{s})^{-\xi} \).
2. \( G = \mathbb{E}[b(Z, X)b(Z, X)'] \) is nonsingular with largest eigenvalue uniformly bounded in \( n \).
3. Denote \( J_\rho = \text{support}(\rho) \). There exists \( k > 3 \) such that for \( \rho \in \{\rho_L, \bar{\rho}\} \)

\[
RE(k) = \inf_{\delta : \delta \neq 0, \sum_{j \in J_\rho} |\delta_j| \leq k, \sum_{j \in J_\rho} |\delta_j|} \frac{\delta' G \delta}{\sum_{j \in J_\rho} \delta_j^2} > 0.
\]

\(^2\)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.

\(^3\)Likewise, it is possible to allow the bound on \( |\rho_n|_1 \) to be a sequence \(B_n\) that increases in \( n \).
\[(4) \ln p = O(\ln n).\]

Assumption 4 is a statement about the quality of approximation of the balancing weight \(\alpha_0\) by a subset of dictionary \(b\). It is satisfied if, for example, \(\alpha_0\) is sparse or approximately sparse [Chernozhukov et al., 2018c]. The uniform bound on the largest eigenvalue of \(G\) rules out the possibility that \(G\) is an equal correlation matrix. \(RE\) is the population version of the restricted eigenvalue condition [Bickel et al., 2009]. It generalizes the familiar notion of “no multicollinearity” to the high dimensional setting. The final condition \(\ln p = O(\ln n)\) rules out the possibility that \(p = \exp(n)\); dimension cannot grow too much faster than sample size.

We adapt convergence guarantees from Chernozhukov et al. [2018c] for the balancing weight estimator \(\hat{\alpha}\) in Estimator 6.2. 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 \(\sqrt{\frac{\ln p}{n}}\).

**Assumption 5 (Regularization).** \(\lambda_n = a_n \sqrt{\frac{\ln p}{n}}\) for some \(a_n \to \infty\).

For example, one could set \(a_n = \ln(\ln(n))\) [Chatterjee and Jafarov, 2015]. In Appendix D, we provide and justify an iterative tuning procedure to determine data-driven regularization parameter \(\lambda_n\). The guarantees are as follows.

**Lemma 7.1 (Dense balancing weight rate).** Under Assumptions 1, 2, 3, and 5,

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

**Lemma 7.2 (Sparse balancing weight rate).** Under Assumptions 1, 2, 4, and 5,

\[
\|\hat{\alpha} - \alpha_0\|^2 = O_p \left( a_n^2 \left( \frac{\ln p}{n} \right)^{\frac{2\xi}{1+2\xi}} \right), \quad |\hat{\rho}|_1 = O_p(1).
\]
See Appendix E for the proofs. Whereas Lemma 7.1 does not require an explicit sparsity condition, Lemma 7.2 does. When $\xi > \frac{1}{2}$, the rate in Lemma 7.2 is faster than the rate in Lemma 7.1 for $a_n$ growing slowly enough. Interpreting the rate in Lemma 7.2, $n^{-\frac{2\xi}{1+2\xi}}$ is the well-known rate of convergence if the identity of the nonzero components of $\hat{p}$ were known. The fact that their identity is unknown introduces a cost of $(\ln p)^{\frac{2\xi}{1+2\xi}}$. The cost $a_n^2$ can be made arbitrarily small.

We place a rate assumption on the machine learning 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 machine learning estimators. In our empirical application, we choose neural network and Lasso. Farrell et al. [2021] provide a rate for the former, while Lemmas 7.1 and 7.2 provide rates for the latter (using the functional $b \mapsto \mathbb{E}[b(Z,X)V']$ instead).

**Assumption 6** (Regression rate). $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-d_\gamma})$ where

1. In the dense balancing weight regime, $d_\gamma \in \left( \frac{1}{4}, \frac{1}{2} \right)$;
2. In the sparse balancing weight 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 7.3** (Product condition). Suppose the conditions of Lemma 7.1 or Lemma 7.2 hold as well as Assumption 6. Then $\|\hat{\alpha} - \alpha_0\| \cdot \|\hat{\gamma} - \gamma_0\| = o_p(n^{-\frac{1}{2}})$.

The product condition in Corollary 7.3 formalizes the trade-off in estimation error permitted in estimating $(\gamma_0, \alpha_0)$. In particular, faster convergence of $\hat{\alpha}$ permits slower convergence
of \( \hat{\gamma} \). Prior information about the balancing weight \( \alpha_0 \) 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 statistical guarantees for complier parameters.\(^4\)

7.2. **Complier parameter.** We now present our second theorem. We prove our Auto-\( \kappa \) estimator for complier parameters is consistent, asymptotically normal, and semiparametrically efficient. We build on the theoretical foundations in Chernozhukov et al. [2016] to generalize the main result in Chernozhukov et al. [2018c]. We consider a complier parameter implicitly characterized by a moment function involving a vector valued regression rather than a total population parameter explicitly characterized as a functional of a scalar valued regression. To do so, we place weak regularity conditions on the propensity score \( \pi_0 \), conditional variance \( \text{var}(V|Z,X) \), Jacobian \( J \), and target parameter space \( \Theta \).

**Assumption 7** (Regularity conditions for complier parameter estimation). The following conditions hold:

1. \( \pi_0(X) \in (\bar{c}, 1 - \bar{c}) \) for some \( \bar{c} > 0 \) uniformly over the support of \( X \);
2. \( \text{var}(V|Z,X) \) is bounded uniformly over the support of \( (Z,X) \);
3. Jacobian \( J = \mathbb{E} \left[ \frac{\partial \psi(W,\gamma_0,\alpha_0,\theta)}{\partial \theta} \right]_{\theta=\theta_0} \) is nonsingular;
4. \( \theta_0, \hat{\theta} \in \Theta \), a compact parameter space.

**Theorem 7.4** (Consistency and asymptotic normality). Suppose the conditions of Corollary 7.3 hold as well as Assumption 7. Then \( \hat{\theta} \xrightarrow{p} \theta_0 \), \( \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,C) \), and \( \hat{C} \xrightarrow{p} C \).

\(^4\)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 semiparametric lower bound.
where

\[
J = \mathbb{E}\left[ \frac{\partial \psi_0(W)}{\partial \theta} \right], \quad \hat{J} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \frac{\partial \hat{\psi}_i(\hat{\theta})}{\partial \theta}, \quad \Omega = \mathbb{E}[\psi_0(W)\psi_0(W)^\prime], \quad \hat{\Omega} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i(\hat{\theta})\hat{\psi}_i(\hat{\theta})^\prime
\]

\[
C = J^{-1}\Omega J^{-1}, \quad \hat{C} = \hat{J}^{-1}\hat{\Omega}\hat{J}^{-1}, \quad \psi_0(W) = \psi(W, \gamma_0, \alpha_0, \theta_0), \quad \hat{\psi}_i(\hat{\theta}) = \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \hat{\theta}).
\]

See Appendix E for the proof, which we divide into two parts: lemmas from previous work, and original arguments necessary for our richer setting. For the complier parameters considered in Definition 5.1, the Auto-κ estimator \( \hat{\theta} \) achieves semiparametric efficiency because the doubly robust moment functions derived in Theorem 5.2 coincide with the semiparametrically efficient score [Chernozhukov et al., 2016, Farrell, 2015, Hahn, 1998]. Moreover, it follows that the proposed hypothesis test is valid.

**Corollary 7.5** (Hypothesis test for difference of complier subpopulations). Under the conditions of Theorem 7.4, the hypothesis test in Estimator 6.3 falsely rejects the null hypothesis with probability approaching the nominal level, i.e., \( \lim_{n \to \infty} \mathbb{P}(T > c_a|H_0) = 1 - a \).

Finally, we present the third theorem of this paper: validity of the bootstrap procedure presented in Estimator 6.4 for simultaneous inference on the counterfactual distributions \( \theta_0 \).

**Theorem 7.6** (Simultaneous confidence band). Suppose the conditions of Theorem 7.4 hold. Then for a fixed and finite grid \( U \), the confidence band in Estimator 6.4 jointly covers the true counterfactual distributions \( \theta_0 \) at all grid points \( y \in U \) with probability approaching the nominal level, i.e., \( \lim_{n \to \infty} \mathbb{P}((\theta_0)_j \in [l_j, u_j] \ \forall j) = 1 - a \).

See Appendix E for the proof.

