Sharp tests of instrument validity for general outcome distributions

Ricardo E. Miranda*†‡

November 14, 2026

Please click here for the most recent version

Abstract

In this paper I study nonparametric potential outcome instrumental variable models with multi-valued treatments and shape restrictions. I develop novel identification and inference results to obtain the first specification test that incorporates all observable implications in this class of models when outcomes are continuous. The main identification result shows that, even if the outcome has infinite support, the sharp testable implications of the exclusion restriction are equivalent to a finite system of linear inequalities. The resulting linear system involves parameters defined by nondifferentiable functions of the data for which existing inference procedures are not applicable. I prove that these parameters are equivalent to the expected loss of nonparametric classifiers. Leveraging this equivalence, I develop a joint inference procedure and obtain a formal test of instrument validity. I apply the proposed test to the Oregon Health Experiment (OHE). While previous research on the OHE demonstrates an impact on insurance take-up, I reject the hypothesis that the intervention affected insurance coverage only on the extensive margin. Leveraging a shape constraint that models this empirical finding, I obtain estimates of the effect of insurance on health and hospital utilization that are qualitatively consistent with previous findings but reflect the intensive margin effect of the intervention.

JEL classification: C12, C21, C25, C26, C31, C36.

*Department of Economics, Duke University, ricardo.e.miranda@duke.edu

†I am grateful to Adam Rosen and Michael Pollmann for their encouragement, and guidance throughout the writing of this paper. I thank my committee members Matt Masten and Arnaud Maurel, as well as Paul Gertler and Christopher Walker for their invaluable suggestions. I thank participants at the Duke Microeconometrics Breakfast, the Triangle Econometrics Conference, and the ITAM Alumni Conference for their helpful comments. All errors are my own.

‡This paper supersedes the working paper *On the Falsification of Discrete Instrumental Variables for Heterogeneous Treatment Effects*, posted on arXiv on January 20, 2026 (https://arxiv.org/abs/2601.14464).

1 Introduction

Instrumental variables (IV) is one of the most widely used methods in empirical economics. Following the seminal work of Imbens and Angrist (1994), a common approach to IV is to leverage potential outcomes models with shape restrictions (such as no-defiers or monotonicity) to identify causal parameters without restricting treatment effect heterogeneity. Existing tests for the validity of IV models with heterogeneous treatment effects either apply to models that impose specific shape restrictions, or require that the outcome be discrete or discretized for implementation. When the outcome is continuous, tests that use discretization don't exhaust the model's observable implications.

In this paper, I combine novel identification and inference results to propose the first validity test for IV models with heterogeneous treatment effects that exhausts the observable implications of outcomes with potentially infinite support and is compatible with a large class of nonparametric shape restrictions. The test is sharp: it employs a set of observable implications that is logically equivalent to the constraints that the model's assumptions impose on the data generating process.

The observable implications are a linear system of inequalities that involves integrals of the pointwise minimum of two or more densities, a class of parameters for which no valid joint inference methods exist. I therefore develop a method for inference on these parameters based on nonparametric classification. This inferential problem also arises in econometric characterizations beyond IV validity, as well as in other disciplines such as medicine, political science, and biology. The proposed method may therefore be of independent interest.

Finally, for IV models with binary instruments, I propose a variant of the test based on an observational equivalence between ruling out specific treatment switches and restricting how the instrument might affect unit's potential preferences over treatments. This variant identifies specific behaviors that exist in the population but are ruled-out by the model's assumptions, has a simple microeconomic interpretation, and eliminates nuisance parameters simplifying implementation of the test. I illustrate this framework in an application to the Oregon Health Experiment (OHE), where I reject the assumption that access to Medicaid affected insurance coverage only on the extensive margin, and propose alternative restrictions that are both consistent with the data and allow point identification of the effect of increased insurance coverage on health and hospital utilization

Identification: The key identification result of this paper is to transform the infinite set of inequalities that define the model into an equivalent, finite one.

I consider a potential outcomes model with a discrete instrument Z, a discrete treatment X, and an outcome Y with unrestricted and potentially multi-dimensional support. The model is defined by potential treatments $X = X(Z)$ and potential outcomes $Y = Y(X, Z)$.

The model imposes three assumptions: *i*) Z is independent of potential treatments and potential outcomes as if randomly assigned, *ii*) Z affects X without violating some prespecified restrictions (e.g. no-defiers or monotonicity), and *iii*) an exclusion restriction on Z, which reduces potential outcomes to $Y = Y(X)$. The goal is to obtain a sharp set of observable implications for which a fully nonparametric test is feasible.

The model is not falsified if potential treatment and outcome functions exist that rationalize the observed conditional outcome probabilities:

$$\mathbb{P}[Y \in B_Y, X = x | Z = z] = \mathbb{P}[Y(x) \in B_Y, X(z) = x | Z = z], \tag{1}$$

for every x and z in the support of X and Z, and every measurable subset B_Y of the support of Y. This includes as a particular case the set of equations:

$$\mathbb{P}[X = x | Z = z] = \mathbb{P}[X(z) = x | Z = z], \tag{2}$$

for all possible values of X and Z.

Since X and Z have finite support, selection into treatment (equation 2) is described by a finite set of equations. However, if Y has infinite support, there are infinitely many measurable subsets B_Y and thus, infinitely many instances of equation 1. The insight that enables the test is that the sharp observable implications of the exclusion restriction combined with conditional outcome probabilities are equivalent to a finite set of linear restrictions on the selection part of the model. Thus, I reduce an infinite set of equations to an equivalent finite one.

The set of linear inequalities that expresses the sharp observable implications of the exclusion restriction involves a finite set of parameters that I refer to as subdensity overlaps. I define subdensities as the functions that result from restricting the joint densities of the outcome and treatment conditional on a given instrument realization to a fixed treatment value. That is, the functions $f_{Y,X|Z=z}(\cdot, x)$ for some fixed values of x and z. The overlap of n functions is the area that lies below all of them, or, in

other words, the integral of their pointwise minimum. Thus, if $\mathcal{Y}$ denotes the support of Y, the target parameters take the form:

$$\int_{\mathcal{Y}} \min_k \{f_{Y,X|Z=z_k}(y, x_l)\} dy$$

Inference: Estimation and joint inference on subdensity overlaps is a nonstandard problem for which no previous, valid, fully nonparametric method exists. I note that subdensity overlaps are equivalent to the loss of a nonparametric classifier. This insight allows me to propose a valid, asymptotically unbiased and normal estimator.

The nonstandard nature of the problem arises from the region of the support of the outcome in which two or more subdensities achieve the minimum. This region is referred to as the contact set. The main inferential challenge follows from the fact that the minimizer is not unique in this region.

Ruling the contact set out is problematic. First, it can arise naturally when compliance (i.e. the effect of Z on X) is correlated with the outcome (for example, if the instrument does not affect the treatment choices of the healthiest subjects in a given population). Second, it also arises whenever the instrument has no effect on treatment adoption (i.e. instrument irrelevance), which, in the context of testing, might not be an *ex-ante* plausible assumption. Finally, the assumption that the contact set is negligible is fundamentally untestable. Thus, a test robust to this possibility is more credible.

Furthermore, the contact set is the reason why discretization based tests are not sharp. Discretization is equivalent to integration with a coarse grid, and it is precisely around the points where subdensities cross that coarse grids overestimate or underestimate the integral of the minimum. Hence, discretization is only sharp if the contact set is known or estimated with high precision. To obtain a sharp test, the problems associated with the contact set cannot be avoided.

The contact set is problematic because nondifferentiability of the minimum impedes using the delta method, while the nonparametric nature of the problem implies there are no guarantees that extensions of the delta method for directionally differentiable functions such as Fang and Santos (2019) work.

Moreover, if the contact set is a non-negligible subset of the support of the outcome, the asymptotic behavior of plug-in estimators is non standard. Any finite-sample subdensity estimator will have a statistical noise component that shifts the estimator away from the true value of the target function. If two subdensities are estimated from

samples that grow at the same rate, the magnitude of the noise component of both estimators will be of the same order, and its ratio will not converge to zero. Hence, for any given sample size, the identity of in-sample minimizers inside the contact set is determined by statistical noise. As a consequence, the limit of the estimated minimizer might not converge in probability to a fixed constant complicating the asymptotic analysis of plug-in estimators.

To circumvent these problems, I note that inference on subdensity overlaps is equivalent to inference on the expected loss of a classifier that uses the outcome and treatment values to predict the instrument. The goal is not to learn the identity of the minimizer at every point in the support of Y and X, but the aggregate frequency with which the outcome, treatment and instrument value associated with the smallest subdensity are jointly realized. The main advantage of this approach is that, precisely in the set in which subdensities are equal, any classification rule attains the exact same expected loss. Hence, while inside the contact set the estimated classifier may behave erratically in the limit as the sample size grows, its expected loss, which corresponds to the target parameter, is essentially deterministic.

I propose a consistent, asymptotically normal, and asymptotically unbiased estimator for the overlap of two or more subdensities. To derive its asymptotic properties, I combine the classification framework with uniform rates of convergence for kernel density estimators (Giné and Guillou (2002)) and a margin condition that controls the magnitude of classification errors in the proximity of the contact set. The estimator uses sample splitting. One fraction of the data is used to train a nonparametric classifier based on kernel density estimators, and the rest to evaluate its loss. Sample splitting ensures that the empirical loss of the trained classifier is unbiased for its theoretical loss.

In Monte-Carlo simulations, I show that the overlap estimators I propose have negligible bias, and that the corresponding confidence intervals achieve coverage close to the nominal level in finite samples. Given this valid method for inference on subdensity overlaps, testing instrument validity reduces to verifying that a fine linear system of inequalities with known coefficients admits at least one solution. This is a well studied problem for which inference procedures exist. Using the methodology from Cox and Shi (2023) I formally show that the resulting test has asymptotically correct size. A second set of Monte-Carlo simulations supports the theoretical properties of the test, and demonstrates its power against violations of the null.

Generalized no-defiers and empirical application: A common assumption in IV models with heterogeneous treatment effects is that some treatment switches cannot occur as the instrument changes (e.g. the no defiers assumption). I call restrictions in this class *generalized no-defiers assumptions*. For models with binary instruments that impose generalized no-defiers assumptions, I propose a specialized, simpler test based on a microeconomic interpretation of such restrictions.

Generalized no-defiers restrictions specify that, as the instrument increases, some treatments are weakly promoted (at least in relative terms) with respect to others. Consequently, if the assumption holds, adoption of the promoted treatments should weakly increase. I formalize the notion of *promoted treatments* in terms of restrictions on how the instrument affects the preferences of units in the population, and characterize how adoption of the promoted treatments *weakly increases* using a generalization of first order stochastic dominance (GFOSD) to arbitrary binary relations.

Additionally, the sharp testable implications of any generalized no-defiers restriction combined with the exclusion restriction are still a set of inequalities with a GFOSD interpretation. The difference is that the exclusion restriction imposes an upper bound on the probabilities associated with some of the promoted treatments, which results in a *tightening* of the GFOSD inequalities relative to the inequalities associated with the generalized no-defiers restriction alone.

These results provide a motivation for generalized no-defiers assumptions, an interpretable set of sharp testable implications, and a framework to formulate plausible restrictions that are consistent with the observed data.

I illustrate the proposed framework in an empirical application to the Oregon Health Experiment, a state-level lottery that randomized eligibility to the Medicaid program among low income individuals (Finkelstein et al. (2012)). Specifically, I test and reject the assumption that the intervention only affected insurance on the extensive margin proposed by Rose and Shem-Tov (2024). Instead, I find that the treatment also affected the duration of insurance coverage for households that would otherwise have acquired insurance for less than six months. That is, the intervention also had an effect on the intensive margin. Leveraging a shape restriction that models this empirical finding, I estimate the effect of increasing the length of insurance coverage periods on hospital utilization and health outcomes. My estimates are quantitavely consistent with previous results by Finkelstein et al. (2012) and Rose and Shem-Tov (2024) who found that increasing insurance led to health improvements and increases in hospital utilizations

of 0.2 and 0.1 standard deviations respectively.

The remainder of this paper is organized as follows: In the rest of this section I position this paper and its contribution in the literature. In section 2 I introduce the class of models and restrictions considered. In section 3 I derive the testable implications of the exclusion restriction, provide the finite sharp characterization for binary IV models, and extend the results to multi-valued instruments. In section 4 I specialize the characterization from section 2 to instrument-response-type restrictions and link them to GFOSD and random utility models. In section 5 I describe my method for joint estimation and inference on subdensity overlaps and derive its asymptotic properties. Then, I construct a sharp test for instrument validity from the proposed estimator. In section 6 I illustrate the performance of both the estimator and the test. In section 7, I use my inference method and the GFOSD characterization to study the effect of the OHE on the duration of insurance coverage, and the effects of increasing the duration of insurance coverage on health and hospital utilization. In section 8 I conclude.

Related literature: This paper contributes to the literature on testing the validity of instrumental variables models for heterogeneous treatment effects. The first to consider instrument validity for heterogeneous treatment effects where Balke and Pearl (1997), but without a formal inference procedure and only for binary outcomes. When both the instrument and treatment are binary and monotonicity is assumed, Huber and Mellace (2015) proposed a test based on moment inequalities and Kitagawa (2015) obtained the first sharp characterization. This later approach was refined by Mourifié and Wan (2017) who proposed a formal inference procedure for this particular case. More recently, Kwon and Roth (2024) showed that the instrument validity assumptions are equivalent to the *sharp null of full mediation* in the context of randomized experiments. Furthermore, they proposed a sharp characterization of the observable implications that covers more general shape restrictions as well as multi-valued treatments. For inference, however, their approach requires discretizing the outcome which results in a set of necessary but not sufficient conditions for instrument validity. That is, the resulting test might not be sharp.

The characterizations I propose are based on noting that the exclusion restriction imposes constraints on the distribution of the outcome for always takers (or, more generally, on units that realize the same treatment for at least two different values of

the instrument) and, consequently, constrains their population prevalence. This idea is present in Kwon and Roth (2024), as well as in papers that have studied robust identification and inference for IV models with one binary instruments and one binary treatment such as Kitagawa (2021), Noack (2021) and Kédagni et al. (2024). For binary instruments, my characterization is equivalent to that of Kwon and Roth (2024), but instead of taking a supremum over an infinite collection of sets, I aggregate an infinite set of restrictions into a single parameter using integration. I therefore interpret my results as extending the identification ideas from Kitagawa (2021) to specification tests and to models with multiple treatments and instruments.

For multi-valued instruments, the sharp testable implications of IV designs have been derived on a case-by-case basis (Bai et al., 2024c). Alternatively, general tests that are not sharp such as Sun (2023) or Frandsen et al. (2023) have been proposed. Recently, Kaido and Ponomarev (2025) proposed a general approach to falsifying and testing econometric models based on graph theory. The main advantage of this last approach is its generality. It covers multiple instrumental variables models with binary and multi-valued instruments and treatments, along with many other important econometric methods.

The method proposed by Kaido and Ponomarev (2025), however, is based on combinatorial graph theory and thus, it is naturally tailored for discrete outcome variables and restrictions on the support of latent types. When outcomes are continuous, their characterization relies on verifying their validity conditions among an infinite sequence of discretizing partitions that get increasingly finer, and become infinite on the limit. Instead, I note that shape restrictions on treatment adoption concern the finite part of the model while the exclusion restriction concerns the outcome, whose support might be infinite. Consequently, I approach each of these two components of the model in distinct ways. In particular, using integration, I aggregate the infinite set of restrictions that exclusion imposes into a finite collection of constraints. The result is a linear system of inequalities that is finite and exhausts the observable implications of the model.

The observable implications I propose apply to a wide range of IV methods for heterogeneous treatment effects. Imbens and Angrist (1994) showed how to use nonparametric restrictions on instrument-response heterogeneity to identify the LATE parameter. Subsequent work has extended their approach to other parameters of interest and a wide range of settings and shape restrictions. These extensions in-

clude re-interpretations of monotonicity (Heckman and Pinto (2018), Navjeevan and Pinto (2022)); alternative restrictions on how the instrument affects treatment take-up (Richardson and Robins (2010), De Chaisemartin (2017), Dahl et al. (2023), Goff (2024)); and generalizations to settings with multiple instruments and treatmens (Lee and Salanié (2020), Goff (2020), Mogstad et al. (2021), van't Hoff et al. (2023), Kazemi and Olsen (2024), Fusejima (2024), Bai et al. (2024a)). These methodological advances have been applied in diverse empirical contexts, including education and labor market studies (Kline and Walters (2016), Kirkeboen et al. (2016), Pinto (2021), Mountjoy (2022)). The results of these paper concern all the aforementioned papers.

The GFOSD characterization and its relation to nonparametric random utility models contributes to the literature that investigates the empirical content of response type restrictions (Goff (2024) and Bai et al. (2024b)) and their microeconomic foundations (Vytlacil (2002), Heckman and Pinto (2018), Navjeevan and Pinto (2022)).

To obtain the GFOSD inequalities I show that finding a distribution over potential treatments and potential outcomes that satisfies the sharp instrument validity conditions amounts to solving a well known transportation problem over a directed graph. This technique, and its equivalence to the maximum flow/minimum cut problems, was first used to characterize incomplete econometric models by Ekeland et al. (2010) and Galichon and Henry (2011). My approach differs in two ways: first, instead of considering transportation problems from the space of observed variables to the domain of latent unobservables, my transportation problem is defined between observed treatment probabilities conditional on different instrument values. Second, to incorporate the exclusion restriction, I extend their minimum-cut characterization method to a broader class of capacity constraints.

For binary instruments and discrete treatments, the optimal transport approach from Galichon and Henry (2011) and the combinatorial method by Kaido and Ponomarev (2025) are equivalent. To incorporate outcomes, Kaido and Ponomarev (2025) extend their graph representation of the model to the Cartesian product of the support of the treatment and the outcome. The result is a system of inequalities that grows polynomially or exponentially with he support of the outcome. On the other hand, my approach, which expresses the exclusion restriction as a capacity constraint on the transportation problem, yields a finite set of sharp inequalities whose cardinality is fixed and does not depend on the cardinality of the outcome's support.

All the testable implications I provide are equivalent to the requirement that a

particular linear system of inequalities admits at least one solution. Inference on the feasibility of linear systems and linear programs is an active area of research (Bai et al. (2022), Fang et al. (2023), Cho and Russell (2024), Goff and Mbakop (2025), Bai et al. (2026), Tamer and Walker (2026)) that is tightly related to inference methods for models characterized by moment inequalities such as Cox and Shi (2023) or Andrews et al. (2023).

To test IV validity using inferential methods for linear systems I first require a method for nonparametric joint inference on subdensity overlaps. This is a particular instance of the more general problem of estimating the integral of nondifferentiable transformations of a collection of functions or densities. Two closely related papers are Semenova (2025) and Levis et al. (2025) who consider inference on the integral of the pointwise minimum/maximum of smooth functions, and its applications to IV models. Neither of these papers studies testing instrument validity. For inference, Semenova (2025) assumes that the minimum is unique almost-everywhere, which eliminates the nonregularity issue. Levis et al. (2025) propose a similar solution, and also consider the non-negligible contact set case. Their solution is to restore regularity using a smooth (*soft-min*) approximation to the minimum function. While the soft-min approach would be valid in the context of this paper, the smooth approximation to the minimum shifts the target parameter. Consequently, the resulting inequalities would be a set of necessary conditions for IV validity, but not the sharp testable implications

To the best of my knowledge, only two papers have considered exact inference on parameters defined by the integral of the minimum/maximum of a collection of functions that is robust to non-negligible contact sets. Anderson et al. (2012) consider a plug-in approach for inference on a single overlap parameter. Unfortunately, their variance estimator is not compatible with existing inference methods for linear systems. Luedtke and Van Der Laan (2016) propose using online estimation to conduct inference on the mean outcome under a nonunique optimal policy. While similar nonregularity issues arise in both problems, in subdensity overlap estimation, realized and potential outcomes enter the target parameter through their distribution rather than directly. Thus, the online estimation solution is not applicable without substantial modifications.

The key argument sustaining my inference procedure is that, in the contact set, the loss of a nonparametric classifier is invariant, and hence, an estimator based on the empirical loss of a classifier is robust to the erratic behavior of the minimum inside the contact set. This insight reduces the inference problem to controlling the behavior of

the classifier out of the contact set. To do so, I combine sample splitting, a margin assumption that controls the misclassification rate near the boundary of the contact set (Tsybakov (2004), Audibert and Tsybakov (2007)), and classical results on the rates of uniform convergence for kernel-density estimators by Giné and Guillou (2002).

My method for inference on density and subdensity overlaps might be of independent interest. First, other IV identification methods such as Kitagawa (2021) or Kédagni et al. (2024) obtain identification results that involve or are related to the vector of subdensity overlaps from this paper, but do not have inference methods that allow for continuous outcomes. Second, similar parameters appear in other econometric settings such as Borusyak (2015) or Meango et al. (2025). Finally, density overlaps are often used as a nonparametric measure of similarity between populations in other disciplines such as political science, ecology, and medicine (see Dey et al. (2025) for a recent review).

2 IV models with heterogeneous treatment effects

The model has three observable random variables: an outcome Y, a treatment X and an instrument Z with joint distribution $F_{Y,X,Z}$. Both X and Z have discrete finite supports denoted $\mathcal{X} = \{x_0, x_1, ..., x_{L-1}\}$ and $\mathcal{Z} = \{z_0, z_1, ..., z_{K-1}\}$ respectively. The support of Y which I denote as $\mathcal{Y}$ has unrestricted cardinality and finite dimension $d \geq 1$. I assume that a known measure μ_Y over $\mathcal{Y}$ exists.

I adopt the potential outcomes framework. The joint distribution of (Y, X, Z) is determined by the potential treatment and potential outcome functions $X(z) : \mathcal{Z} \mapsto \mathcal{X}$ and $Y(x, z) : \mathcal{X} \times \mathcal{Z} \mapsto \mathcal{Y}$. These functions are unobserved random variables, and their joint distribution along with the observed marginal distribution of Z fully determine the model.

Formally, for any measurable subset of the support of the outcome $B_Y \subset \mathcal{Y}$, and any fixed treatment and instrument values $x_l \in \mathcal{X}$ and $z_k \in \mathcal{Z}$, the joint distribution of (Y, X, Z) satisfies:

$$\mathbb{P}[Y \in B_Y, X = x_l, Z = z_k] = \mathbb{P}[Y(x_l, z_k) \in B_Y, X(z_k) = x_l, Z = z_k]$$

I assume that all other relevant covariates are exogenous. Since these covariates play no role in the characterizations I provide, all identification arguments are understood to be conditional on them, and I omit them from the formal statements and proofs.

Unless otherwise noted, I present and discuss all the results using the IV notation and terminology. All results would also hold for the sharp null of full mediation from Kwon and Roth (2024) after appropriately relabeling the values of X as mediators and the values of Z as treatments.

I denote the joint density of X and Y conditional on $Z = z_k$ induced by $F_{Y,X,Z}$ as $f_{Y,X|Z=z_k}(y, x)$.

2.1 Instrument validity and response types

For Z to be a valid instrument, two assumptions are typically invoked, instrument independence, and an exclusion restriction:

Assumption 2.1 (Instrument validity)**.**

1. ***Instrument independence****: The potential outcomes and potential treatment functions $Y(\cdot, \cdot)$ and $X(\cdot)$ are independent from the instrument Z.*
2. ***Exclusion restriction****: $Y(x_l, z_k) = Y(x_l)$ almost surely, for every $x_l \in \mathcal{X}$ and $z_k \in \mathcal{Z}$*

Put in words, the instrument must be independent from potential outcomes and treatments as if it was randomly assigned, and it must have no effect on outcomes after accounting for treatment.

Assumption 2.1 implies that the observed joint distribution of (X, Y, Z) is fully determined by the observed distribution of Z, and the unobserved distribution of the potential outcomes and potential treatments functions. This motivates the following definitions:

Definition 2.1.

1. *An instrument-response type $t^Z \in \mathcal{T}_Z$ is a function $t^Z : \mathcal{Z} \longrightarrow \mathcal{X}$ that maps instrument values to treatment values.*
2. *A treatment-response type $t^X \in \mathcal{T}_X$ is a function $t^X : \mathcal{X} \longrightarrow \mathcal{Y}$ that maps treatment values to outcome values.*
3. *A response type t is a pair $t = (t_z, t_x) \in \mathcal{T} = \mathcal{T}_Z \times \mathcal{T}_X$.*

An important feature of response types is that by construction they satisfy the exclusion restriction. For any $x_l \in \mathcal{X}$, a response type $t = (t^Z, t^X)$ assigns a unique outcome value $t^X(x_l)$ when $X = x_l$ regardless of the value that the instrument takes. This is why, under assumption 2.1, the model is fully determined by the underlying distribution of response types and the marginal distribution of Z.

Since Y is μ_y-measurable, and $\mathcal{X}$ is finite, the set $\mathcal{T}$ is measurable with respect to the product measure induced by μ_Y and the count measure on $\mathcal{X}$. I denote the resulting measure space as $(\mathcal{T}, \mathcal{B}_T)$. This allows me to formally define distributions over latent response types.

Definition 2.2.

1. *A distribution over response types is a probability measure $g : \mathcal{B}_T \longrightarrow [0,1]$ over the measure space $(\mathcal{T}, \mathcal{B}_T)$.*

2. *A distribution over instrument-response types is a function $p : \mathcal{T}_Z \longrightarrow [0,1]$ such that $p(t^Z) \geq 0$ for all $t^Z \in \mathcal{T}_Z$, and $\sum_{t^Z \in \mathcal{T}_Z} p(t^Z) = 1$.*

Note that since both $\mathcal{Z}$ and $\mathcal{X}$ are finite, a distribution over response types is just a finite vector of nonnegative numbers that add up to one.

Remark 2.1. *Any distribution over response types g defines a unique distribution over instrument-response types f_g according to the following equation:*

$$\begin{aligned} p_g(t'^Z) :=& g\Big(\big\{t = (t^Z, t^X) \in \mathcal{T} : t^Z = t'^Z, t^X \in \mathcal{Y}^L\big\}\Big), \qquad \forall\ t'^Z \in \mathcal{T}_Z, \\ =& g\Big(\big\{t = (t^Z, t^X) \in \mathcal{T} : t^Z = t'^Z\big\}\Big), \qquad \forall\ t'^Z \in \mathcal{T}_Z. \end{aligned}$$

If assumption 2.1 holds, then an unobserved true distribution over response types g_0 must exist that generates the observed distribution of (Y, X, Z).

Definition 2.3 (Consistency)**.** *A distribution over response types $g \in \Delta(\mathcal{T})$ is consistent with observed conditional outcome and treatment probabilities if:*

$$\mathbb{P}[Y \in B_Y, X = x_l | Z = z_k] = g_0\Big(\big\{(t^Z, t^X) \in \mathcal{T} : (t^Z)_k = x_l, (t^X)_l \in B_Y\big\}\Big), \qquad (3)$$
$$\forall x_l \in \mathcal{X}, z_k \in \mathcal{Z},\ \text{and } B_Y \in \mathcal{B}_Y.$$

In the previous equation, the observed conditional probability on the left hand side must correspond to the mass that g_0 assigns to response types whose potential

treatment is equal to x_l when the instrument takes the value z_k and realize outcomes in B_Y when treatment is equal to x_l. That is, the distribution g_0 rationalizes the observed conditional outcome and treatment probabilities.

Let $\Delta(\mathcal{T})$ be the set of all distributions over response types, assumption 2.1 is falsified if no $g \in \Delta(\mathcal{T})$ exists that satisfies equation 3.

2.2 Shape restrictions

The class of shape restrictions I consider is general. It covers any combination of restrictions that either rule out specific instrument response types, or impose inequality constraints on their prevalence or relative frequency. Formally stated, I allow for any restriction than can be expressed as a set of linear inequality constraints on the admissible distributions over instrument-response types.

The most prominent example of a restriction that rules out specific response types is monotonicity when both X and Z are binary, switcing from $Z = z_0$ to $z = z_1$ is assumed to promote treatment take-up, and thus defiers[1] are ruled out from the population. Subsequent work has explored relaxations and extensions of this assumption to settings with multiple instruments and multiple, ordered and unordered, treatments (Heckman and Pinto (2018),Goff (2020), Mogstad et al. (2021), van't Hoff et al. (2023), Lee and Salanié (2020), Bai et al. (2024c)). Both, monotonicity and its generalizations rely on ruling out particular instrument-response types and thus are covered by my framework.

Alternatively, researchers may impose inequality restrictions on the prevalence or relative frequency of certain instrument-response types—for example, requiring that the proportion of compliers exceeds that of defiers (De Chaisemartin (2017)), or imposing upper bounds on the share of defiers for sensitivity analysis as Huber (2014), Noack (2021) or Yap (2025)). Assumptions of this type are also covered by my framework.

To formalize the shape restrictions allowed, note that the total number of possible instrument-response types is $|\mathcal{T}_Z| := |\mathcal{X}|^{|\mathcal{Z}|} = (L)^K$. Therefore, a distribution p over $\mathcal{T}_Z$ can be represented as a vector $\boldsymbol{p}$ in $\mathbb{R}_+^{|\mathcal{T}_Z|}$, where each coordinate corresponds to the probability that $p \in \Delta(\mathcal{T}_Z)$ assigns to a different $t \in \mathcal{T}_Z$.[2]

[1] i.e. units that switch from being treated to being untreated as the instrument changes from 0 to 1.

[2] To formally define $\boldsymbol{p}$ is suffices to note that the set of response types is finite and thus can be well ordered. Then, the $n-th$ coordinate of $\boldsymbol{p}$ represents the probability that the distribution $p \in \Delta(\mathcal{T}_Z)$ assigns to the $n-th$ response type. The specific ordering of response types is not relevant as long as

I consider any shape restriction that takes the form $A_Z \boldsymbol{p} \leq \boldsymbol{r}$ where A_Z is an M by $|\mathcal{T}_Z|$ matrix with known coefficients for some fixed $M \geq 1$, and $\boldsymbol{r}$ is an M dimensional vector with entries that are either known or point identified from the data. I use the subscript Z to emphasize that the matrix A_Z encodes restrictions that pertain to treatment assignment as a function of the instrument.

Assumption 2.2. *Let $M \geq$ be a fixed integer, let A_Z be a known fixed matrix in $\mathbb{R}^{M \times |\mathcal{T}_Z|}$, and $\boldsymbol{r}_Z \in \mathbb{R}^M$ a known vector that depends on the joint distribution of (Y, X, Z).*

A distribution over response types $g_0 \in \Delta(\mathcal{T})$ exists such that if $\boldsymbol{p}_{g_0}$ denotes the vector representation of p_{g_0}, the following system of inequalities holds:

$$A_Z \boldsymbol{p}_{g_0} \leq \boldsymbol{r}_Z. \tag{4}$$

Note that strict equality restrictions can be incorporated by representing them as pairs of inequality constraints. In particular, if the k-th entry of $\boldsymbol{p}$ corresponds to the instrument-response type $t^Z \in \mathcal{T}_Z$, the restriction that t^Z occurs with zero probability can be imposed by including the canonical basis vectors e_k and $-e_k$ as rows of A_Z, and setting the corresponding entries of $\boldsymbol{r}_Z$ to zero.[3] Thus, assumption 2.2 encompasses any restriction that rules out specific response types. In the online appendix to this paper I provide examples of models and restrictions encompassed by assumption 2.2[4].

To conclude the discussion of shape restrictions, note that if no distribution over instrument-response types $p \in \Delta(\mathcal{T}_Z)$ exists such that $\boldsymbol{p}$ satisfies inequality 4 and is consistent with observed conditional treatment probabilities, then, the restrictions on instrument-response heterogeneity imposed by assumption 2.2 are falsified.

Proposition 2.1 (**Sharp observable implications of instrument independence and assumption 2.2**). *A distribution over instrument-response types $p \in \Delta(\mathcal{T}_Z)$ is consistent with assumption 2.2 and observed conditional treatment probabilitites, if and only if it satisfies the following system of linear equalities and inequalities:*

$$A_Z \boldsymbol{p} \leq \boldsymbol{r}_Z,$$
$$A_X \boldsymbol{p} = \boldsymbol{r}_X.$$

it is well defined.

[3] e_k denotes the vector that takes the value 0 in every entry except the k-th which is equal to 1.

[4] The online appendix can be found at https://ricardomirandam.github.io/files/OnlineAppendix.pdf

Where A_Z and $\boldsymbol{r}_Z$ are defined as in assumption 2.2 and the system $A_X\boldsymbol{p} = \boldsymbol{r}_X$ stacks the following set of equations:

$$\mathbb{P}[X_i = x_l | Z_i = z_k] = \sum_{t^Z \in \mathcal{T}_Z : t^Z(z_k) = x_l} p(t^Z), \qquad \forall z_l \in \mathcal{Z}, x_k \in \mathcal{X}.$$

The proof of proposition 2.1 is straightforward and follows from the previous discussion.

3 Sharp testable implications

In this section I show how to express the restrictions implied by the observed conditional distribution of the outcome combined with the exclusion restriction as a finite set of linear restrictions on instrument-response probabilities. More specifically, a collection of upper bounds on the joint population prevalence of different groups of instrument-response types. This result, combined with proposition 2.1 transforms the potentially infinite set of linear inequalities that describe the original model into an equivalent finite one. First, I consider binary instruments for which I prove the sharpness of my characterization. Then, I extend the results to multi-valued instruments and prove that the extension is sharp in the three instruments case.

3.1 Binary instruments

In the binary instrument case, the sharp testable implications are just a collection of upper bounds on the share of units that realize the same treatment at every instrument value, that is, always takers. To define these bounds, I first introduce additional notation.

For any $x_l \in \mathcal{X}$, I refer to units that realize treatment x_l at all possible instrument values as x_l-always takers. The population frequency of always-x_l-takers is given by the quantity $\mathbb{P}[X(z_0) = X(z_1) = x_l]$.

Let $p\left(AT_{x_l}\right)$ denote the probability that a distribution over response types $p \in \Delta(\mathcal{T}_Z)$ assigns to always-x_l takers. Formally:

$$p\left(AT_{x_l}\right) := \sum_{t^Z \in \mathcal{T}_Z} p(t^Z) \boldsymbol{I}_{\{t^Z(z_0) = t^Z(z_1) = x_l\}},$$

where $\boldsymbol{I}_{\{t(z_0)=t(z_1)\}}$ is an indicator that response type t^Z realizes treatment x_l at both values of Z. Note that this equation is linear in $\boldsymbol{p}$ with known, deterministic coefficients.

Having defined always-takers and their population prevalence, I now define an observable upper bound on $p\left(AT_{x_l}\right)$.

Definition 3.1 (**Subdensity overlap associated with** x_l)**.** *Let $f_{Y,X|Z=z_k}$ denote the joint distribution of Y and X conditional on $Z=z_k$. Then, for any fixed $x_l \in \mathcal{X}$, define the function $\Psi_{x_l}: \mathcal{Y} \longrightarrow \mathbb{R}_+$ as follows:*

$$\psi_{x_l}(y) = \min_{k=0,1} \left\{ \left[f_{Y,X|Z=z_k}(y, x_l) \right] \right\}.$$

Then, the subdensity overlap associated with x_l is:

$$\Psi_{x_l} := \int_{\mathcal{Y}} \psi_{x_l}(y) dy. \tag{5}$$

As the next theorem shows, the sharp observable implications of the exclusion restriction are equivalent to the condition $p(AT_{x_l}) \leq \Psi_{x_l}$ for every $x_l \in \mathcal{X}$. To express this condition in matrix form, let t_j^Z denote the j-th instrument-response type according to some fixed, arbitrary, pre-specified order over instrument-response types. Let A_Y be a $L \times |\mathcal{T}_Z|$ matrix with entires $(A_Y)_{l,j} = \boldsymbol{I}_{\{t_j^Z(z_0)=t_j^Z(z_1)=x_l\}}$ and let $\boldsymbol{\Psi}$ denote the vector that has Ψ_{x_l} in its $l+1$-th coordinate. Then, the sharp observable implications are expressed by the matrix inequality $A_Y \boldsymbol{p} \leq \boldsymbol{\Psi}$.

Theorem 3.1. *[Sharp observable implications of binary instruments]*
The following statements are equivalent:

1. *A distribution over response types $g \in \Delta(\mathcal{T})$ that satisfies, assumption 2.1 (instrument validity), assumption 2.2 (shape restrictions) and definition 2.3 (consistency with observed probabilities) exists.*

2. *A distribution over instrument-response types $p \in \Delta(\mathcal{T}_Z)$ exists such that p satisfies:*

$$\begin{aligned} A_Z \boldsymbol{p} &\leq \boldsymbol{r}_Z, \\ A_X \boldsymbol{p} &= \boldsymbol{r}_X, \\ A_Y \boldsymbol{p} &\leq \boldsymbol{\Psi}. \end{aligned}$$

Where A_Z, A_X, r_Z and r_X are finite dimensional matrices as defined in assumption 2.2 and proposition 2.1.

The formal proof of theorem 3.1 is appendix A Here, I provide a graphical sketch of the proof and compare this characterization to existing sharp falsification procedures for binary instrument models.

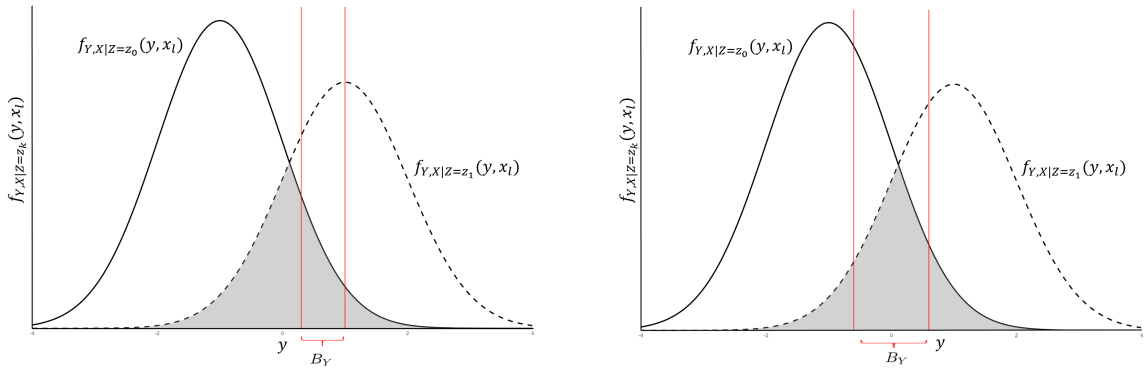
To see why the inequality $p(AT_{x_l}) \leq \Psi_{x_l}$ must hold for every $x_l \in \mathcal{X}$ note that always-x_l takers are a potentially strict subset of all the units that realize treatment x_l both when $Z = z_0$ and when $Z = z_1$. Therefore, if the exclusion restriction holds, the group of always-x_l takers that realize outcomes in any measurable set $B_Y \subset \mathcal{B}_Y$ must also be a (potentially strict) subset of the units that realize outcomes in B_Y when $X = x_l$ and $Z = Z_k$ for both values of $z_k \in \{z_0, z_1\}$. Thus, the fraction of always takers with outcomes in B_Y must be weakly dominated by the integral of $\psi_{x_l}(y)$ over B_Y (see figure 1 for a graphical representation). Setting $B_Y = \mathcal{Y}$ yields the desired result.

To see the sufficiency of the restrictions, suppose that some $p \in \mathcal{T}_Z$ is consistent with observed conditional treatment probabilities and satisfies $p(AT_{x_l}) \leq \Psi_{x_l}$ for every $x_l \in \mathcal{X}$. To construct a distribution over response types consistent with the observed joint distribution of (Y, X, Z), the unobserved marginal distribution of potential outcomes for always-x_l takers when $X = x_l$ can be set to $\frac{\psi_{x_l}(y)}{\Psi_{x_l}}$ so that always-x_l takers account for a share $\frac{p(AT_{x_l})}{\Psi_{x_l}}$ of the mass represented by Ψ_{x_l}. This construction guarantees that the unconditional subdensity associated to the outcomes of always-x_l-takers is dominated by $f_{Y,X|Z=z_0}(y, x_l)$ and $f_{Y,X|Z=z_1}(y, x_l)$. Moreover, this construction ensures that the distribution of the outcome for always takers is invariant to instrument changes, which implies that the exclusion restriction is satisfied.

It remains to distribute a share $1 - \frac{p(AT_{x_l})}{\Psi_{x_l}}$ of the area Ψ_{x_l} as well as any mass strictly above at least one subdensity across response types that realize x_l at just one instrument value (i.e. x_l-compliers and x_l-defiers). Assigning potential outcomes to these units is simple because –since they realize treatment x_l at just one instrument value– the exclusion restriction does not bind their potential outcomes. Figure 2a provides a visual illustration of this procedure. The formal construction is described in appendix A.

A benchmark case is when assumption 2.2 imposes that take-up of treatment x_l can only increase (or decrease) as the instrument changes. This case corresponds to the classical monotone binary instrument and treatment model from Imbens and Angrist (1994), Kitagawa (2015) and Mourifié and Wan (2017). It also corresponds to maximal

Figure 1: Upper bounds on always takers.



