Sharp Instruments for Classifying Compliers and Generalizing Causal Effects

Written by: Edward H. Kennedy[1], Sivaraman Balakrishnan[1], Max G'Sell[1], and Reproduced by: Ethan Ashby[2]

[1] *Department of Statistics and Computer Science, Carnegie Mellon University*
[2] *Department of Biostatistics, University of Washington*

1 Abstract

Instrumental variable (IV) methods are an appealing solution to the problem of identifying causal effects under unmeasured confounding. However, even when in possession of a valid instrument, IV methods require additional unverifiable assumptions to point identify causal effects. The popular assumption of monotonicity identifies a "local" causal effect in the a subset of the population known as compliers, or those who would only take the treatment when encouraged by the instrument. Since the complier subgroup is unobserved, local effects have been criticized on the grounds of poor interpretation and limited relevance to science and policy. We propose that, in some cases, it is possible to predict who compliers are and obtain bounds on causal effects in observable subgroups. We propose methods for complier prediction and study their estimation errors. We derive nonparametric estimators of bounds on the average causal effect in observable subgroups and study their asymptotic properties. We introduce a new summary measure of instrument quality termed "sharpness" which reflects the variation in compliance probabilities over the distribution of observed covariates. Sharp instruments improve the interpretation of complier effects by (a) enabling accurate prediction of compliers from observed covariates and (b) yielding narrow bounds on the average causal effect in observable subgroups. We propose a nonparametric estimator of sharpness and show it is efficient under conditions. We explore the finite sample properties of the proposed approaches via simulation and apply the methods to a field experiment to assess the effect of door-to-door canvassing on voter turnout.

Key words: causal inference, instrumental variables, monotonicity, principal stratification, compliance.

2 Introduction

Instrumental Variable (IV) methods (J. D. Angrist, Guido W. Imbens, and Rubin 1996) are a popular framework for identifying causal effects under unmeasured confounding. Informally, an instrumental variable is a variable associated with treatment, associated with the outcome only through the treatment, and is itself unconfounded (see Figure 1). IV methods are applicable in a broad array of settings including randomized and natural experiments. A canonical use case for IV methods is a randomized unblinded trial with noncompliance, where the goal is to estimate the effect of treatment *received* on outcome despite participants not complying with their randomized treatment assignment. Here, there may exist latent factors associated with both propensity to take treatment and the outcome, a phenomenon known as unmeasured confounding. Unmeasured confounding will distort estimates of the causal effect of treatment received on the outcome. IV methods address the unmeasured confounding problem by considering treatment *assignment*, which is unconfounded by randomization, as a starting point for identifying and estimating the causal effect of interest. Other use cases of IV methods exploit natural randomness in quantities such as treatment preference, access (distance or time), or inheritance of genetic factors to identify causal effects (Hernán and J. M. Robins 2020). In settings where the assumption of no unmeasured confounding may be infeasible due to the presence of many confounders, inadequate data collection, or measurement error, IV methods shift the burden of the unconfoundedness assumption from the treatment to the presumed randomized instrument.

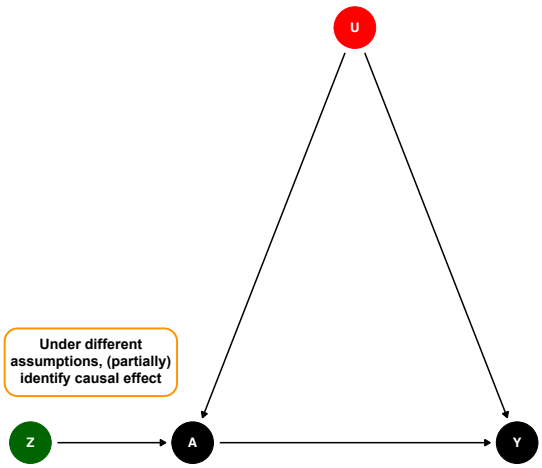


Figure 1: A simple causal diagram illustrating an instrumental variable Z alongside treatment A, outcome Y, and unmeasured confounders U.

However, instrumental variable methods raise complications that do not arise in studies with unconfounded treatments. A valid instrument is insufficient to identify a causal effect of interest. Several approaches to causal effect identification using IV methods are summarized in Table 1 below.

One option is to pursue bounds on the average treatment effect. Pursuing a partial identification approach is appealing in that it focuses on bounding the global average causal effect, a very relevant estimand, and does not require additional assumptions. Under only the counterfactual consistency assumption, bounds can be obtained where one potential outcome is replaced by the observed, factual outcome and the other unobserved potential outcome is replaced by its extreme values ($\{0, 1\}$). Manski-bounds that leverage assumptions on the validity of the instrumental variable can achieve narrower bounds than bounds that make minimal assumptions (Robins 1989; Manski 1990). Typically, narrower bounds are achieved when additional assumptions are made (see Swanson et al. (2018) for a review). Balke and Pearl (1997) famously used linear programming techniques to achieve and prove tight bounds on the average treatment effect in randomized trials with noncompliance and no covariate information. Recent work by Levis et al. (2023) extended the approach to "covariate assisted bounds", which are relevant to observational studies with confounders of the instrument or randomized trials with measured baseline covariates. However, while partial identification accommodate flexible assumption burdens, they often yield bounds on the average causal effect that are so wide that they are uninformative for science and policy.

In many cases, instead of bounding the global average treatment effect, investigators impose additional structural assumptions to point identify a causal effect. An alternative involves restricting the heterogeneity of treatment effects between individuals in a causal structural model. In the most extreme version, one may assume treatment effect homogeneity, which assumes that the effect of treatment on outcome is constant across individuals (Hernán and J. M. Robins 2020). Homogeneity is scientifically implausible for most treatments and outcomes. One may consider weaker versions of the homogeneity assumption, including no interaction between treatment and unmeasured confounders, that the instrument-treatment association is constant across levels of confounders, and parametric assumptions (Hernán and J. M. Robins 2020). However, each of these assumptions *a priori* imposes structure on the effect of interest, relying on functional form knowledge not typically known in practice.

A third strategy assumes monotonicity, which excludes the possibility that the instrument can "discourage" someone to take the treatment. This rules out the principal strata of "defiers" in the study population. In many cases, this is a reasonable assumption. For example, in a randomized, unblinded trial with noncompliance, monotonicity means that no participants would defy their randomization assignment — would refuse treatment if randomized to treatment, yet would take treatment if randomized to control. Monotonicity is also an appealing assumption as it permits nonparametric identification of a causal effect. However, monotonicity identifies a local average treatment effect in a subpopulation of "compliers", subjects who would only take the treatment when encouraged by the instrument (J. D. Angrist, Guido W. Imbens, and Rubin 1996). Average treatment effects among compliers have been crit-

	Bound ATE	Monotonicity	Restrict Trt Effect Heterogeneity
Additional Assumptions (Burden)	None Required (↓)	Monotonicity (↔)	Restrict Trt Effect Heterogeneity (↑)
Point Identification (Precision)	No (↓)	Yes (↑)	Yes (↑)
Estimand (Generalizability)	ATE in Population (↑)	ATE among Compliers (↓)	ATE in Population (↑)

Table 1: Summary of different IV approaches based on assumptions, whether they point identify causal effects, and which estimands they identify. Also summarized in parentheses are the burden of assumptions, the precision of the estimates, and the generalizability of the estimand.

icized in the econometrics and statistics literatures due to lack of interpretability, since it corresponds to an effect in an unknown subgroup (Deaton 2010; Heckman and Urzua 2009). A causal effect among compliers is of questionable scientific or policy relevance, as it is unclear who the compliers are. Also the interpretation of the complier causal effect is instrument specific, and thus lacks external validity to changes in the treatment induced by other instruments. Others have defended complier effects for the plausibility of the monotonicity assumption and that they illustrate the inherent fundamental limitations of causal analyses in nonrandomized studies subject to unmeasured confounding (Guido W Imbens 2010).

Previous research has focused on improving the interpretation of the complier effect by characterizing the complier population. The weighting theorem of Abadie (2003) identified any moment-based summary of the complier population. Work by J. Angrist and Fernandez-Val (2010) assessed the stability of IV estimates across different instruments, and imposed a restriction on treatment heterogeneity that permitted extrapolation of instrument-specific complier effects to other treated or complier populations. Singh and Sun (2022) proposed semiparametric tests to compare the covariate distribution of the complier population between instruments or to the broader study population. However, these approaches fail to characterize the complier probabilities of individual participants.

An approach to characterizing compliance status for individual units is the *compliance score* (Follmann 2000), which measures how well the instrument predicts observed treatment, conditional on pre-treatment covariates. Under a valid instrument and monotonicity, the compliance score is also the subject-level probability of being a complier conditional on observed covariates. Joffe, Ten Have, and Brensinger (2003) proposed using the group-level compliance score as a regressor in models for treatment efficacy in randomized trials with noncompliance, but note that valid estimation of average treatment effect only holds for regression models with identity link. Roy, Hogan, and Marcus (2008) proposed using compliance-predictive covariates to estimate causal effects in principal strata, although their approach required correct specification of two compliance models (one for each value of the instrument) and an association model linking the two. Aronow and Carnegie (2013) proposed weighting by the inverse of the compliance score to generalize the ATE among compliers to the global ATE, however, the approach required parametric assumptions to model the probability of compliance, along with a strong assumption of ignorability of complier status on ATE given covariates $\mathbf{X} = \mathbf{x}$. Liu et al. (2022) proposed using flexible machine learning methods to predict compliance scores and classify likely compliers based on pre-treatment covariates, with the goal of increasing statistical power and reducing sample size of IV studies.

3 Methods

The following section considers whether interpretation of complier average treatment effects identified under monotonicity can be improved. Kennedy, Balakrishnan, and G'Sell (2020) conclude that interpretation can be improved for "sharp" causal instruments, which (a) enable accurate prediction of compliers based on observed covariates and (b) produce tight bounds on average treatment effects in covariate-defined subgroups of predicted compliers. At a high level, sharp instruments produce compliance scores that vary over the distribution of observed covariates. The authors illustrate that sharpness is an fundamental aspect of instrument quality totally separate from the commonly reported summary of instrument strength.

This section is organized as follows. We introduce relevant notation and present assumptions under which the average treatment effect within the subpopulation of compliers is identified. Subsequently,

we will introduce the first property of sharp instruments: accurate prediction of compliers using observed covariates. We introduce three classification rules for predicting complier status, demonstrate classification error optimality in their respective classes, and discuss estimation of the rules. Then, we discuss the second property of sharp instruments: informative bounds on causal effects in identifiable subgroups. We demonstrate that the tightest bounds are generated by the quantile-threshold classifiers based on the compliance score. We also discuss estimating these bounds and highlight their asymptotic properties. Lastly, we propose the summary measure of instrument sharpness, which unifies the two previously described elements of sharp instruments. We discuss how sharpness is fundamental measure of instrumental variable quality.

3.1 Notation and definitions

Consider the case where treatment and outcome are both binary. Let $A \in \{0,1\}$ denote the binary treatment, and $Y \in \{0,1\}$ denote the binary outcome of interest. Let $Z \in \{0,1\}$ denote a binary instrumental random variable. Let $\mathbf{X} \subset \mathbb{R}^p$ be a collection of covariates defined for each participant. Suppose we observe n independent and identically distributed (i.i.d.) draws from a distribution $\{\mathbf{O}_1, \ldots, \mathbf{O}_n\} \sim \mathbb{P}$ with

$$\mathbf{O} = (\mathbf{X}, Z, A, Y)$$

For each participant, we define their potential outcomes, $\{Y^{a=0}, Y^{a=1}\}$, as the possibly counter-to-fact outcome had we set their treatment equal to $A = 0$ and $A = 1$ respectively. The individual causal effects are defined as the difference between the potential outcomes, $Y^{a=1} - Y^{a=0}$. Similarly, for each participant, we define the potential treatments $\{A^{z=0}, A^{z=1}\}$, as the possibly counter to fact treatments had we set their instrument equal to $Z = 0$ and $Z = 1$ respectively. Define $Y^z := Y^{zA^z}$ to be the potential outcome had we intervened on the *instrument* by setting $Z = z$, and therefore induced the potential treatment A^z.

Under causal assumptions of deterministic potential treatments, no treatment interference, and the absence of multiple versions of treatment, we can stably partition the collection of study participants into four categories according to the values of their potential treatments, $\{A^{z=0}, A^{z=1}\}$: always takers (for whom $(A^{z=0}, A^{z=1}) = (1,1)$), never takers (for whom $(A^{z=0}, A^{z=1}) = (0,0)$), compliers (for whom $(A^{z=0}, A^{z=1}) = (0,1)$), and defiers (for whom $(A^{z=0}, A^{z=1}) = (1,0)$). We interchangeably refer to these groupings as compliance categories or principal strata. An visual illustration of these compliance categories is offered in Figure 2. Let $C := \mathbb{1}(A^{z=1} > A^{z=0})$ be the latent variable denoting whether a participant is a complier, or only takes treatment when encouraged by their instrument.

Let $\pi_z(\mathbf{x}) := \mathbb{P}(Z = z|\mathbf{X} = \mathbf{x})$ denote the instrumental propensity score, which is a function of the observed covariates, $\mathbf{X} = \mathbf{x}$. Let $\lambda_z(\mathbf{x}) := \mathbb{P}(A = 1|\mathbf{X} = \mathbf{x}, Z = z)$ denote the treatment propensity score, a function of the observed covariates $\mathbf{X} = \mathbf{x}$ and observed value of the instrument $Z = z$.

Define the *compliance score* as the difference in treatment propensity scores under the two values of the instrument:

$$\begin{aligned} \gamma(\mathbf{x}) &:= \lambda_1(\mathbf{x}) - \lambda_0(\mathbf{x}) \\ &\equiv P(A = 1|\mathbf{X} = \mathbf{x}, Z = 1) - P(A = 1|\mathbf{X} = \mathbf{x}, Z = 0) \end{aligned} \tag{1}$$

The compliance score measures how strongly the instrument Z encourages treatment $A = 1$ conditional on the observed covariates $\mathbf{X} = \mathbf{x}$. Under assumptions presented in the ensuing section, the compliance score can be connected to the probability of being a complier conditional on observed covariates $\mathbf{X} = \mathbf{x}$ (see Result 3.2).

3.2 Assumptions and identification

We require two standard causal assumptions.