---

\( ^5 \)If we allow the dimension of the grid \( U \) to grow with the sample size, even when \( \hat{\theta} \) is no longer asymptotically normal, with appropriate growth conditions and high quality estimators for the Jacobian \( J \) and the asymptotic variance \( C \), the bootstrap procedure should still guarantee simultaneous coverage based on Chernozhukov et al. [2013]. The formal justification seems to be an interesting question for future research.
8. Conclusion

We propose an automatic $\kappa$ weighting (Auto-$\kappa$) procedure for estimating complier parameters with high dimensional covariates. The procedure is easily implemented and semiparametrically efficient. As a contribution to the instrumental variable literature, we characterize the doubly robust moment function for a broad class of complier parameters, combining the well known Wald and $\kappa$ weight formulations. This new characterization leads to a new hypothesis test of whether two different instruments induce the same subpopulation of compliers. As a contribution to the debiased machine learning literature, we provide original arguments proving that the guarantees known for population parameters identified by selection on observables generalize for complier parameters identified by instrumental variables. In simulations, our proposed estimator outperforms $\kappa$ weighting and DML, both of which need to explicitly estimate and invert the propensity scores. In summary, we provide a practical method for selecting covariates (and their transformations) that also eliminates the need for ad hoc trimming or censoring in economic research using a binary instrument and high dimensional covariates.

References

Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292, March 2002. ISSN 0162-1459.

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Alberto Abadie, Joshua D Angrist, and Guido W Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70 (1):91–117, 2002.
Atila Abdulkadiroğlu, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak. Research design meets market design: Using centralized assignment for impact evaluation. *Econometrica*, 85(5):1373–1432, 2017.

Joshua Angrist and Brigham Frandsen. Machine labor. Technical report, National Bureau of Economic Research, 2019.

Joshua D Angrist. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics*, 19(1):2–28, 2001.

Joshua D Angrist and William N Evans. Children and their parents’ labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

Joshua D Angrist and Ivan Fernández-Val. ExtrapoLATE-ing: External validity and overidentifi- cation in the LATE framework. In *Advances in Economics and Econometrics*, pages 401–434, 2013.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, 2008.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Joshua D Angrist, Sarah R Cohodes, Susan M Dynarski, Parag A Pathak, and Christopher R Walters. Stand and deliver: Effects of Boston’s charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34(2):275–318, 2016.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.
Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian B Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian B Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian B Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Peter J Bickel. On adaptive estimation. *The Annals of Statistics*, pages 647–671, 1982.

Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated lasso. *arXiv:1502.06291*, 2015.

Victor Chernozhukov and Christian B Hansen. The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *Review of Economics and Statistics*, 86(3):735–751, 2004.

Victor Chernozhukov and Christian B Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005. ISSN 1468-0262.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, December 2013.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv:1608.00033*, 2016.
Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian B Hansen, Whitney K Newey, and James M Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1):C1–C68, 2018a.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. De-biased machine learning of global and local parameters using regularized Riesz representers. *arXiv:1802.08667*, 2018b.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *arXiv:1809.05224*, 2018c.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. A simple and general debiased machine learning theorem with finite sample guarantees. *arXiv preprint*, 2021.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*(1):187–199, March 2009.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics, 189*(1):1–23, 2015.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica, 89*(1):181–213, 2021.

Markus Frölich. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics, 139*(1):35–75, 2007.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica, 66*:315–331, 1998.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423, 2017.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *Annals of Statistics*, forthcoming.
Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382, 1994.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Elizabeth L Ogburn, Andrea Rotnitzky, and James M Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

James M Poterba and Steven F Venti. 401(k) plans and tax-deferred saving. In *Studies in the Economics of Aging*, pages 105–142. University of Chicago Press, 1994.

James M Poterba, Steven F Venti, and David A Wise. Do 401(k) contributions crowd out other personal saving? *Journal of Public Economics*, 58(1):1–32, 1995.

Yumou Qiu, Jing Tao, and Xiao-Hua Zhou. Inference of heterogeneous treatment effects using observational data with high-dimensional covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James M Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, Aad van der Vaart, and HEI Health Review Committee. New statistical approaches to semiparametric regression with application to air pollution research. Technical report, Health Effects Institute, 2013.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.
Rahul Singh and Liyang Sun. De-biased machine learning in instrumental variable models for treatment effects. *arXiv:1909.05244*, 2019.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.

Tymon Sloczyński and Jeffrey M Wooldridge. A general double robustness result for estimating average treatment effects. *Econometric Theory*, 34(1):112–133, 2018.

Baoluo Sun and Zhiqiang Tan. High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *arXiv:2009.09286*, 2020.

Zhiqiang Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

Mark J Van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011.

Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.
9. Tables and figures

Table 1. Bias, RMSE, and coverage simulation.

(a) $P(Y^{(0)} \leq y \mid D^{(1)} > D^{(0)})$

| y    | Bias       | RMSE       |
|------|------------|------------|
|      | $\kappa$ weight | DML | Auto-$\kappa$ | $\kappa$ weight | DML | Auto-$\kappa$ |
| -2   | -0.003     | -0.138     | -0.037       | 0.099            | 3.070     | 0.075     |
| -1.5 | -0.001     | -0.119     | -0.032       | 0.172            | 2.576     | 0.076     |
| -1   | 0.003      | -0.045     | -0.020       | 0.250            | 2.040     | 0.079     |
| -0.5 | 0.002      | -0.035     | 0.002        | 0.384            | 1.853     | 0.080     |
| 0    | -0.017     | 0.018      | 0.021        | 0.556            | 1.738     | 0.092     |
| 0.5  | -0.012     | 0.002      | 0.034        | 0.638            | 3.072     | 0.098     |
| overall | -0.005     | -0.053     | -0.005       | 0.350            | 2.391     | 0.083     |

(b) $P(Y^{(1)} \leq y \mid D^{(1)} > D^{(0)})$

| y    | Bias       | RMSE       |
|------|------------|------------|
|      | $\kappa$ weight | DML | Auto-$\kappa$ | $\kappa$ weight | DML | Auto-$\kappa$ |
| -2   | 0.002      | -0.115     | 0.013        | 0.028            | 0.444     | 0.015     |
| -1.5 | 0.004      | -0.114     | 0.012        | 0.039            | 0.441     | 0.016     |
| -1   | 0.008      | -0.110     | 0.011        | 0.057            | 0.432     | 0.020     |
| -0.5 | 0.016      | -0.103     | 0.011        | 0.078            | 0.410     | 0.026     |
| 0    | 0.021      | -0.093     | 0.016        | 0.090            | 0.379     | 0.035     |
| 0.5  | 0.021      | -0.078     | 0.027        | 0.092            | 0.315     | 0.044     |
| overall | 0.012      | -0.102     | 0.015        | 0.064            | 0.403     | 0.026     |

Notes: Simulation performance of $\kappa$ weight, DML and Auto-$\kappa$ estimators for the complier counterfactual outcome distribution, where the grid point for the distribution is specified in the first column. The last row averages the performance across grid points. Columns 1–3 compare the bias of different estimators and columns 4–6 compare the RMSE of different estimators. The best performing estimators are in bold. The simultaneous confidence bands based on the Auto-$\kappa$ estimator have coverage probability of 98.4% for the CDF of $Y^{(0)}$ and 93.6% for the CDF of $Y^{(1)}$, which are quite close to the nominal level of 95%. The propensity scores are specified in (B.1).
Figure 1. Numerical stability simulation

Notes: Simulation performance of $\kappa$ weight, DML and Auto-$\kappa$ estimators for estimating the counterfactual distribution, where the grid point is specified on the horizontal axis. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid points mark the median. The propensity scores are specified in (B.1).
Figure 3. Effect of 401(k) on the distribution of net financial assets for compliers

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

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

Notes: Point estimates (line) and simultaneous 95% confidence bands (shaded region) based on the Auto-\( \kappa \) estimator using the SIPP data. The point estimation consists of two steps. In the first step, we estimate the regression functions with \( \hat{\gamma} \) by either (panel A) Lasso according to Estimator 6.2 or (panel B) neural network with a single hidden layer of eight neurons as in Chernozhukov et al. [2018a]. We estimate the balancing weight with \( \hat{\alpha} \) by Lasso according to Estimator 6.2. In the second step, we estimate the complier counterfactual distribution using on the doubly robust moment function according to Estimator 6.1. We construct simultaneous confidence bands by steps outlined in Estimator 6.4.
ONLINE APPENDIX FOR "AUTOMATIC KAPPA WEIGHTING FOR INSTRUMENTAL VARIABLE MODELS OF COMPLIER TREATMENT EFFECTS"