(a) (b)

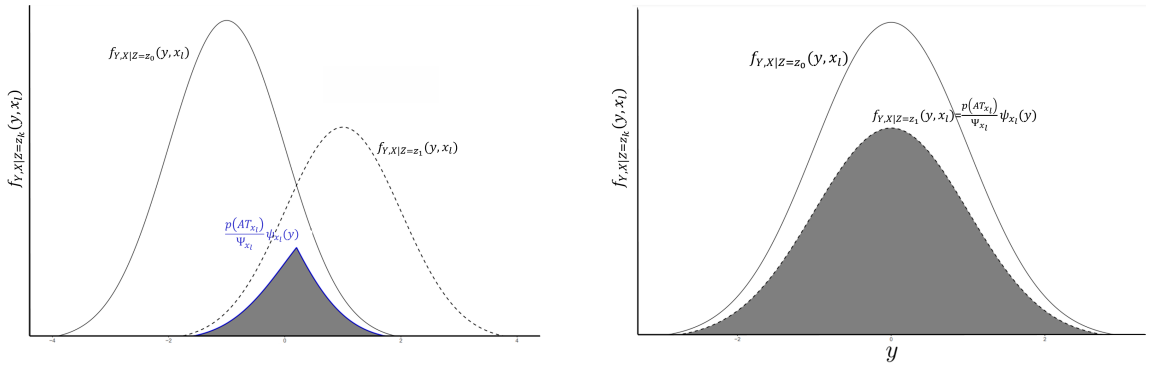
Notes: *In both figures the solid and dashed black lines respectively represent hypothetical subdensities* $f_{Y,X|Z=z_0}(y,x_l)$ *and* $f_{Y,X|Z=z_1}(y,x_l)$*. The constant* Ψ_{x_l} *is defined as the integral over* $\mathcal{Y}$ *of the pointwise minimum of these two subdensities, thus, its magnitude corresponds to the gray colored area. The two vertical red lines delimit a measurable interval* $B_Y \in \mathcal{B}_Y$*.*
The exclusion restriction implies that the distribution of outcomes for the group AT_{x_l} *of subjects that realize treatment* x_l *at both instrument values (*z_0 *and* z_1*) must be the same conditional on* $Z_i = z_0$ *or* $Z_i = z_1$*. Moreover, the always-*x_l*-takers whose outcomes belong to some measurable* $B_Y \in \mathcal{B}_Y$ *are a (potentially strict) subset of all the units that would realize treatment* x_l *and outcomes in* B_Y*. As a consequence, conditional on* $Z_i = z_k$*, the probability that the outcomes of such group belong to any measurable set* $B_Y \in \mathcal{B}_Y$ *must be lower that the probability that the event* $\{Y \in B_Y, X_i = x_l\}$ *is realized. For instance, in subfigure 1a, the fraction of units that realize treatment* x_l *and outcomes in* B_Y *conditional on* $Z = z_0$ *strictly coincides with the intersection of the gray area and the rectangle delimited by the red lines. On the other hand, conditional on* $Z = z_1$*, there is an additional group of units (whose mass corresponds to the area delimited by the red lines, the dashed black line, and the contour of the gray area) who realize outcomes in* B_Y*. This additional group cannot contain always-*x_l*-takers, otherwise, their outcomes would change without their treatment changing, which would be a violation of the exclusion restriction.*
Observe that in subfigure 1b none of the two subdensities delimits the upper contour of the gray area, and therefore, the integral of the pointwise minimum yields a tighter bound on always takers that realize outcomes in B_Y *than the minimum of the integrals of* $f_{Y,X|Z=z_0}(y,x_l)$ *and* $f_{Y,X|Z=z_1}(y,x_l)$ *over* B_Y*. This argument is formalized in appendix A.*

elements in the ordered or unordered case as in Sun (2023).

Suppose that addoption of x_l weakly increases as Z changes from z_0 to z_1. Note that in this special case, the function $f_{Y,X|Z=z_1}(y,x_l)$ weakly dominates the function $f_{Y.X|Z=z_0}(y,x_l)$, therefore, $\psi_{x_l}(y)$ is equal to $f_{Y,X|Z=z_0}(y,x_l)$. Furthermore, the fraction of always x_l takers is identified and coincides with $\mathbf{\Psi}_{x_l}$ and $\mathbb{P}[X = x_l|Z = z_0]$. This particular case is illustrated in figure 2b.

Figure 2b illustrates why in the binary monotone case my approach is equivalent to that of Kitagawa (2015). The integral of $\psi_{x_l}(y)$ over $\mathcal{Y}$ constrains the frequency of always takers, this statement is true even if monotonicity does not hold. The test of

Figure 2: Sufficiency of the necessary conditions



(a) General case (b) Monotone IV

Notes: *In both subfigures the solid and dashed black lines represent hypothetical subdensities* $f_{Y,X|Z=z_0}(y,x_l)$ *and* $f_{Y,X|Z=z_1}(y,x_l)$ *respectively. The function* $\psi_{x_l}(y)$ *is defined as the pointwise minimum* $\psi_{x_l}(y) := min\{f_{Y,X|Z=z_0}(y), f_{Y,X|Z=z_1}(y)\}$*, and the constant* Ψ_{x_l} *as its integral over* $\mathcal{Y}$*, thus, its magnitude corresponds to the area that lies strictly below both subdensities. In this figure, the gray area equals the quantity* $p(AT_{x_l})$*. The blue solid line represents the subdensity* $\frac{p(AT_{x_l})}{\Psi_{x_l}}\psi_{x_l}(y)$ *which is proportional to the distribution of potential outcomes for always-*x_l*-takers. Then, the distribution of potential outcomes for* x_l*-compliers and* x_l*-defiers can be defined to be proportional to the subdensity* $f_{Y,X|Z=z_k}(y,x_l) - \frac{p(AT_{x_l})}{\Psi_{x_l}}\psi_{x_l}(y)$ *for* $k = 0, 1$ *respectively.*
Subfigure (a) corresponds to the general case without monotonicity. It also illustrates a situation in which a valid distribution over instrument response types $p \in \Delta(\mathcal{T}_Z)$ *is such that* $p(AT_{x_l}) < \Psi_{x_l}$*. The monotone case is illustrated in subfigure (b). In this case, monotonicity enforces that* $p(AT_{x_l}) = \Psi_{x_l}$*, furthermore, the distribution of potential outcomes for always-*x_l*-takers is identified and is proportional to* $\psi_{x_l}(y)$*.*

Kitagawa (2015) takes the function $f_{Y,X|Z=z_0}(y,x_l)$ and asks if, indeed, it coincides with ψ_{x_l} almost everywhere. Instead, I first compute the integral of $\psi_{x_l}(y)$ without invoking monotonicity. Then, I note that if violations of monotonicity occur with positive probability, $\psi_{x_l}(y)$ will be strictly smaller than $f_{Y,X|Z=z_0}(y,x_l)$ in some positive measure set. Thus, the two integrals are equal if and only if random assignment, exclusion, and monotonicity are not violated.

For the general, nonmonotone case, Kwon and Roth (2024) note that the population frequency of x_l-defiers (units that realize $X_i = x_l$ only when $Z_i = z_0$) is bounded below by the supremum over measurable sets $B_Y \subset \mathcal{Y}$ of the difference $\mathbb{P}[Y_i \in B_Y, X_i = x_l | Z_i = z_0] - \mathbb{P}[Y_i \in B_Y, X_i = x_l | Z_i = z_0]$. This supremum is attained in the set $\{y \in \mathcal{Y} : f_{Y,X|Z=z_0}(y,x_l) \geq f_{Y,X|Z=z_1}(y,x_l)\}$ and its value coincides with the quantity $\mathbb{P}[X_i = x_l, Z_i = z_0] - \Psi_{x_l}$. Since at $Z = z_0$ only x_l-defiers and x_l-always takers can realize x_l, the upper bound on always takers I derive implies and is implied by the lower bound on compliers from Kwon and Roth (2024). The difference is that instead

of taking a supremum over a potentially infinite collection of measurable sets, I obtain an equivalent bound by integrating the pointwise minimum of two subdensities.

3.2 Multi-valued instruments

Now I extend the results to the multivalued instrument case. The key insights are three, first, every nonempty subset $S \subset \mathcal{Z}$ of instrument values, is associated with a unique set of instrument response types. Second, for every collection $\mathfrak{S}$ of nonempty subset $S \subset \mathcal{Z}$ of instrument values (that is, a collection $\mathfrak{S} = \{S_1, ..., S_j\} \in \mathcal{P}(\mathcal{Z})$) there is a bound on the join population prevalence of all the response types associated with at least one set in $\mathfrak{S}$. Third, these bounds can always be expressed as integrals of the pointwise minimum of a finite number of observable functions. In this section I derive these bounds, and prove that when $\mathcal{Z}$ has cardinality three they amount to the sharp testable implications of the IV model.

In this subsection I assume that the instrument has finite, discrete support $\mathcal{Z} = \{z_0, ..., z_{K-1}\}$. This setting covers multi-valued instruments as well as the case in which multiple, different, discrete valued instruments are available.

Definition 3.2 (S**-sufficient-takers**)**.**
Let $t^Z \in \mathcal{T}_Z$ be an instrument-response type and let S be a fixed subset of $\mathcal{Z}$. The instrument-response type t^Z belongs to the set of S-sufficient-x_l-takers if:

$$t^Z(z) = x_l \ \ \forall z \in S,$$

I denote this set as $\mathcal{S}_{S,x_l}$.

That is, S-sufficient-x_l-takers are the set of instrument-response types for which $z \in S$ is a sufficient condition for realizing treatment x_l. Observe that for every nonempty subset S of the support of the instrument, a nonempty set of sufficient takers exists. Note further that $\mathcal{Z}$-sufficient-x_l-takers correspond to always-x_l-takers. Finally, note that all the groups of S'-sufficient-x_l-takers associated with some $S' \supset S$ are contained in the set of S-sufficientx_l-takers.

The bound on always takers from the two instruments case admits two natural extensions. One is to use the integral of the pointwise minimum of all subdensities $\{f_{Y,X|Z=z_k} : z_k \in \mathcal{Z}\}$ as a bound on the prevalence of always-x_l takers. A second, strictly more general extension is to note that for every $S \subset \mathcal{S}$, the share of S-Sufficient-

x_l-takers is bounded above by the integral over $\mathcal{Y}$ of the function $\min_{z_k \in S}\{f_{Y,X|Z=z_k}(y)\}$. That is:

$$\mathbb{P}[X(\cdot) \in \mathcal{S}_{S,x_l}] \leq \int_{\mathcal{Y}} \min_{z_k \in S}\{f_{Y,X|Z=z_k}(y)\}dy, \;\; \forall S \subset \mathcal{Z}, \;\; \forall x_l \in \mathcal{X} \tag{6}$$

That is, the area that lies strictly below all the subdensities associated with treatment x_l and instrument values in S.

The bounds from inequality 6, to the best of my knowledge, do not currently exist in the literature. These bounds, however, ignore the fact that the exclusion restriction also imposes joint bounds on groups of sufficient-takers groups associated with the same treatment value, but different sets $S \subset \mathcal{Z}$. In fact, for every collection of sufficient-taker groups $\mathfrak{S} = \{S_1, ...S_j, ..., S_J : S_j \subset \mathcal{Z}\}$, analogous bounds exist.

The following definitions are the first step toward these more general bounds.

Definition 3.3 (**$\mathfrak{S}$-Sufficient-x_l-takers**). *Let $t^Z \in \mathcal{T}_Z$ be an instrument-response type and $\mathfrak{S}$ a collection of nonempty subsets of $\mathcal{Z}$.*

t^Z belongs to the set of $\mathfrak{S}$-sufficient-x_l-takers if t^Z belongs to the set of S-sufficient-x_l-takers for some $S \in \mathfrak{S}$.

Furthermore, I denote the set of all instrument-response types that satisfy this definition as $\mathcal{S}_{\mathfrak{S},x_l}$. Consequently, I denote the event that an instrument-response t^Z belongs to the set of $\mathfrak{S}$-sufficient-x_l-takers as $t^Z \in \mathcal{S}_{\mathfrak{S},x_l}$.

For any set $\mathfrak{S}$ of sufficient-x_l-takers, a bound can be constructed from a set of instrument values $C \subset \mathcal{Z}$ such that for every $S \in \mathfrak{S}$, at least one element of C belongs to S. I formalize this condition in the following definition.

Definition 3.4. *Let $\mathfrak{S}$ be a nonempty subset of the power set of the instrument's support $\mathcal{P}(\mathcal{Z})$. A set of instrument values $G \subset \mathcal{Z}$ covers the set $\mathfrak{S}$ if and only if:*

$$\forall S \in \mathfrak{S}, \exists z_k \in C \;\; s.t. \;\; z_k \in S$$

And:

$$\forall z_k \in C, \;\; \exists S \in \mathfrak{S} \;\; s.t. \;\; z_k \in S$$

I denote the collection of all sets that cover $\mathfrak{S}$ as $\mathfrak{C}(\mathfrak{S})$.

Every set C that covers a collection $\mathfrak{S}$ is associated with a function whose integral over $\mathcal{Y}$ bounds the prevalence of $\mathfrak{S}$-sufficient-x_l-takers. For any given C, this function takes the following form:

$$\psi_C(y, x_l) := \sum_{z_k \in C} f_{Y,X|Z=z_k}(y, x_l)$$

Theorem 3.2 (**Observable implications of multi-valued instruments**). *For every* $\mathfrak{S} \subset \mathcal{P}(\mathcal{Z})$*:*

$$\mathbb{P}[t^Z \in \mathcal{S}_{\mathfrak{S},x_l}] \leq \int_{\mathcal{Y}} \min_{C \in \mathfrak{C}(\mathfrak{S})} \{\psi_C(y, x_l)\} dy \tag{7}$$

Theorem 3.2 generalizes the bounds on always takers from the binary instrument case to multi-valued instruments. The set of inequalities in expression 7 are upper bounds on the joint prevalence of different groups of instrument-response types. That is, bounds on the share of units whose behavior is affected in specific ways by the instrument. In addition to providing a falsification procedure for IV models, these bounds can be used to partially identify the share of units that a specific set of treatments induces into taking one particular treatment. One setting in which these bounds might be of particular interest is intent to treat designs with multiple treatment types.

Regarding falsification of IV models, the bounds from theorem 3.2 are only implied by sharp characterizations of the model. That is, they are not implied by the observable implications from Sun (2023) or Frandsen et al. (2023). At the population level, they are implied by the characterization from Kaido and Ponomarev (2025). In finite samples, however, they retain features of the distribution of the outcome that discretization might loose. Hence, in practice, they can be used to complement discretization-based tests rather than as an alternative.

Some of the inequalities from theorem 3.2 are redundant. The reason is that there is a many to one correspondence between colllections of sets of instrument values and sets of instrument-response types. For example, if $S_1 = \{0, 1, 2\}$ and $S_2 = \{0, 1\}$. The bound associated with $\mathfrak{S}_1 = \{S_1\}$ implies the bounds associated with $\mathfrak{S}_{1,2}\{S_1, S_2\}$. The reason is that the bounds associated with $\mathfrak{S}_1$ come from integrating the function $\min\{f_{Y,X|Z=z_0}(y, x_l), f_{Y,X|Z=z_1}(y, x_l)\}$ while the bounds associated with $\mathfrak{S}_{1,2}$ integrate $\min\{f_{Y,X|Z=z_0}(y, x_l), f_{Y,X|Z=z_1}(y, x_l), f_{Y,X|Z=z_0}(y, x_l)+f_{Y,X|Z=z_2}(y, x_l), f_{Y,X|Z=z_1}(y, x_l)+f_{Y,X|Z=z_2}(y, x_l)\}$. But the last two terms inside the minimum operator in the second

expression are redundant. An alternative explanation is that all the response types associated with S_1 are also associated with S_2 while the converse is not true.

In general, the elements of any collection $\mathfrak{S}$ can be partially ordered using the set-inclusion ordering. The bound associated with the collection of minimal elements of $\mathfrak{S}$ with respect to this ordering implies the bound associated with $\mathfrak{S}$. This eliminates some, but not all of the redundant inequalities.

In the next subsection I focus on the three instrument values case. For this particular model, I show that the bounds in theorem 3.2 amount to the sharp testable implications of instrument validity (assumption 2.1), and any shape restriction consistent with assumption 2.2.

3.3 Sharp testable implications in the three instruments case

Suppose that $\mathcal{Z} = \{z_0, z_1, z_2\}$. The collection of nonempty subsets of $\mathcal{Z}$ (i.e. $\mathcal{P}(\mathcal{Z})$) comprises the sets $\{z_0\}$, $\{z_1\}$, $\{z_2\}$, $\{z_0, z_1\}$, $\{z_0, z_2\}$, $\{z_1, z_2\}$ and $\mathcal{Z}$ itself. For a fixed $x_l \in \mathcal{X}$, each of this sets is associated with a specific set of sufficient-takers. For example, the sets $\{z_k\}$ correspond to units that only realize $X = x_l$ when $Z = z_k$. The set $\mathcal{Z}$ corresponds to always-x_l-takers.

The sharp testable implications do not require to calculate bounds for every element of $\mathcal{P}(\mathcal{Z})$. In what follows, I show that the bounds obtained from the collection that includes all singletons $\big\{\{z_0\}\big\}$, $\big\{\{z_1\}\big\}$, $\big\{\{z_2\}\big\}$, $\big\{\{z_0, z_1\}\big\}$, $\big\{\{z_0, z_2\}\big\}$, $\big\{\{z_1, z_2\}\big\}$, and $\big\{\mathcal{Z}\big\}$, as well as all collections that contain two sets of cardinality two $\big\{\{z_0, z_1\}, \{z_0, z_2\}\big\}$, $\big\{\{z_0, z_1\}, \{z_1, z_2\}\big\}$, $\big\{\{z_0, z_2\}, \{z_1, z_2\}\big\}$ is sufficient to construct a finite set of linear inequalities that is equivalent to the sharp testable implications of the model. Each of these *sets of sets* corresponds to a different collection $\mathfrak{S}$ associated with a specific set of $\mathfrak{S}$-sufficient-x_l-takers according to definition 3.3.

From the singletons, the following bounds arise. First, trivial, single-density bounds:

$$\mathbb{P}[X(\cdot) \in \mathcal{S}_{\{z_k\},x_l}] = \int_{\mathcal{Y}} f_{Y,X|Z=z_k}(y, x_l)dy =: \Psi(C_k), \quad \forall \{z_k\} \subset \mathcal{Z}, \tag{8}$$

Where equality holds since $\mathcal{S}_{\{z_k\},x_l}$ includes all units that realize x_l when $Z = z_k$.

Second, bounds on $\{z_k, z_{k'}\}$-sufficient-x_l-takers:

$$\mathbb{P}[X(\cdot) \in \mathcal{S}_{\{z_k,z_{k'}\},x_l}] \leq \int_{\mathcal{Y}} \min_{z=z_k,z_{k'}} \{f_{Y,X|Z=z}(y,x_l)\}dy =: \Psi(C_{kk'}), \quad \forall\{z_k, z_{k'}\} \subset \mathcal{Z}. \tag{9}$$

And third, bounds on always-x_l-takers

$$\mathbb{P}[X(\cdot) \in \mathcal{S}_{\mathcal{Z},x_l}] \leq \int_{\mathcal{Y}} \min_{z=\mathcal{Z}} \{f_{Y,X|Z=z}(y,x_l)\}dy =: \Psi(C_{012}). \tag{10}$$

From the nonsingleton sets that contain two distinct pairs of instrument values each, the following bounds arise:

$$\begin{aligned}\mathbb{P}\Big[\Big(X(\cdot) \in \mathcal{S}_{\{z_k,z_{k'}\},x_l}\Big) \vee \Big(X(\cdot) \in \mathcal{S}_{\{z_k,z_{k''}\},x_l}\Big)\Big] \\ \leq \int_{\mathcal{Y}} \min\left\{f_{Y,X|Z=z_k}, f_{Y,X|Z=z_{k'}} + f_{Y,X|Z=z_{k''}}\right\} =: \Psi(C_{kk',kk''}), \\ \forall\{z_k, z_{k'}\}, \{z_k, z_{k''}\} \subset \mathcal{Z}\end{aligned} \tag{11}$$

The inequalities in 8, 9, 10 and 11 are linear inequality constrains on the vector of latent instrument-response types. Consequently, as in the binary instrument case, they admit a matrix form representation as $A_{Y,3IV}\boldsymbol{p} \leq \Psi_{3IV}$ where $A_{Y,3IV}$ is a matrix with known, deterministic entries and Ψ_{3IV} comprises the parameters on the right hand side of the inequalities.

Theorem 3.3 (**Sharp testable implications**).
Suppose that $\mathcal{Z} = \{z_0, z_1, z_2\}$. The following statements are equivalent:

1. *A distribution over response types $g \in \Delta(\mathcal{T})$ that satisfies, assumption 2.1 (instrument validity), assumption 2.2 (shape restrictions) and definition 2.3 (consistency with observed probabilities) exists.*

2. *A distribution over instrument-response types $p \in \Delta(\mathcal{T}_Z)$ exists such that p sat-*

isfies:

$$A_Z \boldsymbol{p} \leq \boldsymbol{r}_Z,$$
$$A_X \boldsymbol{p} = \boldsymbol{r}_X,$$
$$A_{Y,3IV} \boldsymbol{p} \leq \Psi_{,3IV}.$$

The proof ot the theorem proceeds in three steps. First, I show that the falsification problem does not require enumerating all response types and specifying their potential outcomes. It suffices to specify the marginals associated with every treatment $x_l \in \mathcal{X}$ and just ensure global consistency with instrument-response types. In the continuous outcomes case, this facilitates the proof. For discrete outcomes, this result yields a system linear in the cardinality of the outcome's support, as opposed to exponential if full response-type enumeration is done[5].

The second, and key step is to partition the areas below the functions $f_{Y,X|Z=z_k}(y, x_l)$ into ten different regions for every x_l. This partition addresses a problem: units with instrument-response types that realize treatment x_l at two distinct instrument values z_k and $z_{k'}$ simultaneously contribute to the areas below the function $f_{Y,X|Z=z_k}$ and $f_{Y,X|Z=z_{k'}}$. This contribution, however, has to simultaneously be consistent with the distributions of potential outcomes associated with all other response types that realize x_l when Z takes either value z_k or value $z_{k'}$.

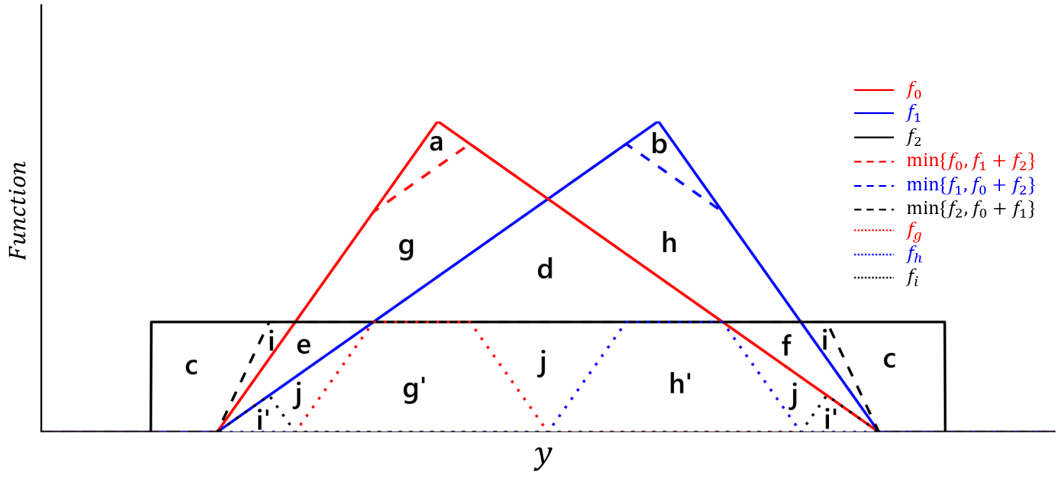
Figure 3 depicts the partitions of the areas below the subdensities $f_{Y,XZ|z_k}$. Not all areas can be constructed from all instrument types. Por example, area a can only be constructed with units that realize x_l if and only if $Z = z_0$. The most important insight is that areas g, h and i exist that must be filled twice. To illustrate why, let I_{01,x_l}, I_{02,x_l} and I_{12,x_l} denote instrument-response types that realize x_l if and only if $z \in \{0, 1\}$, $z \in \{0, 2\}$ and $z \in \{1, 2\}$ respectively. Suppose that all of these types have mass equal to $A_{g,x_l} > 0$ and that their potential outcomes associated with $X = x_l$ follow a distribution proportional to the function f_{g,x_l} depicted in the same figure. Then, at instrument values $Z = z_1$ or $Z = z_2$ these types contribute a total mass of $A_{g,x_l} > 0$, but, when $Z = z_0$, they contribute twice as much, that is $2A_{g,x_l} > 0$.

The partition I propose is constructed in a way that facilitates assigning mass to the instrument-response types I_{01,x_l}, I_{02,x_l} and I_{12,x_l} in a way that joint consistency

[5]This result might be used to reduce the computational cost of discretization-based methods for IV models and thus, might be of independent interest.

with all subdensities is guaranteed to hold. In particular, it allows some areas to be filled once or twice depending on the instrument value. The bounds from 3.3 ensure that such an assignment is feasible.

Figure 3



Notes: *The solid lines in this figure depict the subdensities* $f_{Y,X|Z=z_k}(y, x_l)$ *for* $k = 0, 1, 2$*. Each of the letteres* $a, b, c, d, e, f, g, h, i$ *and* j *label a region that contributes to the partitions of the area below the functions* $f_{Y,X|Z=z_k}(y, x_l)$*. These partitions allow to construct a distribution of potential outcomes that is consistent with the observed joint distribution of* Y, X *and* Z*, the instrument validity assumptions, and any additional shape restriction. The label* f_k *denotes the subdensity* $f_{Y,X|Z=z_k}(y, x_l)$*. The dashed lines represent functions that are obtained from transforming the primitive functions* f_k*. The long dashes represent functions that take the form* $\min\{f_k, \sum_{k' \neq k} f_{k'}\}$ *. The functions* f_g*,* f_h *and* f_i *are represented by the lines with short dashes and are defined as* $f_g = \min\{f_0, f_1 + f_2\} - \min\{f_0, f_1, f_2\}$*,* $f_h = \min\{f_1, f_0 + f_2\} - \min\{f_0, f_1, f_2\}$ *and* $f_i = \min\{f_2, f_0 + f_1\} - \min\{f_0, f_1, f_2\}$*.*
To construct a valid distribution of potential outcomes, I assign sufficient-taker types to each of the areas depicted. Then, I assign outcomes to such groups that are proportional to the heights of the area to which they where assigned.
All areas are delimited by one or more of the depicted functions. The areas g *and* g' *have exact same area and height at any value of* y*. Equivalently, the areas* h *and* h'*; and the areas* i *and* i' *have the same area and height. This is what allows assigning the same area twice, when necessary.*

4 Generalized no-defiers restrictions

Now I specialize theorem 3.1 to binary instruments combined with restrictions that rule out specific treatment switches from the population. This class of restrictions includes ordered and unordered monotonicity (Imbens and Angrist (1994), Angrist and Imbens (1995), Heckman and Pinto (2018)), and extensions or relaxations of these assumptions.

I refer to restrictions in this class as *generalized no-defiers restrictions*. The specialized characterization is a set of sharp inequalities with a clear microeconomic interpretation.

The essence of generalized no-defiers restrictions is that as the instrument changes, some treatments are promoted relative to others. I formalize this notion as follows: If alternative x_1 is preferred over x_0 absent any exogenous incentive, then, it must also be preferred over x_0 when the instrument promotes x_1. Hence, no units that switch from x_1 to x_0 exist in the population.

Furthermore, if generalized no-defiers restriction assume that some treatments are promoted relative to others by the instrument, then, the promoted treatments should be more prevalent at $Z = z_1$ than at $Z = z_0$.

In the ordered monotone case, treatment values are ordered (e.g. by intensity), and $Z = z_1$ is assumed to promote adoption of higher ranked treatments. The sharp observable implications of this restriction are known to be first order stochastic dominance. That is, for any given value of the treatment x_l, treatments ranked at least as high as x_l should be more likely when $Z = z_1$.

Generalized no-defiers restrictions encompass a large class of assumptions on how Z affects realizations of X other than ordered monotonicity. Examples are, the effect on the extensive margin assumption from Rose and Shem-Tov (2024), versions of the unordered monotonicity assumption by Heckman and Pinto (2018), or other restrictions that allow for richer heterogeneity such as the examples I provide in the online appendix to this paper. For these restrictions it is not obvious neither which treatments are being promoted, nor exactly how the promoted treatments should be more prevalent when $Z = z_1$.

I answer these two questions by showing that any generalized no-defiers restriction is uniquely associated with a nonparametric assumption on how the instrument affects the potential preferences over treatments of units in the population, and that this assumption is characterized by a unique binary relation $\geq^R$ over $\mathcal{X}$. Then, I prove that the sharp testable implications of the restriction are equivalent to *generalized first order sotchastic dominance (GFOSD)*, an extension of first order stochastic dominance to arbitrary binary relations, with respect to $\geq^R$.

Furthermore, I show that the sharp testable implications of any *generalized no-defiers restriction* combined with the exclusion restriction from assumption 2.1 is also a set of inequalities with a GFOSD interpretation. The reason is that the exclusion restriction on its own does not rule-out any instrument-response type. But, combined

with the observed distribution of the outcome, it imposes upper bounds on the prevalence of always-takers, which in turn imposes upper bounds on the total share of units that can realize certain treatments.

The GFOSD characterization is a set of inequalities rather than a system of linear equations. Having a characterization in terms of inequalities has two advantages, first, it eliminates nuisance parameters which allows for simpler inference, second, every inequality is linked to a set of behaviors that the imposed shape restriction (on its own or combined with the exclusion restriction) rules out. Hence, if an inequality is violated , it reveals a set of behaviors that exist in the population but are inconsistent with the model's assumptions.

4.1 Sharp testable implications

First, I formally define *generalized no-defiers restrictions* and show that any such restriction can be represented as a binary relation.

Assumption 4.1 (Generalized no-defiers)**.**
Let $R \subset \mathcal{T}_Z = \mathcal{X} \times \mathcal{X}$ be a fixed and known collection of instrument-response types. The underlying distribution over response types $g_0 \in \Delta(\mathcal{T})$ is such that $p_{g_0}(t^Z) = 0$ for all $t^Z \in R$. That is, instrument-response types in R are ruled-out.

Since R is a subset of $\mathcal{X} \times \mathcal{X}$, by definition, it is a binary relation.

Definition 4.1 (Binary relation associated to $R \subset \mathcal{T}_Z$)**.**
Let $R \subset \mathcal{T}_Z$ be a collection of response types. Define the binary relation $\geq^R \subset \mathcal{X} \times \mathcal{X}$ as follows:

$$x_l \geq^R x_{l'} \iff (x_l, x_{l'}) \in R$$

Put in words, a *generalized no-defiers restrictions* simply imposes that instrument-response types contained in the set R occur with zero probability. Moreover, x_l is related to $x_{l'}$ according to $\geq^R$ if switches from x_l to $x_{l'}$ when the instrument goes from z_0 to z_1 are ruled out by the *generalized no-defiers restrictions* associated with R.

Suppose that a natural ordering over treatments exist (e.g. the number of months an individual had medical insurance during a given period of time), and suppose that an exogenous intervention promotes the adoption of weakly higher treatments (e.g. having insurance for longer periods). Then, no unit should realize a strictly lower

treatment as a consequence of the intervention, or, put formally, no unit should realize a a treatment in the lower contour of its initial one.

The generalization of the previous idea is that, for any given set of treatments S, there is a unique second set of treatments $L_{\geq^R}(S)$ such that take-up of treatments in S is promoted relative to take-up of treatments in $L_{\geq^R}(S)$. Consequently, any unit that realizes a treatment in S when $Z = z_0$, cannot realize a treatment in $L_{\geq^R}(S)$ when $Z = z_1$. For a *generalized no-defiers restrictions* associated with a binary relation R, the set $L_{\geq^R}(S)$ is the *common lower contour* of the set S according to $\geq^R$.

Definition 4.2 (Common lower contour of $\geq^R$)**.**
Let $\geq^R$ be a binary relation over $\mathcal{X}$ associated with the generalized no-defiers restriction R. For any $S \subset \mathcal{X}$, the common lower contour of S according to $\geq^R$ is:

$$L_{\geq^R}(S) = \{x_k \in \mathcal{X} : s \geq^R x_k, \forall s \in S\}$$

Having defined common lower contours, I state the sharp observable implications of *generalized no-defiers restrictyions* and binary instruments. The sharp testable implications can be summarized as follows: For every $S \subset \mathcal{X}$, no unit can have potential treatments $X(z_0)$ and $X(z_1)$ such that $X(z_0) \in S$ and $X(z_1) \in L_{\geq^R}(S)$. Consequently, the probability of observing treatments in the complement of $L_{\geq^R}(S)$ conditional on $Z = z_1$ should be at least as large as the probability of observing treatments in S conditional on $Z = z_0$

Theorem 4.1.
A distribution over response types $p \in \Delta(\mathcal{T}_Z)$ that satisfies assumption 2.1 (instrument validity), assumption 4.1 (instrument-response type restriction), and definition 2.3 (consistency) exists, if and only if:

1. ***Shape restriction+Independence:*** *The following inequality holds for every $S \subset \mathcal{X}$.*

$$\mathbb{P}[X_i \in S | Z_i = z_0] \leq \mathbb{P}[X_i \in L_{\geq^R}(S)^C | Z_i = z_1] \tag{12}$$

2. ***Shape restriction+Independence+Exclusion restriction:*** *For any $S \subset \mathcal{X}$ and any partition (Λ_S, Λ'_S) of $L_{\geq^R}(S)^C$ satisfying (i) $\emptyset \neq \Lambda'_S \subset S$, (ii)*

$L_{\geq^R}(x_{l'})^C \setminus \{x_{l'}\} \subset \Lambda_S$ *for every* $x_{l'} \in \Lambda'_S$*, and (iii)* $L_{\geq^R}(S \setminus \Lambda'_S)^C \subset \Lambda_S$*:*

$$\mathbb{P}[X_i \in S | Z = z_0] \leq \mathbb{P}[X_i \in \Lambda_S | Z = z_1] + \sum_{x_{l'} \in \Lambda'_S} \Psi_{x_{l'}} \tag{13}$$

Where Ψ_{x_l} *is defined in equation 5.*

The formal proof of theorem 4.1 is in the online appendix[6]. The key idea is to represent the instrument validity conditions as a flow problem: If the observed conditional treatment probability $\mathbb{P}[X_i = x_l | Z_i = z_0]$ represents the share of the population that realizes treatment x_l at instrument value z_0, then, finding a valid distribution over response types is equivalent to finding a valid set of flows that, for every $x_l \in \mathcal{X}$, takes the share $\mathbb{P}[X = x_l | Z = z_0]$ of the population that starts with treatment x_l and distributes it across potential counterfactual treatments in a way that matches the observed treatment probabilities conditional on $Z = z_1$ without allocating mass to the flows, or treatment switches that the *generalized no-defiers restriction* rules-out.

This flow representation of the falsification problem is equivalent to a pair of well known combinatorial problems: the maximum flow and minimum cut problems.[7] The minimum cut formulation is particularly useful to derive the sharp observable implications of incomplete models in terms of inequalities that only depend on observable quantities as first noted by Ekeland et al. (2010) and Galichon and Henry (2011).

The first part of theorem 4.1 concerns falsification of instrument independence and *generalized no-defiers restrictions* alone (i.e. ignoring the outcome and the exclusion restriction). An interpretation for this part of the theorem is that mutually exclusive events cannot occur in the population with probability greater than one (the events $\{X_i(z_0) \in S\}$ and $X_i(z_1) \in L_{\geq^R(S)}$ are mutually exclusive). Kaido and Ponomarev (2025) recently generalized a logically equivalent idea beyond the binary instrument case.

The second part of theorem 4.1 incorporates the exclusion restriction which, as I show in theorem 3.1, binds the prevalence of always-x_l takers. In some cases, these bounds on always takers *tighten* the GFOSD inequalities by further limiting the max-

[6]The online appendix is at https://ricardomirandam.github.io/files/OnlineAppendix.pdf

[7]A proof of this equivalence is in the online appendix A.2. The argument also applies to more general restrictions such as those considered in assumption 2.2. The only difference is that more general restrictions cannot necessarily be stated as capacity constraints on the edges of the graph. This would complicate using the equivalence with minimum-cut problem to obtain observable inequalities that do not depend on unobservables.

imum of units that could potentially realize treatments in certain upper contours.

To incorporate an outcome along with the exclusion restriction, Kaido and Ponomarev (2025) expand the graph representation of the model to the space of treatment-outcome pairs (i.e. $(x, y) \in \mathcal{X} \times \mathcal{Y}$). For discrete outcomes, this expanded space substantially increases the total number of inequalities to verify, and for continuous outcomes results in an infinite number of inequalities making statistical inference impractical. In contrast, under my approach, the number of inequalities remains the same when an outcome and an exclusion restriction are added.

The fact that the exclusion restriction *tightens* the GFOSD inequalities by replacing some conditional treatment probabilities $\mathbb{P}[X_i = x_{l'}|Z = z_1]$ with the weakly smaller quantities $\Psi_{x_{l'}}$ allows to distinguish violations of the exclusion restriction from violations of instrument-response restrictions. In particular, one of the following three cases must occur:

1. All inequalities are satisfied. Hence, neither the exclusion restriction nor the response-type restriction is falsified.

2. Only inequalities from part 2 are violated. Hence, the response type restriction is not falsified on its own, but only when combined with the exclusion restriction.

3. At least one inequality from part 1 is violated. Hence, the response type restriction is falsified

Additionally, every inequality is linked to exactly one set S. Thus, any violation of an inequalities associated with S points-out to a specific subset of treatment values at which the imposed assumptions are inconsistent with the observed joint distribution of Y, X and Z.

To illustrate its usefulness, I use theorem 4.1 to characterize the binary monotone instrument IV model from Angrist and Imbens (1995).

Corollary 4.1 (**Sharp observable implications of binary monotone instruments**). *Suppose that $\mathcal{X} = \{x_0, x_1, ..., x_{L-1}\}$, with $x_l < x_{l'}$ for every $l < l'$. A distribution over response types exists that satisfies assumptions 2.1 and:*

$$\mathbb{P}[X(z_0) \geq X(z_1)] = 0$$

If and only if:

$$\mathbb{P}[X \geq x_l | Z = z_0] \leq \Psi_{x_l} + \mathbb{P}[X > x_{l+1} | Z = z_1] \qquad \forall x_l \in \mathcal{X}$$

In the ordered monotone case, the set $L_{\geq}(x_l)^C$ corresponds to the standard upper contour of x_l (i.e. $\{x_l, x_{l+1}, ..., x_L\}$), thus, the sharp inequalities reduce to FOSD but *tightened* by the exclusion restriction in the sense that the probability $\mathbb{P}[X = x_l | Z = z_1]$ associated to the minimal element of every $L_{\geq}(x_l)^C$ (i.e. x_l) is replaced by the weakly smaller quantity Ψ_{x_l}.

4.2 A submonotone interpretation

I conclude this section by formally deriving *generalized no-defiers restrictions* from nonparametric restrictions on how the instrument affects the preferences or utility of units in the population. In particular, I assume that selection into treatment follows a utility maximization model, and, rather than restricting potential treatment choices, I restrict potential preferences over the set of treatments $\mathcal{X}$.

Assumption 4.2 ($\geq^*$-Submonotonicity)**.** *Let $\geq^*$ be an irreflexive*[8] *binary relation over $\mathcal{X}$. Suppose that:*

1. *Units are endowed with strict preference profiles over $\mathcal{X}$ that depend on Z_i denoted $\succ_i^{(z_k)}$ for $k = 0, 1$ where the function $\succ_i^{(z_k)}$ maps $\mathcal{Z}$ to strict preference profiles over $\mathcal{X}$. In other words, units have potential preferences.*

2. *Units are rational in the sense that $X_i(z_k) = x_l$ implies $x_l \succ_i^{(z)} x_{l'}$ almost surely for all $l, l' \in \{0, 1, ..., L-1\}$ such that $l \neq l'$.*

3. *There are no preference reversals against $\geq^*$ when the instrument changes from z_0 to z_1. That is, if $x_l \geq^* x_{l'}$ and $x_l \succ_i^{(z_0)} x_{l'}$, then, $x_l \succ_i^{(z_1)} x_{l'}$ almost surely.*

The first two points of assumption 4.2 simply posit that response types arise from latent utility maximization. The third point captures the the notion that if $x_l \geq^* x_{l'}$, then alternative x_l is weakly incentivized or promoted by the instrument. Note that this promotion of x_l is not necesarilly relative to all other treatments, but relative to $x_{l'}$.

[8]The extension to reflexive binary relations is conceptually simple. The reasoning remains the same but requires special treatment of response type restrictions that rule out "always takers" of one particular treatment. In the interest of brevity I address this case in appendix A.2.2.