Assumption 1 (Counterfactual Consistency of Treatment and Outcome):

$$A = \begin{cases} A^{z=1} & \text{when } Z = 1 \\ A^{z=0} & \text{when } Z = 0 \end{cases}$$

$$Y = \begin{cases} Y^{z=1} & \text{when } Z = 1 \\ Y^{z=0} & \text{when } Z = 0 \end{cases}$$

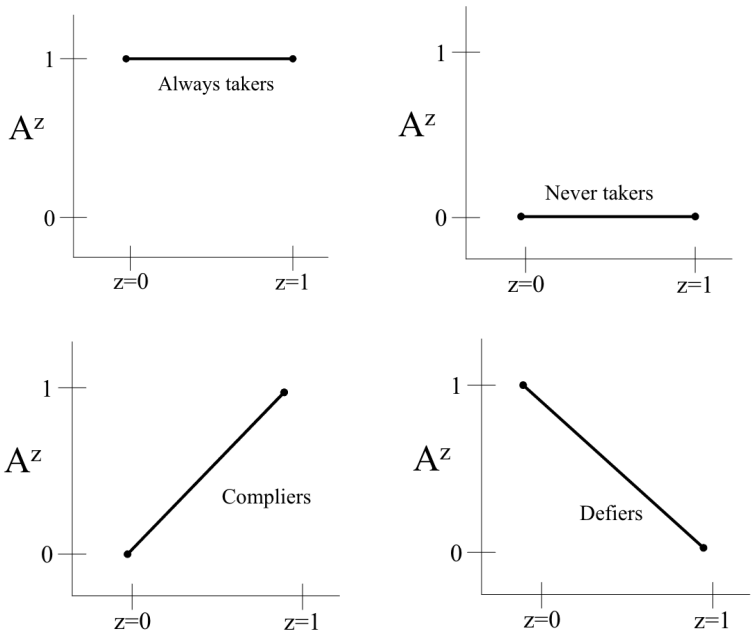


Figure 2: Definitions of compliance categories or principal strata according to instrument values $z = 0, 1$ and potential treatments $A^z = 0, 1$. Figure adapted from Hernán and J. M. Robins (2020).

In words, Assumption 1 ensures that when the instrument Z takes value z, the observed treatments (A) and outcomes (Y) equal the potential treatment or outcome had we intervened on Z and set its value to z.

Assumption 2 (Instrument Positivity):

$$\mathbb{P}(\epsilon \leq \pi_z(\mathbf{X}) \leq 1 - \epsilon) = 1 \quad \text{for some } \epsilon > 0$$

In words, Assumption 2 requires that among participants with covariates $\mathbf{X}$, the probability of having instrument value $Z = 0$ and $Z = 1$ are both bounded away from 0 and 1.

The following three assumptions are the main instrumental variable (IV) conditions. The main IV conditions ensure the validity of a given instrument.

Assumption 3 (Relevance):

$$\mathbb{P}(A|Z = 1) - \mathbb{P}(A|Z = 0) > 0$$

In words, Assumption 3 requires that the instrument Z is associated with treatment A. In the DAG shown in Figure 1, Assumption 3 requires that there be an arrow from Z to A. Note that this is the only empirically verifiable assumption listed, and can be assessed by replacing $\mathbb{P}$ by its empirical plug-in estimator $\mathbb{P}_n$. In practice, instruments quality is judged by its *strength*, the plug-in estimate of the left hand side above. Strength measures proportion of the people in the population who only take treatment when encouraged by the instrument, i.e., the proportion of the study population who are compliers.

Assumption 4 (Exclusion Restriction):

$$Y^{za} = Y^a$$

In words, Assumption 4 requires that the effect of instrument Z on the outcome Y is only exerted through its potential effect on the treatment A. In Figure 1, Assumption 4 means that there are no arrows from the instrument Z to Y that do not flow through A on the causal diagram.

Assumption 5 (Unconfounded IV):

$$Z \perp (A^z, Y^z)|\mathbf{X}$$

In words, Assumption 5 requires that the instrument Z can be considered randomly assigned conditional on covariates $\mathbf{X}$, and is therefore outside the confounding structure on A and Y. In the DAG in Figure 1, Assumption 5 means that the instrument Z and outcome Y do not share common causes outside of the recorded $\mathbf{X}$.

Unfortunately, Assumptions 1-5 are insufficient to identify the average causal effect. There are three strategies typically pursued to estimate causal effects using instrument variables: (1) partial identification of the average treatment effect in the global population, (2) assuming restrictions on treatment effect heterogeneity and obtaining a point estimate of the average causal effect in the global population, and (3) assuming strong monotonicity and obtaining a point estimate in a subgroup of compliers. The assumptions, partial/point identification results, estimands, and generalizability of the estimands under each approach is summarised concisely in Table 1.

For the remainder of this paper, we elect the third strategy and choose to make the assumption of strong monotonicity.

Assumption 6 (Strong Monotonicity):
Let $C = 1$ be an indicator variable denoting membership in the principal strata of compliers (i.e., $\{A^{z=0}, A^{z=1}\} = \{0, 1\}$)

$$\mathbb{P}(A^{z=1} < A^{z=0}) = 0 \text{ and } \mathbb{P}(C = 1) \geq \epsilon > 0$$

In words, Assumption 6 precludes the existence of the defier compliance category in the study population, and requires that a non-zero fraction of the study population is are compliers. Ruling out the principal strata of defiers assumes that no participants would only take treatment when discouraged by the instrument, and would not take treatment when encouraged.

Under Assumptions 1-6, the usual IV estimand equals the average treatment effect among the subgroup of *compliers*.

Result 3.1 (Identification of Complier ATE under Monotonicity)**.** *Under assumptions 1-6,*

$$\frac{\mathbb{E}[\mathbb{E}(Y|Z=1,\boldsymbol{X}) - \mathbb{E}(Y|Z=0,\boldsymbol{X})]}{\mathbb{E}[\mathbb{E}(A|Z=1,\boldsymbol{X}) - \mathbb{E}(A|Z=0,\boldsymbol{X})]} = \mathbb{E}\left[Y^{a=1} - Y^{a=0}|C=1\right]$$

A proof of these identification result is provided in the Appendix. Intuitively, the numerator of the IV estimand is the effect of Z on Y (also known as the intention to treat (ITT) effect). The ITT effect can be considered as a weighted average of the ITT effects in each of the principal strata. However, monotonicity assumes that the set of defiers is empty, so the defiers will not contribute to the estimand. And the effect of Z on Y is zero in the always and never takers, since by Assumption 4, the effect of Z on Y is entirely mediated through A, and A is fixed in always and never takers under different values of the instrument. Thus, the numerator of the IV estimand is the ITT effect among the compliers. Since the instrument Z and treatment A are equal for compliers, the ITT effect equals the treatment effect of A on Y among the compliers.

Assumptions 1,2,5, and 6 also give the instrument *strength* as a function of the compliance score, $\gamma(\mathbf{x})$:

Result 3.2 (Instrument Strength and Compliance score)**.**

$$\begin{aligned}\mathbb{P}(C=1|\boldsymbol{X}=\boldsymbol{x}) &= \gamma(\boldsymbol{x}) \\ \mu = \mathbb{P}(C=1) &= \mathbb{E}(\gamma(\boldsymbol{x}))\end{aligned}$$

Where $\gamma(\boldsymbol{x})$ is the compliance score. We refer to μ is the strength of the instrument Z.

A proof of this result is provided in the Appendix. In words, the strength of an instrument is defined as the average probability of compliance ($C = 1$) marginalized over the distribution of covariates $\mathbf{X}$ in the population.

3.3 Complier classification

Kennedy, Balakrishnan, and G'Sell (2020) propose the idea that instruments, regardless of their marginal strength, may yield subgroups of likely compliers defined observed covariates values. Instruments that

produce variable compliance probabilities over the distribution of observed covariates are deemed *sharp instruments*. Sharp instruments can aid the prediction of compliers ($C = 1$) using observed covariate information ($\mathbf{X} = \mathbf{x}$). This motivates the need to develop rules that classify likely compliers on the basis of observed covariates.

3.3.1 Complier classification rules

Rather than considering raw compliance scores, $\gamma(\mathbf{x})$, predicting complier status is a worthwhile task for a few reasons. First, complier classification should be at worst as difficult of as predicting the compliance scores since miscalibrated compliance scores could still yield correct classifications. Second, complier classification generates a list of participants who are likely compliers, which may improve interpretation of complier effects.

For a given classification rule mapping the support of the observed covariates to a binary prediction of compliance status, $h : \mathcal{X} \to \{0, 1\}$, we can judge the quality of the classifier using the classification error, $\mathcal{E}(h) = \mathbb{P}(h \neq C)$.

Result 3.3 (Classification Error)**.** *For a given classification rule, $h : \mathcal{X} \to \{0, 1\}$, the classification error $\mathcal{E}(h) = \mathbb{P}(h \neq C)$ is identified under assumptions 1,2,5, and 6.*

$$\mathcal{E}(h) = \mathbb{E}[\gamma(\boldsymbol{X})(1 - h(\boldsymbol{X})) + (1 - \gamma(\boldsymbol{X}))h(\boldsymbol{X})] \tag{2}$$

A proof can be located in the appendix. Note that the expectation is implicity over both the randomness in the covariates, $\mathcal{X}$, and the randomness in the classification decision. In words, the classification error of a classification rule is a summary of when the compliance score, $\gamma(\mathbf{X})$, and classification rule, $h(\mathbf{X})$, disagree. When the classification error is small, the compliance score and predicted complier status are highly correlated, meaning that the rule can classify compliers accurately.

Minimal classification error of h is one criterion we may use to select a classifier. However, minimizing classification error may not be the only criterion we wish a classification rule for complier status to satisfy. It is possible that the Bayes classifier (shown in Equation 3), while optimal with respect to classification error, may have the undesirable property of predicting an empty set of compliers if all compliance scores $\gamma(\mathbf{X}) < 1/2$. Thus, other criteria we may desire a classifier to satisfy are the properties of *strength-calibration* and *distribution-matching*.

Property 3.1 (Strength-calibration)**.** *A classification rule $h : \mathcal{X} \to \{0, 1\}$ is strength-calibrated if:*

$$\mathbb{P}(h(\boldsymbol{X}) = 1) = \mathbb{P}(C = 1)$$

In words, strength-calibrated rules ensure that the set of compliers predicted by h is, on average, equal to the size of the true set of compliers ($C = 1$). Strength-calibration is a desirable property in a classification rule, since strength is such a fundamental summary of instrumental variable quality. One may be willing to sacrifice some classification error to ensure that their classifier h produces a set of predicted compliers equal in size to the set of true compliers. A strength-calibrated classifier avoids the situation where an empty set of predicted compliers is produced when $\gamma(\mathbf{X}) < 1/2$ for all units.

Property 3.2 (Distribution-matching)**.** *A classification rule $h : \mathcal{X} \to \{0, 1\}$ is distribution-matched if:*

$$\mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x} \mid h(\boldsymbol{X}) = 1) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x} \mid C = 1) \quad \forall \boldsymbol{x}$$

In words, a distribution-matched rule ensures that the distribution of covariates is equal between predicted and true compliers. This allows practitioners to inspect the distribution of observed covariates or estimate complier characteristics for the predicted compliers ($h = 1$) to learn something about the population of true compliers ($C = 1$). This is an alternative to the identification results and weighting approach developed by Abadie (2003) for estimating complier characteristics.

The authors present three candidate rules which are optimal with respect to classification error within their respective classes. Proofs of optimality of each estimator can be located in the Appendix.

Theorem 3.4 (Classification-error optimality of three classification rules)**.**

1. *The Bayes classifier, h_0, predicts compliers as those with compliance scores greater than one-half.*

$$h_0(\boldsymbol{x}) := \underset{h:\mathcal{X}\to\{0,1\}}{argmin}\ \mathcal{E}(h) \equiv \mathbb{1}(\gamma(\boldsymbol{x}) > 1/2) \tag{3}$$

The Bayes classifier is optimal with respect to classification error among all classifiers.

2. *The quantile-threshold classifier, h_q, selects among units with the highest compliance scores.*

$$h_q(\boldsymbol{x}) := \mathbb{1}(\gamma(\boldsymbol{x}) > q) \tag{4}$$

where $q = F^{-1}(1-\mu)$ where F is the CDF of the compliance score and μ is the instrument strength. q is the $(1-\mu)$-quantile of the distribution of compliance scores, meaning that the rule classifies the observations with the top $100\mu\%$ of compliance scores as compliers.

The quantile-threshold classifier is optimal with respect to classification error among all strength-calibrated classifiers.

3. *The stochastic classifier, h_s, selects compliers with probability equal to the compliance score.*

$$h_s(\boldsymbol{x}) := \mathbb{1}(\gamma(\boldsymbol{x}) > U) \sim Bernoulli(\gamma(\boldsymbol{x})) \tag{5}$$

Where $U \sim N(0,1)$. Thus, h_s predicts that a participant with characteristics $\boldsymbol{X} = \boldsymbol{x}$ is a complier with probability $\gamma(\boldsymbol{x})$.

The stochastic classifier is the only strength-calibrated and distribution-matched classifier, and is therefore optimal with respect to classification error in this class.

Distribution-matching and strength-calibration respectively can be written as

$$\begin{aligned} \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid h(\mathbf{X}) = a) &= \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = a) \quad \forall\, \mathbf{x} \text{ and } a \in \{0,1\} \\ \mathbb{P}(h(\mathbf{X}) = a) &= \mathbb{P}(C = a) \quad \text{for } a \in \{0,1\} \end{aligned}$$

Implying by law of total probability

$$\frac{\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid h(\mathbf{X}) = 1)}{\sum_{a=1}^{2} \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid h(\mathbf{X}) = a) \times \mathbb{P}(h(\mathbf{X}) = a)} = \frac{\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = 1)}{\sum_{a=1}^{2} \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = a) \times \mathbb{P}(C = a)}$$

$$\frac{\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid h(\mathbf{X}) = 1)}{\mathbb{P}(\mathbf{X} \leq \mathbf{x})} = \frac{\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = 1)}{\mathbb{P}(\mathbf{X} \leq \mathbf{x})} \quad \forall \mathbf{x}$$

Meaning strength-calibrated and distribution-matched rules permit estimation of prevalence ratios from predicted compliers (Baiocchi, Cheng, and Small 2014), which measure the frequency of a characteristic among the compliers relative to the general population.

3.3.2 Comparing classifiers

A natural question is how the classification errors of the three proposed classifiers compare, and whether quantile-threshold and stochastic classifiers give up much in terms of classification error relative to the Bayes classifier. The following theorem addresses this question by relating the classification errors of each of the classifiers to one another.