Contents

1. Introduction 2
2. Related work 4
3. Preview 7
  3.1. Classic approaches: Wald formula and \( \kappa \) weight 7
  3.2. Doubly robust moment 8
  3.3. A new synthesis with machine learning: Auto-\( \kappa \) 9
4. Simulations and program evaluation 10
  4.1. Simulations 10
  4.2. Effect of 401(k) on net financial assets 12
5. Framework 13
  5.1. Complier parameters 13
  5.2. Doubly robust moment function 15
6. Estimation 18
  6.1. Automatic debiased machine learning 18
  6.2. Extensions: Hypothesis test and simultaneous confidence band 20
  6.3. Finite sample balance 21
  6.4. Finite sample orthogonality 22
7. Consistency and asymptotic normality 24
  7.1. Regression and balancing weight 24
  7.2. Complier parameter 27
8. Conclusion 29
References 29
9. Tables and figures 35
Appendix A. Notation

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 $X \in \mathcal{X} \subset \mathbb{R}^{\dim(X)}$ 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 Angrist et al. [1996], 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 Chernozhukov et al. [2018c], we denote the regression of random vector \( \mathbf{V} = (V_1, \ldots, V_J)' \) conditional on \((Z, \mathbf{X})\) as

\[
\gamma_0(Z, \mathbf{X}) = \begin{bmatrix}
\gamma_{V_1}^0(Z, \mathbf{X}) \\
\cdots \\
\gamma_{V_J}^0(Z, \mathbf{X})
\end{bmatrix} = \mathbb{E}[\mathbf{V}|Z, \mathbf{X}]
\]

where \( \gamma_{V_j}^0(Z, \mathbf{X}) = \mathbb{E}[V_j|Z, \mathbf{X}] \). The random vector \( \mathbf{V} \) is observable and depends on the complier parameter of interest; we specify its components for LATE, complier characteristics, and counterfactual outcome distributions in Theorem 5.2.

We denote the propensity score \( \pi_0(\mathbf{x}) = \mathbb{P}(Z = 1|\mathbf{X} = \mathbf{x}) \). We denote the classic Horvitz-Thompson weight with

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

We denote by \( |\cdot|_q \) the \( \ell_q \) norm of a vector. We denote by \( \| \cdot \| \) the \( L_2 \) norm of a random variable \( V_j \), i.e. \( \|V_j\| = \sqrt{\mathbb{E}[V_j^2]} \). For random vector \( \mathbf{V} = (V_1, \ldots, V_J)' \), we slightly abuse notation by writing

\[
\|\mathbf{V}\| = \begin{bmatrix}
\|V_1\| \\
\|V_2\| \\
\vdots \\
\|V_J\|
\end{bmatrix}, \quad \mathbb{E}[\mathbf{V}^2] = \|\mathbf{V}\|^2 = \begin{bmatrix}
\|V_1\|^2 \\
\|V_2\|^2 \\
\vdots \\
\|V_J\|^2
\end{bmatrix}.
\]

Likewise, we write the element-wise absolute value as

\[
|\mathbf{V}| = \begin{bmatrix}
|V_1| \\
|V_2| \\
\vdots \\
|V_J|
\end{bmatrix}.
\]
Consider a causal parameter $\theta_0 \in \Theta$, where $\Theta$ is some compact parameter space. It is implicitly defined by moment function $m$:

$$\mathbb{E}[m(W, \gamma_0, \theta)] = 0 \text{ if and only if } \theta = \theta_0.$$ 

We denote the doubly robust moment function for $\theta_0$ by

$$\mathbb{E}[\psi(W, \gamma_0, \alpha_0, \theta)] = 0 \text{ if and only if } \theta = \theta_0$$

$$\psi(w, \gamma, \alpha, \theta) = m(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta).$$

$\phi$ is called the debiasing 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 balancing weight parameter $\rho_L$ is the solution to

$$\rho_L = \arg\min_{\rho} \mathbb{E}[(\alpha_0(Z, X) - \rho^T b(Z, X))^2] + 2\lambda_L |\rho|_1$$

where $\lambda_L = \sqrt{\frac{\ln p}{n}}$ is the theoretical regularization parameter.

The sample balancing weight parameter $\hat{\rho}_{-\ell}$ estimated from $I_{-\ell}$ is the solution to

$$\hat{\rho}_{-\ell} = \arg\min_{\rho} \rho^T \hat{G}_{-\ell} \rho - 2\rho^T \hat{M}_{-\ell} + 2\lambda_n |\rho|_1$$

where

$$\hat{G}_{-\ell} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b(Z_i, X_i)b(Z_i, X_i)^T, \quad \hat{M}_{-\ell} = \frac{1}{n - n_\ell} \sum_{i \in I_{-\ell}} b(1, X_i) - b(0, X_i)$$

and $\lambda_n$ is a data-driven regularization parameter. We denote $\hat{\alpha}_{-\ell}(z, x) = b(z, x)^T \hat{\rho}_{-\ell}$.

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 $\hat{\theta}$. Let $\hat{S} = \text{diag}(\hat{C})$ 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 7.

Appendix B. Simulations and program evaluation

B.1. Trimming and censoring. We assess the effectiveness of trimming and censoring in the simulation considered in Section 4 where the propensity scores is specified in (B.1) below. We impose trimming according to Belloni et al. [2017], dropping observations with \( \hat{\pi} \not\in [10^{-12}, 1-10^{-12}] \). We impose censoring by setting \( \hat{\pi} < 10^{-12} \) to be \( 10^{-12} \) and \( \hat{\pi} > 1-10^{-12} \) to be \( 1-10^{-12} \). As shown in Figure A.5 and A.3, trimming and censoring improve the stability considerably for DML, though sometimes at the cost of introducing bias.

B.2. Simulation design. Each simulation consists of a sample of \( n = 1000 \) observations. A given observation is generated from the following IV model:

\[
X \overset{i.i.d.}{\sim} \text{Unif}[0, 1]
\]

\[
Z \mid X = x \overset{i.i.d.}{\sim} \text{Bern}(\pi_0(x))
\]

\[
D \mid Z = z, X = x \overset{i.i.d.}{\sim} \text{Bern}(\gamma_0^D(z, x))
\]

\[
Y \mid Z = z, X = x \overset{i.i.d.}{\sim} \mathcal{N}(\gamma_0^Y(z, x), 1)
\]

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, the reduced form and first stage regression functions are specified as

\[
\gamma_0^Y(z, x) = 2zx^2, \quad \gamma_0^D(z, x) = zx.
\]

The propensity to receive the instrument is specified as a step function of \( x \):

\[
\pi_0(x) = 0.05 \cdot 1_{x \leq 0.5} + 0.95 \cdot 1_{x > 0.5}.
\]
or a continuous function of $x$:

\[(B.2) \quad \pi_0(x) = (0.95 - 0.05) \frac{\exp(x)}{1 + \exp(x)} + 0.05.\]

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

\[\beta_0^y = \frac{\int_0^1 [\Phi(y - 2x^2)(x - 1) + \Phi(y)]dx}{\int_0^1 xdx}, \quad \delta_0^y = \frac{\int_0^1 [\Phi(y - 2x^2)x]dx}{\int_0^1 xdx}.\]

In the main text, we present the simulation results for propensity scores specified in (B.1). In Table A.1 and Figure A.5 we present the simulation results for propensity scores specified in (B.2). The simultaneous confidence bands based on the Auto-$\kappa$ estimator have coverage probability of 98.3% for the CDF of $Y^0$ and 97.8% for the CDF of $Y^1$, which are quite close to the nominal level of 95%.

**APPENDIX C. FRAMEWORK**

C.1. $\kappa$ weight as Riesz representer to Wald formula.

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

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

Similarly, $\frac{Z}{\pi_0(X)}$ is the Riesz representer to the continuous linear functional $\gamma \mapsto \mathbb{E}[\gamma(1, X)]$, and $\frac{1 - Z}{1 - \pi_0(X)}$ is the Riesz representer to the continuous linear functional $\gamma \mapsto \mathbb{E}[\gamma(0, X)]$. 