I refer to assumption 4.2 submonotonicity because it encodes a notion weaker than monotonicity. It rules out two-way flows between certain pairs of treatments, without necessarily involving all alternatives or treatments. In this sense, any submonotonicity assumption can always be extended to a full monotonicity condition with respect to a total order (or the linear extension of a total order) by incorporating additional restrictions. For instance, both ordered and unordered monotonicity correspond to special cases of submonotonicity where the binary relation is transitive. The notion of preventing *two-way flows* from Navjeevan and Pinto (2022) corresponds to assuming that the binary relation in question is asymmetric.

Theorem 4.2. *Let $\geq^*$ be an irreflexive binary relation over $\mathcal{X}$ and suppose that $R \subset \mathcal{T}$ is such that $\geq^R = \geq^*$. Then:*

1. *If $\geq^*$-Submonotonicity holds, then, $p_{g_0}(t^Z) = 0$ for every $t^Z \in R$.*
2. *R is the smallest cardinality subset of $\mathcal{T}_Z$ such that if $p_{g_0}(t) = 0$ for all $t \in R$, then, $\geq^*$-Submonotonicity cannot be falsified.*

The assumption that $\geq^*$ is irreflexive is not necessary, but it substantially simplifies exposition. The proof of theorem 4.2 is in the online appendix where the irreflexive case is also covered.

Theorem 4.2 shows that $\geq^R$-submonotonicity implies that instrument-response types in of R occur with zero probability. Moreover, R is the minimal (in the sense of set inclusion) restriction over instrument-response types, such that $\geq^R$-submonotonicity is not falsified. Thus, the set of inequalities in 12 comprises the sharp observable implications of $\geq^R$-submonotonicity.

In the online appendix I discuss in depth the interpretation of different assumptions within the submonotonicity class and provide multiple motivating examples. These examples include extensions of unordered monotonicity (Heckman and Pinto (2018)); assumptions that draw from both ordered and unordered monotonicity; specific cases of targeting and encouragement designs along with relaxations and generalizations of these ideas (Lee and Salanié (2020), Bai et al. (2022)); and designs in which the instrument encourages variety and experimentation, or its effect depends on *status-quo* ($Z = z_0$) choices.

In section 7 I use this framework to assess the *compliers in the extensive margin* assumption of Rose and Shem-Tov (2024) in the Oregon Health Experiment, and for-

mulate alternative restrictions that are not rejected by the data but still allow to point identify the local average treatment affect among units affected by the intervention.

5 Inference

In this section I present the inference framework that yields a sharp test of IV validity. In subsection 5.1 I discuss why robustness to the contact set is desirable, and the issues that arise from it. In subsection 5.2 I introduce a classification-based estimator and explain why the classification approach yields an estimator with a tractable characterization. In subsection 5.3 I use this characterization to analyze the asymptotic properties of the estimator, and state formal asymptotic inference results for subdensity overlaps. Finally, in subsection 5.4 I combine the proposed overlap estimators with the inference method for linear systems of Cox and Shi (2023) to obtain a valid, sharp test of IV validity.

Throughout this entire section, I asume that the data is an *i.i.d.* random sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ drawn from the distribution $F_{Y,X,Z}$. In what follows, I assume that the outcome is a real-valued vector of finite dimension d so that $\mathcal{Y} = \mathbb{R}^d$.

In this section I consider the two instruments case, but the results of this section extend naturally to estimation of the overlap of any finite number of densities or subdensities.

5.1 Contact set, nondifferentiability, and plug-in estimators

The main challenge for inference on the vector of subdensity overlaps arises from the region of the support of the outcome in which two or more subdensities are equal. Let $\mathcal{C}_{x_l} = \{y \in \mathbb{R}^d : f_{Y,X|Z=z_0}(y, x_l) = f_{Y,X|Z=z_1}(y, x_l)\})$ denote this region to which I refer as the contact set. When the contact set has positive measure, nondifferentiability of the minimum implies that plug-in estimators are biased and might not have a limiting distribution. These problems are exacerbated if the goal is to have a fully nonparametric method.

Before elaborating on the technical challenges, I argue that, in the context of testing IV validity, robustness to the possibility that the contact set has positive measure is desirable. The first and simplest reason is that an IV validity test should be robust to the possibility that the instrument has no effect on realized treatments, in other words, it should allow for instrument irrelevance. If the instrument has no effect on realized

treatments and the exclusion restriction holds, subdensities will not change with the instrument and therefore the contact set is equal to the entire support of the outcome.

A second, related but subtler reason comes from the quasi-experimental nature of IV models. Consider a randomized intervention designed to encourage individuals to acquire health insurance, with the goal of estimating its effect on health outcomes. If the intervention had a 100% compliance rate, the setup would correspond to a randomized experiment. It is precisely because of noncompliance that an IV approach is needed.

But noncompliance is likely to be correlated with the outcome. The healthiest subjects in the population might not get insurance even if it is promoted by the instrument. If healthier subjects do not respond to the instrument, their treatment and outcomes do not change. Thus, the joint distribution of the outcome and treatment is the same at different instrument values which results in a non-negligible contact set.

Finally, in finite samples, the possibility that the contact set has positive measure cannot be ruled out unless the relevant subdensities do not cross. The reason is that, for any given sample size, around the true crossing points there is a non-negligible set in which the difference between densities is small and statistically indistinguishable from zero. Hence, equality on a non-negligible set cannot be rejected. If the existence of a non-negligible contact set is fundamentally untestable, robustness to this possibility is a desirable property for both overlap estimators and instrument validity tests.

To illustrate the challenges that arise from nondifferentiability, recall that the target parameter is a vector whose $l+1$-th entry, for $l \in \{0, ..., L-1\}$ is:

$$\Psi_{x_l} = \int_{\mathbb{R}^d} \min\{f_{Y,X|Z=z_0}(y, x_l), f_{Y,X|Z=z_1}(y, x_l)\} dy \tag{14}$$

A plug-in estimator for Ψ_{x_l} replaces the conditional densities $f_{Y,X|Z=z_k}(y, x_l)$ for $k = 0, 1$ in equation 14 with nonparametric density estimators $\hat{f}_{Y,X|Z=z_k}(y, x_l)$ and integrates the function $\hat{\psi}_{x_l}(y) := \min\{\hat{f}_{Y,X|Z=z_0}(\tilde{y}, x_l), \hat{f}_{Y,X|Z=z_1}(\tilde{y}, x_l)\}$.

The density estimators admit a decomposition into a signal component $f_{Y,X|Z=z_k}(y, x_l)$, and a noise component $\hat{f}_{Y,X|Z=z_k}(y, x_l) - f_{Y,X|Z=z_k}(y, x_l)$. For any $y \in \mathcal{C}_{x_l}$, the signal components are the same for $Z = z_0$ and $Z = z_1$. Thus, for any given sample size, the estimated minimizer at y is entirely determined by statistical noise.

If the minimum is determined by statistical noise regardless of the sample size, then, for any fixed $y \in \mathcal{C}_{x_l}$, the minimizer can switch back and forth from z_0 to z_1 as

the sample size increases. Consequently, the identity of the estimated minimizer at y need not converge in probability to any fixed limit.

The asymptotic distribution of a plug-in estimator depends on the random process $\{\hat{\psi}_{x_l}(y) : y \in \mathbb{R}^d\}$ which in turn depends on the estimated minimizers. The randomness in the minimizer's identity makes characterizing the variance-covariance of this process and understanding how it affects plug-in estimators technically challenging.

Moreover, aplying the mimimum operator in the contact set creates a bias problem. Instead of averaging out statistical noise, the minimum consistently picks negative noise which results in a downward biased estimator.

Finally observe that nonparametric density estimators, and more specifically kernel density estimators, do not satisfy the tightness assumption required for the *m out of n* boostrap method proposed by Fang and Santos (2019). Hence, even if the minimum is a directionally differentiable function, there are no guarantees that this bootstrap approach is valid.

To circumvent these technical challenges, I recast each subdensity overlap as the loss of an asymmetric classifier. This alternative representation of the target parameters has the advantage of eliminating most of the randomness that arises from the contact set.

5.2 A classification-based estimator

Given the functions $f_{Y,X|Z=z_0}(y, x_l)$ and $f_{Y,X|Z=z_1}(y, x_l)$, calculating Ψ_{x_l} amounts to identifying the regions in which each of these two functions is less or equal than the other, and then computing their integrals over such sets. Thus, the core problem is to identify the points in $\mathbb{R}^d$ that belong to the set $\{y \in \mathbb{R}^n : f_{Y,X|Z=z_1}(y, x_l) \geq f_{Y,X|Z=z_0}(y, x_l)\}$. Identifying this set is a classification problem.

Consider the random vector (Y, X, Z). The classification problem consists of predicting Z given Y and X. A classifier is a function $\eta : \mathcal{Y} \times \mathcal{X} \longrightarrow \{0, 1\}$ that maps pairs (y, x) to an instrument label $k = 0, 1$. A classification mistake occurs on the event $\{\eta(Y, X) \neq k, Z = z_k\}$ for $k = 0, 1$.

Given a classifier η, the probability that it makes a classification error is given by:

$$\mathbb{E}[\mathbf{1}_{\{\eta(Y,X)\neq k, Z=z_k\}}] = \mathbb{P}[\eta(Y, X) \neq k, Z = z_k],$$

This probability is a measure of the quality of the classifier η.

For overlap estimation, the functions to which the minimum are applied are the joint distributions of Y and X conditional on different values of Z. If one value of Z is more prevalent than other in the population, a classifier might attain a low loss by always returning the most likely value of Z. This can be avoided by penalizing classification mistakes differently when $Z = z_0$ than when $Z = z_1$. One simple alternative is to consider the sum of conditional missclassification probabilities $\mathbb{P}[\eta(Y,X) \neq k | Z = z_0] + \mathbb{P}[\eta(Y,X) \neq k | Z = z_1]$. As the next theorem shows, for the oracle classifier η^* that knows the exact values of the functions $f_{Y,X|Z=z_0}(y, x_l)$ and $f_{Y,X|Z=z_1}(y, x_l)$, this quantity coincides with Ψ_{x_l}.

Proposition 5.1. *Define the classifier* $\eta^*(y, x_l) := \mathbf{1}\{f_{Y,X|Z=z_1}(y, x_l) \geq f_{Y,X|Z=z_0}(y, x_l)\}$ *and the loss function* $L^*(\eta^*, x_l)$ *as:*

$$L^*(\eta^*, x_l) := \sum_{k=0,1} \mathbb{P}\big[\eta^*(Y,X) \neq k, Z = z_k, X = x_l \big| Z = z_k\big]. \tag{15}$$

Then:

$$\Psi_{x_l} = L^*(\eta^*, x_l).$$

Furthermore, for any classifier $\eta'(y, x_l)$ *such that* $\eta'(y, x_l) = \eta^*(y, x_l)$ *for every* $y \in \mathbb{R}^n \setminus \mathcal{C}_{k_l}$*:*

$$L^*(\eta^*, x_l) = L^*(\eta', x_l).$$

Proposition 5.1 shows that the overlap parameter is equivalent to the loss of the classifier $L^*(\eta^*, x_l)$. Furthermore, it shows that any classifier that only differs from $L^*(\eta^*, x_l)$ in the contact set attains the same loss as $L^*(\eta^*, x_l)$, resulting in an alternative valid expression for Ψ_{x_l}. Thus, at the population level, the classification approach is invariant to the behavior of classifiers in the contact set. This invariance property makes analyzing the asymptotic behavior of a classification-based estimator tractable.

I propose the following estimation procedure: First, split the sample in two subsamples S_1 and S_2 of size n_1 and n_2 respectively. In the first split, estimate the classifier η^*. Then, evaluate the empirical loss of the estimated classifier in the second split. The resulting empirical loss is an estimator for Ψ_{x_l}.

To formalize the estimator, let $K(\cdot)$ be a kernel function, and h_n a bandwidth

sequence with $K_n(y - Y_i) := K\big(\frac{y-Y_i}{h_n}\big)$. The subdensity estimators are:

$$f_{Y,X|Z=z_k,n}(y, x_l) := P_{k,n_1}^{-1}\Big[\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1}\mathbf{1}\{X_i = x_l, Z_i = z_k\}K_n\big(y - Y_i\big)\Big],$$

where the sum is over all units in the first split, $n_{k,1} := \sum_{i=1}^{n_1}\mathbf{1}\{Z_i = z_k\}$ and $P_{k,n_1} := n_{k,1}/n_1$.

Then, define $\Delta_{n_1}(y, x_l) := f_{Y,X|Z=z_1,n}(y, x_l) - f_{Y,X|Z=z_0,n}(y, x_l)$ and $\eta^*_{n_1}(y, x_l) := \mathbf{1}\{\Delta_{n_1}(y, x_l) \geq 0\}$. The empirical loss of the estimated classifier $\eta^*_{n_1}(y, x_l)$ evaluated in the second split is the empirical analogue of the last expression in (15):

$$L_{n_2}(\eta^*_{n_1}, x_l) := \frac{1}{n_2}\sum_{i=1}^{n_2}\Big(\frac{\mathbf{1}\{\eta^*_{n_1}(Y_i, X_i) = 1, Z_i = z_0, X_i = x_l\}}{P_{0,n_2}} + \frac{\mathbf{1}\{\eta^*_{n_1}(Y_i, X_i) = 0, Z_i = z_1, X_i = x_l\}}{P_{1,n_2}}\Big),$$

where the sum is over all observations that belong to the second split and P_{k,n_2} denotes the sample probability that Z is equal to z_k in the second split. The estimator is $L_{n_2}(\eta^*_{n_1}, x_l)$.

By construction, the estimated classifier $\eta^*_{n_1}$ yields a higher theoretical loss than the oracle classifier η^*. Thus, the estimator is upward biased. In the next subsection I provide a set conditions under which this bias is asymptotically negligible.[9]

Sample splitting plays a crucial role in deriving the asymptotic properties of the estimator. First, it eliminates the finite sample, downward bias that arises from the contact set. The reason is that, for any value of y inside the contact set, the expected empirical loss in the second sample, is invariant to classification decisions made in the first sample. Furthermore, sample splitting implies that the difference between the theoretical loss of the oracle classifier η^* and its estimator $\eta^*_{n_1}$ is uncorrelated with the difference between the empirical and theoretical loos of $\eta^*_{n_1}$. This last insight simplifies the analysis of the asymptotic variance of the estimator.

Two natural questions are whether a full sample approach would be valid and whether K-fold cross-fitting could recover full-sample asymptotics. I investigate both alternatives in a set of Monte Carlo simulations presented in the online appendix to

[9] In simulations I find that this bias is negligible even in small and medium samples.

this paper[10]. I find that both alternatives perform well as long as the contact set is negligible. However, when the contact set is non-negligible, the full-sample estimator is biased, and both methods yield confidence intervals with coverage rates below the nominal level. In the simulations, realized coverage rates range from 23% and 85%, substantially below the 95% target.

The IV validity conditions involve the vector of overlaps $\mathbf{\Psi}$ along with conditional treatment probabilities $\mathbb{P}[X = x_l|Z = z_k]$. Thus, in the evaluation split I also compute estimates of conditional treatment probabilities:

$$P_{x_l|z_k,n_2} = \frac{\sum_{i=1}^{n_2} \mathbf{1}\{X_i = x_l, Z_i = z_k\}}{\sum_{i=1}^{n_2} \mathbf{1}\{Z_i = z_k\}}.$$

Conditional treatment probabilities are linearly deficient, to avoid degenerate asymptotic distributions, I restrict the target parameter to all but one observed treatment values. Let $\mathcal{X}_k = \{l : \mathbb{P}[X = x_l|Z = z_k] > 0\}$ denote the set of treatment indices that occur with strictly positive probability conditional on $Z = z_k$[11]. Define $\boldsymbol{P}_{\boldsymbol{\mathcal{X}_k,n}}$ as the vector that stacks all the estimators $P_{x_0|z_k,n_2}$ for every $l \in \mathcal{X}_k$ except the one associated with the highest treatment index in $\mathcal{X}_k$. Finally, let $\boldsymbol{P}_{\boldsymbol{\mathcal{X}_k}}$ be the vector that stacks the associated target conditional treatment probabilities.[12]

5.3 Asymptotic properties

To analyze the asymptotic properties of the estimator, I decompose its scaled difference with respect to the target parameter into two terms. One that corresponds to the theoretical difference between the loss of the oracle and the estimated classifiers, and another that corresponds to the difference between the empirical and theoretical loss

[10]The online appendix is at https://ricardomirandam.github.io/files/OnlineAppendix.pdf

[11]In some empirical applications, this set might be known. In some others this set might not be obvious ex-ante. However, with a sufficiently large sample, the identity of all the treatment values that occur with positive probability can be identified with high probability.

[12]I restrict the conditional probability estimators to treatment values in $\mathcal{X}_k$ to ensure that the limit joint distribution of the estimators is nondegenerate.

of the estimated classifier.

$$\sqrt{n_2}\Big[L_{n_2}(\eta^*_{n_1}, x_l) - \Psi_{x_l}\Big] = \sqrt{n_2}\Big[L_{n_2}(\eta^*_{n_1}, x_l) - L(\eta^*_{n_1}, x_l)\Big] + \sqrt{n_2}\Big[L(\eta^*_{n_1}, x_l) - L(\eta^*, x_l)\Big] \tag{16}$$

To obtain valid inference, I note that the asymptotic behavior of the first term depends on neither the distribution of the classification decision at particular values of $y \in \mathbb{R}^d$, nor the correlation between classification decisions at different values of $y \in \mathbb{R}^d$. It only depends on aggregate missclassification rates, which substantially simplifies analyzing and computing its variance. Combining this insight with a set of sufficient conditions that make the second term $o_p(1)$ I prove that the estimator is asymptotically normal and propose a variance estimator.

To make the second term in equation 16 $o_p(1)$ I first use the fact that the population level loss is invariant to the behavior of $\eta^*_{n_1}$ in the contact set (see proposition 5.1).

It remains to analyze the loss of $\eta^*_{n_1}$ out of the contact set. The key insight is that, under mild regularity conditions satisfied by most standard kernel functions, the rate of uniform convergence for KDEs is $r_n = \sqrt{\frac{log(h_n^{-1})}{nh_n^d}}$ (Giné and Guillou (2002)). Therefore, in the region of $\mathbb{R}^d$ in which the absolute difference $|\Delta_{x_l}(y)| := |f_{Y,X|Z=z_1}(y, x_l) - f_{Y,X|Z=z_0}(y, x_l)|$ is greater than r_n, classification mistakes occur with very low probability. Consequently, while the additional loss in which $\eta^*_{n_1}$ might incur relative to η^* in this region can be large, the expected additional loss is negligible.

The only remaining region is the set in which $0 < |\Delta_{x_l}(y)| \leq r_n$. Since r_n tends to zero, both the probability of realizing an outcome in this region and the magnitude of the additional loss from suboptimal classification decisions in this region tend to zero as the sample grows. These two properties, however, are not sufficient to guarantee that the additional loss is negligible at the $n^{-1/2}$ rate. I therefore make an additional assumption that forces the region $\{y \in \mathbb{R}^d : 0 < |\Delta_{x_l}(y)| \leq r_n\}$ to shrink at a sufficiently fast rate, or, in other words, I impose a margin condition (see assumption 5.1.3).

Now I formalize the assumptions I make on the functions $f_{Y,X|Z=z_k}$, the kernel function, and bandwidth sequence:.

Assumption 5.1 (**Data generating process**).

1. ***Nonzero instrument probabilities:*** *A constant* $C \in (0,1)$ *exists such that:*

$$\mathbb{P}[Z = z_k] > C \ \ \forall z_k \in \mathcal{Z}.$$

2. ***Bounded continuous Hessian :*** *For every* $x_l \in \mathcal{X}$ *and* $z_k \in \mathcal{Z}$*, the function* $f_{Y,X|Z=z_k}(y, x_l)$ *is twice continuously differentiable in* y *with Hessian matrix* $\nabla^2 f_{Y|X=x_l,Z=z_k}$*. Furthermore, fixed constants* $M > 0$ *and* $M' > 0$ *exist such that:*

$$\sup_{y \in \mathbb{R}^d} f_{Y,X|Z=z_k}(y, x_l) \leq M,$$

and

$$\sup_{||u||=1} ||(\nabla^2 f_{Y|X=x_l,Z=z_k})u|| \leq M'$$

3. ***Margin condition:*** *For every* $x_l \in \mathcal{X}$ *and* $z_k \in \mathcal{Z}$*, positive constants* B, α*, and* $\delta > 0$ *exist such that* $\forall t \in (0, \delta)$

$$\boldsymbol{\mu}\{y \in \mathcal{Y} : 0 < |f_{Y,X|Z=z_1}(y, x_l) - f_{Y,X|Z=z_0}(y, x_l)| \leq t\} \leq Bt^{\alpha}$$

Where $\boldsymbol{\mu}$ *denotes the Lebesgue measure over* $\mathbb{R}^d$*.*

Assumption 5.1 imposes mild conditions on the data generating process. 5.1.1 requires that all instrument values occur with positive probability in the population, which is guaranteed to occur in most empirical settings. 5.1.2 requires that the distribution of the outcome be bounded, and twice continuously differentiable with bounded Hessian. I combine this assumption with standard properties of kernel density estimators (KDE) to argue that the contribution of density estimation bias is negligible for classification mistakes. The margin condition from 5.1.3 is typically imposed in the nonparametric classification literature to obtain fast rates of convergence for estimated classifiers (Tsybakov (2004)). In the one-dimensional outcome case, to satisfy the margin condition, values of α as low as $1/4$ suffice, depending on the bandwidth choice.

Assumption 5.2 (**Kernel function and bandwidth sequence**).

1. ***Kernel regularity:*** *The kernel function* $K : \mathcal{Y} \mapsto \mathbb{R}$ *is positive, symmetric*

around zero, bounded, square integrable and has finite second moment. Furthermore, it satisfies assumption (K_1) *from Giné and Guillou (2002).*[13]

2. ***Bandwidth sequence restrictions:*** *The bandwidth sequence* h_n *satisfies:*

   (a) $h_n = n^{-\gamma}$ *for some* $\gamma > \frac{1}{d+4}$.

   (b) $nh_n^d / ln(n) \longrightarrow \infty$.

   (c) The constant α *from assumption 5.1.3 and* γ *satisfy:* $\alpha > \frac{d\gamma}{1-d\gamma}$.

Assumption 5.2 is a set of common requirements for kernel functions and bandwidth sequences. 5.2.1 is required to control bias and to use the results from Giné and Guillou (2002) to bound the probability that KDEs take extreme values and thus incur in classification mistakes. Typical kernel choices such as gaussian, triangular, uniform or Epanechnikov satisfy this assumption. Part (a) of 5.2.2 simply requires some degree of undersmoothing so that KDEs bias is negligible, part (b) requires that the bandwith decreases at a rate slower than $n^{-1/d}$. Part (c) is necessary to invoke the margin condition from assumption 5.1.3 and guarantee that the bias caused by classification mistakes in the proximity of the contact set is asymptotically negligible.

Assumption 5.3 (Nondegenerate joint inference)**.**

1. ***Strictly positive-overlaps:*** *A fixed constant* $\epsilon > 0$ *exist such that, for every* $x_l \in \mathcal{X}$*:*

$$\Psi_{x_l} > \epsilon.$$

2. ***Nontrivial classification problem*** *A constant* $\epsilon > 0$ *exist such that for every* $x_l \in \mathcal{X}$ *and every* $z_k \in \mathcal{Z}$*:*

$$\operatorname*{plim}_{n\to\infty} \int_{D_k} f_{Y,X|Z=z_k}(y)dy > \epsilon$$

Where $D_k = \mathbb{R}^d \setminus \{y \in \mathbb{R}^d : \eta^*_{n_1}(y, x_l) \neq k\}$.

[13]The assumption requires that K lies in the linear span of functions $\{g : \mathcal{Y} \mapsto \mathbb{R}_+\}$ such that the set $\{(s,u) : g(s) \geq u\}$ can be represented as a finite number of Boolean operations among sets of the form $\{(s,u) : p(s,u) \geq \phi(u)\}$ where p is a polynomial and ϕ an arbitrary real function. See Giné and Guillou (2002) for a discussion of this assumption.

Assumption 5.3 is only necessary for testing IV validity, it is not required for inference on subdensity overlaps. The condition $\Psi_{x_l} > 0$ from assumption 5.3.1 ensures that the limit distribution of the overlap estimator is nondegenerate asymptotically. This condition is testable using the proposed inference framework on the overlap parameters alone.

Assumption 5.3.2 rules out the possibility that, as the sample grows, the classifier estimated in the first split always selects the same instrument value. This condition guarantees that the overlap estimator is not perfectly correlated with conditional treatment probability estimators.

Assumption 5.3.2 can only be violated in two cases, the first is when the contact set is equal to $\mathbb{R}^d$ and yet, even as the sample size tends to infinity, the estimated classifier picks the same fixed instrument value z_k for every $y \in \mathbb{R}^d$. Since this first case appears implausible in practice, I rule it out. The second case is when one subdensity strictly dominates the other almost everywhere. This second condition is testable and, if verified, both the classification and overlap estimation problems become trivial with a sufficiently large sample.

Under these assumptions, I obtain valid joint inference on conditional treatment probabilities and subdensity overlaps.

Theorem 5.1. *Suppose assumptions 5.1, 5.1.2, and 5.3 hold.*

Define the vectors $\boldsymbol{\theta}_n^T = (\boldsymbol{P}_{\boldsymbol{\mathcal{X}_0},\boldsymbol{n}}^T, \boldsymbol{P}_{\boldsymbol{\mathcal{X}_1},\boldsymbol{n}}^T, \boldsymbol{\Psi}_n^T)$ *and* $\boldsymbol{\theta}_{\mathbf{0}}^T = (\boldsymbol{P}_{\boldsymbol{\mathcal{X}_0}}^T, \boldsymbol{P}_{\boldsymbol{\mathcal{X}_1}}^T, \boldsymbol{\Psi}^T)$ *and let* m *denote their dimension. Let* $\boldsymbol{\Omega}_n$ *denote the symmetric positive definite matrix* $n_2 Var(\boldsymbol{\theta}_n | S_1)$.

Then:

$$\sqrt{n_2}\boldsymbol{\Omega}_n^{-1/2}\big(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\big) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_m)$$

Furthermore, the formulas from lemma B.8 in appendix B show how to construct an estimator $\boldsymbol{\Sigma}_n$ *for the matrix* $\boldsymbol{\Omega}_n$. *The matrix* $\boldsymbol{\Sigma}_n$ *satisfies:*

$$\big|\big|\boldsymbol{\Sigma}_n - \boldsymbol{\Omega}_n\big|\big| \xrightarrow{p} 0.$$

In lemma B.5 from appendix B I show that sample splitting implies that the expectation of the empirical loss $L_{n_2}(\eta_{n_1}^* m, x_l)$ is equal to the theoretical loss $L(\eta_{n_1}^* m, x_l)$. Hence, since $L(\eta_{n_1}^* m, x_l)$ converges to Ψ_{x_l}, the overlap estimators I propose are asymptotically unbiased.

Note that theorem 5.1 does not yield asymptotic normality for $\boldsymbol{\theta}_n$ but for the normalized sequence $\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{\theta}_n$. The reason is that, even if the overlap estimator Ψ_{x_l,n_1} converges to a fixed constant, it is obtained by using the second split to evaluate the empirical loss of the classifier trained in the first split $\eta^*_{n_1}$. Therefore, its variance conditional on the first split depends on $\eta^*_{n_1}$. If the contact set has positive measure, the estimated classifier $\eta^*_{n_1}$ has no stable limit and thus $\sqrt{n_2}(\Psi_{x_l,n} - \Psi_{x_l})$ has no stable asymptotic distribution.

In particular, the conditional variance of the empirical loss in the second split is given by:

$$n_2 Var\big(\Psi_{x_l,n} - \Psi_{x_l}\big|S_1\big) = \sum_{k=0,1} \frac{P_{\xi,x_l|z_k,n_2}(1 - P_{\xi,x_l|z_k,n_2})}{\mathbb{P}[Z = z_k]}$$

Where $P_{\xi,x_l|z_k,n_2} := \mathbb{P}[\eta^*_{n_1}(Y, X_l, Z; z_k, x_l) \neq k, X = x_l|Z = z_k]$.

If classification decisions in the contact set are dictated by statistical noise, then, $\eta^*_{n_1}$ randomly re-classifies values inside the contact set from one instrument value to another. Furthermore, this re-classification does not stop as the sample size tends to infinity. Hence, $P_{\xi,x_l|z_k,n_2}$ does not converge to a fixed constant and as a consequence the limit of the scaled variance does not converges in probability to a fixed constant. The same phenomenon occurs with the covariance between overlap estimators associated with different treatment values, and their covariance with conditional treatment probabilities.

Variance instability does not prevent inference on the vector of overlap parameters, it only implies that it has no fixed limit distribution. That is why normalization by $\boldsymbol{\Omega}_n^{-1/2}$ is required to obtain asymptotic normality.

5.4 Formal sharp tests of IV validity

Now I discuss how to implement a test of instrument validity based on the estimators from subsection 5.3 and the characterizations from theorems 3.1 and 4.1. In particular, I show that the testing problem satisfies the assumptions from Cox and Shi (2023) and thus their method is applicable.

Let $\boldsymbol{\theta_0}$ denote the vector of nonredundant conditional treatment probabilities and subdensity overlaps. Theorem 3.1 shows that the null hypothesis that assumptions 2.1 (IV independence and exclusion) and 2.2 (shape restriction) are jointly satisfied is

equivalent to the hypothesis that the linear system:

$$\boldsymbol{H}_0 : A\boldsymbol{p} \leq \boldsymbol{\theta}_0, \quad \boldsymbol{p}_0 \in \Delta(\mathcal{T}), \tag{17}$$

admits at least one solution. Here, A_{Sharp} is a deterministic matrix given by $A_{Sharp}^T = (A_Z^T, A_X^T, A_Y^T)$, the matrices A_Z^T, A_X^T and A_Y^T are as in theorem 3.1, and $\boldsymbol{p}_0$ is a distribution over instrument-response types. That is, $\boldsymbol{p}_0$ is a nonnegative vector of dimension L^K such that the sum of its coordinates is equal to one.

In the case of *generalized no-defiers restrictions* from section 4, theorem 4.1 shows that the hypothesis can also be cast as:

$$A_{GFOSD}\boldsymbol{\theta}_0 \leq 0 \tag{18}$$

Where A_{GFOSD} is a matrix with deterministic, known entries, obtained from stacking all the equations from theorem 4.1.

Cox and Shi (2023) show that in both cases (inequality 17 or inequality 18) the null hypotheses can be recast as

$$\boldsymbol{H}_0 : T := \min_{\boldsymbol{b}: A\boldsymbol{b} \leq 0} (\boldsymbol{\theta} - \boldsymbol{b})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{\theta} - \boldsymbol{b}) = 0, \tag{19}$$

for some fixed, deterministic matrix A with known coefficients and a positive-definite matrix $\boldsymbol{\Omega}$. Furthermore, they propose a testing procedure based on comparing the sample analogue of T to the critical values of a χ-squared distribution with degrees of freedom equal to the rank of the rows of A associated with the inequalities that bind at the optimal $\boldsymbol{b}$.

The following definition establishes a class of distributions for which the assumptions I impose for inference on the vector of subdensity overlaps hold uniformly.

Definition 5.1 (**Admissible nulls**). *For some fixed kernel and bandwidth that satisfy assumption 5.2.2, and deterministic sample-split size sequences n_1, n_2, let $\mathcal{F}$ be the class of distributions satisfying:*

1. ***IV validity and shape restriction:*** *Assumptions 2.1 and 2.2 .*

2. ***Uniform nondegeneracy:*** *Assumption 5.1 holds for every $F \in \mathcal{F}$, with fixed*

constants $C, M, M', B, \alpha, \delta$. *So that:*

$$\inf_{F\in\mathcal{F}} \min_k \mathbb{P}_F(Z = z_k) \geq C,$$
$$\inf_{F\in\mathcal{F}} \min_k \mathbb{P}_F(X = x_l | Z = z_k) \geq C_X, \quad \forall x_l \in \boldsymbol{\mathcal{X}}_k$$
$$\sup_{F\in\mathcal{F}} \max_l \mu\{y : 0 < |\Delta_{F,l}(y)| \leq t\} \leq Bt^\alpha, \qquad 0 < t < \delta.$$

And a fixed $\varepsilon_\Psi > 0$ *exists such that* $\inf_{F\in\mathcal{F}} \min_l \Psi_{F,l} \geq \varepsilon_\Psi$.

3. ***Uniform nontrivial classification:*** *Define* $q_{F,n,k,l}(S_1)$ *as:*

$$q_{F,n,k,l}(S_1) := \mathbb{P}_F\big(X = x_l, \eta_{n_1}(Y, x_l) = k \mid Z = z_k, S_1\big).$$

There exists $\varepsilon_D > 0$ *such that*

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{P}_F \left\{ \min_{k=0,1} \min_{0\leq l<L-1} q_{F,n,k,l}(S_1) < \varepsilon_D \right\} = 0.$$

Where $\mathbb{P}_F(A)$ *denotes the probability of event* A *under* F.

The first part of 5.1 simply requires that the null hypothesis of IV validity holds. The second restricts the class to data generating processes for which the inference procedure for subdensity overlaps is valid and ensures that all data generating processes yield nondegenerate variance-covariance matrices. The third condition is a strengthening of the second part of assumption 5.3.2. It rules out cases in which, for a sufficiently large sample, the classifier returns the same instrument value for every y making overlap estimators and conditional treatment probability estimators collinear.

Theorem 5.2. *Suppose that* $\boldsymbol{\theta}_n$ *is the estimator from theorem 5.1 with variance estimator* $\boldsymbol{\Sigma}_n$ *and let* A *be the matrix from equation (19).*

Define:

$$T_n := \min_{\boldsymbol{b}: A\boldsymbol{b}\leq 0} (\boldsymbol{\theta}_n - \boldsymbol{b})^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} (\boldsymbol{\theta}_n - \boldsymbol{b}), \tag{20}$$

and let $\hat{r}$ *be the rank of the matrix comprised by the rows of* A *associated with the inequalities that bind at the optimal* $\boldsymbol{b}$ *in equation 20.*

For $q \in [0, 1]$, *let* $\chi^2_{m,1-q}$ *be the* $1-q$*-quantile of a* χ*-squared distribution with* m *degrees of freedom. Define* $\phi_n^{CC}(q) := \mathbf{1}_{\{T_n > \max\{\chi^2_{\hat{r},1-q}, tol\}\}}$ *and some tolerance* $tol > 0$.

Then:

$$\limsup_{n\to\infty} \sup_{F_{Y,X,Z}\in\mathcal{F}} \mathbb{P}_{F_{Y,X,Z}}\left[\phi_n^{CC}(q) = 1\right] \leq q$$

To prove theorem 5.2 I proceed by contradiction. If the inequality in 20 is violated, there exists a sequence of data generating processes and sample sizes with a subsequence for which the rejection probability exceeds q by a fixed positive amount. This implies the existence of a sequence of training-sample realizations such that, conditional on these realizations and the classifier estimated on them, the rejection probability in the evaluation sample remains bounded away from q. But, conditional on the training subsamples, the assumptions from Lemma 2 in Cox and Shi (2023) are satisfied. Thus, the rejection probability cannot be larger than q.

In the previous argument, the role of definition 5.1 is to define a class in which the method for inference on subdensity overlaps is uniformly valid. Since establishing uniform inference under the weakest possible regularity conditions is not an objective of this paper, I simply impose conditions that are sufficient for this application.

6 Simulations

In this section, I conduct a series of Monte Carlo simulations to evaluate the performance of the test of IV validity and the method for inference on subdensity overlaps that I propose. I find that the theoretical properties established in section 5 are supported by the simulation results, and that both the test and the estimator are robust to choices of tuning parameters such as the kernel function, the bandwidth sequence, and which fraction of the data is used to estimate the kernel-based classifiers. Moreover, I show that, for some data generating processes, my test outperforms existing, discretization-based tests of IV validity.

In subsection 6.1 I show simulations that assess the performance of the method for inference on subdensity overlaps. In subsection 6.2 I assess the performance of the test of IV validity.

Tuning parameters: The classification-based approach to overlap estimation I propose has three main tuning parameters: the kernel function, the bandwidth sequence, and the fraction of the data that is used to estimate the classifier. I denote these tuning parameters as K, h_n and c_p respectively. Furthermore, for $p \in (0,1)$, c_p

denotes that a share p of the data is used to estimate the classifier, and thus, a share $1-p$ is used to evaluate its loss.

Throughout all simulations I consider three kernel functions: Uniform, Gaussian and Epanechnikov. For the bandwidth sequence, I consider three different rates: $h_n = s_n n^{-1/3}$, $h_n = s_n n^{-1/4}$ and $h_n = s_n n^{-1/5}$ where the scaling factor s_n is defined so that, when $h_n \propto n^{-1/5}$, the bandwidth coincides with Silverman's rule of thumb[14] (Silverman (2018)). Finally, I consider the splits $c_{1/3}$, $c_{1/2}$ and $c_{2/3}$.

6.1 Joint estimation and inference on subdensity overlaps

Data generating processes: The main motivation for the classification-based estimator I propose is its robustness to the contact set. Moreover, the margin condition from assumption 5.1.3 plays a key role in deriving the asymptotic properties of the estimator. Thus, in addition assessing how different choices of tuning parameters affect the estimator, in the simulations, I also examine how the estimator's performance is affected by the measure of the contact set, and the slopes of the subdensities $f_{Y,X|Z=z_0}$ and $f_{Y,X|Z=z_1}$ near the boundary of the contact set.

The data generating processes of this subsection are tailored to study the performance and robustnees of the subdensity overlap estimator without depending or relying on the IV structure for interpretation. I consider simulations with two instrument, and two treatment values (i.e. $\mathcal{Z} = \{z_0, z_1\}$ and $\mathcal{X} = \{x_0, x_1\}$). The unconditional instrument probabilities are $\mathbb{P}[Z = z_0] = 1/3$ and $\mathbb{P}[Z = z_1] = 2/3$, that is, $Z = z_1$ is more likely and thus, subdensities associated with $Z = z_0$ will be more precisely estimated. Since the focus of these simulations is overlap estimation and not instrument validity, I impose independence between the random variable X and the random vector (Y, Z). Hence, treatment probabilities are given by $\mathbb{P}[X = x_l | Z = z_k] = 0.5$, for $k = 0, 1$.

As a baseline case, I assume that $Y|Z = z_k$ follows a normal distribution with standard deviation $\sigma = 1$ and let the mean depend on Z. More specifically, I assume that $Y|Z = z_k \sim \mathcal{N}(\mu_{z_k}, 1)$ with $\mu_{z_0} = 0$ and $\mu_{z_1} = \tau \geq 0$. Observe that when $\mu_1 = 0$, the subdensities $f_{Y,X|Z=z_0}$ and $f_{Y,X|Z=z_1}$ are identical and thus the contact set correspond to the entire real line. On the other hand, if $\mu_1 > 0$, the subdensities differ everywhere except in one point, and the contact set has zero measure.

Since μ_1 affects the point in which $f_{Y,X|Z=z_0}$ and $f_{Y,X|Z=z_1}$ cross, it is related to

[14] That is, $s_n = 0.9 \min\left\{\hat{\sigma}, \frac{IQR}{1.34}\right\}$, where $\hat{\sigma}$ is the sample standard deviation

the margin condition. For example, suppose that μ_1 is small, then, the two densities will cross near the point $y = 0$ in which the derivative of the normal density is zero. Consequently, in a neighborhood of the crossing point, the derivatives of $f_{Y,X|Z=z_0}$ and $f_{Y,X|Z=z_1}$ are also close to zero. Thus, separation between subdensities is locally slow and an estimated classifier will require a large sample to confidently identify the true minimizer in such neighborhood. Furthermore, since $y = 0$ is the global maximizer of the standard normal distribution, realizations of Y near the crossing point occur with non-negligible probability. On the other hand, if μ_1 is large, the densities will cross at a point in which either their derivatives are larger, or their mass is close to zero. In both of these two cases, classification mistakes have a smaller impact on the estimator than when μ_1 is close to zero.

Additionally, I consider DGPs in which the contact set is non-negligible, but different from the entire support of Y. Specifically, I modify the normal densities so that exactly in the crossing point, a region in which both densities are flat and identical appears. I denote the width of this flat contact region by w. Thus, different combinations of μ_1 and w allow me to assess how the performance of the estimator is affected as the measure of the contact set, and the separation rate near its boundary.

In figure 5 I illustrate the data generating process for different values of μ_1 and w. Below every figure I report the parameters μ_1 and w, as well as the integral of $f_{Y,X|Z=z_k}(y, x_l)$ over the contact set (denoted Φ_C), and the absolute difference of the derivatives of the normal densities at the crossing points, that is $\Delta_C := |f_{Y,X|Z=z_0}(\tilde{y}, x_l) - f_{Y,X|Z=z_1}(\tilde{y}, x_l)|$[15], where $\tilde{y}$ is the crossing point. The data generating process is formally described in appendix C.