Theorem 3.5 (Relating classification errors)**.** *Suppose there is a unique quantile $(1-\mu)$-quantile, q, s.t. $\mathbb{P}(\gamma > q) = \mu$. The classification errors of the Bayes (optimal) classifier, quantile-threshold classifier, and stochastic classifier — $\mathcal{E}(h_0)$, $\mathcal{E}_q$, and $\mathcal{E}_s$ respectively — can be related as follows*

$$\begin{gathered} \frac{1}{2}(1 - \sqrt{1 - 2\mathcal{E}_S}) \leq \frac{1}{2}(1 - \sqrt{1 - 2\mathcal{E}_q}) \leq \mathcal{E}(h_0) \leq \mathcal{E}_q \leq \mathcal{E}_S \\ \mathcal{E}_q \leq \mathcal{E}_S \leq 2\mathcal{E}(h_0)(1 - \mathcal{E}(h_0)) \leq 2\mathcal{E}(h_0) \end{gathered} \tag{6}$$

Proofs of these results can be located in the appendix. Theorem 3.5 shows that the quantile and stochastic rules can yield informative bounds on the Bayes classification error. For example, suppose that the quantile classifier is 70% accurate (i.e., $\mathcal{E}_q = 0.30$). Then, the Bayes classifier has risk $0.18 \leq \mathcal{E}(h_0) \leq 0.30$, meaning that the Bayes classifier will be at least 70% accurate and at most 82% accurate. The second expression says that the best unconstrained classifier has a risk lower bounded by one-half the quantile and stochastic risks, which can be informative when these quantities are small.

3.3.3 Estimating classifiers

In practice, the true classification rules are not known and must be estimated from data. A simple approach for estimating the classification rules is a plug-in approach. For example, the plug-in estimator of the Bayes classifier h_0 is given by the following

$$\hat{h}_0(\mathbf{x}) = \mathbb{1}(\hat{\gamma}(\mathbf{x}) > 1/2) \tag{7}$$

where the true compliance score γ is replaced by it's estimate $\hat{\gamma}$. We can relate the error of $\hat{h}_0$ to the error in estimating $\hat{\gamma}$ via Theorem 2.2 in Devroye, László, and Gábor (1996):

$$\mathcal{E}(\hat{h}) - \mathcal{E}(h_0) \leq 2\mathbb{E}|\hat{\gamma} - \gamma| \leq 2||\hat{\gamma} - \gamma||_{L_2(\mathbb{P})}$$

This shows that consistent estimation of γ also yields consistent estimation of h_0 in terms of classification error.

A plug-in estimator of the quantile rule, h_q, is:

$$\hat{h}_q(\mathbf{x}) = \mathbb{1}(\hat{\gamma}(\mathbf{x}) > \hat{q}) \tag{8}$$

where $\hat{q}$ is an estimate of the $(1-\mu)$ quantile of γ. One could use $\hat{q} = \hat{F}^{-1}(1-\hat{\mu})$, i.e., the $(1-\hat{\mu})$-quantile of the empirical distribution of the compliance score where $\hat{\mu}$ is the estimated instrument strength.

A plug-in estimator of the stochastic rule, h_s, is:

$$\hat{h}_s(\mathbf{x}) = \mathbb{1}(\hat{\gamma}(\mathbf{x}) > U) \tag{9}$$

for $U \sim \text{Unif}(0,1)$.

For the plug-in estimators in Equations 8 and 9, we can upper bound the error in estimating classification error via errors in estimating the nuisance parameters – the compliance score γ and in the case of the quantile classifier, the quantile q. For the quantile classifier, h_q, we require a margin assumption (Audibert and Tsybakov 2007) that ensures that the compliance score is not too concentrated about the quantile q.

Assumption 7 (Margin Assumption): There exists constant $\alpha > 0$ s.t.

$$\mathbb{P}(|\gamma - q| \leq t) \lesssim t^\alpha \qquad \forall t > 0$$

Where $\lesssim$ indicated bounded to a constant. This assumption is critical for ensuring the convergence of the classifier by controlling the local behavior of the compliance score around q (Audibert and Tsybakov 2007).

The following theorem describes bounds on the excess error under the plug-in quantile and stochastic classifiers as described in Equations 8 and 9.

Theorem 3.6 (Excess errors of plug-in classifiers)**.** *Let $\hat{h}_q$ and $\hat{h}_s$ be the plugin-classifiers in Equations 8 and 9. Then an upper bound on the excess error for the plug-in stochastic classifier is:*

$$|\mathcal{E}(\hat{h}_s) - \mathcal{E}_S| \leq \sqrt{1 - \mathcal{E}_S}||\hat{\gamma} - \gamma||$$

Under the margin assumption, then the excess error for the plug-in quantile classifier is:

$$|\mathcal{E}(\hat{h}_q) - \mathcal{E}_q| \lesssim (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^\alpha$$

Thus, under accurate estimates of the nuisance parameters (compliance score γ and quantile q), the plug-in classifiers have small excess classification error.

Like the true stochastic classifier rule, the plug-in stochastic classifier, $\hat{h}_S$, can be used to estimate the complier characteristics of the form $\theta = \mathbb{E}(f(\mathbf{X})|C = 1)$ by computing empirical averages among the group of predicted compliers, $\hat{h}_S = 1$.

$$\hat{\theta} = \mathbb{P}_n(f(\mathbf{X})|\hat{h}_S = 1) = \frac{\mathbb{P}_n(f(\mathbf{X}), \hat{h}_S = 1)}{\mathbb{P}_n(\hat{h}_S = 1)}$$

We can upper bound the error in estimating θ via the plug-in stochastic classifier according to estimating the compliance score γ.

Theorem 3.7 (Estimating complier characteristics via plug-in stochastic classifier)**.** *Assume the complier characteristic function f is bounded, then for the estimator described above, we have that:*

$$|\hat{\theta} - \theta| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + ||\hat{\gamma} - \gamma||\right)$$

When $\hat{\gamma}$ is constructed from an independent sample.

Theorem 3.7 states when we estimate γ at a root-n rate, we estimate θ consistently at a root-n rate too. However, the plug-in estimates of θ will not be fully efficient, because it was built using one sample split to estimate γ and because the estimator was not targeted to estimate θ (evidenced by the $||\hat{\gamma} - \gamma||$ term in its convergence rate). Indeed, flexible estimates of the compliance score γ will be consistent at slower rates, which will be inherited in the estimate of θ. While the proposed plug-in estimator for θ could be preferred for its simplicity, a doubly-robust influence function based estimator of θ could be pursued along with cross fitting to achieve full nonparametric efficiency. This is not the focus of this paper and is left to future work.

In summary, a sharp causal instrument aids the accurate prediction of compliers based on observed covariate information. In the previous subsection, we introduced three classification rules and related them to one another in terms of their classification errors. We described plug-in estimators of the classification rules and showed the classification error under the plug-in classifiers were consistent for the true classification error under accurate specification of nuisances. In the case of the quantile-threshold classifier, a margin condition restricting the density of compliance scores around the threshold was also required to ensure consistency. The classifiers examined in this subsection will motivate the choice of an optimal rule with respect to minimizing bound length on a causal effect which is explored in the next subsection.

3.4 Bounding effects in identifiable subgroups

In this section, the second feature of sharp instruments is considered: sharp instruments yield informative bounds in observable subgroups of likely compliers defined by observed covariates. We derive causal effect bounds and the corresponding bound length in observable subgroups and characterize the optimal group of participants of a given size that achieves the narrowest bound. We propose nonparametric estimators of the bounds that are efficient under conditions.

3.4.1 Bounds and bound length

Define an observable subgroup of participants as all participants with $g(\mathbf{x}) = 1$ for some binary indicator function $g : \mathcal{X} \to \{0, 1\}$. Define the average treatment effect in an observable subgroup as

$$\beta(g) := \mathbb{E}[Y^{a=1} - Y^{a=0} | g = 1]$$

The following theorem gives us bounds on the average treatment effect in an identifiable subgroup. Before stating the result, we define relevant notation for lower and upper bounds. Define

$$\beta_j(g) := \mathbb{E}\left[\mathbb{E}(V_{j,1}|\mathbf{X}, Z = 1) - \mathbb{E}(V_{j,0}|\mathbf{X}, Z = 0)|g = 1\right] \tag{10}$$

For $j = \ell, u$, where

$$\begin{aligned} V_{u,1} &= YA + 1 - A \qquad V_{u,0} = Y(1 - A) \\ V_{\ell,1} &= YA \qquad\quad V_{\ell,0} = Y(1 - A) + A \end{aligned}$$

Given these definitions, we present the following result.

Theorem 3.8 (Bounding causal effects in identifiable subgroups)**.** *Under Assumptions 1-6, the treatment effect in the identifiable subgroup defined by $g : \mathcal{X} \to \{0, 1\}$, $\beta(g)$, is bounded as*

$$\beta_\ell(g) \leq \beta(g) \leq \beta_u(g) \tag{11}$$

Theorem 3.8 generalizes partial identification results of Balke and Pearl (1997), Manski (1990), and Robins (1989). A full proof of the theorem can be found in the appendix. Heuristically, we can express $\beta(g)$ in terms of observed data quantities and two unobserved quantities that correspond to expected

potential outcome. By setting the unobserved quantities to their extreme values in $\{0, 1\}$, we obtain the bounds on $\beta(g)$ presented above.

We present a corollary of Theorem 3.8 which describes length of the bound on the causal effect in the observable subgroup.

Corollary 3.9 (Bound length in identifiable subgroups)**.** *The length of the bounds in Theorem 3.8 for any subgroup $g = 1$ is*

$$\ell(g) \equiv \beta_u(g) - \beta_\ell(g) = \mathbb{E}[1 - \gamma(\boldsymbol{X})|g = 1]$$

And under Assumptions 1-6, yields

$$\ell(g) = \mathbb{P}(C = 0|g = 1)$$

Importantly, under Assumptions 1-6, the bound length is interpreted as the proportion of noncompliers in the subgroup defined by $g(\mathbf{x}) = 1$. This fact was noted by Balke and Pearl (1997) for bounds on the global effect ($g = 1$ for all participants). Thus, the bounds on the effect in the observable subgroup are narrower than bounds on the global average treatment effect when $\mathbb{P}(C = 0|g = 1) < \mathbb{P}(C = 0)$, or when the proportion of noncompliers in the subgroup is less than in the total population.

Corollary 3.9 suggests that the task of seeking the tightest bound on an observable subgroup effect is equivalent to finding the subgroup with the highest probabilities of being compliers. Recalling Result 3.2, this is the subgroup of participants with the highest compliance scores, $\gamma(\mathbf{x})$. However, the strictly narrowest bound length is achieved by selecting participants with $\gamma = \gamma_{\max}$ for $\gamma_{\max} := \sup_{\mathbf{x}\in\mathcal{X}} \gamma(\mathbf{x})$, which may have negligible size. Indeed, unless there are multiple participants with the maximum compliance score, the subgroup of participants with the maximum complier scores will be of size one, for which an average causal effect does not exist. Even if there a small number of participants with the maximal compliance score, restricting focus to few participants will yield a estimand describing a narrow subset of the study population and produce estimators with high finite-sample error. We instead elect to choose among subgroups of participants of a particular size. Let

$$\mathcal{G}(t) := \{g : \mathbb{P}(g = 1) = t\}$$

correspond to the collection of all observable subgroups of a given size t. The following result gives the form of the subgroup of a certain size t that minimizes the bound length.

Result 3.10 (Subgroup of size t with minimal bound length)**.** *Let $F(t) = \mathbb{P}(\gamma \leq t)$ be the distribution function of the compliance score. Then the subgroup that minimizes bound length among all those of size at least t is given by*

$$\underset{g\in\mathcal{G}(t)}{\textit{argmin}}\, \ell(g) = \mathbb{1}(\gamma(\boldsymbol{x}) > F^{-1}(1 - t))$$

In words, among all observable subgroups of size t, the subgroup with the narrowest bounds on the average causal effect corresponds to the subgroup with the top $100t\%$ of compliance scores. This is reminiscent of the form of the quantile threshold classifier in Equation 4. In fact, when we restrict to subgroups of the size $t = \mu$, the quantile threshold classifier h_q minimizes both classification error and bound length of the causal effect in the observed subgroup.

The goal of obtaining tight bounds on causal effects in observable subgroups now involves estimating causal effects conditional on the compliance score.

$$\mathbb{E}[Y^{a=1} - Y^{a=0}|\gamma(\mathbf{X}) > F^{-1}(1 - t)]$$

The average causal effect conditional on a compliance score was developed by Follmann (2000) and Joffe, Ten Have, and Brensinger (2003), but their approaches utilized parametric models instead of quantiles. Thus, the proposed approach can be considered a nonparametric alternative.

3.4.2 Estimation and Inference

Consider the goal estimate and perform inference on bounds on $\beta(g)$, the causal effect in an identified subgroup defined by indicator function $g : \mathcal{X} \to \{0, 1\}$. In the last subsection, we showed the narrowest bound on the average causal effect among a subgroup of size μ is obtained for $g := h_q$, the quantile

threshold classifier which selects units with the top $100\mu\%$ of compliance scores. We focus on estimation and inference on $\beta_\ell(h_q), \beta_u(h_q)$, bounds on the average causal effect among compliers predicted by the quantile-threshold classifier.

The true compliance score, $\gamma(\mathbf{X})$, and $(1-\mu)$-th quantile of the compliance score distribution, q, are unknown and must be estimated from data. γ and q as *nuisances* for estimating h_q. The estimators developed in the following subsection use efficient influence functions to combat bias inherited by estimation of the nuisances and use sample splitting to control empirical process terms without imposing complexity restrictions on the forms of the nuisances (e.g., restricting the nuisances to lie in Donsker classes).

We provide a short primer on semiparametric inference. Efficient influence functions (EIFs) provide benchmarks for nonparametric and semiparametric efficiency. EIFs correspond to the score function in the least favorable parametric submodel, the submodel with minimal Fisher information about the parameter of interest. Since estimation of a parameter in the infinite-dimensional model should be at least as hard as any parametric submodel, the variance of the efficient influence function provides a lower bound on the variance of an regular estimator in the infinite-dimensional model. Estimators based on the EIF are efficient and may have other useful properties such as double robustness, where the estimator may exhibit consistency under misspecification of one of the nuisances. Many influence-function-based estimators also do not have first order bias, allowing them to achieve parametric rates of convergence even when the nuisances are estimated flexibly at slower rates. We refer the reader elsewhere for further discussion (Kennedy 2023; Kennedy 2017; Hines et al. 2022).