Proof. Observe that
\[
\mathbb{E}\left[ \gamma(Z, X) \frac{Z}{\pi_0(X)} \right] = \mathbb{E}\left[ \gamma(Z, X) \frac{1}{\pi_0(X)} \right| Z = 1, X] \mathbb{P}(Z = 1|X) \\
= \mathbb{E}\left[ \gamma(Z, X) \frac{1}{\pi_0(X)} \right] \pi_0(X) \\
= \gamma(1, X)
\]
and likewise
\[
\mathbb{E}\left[ \gamma(Z, X) \frac{1 - Z}{1 - \pi_0(X)} \right] = \gamma(0, X).
\]
Combining these two terms, we have by the law of iterated expectations
\[
\mathbb{E}[\gamma(1, X) - \gamma(0, X)] = \int \{ \gamma(1, x) - \gamma(0, x) \} d\mathbb{P}(x) \\
= \int \left\{ \mathbb{E}\left[ \gamma(Z, X) \frac{Z}{\pi_0(X)} \right| X = x \right\} d\mathbb{P}(x) \\
- \mathbb{E}\left[ \gamma(Z, X) \frac{1 - Z}{1 - \pi_0(X)} \right] \pi_0(X) \\
= \mathbb{E}\left[ \gamma(Z, X) \frac{Z}{\pi_0(X)} \right] - \mathbb{E}\left[ \gamma(Z, X) \frac{1 - Z}{1 - \pi_0(X)} \right].
\]

\[\square\]

Proposition C.2. The \( \kappa \) weights can be rewritten as
\[
\kappa^{(0)}(w) = \alpha_0(z, x)[d - 1] \\
\kappa^{(1)}(w) = \alpha_0(z, x)d \\
\kappa(w) = 1 - \frac{d(1 - z)}{1 - \pi_0(x)} - \frac{(1 - d)z}{\pi_0(x)}.
\]
Proof. Recall the $\kappa$ weights are defined as

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

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

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

Observe that

$$
\alpha_0(z, x) = \frac{z}{\pi_0(x)} - \frac{1 - z}{1 - \pi_0(x)} = \frac{z - \pi_0(x)}{\pi_0(x) \{1 - \pi_0(x)\}}
$$

which proves the expression for $\kappa^{(0)}$ and $\kappa^{(1)}$. Plugging in these expressions, we have

$$
\kappa(w) = \{1 - \pi_0(x)\} \alpha_0(z, x) [d - 1] + \pi_0(x) \alpha_0(z, x) d = 1 \frac{d(1 - z)}{1 - \pi_0(x)} - \frac{(1 - d)z}{\pi_0(x)}.
$$

\[\square\]

C.2. General doubly robust moment function.

**Theorem C.1.** Suppose Assumption 1 holds. Let $g(y, d, x, \theta)$ be a measurable, real-valued function such that $E[g(Y, D, X, \theta)]^2 < \infty$ 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

$$
v(w, \theta) = d \cdot g(y, x, \theta).
$$
Then the doubly robust moment function for $\theta_0$ is of the form

$$ \psi(w, \gamma, \alpha, \theta) = \tilde{m}(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta) $$

$$ \tilde{m}(w, \gamma, \theta) = \gamma(1, x, \theta) - \gamma(0, x, \theta) $$

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

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

(3) If $\theta_0$ is defined by the moment condition $E[g(Y, D, X, \theta_0)|D^{(1)} > D^{(0)}] = 0$, the doubly robust moment function for $\theta_0$ is of the form

$$ \psi(w, \gamma, \alpha, \theta) = \tilde{m}(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta) $$

$$ \tilde{m}(w, \gamma, \theta) = \gamma(z, x, \theta) - \gamma_0(1, x, \theta) - \gamma_1(0, x, \theta) $$

$$ \phi(w, \gamma, \alpha, \theta) = [g(y, d, x, \theta) - \gamma(z, x, \theta)] $$

$$ - \frac{z}{\pi(x)}[(1 - d) \cdot g(y, d, x, \theta) - \gamma_0(z, x, \theta)] $$

$$ - \frac{1 - z}{1 - \pi(x)}[d \cdot g(y, d, x, \theta) - \gamma_1(z, x, \theta)] $$

where $\gamma_0(z, x, \theta) := E[g(y, d, x, \theta_0)|z, x]$, $\gamma_1(z, x, \theta) := E[d \cdot g(y, d, x, \theta_0)|z, x]$ and $\gamma_0^0(z, x, \theta) := E[(1 - d) \cdot g(y, d, x, \theta_0)|z, x]$. Here we abuse notation by letting $\alpha(z, x) = \begin{pmatrix} \frac{z}{\pi(x)} & 1 - \frac{z}{\pi(x)} \end{pmatrix}$ denote a vector-valued function like $\gamma(z, x)$.

Proof. Consider the first case. Under Assumption 1, we can appeal to Abadie [2003, Theorem 3.1].

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

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

appealing to Assumption 1, Proposition C.2, and the fact that \(\alpha_0\) is the Riesz representer for \(\gamma \mapsto \mathbb{E}[\gamma_0(1, X, \theta_0) - \gamma_0(0, X, \theta_0)]\). Likewise for the second case.

A similar argument extends to the third case. Under Assumption 1, we can appeal to Abadie [2003, Theorem 3.1].

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

Hence

\[
0 = \mathbb{E}[\kappa(W)g(Y, D, X, \theta_0)] \\
= \mathbb{E}[g(Y, D, X, \theta_0) - \frac{Z}{\pi_0(X)}(1 - D) \cdot g(Y, D, X, \theta_0) - \frac{1 - Z}{1 - \pi_0(X)} D \cdot g(Y, D, X, \theta_0)] \\
= \mathbb{E}[\gamma_0(Z, X, \theta) - \gamma_0(1, X, \theta_0) - \gamma_0(0, X, \theta_0)]
\]

appealing to Assumption 1 and Proposition C.2. \(\square\)

C.3. Proof of main result.

Proof of Theorem 5.2. 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 \(\tilde{m}\) 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 
\[
\mathbb{E}[Y^{(1)} - \delta_0|D^{(1)} > D^{(0)}] = 0
\]
and \( \beta_0 \) is defined by the moment condition 
\[
\mathbb{E}[Y^{(0)} - \beta_0|D^{(1)} > D^{(0)}] = 0.
\]
Applying Case 2 of Theorem C.1 to \( \delta_0 \), we have 
\[
v(w, \delta) = d \cdot (y - \delta).
\]
Applying Case 1 of Theorem C.1 to \( \beta_0 \), we have
\[
v(w, \beta) = (d - 1) \cdot (y - \beta) \propto (d - 1) \cdot y - d \cdot \beta.
\]
Writing \( \theta = \delta - \beta \), 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 
\[
A(\theta) = \begin{bmatrix} 1 & -\theta \end{bmatrix}
\]
in Theorem 5.2.

(2) For average complier characteristics, \( \theta_0 \) is defined by the moment condition
\[
\mathbb{E}[f(X) - \theta_0|D^{(1)} > D^{(0)}] = 0.
\]
Applying Case 2 of Theorem C.1 setting \( g(Y^{(1)}, X, \theta_0) = f(X) - \theta_0 \), we have 
\[
v(w, \theta) = d \cdot (f(x) - \theta).
\]
This expression decomposes into \( V = (Df(X)', D)' \) and 
\[
A(\theta) = \begin{bmatrix} I & -\theta \end{bmatrix}
\]
in Theorem 5.2.

(3) For complier distribution of \( Y^{(0)} \), \( \beta_0 \) is defined by the moment condition
\[
\mathbb{E}[1_{Y^{(0)} \leq \bar{y}} - \beta_0|D^{(1)} > D^{(0)}] = 0.
\]
Applying Case 1 of Theorem C.1 to \( \beta_0 \), we have
\[
v(w, \beta) = (d - 1) \cdot (1_{Y^{(0)} \leq \bar{y}} - \beta) \propto (d - 1) \cdot 1_{Y^{(0)} \leq \bar{y}} - d \cdot \beta.
\]
For complier distribution of \( Y^{(1)} \), \( \delta_0 \) is defined by the moment condition
\[
\mathbb{E}[1_{Y^{(1)} \leq \bar{y}} - \delta_0|D^{(1)} > D^{(0)}] = 0.
\]
Applying Case 2 of Theorem C.1 to \( \delta_0 \), we have
\[
v(w, \delta) = d \cdot (1_{Y^{(0)} \leq \bar{y}} - \delta_0).
\]
Concatenating \( v(w, \beta) \) and \( v(w, \delta) \), we arrive at the decomposition in Theorem 5.2.

\[\square\]

Appendix D. Estimation

D.1. Tuning. Estimator 6.2 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 Chernozhukov et al. [2018c], adapted from Chernozhukov et al. [2018b]. Due to its iterative nature, the tuning procedure is most clearly stated as a replacement for Estimator 6.2.