Reported statistics: I report three main summaries of the simulations. First, I report the mean of the quantity $\Psi_{x_0,n} - \Psi_{x_0}$, that is the difference between the estimated and true overlap associated with $X = x_0$. Second, I use the overlap and variance estimators I propose in section 5 to construct confidence intervals at the 95% confidence level for a single overlap. Finally, the instrument validity test requires joint inference on the vector $\boldsymbol{\theta}$ of subdensity overlaps and conditional treatment probabilities, hence, I report the observed coverage frequency of ellipsoidal confidence regions at the 95% confidence level for $\boldsymbol{\theta}$ constructed using the joint estimator and variance from section 5.

[15]Since X does not affect the distribution of Y, I omit the dependence of Δ_C on x_l.

To construct ellipsoidal confidence regions, I consider the set:

$$E_{0.95} = \{\tilde{\boldsymbol{\theta}} : (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_n^{-1} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) \leq \chi^2_{3d,0.95}\}$$

Where $\xi^2_{3L,0.95}$ denotes the .95 quantile of a χ squared random variable with $3L$ degrees of freedom[16]. I count and report in what percent of the simulations the true vector $\boldsymbol{\theta}$ is contained in $E_{0.95}$.

Simulation results: Table 1 reports the results from 500 simulations for sample sizes $n = 500$, $n = 1000$, and $n = 2000$. Each column corresponds to a different combination of the bandwidth rate h_n and the sample-splitting regime c_p. The rows correspond to different combinations of μ_1 and w along with one of the three reported statistics. In the main text, I report the results for uniform Kernels, the results for gaussian and Epanechnikov kernels are quantitatively and qualitatively similar and can be found in appendix (C).

The simulations show that the bias of the classification-based estimator is small and that the confidence intervals yield coverage rates close to the nominal level.

The sample splitting regime $c_{1/3}$ under-performs relative to the other alternatives I consider. This result is consistent with the fact that my inference method does not require inference on the estimated classifier, only to evaluate its expected loss, and that sample splitting implies that the loss of the estimated classifier evaluated in the second sample-split, is unbiased for its theoretical loss. Thus, a larger estimation sample improves the accuracy of the classifier without biasing its empirical loss.

In the simulations, bias and confidence interval coverage do not change in a clear direction as the bandwidth varies. One explanation is that, while under-smoothing (i.e. bandwidths that shrink at a rate faster than $n^{-1/5}$) reduces the bias of the first stage kernels and, consequently, of the classifier, it also requires a stronger margin condition. Thus, the gains from improving the classifier far away from the contact set, are offset by an increase in the measure of the region around the contact set where classification is noisy.

[16] Recall that L is the cardinality of the support of X. The vector $\boldsymbol{\theta}$ comprises d subdensity overlaps, and $2d$ conditional treatment probabilities.

6.2 Testing IV validity

Now I present a set of simulations designed to assess the performance of the test of IV validity I propose. Specifically, I investigate the size and power of the test, and compare its performance with the discretization-based test proposed by Kwon and Roth (2024). The simulations show that the test has correct size control and power against violations of IV validity. Furthermore, I show that some data generating processes exist that violate the null of instrument validity but discretization based tests have limited or no power against. In contrast, the proposed test attains higher rejection rates, and has power against all the alternatives considered.

The comparison with discretization-based tests allows me to illustrate two risks that discretizing entails and my test avoids: under rejection due to loss of information, and under-rejection due the bias that arises from approximating a smooth integral with a finite grid. Both risks are caused by the fact that discretization-based tests use conditions that are implied by the sharp testable implications, but don't imply them.

Discretization risks: To discuss discretization, I first define it. Let $\mathcal{P}_{Disc} = \{P_0, ..., P_{N_{Disc}-1}\}$ denote a finite partition of the support of the outcome $\mathcal{Y}$ where each of the sets P_0 is called a *discretization bin*[17]. Then, a *discretized* outcome Y_{Disc} can be constructed according to the following rule: $Y_{Disc} := \sum_{m=0}^{N_{Disc}-1} m\mathbf{1}\{Y \in P_m\}$. Observe that since Y_{Disc} is a deterministic function of Y, it is a random variable with support $\{0, ..., N_{Disc} - 1\}$.

The first problem that discretizing might cause is that any feature of the original distribution of (Y, X, Z) that is inconsistent with the IV validity assumptions, but occurs within the same discretization bin cannot be detected after discretizing.

For example, suppose that the outcome of interest is a measure of health, that treatment is a binary indicator of insurance coverage, and that the instrument reduces the cost of insurance and thus, it weakly promotes its acquisition. Suppose further that access to cheaper insurance allows units that acquire it to save money, and that healthier subjects are more likely to spend these savings on improving their diet and exercising, while the rest of the population spends them on goods that have no effect on their health. In this hypothetical scenario, violations of the exclusion restriction can only occur and be detected at high values of Y, near the right tail of its distribution.

Now suppose that, as is common practice (Kwon and Roth (2024)), the outcome

[17]Note that in practice, when $\mathcal{Y} = \mathbb{R}$, the sets P_0 typically are intervals.

is partitioned into N discretization bins constructed from the quantiles of the unconditional distribution of the health measure. Unless the rigth tail of the support of Y is spread across different discretization bins, the discretized outcome will not reflect the violations of the exclusion that occur in the right tail, while the true distribution of Y might. Thus, the model can only be falsified by a sufficiently fine partition or by a test that does not discretizes.

Furthermore, while fine discretizations might detect violations of IV validity at the population level, in practice, they impose both computational and statistical costs. The computational cost comes from how the system of equations that describe the discretized IV model increases with the cardinality of the discretizing partition. The statistical cost comes from the fact that discretization requires estimating the conditional probability that every single bin is realized. Large cardinality partitions imply that the average number of observations per bin is small, and thus, bin probabilities are noisily estimated.

Discretizing can not only lead to tests that fail to detect violations of IV validity, it can also lead to tests that reject, even when the null hypothesis holds at the population level. To see this, note that discretization is equivalent to using a finite integration grid to estimate the overlap parameter Ψ_{x_l}. Unless the contact set is known, using a discrete grid will necessarily either over-estimate, or under-estimate Ψ_{x_l}. If Ψ_{x_l} is over-estimated, discretization leads to a test that uses a set of sufficient, but not necesary conditions for IV validity, and thus, under-rejects. If Ψ_{x_l} is under-estimated, discretization leads to a test that uses necessary but not sufficient conditions for IV validity, and thus it over-rejects.

Neither testing necessary, nor sufficient conditions can avoid the problem of failing to detect some violations of IV validity. Thus, a testing procedure that over-estimates Ψ_{x_l} can simulatenously and erroneously over and under reject the null of IV validity depending on the true data generating process.

Data generating processes (Undetectable violations of IV validity): The first data generating process I consider corresponds to a case in which the exclusion restriction is only violated at the right hand tail of the distribution of the outcome.

I consider a binary instrument ($\mathcal{Z} = \{z_0, z_1\}$), three different treatment levels ($\mathcal{X} = \{x_0, x_1, x_2\}$), and violations of the exclusion restriction that only occur when $X = x_2$ and near the rightmost tail of the subdensity $f_{Y,X|Z=z_k}(y, x_2)$.

To simplify the simulations, I impose standard ordered monotonicity and let instru-

ment assignment be independent from the realized potential treatments. Thus, absent violations of the exclusion restriction, the densities $f_{Y,X|Z=z_0}(y,x)$ and $f_{Y,X|Z=z_1}(y,x)$ are identical. When exclusion is violated, they only differ at $X = x_2$ and near the righmost tail of $f_{Y,X|Z=z_0}(y,x_2)$. This difference implies that the overlap parameter is strictly smaller than $\mathbb{P}[X = x_2|Z = z_0]$ (which, under monotonicity, corresponds to the share of always-x_2-takers) and therefore, the model is falsified.

I parametrize violations of exclusion with a scalar $\tau_{ud} \in (0,\infty)$. When $\tau_{ud} = 0$ the IV model is not falsifiable, and when $\tau_{ud} > 0$ it is. The data generating process is formally defined in appendix C. The following figure illustrates the functions $f_{Y,X|Z=z_k}$ for both $k = 0, 1$, evaluated at $X = x_2$, and for two different values of τ_{ud}. Additionally, the figures depict the partitions that result from discretizing using 5 and 25 bins constructed from the quantiles of the unconditional outcome distribution.

Data generating processes (Finite grid integration failure): The second data generating process I consider simply lets $Y|Z, X$ follow a normal distribution so that $f_{Y,X|Z=z_k}(y,x_l)$ is proportional to the density of a normal random variable. In this DGP, the instrument is binary, and the treatment has three support points ($\mathcal{X} = \{x_0, x_1, x_2\}$). I let the mean of $Y|Z, X$ depend on both X and Z. Dependence on X represents treatment effects, and dependence of Z violations of the exclusion restriction.

When testing IV validity against this data generating process, I impose monotonicity. Moreover, I allow for the instrument to affect the outcome only when $X = x_1$, that is, a the intermediate treatment value. These conditions imply three things: One, when $X = x_0$, the function $f_{Y,X|Z=z_0}$ strictly dominates the function $f_{Y,X|Z=z_1}$[18]. Similarly, when $X = x_2$, the function $f_{Y,X|Z=z_1}$ strictly dominates the function $f_{Y,X|Z=z_0}$. Finally, at $X = x_1$, neither of the two functions dominates the other, the two subdensities cross and their overlap (Ψ_{x_1}) potentially binds the prevalence of always-x_1-takers. If Ψ_{x_1} is sufficiently large, it does not bind and the model is not falsifibable. On the other hand, as it shrinks, it constraints the admissible share of always takers in the population and eventually implies that the data generating process is inconsistent with instrument independence, exclusion, and monotonicity. That is, if Ψ_{x_1} is small, the IV model is falsified.

For this data generating process, I parametrize violations of the IV model with the scalar τ_{di}. Depending on the value of τ_{di} the joint distribution of Y, X and Z might

[18]Since x_1 is the minimal treatment and its take-up can only decrease when Z changes from $Z = z_0$ to $Z = z_1$.

be jointly compatible or not with IV validity and mononicity. In particular, $\tau_{di} = 1$ corresponds to the case in which the restriction $\mathbb{P}[\textit{Always takers}] \leq \Psi_{x_2}$ binds exactly, thus, $\tau_{di} \leq 1$ implies that IV validity, and monotonicity are not falsifiable, while $\tau_{di} > 0$ implies falsifiability. The formal description of this data generating process is given in appendix C. The following figure illustrates the functions $f_{Y,X|Z=z_0}$ at different values of X.

Simulation results: I implement my test and the discretization-based test proposed by Kwon and Roth (2024) with a nominal size of 5%, and compare their perfomance in Montecarlo simulations. The simulations cover different combinations of tuning parameters and data generating processes. The three main takeaways are that my test is robust to choices of the tuning parameters, that discretization-based tests can be very sensitive to the cardinality of the discretizing partition, and that, for the data generating processes considered, my overlaps-based test has more power against violations of the null.

In the simulations, I consider samples of size 500, 1000, and 2000. For every sample size and data generating process, I draw 500 samples, in every sample I implement multiple versions of both tests. For each version of the test, I compute its average rejection rate across simulated samples.

All results are summarized in table 2. Each panel corresponds to a different data generating process. In both panels, each column corresponds to a different test, and each row to a different combination of data generating process and sample size. I report results for 14 different tests grouped in two categories, one for my method using different combinations of tuning parameters, and one for the method proposed by Kwon and Roth (2024) varying the number of discretization bins.

The test of Kwon and Roth (2024) requires a method for inference on linear systems. They consider different alternatives such as Andrews et al. (2023), Fang et al. (2023), Cox and Shi (2023), or Cho and Russell (2024). Since the test I propose relies on the procedure by Cox and Shi (2023), I implement their method using the same procedure so that differences are not driven by the procedure used for inference on linear systems. Furthermore, I consider 5 discretization cardinalities, $N_{Disc} = 3$, $N_{Disc} = 5$, $N_{Disc} = 15$, $N_{Disc} = 25$, and $N_{Disc} = 50$. For my test, I consider the bandwidth sequences $h_n \propto n^{-1/3}, n^{-1/4}$ and $n^{-1/5}$, and splits $c_{1/3}, c_{1/2}$ and $c_{2/3}$.

The first 9 columns of table 2 report the rejection rates of my method under different tuning parameter values. The last 5 report rejection rates for the method by Kwon

and Roth (2024) under different discretization regimes. The acronym NF indicates the rows that correspond to data generating processes under which monotonicity, exclusion and exogeneity are not jointly falsified. The capital letter F indicates data generating processes under which these assumptions are jointly falsified.

The asymptotic theory from section 5 depends on the full sample size n. Nonetheless, the proposed overlap estimator, its variance, and conditional treatment probability estimators (that is, all the parameters required to test IV validity) are functions of conditional probabilities, where the conditioning is on the different instrument values. Thus, in the tables, I also report the expected number of observations per instrument values, which I denote by $\mathbb{E}n_{z_k}$.

The first panel (A) of table 2 corresponds to the data generating process in which exclusion is only violated at $X = x_2$ (i.e. the highest treatment), and on the right tail of the distribution of the outcome. Recall that $\tau_{ud} = 0$ corresponds to distributions from which IV validity cannot be falsified. Consistent with this, both, my test and the one proposed by Kwon and Roth (2024) do not reject the null hypothesis more than 5% of the times.

However, when τ_{ud} is different from zero, the two tests differ. For a sample size of 500 (first four columns of panel A in table 2), the rejection rate of my method increases as τ_{ud} grows and violations of the assumptions become easier to detect, while discretization retains the exact same rejection rate of 5%.

Furthermore, as the sample size increases, the rejection rate of my test steadily increases. On the other hand, for the test of Kwon and Roth (2024), the relation between the rejection rate and the sample size depends on the number of bins in which the suport of the outcome is partitioned. For 3, 5 and 15 bins, the rejection rate does not increases with the sample size. Only for 25 and 50 bins the rejection rate increases.

When $n = 1000$ and $\tau_{ud} = 1$, my method rejectes between 14% and 24% of the times, while Kwon and Roth (2024) with 25 bins reject 13% of the times. If instead $\tau_{ud} = 1.5$ my test rejects between 15% to 30% of the times while discretization attains a maximum rate of 16%. That is, the best version of my test rejects almost twice as much as the best performing version of the test by Kwon and Roth (2024).

When the sample size is equal to 2000, a similar patern arises. In general, my test using either $c_{1/2}$ or $c_{2/3}$ has a rejection rate that is between 3/2 and 2 times larger than the discretization-based test. The only exception is when $n = 2000$ and $\tau_{ud} = 1.5$. That is, for the largest simulated violation and with the largest simulated

sample size. However, for that same sample size, and a less severe violation of the IV assumptions, my test with $c_{1/2}$ or $c_{2/3}$ attains rejection rates between 12% and 20% while discretization with 50 bins only rejects 6% of the times.

The second panel (B) of table 2 shows results for the data generating process in which falsifiability depends on how accurately the integral Ψ_{x_l} is estimated. As in panel A, for my tests and when the null of IV validity holds, the rejection rate is not higher than 5%. When the null is violated, the rejection rate is higher than 5%. Furthermore, the sample splitting regimes $c_{1/2}$ or $c_{2/3}$ ouperform the regime $c_{1/3}$.

On the other hand, the results in panel B of table 2 for the test proposed by Kwon and Roth (2024) show how discretization might lead to under-rejection when the null is falsified, and that rejection rates are sensitive to the number of bins.

I find that only the test based on five bins has rejection rates comparable to my proposed test. On one hand, coarse partitions tend to under-reject due to over-estimating the overlap parameter, while fine partitions lose statistical power since, for any fixed sample size, finer partitions imply that the statistical noise that arises from estimating each of the bins increases.

The test based on five bins, however, does not improves its performance as steeply as tests based on finer discretizations. The reason might be that, the observable implications of the model associated with a five bins discretization are weaker than those associated with finer partitions, or with the sharp test based on subdensity overlaps.

Throughout all simulations reported in panel A, three main patterns arise. The first pattern is that, my method has noticeably more power with $c_{1/2}$ or $c_{2/3}$, that is, when at least half the data is used to train the classifier, than under the sample splitting regime $c_{1/3}$. This is consistent with the discussion and results from the previous subsection: additional estimation data improves the classifier, while the empirical loss of the classifier is unbiased for its theoretical loss, and thus, on average, the gains from additional evaluation data are smaller. Nonetheless, $c_{1/2}$ consistently outperforms $c_{2/3}$ and discretization.

The second, is that my test is robust to choices of tuning parameters, specially bandwidth choices. On the other hand, the simulations show that discretization-based tests are sensitive to the cardinality of the discretizting partition and conclusions might substantially differ depending on which value of this tuning parameter is chosen.

Finally, the simulations show that, consistent with theory, the rejection rate of my method increases both with the sample size, and when joint violations of the IV

assumptions become stronger. On the other hand, discretization-based tests might not necesarilly improve with the sample size, and its power against stronger violations of the null depends on the choice of discretizing partition.

Multiple outcomes and other simulations: The identification and inference theory I present in this paper is not restricted to one dimensional outcomes, it extends naturally to a muti-valued Y. In appendix C I show an additional set of simulations in which I implement the test with a two dimensional outcome[19]. The results are qualitatively similar to the ones I show in this section.

Additionally, I assess the robustness of the method to the case in which some treatments occur with low probability in the population and thus, their associated overlaps are effectively estimated from just a few observations. I find that the test retains correct size and some power against nulls that are inconsistent with IV validity.

Finally, to assess the quality of the inputs that the final step of the test receives (i.e. my implementation of the method by Cox and Shi (2023)), for all the data generating processes I consider in this section, in appendix C I also assess the bias, and confidence interval coverage of my overlap estimators. I find low bias, and close to nominal coverage rates.

7 An application to the Oregon Health Experiment

In this section I illustrate the test using data from the Oregon Health Experiment (OHE). A large-scale experiment in which IV methods within the framework of this paper have been used by Finkelstein et al. (2012) and Rose and Shem-Tov (2024) to study the effect of insurance on health and hospital utilization. Applying the test from this paper, I reject the extensive margin compliers only (EMCO) assumption invoked by Rose and Shem-Tov (2024). I find, however, that slight modifications of this assumption that allow for moderate intensive margin effects are not rejected by the data. Leveraging these alternative assumptions, I re-estimate the effect of increasing insurance coverage and obtain results that are quantitatively similar to those of Finkelstein et al. (2012) and Rose and Shem-Tov (2024).

The OHE and plausible shape restrictions: Medicaid is a government program in the United States that provides health insurance to adults and children with limited

[19]The online appendix can be found at https://ricardomirandam.github.io/files/OnlineAppendix.pdf

income and resources. In the state of Oregon, in the United State, access to Medicaid is provided by the State government through the *Oregon Health Plan*. In January 2008, the government of the state of Oregon announced that it would expand the Oregon Health Plan. The goal of this expansion was to enroll 10,000 new individuals into Medicaid. However, the state anticipated that demand would exceed this number, and thus, the opportunity to apply for coverage was randomized through a public lottery. This randomization, makes this empirical setting ideal to study the effects of health insurance among low income individuals.

Individuals selected by the lottery had the option to enroll into medicaid, but were not obliged to do so. Thus, the lottery is exogenous, but insurance is not. The endogeneity of insurance can be addressed using instrumental variables, but for this, independence is not sufficient. It must jointly hold along with the exclusion restriction and any shape restrictions that researchers might want to invoke.

By design, the lottery results are binary (being selected or not). Furthermore, in the United States, the minimum frequency at which insurance is acquired and payed is monthly. Thus, the duration of insurance is, by definition, discrete. That is, the setting and variables in the OHE fall directly into the framework of this paper.

The first paper to study the OHE using IV methods was Finkelstein et al. (2012). In their analysis, they used the binary lottery results as an instrument for a binary treatment variable defined as an indicator that an individual had insurance for at least one month between the randomization and the end of the study. Two of the outcomes they considered were indices that measured health and hospital utilization. Having a binary instrument and a binary treatment, they argued that, under an exclusion restriction plus a no-defiers/monotonicity assumption, the 2SLS estimand could be interpreted as LATE.

In a posterior paper, Rose and Shem-Tov (2024) argued that binarizing the treatment ignores changes in the duration of insurance that might invalidate interpreting the resulting estimates as LATE. Absent the lottery, some households might have acquired insurance for a few months, but not for as many months as they would if given the option to enroll into Medicaid. Binarizing the treatment variable ignores these changes and its potential effects on health and hospital utilization. To recover a LATE interpretation for the 2SLS estimand, Rose and Shem-Tov (2024) proposed a different assumption: Extensive Margin Compliers Only (EMCO). This assumption imposes that winning the lottery can only affect insurance decisions on the extensive margin.

That is, absent the experiment, inviduals would either acquire no insurance, or acquire it for the exact same duration for which they had Medicaid.

Rose and Shem-Tov (2024) argue that the EMCO assumption is plausible in the OHE. Their main argument is that, to be eligible for Medicaid, individuals are required to have been uninsured during at least 6 uninterrupted months. Thus, a person selected by the lottery who already had insurance cannot take advantage of Medicaid in order to retain insurance for a longer time. Nonetheless, shape restrictions in general, and the EMCO assumption in particular, do not only constraint realized treatments, but also potential treatments. For example, in a hypothetical scenario with no lottery, an individual might acquire private insurance only for 3 months (due to late acquisition, or failure to retain it). If that same individual was selected by the lottery, having Medicaid as a cheaper alternative might allow the individual to acquire and retain insurance for an entire year. The Medicaid requirements do not restrict hypothetical behaviors and thus, do not rule out this possibility[20].

An alternative, less restrictive assumption that might hold in the OHE, is ordered monotonicity. Angrist and Imbens (1995) show that, when treatments are ordered and under the assumption that the instrument can only induce units to adopt weakly higher treatments, the 2SLS estimand can be interpreted as a weighted average of the individual treatment effects among the units whose treatment is changed by the instrument (i.e. among compilers). Ordered monotonicity does not rule out the possibility that, absent the lottery, individuals who enroll into Medicaid might have still acquired insurance for at least some months. It simply requires that they remain enrolled into Medicaid for at least as long as they would have retained private insurance. This is plaussible since Medicaid is just an additional, affordable, available alternative for lottery winners.

Figures 4 (a) and (b) show the histogram of the duration of insurance conditional on being selected ($Z = 1$), and not selected ($Z = 0$) by the lottery. Subfigure (a) shows all durations observed in the sample (from 0 to 19 months). This figure reveals two patters, first, a large share of the population was not insured, even when they had the option to enroll into Medicaid, and second, it is clear that the share of individuals who never had insurance is much lower among lottery winners. This second finding

[20]Furthermore, prior to the lottery, the government of Oregon advertised that the Oregon Health Standard program was to be expanded so that all interested individuals could participate from the lottery. It is plausible that this advertising campaign increased awareness about insurance and its benefits which in turn might lead some individuals who where eligible for Medicaid, to acquire private insurance.

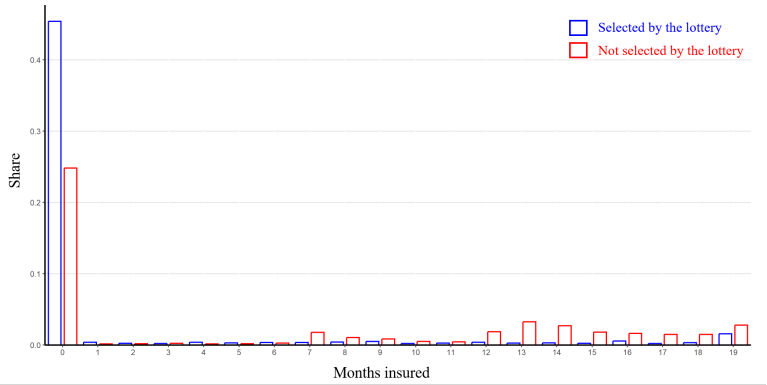
confirms that a large portion of the effect of the lottery was on the extensive margin, that is, a large fraction of the selected individuals went from having no insurance, to having insurance for some positive number of months.

But, for EMCO to hold, it is not sufficient that the lottery had a large effect on the extensive margin. It also requires that there are no intensive margin effects. Subfigure 4 (b) shows the same histogram but now restricted to strictly positive durations of insurance. From this figure, it is clear that durations between 1 and 7 months were less prevalent among lottery winners. This is inconsistent with EMCO, if the intervention only affected the intensive margin, all positive insurance durations should be weakly more prevalent.

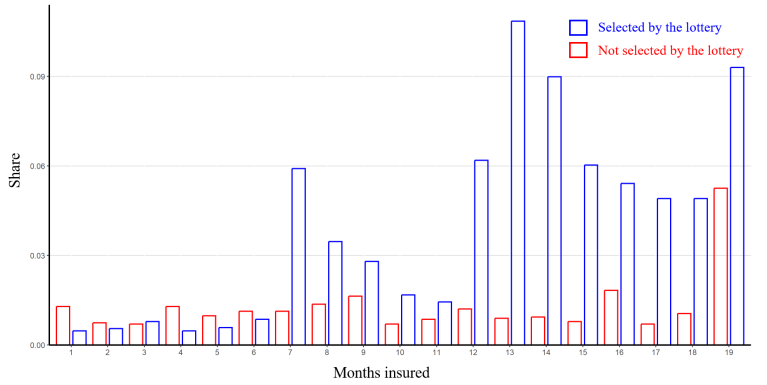
Note that while subfigure 4 (b) suggests that EMCO is not consistent with the data, it is unclear if these differences are statistically meaningful, or attributable to sampling noise. The need to distinguish these two possibilities further motivates testing the validity of shape restrictions on their own, and jointly with exclusion and IV independence.

Figures 4 (a) and (b) suggest one additional, plausible restriction. While most of the effect of the lottery was on the extensive margin, there were also some individuals that, absent the experiment, would have acquired insurance but failed to retain it for more than $1 \leq t \leq 7$ months. On the other hand, it could be that individuals who had private insurance for more than t months are less financially constrained, and thus, as EMCO imposes, would have held Medicaid (if available to them), for the exact same duration that they held private insurance. One advantage of this restriction is that the valid binarization conditions from Rose and Shem-Tov (2024) hold. Consequently, it restores a LATE interpretation for the the 2SLS estimand but now as the effect among individuals who where affected on the extensive margin, or induced to retain insurance for more than t months.

Figure 4: Histograms of insurance duration in the OHE



(a) All durations.



(b) Strictly positive durations.

Notes: *These two figures show the histrograms of duration of insurance across the OHE sample. Duration of insurance is measured in months, and ranges from 0 to 19. The red bars represent the histogram for individuals who where not selected by the lotery, the blue bars represent the histrogram for individuals who where selected by the lottery. Sibfigure (a) shows the histogram including all observed durations. Subfigure (b) restricts the histrogams to strictly positive durations.*

Thus, I test multiple potentially valid shape restrictions: EMCO, Ordered Monotonicity, and the assumption described in the previous paragraph for multiple values of t. The role of these assumptions is to facilitate identification of the LATE, to this end, the must hold jointly along with independence and the exclusion restriction. Hence, I do not only test the shape restrictions, but their joint validity along with independence and the exclusion.

By design, the lottery is randomized and thus, independent from potential treatments and outcomes. However, there are several reasons why the exclusion restriction

might fail in the OHE. The firs reason is income effects. Since Medicaid reduces the cost of health insurance, enrolled individuals may have additional savings which they can spend in ways that improve their health. For example, the additional monetary resources could allow individuals to improve their diets. If these indirect effects are sufficiently large, they could invalidate the IV assumptions.

Furthermore, both Finkelstein et al. (2012) and Rose and Shem-Tov (2024) define the treatment variable as months of insurance. This definition underpins one assumption: that N months of private insurance have the same effect as N months of Medicaid. While plaussible, it is not obvious that this assumption hold. If it doesn't, then, the same unit might have different outcomes when $X = N$ and $Z = 0$ than when $X = N$ and $Z = 0$. This would be a violation of the exclusion restriction.

Test implementation: The outcomes of interest in Finkelstein et al. (2012) are multiple measures of health and hospital utilization. In the main analysis of Rose and Shem-Tov (2024), these outcomes are aggregated into two Z-score variables, one for health, and one for hospital utilization. In my implementations of the test, I consider these two variables as the outcome Y. I first apply the test to each of the two scores separately, and then, taking advantage of the fact that both the identification and inference results of this paper apply to multi-dimensional outcomes, I jointly apply the test to both indices.

I report the results of the test for gaussian kernels with a bandwidth of $h_n \propto n^{-1/5}$. I use half the sample to estimate the classifier, and the other half to evaluate its loss and estimate conditional treatment probabilities. The reason I report the results for these values of the tuning parameters is that they achieved the best performance in the simulations from section 6. For alternative choices of the tuning parameters, I obtained qualitatively similar results which I omit in the interest of brevity.

Two features of the OHE make it different from the simulations from section 6. One is the fact that the outcome is two-dimensional. Another is the fact that the treatment takes 20 different values, and some of them occur with very low frequency in the sample. In appendix I assess the robustness of my method to data generating processes with these features. I find that in boths cases the test has adequate size control, and power against alternatives under which IV validity is falsified.

The expansion of the Oregon Health Standard occurred in 7 different waves. Across waves, the probability of wining the lottery was different. Hence, I apply the test separately to each wave. Additionally, I pool together waves 1, 2 and 3, as well as

waves 4, 5 and 6. I choose this particular pooling because the hypothesis that the probability of wining the lottery is identical across all waves within each of the two the pools is not rejected. Pooling allows to implement the test on a larger sample. The results of the test are consistent across waves and pooled waves.

While the lottery was at the individual level, the option to enroll into Medicaid was given at the household level, thus, individuals in larger households have a higher probability of effectively wining the lottery. Moreover, some observations were assigned to an intensive-follow up survey program and thus, their outcomes are more like to be observed in the final sample. I restrict the testing sample to one person households who were not part of the intensive follow up survey. The resulting testing sample consist of 8,625 different individuals, and represents 58% of the entire sample. Since, within this subsample the lottery remains independent from potential treatments and outcomes, if the IV validity assumptions hold for the entire population, they must also hold for this subsample.

Restricting to households who were not part of the intensive follow up survey simplifies implementation, but is not necessary to apply the test. Both the kernel density estimators used in the classifier, and the second sample empirical loss could be estimated using weights that reflect the different sample probabilities.

Results: I present the results of the test in table 3. I do not jointly reject ordered monotonicity along with the exclusion restriction and instrument independence. I reject however the EMCO assumptions. That is, I find that the lottery had an effect on the intensive margin too.

Additionally, for each t between 2 and 7, I fail to reject the hypothesis that the lottery did not affect insurance duration among individuals who would otherwise retained it for more than t months, but that it allowed individuals that would have been insured for less that t months to remain insured for more than t months.

The results also show how accounting for the fact that the outcome is multivalued shrinks the set of admissible distributions which in turn increases the power of the test. This is clear in the case of the monotonicity assumption, where, separately testing IV validity for one outcome yields p-values as high as 1. While, jointly testing validity for the two outcomes substantially reduces the obtained p-values.

Finally, I report weighted LATE estimates under all the assumptions that were not rejected in at least one of the tests. For the EMCO assumption, and each assumption associated with a different threshold t, the estimands where obtained by recoding the

original, multi-valued treatment variable according to Rose and Shem-Tov (2024) and then running 2SLS using the binarized treatment variable.

For ordered monotonicity, the estimator was obtained using 2SLS on the original treatment variable. In this monotone case, Angrist and Imbens (1995) show that the 2SLS estimand is a weighted average of local treatment effects among compliers, but, units whose treatment increases the most (in terms of duration) as a result of winning the lottery, receive larger weight in the final estimand. All estimation results are shown in table 4 for both the full sample, and the subsample in which I implemented the test.

In all cases, I find that increasing insurance (either on the intensive margin, extensive margin, or both) has a positive and significant effect on health and hospital utilization. However, under ordered monotonicity the magnitude of the effect is about 10 times smaller than under the rest of constraints.

Finally, I find that for all the restrictions associated with a threshold $t = 1, ..., 7$, the LATE estiamtes are similar in magnitude to the estimate under EMCO, which corresponds to the original numbers reported both by Finkelstein et al. (2012) and Rose and Shem-Tov (2024). The reason is that, while the lottery had an effect on the intensive margin, the share of intensive-margin compliers is small relative to the share of extensive-margin compliers. Consequently, the LATE estimator does not change substantially after accounting for the intensive-margin compliers.

Leveraging the restrictions associated with a threshold $t = 1, ..., 7$, I find that the the effect of increasing insurance among all units that were either switched on the extensive margin, or allowed to retain insurance from less than t to at least $t + 1$ months was between 0.15 and 0.160 standard deviations for the health index, and between 0.24 and 0.26 standard deviations for hospital utilization.

8 Conclusion

This paper combines two novel results to obtain a sharp, general test of instrument validity. The first result is a novel characterization of the sharp observable implication of the IV model. This characterization reduces the set of infinite restrictions that the exclusion restriction combined with the distribution of the outcome impose on the model to an equivalent, finite set of linear constrains on the selection model. That is, a set of additional inequalities on the linear, finite, discrete model that describes how the instrument affects treatment adoption.

The second novel result is a valid joint inference procedure for the quantities involved in the finite characterization of the observable implications of the model. Some of these quantities are subdensity overlaps, that is, integrals of the pointwise minimum of conditional densities over some strict subset of their support. Naive plug in approaches to subdensity overlap estimation suffer two problems, one, whenever the two densities are equal, applying the minimum operator biases the estimator. Second, analyzing the behavior of the in-sample minimizers is technically challenging. I avoid both problems by noting that the target parameters can be recast as the expected loss of nonparametric classifiers. This insight allows me to obtain a valid, root-n consistent, asymptotically normal, asymptotically unbiased estimator.

These two novel results, combined with the inference method for linear systems proposed by Cox and Shi (2023) allow me to derive a sharp test of instrument validity that jointly accommodates continuous, potentially multi-valued outcomes, along with a general class of restrictions on how the instrument affects treatment adoption.

In an application to the OHE, I illustrate the test, and how it can be used to interpret and formulate plausible restrictions on how a lottery that randomized access to Medicaid affected insurance on the extensive and intensive margin. Leveraging these restrictions, I obtain estimates for the effect of increasing the duration of insurance coverage on health and hospital utilization. My findings are similar to those of Finkelstein et al. (2012) and Rose and Shem-Tov (2024). I emphasize, however, that part of the effect came from changes in the duration of insurance on the intensive margin.

References

Gordon Anderson, Oliver Linton, and Yoon-Jae Whang. Nonparametric estimation and inference about the overlap of two distributions. *Journal of Econometrics*, 171(1):1–23, 2012.

Isaiah Andrews, Jonathan Roth, and Ariel Pakes. Inference for linear conditional moment inequalities. *Review of Economic Studies*, 90(6):2763–2791, 2023.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. 2007.

Yuehao Bai, Andres Santos, and Azeem M Shaikh. On testing systems of linear inequalities with known coefficients. Technical report, Working Paper, 2022.

Yuehao Bai, Shunzhuang Huang, Sarah Moon, Andres Santos, Azeem M Shaikh, and Edward J Vytlacil. Inference for treatment effects conditional on generalized principal strata using instrumental variables. *arXiv preprint arXiv:2411.05220*, 2024a.

Yuehao Bai, Shunzhuang Huang, Sarah Moon, Azeem M Shaikh, and Edward J Vytlacil. On the identifying power of monotonicity for average treatment effects. *arXiv preprint arXiv:2405.14104*, 2024b.

Yuehao Bai, Shunzhuang Huang, and Max Tabord-Meehan. Sharp testable implications of encouragement designs. *arXiv preprint arXiv:2411.09808*, 2024c.

Yuehao Bai, Kirill Ponomarev, Andres Santos, Azeem M Shaikh, Max Tabord-Meehan, and Alexander Torgovitsky. Inference for linear systems with unknown coefficients. *arXiv preprint arXiv:2604.24904*, 2026.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176, 1997.

Kirill Borusyak. Bounding the population shares affected by treatments. *Available at SSRN 2473827*, 2015.

JoonHwan Cho and Thomas M Russell. Simple inference on functionals of set-identified parameters defined by linear moments. *Journal of Business & Economic Statistics*, 42(2): 563–578, 2024.

Gregory Cox and Xiaoxia Shi. Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *The Review of Economic Studies*, 90(1):201–228, 2023.

Christian M Dahl, Martin Huber, and Giovanni Mellace. It is never too late: a new look at local average treatment effects with or without defiers. *The Econometrics Journal*, 26(3): 378–404, 2023.

Clement De Chaisemartin. Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2):367–396, 2017.

Raju Dey, Arne C Bathke, and Somesh Kumar. Inference on overlap index: with an application to cancer data. *The International Journal of Biostatistics*, (0), 2025.

Ivar Ekeland, Alfred Galichon, and Marc Henry. Optimal transportation and the falsifiability of incompletely specified economic models. *Economic Theory*, 42:355–374, 2010.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019.

Zheng Fang, Andres Santos, Azeem M Shaikh, and Alexander Torgovitsky. Inference for large-scale linear systems with known coefficients. *Econometrica*, 91(1):299–327, 2023.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012.

Brigham Frandsen, Lars Lefgren, and Emily Leslie. Judging judge fixed effects. *American Economic Review*, 113(1):253–277, 2023.

Koki Fusejima. Identification of multi-valued treatment effects with unobserved heterogeneity. *Journal of Econometrics*, 238(1):105563, 2024.

Alfred Galichon and Marc Henry. Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298, 2011.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

Leonard Goff. A vector monotonicity assumption for multiple instruments. *arXiv preprint arXiv:2009.00553*, 2020.

Leonard Goff. When does iv identification not restrict outcomes? *arXiv preprint arXiv:2406.02835*, 2024.

Leonard Goff and Eric Mbakop. Inference on the value of linear programs. *arXiv preprint arXiv:2506.06776*, 2025.

James J Heckman and Rodrigo Pinto. Unordered monotonicity. *Econometrica*, 86(1):1–35, 2018.

Martin Huber. Sensitivity checks for the local average treatment effect. *Economics Letters*, 123(2):220–223, 2014.

Martin Huber and Giovanni Mellace. Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics*, 97(2):398–411, 2015.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Hiroaki Kaido and Kirill Ponomarev. Testing exclusion and shape restrictions in potential outcomes models. *arXiv preprint arXiv:2512.20851*, 2025.

Amir Aamodt Kazemi and Inge Christoffer Olsen. Instrumental variable analysis with categorical treatment. *Statistical Methods in Medical Research*, 33(11-12):2043–2061, 2024.

Désiré Kédagni, Huan Wu, and Yi Cui. Robust identification in randomized experiments with noncompliance. *arXiv preprint arXiv:2408.03530*, 2024.

Lars J Kirkeboen, Edwin Leuven, and Magne Mogstad. Field of study, earnings, and self-selection. *The quarterly journal of economics*, 131(3):1057–1111, 2016.

Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015.

Toru Kitagawa. The identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics*, 225(2):231–253, 2021.

Jon Kleinberg and Éva Tardos. *Algorithm Design*. Addison Wesley, 2006.

Patrick Kline and Christopher R Walters. Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics*, 131(4):1795–1848, 2016.

Soonwoo Kwon and Jonathan Roth. Testing mechanisms. *arXiv preprint arXiv:2404.11739*, 2024.

Sokbae Lee and Bernard Salanié. Treatment effects with targeting instruments. *arXiv preprint arXiv:2007.10432*, 2020.

Alexander W Levis, Matteo Bonvini, Zhenghao Zeng, Luke Keele, and Edward H Kennedy. Covariate-assisted bounds on causal effects with instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf028, 2025.

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Romuald Meango, Marc Henry, and Ismael Mourifie. Combining stated and revealed preferences. *arXiv preprint arXiv:2507.13552*, 2025.

Magne Mogstad, Alexander Torgovitsky, and Christopher R Walters. The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–3698, 2021.

Jack Mountjoy. Community colleges and upward mobility. *American Economic Review*, 112 (8):2580–2630, 2022.

Ismael Mourifié and Yuanyuan Wan. Testing local average treatment effect assumptions. *Review of Economics and Statistics*, 99(2):305–313, 2017.

Manu Navjeevan and Rodrigo Pinto. Ordered, unordered and minimal monotonicity criteria. Technical report, UCLA, mimeo, https://www. rodrigopinto. net, 2022.

Claudia Noack. Sensitivity of late estimates to violations of the monotonicity assumption. *arXiv preprint arXiv:2106.06421*, 2021.

Rodrigo Pinto. Beyond intention to treat: Using the incentives in moving to opportunity to identify neighborhood effects. *NBER Working Paper*, 2021.

Thomas S Richardson and James M Robins. Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 25:415–444, 2010.

Evan K Rose and Yotam Shem-Tov. On recoding ordered treatments as binary indicators. *Review of Economics and Statistics*, pages 1–32, 2024.

Vira Semenova. Debiased machine learning of aggregated intersection bounds and other causal parameters. *Available at SSRN 5134514*, 2025.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Zhenting Sun. Instrument validity for heterogeneous causal effects. *Journal of Econometrics*, 237(2):105523, 2023.

Elie Tamer and Christopher D Walker. Nonparametric bayesian inference for partially identified discrete response models. *arXiv preprint arXiv:2608.25814*, 2026.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Nadja van't Hoff, Arthur Lewbel, and Giovanni Mellace. *Limited Monotonicity and the Combined Compliers LATE*. University of Southern Denmark, Faculty of Business and Social Sciences . . . , 2023.