We begin with general notation. Let T be a general random variable (e.g., T could be the treatment A, outcome Y, or a function of them). Consider the two nuisance functions $\eta := \{\pi_z(\mathbf{X}), \mathbb{E}[T|\mathbf{X}, Z=z]\}$, the instrument propensity score and the regression function. Suppose we are interested in estimating the G-computation mean form T conditional on $Z=z$, which is given by

$$\Psi(P) := \mathbb{E}[T|Z=z] \equiv \mathbb{E}\{\mathbb{E}[T|\mathbf{X}, Z=z]\}$$

The uncentered efficient influence function of $\Psi(P) = \mathbb{E}\{\mathbb{E}[T|\mathbf{X}, Z=z]\}$ in a nonparametric model is given by

$$\varphi_z(T;\eta) = \frac{\mathbb{1}(Z=z)}{\pi_z(\mathbf{X})}(T - \mathbb{E}[T|\mathbf{X}, Z=z]) + \mathbb{E}[T|\mathbf{X}, Z=z] \tag{12}$$

We propose using cross-fitting to estimate the nuisances $\hat{\eta}$ flexibly without imposing Donsker restrictions. The idea behind sample splitting is to estimate η values for a group of units using data from the other held-out units. Split the data into K groups by drawing n independent identical uniform draws over $\{1, \ldots, K\}$. This splits are denoted by $(B_1, \ldots, B_n)$, where $B_i = b$ denoting unit i was split into group b. The first quantity of interest we estimate is the strength of the instrument, μ.

$$\hat{\mu} = \sum_{k=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(B_i = b) \right\} \mathbb{P}_n^b\{\phi(\mathbf{O}; \hat{\eta}_{-b})\} \equiv \mathbb{P}_n\{\phi_\mu(\mathbf{O}; \hat{\eta}_{-B})\}$$

where $\mathbb{P}_n^b$ is the empirical distribution over the units in group b, and $\phi_\mu(\mathbf{O}, \eta) = \varphi_1(A;\eta) - \varphi_0(A;\eta)$ is the EIF of instrument strength $\mu = \mathbb{E}(\gamma) = \mathbb{E}[\mathbb{E}[A=1|\mathbf{X}=\mathbf{x}, Z=1] - \mathbb{E}[A=1|\mathbf{X}=\mathbf{x}, Z=0]]$. $\hat{\eta}_{-b}$ denotes the estimates of the nuisances $\eta = (\pi_z, \lambda_z)$ based on all the units not in b.

We propose estimating bounds on average causal effect among the observable subgroup identified by the quantile-classifier, $\beta_\ell(h_q)$ and $\beta_u(h_q)$, with $\hat{\beta}_\ell(\hat{h}_q)$ and $\hat{\beta}_u(\hat{h}_q)$, where

$$\hat{\beta}_\ell(\hat{h}_q) = \mathbb{P}_n[\{\varphi_1(V_{\ell,1}; \hat{\eta}_{-B}) - \varphi_0(V_{\ell,0}; \hat{\eta}_{-B})\}\, \hat{h}_{q,-B}]/\mathbb{P}_n[\hat{h}_{q,-B}]$$
$$\hat{\beta}_u(\hat{h}_q) = \mathbb{P}_n[\{\varphi_1(V_{u,1}; \hat{\eta}_{-B}) - \varphi_0(V_{u,0}; \hat{\eta}_{-B})\}\, \hat{h}_{q,-B}]/\mathbb{P}_n[\hat{h}_{q,-B}]$$

For $V_{\cdot,\cdot}$ defined as in Equation 10. We also define $\hat{h}_{q,-B} := \mathbb{1}(\hat{\gamma}_{-b} > \hat{q}_{-b})$ and $\hat{q}_{-b}$ is the $(1-\hat{\mu})$-th quantile of $\hat{\mu}$ solving $\mathbb{P}_n^b\{\mathbb{1}(\hat{\gamma}_{-b} > \hat{q}_{-b})\} = \hat{\mu}$.

Define the following remainder terms.

$$R_{1,n} = ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z|| + \max_z ||\hat{v}_{j,z} - v_{j,z}|| \right)$$
$$R_{2,n} = (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^\alpha$$

Where $\alpha > 0$ is the exponent from the margin from the Margin Assumption in Assumption 7, assuring not too many compliance scores concentrate about the quantile q. The following result establishes the convergence rate and conditions that achieve asymptotic normality

Theorem 3.11 (Convergence rate and ASN of bound estimators)**.** *Assume instrumental propensity score positivity.*

$$\mathbb{P}(\epsilon \leq \hat{\pi}_z(\boldsymbol{X}) \leq 1-\epsilon) = 1 \text{ for } z = 0, 1 \text{ and some } \epsilon > 0$$

Assume the nuisances are all estimated consistently.

$$||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| + \max_z ||\hat{v}_{j,z} - v_{j,z}|| + \mathbb{P}(\hat{h}_q \neq h_q) = o_{\mathbb{P}}(1)$$

1. *If the margin condition in Equation 10 holds for some $\alpha > 0$, then*

$$\hat{\beta}_j(\hat{h}_q) - \beta_j(h_q) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + R_{1,n} + R_{2,n}\right) \tag{13}$$

2. *If it also holds if each component of the remainder terms are estimated with root-n consistency, that is, $R_{1,n} + R_{2,n} = o_{\mathbb{P}}(1/\sqrt{n})$, then*

$$\sqrt{n}(\hat{\beta}_j(\hat{h}_q) - \beta_j(h_q)) \rightsquigarrow N(0, Var[\{\varphi_1(V_{j,1}) - \varphi_0(V_{j,0})\}h_q - \beta_j(h_q)\phi_\mu]/\mu^2) \tag{14}$$

In words, the convergence rate of the estimated bounds described in Equation 13 are, at best, root-n consistent. The convergence rate depends on a doubly robust term, $R_{1,n}$, which is small if the instrumental propensity score π_1 is estimated accurately or if the treatment regression λ_z and bound parameters $v_{j,z}$ are both estimated accurately. Each of the constituent nuisances may be estimated at slower than root-n rates and still $R_{1,n}$ will be root-n consistent and not effect the convergence of the bound estimator. The convergence rate also depends on the $R_{2,n}$ term which will be small if the compliance score γ is estimated accurately. In particular, when γ is estimated accurately and α from the margin condition is large, the contribution of $R_{2,n}$ to the convergence rate will be negligible. For sharp instruments which have variable compliance scores γ, we anticipate that α will be larger as less mass will be concentrated around the quantile q.

Using the approach of G. Imbens and C. Manski (2004), confidence intervals for the partially identified parameter $\beta(h_q)$ can be constructed that cover the parameter uniformly over the widths of the partial identification bounds.

3.5 Summarizing instrument sharpness

The previous two subsections describe the two properties intrinsic to sharp instruments: they enable accurate prediction of compliers and yield narrow bounds on average causal effects in observable subgroups. The following subsection formalizes the definition of sharpness which captures these two properties. We describe the relationship between sharpness, instrument strength, complier classification error, and bound length of causal effects observable subgroup effects. We propose a nonparametric estimator of the sharpness statistic, and advocate for its reporting alongside instrument strength to describe instrument quality.

3.5.1 A summary measure of instrument sharpness

The proposed summary measure of instrument sharpness is the proportion of variance in compliance status that is explained by the set of observed covariates $\mathbf{X}$, particularly by the highest compliance score values. This is equivalent to the correlation between the true and predicted compliance statuses.

Definition 3.1 (Instrument Sharpness)**.** *The sharpness ψ of an instrument Z with latent compliance indicator C and compliance score γ is given by*

$$\psi = \frac{cov(C, h_q)}{var(C)} = corr(C, h_q) \tag{15}$$

Where h_q is the quantile-threshold classifier described in Equation 4 which selects units with the top $100\mu\%$ of compliance scores.

Sharpness can be intuitively interpreted as the variation in true compliance status explained by top $100\mu\%$ of compliance scores. It is also the slope of the regression of true compliance indicators C on predicted compliance indicators h_q. The higher the agreement between true and predicted compliance

labels, the sharper the instrument. If the two perfectly agree, the sharpness measure equals 1. If the compliance scores do not predict compliance status at all then the sharpness measure would be 0. The sharpness measure can be considered a model free measure of how well compliance status can be predicted from observed covariates.

We focus on the h_q classifier because in the definition of sharpness because h_q is optimal among classifiers of size μ, meaning

$$\underset{h\in\mathcal{G}(\mu)}{\operatorname{argmin}}\ \mathcal{E}(h) = \underset{h\in\mathcal{G}(\mu)}{\operatorname{argmin}}\ \ell(h) = \underset{h\in\mathcal{G}(\mu)}{\operatorname{argmax}}\ \psi(h) = h_q$$

h_q also is a preferred choice to classify compliers as it yields simple relationships between classification error, bound length, and sharpness, as we will discuss shortly.

For any $h : \mathcal{X} \to \{0, 1\}$, sharpness also measures the difference between true positive and false positive rates.

$$\psi = \frac{\operatorname{cov}(C, h)}{\operatorname{var}(C)} = \mathbb{P}(h = 1|C = 1) - \mathbb{P}(h = 1|C = 0)$$

Specifically for the choice of h_q, sharpness can also be interpreted as the Youden Index, which is a popular summary of classifier performance.

$$\psi = \mathbb{P}(\gamma > F^{-1}(1-\mu)|C = 1) - \mathbb{P}(\gamma > F^{-1}(1-\mu)|C = 0)$$

Why is instrument sharpness a useful measure to report aside from classification error $\mathcal{E}(h)$ and bound length $\ell(h)$? Importantly, sharpness ψ is a measure of instrumental quality wholly separate from instrument strength in the sense of variation independence. This means while quantities like $\mathcal{E}(h)$ and $\ell(h)$ depend on instrument strength, ψ does not. For any strength $\mu \in [\epsilon, 1-\epsilon]$, we can construct a compliance score γ that produces any sharpness $\psi \in [0, 1]$. And vice versa, for any sharpness $\psi \in [0, 1]$, we can construct a compliance score γ that produces any strength value $\mu = [\epsilon, 1-\epsilon]$.

Although compliance status C is latent, we may still still identify sharpness under instrumental conditions.

Result 3.12 (Sharpness identification)**.** *Under assumptions 1-4 and 6, sharpness is identified.*

$$\psi = \mathbb{E}[\gamma(\boldsymbol{X})h_q(\boldsymbol{X}) - \mu^2]/\mu(1-\mu)$$

Next, we describe the relationship to classification error and bound length. For instruments of a fixed length, sharper instruments yield a more accurate complier classification and narrower bounds on the causal effects in observable subgroups.

Theorem 3.13 (Relationship of sharpness to classification error and bound length)**.** *The relationship between classification error $\mathcal{E}(h_q)$, bound length $\ell(h_q)$, and sharpness ψ can be described by*

$$\begin{aligned}\mathcal{E}(h_q) &= 2\mu(1-\mu)(1-\psi)\\ \ell(h_q) &= (1-\mu)(1-\psi)\end{aligned}$$

Theorem 3.13 demonstrates that strength and sharpness are fundamental aspects of instrument quality. In particular, the two quantities fully determine the classification error and bound length. Instruments of fixed strength yield lower classification errors and and narrower bounds. Instruments of perfect sharpness yield perfect complier prediction, $\mathcal{E}(h_q) = 0$ and point identification of the causal effect in the identifiable subgroup, $\ell(h_q) = 0$. Also, classification error and bound length decay linearly with sharpness for an instrument of fixed strength.

The theorem also shows that nonzero sharpness improves complier prediction and narrows bounds on causal effects in observable subgroups. For complier prediction, consider the naive strength-calibrated classifier which flips a coin with probability μ. When an instrument has nonzero sharpness, $\psi > 0$, $\mathcal{E}(h_q) < 2\mu(1-\mu)$, which is the error of the rule $h \sim \text{Bern}(\mu)$. When we consider bound lengths, when we have nonzero sharpness $\psi > 0$ implies $\ell(h_q) < 1-\mu$, which is the length on the bounds on the average treatment effect as developed by Balke and Pearl (1997), Manski (1990), and Robins (1989). Thus, ψ represents the percent reduction in the length of the bounds.

In summary, the sharpness measure proposed in Definition 3.1 can be interpreted as the variance in compliance statuses explained by the highest compliance scores. The measure captures both the accuracy of complier classification and the tightness on the bounds on the causal effects.

3.5.2 Estimation and Inference

This section focuses on developing an estimator of sharpness ψ. This estimator uses influence functions to correct bias inherited by the nuisances and incorporates sample splitting to allow flexible estimation of nuisances and avoid complexity restrictions.

The sharpness estimator reflects the expressioon in the identification Result 3.12. From the previous section, recall the estimator of strength $\hat{\mu} = \mathbb{P}_n\{\phi_\mu(\mathbf{O}, \hat{\eta}_{-B})\}$. The sharpness estimator also depends on an estimator $\hat{\xi} = \mathbb{P}_n(\hat{\phi}_{\xi,-B})$ of

$$\xi = \mathbb{E}[\gamma h_q]$$

$\hat{\phi}_\xi = \phi_\mu(\mathbf{O}; \hat{\eta})\hat{h}_q(\mathbf{X})$ are the influence functions for $\hat{\xi}$. We estimate sharpness via

$$\hat{\psi} = \frac{(\hat{\xi} - \hat{\mu}^2)}{\hat{\mu}(1 - \hat{\mu})} \tag{16}$$

which combines the influence-function-based estimators of the terms from Result 3.12. For example, the numerator is the estimator of the covariance between compliance status C and classifier h_q. Define the following remainder terms.

$$R_{1,n} = ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right)$$
$$R_{2,n} = (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^{1+\alpha}$$

Where α is the exponent in the margin assumption. Compared to the remainder for estimating subgroup effects in section 3.4, the following remainder is of higher order, meaning that sharpness can be measured at faster rates.

The following theorem gives convergence rates and asymptotic normality of the estimated sharpness measure.