Recall that the inputs to Estimator 6.2 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 Estimator 6.2, the output of the tuning procedure is \( \hat{\alpha}_{-\ell} \), an estimator of the balancing weight trained only on observations in \( I_{-\ell} \).

The tuning procedure is as follows.

**Estimator D.1** (Regularized balancing weight with tuning). *For observations in \( I_{-\ell} \),*

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

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

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

\[
\hat{\rho}_{-\ell} = \begin{bmatrix} (\hat{G}_{-\ell})^{-1} \\ \hat{M}_{-\ell} \end{bmatrix}.
\]

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

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

\[
\hat{\rho}_-\ell = \text{argmin}_{\rho} \rho' \hat{G}_-\ell \rho - 2\rho' \hat{M}_-\ell + 2\lambda_n c_3 |\hat{D}_{-\ell,11} \cdot \rho_1| + 2\lambda_n \sum_{j=2}^p |\hat{D}_{-\ell,jj} \cdot \rho_j|;
\]

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

In step 3, \((c_1, c_2, c_3)\) are hyperparameters taken as \((1, 0.1, 0.1)\) in practice. We implement the optimization via generalized coordinate descent with soft thresholding. See Chernozhukov et al. [2018c] 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 Estimator D.1 in the same manner as Chernozhukov et al. [2018b, Section 5.1]. Specifically, we appeal to Belloni and Chernozhukov [2013, Theorem 8] for the homoscedastic case and Belloni et al. [2012, Theorem 1] for the heteroscedastic case.

D.2. Balancing and orthogonality.

Proof of Proposition 6.1. The Lasso first order condition gives \(|\hat{M}_{-\ell} - \hat{G}_{-\ell} \hat{\rho}_{-\ell}|_\infty \leq \lambda_n\). \(\square\)

Proof of Proposition 6.2. The derivatives are

\[
\frac{\partial}{\partial \beta} \psi(W_i, \beta, \rho, \theta_0) = \{b(1, X_i) - b(0, X_i) - \rho' b(Z_i, X_i) b(Z_i, X_i)\} A(\theta_0);
\]

\[
\frac{\partial}{\partial \rho} \psi(W_i, \beta, \rho, \theta_0) = b(Z_i, X_i) A(\theta_0) \{V_i - \beta' b(Z_i, X_i)\}.
\]
We analyze the former expression; the argument for the latter expression is analogous, replacing $b \mapsto b(1, x) - b(0, x)$ with $b \mapsto b(z, x)v'$. In the former expression, we have,

$$\frac{1}{n - n_\ell} \sum_{i \in I_\ell} \frac{\partial}{\partial \beta} \psi(W_i, \hat{\beta}_{-\ell}, \hat{\rho}_{-\ell}, \theta_0) = \{\hat{M}_{-\ell} - \hat{G}_{-\ell}\hat{\rho}_{-\ell}\} A(\theta_0).$$

Finally we appeal to the Lasso first order condition that is satisfied by $\hat{\rho}_{-\ell}$. □

**Appendix E. Consistency and asymptotic normality**

Consider the notation

$$\psi(w, \gamma, \alpha, \theta) = m(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta);$$

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

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

**E.1. Lemmas from previous work.**

**Definition E.1.** Define the following $p \times p$ matrix $G$ and the vector $M$:

$$G = E[b(Z, X)b(Z, X)'],$$

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

**Proposition E.1.** Under Assumption 2, we have $|\hat{G} - G|_{\infty} = O_p \left(\sqrt{\frac{\ln p}{n}}\right).$

*Proof.* The proof follows Chernozhukov et al. [2018c, Lemma C1]. □

**Proposition E.2.** Under Assumptions 1 and 2, we have $|\hat{M} - M|_{\infty} = O_p \left(\sqrt{\frac{\ln p}{n}}\right).$

*Proof.* The proof follows Chernozhukov et al. [2018c, Lemma 4]. □

**Proposition E.3.** Denote $\tilde{m}(w, \gamma) := \gamma(1, x) - \gamma(0, x)$. Under Assumptions 1 and 7, the following holds.

1. $E[\tilde{m}(W, \gamma_0)^2] < \infty,$
(2) $\mathbb{E}[\{\tilde{m}(W, \gamma) - \tilde{m}(W, \gamma_0)\}^2] \text{ is continuous at } \gamma_0 \text{ with respect to } \|\gamma - \gamma_0\|,$

(3) $\max_j |\tilde{m}(W, b_j) - \tilde{m}(W, 0)| \leq C.$

Proof. The proof follows Chernozhukov et al. [2018c, Theorem 6].

Proposition E.4. Under Assumption 7, the following holds.

(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 $z \in \{0, 1\}.$

Proof. The proof follows Chernozhukov et al. [2018c, Theorem 6].

Proof of Lemma 7.1. Applying Proposition E.1 and Proposition E.2, the proof follows Chernozhukov et al. [2018c, Theorem 1].

Proof of Lemma 7.2. Applying Proposition E.1 and Proposition E.2, the proof follows Chernozhukov et al. [2018c, Theorem 3]. The argument that $|\hat{\rho}|_1 = O_p(1)$ is analogous to Chernozhukov et al. [2018c, Lemmas 2 and 3].

Proposition E.5. Suppose the conditions of Theorem 7.4 hold. Then we have

$$\frac{1}{\sqrt{n}} \sum_{\ell=1}^L \sum_{i \in I_\ell} \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \hat{\theta}) \xrightarrow{p} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_0(W_i), \quad \psi_0(W_i) := \psi(W_i, \gamma_0, \alpha_0, \theta_0).$$

Proof. The proof follows from Chernozhukov et al. [2018c, Theorem 5].

Proposition E.6. Consider the estimator $\hat{\theta} = \text{argmin}_{\theta \in \Theta} \hat{Q}(\theta)$, where $\hat{Q} : \Theta \rightarrow \mathbb{R}$ estimates $Q_0 : \Theta \rightarrow \mathbb{R}$. If the following conditions hold

(1) $\Theta$ is compact,

(2) $Q_0$ is continuous in $\theta \in \Theta$,

(3) $Q_0$ is uniquely maximized at $\theta_0$,

(4) $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q_0(\theta)| \xrightarrow{p} 0,$

then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. The proof follows Newey and McFadden [1994, Theorem 2.1].
E.2. Proof of main result.

Proposition E.7. Suppose the conditions of Theorem 7.4 hold. Then for each fold $I_\ell$ the following holds

1. $E\left[\{m(W, \hat{\gamma}_{-\ell}, \theta_0) - m(W, \gamma_0, \theta_0)\}^2 | I_{-\ell}\right] \xrightarrow{p} 0,$

2. $E\left[\{\phi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0) - \phi(W, \gamma_0, \alpha_0, \theta_0)\}^2 | I_{-\ell}\right] \xrightarrow{p} 0,$

3. $E\left[\{\phi(W, \gamma_0, \hat{\alpha}_{-\ell}, \theta_0) - \phi(W, \gamma_0, \alpha_0, \theta_0)\}^2 | I_{-\ell}\right] \xrightarrow{p} 0.$

The notation $E[\cdot | I_{-\ell}]$ means conditional on $W_{-\ell} := \{W_i\}_{i \notin I_\ell}$, i.e. observations not in fold $I_\ell$.

Proof. First note that

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

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

To lighten the proof, we slightly abuse notation as follows:

$$\|\gamma_0 - \hat{\gamma}_{-\ell}\|^2 = E\left[\{\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)\}^2 | I_{\ell}\right];$$

$$\|\alpha_0 - \hat{\alpha}_{-\ell}\|^2 = E\left[\{\alpha(Z, X) - \hat{\alpha}_{-\ell}(Z, X)\}^2 | I_{\ell}\right].$$

(1) By Proposition E.4, the convergence holds due to $\|\gamma_0 - \hat{\gamma}_{-\ell}\| \xrightarrow{p} 0$.