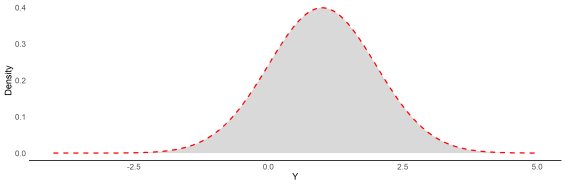
Edward Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

Luther Yap. Sensitivity of policy-relevant treatment parameters to violations of monotonicity. *Journal of Applied Econometrics*, 2025.

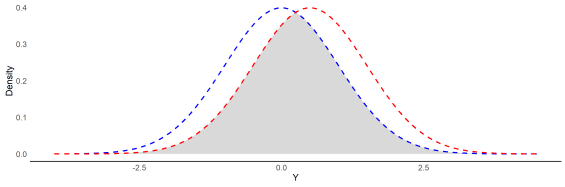
Figures

Figures for section 6

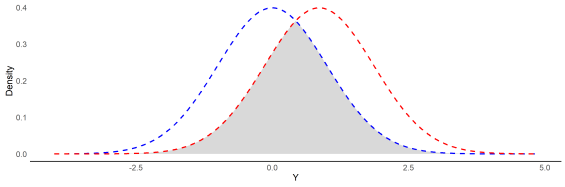
Figure 5: Outcome distributions, contact set, and density separation under different DGPs.



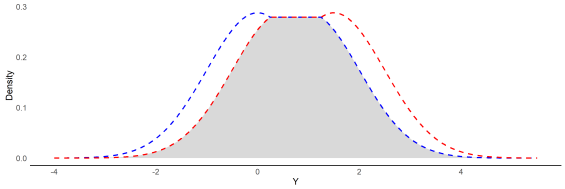
(a) $\mu_1 = 0, \Phi_C = .5$



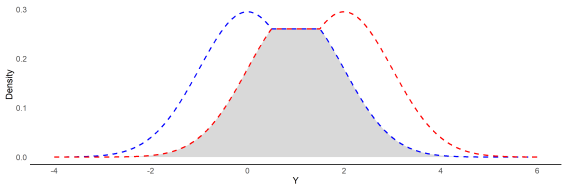
(b) $\mu_1 = 0.5, \Phi_C = 0, |\Delta_B| = 0.31$



(c) $\mu_1 = 1, \Phi_C = 0, |\Delta_B| = 0.31$



(d) $\mu_1 = 0.5, \Phi_C = .28, |\Delta_B| = 0.16$



(e) $\mu_1 = 1, \Phi_C = .27, |\Delta_B| = 0.31$

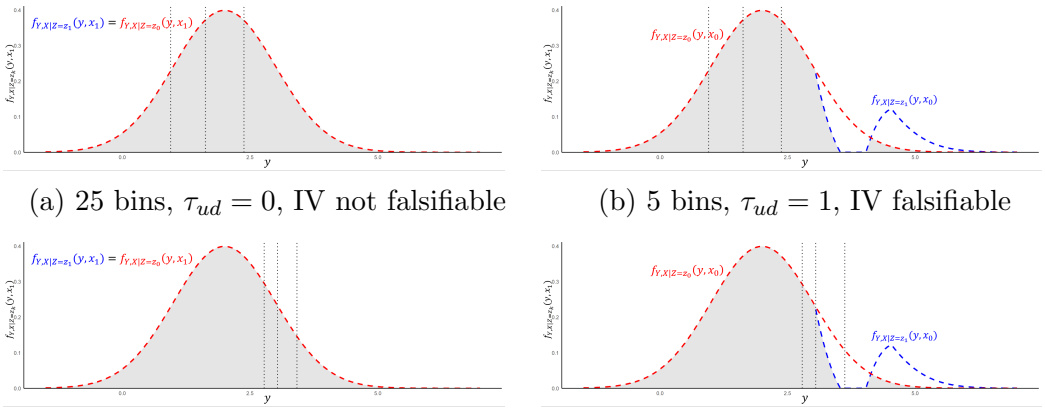
Notes: *These figures depict the functions* $f_{Y,X|Z=z_k}(y, x_l)$ *fixing* $x_l = x_2$*, for* $k = 0, 1$*, for the data generating process I use to assess the performance of the method for inference on subdensity overlaps.*

Subfigure (a) represents the baseline case in which the functions $f_{Y,X|Z=z_0}(y, x_l)$ *and* $f_{Y,X|Z=z_1}(y, x_l)$ *are proportional to a standard normal distribution. In this case, the contact set is equal to the entire real line.*

Subfigures (b) and (c) represent the case in which $f_{Y,X|Z=z_0}(y, x_l)$ *is proportional to a standard normal while the mean associated with* $f_{Y,X|Z=z_1}(y, x_l)$ *is shifted by* μ_1*. In this case, the contact set has zero measure. Smaller values of* μ_1 *imply that the difference between subdensities is harder to estimate and thus the classifier is more likely to make mistakes.*

Finally, subfigures (d) and (e) correspond to the case in which, at the crossing point, a flat region exists in which both subdensities are equal. In this case, the contact set is nonnegligible, but different from the entire real line.

Figure 6: Data generating process: Undetectable IV validity violations



(a) 25 bins, $\tau_{ud} = 0$, IV not falsifiable

(b) 5 bins, $\tau_{ud} = 1$, IV falsifiable

(c) 5 bins, $\tau_{ud} = 0$, IV not falsifiable

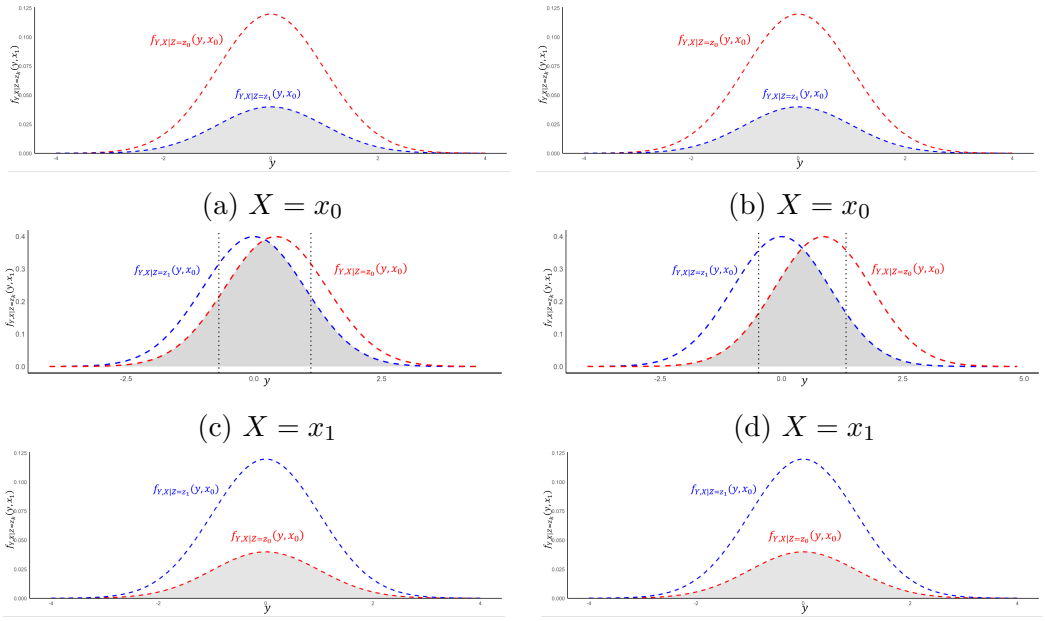
(d) 25 bins, $\tau_{ud} = 0$, IV falsifiable

Notes: *These figures depict the functions $f_{Y,X|Z=z_k}(y, x_l)$ fixing $x_l = x_2$, for $k = 0, 1$, for the data generating process in which violations of the exclusion restriction only occur at the rightmost tail of the distribution. The instrument is binary (i.e. $\mathcal{Z} = \{z_0, z_1\}$), and the treatment takes three values $\mathcal{X} = \{x_0, x_1, x_2\}$. The outcome Y is continuous with support $\mathbb{R}$. The scalar τ_{ud} parametrizes violations of IV validity, $\tau_{ud} = 0$ implies that the IV assumptions are not falsified.*

The figures in the first column (subfigures (a) and (c)) show the baseline case in which the two densities are equal and thus, there are no violations of the exclusion restriction, and outcomes only depend on realized treatment. These two figures depict the subdensity that results from fixing $X = x_2$, that is, $f_{Y,X|Z=z_k}(y, x_2)$ for every $y \in \mathbb{R}$. In this baseline case, $Y|X, Z$ follows a normal distribution and X affects its mean linearly.

The figures in the second column (subfigures (b) and (d)) depict a case with violations of the exclusion restriction. In particular, at treatment value $X = x_2$, some of the mass in the right tail of the normal distribution is shifted further right, but only if $Z = z_1$. This results in an overlap area that is strictly smaller than in the benchmark case, which in turn leads to independence, exclusion, and monotonicity being jointly falsified.

The vertical lines in the subfigures in the top (subfigures (a) and (b)) row depict the limits of the three rightmost discretization bins that would result if discretization was done using 5 bins constructed from unconditional quantiles of the observed density f_Y. The figures in the bottom row (subfigures (b) and (d)) only differ in that the vertical lines depict the three rightmost bins that would result if 25 discretization bins were used instead of 5.



(a) $X = x_0$ (b) $X = x_0$

(c) $X = x_1$ (d) $X = x_1$

(e) $X = x_2$ (f) $X = x_2$

Notes: *These figures depict the functions $f_{Y,X|Z}(y, x_l)$ over $\mathcal{Y} = \mathbb{R}$, for different fixed values of x_l. The first row (subfigures (a) and (b)) corresponds to x_0, the second (subfigures (c) and (d)) to x_1 and the third (subfigures (e) and (f)) to x_2. The data generating process satisfies monotonicity, that is, when Z changes from z_0 to z_1, units either realize the same treatment, or switch to one with a strictly higher index.*

In the first row, in both columns, treatment $X = x_0$ is more likely when $Z = z_0$, thus, the function $f_{Y,X|Z=z_0}$ dominates $f_{Y,X|Z=z_1}$. Equivalently, $X = x_2$ is more likely when $Z = z_1$ and thus, in the third row, $f_{Y,X|Z=z_1}$ dominates $f_{Y,X|Z=z_0}$. This is because x_0 and x_2 are the minimal and maximal treatment respectively and thus, the former is weakly disincentivized by the instrument, while the later is weakly incentivized.

The second row corresponds to $X = x_1$. Since this is not a minimal or maximal treatment, units that switch from both x_0 to x_1, and from x_1 to x_2 might exist. Nonetheless, the population frequency of always-x_1-takers is bounded above by the overlap Ψ_{x_2}. This constraint might be binding, and if Ψ_{x_2} is sufficiently small, imply that the model is falsified. The figure in the second row, in the first column, corresponds to the case $\tau_{di} = 0.5$, while the one to the left to $\tau_{di} = 1.5$. In the first case, the restriction on always takers is not binding and thus the data generating process is compatible with independence, exogeneity and ordered monotonicity. In the second case, the restriction is binding and IV validity is falsified.

Tables

Tables for section 6

Table 1: Bias, overlap coverage, and joint overlaps and CTPs coverage for sample splitting estimators.

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$		
$n = 500$											
$\mathbb{E}n_{z_0} = 167$, $\mathbb{E}n_{z_1} = 333$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.94	0.94	0.95	0.95	0.95	0.94	0.93	0.94	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.93	0.93	0.93	0.92	0.93	0.92	0.89	0.91	0.92
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.04	0.03	0.02	0.03	0.02	0.02	0.03	0.02	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.85	0.91	0.92	0.87	0.91	0.93	0.88	0.92	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.73	0.82	0.87	0.74	0.83	0.89	0.73	0.82	0.90
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.93	0.94	0.94	0.94	0.93	0.93	0.94	0.93	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.93	0.92	0.90	0.92	0.92	0.90	0.92	0.91	0.91
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.03	0.02	0.01	0.02	0.01	0.01	0.02	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.90	0.90	0.93	0.88	0.90	0.93	0.90	0.92	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.81	0.88	0.90	0.80	0.87	0.91	0.79	0.87	0.89
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.93	0.95	0.93	0.94	0.94	0.93	0.94	0.94	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.92	0.91	0.91	0.92	0.91	0.92	0.92	0.91	0.92
$n = 1000$											
$\mathbb{E}n_{z_0} = 333$, $\mathbb{E}n_{z_1} = 666$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.93	0.96	0.95	0.94	0.96	0.95	0.94	0.96	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.94	0.94	0.93	0.94	0.95	0.94	0.93	0.94	0.93
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.03	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.84	0.89	0.92	0.86	0.91	0.92	0.88	0.92	0.92
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.72	0.85	0.89	0.77	0.88	0.91	0.79	0.89	0.92
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.96	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.93	0.94	0.92	0.93	0.94	0.93	0.92	0.94	0.93
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.89	0.92	0.93	0.90	0.92	0.93	0.89	0.92	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.80	0.88	0.90	0.81	0.89	0.90	0.81	0.87	0.91
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.94	0.96	0.95	0.94	0.96	0.93	0.95	0.94	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.94	0.95	0.94	0.94	0.96	0.95	0.95	0.96	0.95
$n = 2000$											
$\mathbb{E}n_{z_0,x_l} = 666$, $\mathbb{E}n_{z_1,x_l} = 1333$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.96	0.96	0.95	0.95	0.95	0.95	0.94	0.94	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.94	0.96	0.95	0.95	0.96	0.95	0.94	0.94	0.95
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.87	0.92	0.93	0.92	0.93	0.94	0.92	0.93	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.79	0.89	0.92	0.86	0.91	0.92	0.87	0.91	0.92
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.96	0.95	0.94	0.96	0.96	0.94	0.96	0.95	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.95	0.93	0.94	0.95	0.94	0.94	0.94	0.94	0.94
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.91	0.93	0.94	0.92	0.94	0.94	0.92	0.93	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.85	0.91	0.93	0.87	0.92	0.93	0.86	0.92	0.94
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.94	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.93

Notes: *This table shows the average bias ($\Psi_{x_l,0} - \Psi_{x_0}$), confidence interval coverage rate ($\Psi_{x_l,n} \in C_{0.95}$), and confidence ellipsoid coverage rates ($\boldsymbol{\theta}_0 \in E_{0.95}$) for overlap estimators obtained using Gaussian kernels with different combinations of tuning parameters. The sample size is represented by n at the top of each panel along with $\mathbb{E}n_{z_k}$ which represents the expected number of observations per instrument value. The bandwidth*

rate is given by h_n

The parameters w *and* μ_1 *define different data generating processes.* μ_1 *determines the difference in means between the subdensities* $f_{Y,X|Z=z_0}$ *and* $f_{Y,X|Z=z_1}$*, and consequently, the point in which the subdensities cross. The parameter* w *determines the width of the contact set at the point in which the subdensities cross. If* w *is equal to zero, the contact set has zero measure. If* $w > 0$ *then, near the crossing points, a region in which the two densities are equal exist. When* $\mu_1 = 0$ *both subdensities are equal throughout the entire support of* Y *and thus, the contact set corresponds to the entire real line.*

Table 2: Null hypothesis of IV validity rejection rates.

Panel A: Undetectable violations of the assumptions

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
		KDE overlaps									KR Discretization				
DGP		$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$							
$n = 500$ $\mathbb{E}n_{z_k} = 250$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{ud} = 0$	NF	0.04	0.03	0.03	0.04	0.03	0.04	0.05	0.03	0.03	0.06	0.03	0.02	0.01	0.01
$\tau_{ud} = 0.5$	F	0.05	0.06	0.07	0.06	0.05	0.07	0.07	0.06	0.06	0.06	0.03	0.02	0.03	0.02
$\tau_{ud} = 1$	F	0.08	0.10	0.11	0.09	0.12	0.10	0.09	0.10	0.10	0.06	0.03	0.02	0.05	0.04
$\tau_{ud} = 1.5$	F	0.11	0.14	0.14	0.11	0.17	0.14	0.12	0.17	0.14	0.06	0.03	0.02	0.06	0.05
$n = 1000$ $\mathbb{E}n_{z_k} = 500$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{ud} = 0$	NF	0.05	0.05	0.03	0.05	0.05	0.03	0.05	0.05	0.04	0.07	0.04	0.04	0.01	0.00
$\tau_{ud} = 0.5$	F	0.09	0.10	0.08	0.08	0.10	0.08	0.09	0.11	0.09	0.07	0.04	0.03	0.06	0.02
$\tau_{ud} = 1$	F	0.14	0.20	0.17	0.14	0.21	0.16	0.14	0.24	0.18	0.07	0.04	0.03	0.13	0.07
$\tau_{ud} = 1.5$	F	0.15	0.24	0.23	0.17	0.28	0.22	0.19	0.30	0.24	0.07	0.04	0.03	0.13	0.16
$n = 2000$ $\mathbb{E}n_{z_k} = 1000$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{ud} = 0$	NF	0.04	0.05	0.04	0.03	0.05	0.04	0.03	0.03	0.05	0.06	0.05	0.03	0.02	0.01
$\tau_{ud} = 0.5$	F	0.09	0.17	0.12	0.08	0.20	0.14	0.09	0.18	0.15	0.06	0.05	0.03	0.18	0.06
$\tau_{ud} = 1$	F	0.19	0.38	0.28	0.19	0.40	0.30	0.23	0.43	0.31	0.06	0.05	0.03	0.27	0.51
$\tau_{ud} = 1.5$	F	0.24	0.48	0.36	0.26	0.51	0.39	0.28	0.52	0.39	0.06	0.05	0.03	0.27	0.83

Panel B: Discrete integration failure

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
		KDE overlaps									KR Discretization				
DGP		$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$							
$n = 500$ $\mathbb{E}n_{z_k} = 250$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$	NF	0.02	0.01	0.02	0.01	0.01	0.01	0.03	0.01	0.02	0.00	0.01	0.01	0.00	0.00
$\tau_{di} = 0.5$	F	0.06	0.07	0.08	0.06	0.08	0.08	0.06	0.08	0.07	0.02	0.05	0.03	0.00	0.00
$\tau_{di} = 1$	F	0.14	0.20	0.15	0.13	0.21	0.18	0.13	0.21	0.17	0.05	0.19	0.10	0.01	0.00
$n = 1000$ $\mathbb{E}n_{z_k} = 500$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$	NF	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.03	0.03	0.00	0.01	0.00	0.00	0.00
$\tau_{di} = 0.5$	F	0.08	0.12	0.09	0.08	0.12	0.10	0.08	0.12	0.10	0.01	0.10	0.05	0.00	0.00
$\tau_{di} = 1$	F	0.21	0.39	0.31	0.22	0.39	0.33	0.21	0.42	0.32	0.10	0.36	0.29	0.05	0.00
$n = 2000$ $\mathbb{E}n_{z_k} = 1000$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$	NF	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0.5$	F	0.12	0.20	0.17	0.12	0.22	0.18	0.12	0.22	0.18	0.01	0.09	0.12	0.01	0.00
$\tau_{di} = 1$	F	0.36	0.62	0.52	0.35	0.65	0.52	0.38	0.65	0.54	0.10	0.54	0.59	0.21	0.02

Notes: *This table shows the rejection rate across 500 simulations for different joint tests of instrument validity and monotonicity. The first 9 columns correspond to versions of the test I propose with different tuning parameters. The last 5 correspond to implementations of the test proposed by Kwon and Roth (2024) with different cardinalities of the discretizing partition. The parameter c_{prop} for prop $\in \{1/3, 1/2, 2/3\}$ denotes that a proportion prop of the data was used to estimate the subdensities and classifier while a fraction $1-$prop was used to evaluate the empirical loss of the estimated classifier and to estimate conditional treatment probabilities. h_n denotes the bandwidth rate. For the discretization based tests, b_{np} denotes that*

the support of the outcome was partitioned into np bins to implemet the test. The sample size is represented by n at the top of each panel along with $\mathbb{E}n_{z_k}$ which represents the expected number of observations per instrument value. The acronym NF stands for not-falsifiable and indicates data generating processes for which the null hypothesis of instrument validity and monotonicity cannot be jointly falsified. On the other hand, F indicates data generating processes for which the null is falsifiable.

Tables for section 7

Table 3: Joint tests of IV validity.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Shapre R.	Mntn.	Ext.	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$
Waves 1,2 and 3 (n=2615)													
Health	1	0.64	0.658	0.658	0.658	0.547	0.547	0.739	0.578	0.164	0	0	0
Hospital	1	0.734	0.851	0.898	0.849	0.74	0.377	0.64	0.012	0.002	0	0	0
Joint	0.188	0.563	0.455	0.455	0.455	0.437	0.698	0.671	0.041	0.006	0	0	0
Waves 4,5 and 6 (n=3920)													
Health	1	0.049	0.254	0.254	0.174	0.24	0.14	0.367	0.041	0.001	0	0	0
Hospital	1	0.044	0.235	0.172	0.11	0.147	0.081	0.078	0	0	0	0	0
Joint	0.729	0.007	0.012	0.023	0.041	0.18	0.15	0.237	0	0	0	0	0
Wave 1 (n=890)													
Health	1	0.252	0.252	0.217	0.316	0.979	1	1	0.048	0.004	0	0	0
Hospital	0.155	0.278	0.156	0.142	0.187	0.104	0.048	0.152	0.001	0	0	0	0
Joint	0.245	0.302	0.302	0.283	0.629	0.42	0.692	0.698	0.076	0.008	0	0	0
Wave 2 (n=876)													
Health	0.158	0.182	0.25	0.387	0.384	0.216	0.216	0.382	0.061	0.051	0	0	0
Hospital	0.158	0.205	0.282	0.159	0.266	0.328	0.421	0.185	0.01	0.014	0	0	0
Joint	0.082	0.054	0.067	0.017	0.049	0.138	0.168	0.123	0.009	0.018	0	0	0
Wave 3 (n=849)													
Health	0.711	0.683	0.683	0.583	0.622	0.529	0.433	0.377	0.12	0.017	0	0	0
Hospital	0.871	0.368	0.368	0.381	0.448	0.359	0.273	0.461	0.18	0.032	0	0	0
Joint	0.301	0.347	0.347	0.278	0.298	0.237	0.184	0.134	0.033	0.004	0	0	0
Wave 4 (n=1105)													
Health	0.212	0.207	0.277	0.277	0.254	0.362	0.903	0.903	0.91	0.833	0	0	0
Hospital	0.212	0.128	0.168	0.172	0.154	0.203	0.279	0.357	0.051	0.039	0	0	0
Joint	0.584	0.127	0.166	0.166	0.132	0.156	1	1	0.658	0.816	0	0	0
Wave 5 (n=1149)													
Health	1	0.133	0.209	0.285	0.389	0.317	0.465	0.548	0.017	0.003	0	0	0
Hospital	1	0.179	0.294	0.386	0.357	0.308	0.412	0.406	0.001	0	0	0	0
Joint	0.549	0.392	0.495	0.63	0.596	0.552	0.712	0.66	0.028	0.005	0	0	0
Wave 6 (n=1666)													
Health	0.139	0.108	0.07	0.054	0.03	0.026	0.026	0.026	0.035	0.034	0	0	0
Hospital	0.648	0.251	0.204	0.343	0.459	0.485	0.328	0.428	0.098	0.121	0	0	0
Joint	0.377	0.261	0.196	0.452	0.35	0.365	0.365	0.365	0.356	0.349	0	0	0

Notes: *This table shows the p-values for the test of joint validity of instrument independence, the exclusion restriction, and a different shape restrictions. Each column corresponds to a different shape restriction. The firs column corresponds to ordered monotonicity, the second to the EMCO assumption, columns 3 to 13 represent different versions of the assumption that winning the lottery promoted insurance duration strictly above t, for t ranging from 2 to 12. The table reports the reslts of applying the test wave by wave, as well as pooling waves 1,2 and 3 toghether, as well as waves 4,5 and 6. Within each wave or group of waves I report three versions of the test: first, applied just to the health index as the outcome, second, applied to just the hospital utilization index, and third, applied jointly to both the health and hospital utilization indices, treated as a two-dimensional outcome. Each row corresponds to a different wave-outcome(s) combination. The test was implemented using Gaussian kernels, a bandwidth $h_n \propto n^{-1/5}$ and Silverman's rule, and using half the data to estimate kernel densities and the nonparametric classifier.*

Table 4: Weighted local average treatment effect estimates.

Panel A: Full sample										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Shape R	Mntn.	Ext.	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$
Health										
WLATE	0.011***	0.150***	0.147***	0.147***	0.146***	0.145***	0.145***	0.144***	0.160***	0.170***
s.e.	(0.004)	(0.049)	(0.049)	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)	(0.053)	(0.056)
Compliers share		0.292	0.296	0.297	0.298	0.300	0.301	0.301	0.273	0.257
Wave-Hh size FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hospital Utilization										
WLATE	0.007**	0.099**	0.098**	0.097**	0.097**	0.096**	0.096**	0.096**	0.106**	0.113**
s.e.	(0.003)	(0.045)	(0.045)	(0.045)	(0.044)	(0.044)	(0.044)	(0.044)	(0.049)	(0.052)
Compliers share		0.028	0.027	0.027	0.026	0.026	0.025	0.025	0.026	0.028
Observations	14,782	14,782	14,782	14,782	14,782	14,782	14,782	14,782	14,782	14,782
Wave-Hh size FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: One person households, standard follow up										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Shape R	Mntn.	Ext.	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$
Health										
WLATE	0.013***	0.182***	0.180***	0.179***	0.180***	0.177***	0.176***	0.176***	0.192***	0.201***
s.e.	(0.004)	(0.050)	(0.049)	(0.049)	(0.049)	(0.049)	(0.048)	(0.048)	(0.053)	(0.055)
Compliers share		0.338	0.342	0.343	0.342	0.346	0.348	0.349	0.318	0.304
Wave-Hh size FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hospital Utilization										
WLATE	0.009***	0.122***	0.120***	0.120***	0.120***	0.119***	0.118***	0.117***	0.128***	0.134***
s.e.	(0.003)	(0.042)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.040)	(0.044)	(0.046)
Compliers share		0.027	0.026	0.025	0.025	0.024	0.023	0.022	0.024	0.025
Observations	8,625	8,625	8,625	8,625	8,625	8,625	8,625	8,625	8,625	8,625
Wave-Hh size FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: *This table shows LATE type estimates. That is, estimates of weighted averages of individual treatment effects among the complier groups that different shape restrictions allow. Column one is estimated by 2SLS of the outcomes of interest on duration of insurance, using the lottery as an instrument. Under ordered monotonicity, this can be interpreted as an estimate for a weighted average of the treatment effect among individuals who increased their duration of insurance as a result of the lottery (see Angrist and Imbens (1995)). Columns (2) to (10) report estimates for average treatment effect under the assumption that the lottery incentivized individuals to swich from an insurance level below a threshold t to an insurance strictly higher than t, but induced no switches within durations below t, nor within durations above t. The first column corresponds to $t = 1$, that is, the EMCO assumption. The estimates are obtained using 2SlS on a binarized constructed treatment variable as described in Rose and Shem-Tov (2024).*

Appendix

This appendix contains the proofs for sections 3 (except sharpness in the three instruments case) and 5 as well as the main simulation results. For all other proofs, discussions, and simulations, please visit the online appendix at https://ricardomirandam.github.io/files/OnlineAppendix.pdf.

A Proofs for section 3

A.1 Proofs for subsection 3.1

Proof of theorem 3.1.

1⟹2:

The proof follows immediately from theorem 3.2 setting $\mathfrak{S} = \big\{\{z_0, z_1\}\big\}$. Here, I provide a simplified proof.

First note that for every $z_{k'} \in \{z_0, z_1\}$ and every measurable $B_Y \subset \mathcal{Y}$:

$$\begin{aligned}\mathbb{P}\big[X(z_k) = x_l,\ \ Y(x_l) \in B_Y,\ \ \forall z_k \in \{z_0, z_1\}\big] &\leq \mathbb{P}\big[X(z_{k'}) = x_l,\ \ Y(x_l) \in B_Y\big] \\ &= \mathbb{P}[X = x_l, Y \in B_Y | Z = z_{k'}] \\ &= \int_{B_Y} f_{Y,X|Z=z_k}(y, x_l)dy\end{aligned}$$

Thus, for every measurable $B_Y \subset \mathcal{Y}$:

$$\mathbb{P}\big[X(z_k) = x_l,\ \ Y(x_l) \in B_Y,\ \ \forall z_k \in \{z_0, z_1\}\big] \leq \int_{B_Y} f_{Y,X|Z=z_k}(y, x_l)dy$$

Now set $B_0 = \{y \in \mathcal{Y} : f_{Y,X|Z=z_0}(y, x_l) \leq f_{Y,X|Z=z_1}(y, x_l)\}$ and $B_1 = \mathcal{Y} \setminus B_0$ and

note that:

$$\begin{aligned}\mathbb{P}\big[X(z_k)=x_l,\ \ \forall z_k\in\{z_0,z_1\}\big] =&\mathbb{P}\big[X(z_k)=x_l,\ \ Y(x_l)\in B_0,\ \ \forall z_k\in\{z_0,z_1\}\big]\\ &+\mathbb{P}\big[X(z_k)=x_l,\ \ Y(x_l)\in B_1,\ \ \forall z_k\in\{z_0,z_1\}\big]\\ \leq&\int_{B_0} f_{Y,X|Z=z_0}(y,x_l)dy+\int_{B_1} f_{Y,X|Z=z_1}(y,x_l)dy \qquad (21)\\ =&\int_{\mathcal{Y}}\min_{z_k\in\{z_0,z_1\}}\big\{f_{Y,X|Z=z_k}(y,x_l)\big\}dy\end{aligned}$$

2$\Longrightarrow$1:

Let $\boldsymbol{p}$ be a distribution over instrument response types such that:

$$\begin{aligned}A_Z\boldsymbol{p}&\leq \boldsymbol{r}_Z\\ A_X\boldsymbol{p}&=\boldsymbol{r}_X\\ A_Y\boldsymbol{p}&\leq\boldsymbol{\Psi}\end{aligned}$$

The goal is to define a distribution over response types that satisfies equation 3 and proposition 2.1. To this end, it is necessary to specify potential treatments and outcomes for every possible response types, including potential outcomes for types that never realize certain treatments and thus never realize the potential outcome associated to such treatment.

To simplify exposition, let $\tilde{p}_{x_l}$ denote the quantity $p\big(AT_{x_l}\big)$.

Let $t\in\Delta(\mathcal{T})$ be a response type. Define the set of treatments that response type $t=(t^Z,t^X)$ realizes as $R(t)=\{x_l\in\mathcal{X}:(t^Z)_{k'}=x_l \text{ for some } k'=1,2\}$. Note that by construction $|R(t)|\leq 2$

Let λ be an arbitrary fixed distribution over $\mathcal{Y}$. Then, define $g':\Delta(\mathcal{T})\longrightarrow\mathbb{R}$ as follows:

$$g'(t^Z, t^X) = \begin{cases} \Big[\frac{1}{\Psi_{x_l}}\psi_{x_l,\mathcal{Z}}\big((t^X)_l\big)\Big]\Big[\prod_{l':x_{l'}\notin R(t)}\lambda\big((t^X)_{l'}\big)\Big], & \text{if } R(t)=\{x_l\} \\ & \text{and } \Psi_{x_l}>0 \\ \Big[\prod_{l,k:x_l=(t^Z)_k}\frac{1}{P_{x_l|z_k}-\tilde{p}_{x_l}}\xi_{x_l}\big((t^X)_l\big)\Big]\Big[\prod_{l':x_{l'}\notin R(t)}\lambda\big((t^X)_{l'}\big)\Big], & \text{if } |R(t)|=2 \\ & \text{and } P_{x_l|z_k}-\tilde{p}_{x_l}>0 \\ 0, & \text{otherwise} \end{cases}$$

Where $P_{x_l|z_k} = \mathbb{P}[X = x_l|Z = z_k]$ and $\xi_{x_l}\big((t^X)_l\big) = f_{Y,X|Z=z_k}((t^X)_l, x_l) - \frac{\tilde{p}_{x_l}}{\Psi_{x_l}}\psi_{x_l}\big((t^X)_l\big)$ if $\psi_{x_l} > 0$ and $f_{Y,X|Z=z_k}((t^X)_l, x_l)$ otherwise.

Integrating the function g' allows to construct the desired distribution over response types. To this end, recall the following result from measure theory:

Lemma A.1. *Any measurable set of response types $B_{\mathcal{T}} \in \mathcal{B}_{\mathcal{T}}$ can be partitioned into slices as follows:*

$$B_{\mathcal{T}} = \bigcup_{t^Z\in\mathcal{T}_Z}\{t^Z\}\times B_{T_X}^{t^Z}$$

Where $B_{T_X}^{t^Z} \in \mathcal{Y}^L$ is potentially empty and measurable with respect to the product measure over the product of the L spaces that correspond to the potential outcomes at each different treatment value $x_l \in \mathcal{X}$.

Moreover, each individual set $= B_{Y_l}^{t^Z}$ for $l = 0, 1..., L-1$ is measurable with respect to μ_Y.

Then, using the lemma, define:

$$g\Big(B_{\mathcal{T}}\Big) = \sum_{t^Z\in\mathcal{T}_Z}\Big[p(t^Z)\int\limits_{B_{T_X}^{t^Z}} g'(t^Z, t^X)d\mu_{T^X}\Big]$$

Where μ_{T^X} denotes the product measure $\mu_{T^X} = \mu_Y \otimes \mu_Y \otimes ... \otimes \mu_Y$ over the space of potential treatment-responses.

It remains to prove that $g\Big(B_{\mathcal{T}}\Big)$ is a valid distribution over reponse types, this is, it is nonnegative, adds up to 1, satisfies countable additivity, and is consistent with the observed joint distribution of outcomes, treatments and instruments..

When $|R(t)| = 1$, the function $g'(t^Z, t^X)$ is the product of nonnegative functions, therefore, it must be nonnegative itself. When $|R(t)| = 2$ it is either the product of nonnegative functions, or zero. Therefore, it must be nonnegative. From here, it follows that $g\Big(B_{\mathcal{T}}\Big)$ is nonnegative for every measurable $B_{\mathcal{T}} \in \mathcal{B}_{\mathcal{T}}$.

Moreover, note that:

$$\int_{\mathcal{Y}} \frac{p(t^Z)}{\Psi_{x_l,\mathcal{Z}}} \psi_{x_l,\mathcal{Z}}\big(y\big) d\mu_Y = p(t^Z) \tag{22}$$

$$\int_{\mathcal{Y}} \Big[\frac{p(t^Z)}{P_{x_l|z_k} - \tilde{p}_{x_l}} \big(f_{Y,X|Z=z_k}(y, x_l) - \frac{\tilde{p}_{x_l}}{\Psi_{x_l,\mathcal{Z}}} \psi_{x_l,\mathcal{Z}}\big((y)\big)\Big] = p(t^Z)$$

From which follows that:

$$\begin{aligned} g(\mathcal{T}) &= \sum_{t^Z \in \mathcal{T}_Z} \Big[p(t^Z) \int_{\mathcal{T}_X} g'(t^Z, t^X) d\mu_{T^X}\Big] \\ &= \sum_{t^Z \in \mathcal{T}_Z} \Big[p(t^Z)\Big] \\ &= 1 \end{aligned}$$

Countable additivity follows mechanically. Let $\{B^m_{\mathcal{T}}\}_{m \in M}$ be a countable collection of disjoint subsets of $\mathcal{T}$ indexed by $m \in M$. Applying lemma A.1.3, each sets $B^m_{\mathcal{T}}$ can be sliced as follows:

$$B^m_{\mathcal{T}} = \bigcup_{t^Z \in \mathcal{T}_Z} \{t^Z\} \times B^{t^Z,m}_{T_X}$$

Since $B^m \cap B^{m'} = \emptyset$ for all $m, m' \in M$ such that $m \neq m'$ then its must be that $B^{t^Z,m}_{T_X} \cap B^{t^Z,m'}_{T_X} = \emptyset$.

Then:

$$
\begin{aligned}
g\Big(\bigcup_{m\in M} B^m_{\mathcal{T}}\Big) &= g\Big(\bigcup_{m\in M}\Big[\bigcup_{t^Z\in\mathcal{T}_Z}\Big\{\{t^Z\}\times B^{t^Z,m}_{T_X}\Big\}\Big]\Big)\\
&= g\Big(\bigcup_{t^Z\in\mathcal{T}_Z}\Big[\bigcup_{m\in M}\Big\{\{t^Z\}\times B^{t^Z,m}_{T_X}\Big\}\Big]\Big)\\
&= g\Big(\bigcup_{t^Z\in\mathcal{T}_Z}\Big[\{t^Z\}\times\Big\{\bigcup_{m\in M} B^{t^Z,m}_{T_X}\Big\}\Big]\Big)\\
&= \sum_{t^Z\in\mathcal{T}_Z} p(t^Z)\int_{\bigcup_{m\in M} B^{t^Z,m}_{T_X}} g'(t^Z,t^X)d\mu_{T_X}\\
&= \sum_{t^Z\in\mathcal{T}_Z}\sum_{m\in M} p(t^Z)\int_{B^{t^Z,m}_{T_X}} g'(t^Z,t^X)d\mu_{T_X}\\
&= \sum_{m\in M}\sum_{t^Z\in\mathcal{T}_Z} p(t^Z)\int_{B^{t^Z,m}_{T_X}} g'(t^Z,t^X)d\mu_{T_X}\\
&= \sum_{m\in M} g(B^m_{\mathcal{T}})
\end{aligned}
$$

The first line is comes from application of lemma A.1.3 to every set $B^m_{\mathcal{T}}$, the second simply inverts the order of the unions, the third is a property of cartesian products, the fourth simply applies the definition of g and takes advantage of the fact that the set is already sliced, the fifth line uses countable additivity of the integral, the sixth changes the order of the sums, and the last one uses the definition of g but now applied to each set B^m_Y separately. This proves countable additivity.

It only remains to verify consistency with observed probabilities, let $B^{t^Z}_{T_X} = \mathcal{Y}\times$ $....\times B_Y\times ...\times\mathcal{Y}$ be the set obtained by taking the product of the outcome's support $L-1$ times and letting the $l'-th$ element of the product coincide with B_Y instead of $\mathcal{Y}$. Then:

$$g\Big(\big\{(t^Z,t^X)\in\mathcal{T}:(t^Z)_k=x_{l'},(t^X)_l\in B_Y\big\}\Big)=\sum_{t^Z:(t^Z)_k=x_l}\Big[p(t^Z)\int\limits_{B^{t^Z}_{T_X}}g'(t^Z,t^X)d\mu_{T^X}\Big]$$

$$=\frac{\tilde{p}_{x_l}}{\Psi_{x_l,\mathcal{Z}}}\int\limits_{B_Y}\psi_{x_l,\mathcal{Z}}\big(y\big)d\mu_Y$$

$$+\frac{P_{x_l|z_k}-\tilde{p}_{x_l}}{P_{x_l|z_k}-\tilde{p}_{x_l}}\int\limits_{B_Y}\Big(f_{Y,X|Z=z_k}(y,x_l)-\frac{\tilde{p}_{x_l}}{\Psi_{x_l,\mathcal{Z}}}\psi_{x_l,\mathcal{Z}}(y)\Big)d\mu_Y$$

$$=\int\limits_{B_Y}f_{Y,X|Z=z_k}(y,x_l)d\mu_Y$$

The first equality is just the definition of g, the second uses the fact that at instrument value z_k there are only two instrument-response types that realize x_l (always-x_l takers, and $\{z_k\}$-IFF-x_l-takers) combined with the definition of g'. The last line simply re-arranges terms.

This completes the proof.

□

A.2 Proofs for subsection 3.2

I first state two lemmas that I will use to prove theorem 3.2.

Lemma A.2. *Let $t=(t^Z,t^X)$ be a response type. For every $\mathfrak{S}\subset\mathcal{P}(\mathcal{Z})$, any $C\in\mathfrak{C}(\mathfrak{S})$, and any measurable $B_Y\subset\mathcal{Y}$:*

$$\mathbb{P}[t^Z\in\mathcal{T}_{\mathfrak{S},x_l},t^X\in B_Y]\leq\int\limits_{B_Y}\psi_C(y,x_l)dy$$

Lemma A.3. *Let $\mathcal{T}_{\mathfrak{S},x_l}$ be a collection of $\mathfrak{S}$-sufficient-x_l-takers. Let $\mathfrak{C}=\{C_1,...,C_n\}$ be a collection of function indices such that for every $C\in\mathfrak{C}$ and every measurable set $B_Y\subset\mathcal{Y}$, the function $\psi_C(y)$ is such that:*

$$\mathbb{P}[t=(t^Z,t^X),\ \ t^Z\in\mathcal{T}_{\mathfrak{S},x_l},t^X(x_l)\in B_Y]\leq\int\limits_{B_Y}\psi_C(y)dy$$

Then:

$$\mathbb{P}[t = (t^Z, t^X),\ \ t^Z \in \mathcal{T}_{\mathfrak{S},x_l}] \leq \int_{\mathcal{Y}} \min_{C \in \mathfrak{C}} \{\psi_C(y)\} dy$$

Proof of lemma A.2.