Theorem 3.14 (Convergence rates and asymptotic normality of sharpness)**.** *Assume instrumental propensity score positivity.*

$$\mathbb{P}(\epsilon \leq \hat{\pi}_z(\boldsymbol{X}) \leq 1 - \epsilon) = 1 \text{ for } z = 0, 1 \text{ and some } \epsilon > 0$$

Assume the nuisances are all estimated consistently.

$$||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| + \mathbb{P}(\hat{h}_q \neq h_q) = o_{\mathbb{P}}(1)$$

1. *If the margin assumption holds for some $\alpha > 0$,*

$$\hat{\psi} - \psi = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + R_{1,n} + R_{2,n}\right) \tag{17}$$

2. *If it also holds the remainder terms are estimated at root-n consistency, $R_{1,n} + R_{2,n} = o_{\mathbb{P}}(1/\sqrt{n})$, then*

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N\left(0, var\left[\frac{\phi_\mu h_q + q(\phi_\mu - h_q) - \xi}{(\mu - \mu^2)} + \frac{(2\mu\xi - \xi - \mu^2)}{(\mu - \mu^2)^2}(\phi_\mu - \mu)\right]\right) \tag{18}$$

Theorem 3.14 shows that the proposed sharpness estimator can achieve root-n consistency at best when the nuisances are estimated accurately. The true convergence rate is second order in the nuisance estimators. Thus, sharpness can be estimated at faster rates than the nuisances. Theorem 3.14 also indicates if the remainder terms are estimated at a root-n rate, the estimator is asymptotically normal. The conditions in the theorem will hold if the nuisances are estimated at $n^{1/4}$ rates and $\alpha = 1$, which is achievable under flexible nonparametric estimation.

4 Results

The following section describes results of simulations and an application of the proposed methods to a field experiment assessing the effect of door-to-door canvassing on voter turnout.

4.1 Simulations

We use the proposed methods to classify compliers, estimate bound lengths, and estimate sharpness in simulations. We examine the classification error of plug-in Bayes, plug-in quantile threshold, and plug-in stochastic classifiers under varying levels of instrument sharpness and sample size. We estimate performance using classification error. We evaluate the bound lengths on the average treatment effect and the causal effect in the identifiable subgroup produced by the quantile-threshold classifier $\beta(h_q)$ under varying sample sizes and instrument sharpness values. Evaluating classification error and bound length under differing instrument sharpness values serves to numerically support Theorem 3.13 which relates instrument sharpness to classification error and causal effect bound length. We also assess the bias and coverage of the proposed influence-function-based sharpness estimator given in Equation 16.

Summary of simulation parameters are presented in Table 2. Note that instrument strength is fixed at $\mu = 0.30$, yet sharpness and sample size are varied across replicates. The average treatment effect is assumed to be $\beta = 0.2$. We estimate nuisances with correctly-specified logistic regression models with $K = 2$ sample splits. All simulation were run on the UW Biostatistics Bayes Computing cluster, a computer with 2 6-core Intel Xeon CPU E5- 2620 @ 2.00 GHz 128GB RAM. Simulations were run using the R package `SimEngine`. Proposed methods were implemented using the `npcausal` package.

Parameter	Value/Distribution
Sample size: n	$\{500, 1000, 5000\}$
Sharpness: ψ	$\{0.2, 0.5, 0.8\}$
Covariate: X	$\sim N(0, 1)$
Compliance Label: $C\|X$	$\sim \text{Bern}(\gamma(X))$
Compliance Score: $\gamma(X)$	$\Phi(b_0 + b_1 X)$
Compliance Score Parameter: (b_0, b_1)	Chosen such that $\mu = 0.30$ and ψ as specified
Instrument Propensity: $\pi_z(x)$	$\text{expit}(x)$
Instrument: $Z\|X, C$	$\sim \text{Bern}(\pi_{z=1})$
Treatment: A	$= CZ + (1 - C)A^*$
Variable creating other principal strata: A^*	$A^*\|X, C, Z \sim \text{Bern}(0.5)$
Outcome: Y	$= AY^{a=1} + (1 - A)Y^{a=0}$
CF Outcome: $Y^a\|X, C, Z, A$	$\sim \text{Bern}(0.5 + (a - 0.5)\beta)$
Treatment Effect: β	$= 0.20$
Simulation runs (per setting)	500

Table 2: Summary of simulation parameters. Only sample size n and sharpness ψ were intentionally varied between replicates.

The effect of sharpness on complier classification error of the plugin Bayes classifier h_0, Quantile Threshold classifier h_q, and Stochastic classifier h_s are shown in Figure 3. In support of the theoretical result in Theorem 3.13, accurate prediction of compliers based on observed covariates improves under sharper instruments, even when the instrument has a fixed strength. Results support that the Quantile Threshold classifier h_q especially, and Stochastic classifier h_s to a lesser degree, don't give up much relative to the Bayes classification error when the instrument is sharp. In fact, the quantile-threshold classifier achieves near Bayes-optimal classification error when the instrument is sharp. The decay in classification error for the quantile threshold classifier h_q as a function of increasing sharpness ψ appears linear, supporting the functional form of the classification error presented in Theorem 3.13. At a high level, these results confirm that accurate prediction of compliers is possible for a relatively weak instrument when the instrument is sufficiently sharp.

The effect of sharpness on bounds on causal effects in observable subgroups was assessed in Figure 4a. Findings supported Theorem 3.13. Even for weak instruments of fixed strength, the bounds on the ATE among predicted compliers $\beta(h_q)$ were considerably narrower than the bounds on the global average treatment effect. Additionally, nonzero sharpness was sufficient to guarantee a reduction in the width of the bounds on the causal effect. The decay in bound length of $\beta(h_q)$ appears linear in ψ, supporting the functional form that bound length depends on ψ.

Bias, standard error, and coverage of the influence-function-based estimators of sharpness ψ as given in Equation 16 and Theorem 3.14 are summarized in Figure 4b. As the sample size increased, bias in the estimated instrument sharpness was reduced, reflecting consistency of the sharpness estimator $\hat{\psi}$ under correct nuisance specification. Confidence interval coverage was achieved the nominal 95% level across all simulations, suggesting validity of the asymptotic results when nuisances were correctly specified.

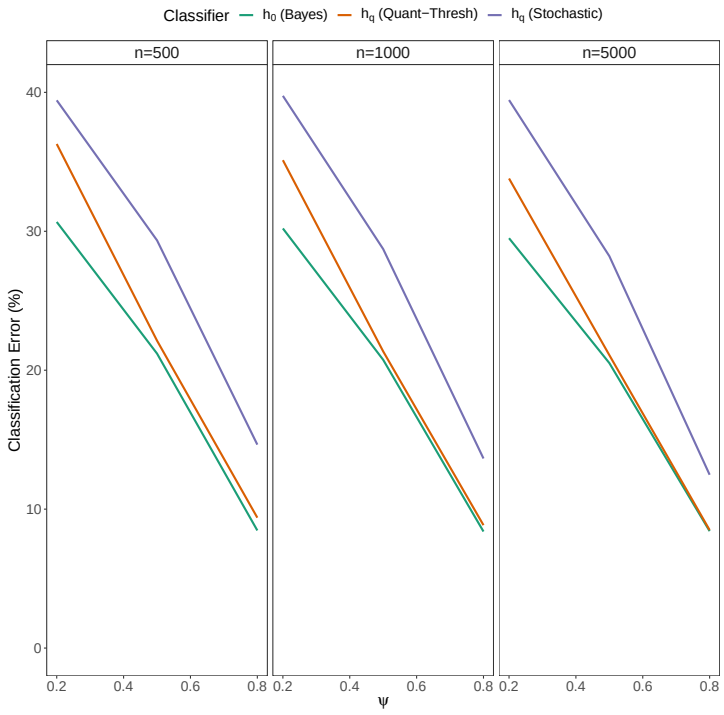


Figure 3: Mean classification errors for the plugin Bayes classifier h_0, Quantile Threshold classifier h_q, and Stochastic classifier h_s under varying sample sizes and levels of instrument sharpness ψ. Note that instrument strength is assumed constant for all simulations $\mu = 0.30$.

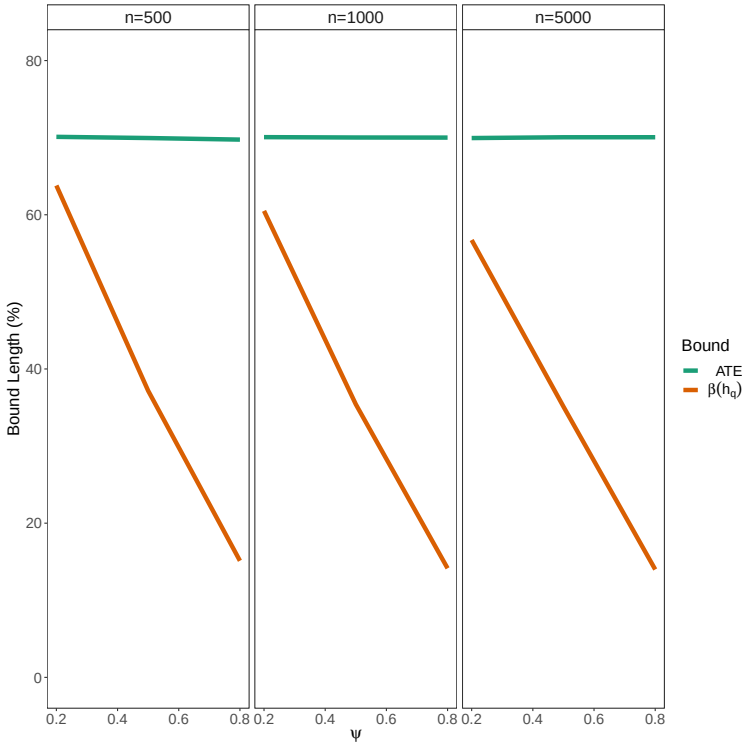
4.2 Application to voting turnout data

In this subsection, we illustrate the proposed methods by applying them to a study assessing whether a voter encouragement intervention could improve voter turnout in six US cities in 2001 local elections. Green, Gerber, and Nickerson (2003) conducted a study among $n = 18,933$ voters in six US cities - Bridgeport, Columbus, Detroit, Minneapolis, Raleigh, and St Paul. Voters were randomly assigned to receive or not receive encouragement to vote in local elections. However, not every voter *assigned* to receive voter encouragement was actually contacted and *received* encouragement. Thus, the trial represents a randomized trial with noncompliance. Green, Gerber, and Nickerson (2003) pursued an IV approach, assumed monotonicity, and estimated the average treatment effect among compliers (a local average treatment effect or LATE). Since the causal effect corresponds to a latent subgroup of voters, the causal effect is of questionable policy interest.

The methods of Kennedy, Balakrishnan, and G'Sell (2020) were applied to classify compliers and estimate bounds on the effect in the subgroup of likely compliers ($\beta(h_q)$) based on observed covariates. Estimates of the sharpness summary of the instrument (randomization) using the proposed methods was also provided. Observed covariates in the study include city, party affiliation, prior voting history in the 1999 and 2000 elections, voting history in the 2021 primary, voter age, family size, and race. Columns were also generated for missing responses to each covariate. We flexibly estimated the nuisances using random forests with 2-fold sample splitting using the R package, `ranger`. We also provided bounds on the global average treatment effect and an estimate of the local/complier average treatment effect (LATE). Nonparametric estimates of the instrument strength and sharpness were also produced.

Plotted in Figure 5 are the estimated compliance score $\hat{\gamma}(\mathbf{X})$ versus age and city. Figure 5 can aid interpretation of who the likely compliers are in the study population. Estimated compliance probabilities varied between 6.3% and 82.7% between participants. The most likely compliers were voters in Raleigh and select older voters from Detroit. Results were consistent with the analysis by Kennedy, Balakrishnan, and G'Sell (2020), save for an anonymization issue with the dataset's encoding of age preventing stratifying Columbus voters by age.

Figure 6 shows bounds on the causal effects and the summary sharpness measure. Compared to bounds on the global average treatment effect, the bounds on the effect on predicted compliers were narrower. The bounds on $\beta(h_q)$ were 20% shorter relative to the bounds on the global average treat-



(a)

Setting	Sharpness		
	Bias	SE	Coverage
n= 500			
ψ=0.2	-7.2	13.5	96.4
ψ=0.5	-2.8	14.0	97.4
ψ=0.8	-1.7	11.2	92.6
n= 1000			
ψ=0.2	-4.8	10.0	96.4
ψ=0.5	-0.9	7.8	97.2
ψ=0.8	-0.2	7.2	96.0
n= 5000			
ψ=0.2	-1.0	4.0	93.8
ψ=0.5	-0.2	3.1	95.2
ψ=0.8	0.0	3.1	96.0

(b)

Figure 4: Summary of 4a illustrating the ATE and 'likely complier ATE' bound length and 4b illustrating bias, SE, coverage of the estimator of sharpness $\hat{\psi}$.

ment effect, although the bounds still covered 0 (bounds: $[-18.1\%, 37.7\%]$, 95% CI: $[-20.0\%, 40.2\%]$). Additionally, the instrument (randomization to encouragement) was stronger than it was sharp. While the instrument had 30.2% strength (95% CI: $[29.2\%, 31.2\%]$), it had only 20.0% sharpness (95% CI: $[17.7\%, 22.4\%]$). The data application affirms the theoretical result that nonzero instrument sharpness is a sufficient condition to achieve narrower causal effect bounds. However, the relative improvement in bound length depends how sharp an instrument is. Given that the instrument was a relatively blunt instrument, the improvement in bound length was relatively modest.

5 Discussion

This paper presents sharpness, a new measure of instrumental variable quality that measures the variation in an instrument's compliance probabilities over the distribution of observed covariates. Sharpness is a new measure of instrumental variable quality that is completely separate from to instrument strength (in the sense of variation independence). Sharp causal instruments permit accurate prediction of compliers using observed covariates and achieve narrow partial identification results for causal effects in observable subgroups defined by covariates. We propose classification rules to predicting compliance status and characterize their errors in large samples. We define bounds on average causal effects in subgroups defined by observed covariates, and illustrate that for a given subgroup size, bound length is minimized among the subgroup with the highest compliance scores. Nonparametric methods are used to estimate all these quantities — classifiers, bounds, and sharpness — and conditions required for consistency and asymptotic normality are discussed. Performance of the methods were evaluated in simulation and in an application to a field experiment examining the effect of door-to-door canvassing on voter turnout.

There are several limitations and future avenues of research motivated by this work.