(2) By Assumption 6 and Assumption 7, we have

$$\|\alpha_0 A(\theta_0)[\gamma_0 - \hat{\gamma}_{-\ell}]\| \leq CA(\theta_0)\|\gamma_0 - \hat{\gamma}_{-\ell}\| \xrightarrow{p} 0.$$ 

(3) By Lemma 7.1 or Lemma 7.2, Assumption 7, and law of iterated expectations with respect to $I_{-\ell}$, we have

$$\|\hat{\alpha}_{-\ell} - \alpha_0\|A(\theta_0)[v - \gamma_0(z, x)]\| \leq \|\hat{\alpha}_{-\ell} - \alpha_0\|A(\theta_0)C \cdot \bar{1} \xrightarrow{p} 0.$$ 

$\square$
Proposition E.8. Suppose the conditions of Theorem 7.4 hold. Then

\[
\frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \left[ \phi(W_i, \hat{\gamma}_-\ell, \hat{\alpha}_-\ell, \theta_0) - \phi(W_i, \hat{\gamma}_-\ell, \alpha_0, \theta_0) \right] - \phi(W_i, \gamma_0, \hat{\alpha}_-\ell, \theta_0) + \phi(W_i, \gamma_0, \alpha_0, \theta_0) \xrightarrow{P} 0.
\]

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

\[
\mathbb{E} \left[ \frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} -[\hat{\alpha}_-\ell(Z_i, X_i) - \alpha_0(Z_i, X_i)] A(\theta_0)[\hat{\gamma}_-\ell(Z_i, X_i) - \gamma_0(Z_i, X_i)] \right]
\]

\[
\leq \sum_{\ell=1}^{L} \mathbb{E} \left[ \sqrt{n} \frac{1}{n} \sum_{i \in I_\ell} -[\hat{\alpha}_-\ell(Z_i, X_i) - \alpha_0(Z_i, X_i)] A(\theta_0)[\hat{\gamma}_-\ell(Z_i, X_i) - \gamma_0(Z_i, X_i)] \right]
\]

\[
= \sum_{\ell=1}^{L} \mathbb{E} \left[ \sqrt{n} \frac{1}{n} \sum_{i \in I_\ell} [\hat{\alpha}_-\ell(Z_i, X_i) - \alpha_0(Z_i, X_i)] A(\theta_0)[\hat{\gamma}_-\ell(Z_i, X_i) - \gamma_0(Z_i, X_i)] \right] I_{-\ell}
\]

\[
= \sum_{\ell=1}^{L} \mathbb{E} \left[ \sqrt{n} \frac{1}{n} [\hat{\alpha}_-\ell(Z_i, X_i) - \alpha_0(Z_i, X_i)] A(\theta_0)[\hat{\gamma}_-\ell(Z_i, X_i) - \gamma_0(Z_i, X_i)] I_{-\ell} \right].
\]

Applying the Hölder’s inequality element-wise and Corollary 7.3, we have convergence for each summand as follows:

\[
\mathbb{E} \left[ \sqrt{n} \frac{1}{n} [\hat{\alpha}_-\ell(Z_i, X_i) - \alpha_0(Z_i, X_i)] A(\theta_0)[\hat{\gamma}_-\ell(Z_i, X_i) - \gamma_0(Z_i, X_i)] I_{-\ell} \right]
\]

\[
\leq \mathbb{E} \left[ \sqrt{n} [\hat{\alpha}_-\ell(Z_i, X_i) - \alpha_0(Z_i, X_i)] A(\theta_0)[\hat{\gamma}_-\ell(Z_i, X_i) - \gamma_0(Z_i, X_i)] I_{-\ell} \right]
\]

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

\( \xrightarrow{P} 0. \)
In the penultimate step, we slightly abuse notation, using

\[ \| \gamma_0 - \hat{\gamma}_{-\ell} \|^2 = \mathbb{E}\left[ \{ \gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X) \}^2 \right] | I_{\ell}; \]

\[ \| \alpha_0 - \hat{\alpha}_{-\ell} \|^2 = \mathbb{E}\left[ \{ \alpha(Z, X) - \hat{\alpha}_{-\ell}(Z, X) \}^2 \right] | I_{\ell}. \]

\[ \square \]

**Proposition E.9.** Under Assumption 1, for each fold \( I_{\ell} \), the following holds:

1. \( \sqrt{n} \mathbb{E} [\psi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0)] \overset{p}{\to} 0; \)
2. \( \sqrt{n} \mathbb{E} [\phi(W, \gamma_0, \hat{\alpha}_{-\ell}, \theta_0)] \overset{p}{\to} 0. \)

**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. By Proposition C.1, \( \mathbb{E} \left[ \psi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0) \left| I_{-\ell} \right. \right] = 0. \) Applying the law of iterated expectations, we have \( \mathbb{E}[\psi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0)] = 0. \)
2. By law of iterated expectations, \( \mathbb{E} \left[ \phi(W, \gamma_0, \hat{\alpha}_{-\ell}, \theta_0) \left| I_{-\ell} \right. \right] = 0. \) Applying the law of iterated expectations, we have \( \mathbb{E}[\psi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0)] = 0. \)

\[ \square \]

**Proposition E.10.** Suppose the conditions of Theorem 7.4 hold. Then

1. The Jacobian \( J \) exists.
2. There exists a neighborhood \( N \) of \( \theta_0 \) with respect to \( | \cdot |_2 \) such that
   
   (a) \( \| \hat{\gamma}_{-\ell} - \gamma_0 \| \overset{p}{\to} 0; \)
   
   (b) \( \| \hat{\alpha}_{-\ell} - \alpha_0 \| \overset{p}{\to} 0; \)
   
   (c) For \( \| \gamma - \gamma_0 \| \) and \( \| \alpha - \alpha_0 \| \) small enough, \( \psi(W_i, \gamma, \alpha, \theta) \) is differentiable in \( \theta \) with probability approaching one.
(d) There exists $\zeta > 0$ and $d(W)$ such that $E[d(W)] < \infty$ and for $\|\gamma - \gamma_0\|$ small enough,

$$\left| \frac{\partial \psi(w, \gamma, \alpha, \theta)}{\partial \theta} - \frac{\partial \psi(w, \gamma, \alpha, \theta_0)}{\partial \theta} \right|_2 \leq d(w)\|\theta - \theta_0\|_2^\zeta.$$ 

(3) For any fold $I_\ell$ and any components $(j,k)$,

$$E\left[ \left| \frac{\partial \psi_j(W, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_0)}{\partial \theta_k} - \frac{\partial \psi_j(W, \gamma_0, \alpha_0, \theta_0)}{\partial \theta_k} \right| \right] \overset{p}{\to} 0.$$

Proof. Note that

$$\frac{\partial \psi(w, \gamma, \alpha, \theta)}{\partial \theta} = \frac{\partial A(\theta)}{\partial \theta} [\gamma(1, x) - \gamma(0, x)] + \alpha(z, x) \frac{\partial A(\theta)}{\partial \theta} [v - \gamma(z, x)]$$

where $\frac{\partial A(\theta)}{\partial \theta}$ is a tensor consisting of 1s and 0s.

To lighten the proof, we slightly abuse notation as follows:

$$\|\gamma_0 - \hat{\gamma}_{-\ell}\|^2 = E[\{\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)\}^2 | I_\ell];$$

$$\|\alpha_0 - \hat{\alpha}_{-\ell}\|^2 = E[\{\alpha(Z, X) - \hat{\alpha}_{-\ell}(Z, X)\}^2 | I_\ell].$$

(1) It suffices to show the second moment of the derivative is finite. By triangle inequality and Assumption 7 we have

$$\left\| \frac{\partial A(\theta_0)}{\partial \theta} [\gamma_0(1, x) - \gamma_0(0, x)] + \alpha_0(z, x) \frac{\partial A(\theta)}{\partial \theta} [v - \gamma_0(z, x)] \right\| \leq \frac{\partial A(\theta_0)}{\partial \theta} \{\|\gamma_0(1, x) - \gamma_0(0, x)\| + CC'\}.$$ 

To bound the RHS, by Proposition E.4 we have

$$\|\gamma_0(1, x) - \gamma_0(0, x)\| \leq \|\gamma_0(1, x)\| + \|\gamma_0(0, x)\| \leq C\|\gamma_0\| < \infty.$$

(2) (a) The convergence holds due to Assumption 6.

(b) The convergence holds due to Lemma 7.1 or Lemma 7.2.
(c) Differentiability holds since \( \frac{\partial \psi(w, \gamma, \alpha, \theta)}{\partial \theta} \) does not depend on \( \theta \).