First note that for any $\mathfrak{S} \subset \mathcal{P}(\mathcal{Z})$, a function g_1 that maps $\mathcal{T}_{\mathfrak{S},x_l}$ to $\mathfrak{S}$ exists. This is trivial since $t^Z \in \mathcal{T}_{\mathfrak{S},x_l}$ implies that t^Z belongs to set of S-sufficient-x_l-takers for some $S \in \mathfrak{S}$. Setting $g_1(t^Z)$ equal to any such S proves that such a function exists.

Equivalently, for any $C \in \mathfrak{C}(\mathfrak{S})$ a function g_2 that maps $\mathfrak{S}$ to C exists. This is also trivial since $S \in \mathfrak{S}$ and $C \in \mathfrak{C}(\mathfrak{S})$ imply that some $z_k \in C$ exists such that $z_k \in S$. Setting $g_2(S)$ equal to any such z_k proves that such function exists.

Fix two arbitrary functions g_1 and g_2 that obey the previous constructions. Then:

$$\begin{aligned}
\mathbb{P}[t^Z \in \mathcal{T}_{\mathfrak{S},x_l}, t^X \in B_Y] &= \sum_{t': t'^Z \in \mathcal{T}_{\mathfrak{S},x_l}, t'^X(x_l) \in B_Y} \mathbb{P}[t = t'] \\
&= \sum_{S \in \mathfrak{S}} \Big(\sum_{t': g_1(t'^z) = S, t'^X(x_l) \in B_Y} \mathbb{P}[t = t'] \Big) \\
&= \sum_{z_k \in C} \Big[\sum_{S: g_2(S) = z_k} \Big(\sum_{t': g_1(t'^z) = S, t'^X(x_l) \in B_Y} \mathbb{P}[t = t'] \Big) \Big] \\
&= \sum_{z_k \in C} \Big(\sum_{t': g_2\left(g_1(t'^z)\right) = S, t'^X(x_l) \in B_Y} \mathbb{P}[t = t'] \Big) \\
&\leq \sum_{z_k \in C} \sum_{t': t'^Z(z_k) = k_l, t'^X(x_l) \in B_Y} \mathbb{P}[t = t'] \\
&= \sum_{z_k \in C} \int_{B_Y} f_{Y,X|Z=z_k}(y, x_l)
\end{aligned}$$

The first line simply sums over all response types that belong to $\mathcal{T}_{\mathfrak{S},x_l}$. The second uses the fact that g_1 maps response types t^Z in $\mathcal{T}_{\mathfrak{S},x_l}$ to sets S in $\mathfrak{S}$ such that t^Z belongs to the set of S-sufficient-x_l-takers. The third line uses the fact that g_2 maps sets S in $\mathfrak{S}$ to elements x_k of C such that z_k belongs to S. The fourth line simply composes the functions g_1 and g_2. The fifth line notes that if t^Z belongs to the S-sufficient-x_l-takers and $z_k \in S$, then t^Z belongs to the set of instrument-response types such that $t^Z(z_k) = x_l$. The last line notes that the set of all types such that $t^Z(z_k) = x_l$ and $t^X(x_l) \in B_Y$ must correspond to the integral of $f_{Y,X|Z=z_k}(y, x_l)$ over B_Y. □

Proof of lemma A.3.

First define the set:

$$B_1 = \{y \in \mathcal{Y} : f_{C_1}(y) \leq f_{C_{j'}}, \ \ \forall C_{j'} \in \mathfrak{C}\}$$

Then, inductively define the sets:

$$\begin{aligned} B_j = \big\{y \in \mathcal{Y} : f_{C_j}(y) < f_{C_{j'}} \ \ \forall C_j \in \{C_1, ..., C_j\}, \\ \textit{and } f_{C_j}(y) \leq f_{C_{j'}} \ \ \forall C_j \in \{C_{j+1}, ..., C_n\}\big\} \end{aligned}$$

Note that for every set B_j with $j = 1, ..., n$:

$$\begin{aligned} \mathbb{P}[t = (t^Z, t^X), \ \ t^Z \in \mathcal{T}_{\mathfrak{S},x_l}, t^X(x_l) \in B_Y] &\leq \int_{B_j} \psi_{C_j}(y)dy \\ &= \int_{B_j} \min_{C \in \mathfrak{C}}\{\psi_C(y)\}dy \end{aligned} \tag{23}$$

Observe that the sets B_j with $j = 1, ..., n$ comprise a partition of $\mathcal{Y}$, hence:

$$\begin{aligned} \mathbb{P}[t = (t^Z, t^X), \ \ t^Z \in \mathcal{T}_{\mathfrak{S},x_l}] &= \sum_{j=1}^{n} \mathbb{P}[t = (t^Z, t^X), \ \ t^Z \in \mathcal{T}_{\mathfrak{S},x_l}, t^X(x_l) \in B_Y] \\ &\leq \sum_{j=1}^{n} \int_{B_j} \min_{C \in \mathfrak{C}}\{\psi_C(y)\}dy \\ &= \int_{\mathcal{Y}} \min_{C \in \mathfrak{C}}\{\psi_C(y)\}dy \end{aligned}$$

Where the second line comes from 23. The first and third line use the fact that the sets B_j with $j = 1, ..., n$ comprise a partition of $\mathcal{Y}$. This completes the proof. □

Proof of theorem 3.2.

The proof follows immediately from lemmas A.2 and A.3. □

B Proofs for section

B.1 Estimation algorithm 1

1. Partition the sample in two splits: S_1 and S_2 of size n_1 and n_2 respectively.

2. Using the first sample, train a nonparametric classifier $\eta^*_{n_1}(y, x)$ as follows:

   (a) Define:

   $$\phi_{z_k,n}(y, x_l) := P^{-1}_{k,n_1}\Big[\frac{1}{n_1 h^d_n}\sum_{i=1}^{n_1}\mathbf{1}\{X_i = x_l, Z_i = z_k\}K_n\big(y - Y_i\big)\Big]$$

   Where $n_{k,1} = |\{i \in S_1 : Z_i = z_k\}|$ and $P_{k,n_1} = n_{k,1}/n_1$.

   (b) Then, define:

   $$\Delta_{n_1}(y, x_l) := \phi_{z_1,n}(y, x_l) - \phi_{z_0,n}(y, x_l)$$

   (c) And:

   $$\eta^*_{n_1}(y, x_l) := \mathbf{1}\{\Delta_{n_1}(y, x_l) \geq 0\}.$$

3. In the second sample, estimate the probabilities $\mathbb{P}[Z = z_k]$ for $k = 0, 1$. Denote the estimators as P_{k,n_2}.

4. In the second sample, evaluate the weighted empirical loss:

   $$L_{n_2}(\eta^*_{n_1}, x_l) := \frac{1}{n_2}\sum_{i=1}^{n_2}\Big(\frac{\mathbf{1}\{\eta^*_{n_1}(Y_i, X_i) = 1, Z_i = z_0, X_i = x_l\}}{P_{0,n_2}} + \frac{\mathbf{1}\{\eta^*_{n_1}(Y_i, X_i) = 0, Z_i = z_1, X_i = x_l\}}{P_{1,n_2}}\Big)$$

B.2 Proofs for subsection 5.1

Decompose:

$$\sqrt{n_2}\big(L_{n_2}(\eta^*_{n_1}, x_l) - \Psi_{x_l}\big) = \underbrace{\sqrt{n_2}\big(L_{n_2}(\eta^*_{n_1}, x_l) - L(\eta^*_{n_1}, x_l)\big)}_{:=A_{1,n}(x_l)} + \underbrace{\sqrt{n_2}\big(L(\eta^*_{n_1}, x_l) - L(\eta^*, x_l)\big)}_{:=A_{2,n}(x_l)}$$

The first term corresponds to how well the empirical loss evaluated in the subsample S_2 approximates the expected loss of the classifier $\eta^*_{n_1}$, the second term corresponds to the loss difference between the estimated classifier $\eta^*_{n_1}$ and the optimal unfeasible classifier η^*.

Consequently:

$$n_2 Var\Big(L_{n_2}(\eta^*_{n_1}, x_l) - \Psi_{x_l}\Big) = \underbrace{n_2 Var\Big(L_{n_2}(\eta^*_{n_1}, x_l) - L(\eta^*_{n_1}, x_l)\Big)}_{:=V_1} + \underbrace{n_2 Var\Big(L(\eta^*_{n_1}, x_l) - L(\eta^*, x_l)\Big)}_{:=V_2} + \underbrace{2n_2 Cov\Big(L_{n_2}(\eta^*_{n_1}, x_l) - L(\eta^*_{n_1}, x_l), L(\eta^*_{n_1}, x_l) - L(\eta^*, x_l)\Big)}_{:=V_3}$$

Propositions B.1 shows that, under assumptions 5.1 and 5.2, the classifier trained in subsample S_1 is sufficiently good so that its expected missclassification loss (and its variance) are small relative to the expected loss of the optimal classifier. Moreover, sample splitting ensures that $V_3 = 0$. Thus, the asymptotic behavior of the estimator is governed by the term $A_{1,n}(x_l)$ with variance V_1.

Proposition B.1.

1. $A_{2,n}(x_l) = o_p(1)$.
2. $Cov(A_{1,n}(x_l), A_{2,n}(x_l)) = 0$.
3. $Var(A_{2,n}(x_l)) = o(1)$.

Proof of proposition 5.1.

$$
\begin{aligned}
L(\eta^*, x_l) =& \frac{\mathbb{P}[\eta^*(Y,X)=0, Z=z_1, X=x_l]}{\mathbb{P}[Z=z_1]} + \frac{\mathbb{P}[\eta^*(Y,X)=1, Z=z_0, X=x_l]}{\mathbb{P}[Z=z_0]} \\
=& \mathbb{P}[X=x_l]\Big(\frac{\mathbb{P}[\eta^*(Y,X)=0, Z=z_1|X=x_l]}{\mathbb{P}[Z=z_1]} + \frac{\mathbb{P}[\eta^*(Y,X)=1, Z=z_0|X=x_l]}{\mathbb{P}[Z=z_0]}\Big) \\
=& \mathbb{P}[X=x_l]\int_{\mathcal{Y}} \Big(\frac{\mathbb{P}[\eta^*(y,x_l)=0, Z=z_1|Y=y, X=x_l]}{\mathbb{P}[Z=z_1]} \\
& \qquad + \frac{\mathbb{P}[\eta^*(y,x_l)=1, Z=z_0|Y=y, X=x_l]}{\mathbb{P}[Z=z_0]}\Big) f_{Y|X=x_l}(y)dy \\
=& \mathbb{P}[X=x_l]\int_{\mathcal{Y}} \Big(\mathbf{1}\{\eta^*(y,x_l)=0\}\frac{\mathbb{P}[Z=z_1|Y=y, X=x_l]}{\mathbb{P}[Z=z_1]} \\
& \qquad + \mathbf{1}\{\eta^*(y,x_l)=1\}\frac{\mathbb{P}[Z=z_0|Y=y, X=x_l]}{\mathbb{P}[Z=z_0]}\Big) f_{Y|X=x_l}(y)dy \\
=& \mathbb{P}[X=x_l]\int_{\mathcal{Y}} \Big(\mathbf{1}\{\eta^*(y,x_l)=0\}\frac{f(y,x_l,z_1)}{\mathbb{P}[Z=z_1]f(y,x_l)} \\
& \qquad + \mathbf{1}\{\eta^*(y,x_l)=1\}\frac{f(y,x_l,z_0)}{\mathbb{P}[Z=z_0]f(y,x_l)}\Big) f(y,x_l)dy \\
=& \int_{\mathcal{Y}} \Big(\mathbf{1}\{\eta^*(y,x_l)=0\}f_{Y,X|Z=z_1}(y,x_l) + \mathbf{1}\{\eta^*(y,x_l)=1\}f_{Y,X|Z=z_0}(y,x_l)\Big)dy \\
=& \Psi_{x_l}
\end{aligned}
$$

□

Proof of proposition B.1.

Proof of part 1: Consider the difference between the loss of the estimated and target classifier:

$$
\begin{aligned}
A_{2,n}(x_l) =& L(\eta^*_{n_1}, x_l) - L(\eta^*, x_l) \\
=& \int_{\mathcal{Y}} \Big(\mathbf{1}\{\eta^*_{n_1}(y, x_l) = 0\} f_{Y,X|Z=z_1}(y, x_l) + \mathbf{1}\{\eta^*_{n_1}(y, x_l) = 1\} f_{Y,X|Z=z_0}(y, x_l) \Big) \\
& - \Big(\mathbf{1}\{\eta^*(y, x_l) = 0\} f_{Y,X|Z=z_1}(y, x_l) + \mathbf{1}\{\eta^*(y, x_l) = 1\} f_{Y,X|Z=z_0}(y, x_l) \Big) dy \\
=& \int_{\mathcal{Y}} \mathbf{1}\{\eta^*_{n_1}(y, x_l) \neq \eta^*(y, x_l)\} \big| \Delta(y, x_l) \big| dy \qquad (24)
\end{aligned}
$$

Where $\Delta(y, x_l) := f_{Y,X|Z=z_1}(y, x_l) - f_{Y,X|Z=z_0}(y, x_l)$.

Partition the support of the outcome as follows:

$$
\mathcal{Y} = \mathcal{S}_{x_l,n_1} \cup \mathcal{M}_{x_l,n_1} \cup \mathcal{C}_{x_l}
$$

Where:

$$
\begin{aligned}
\mathcal{S}_{x_l,n_1} &= \Big\{ y \in \mathcal{Y} : |\Delta(y, x_l)| > C_0 \sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}} \Big\} \\
\mathcal{M}_{x_l,n_1} &= \Big\{ y \in \mathcal{Y} : 0 < |\Delta(y, x_l)| \leq C_0 \sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}} \Big\} \\
\mathcal{C}_{x_l} &= \Big\{ y \in \mathcal{Y} : |\Delta(y, x_l)| = 0 \Big\}
\end{aligned}
$$

Then, equation 24 becomes:

$$
\begin{aligned}
A_{2,n}(x_l) = & \int_{\mathcal{S}_{x_l,n_1}} \Big(\mathbf{1}\{\eta^*(y, x_l) \neq \eta^*_{n_1}(y, x_l)\} |\Delta(y, x_l)| \Big) dy \\
& + \int_{\mathcal{M}_{x_l,n_1}} \Big(\mathbf{1}\{\eta^*(y, x_l) \neq \eta^*_{n_1}(y, x_l)\} |\Delta(y, x_l)| \Big) dy \\
& + \int_{\mathcal{C}_{x_l}} \Big(\mathbf{1}\{\eta^*(y, x_l) \neq \eta^*_{n_1}(y, x_l)\} |\Delta(y, x_l)| \Big) dy
\end{aligned}
$$

Now I show that $A_{2,n}(x_l) = o_p(n^{-\frac{1}{2}})$. For the first term I use the classic results on rates of uniform convergence for kernel density estimators from Audibert and Tsybakov and Talagrand to prove that, when the subdensities are well separated, misslasification

occurs with low probability. For the second term I invoke the margin condition from assumption 5.1.3 combined with the fact that miss-classification loss near the boundary of the contact set is small. The third term is exactly zero since $|\Delta(y, x_l)| = 0$ for every $y \in \mathcal{C}_{x_l}$.

Consider the term associated with $\mathcal{S}_{x_l,n_1}$:

$$\begin{aligned}
&\int_{\mathcal{S}_{x_l,n_1}} \Big(\mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\}|\Delta(y,x_l)|\Big)dy \\
&\quad\leq \int_{\mathcal{S}_{x_l,n_1}} \mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\}\Big(f_{Y,X|Z=z_1}(y,x_l) + f_{Y,X|Z=z_0}(y,x_l)\Big)dy \\
&\quad\leq B^{-1} \int_{\mathcal{S}_{x_l,n_1}} \mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\} f_Y(y)dy \\
&\quad= B^{-1}\mathbb{P}[\eta^*(Y,x_l) \neq \eta^*_{n_1}(Y,x_l), Y \in \mathcal{S}_{x_l,n_1}|S_1] \qquad (25)
\end{aligned}$$

Where the first line follows from the triangle inequality, the second from the definition of $f_{Y,X|Z=z_k}$ combined with assumption 5.1.1, and the third is a standard property of indicator expectations.

Note that for any $y \in \mathcal{S}_{x_l,n_1}$, $|\Delta(y,x_l)| > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}$. Hence, if $\|\Delta_{n_1}(\cdot,x_l) - \Delta(\cdot,x_l)\|_\infty \leq r_{n_1}$, then $\Delta_{n_1}(y,x_l)$ has the same sign as $\Delta(y,x_l)$ for every $y \in \mathcal{S}_{x_l,n_1}$. Therefore:

$$\mathbb{P}\big[\eta^*(Y,x_l) \neq \eta^*_{n_1}(Y,x_l), Y \in \mathcal{S}_{x_l,n_1} \mid S_1\big] \leq \mathbf{1}\Big\{\|\Delta_{n_1}(\cdot,x_l) - \Delta(\cdot,x_l)\|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}\Big\}.$$

By Lemma B.3, the right-hand side is $o_p(n^{-1})$. Hence

$$\mathbb{P}\big[\eta^*(Y,x_l) \neq \eta^*_{n_1}(Y,x_l), Y \in \mathcal{S}_{x_l,n_1}|S_1\big] = o_p(n^{-1}). \qquad (26)$$

Combining 25 and 26 yields:

$$\int_{\mathcal{S}_{x_l,n_1}} \Big(\mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\}|\Delta(y,x_l)|\Big)dy = o_p(n^{-1})$$

Now consider the term associated with $\mathcal{M}_{x_l,n_1}$. Assumption 5.1.3 implies that the conditions from the first part of lemma B.4 is satisfied for $r_n = C_0\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}$ and

$\mathcal{M}_{x_l}(r_n) = \mathcal{M}_{x_l,n_1}$. Therefore:

$$\int\limits_{\mathcal{M}_{x_l,n_1}} \Big(\mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\}|\Delta(y,x_l)|\Big)dy = o_p(n^{-1/2})$$

Thus, $A_{2,n}(x_l) = o_p(n^{-1/2})$. This completes the proof of part 1.

Proof of part 2:

Lemma B.5 implies that:

$$\begin{aligned} Cov\Big(L_{n_2}(\eta^*_{n_1},x_l), L(\eta^*_{n_1},x_l)\Big) &= \mathbb{E}\Big[L_{n_2}(\eta^*_{n_1},x_l)L(\eta^*_{n_1},x_l)\Big] - \mathbb{E}\Big[L_{n_2}(\eta^*_{n_1},x_l)\Big]\mathbb{E}\Big[L(\eta^*_{n_1},x_l)\Big] \\ &= Var\Big(L(\eta^*_{n_1},x_l)\Big) \end{aligned}$$

Then:

$$\begin{aligned} &Cov\Big(L_{n_2}(\eta^*_{n_1},x_l) - L(\eta^*_{n_1},x_l), L(\eta^*_{n_1},x_l) - L(\eta^*,x_l)\Big) \\ &\quad = Cov\Big(L_{n_2}(\eta^*_{n_1},x_l), L(\eta^*_{n_1},x_l)\Big) - Var\Big(L_{n_2}(\eta^*_{n_1},x_l)\Big) \\ &\quad = 0 \end{aligned}$$

Proof of part 3:

The argument replicates the idea from part 1, but now applied to the variance. In the contact set, any classification decision yields the exact same loss, therefore, even if the classifier is random inside the contact set, the loss associated to such randomness is deterministic. On a shrinking region near the boundary of the contact set, classification mistakes occur with nonnegligible probability. Nonetheless, the difference in loss between the optimal classifier and a random classifier is small and so is the variance of such loss, furthermore, the margin condition implies that such region shrinks at a sufficiently fast rate. Finally, far from the closure of the contact set, missclassification events are unlikely.

First note that:

$$
\begin{aligned}
Var(A_{2,n}(x_l)) =& Var\Big(\int_{\mathcal{Y}} \mathbf{1}\{\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\}\big|\Delta(y,x_l)\big|dy\Big) \\
=& Cov\Big(\int_{\mathcal{Y}} \mathbf{1}\{\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\}\big|\Delta(y,x_l)\big|dy, \\
& \qquad \int_{\mathcal{Y}} \mathbf{1}\{\eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\}\big|\Delta(y',x_l)\big|dy'\Big) \\
=& \int_{\mathcal{Y}}\int_{\mathcal{Y}} Cov\big(\mathbf{1}\{\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\}\big|\Delta(y,x_l)\big|, \\
& \qquad \mathbf{1}\{\eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\}\big|\Delta(y',x_l)\big|\big)dydy' \\
=& \int_{\mathcal{Y}}\int_{\mathcal{Y}} \Big(\mathbb{E}\big[\mathbf{1}\{\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\}\mathbf{1}\{\eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\}\big] \\
& \qquad - \mathbb{E}\big[\mathbf{1}\{\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\}\big]\mathbb{E}\big[\mathbf{1}\{\eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\}\big]\Big) \\
& \qquad \times \big|\Delta(y,x_l)\big|\big|\Delta(y',x_l)\big|dydy' \\
=& \int_{\mathcal{Y}}\int_{\mathcal{Y}} \Big(\mathbb{P}\big[\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l), \eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\big] \\
& \qquad - \mathbb{P}\big[\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\big]\mathbb{P}\big[\eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\big]\Big) \\
& \qquad \times \big|\Delta(y,x_l)\big|\big|\Delta(y',x_l)\big|dydy' \\
=& \int_{\mathcal{Y}}\int_{\mathcal{Y}} C_p(y,x_l)dydy' \\
\leq& \int_{\mathcal{Y}}\int_{\mathcal{Y}} |C_p(y,x_l)|dydy' \qquad (27)
\end{aligned}
$$

Where:

$$
\begin{aligned}
C_P(y,y',x_l) :=& \mathbb{P}\big[\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l), \eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\big] \\
& - \mathbb{P}\big[\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\big]\mathbb{P}\big[\eta^*_{n_1}(y',x_l) \neq \eta^*(y',x_l)\big]\big|\Delta(y,x_l)\big|\big|\Delta(y',x_l)\big|
\end{aligned}
$$

Now decompose 27 as follows:

$$Var(A_{2,n}(x_l)) \leq \int_{\mathcal{S}_{x_l,n_1}} \int_{\mathcal{Y}} |C_P(y,y',x_l)|dydy' + \int_{\mathcal{M}_{x_l,n_1}} \int_{\mathcal{M}_{x_l,n_1}} |C_P(y,y',x_l)|dydy' + \int_{\mathcal{M}_{x_l,n_1}} \int_{\mathcal{C}_{x_l}} |C_P(y,y',x_l)|dydy' + \int_{\mathcal{C}_{x_l}} \int_{\mathcal{C}_{x_l}} |C_P(y,y',x_l)|dydy' + \int_{\mathcal{M}_{x_l}} \int_{\mathcal{S}_{x_l}} |C_P(y,y',x_l)|dydy'$$

Where the sets $\mathcal{S}_{x_l,n_1}$, $\mathcal{M}_{x_l,n_1}$ and $\mathcal{C}_{x_l,n_1}$ are as defined in part 1 of the proof.

All the terms that involve integrals over $\mathcal{C}_{x_l,n_1}$ are exactly zero, thus:

$$Var(A_{2,n}(x_l)) \leq \int_{\mathcal{S}_{x_l,n_1}} \int_{\mathcal{Y}_{x_l,n_1}} |C_P(y,y',x_l)|dydy' + \int_{\mathcal{M}_{x_l,n_1}} \int_{\mathcal{M}_{x_l,n_1}} |C_P(y,y',x_l)|dydy' + \int_{\mathcal{M}_{x_l}} \int_{\mathcal{S}_{x_l}} |C_P(y,y',x_l)|dydy'$$

Let $M(y,x_l)$ denote the event $\{\eta^*_{n_1}(y,x_l) \neq \eta^*(y,x_l)\}$. Then:

$$\begin{aligned}
&\int_{\mathcal{Y}_{x_l,n_1}} \int_{\mathcal{Y}_{x_l,n_1}} |C_P(y,y',x_l)|dydy' \\
&\qquad \leq \int_{\mathcal{S}_{x_l,n_1}} \int_{\mathcal{S}_{x_l,n_1}} \mathbb{P}[M(y,x_l)]\big|\Delta(y,x_l)\big|\big|\Delta(y',x_l)\big|dydy' \\
&\qquad \leq \Big(\int_{\mathcal{S}_{x_l,n_1}} \big|\Delta(y',x_l)\big|dy'\Big)\Big(\int_{\mathcal{Y}_{x_l,n_1}} \mathbb{P}[M(y,x_l)]\big|\Delta(y,x_l)\big|dy\Big)
\end{aligned}$$

Note that $\big|\Delta(y',x_l)\big| \leq f_Y, X|Z = z_0(y,x_l) + f_{Y,X|Z=z_1}(y,x_l)$ and that both $f_{Y,X|Z=z_0}(y,x_l)$ and $f_{Y,X|Z=z_1}(y,x_l)$ integrate to less than 1. Furthermore, lemma B.3 implies that $\mathbb{P}[M(y,x_l)] = o_{n_1^{-1}}$ for every $y \in \mathcal{S}_{x_l,n_1}$. Thus:

$$\int_{\mathcal{Y}_{x_l,n_1}} \int_{\mathcal{Y}_{x_l,n_1}} |C_P(y,y',x_l)|dydy' = o(n^{-1})$$

A similar argument shows that:

$$\int_{\mathcal{M}_{x_l}} \int_{\mathcal{S}_{x_l}} |C_P(y, y', x_l)| dy dy' = o(n^{-1})$$

Observe that the term $\mathbb{P}\big[\eta^*_{n_1}(y, x_l) \neq \eta^*(y, x_l), \eta^*_{n_1}(y', x_l) \neq \eta^*(y', x_l)\big] - \mathbb{P}\big[\eta^*_{n_1}(y, x_l) \neq \eta^*(y, x_l)\big]\ \mathbb{P}\big[\eta^*_{n_1}(y', x_l) \neq \eta^*(y', x_l)\big]$ is the covariance of the random variables $\upsilon_{y,x_l} := M(y, x_l)$ and $\upsilon_{y',x_l} := M(y', x_l)$ that indicate missclassification events at fixed values of Y and X with respect to the distribution of subsample S_1.

Thus, Cauchy-Schwarz's inequality implies:

$$\begin{aligned} |Cov(\upsilon_{y,x_l}, \upsilon_{y',x_l})| &\leq \sqrt{Var(\upsilon_{y,x_l}) Var(\upsilon_{y',x_l})} \\ &= \sqrt{\big(\mathbb{P}[M(y, x_l)] - \mathbb{P}[M(y, x_l)]^2\big)\big(\mathbb{P}[M(y', x_l)] - \mathbb{P}[M(y', x_l)]^2\big)} \\ &\leq \sqrt{\mathbb{P}[M(y, x_l)]\mathbb{P}[M(y', x_l)]} \end{aligned} \tag{28}$$

Where the first inequality follows from 28, the second from the fact that $\mathbb{P}[M(y', x_l)]$ $\big|\Delta(y, x_l)\big|\big|\Delta(y', x_l)\big|$ is uniformly bounded, and the last from the fact that for any $y \in \mathcal{S}_{x_l,n_1}$, $|\Delta(y, x_l)| > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}$. Hence, if $\|\Delta_{n_1}(\cdot, x_l) - \Delta(\cdot, x_l)\|_\infty \leq r_{n_1}$, then $\Delta_{n_1}(y, x_l)$ has the same sign as $\Delta(y, x_l)$ for every $y \in \mathcal{S}_{x_l,n_1}$. The final identity follows from lemma B.3.

Now consider the term associated with $\mathcal{M}_{x_l,n_1}$:

$$\int_{\mathcal{M}_{x_l,n_1}} \int_{\mathcal{M}_{x_l,n_1}} |C_P(y, y', x_l)| dy dy' \leq \int_{\mathcal{M}_{x_l,n_1}} \int_{\mathcal{M}_{x_l,n_1}} \sqrt{\mathbb{P}[M(y, x_l)]\mathbb{P}[M(y, x_l)]}\big|\Delta(y, x_l)\big|\big|\Delta(y', x_l)\big| dy dy'$$

But, part 2 of lemma B.4 implies that the right hand side is $o(n^{-1})$. This completes the proof.

□

Proof of theorem 5.1.

Proposition B.1 implies that:

$$\sqrt{n_2}\big(\boldsymbol{\theta}_n - \boldsymbol{\mu}_\theta\big) = \sqrt{n_2}\big(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\mu}_\theta\big) + o(1)$$

Where $\tilde{\boldsymbol{\theta}}_n = (\boldsymbol{P}^T_{\boldsymbol{\mathcal{X}_0,n}}, \boldsymbol{P}^T_{\boldsymbol{\mathcal{X}_1,n}}, n_2^{-1/2}\boldsymbol{A}^T_{1,n})$ and $\boldsymbol{A}^T_{1,n} = (n_2^{-1/2}A_{1,n}(x_0), ..., A_{1,n}(x_{L'-1}))$. Furthermore, lemma B.5 and proposition B.1 imply that $\mathbb{E}[\tilde{\boldsymbol{\theta}}_n] - \mu_\theta = o(n^{-1})$.

Define the random variable $\xi(Y, X, Z; z_k, x_l) = \mathbf{1}\{\eta^*_{n_1}(Y, X, Z; z_k, x_l) \neq k, X = x_l, Z = z_k\}$ and note that:

$$A_{1,n}(x_l) = \frac{\sqrt{n_2}}{n_2} \sum_{i=1}^{n_2} \left(\frac{n_2}{n_{2,0}} \xi(Y_i, X_i, Z_i; z_0, x_l) + \frac{n_2}{n_{2,1}} \xi(Y_i, X_i, Z_i; z_1, x_l) - L(\eta^*_{n_1}, x_l) \right) \tag{29}$$

Let the vector $\boldsymbol{P}_{n_2}$ and the function $\boldsymbol{g}$ be as defined in lemmas B.6 and B.7 respectively. Equation 29 implies that $\tilde{\boldsymbol{\theta}}_n$ is a subvector of $\boldsymbol{g}(\boldsymbol{P}_{n_2})$. Formally, if J denotes the set of indices associated with all the entries of $\boldsymbol{g}(\boldsymbol{P}_{n_2})$ that comprise $\tilde{\boldsymbol{\theta}}_n$, and e_j is the $j-th$ canonical vector in $\mathbb{R}^{3L}$. Then, the selector matrix $S_{\boldsymbol{\theta}}$ is the matrix that stacks the vectors e_j for every $j \in J$ so that $\tilde{\boldsymbol{\theta}}_n = S_{\boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{P}_{n_2})$.

By construction, every component of $(\boldsymbol{P}^T_{\boldsymbol{\mathcal{X}_0,n}}, \boldsymbol{P}^T_{\boldsymbol{\mathcal{X}_1,n}})$ is linearly independent from the rest asymptotically, and all the components of $\boldsymbol{A}_{1,n}$ are linearly independent from each other with probability approaching one. Moreover, assumption 5.3.2 guarantees that every component of $A_{1,n}$ of $\boldsymbol{A}_{1,n}$ is linearly independent from $P_{x_l|z_0,n_2}$ and $P_{x_l|z_0,n_2}$ with probability approaching one. Thus, all the elements of $\tilde{\boldsymbol{\theta}}_n$ are asymptotically linearly independent with probability approaching one.

Hence, with probability approaching one, the matrix $S_{\boldsymbol{\theta}} \boldsymbol{V}_{S_1,n} S^T_{\boldsymbol{\theta}}$ is symmetric positive-definite and, for every nonzero $\boldsymbol{c} \in \mathbb{R}^m$, a fixed constant $\epsilon_{\boldsymbol{c}}$ that depends on $\boldsymbol{c}$ exists such that the probability $\mathbb{P}[\boldsymbol{c}^T S_{\boldsymbol{\theta}} \boldsymbol{V}_{S_1,n} S^T_{\boldsymbol{\theta}} \boldsymbol{c} > \epsilon_{\boldsymbol{c}}]$ tends to one.

Then, lemma B.7 implies that:

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left[\frac{\boldsymbol{c}^T \sqrt{n_2} \left(\boldsymbol{\theta}_n - \boldsymbol{\mu}_\theta \right)}{\sqrt{\boldsymbol{c}^T S_{\boldsymbol{\theta}} \boldsymbol{V}_{S_1,n} S^T_{\boldsymbol{\theta}} \boldsymbol{c}}} \leq t \middle| S_1 \right] - \Phi(t) \right| \xrightarrow{p} 0, \ \forall \boldsymbol{c} \in \mathbb{R}^m, \boldsymbol{c} \neq \boldsymbol{0},$$

and the Cramér-Wold device yields the desired result.

□

Lemma B.1 (KDE bias)**.** *Let* $\tilde{\Delta}_{n_1}(Y, x_l) = \frac{1}{n_1 h_n^d} \sum_{i=1}^{n_1} H_n(Y_i, X_i, Z_i; x_l, y)$*, then:*

$$\sup_{y \in \mathcal{Y}} \left| \left| \mathbb{E}[\tilde{\Delta}_{n_1}(Y, x_l)] - \Delta(y, x_l) \right| \right| = O(h_n^2)$$

Lemma B.2. *For any $r_n > 0$ such that $\frac{n_1 r_n^2}{\log(n_1)} \longrightarrow \infty$ and $r_n \longrightarrow 0$:*

$$\mathbb{P}\Big(\Big|\Big|R_n(y, x_l)\Big|\Big|_\infty \geq r_n\Big) = o(n_1^{-1})$$

Lemma B.3. *Let K be a kernel density function satisfying assumption 5.2.1, h_n a bandwidth sequence satisfying assumption 5.2.2 and suppose that $f_{Y|X=x_l,Z=z_k}$ satisfies assumption 5.1.2. Then, a constant $C_0 > 0$ exists such that:*

$$\mathbb{P}\Big[\Big|\Big|\Delta_{n_1}(y, x_l) - \Delta(y, x_l)\ \Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}\Big] = o(n_1^{-1})$$

Lemma B.4 (Margin missclassification loss)**.**
Let r_n be a sequence such that $r_n \longrightarrow 0$ and $r_n^{1+\alpha}\sqrt{n_1} \longrightarrow 0$, and let $\mathcal{M}_{x_l}(r_n)$ be the set:

$$\mathcal{M}_{x_l}(r_n) = \{y \in \mathcal{Y} : 0 < |\Delta(y, x_l)| \leq r_n\}$$

Then:

$$\int\limits_{\mathcal{M}_{x_l}(r_n)} \mathbf{1}\{M(y, x_l)\}|\Delta(y, x_l)|dy = o(n^{-1/2})$$

And:

$$\int\limits_{\mathcal{M}_{x_l}(r_n)} \int\limits_{\mathcal{M}_{x_l}(r_n)} \sqrt{\mathbb{P}[M(y, x_l)]\mathbb{P}[M(y', x_l)]}\big|\Delta(y, x_l)\big|\big|\Delta(y', x_l)\big|dydy' = o(n^{-1})$$

Where $M(y, x_l)$ denotes the event $\eta^*(y, x_l) \neq \eta^*_{n_1}(y, x_l)$

Lemma B.5 (The expected empirical loss of $\eta^*_{n_1}$ is its theoretical loss)**.**

$$\mathbb{E}\Big[L_{n_2}(\eta^*_{n_1}, x_l)|S_1\Big] = L(\eta^*_{n_1}, x_l)$$

Lemma B.6.

For $i \in S_2$, let $\boldsymbol{\pi}_{i,n}$ denote the random vector:

$$\begin{aligned}\boldsymbol{\pi}_{i,n} =&(\mathbf{1}\{Z_i = z_0\}, \mathbf{1}\{Z_i = z_1\},\\ &\mathbf{1}\{Z_i = z_0, X_i = x_0\}), ..., \mathbf{1}\{Z_i = z_0, X_i = x_{L-1},\\ &\mathbf{1}\{Z_i = z_1, X_i = x_1\}), ..., \mathbf{1}\{Z_i = z_1, X_i = x_{L-1}\},\\ &\xi(Y_i, X_i, Z_i; z_0, x_0), ..., \xi(Y_i, X_i, Z_i; z_0, x_{L-1})\\ &\xi(Y_i, X_i, Z_i; z_1, x_0), ..., \xi(Y_i, X_i, Z_i; z_1, x_{L-1}))\end{aligned}$$

With mean $\boldsymbol{\mu}_{S_1,n} := \mathbb{E}[\boldsymbol{\pi}_{i,n}|S_1]$ and variance $\boldsymbol{\Sigma}_{S_1,n} := Var(\boldsymbol{\pi}_{i,n}|S_1)$

Let $\boldsymbol{P}_{n_2}$ denote the sample average:

$$\boldsymbol{P}_{n_2} := \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{\pi}_{i,n}$$

For any measurable vector-valued function $\boldsymbol{b}_{S_1,n} = \boldsymbol{b}_{S_1,n}(S_1, n)$ mapping realizations of S_1 to R^{2+3L} for which constants M_b and c_b exist such that $\mathbb{P}[||\boldsymbol{b}_{S_1,n}||_1 \leq M_b] \longrightarrow 1$ and $\mathbb{P}[\boldsymbol{b}_{S_1,n}^T \boldsymbol{\Sigma}_{S_1,n} \boldsymbol{b}_{S_1,n} > c_b] \longrightarrow 1$:

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left[\sqrt{n_2} \frac{\boldsymbol{b}_{S_1,n}^T \left(\boldsymbol{P}_{n_2} - \boldsymbol{\mu}_{S_1,n}\right)}{\sqrt{\boldsymbol{b}_{S_1,n}^T \boldsymbol{\Sigma}_{S_1,n} \boldsymbol{b}_{S_1,n}}} \leq t \middle| S_1 \right] - \Phi(t) \right| \xrightarrow{p} 0$$

Where $\Phi(\cdot)$ is the cumulative distribution of a standard normal random variable.

Lemma B.7. *Consider the sample averages that comprise the vector $\boldsymbol{P}_{n_2}$ defined in lemma B.6:*

$$\begin{aligned}P_{z_k,n_2} &= \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{1}\{Z_i = z_k\}\\ P_{z_k,x_l,n_2} &= \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{1}\{X_i = x_l, Z_i = z_k\}\\ P_{\xi,z_k,x_l,n_2} &= \frac{1}{n_2}\sum_{i=1}^{n_2} \xi(Y_i, X_i, Z_i; z_0, x_l)\end{aligned}$$

Let $\boldsymbol{P}_{n_2}$, $\boldsymbol{\mu}_{S_1,n}$ and $\boldsymbol{\Sigma}_{S_1,n}$ be defined as in Lemma B.6. Define the map $\boldsymbol{g} : \mathbb{R}^{2+3L} \to$

$\mathbb{R}^{3L-2}$ *as:*

$$\boldsymbol{g}(\boldsymbol{P}_{n_2}) = \Big(r_{0,0}(\boldsymbol{P}_{n_2}), \ldots, r_{0,L-1}(\boldsymbol{P}_{n_2}), \\ r_{1,0}(\boldsymbol{P}_{n_2}), \ldots, r_{1,L-1}(\boldsymbol{P}_{n_2}), \\ g_0(\boldsymbol{P}_{n_2}), \ldots, g_{L-1}(\boldsymbol{P}_{n_2}), \Big)^T,$$

where, for each $l = 0, \ldots, L-1$,

$$g_l(\boldsymbol{P}_{n_2}) := \frac{P_{\xi,z_0,x_l,n_2}}{P_{z_0,n_2}} + \frac{P_{\xi,z_1,x_l,n_2}}{1 - P_{z_0,n_2}},$$

and, for $k \in \{0,1\}$ *and* $l = 0, \ldots, L-1$,

$$r_{k,l}(\boldsymbol{P}_{n_2}) := \begin{cases} \frac{P_{z_0,x_l,n_2}}{P_{z_0,n_2}} & \text{if } k = 0 \\ \frac{P_{z_1,x_l,n_2}}{1-P_{z_0,n_2}} & \text{if } k = 1 \end{cases}.$$

Let $\boldsymbol{J}_{S_1,n}$ *denote the Jacobian of* $\boldsymbol{g}$ *evaluated at* $\boldsymbol{\mu}_{S_1,n}$, *and define* $\boldsymbol{V}_{S_1,n} := \boldsymbol{J}_{S_1,n}\boldsymbol{\Sigma}_{S_1,n}\boldsymbol{J}_{S_1,n}^T$.