First, the framework assumes a binary instrument Z. This is problematic for application of the proposed methods to continuous-valued instruments. These include "access-based" instruments such as distance to hospital or number of clinics in neighborhood, as well as "preference-based" instruments defined on the level of geographic region, hospital/clinic, or individual physician (Brookhart and Schneeweiss 2007; Hernán and J. M. Robins 2020). As noted by the authors, the restriction on the binary instrument

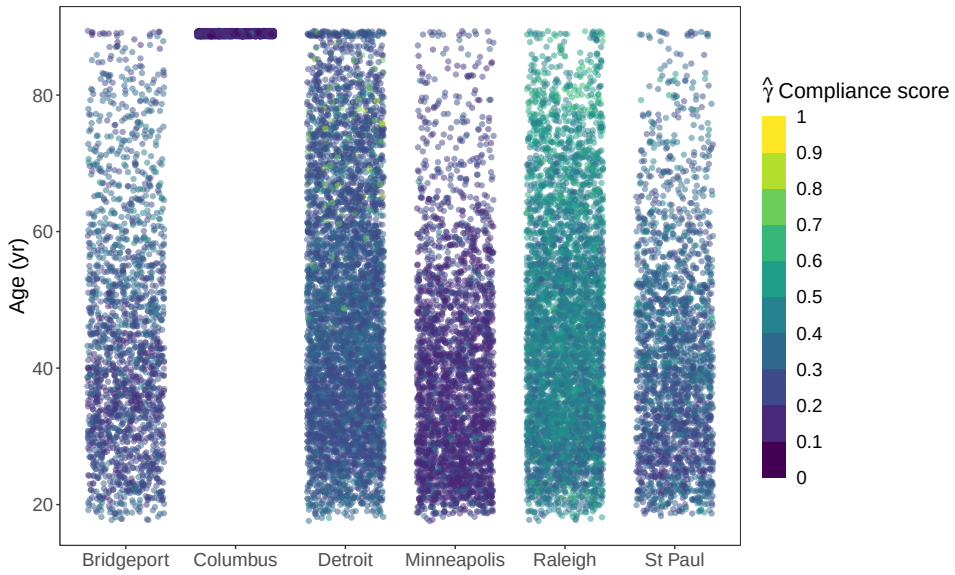


Figure 5: Estimated compliance scores $\hat{\gamma}(\mathbf{X})$ as a function of city and age.

could be removed without changing the proposed approaches. However, accommodating a continuous instrument would requiring estimating of a local instrumental variable (LIV) effect curve that considers an effect among units who would comply right at a threshold of an instrument. There exists a semi-parametric, doubly robust estimator for the LIV curve (Kennedy, Lorch, and Small 2019). Applying sharpness with continuous instruments could help improve interpretation of LIV effects by considering a causal effect in observable subgroup of individuals likely to comply at the threshold. It may also be of interest to assess how stable sharpness is over a varying instrument threshold.

Second, the framework presented here considers a single instrument. While considering a large number of instruments can increase the odds of violating a main IV assumption (exclusion restriction or unconfoundedness), in Mendelian randomization studies, it is commonplace to combine multiple variants into a single allele score or consider weighted causal estimates across several variants (Burgess, Dudbridge, and Thompson 2016). It is probable that genetic variants' abilities to predict an exposure will vary over observed covariates due to gene-environment interactions. However, it is plausible that some covariates may predict compliance well for some instruments and not others. How the notion of sharpness generalizes to the setting of multiple instruments remains an open research question.

Third, a key limitation of the proposed methods is that they do not accommodate multi-valued treatments. Multi-valued treatment prevents identification of the compliance score and treatment effects among compliers, and would require a different approach altogether.

Fourth, the estimation framework presented in this paper assumes that each observation is i.i.d. Data that reflects underlying grouped structure is known to generate anticonservative inference in instrumental variable settings (Shore-Sheppard 1996), and robust 2-stage-least-squares methods have been proposed for IV-based causal effect estimation under heteroskedasticity and autocorrelation (Wooldridge 2012). However, generalizing the nonparametric, influence-function based estimators of causal effect bounds in observed subgroups and summary measures of sharpness presented here to non-independent data is an avenue for further research.

Fifth, the sensitivity of proposed methods to violations of the instrumental conditions (exchangeability of instrument and exclusion restriction) and monotonicity should be assessed in future simulations. Also, performance of the proposed methods to misspecified nuisances should also be assessed via simulations.

Sixth, all results presented rely on the monotonicity assumption, and developing an analog of instrument sharpness for other instrumental variable models — e.g., models that assume effect homogeneity — may be a useful step for future research. In some cases, this is already being pursued (Levis et al. 2023).

Lastly, sharp instruments can produce narrow bounds on causal effects among subgroups of participants that share certain covariates associated with likely compliance with the instrument. The effect among likely compliers is a useful piece of the causal puzzle but is an incomplete view of the casual effect of treatment in the general population. Practitioners should carefully consider whether likely compliers

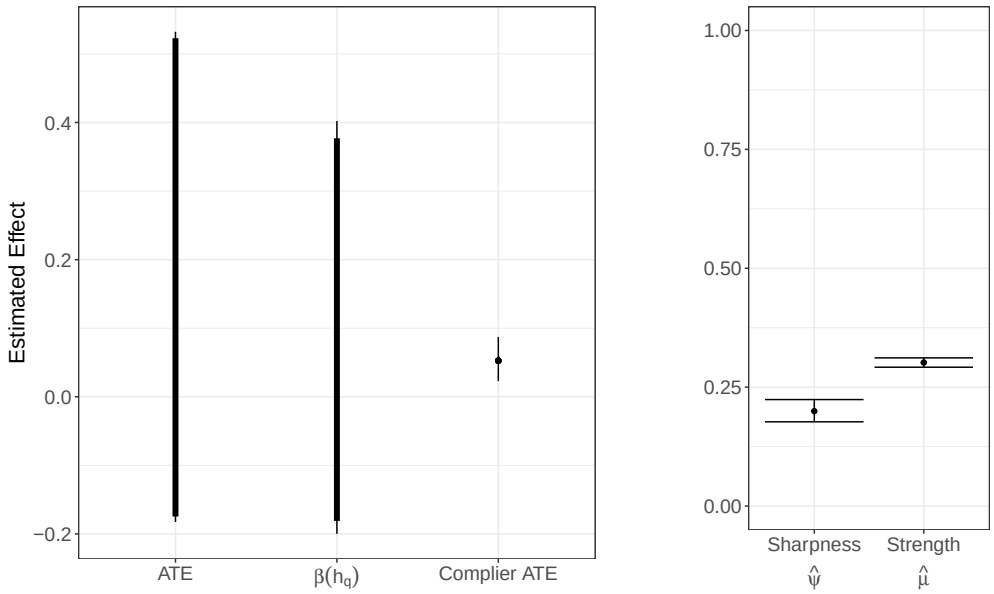


Figure 6: Left: Estimated effect sizes of voting encouragement on turnout in the 2001 elections for bounds on the ATE (ATE), bounds on the ATE among likely compliers (beta(h_q)), and the treatment effect estimated among the latent population of compliers (LATE). Right: Estimated instrument sharpness and strength.

are of primary scientific interest. Also, subgroup-specific causal effects have the potential to generate scientific insight, motivate future interventions, and ultimately bring benefit to specific subpopulations. Issues related to equity and fairness should be discussed in research and policy decisions focused on estimating subgroup-specific causal effects.

In practice, sharpness is a novel measure of instrument quality that should be reported alongside instrument strength in practical settings. Sharpness can improve the interpretation of complier effects identified under monotonicity by aiding prediction of compliers by observed covariates and generating narrow causal effect bounds in observable subgroups. This work also motivates the need for experiments to collect data on factors not only associated with treatment assignment (and thereby ensure the instrument is unconfounded) but also factors associated with compliance. Lastly, sharpness can be considered alongside strength in choosing between multiple instruments. Additionally, sharpness and strength can be estimated without outcome data. If collection of outcome data is costly, one can use sharpness and strength to decide where to collect data to obtain improved compliance and narrow effect bounds *a priori*, which could improve statistical power for IV studies.

6 References

[1] Alberto Abadie. "Semiparametric instrumental variable estimation of treatment response models". en. In: *Journal of Econometrics* 113.2 (Apr. 2003), pp. 231–263. ISSN: 0304-4076. DOI: 10.1016/S0304-4076(02)00201-4. URL: https://www.sciencedirect.com/science/article/pii/S0304407602002014 (visited on 04/18/2023).

[2] Joshua Angrist and Ivan Fernandez-Val. *ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework*. Working Paper. Dec. 2010. DOI: 10.3386/w16566. URL: https://www.nber.org/papers/w16566 (visited on 04/18/2023).

[3] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434 (1996). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 444–455. ISSN: 0162-1459. DOI: 10.2307/2291629. URL: https://www.jstor.org/stable/2291629 (visited on 04/25/2023).

[4] Peter M. Aronow and Allison Carnegie. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable". In: *Political Analysis* 21.4 (2013), pp. 492–506. ISSN: 1047-1987.

[5] Jean-Yves Audibert and Alexandre B. Tsybakov. "Fast learning rates for plug-in classifiers". In: *The Annals of Statistics* 35.2 (Apr. 2007). arXiv:0708.2321 [math, stat]. ISSN: 0090-5364. DOI: 10.1214/009053606000001217. URL: http://arxiv.org/abs/0708.2321 (visited on 05/02/2023).

[6] Michael Baiocchi, Jing Cheng, and Dylan S. Small. "Instrumental variable methods for causal inference". en. In: *Statistics in Medicine* 33.13 (2014), pp. 2297–2340. ISSN: 1097-0258. (Visited on 04/18/2023).

[7] A Balke and J Pearl. "Bounds on treatment effects from studies with imperfect compliance." In: *Journal of the American Statistical Association* 92.439 (1997), pp. 1171–1176.

[8] M. Alan Brookhart and Sebastian Schneeweiss. "Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results". In: *The international journal of biostatistics* 3.1 (2007), p. 14. ISSN: 1557-4679. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2719903/ (visited on 05/21/2023).

[9] Stephen Burgess, Frank Dudbridge, and Simon G. Thompson. "Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods". In: *Statistics in Medicine* 35.11 (May 2016), pp. 1880–1906. ISSN: 0277-6715. DOI: 10.1002/sim.6835. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4832315/ (visited on 05/21/2023).

[10] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967). Conference Name: IEEE Transactions on Information Theory, pp. 21–27. ISSN: 1557-9654. DOI: 10.1109/TIT.1967.1053964.

[11] Angus Deaton. "Instruments, Randomization, and Learning about Development". en. In: *Journal of Economic Literature* 48.2 (June 2010), pp. 424–455. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.424. URL: https://pubs.aeaweb.org/doi/10.1257/jel.48.2.424 (visited on 05/24/2023).

[12] Luc Devroye, Györfi László, and Lugosi Gábor. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics: Stochastic Modeling and Applied Probability 31. New York: Springer, 1996. ISBN: 0-387-94618-7.

[13] Dean A. Follmann. "On the Effect of Treatment among Would-Be Treatment Compliers: An Analysis of the Multiple Risk Factor Intervention Trial". In: *Journal of the American Statistical Association* 95.452 (2000). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 1101–1109. ISSN: 0162-1459. DOI: 10.2307/2669746. URL: https://www.jstor.org/stable/2669746 (visited on 04/25/2023).

[14] Donald P. Green, Alan S. Gerber, and David W. Nickerson. "Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments". en. In: *The Journal of Politics* 65.4 (Nov. 2003), pp. 1083–1096. ISSN: 0022-3816, 1468-2508. DOI: 10.1111/1468-2508.t01-1-00126. URL: https://www.journals.uchicago.edu/doi/10.1111/1468-2508.t01-1-00126 (visited on 05/15/2023).

[15] James M. Heckman and Sergio Urzua. "Comparing IV with Structural Models: What Simple IV can and cannot Identify". In: *National Bureau of Economic Research Working Paper* 14706 (2009).

[16] Miguel A Hernán and James M Robins. *Causal Inference: What If*. en. Boca Raton: Chapman & Hall/CRC, 2020.

[17] Oliver Hines et al. "Demystifying statistical learning based on efficient influence functions". en. In: *The American Statistician* 76.3 (July 2022). arXiv:2107.00681 [math, stat], pp. 292–304. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.2021.2021984. URL: http://arxiv.org/abs/2107.00681 (visited on 05/27/2023).

[18] Guido Imbens and Charles Manski. "Confidence Intervals for Partially Identified Parameters". en. In: *Econometrica* 72.6 (2004), pp. 1845–1857.

[19] Guido W Imbens. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)". en. In: *Journal of Economic Literature* 48.2 (June 2010), pp. 399–423. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.399. URL: https://pubs.aeaweb.org/doi/10.1257/jel.48.2.399 (visited on 05/24/2023).

[20] Marshall M. Joffe, Thomas R. Ten Have, and Colleen Brensinger. "The compliance score as a regressor in randomized trials". eng. In: *Biostatistics (Oxford, England)* 4.3 (July 2003), pp. 327–340. ISSN: 1465-4644. DOI: 10.1093/biostatistics/4.3.327.

[21] Edward H. Kennedy. *Semiparametric doubly robust targeted double machine learning: a review*. en. arXiv:2203.06469 [stat]. Jan. 2023. URL: http://arxiv.org/abs/2203.06469 (visited on 05/27/2023).

[22] Edward H. Kennedy. *Semiparametric theory*. en. arXiv:1709.06418 [stat]. Sept. 2017. URL: http://arxiv.org/abs/1709.06418 (visited on 05/27/2023).

[23] Edward H. Kennedy, Sivaraman Balakrishnan, and Max G'Sell. "Sharp instruments for classifying compliers and generalizing causal effects". en. In: *The Annals of Statistics* 48.4 (Aug. 2020). ISSN: 0090-5364. DOI: 10.1214/19-AOS1874. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-4/Sharp-instruments-for-classifying-compliers-and-generalizing-causal-effects/10.1214/19-AOS1874.full (visited on 04/11/2023).

[24] Edward H. Kennedy, Scott Lorch, and Dylan S. Small. "Robust Causal Inference with Continuous Instruments Using the Local Instrumental Variable Curve". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81.1 (Feb. 2019), pp. 121–143. ISSN: 1369-7412. DOI: 10.1111/rssb.12300. URL: https://doi.org/10.1111/rssb.12300 (visited on 05/21/2023).

[25] Alexander W. Levis et al. *Covariate-assisted bounds on causal effects with instrumental variables*. arXiv:2301.12106 [stat]. Jan. 2023. URL: http://arxiv.org/abs/2301.12106 (visited on 04/18/2023).

[26] Tony Liu et al. "Data-driven exclusion criteria for instrumental variable studies". en. In: *Proceedings of the First Conference on Causal Learning and Reasoning*. ISSN: 2640-3498. PMLR, June 2022, pp. 485–508. URL: https://proceedings.mlr.press/v177/liu22a.html (visited on 04/18/2023).

[27] Manski. "Nonparametric bounds on treatment effects." In: *American Economic Review* 80.2 (1990), pp. 319–323.