(d) The LHS is exactly \( \vec{0} \) since \( \frac{\partial \psi(w, \gamma, \alpha, \theta)}{\partial \theta} \) does not depend on \( \theta \).

(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 - \hat{\gamma}_{-\ell}(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)]
+ \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
\[
\mathbb{E} [\|\xi\|] \leq \mathbb{E} [\|\hat{\gamma}_{-\ell}(1, X) - \gamma_0(1, X)\|]
\]
\[
+ \mathbb{E} [\|\hat{\gamma}_{-\ell}(0, X) - \gamma_0(0, X)\|]
\]
\[
+ \mathbb{E} [\|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)\|]\mathbb{V} - \gamma_0(Z, X)\|
\]
\[
+ \mathbb{E} [\|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)\|]\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)\|
\]
\[
+ \mathbb{E} [\|\alpha_0(Z, X)\|]\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)\|.
\]

Consider the first term. Under Assumption 6, applying law of iterated expectation, Jensen’s inequality, and Proposition E.4, we have
\[
\mathbb{E} [\|\hat{\gamma}_{-\ell}(1, X) - \gamma_0(1, X)\|] = \mathbb{E} \left[ \mathbb{E} \left[ \|\hat{\gamma}_{-\ell}(1, X) - \gamma_0(1, X)\| \mid I_{-\ell} \right] \right]
\]
\[
\leq \mathbb{E} [\|\hat{\gamma}_{-\ell}(1, x) - \gamma_0(1, x)\|]
\]
\[
\leq C \mathbb{E} [\|\hat{\gamma}_{-\ell} - \gamma_0\|]
\]
\[
P \rightarrow 0.
\]

Likewise for the second term. Consider the third term. Under Assumption 7, applying law of iterated expectation, Lemma 7.1 or Lemma 7.2, and Hölder’s inequality we have
\[
\mathbb{E} [\|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)\|]\mathbb{V} - \gamma_0(Z, X)\|
\]
\[
= \mathbb{E} \left[ \mathbb{E} \left[ \|\hat{\alpha}_{-\ell}(Z, X) - \alpha_0(Z, X)\|\mathbb{V} - \gamma_0(Z, X)\| \mid I_{-\ell} \right] \right]
\]
\[
\leq \mathbb{E} [\|\hat{\alpha}_{-\ell} - \alpha_0\|]\mathbb{V} - \gamma_0(z, x)\|
\]
\[
\leq C \mathbb{E} [\|\hat{\alpha}_{-\ell} - \alpha_0\|]
\]
\[
P \rightarrow 0.
\]
Consider the fourth term. By law of iterated expectations, Hölder’s inequality, and Corollary 7.3 we have

\[
\mathbb{E} \left[ \left| \hat{\alpha} - \ell(Z, X) - \alpha_0(Z, X) \right| \gamma_0 - \hat{\gamma} - \ell(Z, X) \right] = \mathbb{E} \left[ \mathbb{E} \left[ \left| \hat{\alpha} - \ell(Z, X) - \alpha_0(Z, X) \right| \gamma_0 - \hat{\gamma} - \ell(Z, X) \right] \right] \leq \mathbb{E} \left[ \left| \hat{\alpha} - \alpha_0 \right| \left| \gamma_0 - \hat{\gamma} \right| \right] \xrightarrow{p} 0.
\]

Consider the fifth term. By law of iterated expectations, Assumptions 6 and 7, and Jensen’s inequality, we have

\[
\mathbb{E} \left[ \left| \alpha_0(Z, X) \right| \gamma_0 - \hat{\gamma} - \ell(Z, X) \right] = \mathbb{E} \left[ \mathbb{E} \left[ \left| \alpha_0(Z, X) \right| \gamma_0 - \hat{\gamma} - \ell(Z, X) \right] \right] \leq C \mathbb{E} \left[ \left| \gamma_0 - \hat{\gamma} - \ell(Z, X) \right| \right] \xrightarrow{p} 0.
\]

\[\square\]

**Proposition E.11.** Suppose the conditions of Theorem 7.4 hold. Then \( \hat{\theta} \xrightarrow{p} \theta_0 \).

**Proof.** We verify the four conditions of Proposition E.6 with

\[
Q_0(\theta) = \mathbb{E}[\psi_0(\theta)] / \mathbb{E}[\psi_0(\theta)],
\]

\[
\hat{Q}(\theta) = \left[ \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \hat{\psi}_i(\theta) \right] / \left[ \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \hat{\psi}_i(\theta) \right],
\]

\[
\psi_0(\theta) = \psi(W, \gamma_0, \alpha_0, \theta),
\]

\[
\hat{\psi}_i(\theta) = \psi(W_i, \hat{\gamma}_-\ell, \hat{\alpha}_-\ell, \theta).
\]
(1) The first condition follows from Assumption 7,
(2) The second condition follows from Theorem 5.2.
(3) The third condition follows from Theorem 5.2.

(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 for \( i \in I_\ell \),

\[ \psi_0(\theta) = A(\theta)\eta_0(W), \quad \mathbb{E}[\psi_0(\theta)] = A(\theta)\mathbb{E}[\eta_0(W)]; \]

\[ \hat{\psi}_i(\theta) = A(\theta)\hat{\eta}_{-\ell}(W_i), \quad \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i(\theta) = A(\theta)\frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\eta}_{-\ell}(W_i). \]

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

We therefore turn to proving 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 initial terms. Denote $\xi_i = \hat{\eta}_\ell(W_i) - \eta_0(W_i)$ as in Proposition E.10 item 3. We prove convergence in mean by

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

$$
= \sum_{\ell=1}^L \mathbb{E} \left[ \frac{n_{\ell}}{n} \mathbb{E} \left[ |\xi_i| I_{-\ell} \right] \right]
$$

$$
= \sum_{\ell=1}^L \mathbb{E} \left[ \mathbb{E} \left[ |\xi_i| I_{-\ell} \right] \right]
$$

$$
\leq \sum_{\ell=1}^L \mathbb{E} \left[ \mathbb{E} \left[ |\xi_i| I_{-\ell} \right] \right]
$$

$$
P \to 0
$$

where the first inequality is due to triangle inequality, the second equality is due to the law of iterated expectations, and the rest echoes the proof of Proposition E.10 item 3.

Consider the latter terms. By the weak law of large numbers, if $\mathbb{E}[\eta_0(W)^2] < \infty$ then

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

To finish the argument, we verify $\mathbb{E}[\eta_0(W)^2] = \|\eta_0\|^2 < \infty$. By triangle inequality, Assumption 7, and Proposition E.4,

$$
\|\eta_0\| = \|\gamma_0(1, x) - \gamma_0(0, x) + \alpha_0(z, x)[v - \gamma_0(z, x)]\| \leq \|\gamma_0(1, x) - \gamma_0(0, x)\| + CC'.
$$

To bound the RHS, appeal to Proposition E.4:

$$
\|\gamma_0(1, x) - \gamma_0(0, x)\| \leq \|\gamma_0(1, x)\| + \|\gamma_0(0, x)\| \leq C\|\gamma_0\| < \infty.
$$
Proposition E.12. Suppose the conditions of Theorem 7.4 hold. Then the following holds.

1. $\hat{\theta} \xrightarrow{p} \theta_0$,
2. $J'J$ is nonsingular,
3. $E[\psi_0(W)^2] < \infty$,
4. $E[\{\phi(W, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta_0) - \phi(W, \hat{\gamma}_{-\ell}, \alpha_0, \theta_0) - \phi(W, \gamma_0, \hat{\alpha}_{-\ell}, \theta_0) + \phi(W, \gamma_0, \alpha_0, \theta_0)\}^2] \xrightarrow{p} 0$.