Suppose that $P[Z = z_k] > c_z > 0$ *for* $k = 0, 1$. *Then, for any fixed* $\boldsymbol{a} \in \mathbb{R}^{3L}$ *such that there exists* $c_a > 0$ *that satisfies* $\mathbb{P}\left[\boldsymbol{a}^T\boldsymbol{V}_{S_1,n}\boldsymbol{a} > c_a\right] \longrightarrow 1$:

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left[\frac{\boldsymbol{a}^T\sqrt{n_2}\left(\boldsymbol{g}(\boldsymbol{P}_{n_2}) - \boldsymbol{g}(\boldsymbol{\mu}_{S_1,n})\right)}{\sqrt{\boldsymbol{a}^T\boldsymbol{V}_{S_1,n}\boldsymbol{a}}} \leq t \right] - \Phi(t) \right| \xrightarrow{p} 0.$$

Lemma B.8.
For every $x_l \in \mathcal{X}$ *and* $z_k \in \mathcal{Z}$:

$$n_2 Cov\big(\Psi_{x_l,n} - \Psi_{x_l}, \Psi_{x_l,n} - \Psi_{x_l'} \big| S_1\big) = \sum_{k=0,1} \frac{P_{\xi,x_l|z_k,n_2}(\mathbf{1}\{l = l'\} - P_{\xi,x_{l'}|z_k,n_2})}{\mathbb{P}[Z = z_k]} + o(1)$$

$$n_2 Cov\big(\Psi_{x_l,n} - \Psi_{x_l}, P_{x_{l'}|z_k,n_2} - P_{x_{l'}|z_k} \big| S_1\big) = \frac{P_{\xi,x_l|z_k,n_2}(\mathbf{1}\{l = l'\} - P_{x_{l'}|z_k,n_2})}{\mathbb{P}[Z = z_k]} + o(1)$$

$$n_2 Cov\big(P_{x_l|z_k,n_2} - P_{x_l|z_k}, P_{x_{l'}|z_{k'},n_2} - P_{x_{l'}|z_{k'}} \big| S_1\big) = \mathbf{1}\{k = k'\}\frac{P_{x_l|z_k}(\mathbf{1}\{l = l'\} - P_{x_l|z_{k'}})}{\mathbb{P}[Z = z_k]} + o(1)$$

Where $P_{\xi,x_l|z_k,n_2} := \mathbb{P}[\eta_{n_1}^*(X,Y) \neq z_k, X = x_l | Z = z_k, S_1]$.

Proof of lemma B.1.

Note that:

$$\begin{aligned}\mathbb{E}[\tilde{\Delta}_{n_1}(y,x_l)] - \Delta(y,x_l) =& \mathbb{P}[X=x_l|Z=z_1]\mathbb{E}\big[\frac{K_n(y-Y)}{h_n^d}|X=x_l,Z=z_1\big] - \Delta(y,x_l) \\ &- \mathbb{P}[X=x_l|Z=z_0]\mathbb{E}\big[\frac{K_n(y-Y)}{h_n^d}|X=x_l,Z=z_0\big] \\ =& \mathbb{P}[X=x_l|Z=z_1]\big(\mathbb{E}\big[\frac{K_n(y-Y)}{h_n^d}|X=x_l,Z=z_1\big] \\ &- f_{Y|X=x_l,Z=z_1}(y)\big) \\ &- \mathbb{P}[X=x_l|Z=z_0]\big(\mathbb{E}\big[\frac{K_n(y-Y)}{h_n^d}|X=x_l,Z=z_0\big] \\ &- f_{Y|X=x_l,Z=z_0}(y)\big)\end{aligned}$$

Where the first equality follows from the law of total probability and the definition of H_n, and the second equality from re-writing $\Delta(y,x_l)$ in terms of conditional distributions.

Then, an exact Taylor expansion of $f_{Y|X=x_l,Z=z_k}$, assumptions 5.2.1, 5.2.2, 5.1.2, and standard properties of kernel density estimators imply that if $\sup_{||u||=1} ||(\nabla^2 f_{Y|X=x_l,Z=z_k})u||$ is bounded, then:

$$\mathbb{E}\big[\frac{K_n(y-Y)}{h_n^d}|X=x_l,Z=z_k\big] - f_{Y|X=x_l,Z=z_k}(y) = O(h_n^2)$$

uniformly over y, x_l and z_k.

□

Proof of lemma B.2. First note that:

$$\begin{aligned}\mathbb{P}\Big[\Big|\Big|R_n(y,x_l)\Big|\Big|_\infty \geq r_n\Big] &\leq \mathbb{P}\Big[\Big|\Big|R_n^1(y,x_l)\Big|\Big|_\infty + \Big|\Big|R_n^0(y,x_l)\Big|\Big|_\infty \geq r_n\Big] \\ &\leq \mathbb{P}\Big[\Big|\Big|R_n^1(y,x_l)\Big|\Big|_\infty \geq \frac{r_n}{2}\Big] + \mathbb{P}\Big[\Big|\Big|R_n^0(y,x_l)\Big|\Big|_\infty \geq \frac{r_n}{2}\Big]\end{aligned}$$

To bound the previous probabilities, define the class of functions:

$$\mathcal{K}_n = \Big\{\kappa_n(y,x,z;t,x_l,z_k) = \mathbf{1}\{X=x_l,Z=z_k\}K_n\big(t-y\big) : x_l \in \mathcal{X}, z_k \in \mathcal{Z}, t \in \mathcal{Y}\Big\}$$

And the associated subdensity estimators:

$$f_{z_k,n}(y,x_l) := \Big[\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1}\mathbf{1}\{X_i = x_l, Z_i = z_k\}K_n\big(y - Y_i\big)\Big]$$

Note that if $K(\cdot)$ is bounded, for any constant $B' > 0$:

$$\begin{aligned}
&\mathbb{P}\Big[\big|\big|f_{z_k,n}(y,x_l) - \mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty \geq B'\Big] \\
&\quad = \mathbb{P}\Big[\Big|\Big|\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1}\big(\kappa_n(Y_i,X_i,Z_i;y,x_l,z_k) - \mathbb{E}[\kappa_n(Y,X,Z;y,x_l,z_k)]\big)\Big|\Big|_\infty \geq B'\Big] \\
&\quad = \mathbb{P}\Big[\Big|\Big|\sum_{i=1}^{n_1}\big(\kappa_n(Y_i,X_i,Z_i;y,x_l,z_k) - \mathbb{E}[\kappa_n(Y,X,Z;y,x_l,z_k)]\big)\Big|\Big|_\infty \geq B' n_1 h_n^d\Big]
\end{aligned}$$

Let $\sigma^2_{\kappa_n} = \sup_{\kappa\in\mathcal{K}_n} Var\big(\kappa(Y,X,Z)\big) = O(h_n^d)$ and $U = \sup_{\kappa\in\mathcal{K}_n}||\kappa||_\infty$. Note that for any $(t,x_l,z_k) \in \mathcal{Y}\times\mathcal{X}\times\mathcal{Z}$ the function $\kappa_n(\cdot,\cdot,\cdot;t,x_l,z_k)$ satisfies the conditions from assumption 5.2.1. Moreover, the map $(y,x,z)\mapsto\kappa_n(y,x,z;t,x_l,z_k)$ is jointly measurable and, for a sufficiently large n, the condition $0 < \sigma^2_{\kappa_n} < U/2$ is satisfied. Thus, corollary 2.2. from Gine and Gillou 2002 implies that for any sufficiently large B' and n, positive constants C and L depending only on the VC class of the family $\mathcal{K}_n$ exist, such that for every λ such that for $B'h_n^d \leq \frac{\lambda}{U}\sigma^2_{\kappa_n}$:

$$\mathbb{P}\Big[\big|\big|f_{z_k,n}(y,x_l) - \mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty \geq B'\Big] \leq L\exp\Big(\frac{-\log(1+\lambda/(4L))}{\lambda L}\frac{n_1(B'h_n^d)^2}{\sigma^2_{\kappa_n}}\Big) = o(n_1^{-1})$$

Moreover, since $K(\cdot)$ and f are bounded, the sequence $\big|\big|\mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty$ is also bounded. Thus, for some $B \geq B'$:

$$\begin{aligned}
\mathbb{P}\Big[\big|\big|f_{z_k,n}(y,x_l)\big|\big|_\infty \geq B\Big] &= \mathbb{P}\Big[\big|\big|f_{z_k,n}(y,x_l) - \mathbb{E}f_{z_k,n}(y,x_l) + \mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty \geq B\Big] \\
&\leq \mathbb{P}\Big[\big|\big|f_{z_k,n}(y,x_l) - \mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty + \big|\big|\mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty \geq B\Big] \\
&\leq \mathbb{P}\Big[\big|\big|f_{z_k,n}(y,x_l) - \mathbb{E}f_{z_k,n}(y,x_l)\big|\big|_\infty \geq B'\Big] \qquad (30)
\end{aligned}$$

Now let P_k denote $\mathbb{P}[Z = z_k]$ and note that for any $B > 0$:

$$\begin{aligned}
\mathbb{P}\Big[\Big|\frac{1}{P_{k,n_1}} - \frac{1}{P_k}\Big| B \geq \frac{r_n}{2}\Big] =& \mathbb{P}\Big[\Big|\frac{P_k - P_{k,n_1}}{P_k P_{k,n_1}}\Big| B \geq \frac{r_n}{2}\Big] \\
=& \mathbb{P}\Big[\Big|\frac{P_k - P_{k,n_1}}{P_k P_{k,n_1}}\Big| B \geq \frac{r_n}{2}, P_{k,n_1} \geq \frac{P_k}{2}\Big] \\
&+ \mathbb{P}\Big[\Big|\frac{P_k - P_{k,n_1}}{P_k P_{k,n_1}}\Big| B \geq \frac{r_n}{2}, P_{k,n_1} < \frac{P_k}{2}\Big] \\
\leq& \mathbb{P}\Big[|P_k - P_{k,n_1}| \geq \frac{P_k^2 r_n}{2B}\Big] + \mathbb{P}\Big[|P_{k,n_1} - P_k| \geq \frac{P_k}{2}\Big]
\end{aligned}$$

Where the second equality follows from the law of total probability and the inequality from the fact that $P_{k,n_1} < \frac{P_k}{2} \Rightarrow |P_{k,n_1} - P_k| \geq \frac{P_k}{2}$.

Thus, for a sufficiently large n_1, Hoeffding's inequality implies:

$$\begin{aligned}
\mathbb{P}\Big[\Big|\frac{1}{P_{k,n_1}} - \frac{1}{P_k}\Big| B \geq \frac{r_n}{2}\Big] \leq& 2\mathbb{P}\Big[|P_k - P_{k,n_1}| \geq \frac{P_k^2 r_n}{2B}\Big] \\
\leq& 4\exp\big(-\frac{P_k^4}{2B^2} n_1 r_n^2\big) = o(n_1^{-1}).
\end{aligned} \tag{31}$$

Then, constants B, B' and λ must exist such that $B' h_n^d \leq \frac{\lambda}{U}\sigma_{\kappa_n}^2$, $B \geq B'$ and:

$$\begin{aligned}
\mathbb{P}\Big[\Big|\Big|R_n^k(y, x_l)\Big|\Big|_\infty \geq \frac{r_n}{2}\Big] &= \mathbb{P}\Big[\Big|\Big|\Big(\frac{1}{P_{k,n_1}} - \frac{1}{P_k}\Big) f_{z_k,n}(y, x_l)\Big|\Big|_\infty \geq \frac{r_n}{2}\Big] \\
&= \mathbb{P}\Big[\Big|\frac{1}{P_{k,n_1}} - \frac{1}{P_k}\Big| ||f_{z_k,n}(y, x_l)||_\infty \geq \frac{r_n}{2}\Big] \\
&\leq \mathbb{P}\Big[\Big|\frac{1}{P_{k,n_1}} - \frac{1}{P_k}\Big| B \geq \frac{r_n}{2}\Big] + \mathbb{P}\Big[||f_{z_k,n}(y, x_l)||_\infty \geq B\Big] \\
&\leq 4\exp\big(-\frac{P_k^4}{2B^2} n_1 r_n^2\big) + L\exp\Big(\frac{-\log(1 + \lambda/(4L))}{\lambda L} \frac{n_1 (B' h_n^d)^2}{\sigma_{\kappa_n}^2}\Big) \\
&= o(n_1^{-1})
\end{aligned}$$

Where the last inequality follows from 30 and 31. □

Proof of lemma B.3.

This lemma shows that the asymptotic behavior of the difference $\Delta_{n_1}(y, x_l)$ is determined by the kernel density estimator component, and that the difference $|\Delta_{n_1}(y, x_l) - \Delta_n(y, x_l)|$ exceeds the rate $\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}$ with probability $o(n^{-1})$.

The result uses lemmas B.1 and B.2 which prove that the bias from kernel density estimation and the randomness from estimating conditional treatment and instrument probabilities are negligible at the rate $\sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}$.

For a given value of $x_l \in \mathcal{X}$, define:

$$H_n(y', x, z; x_l, y) = \Big[\frac{\mathbf{1}\{x = x_l, z = z_1\}}{\mathbb{P}[Z = z_1]} - \frac{\mathbf{1}\{x = x_l, z = z_0\}}{\mathbb{P}[Z = z_0]}\Big] K_n(y - y')$$

Then, $\Delta_{n_1}(y, x_l)$ can be decomposed into a main component depending on $H_n(y', x, z; x_l, y)$ plus a bias term and a residual that depends on the difference $\Big(\frac{1}{P_{1,n_1}} - \frac{1}{\mathbb{P}[Z=z_1]}\Big)$:

$$\begin{aligned}
\Delta_{n_1}(y, x_l) - \Delta(y, x_l) =& P_{1,n_1}^{-1}\Big[\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1} \mathbf{1}\{X_i = x_l, Z_i = z_1\} K_n\big(y - Y_i\big)\Big] \\
&- P_{0,n_1}^{-1}\Big[\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1} \mathbf{1}\{X_i = x_l, Z_i = z_0\} K_n\big(y - Y_i\big)\Big] - \Delta(y, x_l) \\
=& \frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1} H_n(Y_i, X_i, Z_i; x_l, y) - \Delta(y, x_l) \\
&+ \underbrace{\Big(\frac{1}{P_{1,n_1}} - \frac{1}{\mathbb{P}[Z = z_1]}\Big)\Big[\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1} \mathbf{1}\{X_i = x_l, Z_i = z_1\} K_n\big(y - Y_i\big)\Big]}_{:= R_n^1(y, x_l)} \\
&- \underbrace{\Big(\frac{1}{P_{0,n_1}} - \frac{1}{\mathbb{P}[Z = z_0]}\Big)\Big[\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1} \mathbf{1}\{X_i = x_l, Z_i = z_0\} K_n\big(y - Y_i\big)\Big]}_{:= R_n^0(y, x_l)} \\
=& \underbrace{\frac{1}{n_1 h_n^d}\sum_{i=1}^{n_1} H_n(Y_i, X_i, Z_i; x_l, y)}_{:= \tilde{\Delta}_{n_1}(y, x_l)} - \Delta(y, x_l) + R_n(y, x_l) \\
=& \Big(\tilde{\Delta}_{n_1}(y, x_l) - \mathbb{E}[\tilde{\Delta}_{n_1}(Y, x_l)]\Big) \\
&+ \underbrace{\Big(\mathbb{E}[\tilde{\Delta}_{n_1}(Y, x_l)] - \Delta(y, x_l)\Big)}_{Bias} + R_n(y, x_l) \qquad (32)
\end{aligned}$$

Where $R_n(y, x_l) = R_n^0(y, x_l) - R_n^1(y, x_l)$.

Define the class of functions:

$$\mathcal{H}_n = \Big\{ H_n(y, x, z; x, tl) : x_l \in \mathcal{X}, t \in \mathcal{Y} \Big\}$$

Let $\sigma^2_{H_n} = \sup_{H \in \mathcal{H}_n} Var\big(H(Y, X, Z)\big) = O(h_n^d)$ and $U = \sup_{H \in \mathcal{H}_n} ||H||_\infty$. Note that for any $(t, x_l) \in \mathcal{Y} \times \mathcal{X}$ the function $H_n(\cdot, \cdot, \cdot; t, x_l)$ satisfies the conditions from assumption 5.2.1. Moreover, the map $(y, x, z) \mapsto H_n(y, x, z; x_l, t)$ is jointly measurable and, for a sufficiently large n_1, the condition $0 < \sigma^2_{H_n} < U/2$ is satisfied. Thus, part 2 of corollary 2.2. from Gine and Gillou 2002 implies that positive constants C and L depending only on the VC class of the family $\mathcal{H}_n$ exist, such that for every $C' > C$:

$$\begin{aligned}
&\mathbb{P}\Big[\Big|\Big| \sum_{i=1}^{n_1} \big(H_n(Y_i, X_i, Z_i; x_l, y) - \mathbb{E}[H_n(Y_i, X_i, Z_i; x_l, y)]\big)\Big|\Big|_\infty > C' \sqrt{n_1 \sigma^2_{H_n}} \sqrt{\log\big(U \sigma^{-1}_{H_n}\big)}\Big] \\
&\quad \leq L \exp\Big\{ - \frac{C' \log(1 + C'/(4L))}{L} \log(U \sigma^{-1}_{H_n}) \Big\} \\
&\quad = L \exp\Big\{ - K(C') \log(U \sigma^{-1}_{H_n}) \Big\} \\
&\quad = L \big[U \sigma^{-1}_{H_n}\big]^{-K(C')} \\
&\quad = M \big[\sigma^{K(C')}_{H_n}\big] \qquad (33)
\end{aligned}$$

Where M is positive and depends on L and C', and $K'(C')$ denotes the term $\frac{C' \log(1 + C'/(4L))}{L}$.

Moreover, by standard properties of kernel density estimators:

$$\mathbb{E}\big[H_n(Y, X, Z; x_l, y)^2\big] = O(h_n^d)$$

Thus, $\sigma_{H_n} = O(h_n^{d/2})$ and $\sigma^{K(C')}_{H_n} = O(n_1^{-\gamma d K(C')/2})$.

Observe that the function $K(C')$ is strictly increasing in C', and thus, for a sufficiently large C', the upper bound from 33 is $o(n_1^{-1})$.

For any fixed C' such that $\sigma^{K(C')}_{H_n} = o(n_1^{-1})$, elementary algebra shows that a positive constant $C_0 \geq C'$ and a fixed integer N exist such that $\frac{C_0}{2} \sqrt{n_1 h_n^d \log(h_n^{-1})} \geq C' \sqrt{n_1 \sigma^2_{H_n} \log(U \sigma^{-1}_{H_n})}$ for any $n_1 \geq N$.

Fix such C_0 and consider the probability $\mathbb{P}\Big[\Big|\Big|\Delta_{n_1}(y, x_l) - \Delta(y, x_l)\Big|\Big|_\infty > C_0 \sqrt{\frac{\log(h_n^{-1})}{n_1 h_n^d}}\Big]$.

Then, lemma B.1 applied to the bias term in equation 32 yields:

$$\mathbb{P}\Big[\Big|\Big|\Delta_{n_1}(y,x_l)-\Delta(y,x_l)\Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big]$$
$$=\mathbb{P}\Big[\Big|\Big|\Big(\tilde{\Delta}_{n_1}(y,x_l)-\mathbb{E}[\tilde{\Delta}_{n_1}(y,x_l)]\Big)+O(h_n^2)+R_n(y,x_l)\Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big]$$

Lemma B.2 applied with $r_n=\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}$ implies that:

$$\mathbb{P}\Big[\Big|\Big|\Delta_{n_1}(y,x_l)-\Delta(y,x_l)\Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big]$$
$$=\mathbb{P}\Big[\Big|\Big|\Big(\tilde{\Delta}_{n_1}(y,x_l)-\mathbb{E}[\tilde{\Delta}_{n_1}(y,x_l)]\Big)+O(h_n^2)+o\Big(\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big)\Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big]+o(n_1^{-1})$$
$$\leq\mathbb{P}\Big[\Big|\Big|\Big(\tilde{\Delta}_{n_1}(y,x_l)-\mathbb{E}[\tilde{\Delta}_{n_1}(y,x_l)]\Big)\Big|\Big|_\infty+\Big|\Big|o\Big(\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big)\Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big]+o(n_1^{-1})$$

Where the last line follows from assumption 5.2.2 (which implies $h_n^2=o\Big(\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big)$) and the triangle inequality.

Then, for some integer N', if $n_1\geq\max\{N,N'\}$, the bound from 33 yields:

$$\mathbb{P}\Big[\Big|\Big|\Delta_{n_1}(y,x_l)-\Delta(y,x_l)\Big|\Big|_\infty > C_0\sqrt{\frac{\log(h_n^{-1})}{n_1h_n^d}}\Big]$$
$$\leq\mathbb{P}\Big[\Big|\Big|\sum_{i=1}^{n_1}\Big(H_n(Y_i,X_i,Z_i;x_l,y)-\mathbb{E}[H_n(Y_i,X_i,Z_i;x_l,y)]\Big)\Big|\Big|_\infty > \frac{C_0}{2}\sqrt{n_1h_n^d\log(h_n^{-1})}\Big]+o(n_1^{-1})$$
$$=o(n_1^{-1})$$

□

Proof of lemma B.4.

For a sufficiently large n:

$$\int_{\mathcal{M}_{x_l}(r_n)} \Big(\mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\}|\Delta(y,x_l)|\Big)dy \leq \int_{\mathcal{M}_{x_l}(r_n)} \Big(\mathbf{1}\{\eta^*(y,x_l) \neq \eta^*_{n_1}(y,x_l)\}r_n\Big)dy$$
$$= \big(\mu\{\mathcal{M}_{x_l}(r_n)\}\big)r_n$$
$$\leq Br_n^{1+\alpha} = O(r_n^{1+\alpha})$$

Where the firs inequality follows from the definition of the set $\mathcal{M}_{x_l}(r_n)$, and the second inequality from direct application of assumption 5.1.3.

For part 2, observe that:

$$\int_{\mathcal{M}_{x_l}(r_n)} \int_{\mathcal{M}_{x_l}(r_n)} \sqrt{\mathbb{P}[M(y,x_l)]\mathbb{P}[M(y',x_l)]}\big|\Delta(y,x_l)\big|\big|\Delta(y',x_l)\big|dydy'$$
$$\leq \big(\mu\{\mathcal{M}_{x_l}(r_n)\}\big)^2 r_n^2$$
$$= O(r_n^{2\alpha+2})$$

□

Proof of lemma B.5.

$$
\begin{aligned}
&\mathbb{E}\Big[L_{n_2}(\eta^*_{n_1}, x_l)|S_1\Big] \\
&= \mathbb{E}\Big[\frac{1}{n_2}\sum_{i=1}^{n_2}\Big(\frac{\mathbf{1}\{\eta^*_{n_1}(Y_i,X_i)=0, Z_i=z_1, X_i=x_l\}}{P_{1,n_2}} \\
&\qquad + \frac{\mathbf{1}\{\eta^*_{n_1}(Y_i,X_i)=1, Z_i=z_0, X_i=x_l\}}{P_{0,n_2}}\Big)\Big|S_1\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big\{\frac{1}{n_2}\sum_{i=1}^{n_2}\Big(\frac{\mathbf{1}\{\eta^*_{n_1}(Y_i,X_i)=0, Z_i=z_1, X_i=x_l\}}{P_{1,n_2}} \\
&\qquad + \frac{\mathbf{1}\{\eta^*_{n_1}(Y_i,X_i)=1, Z_i=z_0, X_i=x_l\}}{P_{0,n_2}}\Big)\Big|\{Z_i\}_{i=1}^{n_2}\Big\}\Big|S_1\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big\{\frac{1}{n_{2,z_1}}\sum_{i\in S_2, Z_i=z_1}^{n_2}\mathbf{1}\{\eta^*_{n_1}(Y_i,X_i)=0, Z_i=z_1, X_i=x_l\} \\
&\qquad + \frac{1}{n_{2,z_0}}\sum_{i\in S_2, Z_i=z_0}^{n_2}\mathbf{1}\{\eta^*_{n_1}(Y_i,X_i)=1, Z_i=z_0, X_i=x_l\}\Big|\{Z_i\}_{i=1}^{n_2}, S_1\Big\}\Big|S_1\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big\{\mathbb{P}[\eta^*_{n_1}(Y,X)=1, X=x_l|Z=z_0, S_1] \\
&\qquad + \mathbb{P}[\eta^*_{n_1}(Y,X)=0, X=x_l|Z=z_1, S_1]\Big|\{Z\}_{i=1}^{n_2}, S_1\Big\}\Big|S_1\Big] \\
&= \mathbb{P}[\eta^*_{n_1}(Y,X)=1, X=x_l|Z=z_0, S_1] + \mathbb{P}[\eta^*_{n_1}(Y,X)=0, X_i=x_l|Z=z_1, S_1] \\
&= \frac{\mathbb{P}[\eta^*_{n_1}(Y,X)=1, Z=z_0, X=x_l|S_1]}{\mathbb{P}[Z=z_0]} + \frac{\mathbb{P}[\eta^*_{n_1}(Y,X)=0, Z=z_1, X=x_l|S_1]}{\mathbb{P}[Z=z_1]} \\
&= L(\eta^*_{n_1}, x_l)
\end{aligned}
$$

□

Proof of lemma B.6.

Note that all the coordinates of $\boldsymbol{P}_i$ are indicator variables, thus, its support is bounded, and all its conditional moments exist.

Given the measurable function $\boldsymbol{b}_{S_1,n}$, define:

$$U_{i,n} = \boldsymbol{b}^T_{S_1,n}\big(\boldsymbol{\pi}_{i,n} - \boldsymbol{\mu}_{S_1,n}\big)$$

By definition, $\mathbb{E}[U_{i,n}|S_1] = 0$ and $\sigma_{b,S_1,n} := Var(U_i|S_1) = \boldsymbol{b}^T_{S_1,n}\boldsymbol{\Sigma}_{S_1,n}\boldsymbol{b}_{S_1,n}$.

Now consider the standardized sum:

$$\sqrt{n_2}\frac{\boldsymbol{b}_{S_1,n}^T\big(\boldsymbol{\pi}_{n_2}-\boldsymbol{\mu}_{S_1,n}\big)}{\sqrt{\boldsymbol{b}_{S_1,n}^T\boldsymbol{\Sigma}_{S_1,n}\boldsymbol{b}_{S_1,n}}}=\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\frac{U_{i,n}}{\sigma_{b,S_1,n}}$$

Let $s_n^2=\sum_{i=1}^{n_2}\sigma_{b,S_1,n}^2=n_2\sigma_{b,S_1,n}^2$. Then, on the events $\sigma_{b,S_1,n}^2>c_b$ and $||\boldsymbol{b}_{S_1,n}||_1\leq M_b$, for any fixed $\epsilon>0$ and a sufficiently large n:

$$\begin{aligned}\frac{1}{s_n^2}\sum_{i=1}^{n_2}\mathbb{E}\big[U_{i,n}^2\mathbf{1}\{|U_i|\geq\epsilon s_n\}|S_1\big]&\leq\frac{4||\boldsymbol{b}_{S_1,n}||_1^2}{s_n^2}\sum_{i=1}^{n_2}\mathbb{E}\big[\mathbf{1}\{|U_i|\geq\epsilon s_n\}|S_1\big]\\&=\frac{4M_b^2}{s_n^2}\sum_{i=1}^{n_2}\mathbb{P}[|U_i|\geq\epsilon s_n|S_1]\\&=0\end{aligned}$$

Where the inequality follows from the fact that $|U_2|\leq 2||\boldsymbol{b}_{S_1,n}||_1\leq 2M_b$ which in turn implies that $\mathbb{P}[|U_i|\geq\epsilon s_n|S_1]=0$.

Thus, the Lindeberg condition holds which implies that, conditional on S_1, $\sigma_{b,S_1,n}^2>c_b$ and $||\boldsymbol{b}_{S_1,n}||_1\leq M_b$, the CDF of the sum $\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\frac{U_{i,n}}{\sigma_{b,S_1,n}}$ converges uniformly to $\Phi(\cdot)$.

Hence:

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left[\sqrt{n_2}\frac{\boldsymbol{b}_{S_1,n}^T\big(\boldsymbol{P}_{n_2}-\boldsymbol{\mu}_{S_1,n}\big)}{\sqrt{\boldsymbol{b}_{S_1,n}^T\boldsymbol{\Sigma}_{S_1,n}\boldsymbol{b}_{S_1,n}}}\leq t\middle|S_1\right]-\Phi(t)\right|\xrightarrow{p}0\tag{34}$$

□

Proof of Lemma B.7. The proof follows from the Delta method and Lemma B.6. First note that, by construction:

$$g_l(\boldsymbol{\pi}_{n_2})=\frac{\pi_{\xi,z_0,x_l,n_2}}{\pi_{z_0,n_2}}+\frac{\pi_{\xi,z_1,x_l,n_2}}{\pi_{z_1,n_2}}=L_{n_2}(\eta_{n_1}^*,x_l),$$

and

$$g_l(\boldsymbol{\mu}_{S_1,n})=\frac{\mu_{\xi,z_0,x_l,S_1,n}}{P[Z=z_0]}+\frac{\mu_{\xi,z_1,x_l,S_1,n}}{P[Z=z_1]}=L(\eta_{n_1}^*,x_l).$$

Similarly:

$$r_{k,l}(\boldsymbol{\pi}_{n_2}) = \frac{\pi_{z_k,x_l,n_2}}{\pi_{z_k,n_2}}$$

is the sample analogue of $P[X = x_l \mid Z = z_k]$, and:

$$r_{k,l}(\boldsymbol{\mu}_{S_1,n}) = \frac{P[X = x_l, Z = z_k]}{P[Z = z_k]} = P[X = x_l \mid Z = z_k].$$

Since $P[Z = z_k] > c_z > 0$ for $k = 0, 1$, the map $\boldsymbol{g}$ is continuously differentiable in a neighborhood of $\boldsymbol{\mu}_{S_1,n}$ with probability approaching one. Moreover, by the law of large numbers conditional on S_1,

$$\boldsymbol{\Pi}_{n_2} - \boldsymbol{\mu}_{S_1,n} = o_p(1).$$

Therefore, a first-order Taylor expansion yields

$$\sqrt{n_2}\left(\boldsymbol{g}(\boldsymbol{\Pi}_{n_2}) - \boldsymbol{g}(\boldsymbol{\mu}_{S_1,n})\right) = \boldsymbol{J}_{S_1,n}\sqrt{n_2}\left(\boldsymbol{\Pi}_{n_2} - \boldsymbol{\mu}_{S_1,n}\right) + \boldsymbol{R}_n,$$

where, for every fixed $\boldsymbol{a} \in \mathbb{R}^{3L}$, $\boldsymbol{a}^T\boldsymbol{R}_n = o_p(1)$.

Fix $\boldsymbol{a} \in \mathbb{R}^{3L}$ and define

$$\boldsymbol{b}_{S_1,n} := \boldsymbol{J}_{S_1,n}^T\boldsymbol{a}.$$

Because $\boldsymbol{J}_{S_1,n}$ is a function of $\boldsymbol{\mu}_{S_1,n}$, and $\boldsymbol{\mu}_{S_1,n}$ is measurable with respect to S_1, the vector $\boldsymbol{b}_{S_1,n}$ is S_1-measurable. Furthermore, since all coordinates of $\boldsymbol{\mu}_{S_1,n}$ are probabilities and $P[Z = z_k]$ is bounded away from zero, there exists a finite constant M_b such that

$$\mathbb{P}\left[\|\boldsymbol{b}_{S_1,n}\|_1 \leq M_b\right] \longrightarrow 1.$$

Also,

$$\boldsymbol{b}_{S_1,n}^T\boldsymbol{\Sigma}_{S_1,n}\boldsymbol{b}_{S_1,n} = \boldsymbol{a}^T\boldsymbol{J}_{S_1,n}\boldsymbol{\Sigma}_{S_1,n}\boldsymbol{J}_{S_1,n}^T\boldsymbol{a} = \boldsymbol{a}^T\boldsymbol{V}_{S_1,n}\boldsymbol{a}.$$

By assumption,

$$\mathbb{P}\left[\boldsymbol{b}_{S_1,n}^T \boldsymbol{\Sigma}_{S_1,n} \boldsymbol{b}_{S_1,n} > c_a\right] \longrightarrow 1.$$

Thus, Lemma B.6 applies to the random direction $\boldsymbol{b}_{S_1,n} = \boldsymbol{J}_{S_1,n}^T \boldsymbol{a}$ and implies

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left[\frac{\boldsymbol{a}^T \boldsymbol{J}_{S_1,n} \sqrt{n_2}(\boldsymbol{\Pi}_{n_2} - \boldsymbol{\mu}_{S_1,n})}{\sqrt{\boldsymbol{a}^T \boldsymbol{V}_{S_1,n} \boldsymbol{a}}} \leq t \middle| S_1 \right] - \Phi(t) \right| \xrightarrow{p} 0.$$

Finally, since:

$$\boldsymbol{a}^T \sqrt{n_2}\left(\boldsymbol{g}(\boldsymbol{\Pi}_{n_2}) - \boldsymbol{g}(\boldsymbol{\mu}_{S_1,n})\right) = \boldsymbol{a}^T \boldsymbol{J}_{S_1,n} \sqrt{n_2}(\boldsymbol{\Pi}_{n_2} - \boldsymbol{\mu}_{S_1,n}) + o_p(1),$$

and since: $\boldsymbol{a}^T \boldsymbol{V}_{S_1,n} \boldsymbol{a}$ is bounded away from zero with probability approaching one, Slutsky's theorem gives:

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left[\frac{\boldsymbol{a}^T \sqrt{n_2}\left(\boldsymbol{g}(\boldsymbol{\Pi}_{n_2}) - \boldsymbol{g}(\boldsymbol{\mu}_{S_1,n})\right)}{\sqrt{\boldsymbol{a}^T \boldsymbol{V}_{S_1,n} \boldsymbol{a}}} \leq t \middle| S_1 \right] - \Phi(t) \right| \xrightarrow{p} 0. \tag{35}$$

Now set $T_n := \frac{\boldsymbol{a}^T \sqrt{n_2}\left(\boldsymbol{g}(\boldsymbol{\Pi}_{n_2}) - \boldsymbol{g}(\boldsymbol{\mu}_{S_1,n})\right)}{\sqrt{\boldsymbol{a}^T \boldsymbol{V}_{S_1,n} \boldsymbol{a}}}$ and note that:

$$\begin{aligned} \sup_t \left| \mathbb{P}[T_n = t] - \boldsymbol{\Phi}(t) \right| =& \sup_t \left| \mathbb{E}[\mathbb{P}[T_n \leq t | S_1]] - \boldsymbol{\Phi}(t) \right| \\ \leq& \mathbb{E}\left[\sup_t \left| \mathbb{P}[T_n \leq t | S_1] - \boldsymbol{\Phi}(t) \right| \right] \end{aligned}$$

Thus, 35 becomes:

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left[\frac{\boldsymbol{a}^T \sqrt{n_2}\left(\boldsymbol{g}(\boldsymbol{\Pi}_{n_2}) - \boldsymbol{g}(\boldsymbol{\mu}_{S_1,n})\right)}{\sqrt{\boldsymbol{a}^T \boldsymbol{V}_{S_1,n} \boldsymbol{a}}} \leq t \right] - \Phi(t) \right| \xrightarrow{p} 0.$$

This proves the claim. □

Proof of lemma B.8.
First consider $Var(A_{1,n}(x_l))$:

$$\begin{aligned} Var(A_{1,n}(x_l)) &= Var(\mathbb{E}[A_{1,n}(x_l)|S_1]) + \mathbb{E}(Var[A_{1,n}(x_l)|S_1]) \\ &= \mathbb{E}(Var[A_{1,n}(x_l)|S_1]) \end{aligned}$$

Where the first line follows from the law of total variance and the second from lemma 5.1.

Note that:

$$\begin{aligned}
E[L_{n_2}(\eta^*_{n_1}, x_l)|S_1] =& \mathbb{E}\Big[\sum_{i=1}^{n_2}\Big(\sum_{k=0,1}\frac{1}{n_{2,k}}\xi(Y_i, X_i, Z_i; z_k, x_l)\Big)\Big|S_1\Big] \\
=& \mathbb{E}\Big\{\mathbb{E}\Big[\sum_{i=1}^{n_2}\Big(\sum_{k=0,1}\frac{1}{n_{2,k}}\xi(Y_i, X_i, Z_i; z_k, x_l)\Big)\Big|\{Z_i\}_{i\in S_2}, S_1\Big]\Big|S_1\Big\} \\
=& \mathbb{E}\Big\{\sum_{k=0,1}\mathbb{E}[\xi(Y_i, X_i, Z_i; z_k, x_l)|Z_i = z_k, S_1]\Big|S_1\Big\} \\
=& \mathbb{P}\big[\eta^*_{n_1}(Y, X, Z; z_0, x_l) \neq k, X = x_l|Z = z_0, S_1\big] \\
&+ \mathbb{P}\big[\eta^*_{n_1}(Y, X, Z; z_1, x_l) \neq k, X = x_l|Z = z_1, S_1\big] \qquad (36)
\end{aligned}$$

Where the second line comes from conditioning on realized instrument values, the third from the fact that observations are i.i.d. conditional on instrument values, and the last from the definition of $\xi(\cdot)$.