[28] Robins. "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies." In: *Health Services Research Methodology: A Focus on AIDS*. U.S. Public Health Service, 1989, pp. 113–159.

[29] Jason Roy, Joseph W. Hogan, and Bess H. Marcus. "Principal stratification with predictors of compliance for randomized trials with 2 active treatments". en. In: *Biostatistics* 9.2 (Apr. 2008), pp. 277–289. ISSN: 1468-4357, 1465-4644. DOI: 10.1093/biostatistics/kxm027. URL: https://academic.oup.com/biostatistics/article/9/2/277/353596 (visited on 04/25/2023).

[30] Lara D. Shore-Sheppard. "The Precision of Instrumental Variables Estimates With Grouped Data". en. In: *Princeton University* (1996). Number: 753 Publisher: Princeton University, Department of Economics, Industrial Relations Section. URL: https://ideas.repec.org//p/pri/indrel/374.html (visited on 05/21/2023).

[31] Rahul Singh and Liyang Sun. *Double Robustness for Complier Parameters and a Semiparametric Test for Complier Characteristics*. arXiv:1909.05244 [cs, econ, math, stat]. Dec. 2022. URL: http://arxiv.org/abs/1909.05244 (visited on 04/18/2023).

[32] Sonja A. Swanson et al. "Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes". In: *Journal of the American Statistical Association* 113.522 (2018), pp. 933–947. ISSN: 0162-1459. DOI: 10.1080/01621459.2018.1434530. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6752717/ (visited on 04/21/2023).

[33] Jeffery M Wooldridge. *Introductory Econometrics:A Modern Approach*. 5th ed. Mason, OH: South-Western Cengage Learning, 2012. ISBN: 978-1-111-53104-1.

7 Appendix

7.1 Proofs of Theorems and Results

Proof. **Result** 3.1, Identification of Complier ATE under Monotonicity

Start with the definition of the intention to treat (ITT effect), which is the average causal effect of the instrument Z on Y. Let $\mathcal{A} = \{0,1\}^2$ denote all possible permutations of the potential treatments

$\{A^{z=0}, A^{z=1}\}$. We can define the ITT effect as a weighted average of effects in the principal strata.

$$\mathbb{E}(Y^{z=1} - Y^{z=0}) = \sum_{(a_1,a_2)\in\mathcal{A}} \mathbb{E}(Y^{z=1} - Y^{z=0}|A^{z=1} = a_1, A^{z=0} = a_2) \cdot \mathbb{P}(A^{z=1} = a_1, A^{z=0} = a_2)$$

By Exclusion Restriction (Assumption 4), we know that Z only affects Y through A. Among always and never takers – those with potential treatments $\{A^{z=0}, A^{z=1}\} = (1,1), (0,0)$ – we can replace $Y^{z=1} - Y^{z=0}$ with $Y^1 - Y^1$ and $Y^0 - Y^0$ in the estimand above. These quantities are both zero.

By monotonicity (Assumption 6), we assume the set of defiers is empty, and therefore does not contribute to the ITT effect.

Thus, we can ITT estimand rewrites as follows:

$$\mathbb{E}(Y^{z=1} - Y^{z=0}) = \mathbb{E}(Y^{z=1} - Y^{z=0}|A^{z=1} = 1, A^{z=0} = 0) \cdot \mathbb{P}(A^{z=1} = 1, A^{z=0} = 0)$$

By Exclusion Restriction (Assumption 4):

$$\mathbb{E}(Y^{z=1} - Y^{z=0}|A^{z=1} = 1, A^{z=0} = 0) = \mathbb{E}(Y^{a=1} - Y^{a=0}|A^{z=1} = 1, A^{z=0} = 0)$$

Which is the average treatment effect among the subpopulation of compliers. Plugging in the above equality and rearranging, by consistency (Assumption 1) and marginal exchangeability (Assumption 5), we obtain:

$$\begin{aligned}\mathbb{E}(Y^{a=1} - Y^{a=0}|A^{z=1} = 1, A^{z=0} = 0) &= \frac{\mathbb{E}(Y^{z=1} - Y^{z=0})}{\mathbb{P}(A^{z=1} = 1, A^{z=0} = 0)} \\ &= \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0))}{\mathbb{P}(A=1|Z=1) - \mathbb{P}(A=0|Z=0)} \quad \text{(Assumptions 1 and 2)}\end{aligned}$$

Which is the typical IV estimand. Under monotonicity, the IV estimand identifies the ATE among the compliers. □

Proof. **Result** 3.2, Instrument Strength and Compliance Score

Recall the definition of the compliance score in Equation 3.2:

$$\gamma(\mathbf{x}) := P(A = 1, \mathbf{X} = \mathbf{x}, Z = 1) - P(A = 1, \mathbf{X} = \mathbf{x}, Z = 0)$$

Also recall the definition of the complier indicator $C := \mathbb{1}(A^{z=1} > A^{z=0})$.

The conditional compliance probability is writable as:

$$\begin{aligned}P(C = 1|\mathbf{X} = \mathbf{x}) &= \mathbb{E}(\mathbb{1}(A^{z=1} > A^{z=0})|\mathbf{X} = \mathbf{x}) \\ &= P(A^{z=1} = 1|\mathbf{X} = \mathbf{x}) - P(A^{z=0} = 1|\mathbf{X} = \mathbf{x}) \\ &= P(A^{z=1} = 1|\mathbf{X} = \mathbf{x}, Z = 1) - P(A^{z=0} = 1|\mathbf{X} = \mathbf{x}, Z = 0) \quad \text{(Assumption 5, Marg. Exch.)} \\ &= P(A = 1|\mathbf{X} = \mathbf{x}, Z = 1) - P(A = 1|\mathbf{X} = \mathbf{x}, Z = 0) \quad \text{(Assumption 1, Consistency)} \\ &= \gamma(\mathbf{x}) \quad \text{(definition)}\end{aligned}$$

The strength of the instrument $\mu = \mathbb{E}(P(C = 1|\mathbf{X} = \mathbf{x})) = \mathbb{E}(\gamma(\mathbf{x}))$ is proven by law of total expectation. □

Proof. **Result** 3.3, Classification Error

$$\begin{aligned}\mathcal{E}(h) &= \mathbb{P}(h \neq C) \quad \text{(Definition class. error)} \\ &= \mathbb{P}(h = 1, C = 0) + \mathbb{P}(h = 0, C = 1) \\ &= \mathbb{E}[\mathbb{P}(h(\mathbf{X}) = 1, C = 0|\mathbf{X} = \mathbf{x}) + \mathbb{P}(h(\mathbf{X}) = 0, C = 1|\mathbf{X} = \mathbf{x})] \quad \text{(Tower)} \\ &= \mathbb{E}[h(\mathbf{X}) \cdot \mathbb{P}(C = 0|\mathbf{X} = \mathbf{x}) + (1 - h(\mathbf{X})) \cdot \mathbb{P}(C = 1|\mathbf{X} = \mathbf{x})|\mathbf{X} = \mathbf{x}] \\ &= \mathbb{E}[h(\mathbf{X}) \cdot (1 - \gamma(\mathbf{X})) + (1 - h(\mathbf{X})) \cdot \gamma(\mathbf{X})] \quad \text{(Result 3.2)}\end{aligned}$$

□

Proof. **Theorem** 3.4 classification error optimality of three classifiers.

Part 1 (Bayes classifier h_0 is classification-error optimal across all estimators). This proof is adapted from (Devroye, László, and Gábor 1996). To show the Bayes classifier, h_0, is optimal with respect to classification error, it suffices to show that:

$$\mathbb{P}(h_0(\mathbf{X}) \neq C) \leq \mathbb{P}(h(\mathbf{X}) \neq C)$$

For any other binary classifier h. Given any deterministic decision rule h, for $\mathbf{X} = \mathbf{x}$, the conditional classification error can be written as:

$$\begin{aligned}
\mathbb{P}(h(\mathbf{X}) \neq C|\mathbf{X} = \mathbf{x}) &= 1 - \mathbb{P}(h(\mathbf{X}) = C|\mathbf{X} = \mathbf{x}) \\
&= 1 - (\mathbb{P}(h(\mathbf{X}) = 1, C = 1|\mathbf{X} = \mathbf{x}) + \mathbb{P}(h(\mathbf{X}) = 0, C = 0|\mathbf{X} = \mathbf{x})) \\
&= 1 - (\mathbb{1}(h(\mathbf{x}) = 1)\mathbb{P}(C = 1|\mathbf{X} = \mathbf{x}) + \mathbb{1}(h(\mathbf{x}) = 0)\mathbb{P}(C = 0|\mathbf{X} = \mathbf{x})) \\
&= 1 - (\mathbb{1}(h(\mathbf{x}) = 1)\gamma(\mathbf{x}) + \mathbb{1}(h(\mathbf{x}) = 0)(1 - \gamma(\mathbf{x})))
\end{aligned}$$

Comparing the risks of the Bayes rule to any other rule, for all $\mathbf{X} = \mathbf{x}$, we obtain:

$$\begin{aligned}
&\mathbb{P}(h(\mathbf{X}) \neq C|\mathbf{X} = \mathbf{x}) - \mathbb{P}(h_0(\mathbf{X}) \neq C|\mathbf{X} = \mathbf{x}) \\
&= \gamma(\mathbf{x})(\mathbb{1}(h_0(\mathbf{x}) = 1) - \mathbb{1}(h(\mathbf{x}) = 1)) + (1 - \gamma(\mathbf{x}))(\mathbb{1}(h_0(\mathbf{x}) = 0) - \mathbb{1}(h(\mathbf{x}) = 0)) \\
&= \ldots + (-1) \times (1 - \gamma(\mathbf{x})) \times (-1) \times [(\mathbb{1}(h_0(\mathbf{x}) = 0) - 1) - (\mathbb{1}(h(\mathbf{x}) = 0) - 1)] \\
&= \gamma(\mathbf{x})(\mathbb{1}(h_0(\mathbf{x}) = 1) - \mathbb{1}(h(\mathbf{x}) = 1)) + (\gamma(\mathbf{x}) - 1)(\mathbb{1}(h_0(\mathbf{x}) = 1) - \mathbb{1}(h(\mathbf{x}) = 1)) \\
&= (2\gamma(\mathbf{x}) - 1)(\mathbb{1}(h_0(\mathbf{x}) = 1) - \mathbb{1}(h(\mathbf{x}) = 1)) \\
&\geq 0
\end{aligned}$$

Because when $\gamma(\mathbf{x}) > 1/2$, $h_0 = 1$, yielding a non-negative risk difference. And when $\gamma(x) \leq 1/2$, $h_0 = 0$, yielding a non-negative risk difference. Thus, h_0 is risk optimal.

Part 2 (Quantile-threshold classifier is classification-error optimal among strength-calibrated rules). First, show the quantile classifier in Equation 4 is indeed strength-calibrated. Assume there exists a unique quantile q s.t. $F(q) = 1 - \mu$ where $F(t) = \mathbb{P}(\gamma(\mathbf{X}) \leq t)$ is the cumulative distribution function of the compliance score. We will be to show that the predicted complier probability equals the strength, the true complier probability, μ.

$$\begin{aligned}
\mathbb{P}(h_q(\mathbf{x}) = 1) &= \mathbb{P}(\mathbb{1}(\gamma(\mathbf{x}) > F^{-1}(1 - \mu)) = 1) \quad \text{(Def classifier)} \\
&= \mathbb{P}(F(\gamma(\mathbf{x})) > (1 - \mu)) \\
&= 1 - (1 - \mu) \quad (F(\gamma) \text{ is uniform}) \\
&= \mu
\end{aligned}$$

Now we will show that h_q is optimal within the class of strength-calibrated classifiers. Let $h : \mathcal{X} \to \{0, 1\}$ be an arbitrary strength-calibrated classifier such that $\mathbb{E}(h) = \mu$ and $h \neq h_q := \mathbb{1}(\gamma > F^{-1}(1-\mu))$. Thus, we can view h as moving some mass from a region R_1 above $F^{-1}(1 - \mu)$ to a region R_0 below $F^{-1}(1 - \mu)$. Formally, for $j = 0, 1$, let $f_j : \mathcal{X} \to \{0, 1\}$ be two classifiers applied to the "shifted" bits of mass. Then we can write the new rule h:

$$\begin{aligned}
h &= \mathbb{1}(\gamma \geq F^{-1}(1 - \mu)) + (f_1 - 1)\mathbb{1}(\gamma \in R_1) + f_0\mathbb{1}(\gamma \in R_0) \\
&= \begin{cases} h_q & \text{if } \gamma \geq F^{-1}(1 - \mu) \text{ and } \gamma \notin R_1 \\ f_1 & \text{if } \gamma \in R_1 \\ f_0 & \text{if } \gamma \in R_0 \end{cases}
\end{aligned}$$

The strength-calibration constraint on h implies that

$$\mathbb{E}(f_0|\gamma \in R_0)\mathbb{P}(\gamma \in R_0) = \mathbb{E}(f_1 - 1|\gamma \in R_1)\mathbb{P}(\gamma \in R_1)$$

Define $R_{1,\min}$ and $R_{1,\max}$ as the infimimum and supremum of compliance scores in their respective sets.

$$R_{1,\min} := \inf\{\gamma : \gamma \in R_1\} \geq \sup\{\gamma : \gamma \in R_0\} := R_{0,\max}$$

Where the inequality holds because R_1 is a collection of mass above $F^{-1}(1-\mu)$ and R_0 is a collection of mass below.

Note that from Result 3.3,

$$\begin{aligned}\frac{\mathcal{E}(h_q)-\mathcal{E}(h)}{2} &= \frac{1}{2}\left(\mathbb{E}[h_q(\mathbf{X})\cdot(1-\gamma(\mathbf{X}))+(1-h_q(\mathbf{X}))\cdot\gamma(\mathbf{X})]-\mathbb{E}[h(\mathbf{X})\cdot(1-\gamma(\mathbf{X}))+(1-h(\mathbf{X}))\cdot\gamma(\mathbf{X})]\right)\\ &= \frac{1}{2}\mathbb{E}\left[(h_q(\mathbf{X})-h(\mathbf{X}))\,(1-\gamma(\mathbf{X}))-\gamma(\mathbf{X})(1-h_q(\mathbf{X})-(1-h(\mathbf{X})))\right]\\ &= \mathbb{E}(\gamma(1-h_q))-\mathbb{E}(\gamma(1-h))=\mathbb{E}(\gamma(h-h_q))\end{aligned}$$

And plugging in our formulas for h and h_q, we obtain:

$$\begin{aligned}\mathbb{E}(\gamma(h-h_q)) &= \mathbb{E}(\gamma(f_1-1)\mathbb{1}(\gamma\in R_1)+\gamma f_0\mathbb{1}(\gamma\in R_0))\\ &= \mathbb{E}(\gamma f_0|\gamma\in R_0)\mathbb{P}(\gamma\in R_0)-\mathbb{E}(\gamma(1-f_1)|\gamma\in R_1)\mathbb{P}(\gamma\in R_1)\\ &\le R_{0,\max}\mathbb{E}(f_0|\gamma\in R_0)\mathbb{P}(\gamma\in R_0)-R_{1,\min}\mathbb{E}(1-f_1|\gamma\in R_1)\mathbb{P}(\gamma\in R_1)\\ &\le 0\end{aligned}$$

By the constraint. Thus, the classification error of h_q minus the classification error of a general strength-calibrated rule h is non-positive, therefore, h_q is risk optimal.