Proof. As in the proof of Proposition E.8, we can write

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

To lighten the proof, we slightly abuse notation as follows:

$$
\|\gamma_0 - \hat{\gamma}_{-\ell}\|^2 = E\{\gamma_0(Z, X) - \hat{\gamma}_{-\ell}(Z, X)\}^2|I_\ell|;
$$
$$
\|\alpha_0 - \hat{\alpha}_{-\ell}\|^2 = E\{\alpha(Z, X) - \hat{\alpha}_{-\ell}(Z, X)\}^2|I_\ell|.
$$

1. Convergence holds due to Proposition E.11.
2. Nonsingularity holds due to Assumption 7.
3. $E[\psi_0(W)^2] < \infty$ is immediate from $E[\eta_0(W)^2]$, which is proved in Proposition E.11 item 4.
4. It suffices to analyze

$$
E\left[\{\hat{\alpha}_{-\ell}(z, x) - \alpha_0(z, x)\}A(\theta_0)[\hat{\gamma}_{-\ell}(z, x) - \gamma_0(z, x)]^2\right]
= E\left[\left\{E\left[\{\hat{\alpha}_{-\ell}(z, x) - \alpha_0(z, x)\}A(\theta_0)[\hat{\gamma}_{-\ell}(z, x) - \gamma_0(z, x)]^2|I_{-\ell}\right]\right\}\right]
= E\left[\|\hat{\alpha}_{-\ell} - \alpha_0\|A(\theta_0)[\hat{\gamma}_{-\ell} - \gamma_0]\right]
\leq E\left[\|\hat{\alpha}_{-\ell}A(\theta_0)[\hat{\gamma}_{-\ell} - \gamma_0]\| + \|\alpha_0A(\theta_0)[\hat{\gamma}_{-\ell} - \gamma_0]\|\right].
$$
Consider the first term. By Hölder’s inequality, Assumption 2, and either Lemma 7.1 or Lemma 7.2, we have

$$\left| \hat{\alpha} - \ell(z, x) \right| = \left| \hat{\rho} b(z, x) \right| \leq \left| \hat{\rho} \right| \leq 1 |b(z, x)|_\infty = O_p(1).$$

It follows by Assumption 6 that

$$\| \hat{\alpha} - \ell A(\theta_0) [\hat{\gamma} - \gamma_0] \| = O_p(1) \| \hat{\gamma} - \gamma_0 \| = O_p(n^{-d_\gamma}) \xrightarrow{p} 0.$$ 

Consider the second term. By Assumption 6 and Assumption 7, we have

$$\| \alpha_0 A(\theta_0) [\hat{\gamma} - \gamma_0] \| \leq C A(\theta_0) \| \hat{\gamma} - \gamma_0 \| \xrightarrow{p} 0.$$

\(\square\)

**Proof of Theorem 7.4.** The proof now follows from Chernozhukov et al. [2016, Theorems 16 and 17]. In particular, Proposition E.7 verifies Chernozhukov et al. [2016, Assumption 4], Proposition E.8 verifies Chernozhukov et al. [2016, Assumption 5], Proposition E.9 verifies Chernozhukov et al. [2016, Assumption 6], Proposition E.10 verifies Chernozhukov et al. [2016, Assumption 7], and Proposition E.12 verifies the additional conditions in Chernozhukov et al. [2016, Theorems 16 and 17]. Finally, note that parameter \(\theta_0\) is exactly identified; \(J\) is a square matrix, the GMM weighting can be taken as the identity matrix, so the formula for the asymptotic covariance matrix simplifies. \(\square\)

**E.3. Extensions.**

**Proof of Corollary 7.5.** Immediate from Theorem 7.4 and Newey and McFadden [1994, Section 9]. \(\square\)
Proof of Theorem 7.6. Let $c_a$ be the $(1 - a)$-quantile of $|\mathcal{N}(0, \Sigma)|_\infty$ where $\Sigma = S^{-1/2}CS^{-1/2}$ and $S = \text{diag}(C)$. We first show that this critical value ensures correct (asymptotic) simultaneous coverage of confidence bands in the form of the rectangle

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

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 $d = \dim(U)$. Then the simultaneous coverage probability is

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

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

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

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

$$= 1 - a + o(1).$$

Gaussian multiplier bootstrap is operationally equivalent to approximating $c_a$ with $\hat{c}_a$ calculated in Estimator 6.4, which is based on the consistent estimator $\hat{C}$. □
Figure A.1. Numerical stability simulation: Trimming.

(A) $\mathbb{P}(Y(0) \leq y \mid D^{(1)} > D^{(0)})$

(B) $\mathbb{P}(Y(1) \leq y \mid D^{(1)} > D^{(0)})$

Notes: Simulation performance of $\kappa$ weight, DML and Auto-$\kappa$ estimators for estimating the counterfactual distribution, where the grid point is specified on the horizontal axis. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid points mark the median. The propensity scores are specified in (B.1).
Observations with extreme propensity scores $\hat{\pi} \not\in [10^{-12}, 1 - 10^{-12}]$ are dropped.
Figure A.3. Numerical stability simulation: Censoring.

\((a)\ P(Y^{(0)} \leq y \mid D^{(1)} > D^{(0)})\)

\((b)\ P(Y^{(1)} \leq y \mid D^{(1)} > D^{(0)})\)

Notes: Simulation performance of $\kappa$ weight, DML and Auto-$\kappa$ estimators for estimating the counterfactual distribution, where the grid point is specified on the horizontal axis. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid points mark the median. The propensity scores are specified in (B.1). Observations with extreme propensity scores are censored by setting $\hat{\pi} < 10^{-12}$ to be $10^{-12}$ and $\hat{\pi} > 1 - 10^{-12}$ to be $1 - 10^{-12}$.
Table A.1. Bias, RMSE, and coverage simulation. Smooth propensity score.

(A) $P(Y^{(0)} \leq y \mid D^{(1)} > D^{(0)})$

| $y$   | $\kappa$ weight | DML  | Auto-$\kappa$ | $\kappa$ weight | DML  | Auto-$\kappa$ |
|-------|-----------------|------|---------------|-----------------|------|---------------|
| -2    | -0.001          | -0.033 | 0.022         | 0.017           | 0.037 | 0.027         |
| -1.5  | -0.001          | -0.032 | 0.019         | 0.028           | 0.042 | 0.032         |
| -1    | -0.002          | -0.030 | 0.014         | 0.042           | 0.051 | 0.042         |
| -0.5  | 0.000           | -0.022 | 0.009         | 0.055           | 0.058 | 0.053         |
| 0     | 0.000           | -0.016 | 0.000         | 0.061           | 0.061 | 0.057         |
| 0.5   | -0.002          | -0.010 | -0.009        | 0.056           | 0.055 | 0.053         |
| overall | -0.001      | -0.024 | 0.009         | 0.043           | 0.051 | 0.044         |

(B) $P(Y^{(1)} \leq y \mid D^{(1)} > D^{(0)})$

| $y$   | $\kappa$ weight | DML  | Auto-$\kappa$ | $\kappa$ weight | DML  | Auto-$\kappa$ |
|-------|-----------------|------|---------------|-----------------|------|---------------|
| -2    | 0.000           | -0.000 | 0.006         | 0.004           | 0.004 | 0.007         |
| -1.5  | 0.000           | -0.001 | 0.006         | 0.007           | 0.007 | 0.009         |
| -1    | 0.001           | -0.002 | 0.006         | 0.012           | 0.011 | 0.013         |
| -0.5  | 0.001           | -0.005 | 0.005         | 0.016           | 0.016 | 0.017         |
| 0     | 0.001           | -0.009 | 0.003         | 0.022           | 0.023 | 0.022         |
| 0.5   | -0.000          | -0.013 | 0.001         | 0.026           | 0.028 | 0.026         |
| overall | 0.000        | -0.005 | 0.005         | 0.015           | 0.015 | 0.016         |

Notes: Simulation performance of $\kappa$ weight, DML and Auto-$\kappa$ estimators for the complier counterfactual outcome distribution, where the grid point for the distribution is specified in the first column. The last row averages the performance across grid points. Columns 1–3 compare the bias of different estimators and columns 4–6 compare the RMSE of different estimators. The best performing estimators are in bold. The simultaneous confidence bands based on the Auto-$\kappa$ estimator have coverage probability of 98.3% for the CDF of $Y^{(0)}$ and 97.8% for the CDF of $Y^{(1)}$, which are quite close to the nominal level of 95%. The propensity scores are specified in (B.2).
**Figure A.5.** Numerical stability simulation: Smooth propensity score.

(a) \( P(Y^{(0)} \leq y \mid D^{(1)} > D^{(0)}) \)

(b) \( P(Y^{(1)} \leq y \mid D^{(1)} > D^{(0)}) \)

Notes: Simulation performance of \( \kappa \) weight, DML and Auto-\( \kappa \) estimators for estimating the counterfactual distribution, where the grid point is specified on the horizontal axis. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid points mark the median. The propensity scores are specified in (B.2).