And:

$$\begin{aligned}
\mathbb{E}\Big[L_{n_2}(\eta^*_{n_1}, x_l)^2|S_1\Big] =& \mathbb{E}\Big[\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\Big(\sum_{k=0,1}\frac{1}{n_{2,k}}\xi(Y_i, X_i, Z_i; z_k, x_l)\Big) \\
&\Big(\sum_{k=0,1}\frac{1}{n_{2,k}}\xi(Y_j, X_j, Z_j; z_k, x_l)\Big)\Big|S_1\Big] \\
=& \mathbb{E}\Big[\sum_{i=1}^{n_2}\Big(\sum_{k=0,1}\frac{1}{n^2_{2,k}}\xi(Y_i, X_i, Z_i; z_k, x_l) \qquad (37) \\
&+ \sum_{j\neq i, j\in S_2}\sum_{k'=0,1}\frac{1}{n_{2,k}n_{2,k'}}\xi(Y_i, X_i, Z_i; z_k, x_l)\xi(Y_j, X_j, Z_j; z_k, x_l)\Big)\Big|S_1\Big]
\end{aligned}$$

Now expand the second term and condition on instrument realizations:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^{n_2}\sum_{k=0,1}\sum_{j\neq i, j\in S_2}\sum_{k'=0,1}\frac{1}{n_{2,k}n_{2,k'}}\xi(Y_i,X_i,Z_i;z_k,x_l)\xi(Y_j,X_j,Z_j;z_k,x_l)\Big)\Big|S_1\Big]\\
=&\mathbb{E}\Big\{\mathbb{E}\Big[\sum_{i=1}^{n_2}\sum_{k=0,1}\sum_{j\neq i, j\in S_2}\sum_{k'=0,1}\frac{1}{n_{2,k}n_{2,k'}}\xi(Y_i,X_i,Z_i;z_k,x_l)\\
&\qquad\qquad\times\xi(Y_j,X_j,Z_j;z_k,x_l)\Big)\Big|\{Z_i\}_{i\in S_2},S_1\Big]\Big|S_1\Big\}\\
=&\mathbb{E}\Big[\frac{n_{2,0}(n_{2,0}-1)}{n_{2,0}^2}\mathbb{E}\big[\xi(Y,X,Z;z_0,x_l)|Z=z_0,S_1\big]^2\\
&+\frac{n_{2,1}(n_{2,1}-1)}{n_{2,1}^2}\mathbb{E}\big[\xi(Y,X,Z;z_1,x_l)|Z=z_1,S_1\big]^2\\
&+2\big(\mathbb{E}\big[\xi(Y,X,Z;z_0,x_l)|Z=z_0,S_1\big]\mathbb{E}\big[\xi(Y,X,Z;z_1,x_l)|Z=z_1,S_1\big]\big)\Big|S_1\Big] \qquad (38)
\end{aligned}$$

Combining equations 36, 37 and 38 yields:

$$\begin{aligned}
Var(A_{1,n}(x_l)|S_1)=&\mathbb{E}\Big[\frac{n_2}{n_{2,k}}\sum_{k=0,1}\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]\\
&\times\big(1-\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]\big)\Big|S_1\Big]
\end{aligned}$$

Moreover, the terms $\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k]\times\big(1-\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k]\big)4$ for $k=0,1$ are bounded and the difference $\mathbb{E}\big[\frac{n_2}{n_{2,k}}\big]-\frac{1}{\mathbb{P}[Z=z_k]}$ is $o(1)$ and independent from S_1, thus:

$$\begin{aligned}
Var(A_{1,n}(x_l)|S_1)=&\mathbb{E}\Big[\sum_{k=0,1}\frac{\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]}{\mathbb{P}[Z=z_k]}\\
&\times\big(1-\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]\big)\Big|S_1\Big]+o(1)\\
=&\sum_{k=0,1}\frac{\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]}{\mathbb{P}[Z=z_k]}\\
&\times\big(1-\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]\big)+o(1)\\
=&\frac{P_{\xi,x_l|z_0,n_2}(1-P_{\xi,x_l|z_0,n_2})}{\mathbb{P}[Z=z_0]}+\frac{P_{\xi,x_l|z_1,n_2}(1-P_{\xi,x_l|z_1,n_2})}{\mathbb{P}[Z=z_1]}+o(1) \qquad (39)
\end{aligned}$$

Where $P_{\xi,x_l|z_k,n_2}:=\mathbb{P}[\eta^*_{n_1}(X,Y)\neq z_k,X=x_l|Z=z_k,S_1]$ is the missclassification probability for the pre-trained classifier associated with $X=x_l$, conditional on $Z=z_k$

To derive $Cov(A_{1,n}(x_l), A_{1,n}(x_{l'}))$ consider the variance of the sum $A_{1,n}(x_l)+A_{1,n}(x_{l'})$, define $\zeta(Y,X,Z;z_l,x_l,x_l') := \xi(Y,X,Z;z_l,x_l)+\xi(Y,X,Z;z_l,x_l')$ and note that:

$$\mathbb{E}\big[\big(L_{n_2}(\eta^*_{n_1},x_l)+L_{n_2}(\eta^*_{n_1},x_{l'})\big)|S_1\big] = \mathbb{E}\Big[\sum_{i=1}^{n_2}\Big(\sum_{k=0,1}\frac{1}{n_{2,k}}\zeta(Y_i,X_i,Z_i;z_k,x_l,x_{l'})\Big)\Big|S_1\Big]$$

Then, a derivation analogous to the variance of $A_{1,n}(x_l)$ yields:

$$\begin{aligned}Var(A_{1,n}(x_l)+A_{1,n}(x_{l'})|S_1) =& \frac{(P_{\xi,x_l|z_0,n_2}+P_{\xi,x_{l'}|z_0,n_2})(1-P_{\xi,x_l|z_0,n_2}-P_{\xi,x_{l'}|z_0,n_2})}{\mathbb{P}[Z=z_0]}\\ &+\frac{(P_{\xi,x_l|z_1,n_2}+P_{\xi,x_{l'}|z_1,n_2})(1-P_{\xi,x_l|z_1,n_2}-P_{\xi,x_{l'}|z_1,n_2})}{\mathbb{P}[Z=z_1]}+o(1)\end{aligned}$$

Thus:

$$Cov(A_{1,n}(x_l),A_{1,n}(x_{l'})|S_1) = -\frac{P_{\xi,x_l|z_0,n_2}P_{\xi,x_{l'}|z_0,n_2}}{\mathbb{P}[Z=z_0]}-\frac{P_{\xi,x_l|z_1,n_2}P_{\xi,x_{l'}|z_1,n_2}}{\mathbb{P}[Z=z_1]}+o(1) \quad (40)$$

Now consider the conditional treatment probability estimators $P_{x_l|z_k,n_2} := \frac{P_{z_k,n_2}}{P_{z_k,x_l,n_2}}$. Note that:

$$\begin{aligned}&\mathbb{E}\Big[L_{n_2}(\eta^*_{n_1},x_l)P_{x_{l'}|z_k,n_2}\Big]\\ &=\mathbb{E}\Big[\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\Big(\sum_{k'=0,1}\frac{1}{n_{2,k}}\xi(Y_i,X_i,Z_i;z_{k'},x_l)\Big)\Big(\frac{1}{n_{2,k}}\mathbf{1}\{X_j=x_{l'},Z_j=z_k\}\Big)\Big|S_1\Big]\\ &=\mathbb{E}\Big[\sum_{i=1}^{n_2}\Big(\frac{\mathbf{1}\{l=l'\}}{n_{2,k}^2}\xi(Y_i,X_i,Z_i;z_k,x_l)\\ &\quad+\sum_{j\neq i,j\in S_2}\sum_{k'=0,1}\frac{1}{n_{2,k}n_{2,k'}}\xi(Y_i,X_i,Z_i;z_{k'},x_l)\mathbf{1}\{X_j=x_{l'},Z_j=z_k\}\Big)\Big|S_1\Big]\end{aligned}$$

Again, expand the second term and condition on instrument realizations to obtain:

$$\begin{aligned}&\mathbb{E}\Big[\sum_{i=1}^{n_2}\sum_{j\neq i,j\in S_2}\sum_{k'=0,1}\frac{1}{n_{2,k}n_{2,k'}}\xi(Y_i,X_i,Z_i;z_{k'},x_l)\mathbf{1}\{X_j=x_{l'},Z_j=z_k\}\Big|S_1\Big]\\ &\quad=\mathbb{E}\Big[\sum_{k'=0,1}\frac{(n_{2,k}-\mathbf{1}\{k=k'\})n_{2,k'}}{n_{2,k}n_{2,k'}}P_{\xi,x_l|z_{k'},n_2}P_{x_{l'}|z_k,n_2}\Big|S_1\Big]\end{aligned}$$

Therefore:

$$Cov\Big(A_{1,n}(x_l), \sqrt{n_2}\big(P_{x_{l'}|z_k,n_2} - P_{x_{l'}|z_k}\big)\Big) = \frac{P_{\xi,x_l|z_k,n_2}(\mathbf{1}\{l = l'\} - P_{x_l|z_k,n_2})}{\mathbb{P}[Z = z_k]} + o(1) \quad (41)$$

Finally, the delta method yields:

$$\begin{aligned} &Cov\Big(\sqrt{n_2}\big(P_{x_l|z_k,n_2} - P_{x_l|z_k}\big), \sqrt{n_2}\big(P_{x_{l'}|z_{k'},n_2} - P_{x_{l'}|z_{k'}}\big)\Big) \quad (42) \\ &= \mathbf{1}\{k = k'\}\frac{\mathbf{1}\{l = l'\}P_{x_l|z_k} - P_{x_l|z_k}P_{x_l|z_{k'}}}{\mathbb{P}[Z = z_k]} + o(1) \end{aligned}$$

Equations 39, 40, 41 and 42 characterize the entire variance-covariance matrix of the vector $\mathbf{\Pi}_{n_2}$ conditional on S_1.

□

B.3 Proofs for subsection 5.4

Proof of theorem 5.2.

Consider a sequence $\{F_n \in \mathcal{F}\}_{n=1}^{\infty}$ and a sequence of *i.i.d.* samples $S_{1,n}$ and $S_{2,n}$ drawn from F_n.

Suppose toward contradiction that:

$$\limsup_{n\to\infty}\{\mathbb{P}_{F_n}(\phi_n^{CC}(q)=1)\} > q$$

Then, there exist $\epsilon > 0$ and a subsequence n_m such that:

$$P_{F_{n_m}}(\phi_{n_m}^{CC}(q)=1) > q+\epsilon \quad \forall m,$$

Define:

$$R_n(S_1) = \mathbb{P}_{F_n}(\phi_n^{CC}(q)=1|S_{1,n})$$

Then, by iterated expectations:

$$\mathbb{E}_{F_{n_m}}[R_n(S_1)] > q+\epsilon \quad \forall m \tag{43}$$

Note that:

$$\begin{aligned}\mathbb{E}_{F_n}[R_n(S_1)] =& \mathbb{E}_{F_n}[R_n(S_1)\mathbf{1}\{R_n(S_1) > q+\epsilon/2\}] \\ &+ \mathbb{E}_{F_n}[R_n(S_1)\mathbf{1}\{R_n(S_1) \leq q+\epsilon/2\}] \\ \leq& \mathbb{P}_{F_n}[\{R_n(S_1) > q+\epsilon/2\}] \\ &+ \mathbb{P}_{F_n}[\{R_n(S_1) \leq q+\epsilon/2\}](q+\epsilon/2).\end{aligned}$$

Which implies that:

$$\mathbb{P}_{F_n}[R_n(S_1) > q+\epsilon/2] > \epsilon/2$$

Since the preceding event has probability bounded away from zero, select a sequence of training-sample realizations $\{s_{1,n_m}\}_{m=1}^{\infty}$ such that $R_{n_m}(s_{1,n_m}) > q+\epsilon/2$ for every m. In what follows, we consider the distributions conditional on these selected realizations.

First note from lemma B.8 that all the entries of the leading terms in $\mathbf{\Omega}_n = n_2 Var(\boldsymbol{\theta}_n|S_1)$ are ratios in which the numerator depends on the probabilities of the events $\{\eta_{n_1}(Y,X) \neq k, X = x_l, Z = z_k\}$ and $\{X = x_l, Z = z_k\}$, and the denominator takes the form $\mathbb{P}[Z = z_k]$. This is a finite collection of probabilities, thus the sequence indexed by n_m admits a subsequence n_q along which all these probabilities converge. Furthermore, assumption 5.1 guarantees that these probabilities converge to a limit bounded away from both zero and one.

If these probabilities converge, the subsequence of matrices $\mathbf{\Omega}_q$ also converges to a limit $\mathbf{\Omega}$.

Suppose toward contradiction that $\mathbf{\Omega}$ is singular. Then there exists a fixed vector $\boldsymbol{c} = (\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{c}_\Psi)$ such that $||\boldsymbol{c}||_1 = 1$ and

$$\boldsymbol{c}^\top \mathbf{\Omega}_q \boldsymbol{c} \to 0.$$

By the conditional variance representation from lemma B.8:

$$\boldsymbol{c}^\top \mathbf{\Omega}_q \boldsymbol{c} = \sum_{k=0,1} \frac{\text{Var}(H_k \mid Z = z_k)}{p_k},$$

where

$$H_k(Y,X) = \sum_{x_l \in \boldsymbol{\mathcal{X}_k}} c_{k,l} \mathbf{1}\{X = x_l, Z = z_k\} + \sum_{x_l \in \mathcal{X}} c_{\Psi,l} \mathbf{1}\{\eta_{n1,q}(Y,X) \neq k, X = x_l, Z = z_k\}.$$

Since $\text{Var}(H_k \mid Z = z_k)$ is nonnegative it must be that $\text{Var}(H_k \mid Z = z_k) \leq \epsilon * p_k := \epsilon'$.

Let $x_{l^{k*}}$ be the a treatment value excluded from $\boldsymbol{\mathcal{X}}_k$. Note that on the event $A_{k*} = \{\eta_{n1,q}(Y,X) = k, X = x_l^{*k}, Z = z_k\}$, the term $H_k(Y,X)$ is equal to zero. Thus $\mathbb{E}[H_k(Y,X)|Z = z_k] = \mathbb{E}[H_k(Y,X)\mathbf{1}\{A_{k*}^C\}|Z = z_k]$.

Hence: $\big(\mathbb{E}[H_k(Y,X)|Z = z_k]\big)^2 \leq \big(1 - \mathbb{P}[A_{k*}|Z = z_k]\big)\mathbb{E}[H_k(Y,X)^2|Z = z_k]$ and thus $Var\Big(H_k(Y,X)|Z = z_k\Big) \geq \mathbb{P}[A_{k*}|Z = z_k]\mathbb{E}[H_k(Y,X)^2|Z = z_k]$.

Define $A_l = \{\eta_{n1,q}(Y,X) = k, X = x_l, Z = z_k\}$, $A_l^C = \{\eta_{n1,q}(Y,X) \neq k, X = x_l, Z = z_k\}$ and note that:

$$\mathbb{E}[H_k(Y,X)^2|Z = z_k] = \sum_{l=0}^{L-1} \Big(\mathbb{P}[A_l|Z = z_k]c_{k,l}^2 + \mathbb{P}[A_l^C|Z = z_k](c_{k,l} + c_{\Psi,l})^2\Big)$$

The probabilities $\mathbb{P}[A_l|Z = z_k]$ and $\mathbb{P}[A_l^C|Z = z_k]$ are bounded away from zero by

definition 5.1 for every k and l. Thus, it must be that:

$$c_{k,l}^2 < \epsilon'$$
$$(c_{k,l} + c_{\Psi,l})^2 < \epsilon'$$

Which implies that all entries of $\boldsymbol{c} < \epsilon'$ which contradicts $||\boldsymbol{c}||_1 = 1$ for a sufficiently small ϵ. Thus, $\boldsymbol{\Omega}$ must be positive definite.

Note that the convergence results in Proposition B.1 and Lemma B.7 also hold along arbitrary sequences $\{F_n\}_{n=1}^{\infty} \subset \mathcal{F}$. In particular, definition 5.1 imposes common constants for the boundedness, smoothness, and margin conditions across $\mathcal{F}$. Moreover, the inequalities from Giné and Guillou (2002) only require that the densities be bounded. Thus, the inequalities hold uniformly with constants depending only on kernel function . Consequently, the constants appearing in the inequalities of Lemmas C.2 and C.3, as well as the bounds on the classification error in Proposition B.1, can be chosen independently of $F \in \mathcal{F}$. Hence, the corresponding convergence results hold uniformly over $\mathcal{F}$.

Furthermore, proposition B.1 and lemma B.7 imply that:

$$\sqrt{n_{2_q}}\boldsymbol{\Omega}_q^{-1/2}(\boldsymbol{\theta}_q - \boldsymbol{\theta}_{0,q})\big|S_{1,q} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_m)$$

And lemma B.8 and essential properties of count estimators imply that $\boldsymbol{\Sigma}_{n_q}$ converges in probability to $\boldsymbol{\Omega}_{n_q}$ conditional on S_1. Thus

$$\left|\left|\boldsymbol{\Omega}_{n_q}^{-1/2}\boldsymbol{\Sigma}_{n_q}\boldsymbol{\Omega}_{n_q}^{-1/2} - \boldsymbol{I}_m\right|\right| \xrightarrow{p} 0 \tag{44}$$

The previous discussion implies that, given the sequence $\{S_{1,n}\}_{n=1}^{\infty}$, the sequences $\{F_n\}_{n=1}^{\infty}$ and $\{S_{s,n}\}_{n=1}^{\infty}$ are such that the sequence $\{\boldsymbol{\Theta}_n, \boldsymbol{\Omega}_n, \boldsymbol{\Sigma}_n\}_{n=1}^{\infty}$ and the sequence $\{\boldsymbol{\Theta}_{n,0}\}_{n=1}^{\infty}$, where $\boldsymbol{\Theta}_{n,0}$ denotes the true value of the parameters under F_n and $\boldsymbol{\Theta}_n$ the estimator, is such that assumption 2 from Cox and Shi (2023) is satisfied. Thus:

$$\limsup_{j\to\infty} R_{n_j}(s_{1,n_j}) \leq q.$$

Which contradicts $R_{n_q}(s_{1,n_q}) > q + \epsilon/2 \quad \forall q$. Since the sequence $\{F_n\}_{n=1}^{\infty}$ was arbitrary, the desired uniform size result follows. □

C Details of main simulations and supplementary simulation results

C.1 Data generating process for subsection 6.1

The data generating process can be formally expressed as:

$$f_{Y|Z=z_k}(y; w, \mu_2) = \frac{1}{1 + w\phi(\mu_2/2)} \begin{cases} \phi(y - \mathbf{1}\{Z = z_1\}\mu_2), & y < \frac{\mu_2}{2} \quad (\textit{Scaled normal}), \\ \phi\left(\frac{\mu_2}{2}\right), & \frac{\mu_2}{2} \leq y \leq \frac{\mu_2}{2} + w \quad (\textit{Uniform}), \\ \phi(y - w - \mathbf{1}\{Z = z_1\}\mu_2), & y > \frac{\mu_2}{2} + w \quad (\textit{Scaled normal}), \end{cases}$$

$$X \sim U\{0, 1\}, \quad \mathbb{P}[Z = z_1] = 1/3, \quad \mathbb{P}[Z = z_0] = 2/3, \quad X \perp (Y, Z)$$

Where ϕ is the standard normal density.

C.2 Additional simulations for subsection 6.1: Uniform and Epanechnikov kernels

Table 5: Bias, overlap coverage, and joint overlaps and CTPs coverage for sample splitting estimators. Uniform kernel.

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$		
$n = 500$											
$\mathbb{E}n_{z_0} = 167$, $\mathbb{E}n_{z_1} = 333$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.95	0.94	0.92	0.94	0.93	0.92	0.93	0.93	0.92
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.05	0.04	0.04	0.05	0.04	0.03	0.04	0.03	0.03
		$\Psi_{x_0} \in C_{0.95}$	0.81	0.88	0.92	0.84	0.91	0.93	0.86	0.91	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.65	0.79	0.84	0.71	0.83	0.89	0.70	0.83	0.89
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.89	0.93	0.95	0.92	0.94	0.93	0.93	0.93	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.87	0.91	0.91	0.91	0.93	0.90	0.92	0.93	0.90
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.04	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.87	0.91	0.92	0.89	0.92	0.94	0.87	0.90	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.80	0.87	0.89	0.81	0.88	0.91	0.79	0.87	0.90
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.94	0.94	0.94	0.94	0.94	0.93	0.95	0.95	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.87	0.91	0.93	0.91	0.91	0.93	0.92	0.92	0.92
$n = 1000$											
$\mathbb{E}n_{z_0} = 333$, $\mathbb{E}n_{z_1} = 666$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.95	0.97	0.96	0.96	0.94	0.94	0.95	0.96	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.94	0.94	0.94	0.93	0.94	0.93	0.93	0.93	0.93
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.04	0.04	0.03	0.03	0.03	0.02	0.03	0.02	0.02
		$\Psi_{x_0} \in C_{0.95}$	0.76	0.86	0.90	0.79	0.89	0.92	0.83	0.91	0.92
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.58	0.76	0.83	0.66	0.83	0.89	0.72	0.85	0.90
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.90	0.94	0.94	0.93	0.95	0.94	0.94	0.96	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.88	0.93	0.93	0.91	0.93	0.92	0.92	0.93	0.93
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.87	0.91	0.92	0.88	0.92	0.92	0.89	0.92	0.92
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.75	0.85	0.88	0.77	0.89	0.90	0.78	0.89	0.90
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.93	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.89	0.94	0.93	0.93	0.94	0.94	0.93	0.95	0.94
$n = 2000$											
$\mathbb{E}n_{z_0,x_l} = 666$, $\mathbb{E}n_{z_1,x_l} = 1333$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.96	0.94	0.95	0.94	0.95	0.95	0.96	0.96	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.96	0.95	0.94	0.92	0.93	0.93	0.94	0.96	0.93
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.03	0.02	0.02	0.02	0.01	0.01	0.02	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.76	0.85	0.91	0.83	0.91	0.94	0.90	0.92	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.60	0.79	0.86	0.72	0.88	0.91	0.81	0.90	0.92
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.93	0.96	0.94	0.95	0.95	0.94	0.95	0.95	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.93	0.93	0.94	0.94	0.93	0.95	0.94	0.94	0.94
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.86	0.91	0.93	0.91	0.92	0.95	0.90	0.93	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.76	0.86	0.89	0.84	0.91	0.91	0.85	0.91	0.93
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.96	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.93	0.95	0.96	0.94	0.95	0.95	0.95	0.95	0.94

Notes: *This table shows the average bias ($\Psi_{x_l,0} - \Psi_{x_0}$), confidence interval coverage rate ($\Psi_{x_l,n} \in C_{0.95}$), and confidence ellipsoid coverage rates ($\boldsymbol{\theta}_0 \in E_{0.95}$) for overlap estimators obtained using Uniform kernels with different combinations of tuning parameters. The sample size is represented by n at the top of each panel*

along with $\mathbb{E}n_{z_k}$ *which represents the expected number of observations per instrument value. The bandwidth rate is given by* h_n

The parameters w *and* μ_1 *define different data generating processes.* μ_1 *determines the difference in means between the subdensities* $f_{Y,X|Z=z_0}$ *and* $f_{Y,X|Z=z_1}$*, and consequently, the point in which the subdensities cross. The parameter* w *determines the width of the contact set at the point in which the subdensities cross. If* w *is equal to zero, the contact set has zero measure. If* $w > 0$ *then, near the crossing points, a region in which the two densities are equal exist. When* $\mu_1 = 0$ *both subdensities are equal throughout the entire support of* Y *and thus, the contact set corresponds to the entire real line.*

Table 6: Bias, overlap coverage, and joint overlaps and CTPs coverage for sample splitting estimators. Epanechnikov kernel.

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$		
$n = 500$ $\mathbb{E}n_{z_0} = 167$, $\mathbb{E}n_{z_1} = 333$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.95	0.96	0.93	0.95	0.95	0.93	0.95	0.94	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.95	0.94	0.91	0.95	0.94	0.92	0.94	0.94	0.94
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.06	0.05	0.04	0.05	0.04	0.03	0.05	0.03	0.03
		$\Psi_{x_0} \in C_{0.95}$	0.81	0.87	0.90	0.81	0.91	0.91	0.82	0.91	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.62	0.76	0.83	0.66	0.80	0.84	0.68	0.81	0.87
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.88	0.92	0.93	0.92	0.94	0.94	0.95	0.94	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.86	0.90	0.91	0.91	0.92	0.91	0.93	0.92	0.90
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.04	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.02
		$\Psi_{x_0} \in C_{0.95}$	0.86	0.91	0.94	0.87	0.91	0.95	0.88	0.90	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.79	0.85	0.90	0.82	0.87	0.91	0.82	0.89	0.91
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.92	0.94	0.94	0.93	0.94	0.93	0.95	0.94	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.88	0.91	0.92	0.91	0.92	0.93	0.91	0.92	0.92
$n = 1000$ $\mathbb{E}n_{z_0} = 333$, $\mathbb{E}n_{z_1} = 666$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.94	0.96	0.96	0.95	0.96	0.96	0.94	0.96	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.93	0.95	0.94	0.94	0.93	0.93	0.93	0.94	0.94
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.02
		$\Psi_{x_0} \in C_{0.95}$	0.72	0.84	0.88	0.78	0.86	0.91	0.81	0.86	0.92
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.53	0.74	0.82	0.61	0.80	0.88	0.68	0.83	0.91
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.88	0.92	0.94	0.92	0.94	0.93	0.93	0.95	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.85	0.91	0.92	0.89	0.91	0.92	0.93	0.94	0.93
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02
		$\Psi_{x_0} \in C_{0.95}$	0.82	0.89	0.93	0.88	0.91	0.92	0.89	0.92	0.92
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.72	0.85	0.87	0.76	0.87	0.89	0.78	0.88	0.89
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.92	0.94	0.93	0.95	0.96	0.94	0.94	0.95	0.96
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.90	0.95	0.93	0.91	0.95	0.92	0.92	0.95	0.94
$n = 2000$ $\mathbb{E}n_{z_0,x_l} = 666$, $\mathbb{E}n_{z_1,x_l} = 1333$			$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$w = 0$	$\mu_1 = 0$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.96	0.96	0.95	0.95	0.96	0.94	0.95	0.95	0.95
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.96	0.95	0.94	0.95	0.96	0.95	0.94	0.95	0.95
	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.04	0.03	0.02	0.03	0.02	0.01	0.02	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.68	0.82	0.90	0.83	0.89	0.93	0.86	0.92	0.93
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.50	0.73	0.85	0.69	0.85	0.91	0.76	0.87	0.92
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.93	0.96	0.95	0.96	0.94	0.95	0.96	0.95	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.92	0.93	0.93	0.94	0.92	0.93	0.94	0.93	0.94
$w = 1$	$\mu_1 = 0.5$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.03	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01
		$\Psi_{x_0} \in C_{0.95}$	0.84	0.89	0.93	0.89	0.92	0.92	0.92	0.93	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.72	0.80	0.87	0.80	0.88	0.91	0.82	0.91	0.93
	$\mu_1 = \tau$	$\Psi_{x_0,n} - \Psi_{x_0}$	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\Psi_{x_0} \in C_{0.95}$	0.94	0.95	0.95	0.95	0.96	0.95	0.96	0.95	0.94
		$\boldsymbol{\theta}_0 \in E_{0.95}$	0.92	0.94	0.94	0.94	0.95	0.95	0.94	0.95	0.96

Notes: *This table shows the average bias ($\Psi_{x_l,0} - \Psi_{x_0}$), confidence interval coverage rate ($\Psi_{x_l,n} \in C_{0.95}$), and confidence ellipsoid coverage rates ($\boldsymbol{\theta}_0 \in E_{0.95}$) for overlap estimators obtained using Uniform kernels with different combinations of tuning parameters. The sample size is represented by n at the top of each panel*

along with $\mathbb{E}n_{z_k}$ *which represents the expected number of observations per instrument value. The bandwidth rate is given by* h_n

The parameters w *and* μ_1 *define different data generating processes.* μ_1 *determines the difference in means between the subdensities* $f_{Y,X|Z=z_0}$ *and* $f_{Y,X|Z=z_1}$*, and consequently, the point in which the subdensities cross. The parameter* w *determines the width of the contact set at the point in which the subdensities cross. If* w *is equal to zero, the contact set has zero measure. If* $w > 0$ *then, near the crossing points, a region in which the two densities are equal exist. When* $\mu_1 = 0$ *both subdensities are equal throughout the entire support of* Y *and thus, the contact set corresponds to the entire real line.*

C.3 Data generating processes for subsection 6.2

C.3.1 Undetectable violations of IV validity

The data generating process if formally described by the following equations:

$$\underbrace{X \sim U\{0,1,2\}}_{\textit{Random treatment}}, \qquad \underbrace{Y^0 \sim N(0,1)}_{\textit{Normal outcome}}, \qquad \underbrace{Y = \begin{cases} X + Y^0 + \tau_{und} S, & Z = 1, \\ X + Y^0, & Z = 0, \end{cases}}_{\textit{Violation of exclusion w. prob. } p(Y_0) \textit{ in right tail}},$$

$$S \mid Y^0 = y \sim \text{Bernoulli}(p(y)), \qquad p(y) = \begin{cases} 0, & y \le c, \\ y - c, & c < y < c + 1, \\ 1, & y \ge c + 1, \end{cases} \qquad c = 2 + \Phi^{-1}(0.85).$$

C.3.2 Finite-grid integration failure

The data generating process if formally described by the following equations:

$$Y = \underbrace{c(X-2)}_{\textit{Treatment effect}} + \underbrace{\tau_{di}\beta_{IV}\,(Z-1.5)\mathbf{1}\{X=1\}}_{\textit{Violation of exclusion at } X=1} + \varepsilon, \qquad \tau = 0.861, \qquad \varepsilon \sim N(0,1),$$

$$\mathbb{P}[X=0|Z=z_0] = 0.3, \quad \mathbb{P}[X=1|Z=z_0] = 0.6, \quad \mathbb{P}[X=2|Z=z_0] = 0.1,$$
$$\mathbb{P}[X=0|Z=z_1] = 0.1, \quad \mathbb{P}[X=1|Z=z_0] = 0.6, \quad \mathbb{P}[X=2|Z=z_0] = 0.3.$$

C.4 Additional simulations for subsection 6.2: Test performace with Epanechnikov and Uniform kernels.

Table 7: Test performace with Epanechnikov and uniform kernels.

Panel A: Undetectable violations of the assumptions																		
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(18)	(19)
		Uniform kernel									Epanechnikov kernel								
DGP		$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$			$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$		
$n = 500$ $\mathbb{E}n_{z_k} = 250$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$\tau_{ud} = 0$	NF	0.03	0.04	0.04	0.03	0.03	0.05	0.03	0.05	0.04	0.03	0.05	0.04	0.03	0.04	0.05	0.03	0.04	0.05
$\tau_{ud} = 0.5$	F	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.08	0.05	0.04	0.06	0.04	0.06	0.06	0.06	0.05	0.06	0.06
$\tau_{ud} = 1$	F	0.08	0.07	0.06	0.08	0.09	0.08	0.09	0.11	0.08	0.09	0.09	0.05	0.09	0.09	0.07	0.11	0.10	0.10
$\tau_{ud} = 1.5$	F	0.12	0.09	0.08	0.13	0.11	0.08	0.13	0.13	0.11	0.10	0.10	0.07	0.10	0.11	0.09	0.12	0.12	0.11
$n = 1000$ $\mathbb{E}n_{z_k} = 500$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$\tau_{ud} = 0$	NF	0.04	0.05	0.06	0.05	0.04	0.06	0.04	0.05	0.05	0.04	0.03	0.05	0.05	0.03	0.06	0.05	0.04	0.04
$\tau_{ud} = 0.5$	F	0.08	0.08	0.09	0.07	0.09	0.08	0.07	0.07	0.09	0.09	0.08	0.08	0.09	0.07	0.08	0.08	0.08	0.07
$\tau_{ud} = 1$	F	0.14	0.14	0.14	0.18	0.16	0.14	0.15	0.14	0.14	0.17	0.14	0.13	0.17	0.17	0.13	0.16	0.17	0.14
$\tau_{ud} = 1.5$	F	0.18	0.19	0.15	0.23	0.20	0.16	0.22	0.19	0.15	0.23	0.19	0.18	0.22	0.21	0.16	0.22	0.21	0.16
$n = 2000$ $\mathbb{E}n_{z_k} = 1000$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$\tau_{ud} = 0$	NF	0.03	0.04	0.04	0.05	0.04	0.03	0.05	0.03	0.04	0.06	0.05	0.03	0.04	0.05	0.04	0.05	0.04	0.04
$\tau_{ud} = 0.5$	F	0.12	0.11	0.10	0.16	0.14	0.10	0.16	0.14	0.09	0.13	0.09	0.07	0.13	0.13	0.08	0.14	0.11	0.08
$\tau_{ud} = 1$	F	0.28	0.25	0.18	0.35	0.26	0.21	0.36	0.26	0.19	0.30	0.22	0.17	0.31	0.27	0.20	0.36	0.28	0.17
$\tau_{ud} = 1.5$	F	0.38	0.31	0.21	0.47	0.36	0.27	0.48	0.35	0.24	0.40	0.32	0.22	0.45	0.37	0.25	0.47	0.36	0.26
Panel B: Discrete integration failure																			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(18)	(19)
		Uniform kernel									Epanechnikov kernel								
DGP		$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$			$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$		
$n = 500$ $\mathbb{E}n_{z_k} = 250$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$\tau_{di} = 0$	NF	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.02	0.01
$\tau_{di} = 0.5$	F	0.05	0.06	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.04	0.04	0.04	0.04	0.06	0.06	0.05	0.07	0.05
$\tau_{di} = 1$	F	0.15	0.14	0.12	0.18	0.16	0.13	0.19	0.16	0.14	0.14	0.12	0.10	0.16	0.15	0.12	0.18	0.16	0.13
$n = 1000$ $\mathbb{E}n_{z_k} = 500$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$\tau_{di} = 0$	NF	0.00	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.02	0.02	0.01	0.02	0.01
$\tau_{di} = 0.5$	F	0.06	0.07	0.07	0.09	0.09	0.07	0.10	0.10	0.08	0.05	0.06	0.06	0.08	0.09	0.08	0.10	0.10	0.08
$\tau_{di} = 1$	F	0.32	0.28	0.19	0.37	0.30	0.21	0.39	0.31	0.21	0.30	0.24	0.17	0.35	0.29	0.20	0.38	0.30	0.21
$n = 2000$ $\mathbb{E}n_{z_k} = 1000$		$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$
$\tau_{di} = 0$	NF	0.01	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02
$\tau_{di} = 0.5$	F	0.17	0.14	0.10	0.20	0.16	0.12	0.22	0.15	0.12	0.14	0.13	0.10	0.17	0.16	0.11	0.20	0.17	0.12
$\tau_{di} = 1$	F	0.56	0.47	0.33	0.60	0.51	0.36	0.65	0.50	0.36	0.57	0.48	0.34	0.60	0.49	0.36	0.63	0.53	0.37

Notes: *This table shows the rejection rate across 500 simulations for different joint tests of instrument validity and monotonicity. The first 9 columns correspond to versions of the test I propose using uniform kernels and different combinations tuning parameters. The last 9 use Epanechnikov kernels instead. The parameter c_{prop} for $prop \in \{1/3, 1/2, 2/3\}$ denotes that a proportion prop of the data was used to estimate the subdensities and classifier while a fraction $1 - prop$ was used to evaluate the empirical loss of the estimated classifier and to estimate conditional treatment probabilities. h_n denotes the bandwidth rate. The sample size is represented by n at the top of each panel along with $\mathbb{E}n_{z_k}$ which represents the expected number of observations per instrument value. The acronym NF stands for not-falsifiable and indicates data generating processes for which the null hypothesis of instrument validity and monotonicity cannot be jointly falsified. On the other hand, F indicates data generating processes for which the null is falsifiable.*

D Supporting simulations for section 7.

D.1 Two dimensional outcomes

I first assess the performance of the test when the outcomes of interest are more than one. The theory from section 5 is not restricted to one dimensional outcomes, both, the classification based estimator for subdensity overlaps, and the testing procedure based on Cox and Shi (2023) apply to the case in which the outcome is multi-dimensional. In particular, I conduct a set of simulations in which the outcome is two-dimensional, that is $Y = (Y_1, Y_2)$.

In these simulations, the instrument, treatment and first outcome follow the same data generating process than in the simulations designed to compare my test to tests based on finite-grid integration. The outcome Y_2, follows a standard normal distribution and is independent from both X, Z and Y_1. The independence of Y_2 implies that it does not contribute to the falsifiability of the model. Under the data generating process, instrument independence, exclusion and monotonicity are falsifiable only from the joint distribution of Y_1, X and Z. Having an additional outcome, however, complicates the overlap estimation problem since now it is two-dimensional.

The following table shows the rejection rates over 500 simulations. I find that discretization based tests have less power to detect violations of the model's assumptions than overlap based tests. As in the one dimensional case, estimators that use half the data for estimating the classifier, and half the data for evaluating its loss outperform other sample splitting rules.

Table 8: Test performace with two-dimensional outcomes.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	KDE overlaps									KR Discretization				
DGP	$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$							
$n = 500$ $\mathbb{E}n_{z_k} = 250$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$ NF	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0.5$ F	0.07	0.11	0.10	0.12	0.20	0.18	0.15	0.24	0.20	0.08	0.16	0.05	0.04	0.01
$\tau_{di} = 1$ F	0.07	0.11	0.10	0.12	0.20	0.18	0.15	0.24	0.20	0.08	0.16	0.05	0.04	0.01
$n = 1000$ $\mathbb{E}n_{z_k} = 500$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$ NF	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0.5$ F	0.15	0.17	0.16	0.13	0.37	0.31	0.24	0.43	0.38	0.13	0.22	0.12	0.13	0.02
$\tau_{di} = 1$ F	0.15	0.17	0.16	0.13	0.37	0.31	0.24	0.43	0.38	0.13	0.22	0.12	0.13	0.02
$n = 2000$ $\mathbb{E}n_{z_k} = 1000$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$ NF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0.5$ F	0.34	0.44	0.40	0.30	0.68	0.60	0.50	0.80	0.69	0.31	0.30	0.60	0.49	0.29
$\tau_{di} = 1$ F	0.34	0.44	0.40	0.30	0.68	0.60	0.50	0.80	0.69	0.31	0.30	0.60	0.49	0.29

Notes: *This table shows the rejection rate across 500 simulations for different joint tests of instrument validity and monotonicity. The first 9 columns correspond to versions of the test I propose with different tuning parameters. The last 5 correspond to implementations of the test proposed by Kwon and Roth (2024) with different cardinalities of the discretizing partition. The parameter c_{prop} for $prop \in \{1/3, 1/2, 2/3\}$ denotes that a proportion prop of the data was used to estimate the subdensities and classifier while a fraction $1-prop$ was used to evaluate the empirical loss of the estimated classifier and to estimate conditional treatment probabilities. h_n denotes the bandwidth rate. For the discretization based tests, b_{np} denotes that the support of the outcome was partitioned into np bins to implemet the test. The parameter N_{st} indicates the number of subtreatments in which observations realizing each different treatments were randomly split. The sample size is represented by n at the top of each panel along with $\mathbb{E}n_{z_k}$ which represents the expected number of observations per instrument value. The acronym NF stands for not-falsifiable and indicates data generating processes for which the null hypothesis of instrument validity and monotonicity cannot be jointly falsified. On the other hand, F indicates data generating processes for which the null is falsifiable.*

D.2 Low-frequency treatments

In the application to the OHE, treatment is defined as the number of months that a person was covered by insurance during the study, thus, there are 20 distinct treatment values. This is a potential problem since, even if the sample size is large, the number of observations that realize some particular treatment values is potentially small. For instance, no more than 100 observations realize the treatment values $X = x_2$, $X = x_3$, $X = x_4$ and $X = x_5$. This implies that estimators associated to these low-frequency treatments might not be precise and in turn affect the power or size of the test. This is a problem for the test I propose since the subdensity overlap estimators rely on within-treatment kernel density estimation, and also for discretization-based tests since cells associated with these particular treatments are noisily estimated.

To assess how the performance of both tests is affected by this proble, I conduct a series of simulations in which I artificially create multiple low-frequency treatments. In particular, I generate data from the same data generating process than in the simulations designed to compare my test to tests based on finite-grid integration, but then, randomly split all observations into N_{st} different subtreatments. Thus, I end up with $3N_{st}$ different treatments. Then, I test the instrument validity assumptions under the additional restriction that the instrument cannot induce treatment changes within subtreatment categories [21]. Observe that, at the population level, the model with split treatments is only falsifiable when the original, monotone, three treatments model is falsifiable.

The following table shows the rejection rates obtained across 500 simulations. I find that my test has power against violations of the null, even when some treatments occur with very low probability. On the other hand, I find that the power of discretization based tests is severely diminished by the presence of multiple low-frequency treatments.

[21] For examle, if $N_{st} = 2$, I split treatment x_0 into treatments x_{00} and x_{01}, treatment x_1 into x_{10} and x_{11}, and treatment x_{20} and x_{21}. Then, impose that no changes from $x_{kk'}$ to $x_{k''k'''}$ are allowed unless $k > k''$.

Table 9: Test robustness to low-frequency treatments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	KDE overlaps									KR Discretization				
DGP	$h_n = n^{-1/3}$			$h_n = n^{-1/4}$			$h_n = n^{-1/5}$							
$n = 500$ $\mathbb{E}n_{z_k} = 250$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$ $N_{st} = 2$ NF	0.01	0.01	0.02	0.01	0.00	0.01	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00
	0.04	0.03	0.06	0.05	0.04	0.05	0.06	0.05	0.06	0.00	0.00	0.00	0.00	0.00
	0.11	0.15	0.14	0.12	0.16	0.14	0.12	0.17	0.14	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0$ $N_{st} = 3$ F	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00
	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00
	0.04	0.03	0.03	0.04	0.04	0.04	0.05	0.03	0.05	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 1$ $N_{st} = 4$ F	0.02	0.00	0.02	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	0.03	0.01	0.03	0.04	0.02	0.02	0.04	0.02	0.02	0.00	0.00	0.00	0.00	0.00
	0.06	0.08	0.09	0.08	0.09	0.10	0.08	0.10	0.10	0.00	0.00	0.00	0.00	0.00
$n = 1000$ $\mathbb{E}n_{z_k} = 500$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$ $N_{st} = 2$ NF	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00
	0.07	0.08	0.08	0.08	0.11	0.08	0.07	0.10	0.08	0.00	0.00	0.00	0.00	0.00
	0.19	0.35	0.27	0.18	0.34	0.27	0.20	0.34	0.27	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0$ $N_{st} = 3$ F	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	0.07	0.05	0.06	0.08	0.06	0.08	0.07	0.07	0.07	0.00	0.00	0.00	0.00	0.00
	0.17	0.28	0.24	0.18	0.29	0.26	0.20	0.30	0.25	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 1$ $N_{st} = 4$ F	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	0.04	0.04	0.04	0.04	0.05	0.05	0.03	0.05	0.05	0.00	0.00	0.00	0.00	0.00
	0.14	0.20	0.19	0.17	0.22	0.21	0.17	0.25	0.21	0.00	0.00	0.00	0.00	0.00
$n = 2000$ $\mathbb{E}n_{z_k} = 1000$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	$c_{\frac{1}{3}}$	$c_{\frac{1}{2}}$	$c_{\frac{2}{3}}$	b_3	b_5	b_{15}	b_{25}	b_{50}
$\tau_{di} = 0$ $N_{st} = 2$ NF	0.02	0.01	0.02	0.01	0.01	0.02	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00
	0.09	0.15	0.13	0.11	0.18	0.15	0.11	0.19	0.15	0.00	0.00	0.00	0.00	0.00
	0.33	0.58	0.48	0.34	0.62	0.50	0.36	0.61	0.49	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 0$ $N_{st} = 3$ F	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00
	0.09	0.13	0.12	0.09	0.15	0.15	0.11	0.16	0.15	0.00	0.00	0.00	0.00	0.00
	0.33	0.50	0.44	0.31	0.55	0.44	0.32	0.59	0.48	0.00	0.00	0.00	0.00	0.00
$\tau_{di} = 1$ $N_{st} = 4$ F	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	0.07	0.11	0.12	0.09	0.13	0.13	0.08	0.15	0.14	0.00	0.00	0.00	0.00	0.00
	0.30	0.46	0.44	0.31	0.54	0.44	0.29	0.54	0.46	0.00	0.00	0.00	0.00	0.00

Notes: *This table shows the rejection rate across 500 simulations for different joint tests of instrument validity and monotonicity. The first 9 columns correspond to versions of the test I propose with different tuning parameters. The last 5 correspond to implementations of the test proposed by Kwon and Roth (2024) with different cardinalities of the discretizing partition. The parameter c_{prop} for $prop \in \{1/3, 1/2, 2/3\}$ denotes that a proportion prop of the data was used to estimate the subdensities and classifier while a fraction $1-prop$ was used to evaluate the empirical loss of the estimated classifier and to estimate conditional treatment probabilities. h_n denotes the bandwidth rate. For the discretization based tests, b_{np} denotes that the support of the outcome was partitioned into np bins to implemet the test. The sample size is represented by n at the top of each panel along with $\mathbb{E}n_{z_k}$ which represents the expected number of observations per instrument value. The acronym NF stands for not-falsifiable and indicates data generating processes for which the null hypothesis of instrument validity and monotonicity cannot be jointly falsified. On the other hand, F indicates data generating processes for which the null is falsifiable.*