Part 3 (Stochastic classifier is only strength-calibrated and distribution-matched rule.)

Note that any strength-calibrated and distribution-matched classifier has $\mathbb{E}[h(\mathbf{X})]=\mathbb{E}[\gamma(\mathbf{X})]=\mathbb{P}(C=1)=\mu$ and $\mathbb{P}(X\le x|h(\mathbf{X})=1)=\mathbb{P}(X\le x|C=1)$. This together implies that the joint probabilities are equal, and therefore that the marginals are equal:

$$\mathbb{E}(h(\mathbf{X})|\mathbf{X})=\mathbb{P}(C=1|\mathbf{X})$$

Which by Result 3.2 which states $\mathbb{P}(C=1|\mathbf{X})=\gamma(\mathbf{X})$, implies

$$\mathbb{E}(h(\mathbf{X})|\mathbf{X})=\gamma(\mathbf{X})$$

The expected value of decision equals the compliance score, which differs for each unit. Thus, the strength-calibrated and distribution-matched decision rule must be stochastic. $h:\mathcal{U}\to\{0,1\}$ for $\mathcal{U}$ the support of a random variable. However, the only random variable that outputs 0,1 with probability γ is Bernoulli($\gamma(\mathbf{X})$), which is the stochastic classifier as stated.

Noting that $\mathbb{E}(h(\mathbf{X})|\mathbf{X})=\gamma(\mathbf{X})$, the classification error $\mathcal{E}_s$, by the tower law can be written as

$$\begin{aligned}\mathcal{E}_s &= \mathbb{E}[h_s(\mathbf{X})\cdot(1-\gamma(\mathbf{X}))+(1-h_s(\mathbf{X}))\cdot\gamma(\mathbf{X})]\\ &= \mathbb{E}[\mathbb{E}[h_s|\mathbf{X}]\cdot(1-\gamma(\mathbf{X}))+(1-\mathbb{E}[h_s|\mathbf{X}])\cdot\gamma(\mathbf{X})]\\ &= 2\mathbb{E}[\gamma-\gamma^2]\end{aligned}$$

□

Proof. **Theorem** 3.5, relating classification errors

We leverage Theorem 3.1 in Devroye, László, and Gábor (1996) which says for all distributions:

$$\frac{1}{2}(1-\sqrt{1-4\rho^2})\le\mathcal{E}(h_0)$$

Where $\rho=\mathbb{E}\left[\sqrt{\gamma(\mathbf{X})(1-\gamma(\mathbf{X}))}\right]$ is known as the Matushita error.

Recall that $\mathcal{E}_S=2\mathbb{E}[\gamma(\mathbf{X})(1-\gamma(\mathbf{X})]$. By Cauchy-Schwartz:

$$2\mathcal{E}_q\le 2(\mathcal{E}_S)\equiv 4\mathbb{E}[\gamma(\mathbf{X})(1-\gamma(\mathbf{X})]\le 4\mathbb{E}\left[\sqrt{\gamma(\mathbf{X})(1-\gamma(\mathbf{X}))}\right]^2\equiv 4\rho^2$$

Implying:

$$\frac{1}{2}(1-\sqrt{1-2\mathcal{E}_S})\le\frac{1}{2}(1-\sqrt{1-2\mathcal{E}_q})\le\frac{1}{2}(1-\sqrt{1-4\rho^2})$$

Therefore combined with Theorem 3.1.

$$\frac{1}{2}(1-\sqrt{1-2\mathcal{E}_S}) \leq \frac{1}{2}(1-\sqrt{1-2\mathcal{E}_q}) \leq \mathcal{E}(h_0)$$

That $\mathcal{E}(h_0) \leq \mathcal{E}_q \leq \mathcal{E}_S$ follows from the optimality of the Bayes classifier.

To illustrate the upper bound, we rely on the Theorem in Cover and Hart (1967), which says that the asymptotic risk of the nearest neighbor classifier is upper bounded by $2\mathcal{E}(h_0)(1-\mathcal{E}(h_0))$. Since classifiers that make use of the infinite sample set (quantile-threshold and stochastic) will have lower risk that the nearest neighbor classifier but greater than the risk of the Bayes classifier:

$$\mathcal{E}_q \leq \mathcal{E}_S \leq 2\mathcal{E}(h_0)(1-\mathcal{E}(h_0)) \leq 2\mathcal{E}(h_0)$$

□

Proof. **Theorem** 3.6, Excess Error of Plug-in Classifiers

Consider the plug-in quantile-threshold classifier first, $\hat{h}_q$. Note that the classification errors and expectations are with respect to a new observation $\mathbf{O}$ *conditional* on the observed data $\{\mathbf{O}_1, \ldots, \mathbf{O}_n\}$. By the definition of classification error we obtain

$$\begin{aligned}
|\mathcal{E}(\hat{h}_q) - \mathcal{E}_q| &= \left|\mathbb{E}\left[\hat{h}_q + \gamma - 2\hat{h}_q\gamma - (h_q + \gamma - 2h_q\gamma)\right]\right| \\
&= \mathbb{E}\left[|(1-2\gamma)||(\hat{h}_q - h_q)|\right] \qquad \text{(Jensen's inequal)}
\end{aligned}$$

In the next step, we invoke the triangle inequality and a Lemma that says that for real valued f and $\hat{f}$, $|\mathbb{1}(\hat{f} > 0) - \mathbb{1}(f > 0)| \leq \mathbb{1}(|f| \leq |\hat{f} - f|)$. Recalling that $\hat{h}_q := \mathbb{1}(\hat{\gamma} - \hat{q} > 0)$ and $h_q := \mathbb{1}(\gamma - q > 0)$,

$$\begin{aligned}
\mathbb{E}\left[|(1-2\gamma)||(\hat{h}_q - h_q)|\right] &= \mathbb{E}\left[|(1-2\gamma)|\mathbb{1}(|\gamma - q| \leq |\hat{\gamma} - \hat{q} - (\gamma - q)|)\right] \\
&= \mathbb{E}\left[|(1-2\gamma)|\mathbb{1}(|\gamma - q| \leq |\hat{\gamma} - \gamma| + |\hat{q} - q|)\right] \\
&= 2\mathbb{E}\left[|\gamma - q| + |q - 1/2|\mathbb{1}(|\gamma - q| \leq |\hat{\gamma} - \gamma| + |\hat{q} - q|)\right]
\end{aligned}$$

Where the last two steps follow by triangle inequality. Then, recognizing that $|a-b| \leq |(a-c)+(d-b)| \leq |a-c| + |d-b|$ by triangle inequality we have

$$\text{“} = 2\mathbb{E}\left[|\hat{\gamma} - \gamma| + |\hat{q} - q| + |q - 1/2|\mathbb{1}(|\gamma - q| \leq |\hat{\gamma} - \gamma| + |\hat{q} - q|)\right]$$

Which by the Margin Assumption $\mathbb{P}(|\gamma - q| \leq t) \lesssim t^\alpha$ yields

$$|\mathcal{E}(\hat{h}_q) - \mathcal{E}_q| \lesssim (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^\alpha$$

Next, consider the stochastic classifier h_s which has classification error $2\mathbb{E}[\gamma - \gamma^2]$. Then

$$\begin{aligned}
|\mathcal{E}(\hat{h}_s) - \mathcal{E}_s| &= |\mathbb{E}[(\gamma + \hat{h}_s - 2\gamma\hat{h}_s) - 2(\gamma - \gamma^2)] \\
&= |\mathbb{E}[(\gamma + \hat{\gamma} - 2\gamma\hat{\gamma}) - 2(\gamma - \gamma^2)]| \qquad \text{(Tower Law)} \\
&= |\mathbb{E}[(\hat{\gamma} - \gamma)(1-2\gamma)]| \\
&= ||\hat{\gamma} - \gamma|| \cdot ||1 - 2\gamma||_{L^2_{\mathbb{P}_{\mathcal{X}}}} \qquad \text{(Cauchy-Schwarz)}
\end{aligned}$$

Recalling that $||1-2\gamma||_{L^2_{\mathbb{P}_{\mathcal{X}}}} := \sqrt{\mathbb{E}\left[(1-2\gamma)^2\right]} = \sqrt{1 - 4\mathbb{E}[\gamma - \gamma^2]} = \sqrt{1 - 2\mathcal{E}_S}$. This yields

$$|\mathcal{E}(\hat{h}_s) - \mathcal{E}_s| \leq ||\hat{\gamma} - \gamma||\sqrt{1 - 2\mathcal{E}_S}$$

□

Proof. **Theorem** 3.7, Estimating complier characteristics using stochastic classifier

Note that the classification errors and expectations are with respect to a new observation $\mathbf{O}$ *conditional* on the observed data $\{\mathbf{O}_1, \ldots, \mathbf{O}_n\}$. Recall that $\mathbb{E}(h_s|\mathbf{X}) = \gamma$, and suppose we hope to estimate

$\theta = \mathbb{P}(f(\mathbf{X})|C=1) \equiv \frac{\mathbb{P}(f(\mathbf{X}),C=1)}{\mathbb{P}(C=1)}$. Then $|\hat{\theta} - \theta|$ equals

$$\begin{aligned}
\left|\frac{\mathbb{P}_n(f\hat{h}_s)}{\mathbb{P}_n(\hat{h}_s)} - \frac{\mathbb{P}(f\gamma)}{\mathbb{P}(\gamma)}\right| &= \mathbb{P}(\gamma)^{-1}\left|\hat{\theta}(\mathbb{P}(\gamma) - \mathbb{P}_n(\hat{h}_s)) + (\mathbb{P}_n(f\hat{h}_s) - \mathbb{P}(f\gamma))\right| \qquad \text{(Factor)} \\
&= \mathbb{P}(\gamma)^{-1}\left|\hat{\theta}\left[(\mathbb{P} - \mathbb{P}_n)(\hat{h}_s) + \mathbb{P}(\gamma - \hat{h}_s)\right] + \left[(\mathbb{P}_n - \mathbb{P})(f\hat{h}_s) + \mathbb{P}(f(\hat{h}_s - \gamma))\right]\right| \\
&= \mathbb{P}(\gamma)^{-1}\left|\hat{\theta}\left[(\mathbb{P} - \mathbb{P}_n)(\hat{h}_s) + \mathbb{P}(\gamma - \hat{\gamma})\right] + \left[(\mathbb{P}_n - \mathbb{P})(f\hat{h}_s) + \mathbb{P}(f(\hat{\gamma} - \gamma))\right]\right|
\end{aligned}$$

Now we use a lemma stating when $\hat{f}$ is estimated from a prior sample of data, and letting $\mathbb{P}_n$ denote the empirical measure over a new set of observations, then

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = \mathcal{O}_{\mathbb{P}}\left(\frac{||f - \hat{f}||}{\sqrt{n}}\right)$$

Therefore, $(\mathbb{P} - \mathbb{P}_n)(\hat{h}_s) = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ and $(\mathbb{P}_n - \mathbb{P})(f\hat{h}_s) = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$.

By Cauchy-Schwarz, we have that $\mathbb{P}(\gamma - \hat{\gamma}) \leq ||\hat{\gamma} - \gamma||$ and $\mathbb{P}(f(\hat{\gamma} - \gamma)) \leq ||f||_\infty ||\hat{\gamma} - \gamma||$. Rewriting the above

$$|\hat{\theta} - \theta| \leq \mathbb{P}(\gamma)^{-1}\left|\hat{\theta}\left[\mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + ||\hat{\gamma} - \gamma||\right] + \left[\mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + ||f||_\infty ||\hat{\gamma} - \gamma||\right]\right|$$

We know by strong monotonicity (Assumption 6) that $\mathbb{P}(C=1) \equiv \mathbb{E}[\gamma] \geq \epsilon > 0$. Observing that $\hat{\theta} = \mathbb{P}_n(f\hat{h}_s)/\mathbb{P}_n(\hat{h}_s) \leq ||f||_\infty < \infty$ because f is bounded, we obtain

$$|\hat{\theta} - \theta| = \left|\mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + ||\hat{\gamma} - \gamma||\right)\right|$$

□

Proof. **Theorem** 3.8, Bounding causal effects in identifiable subgroups

For any subgroup defined by observed covariates, $g : \mathcal{X} \to \{0, 1\}$, under Assumptions 1-6, we have that

$$\begin{aligned}
\beta(g) &= \mathbb{E}\left(Y^{a=1} - Y^{a=0}|g=1\right) \\
&= \mathbb{E}\left((Y^{a=1} - Y^{a=0})\left\{\left(A^{z=1} - A^{z=0}\right) + (1 - A^{z=1}) + A^{z=0}\right\}|g=1\right) \\
&= \mathbb{E}\left((Y^{a=1} - Y^{a=0})\left\{\mathbb{1}\left(A^{z=1} > A^{z=0}\right) + \mathbb{1}\left(A^{z=1} = A^{z=0} = 0\right) + \mathbb{1}\left(A^{z=1} = A^{z=0} = 1\right)\right\}|g=1\right)
\end{aligned}$$

Where the third equality is by strong monotonicity (Assumption 6). The indicators correspond to causal effects in compliers, never takers, and always takers in the subgroup.

$$\begin{aligned}
\text{“} &= \mathbb{E}\left[\{\mathbb{E}(Y|X, Z=1) - \mathbb{E}(Y|X, Z=0)\} + \left\{(Y^{a=1} - Y^{a=0})((1 - A^{z=1}) + A^{z=0})\right\}\right] \\
&= \mathbb{E}[\{\mathbb{E}(Y|X, Z=1) - \mathbb{E}(Y|X, Z=0)\} \\
&+ \left\{\mathbb{E}[YA|X, Z=0] - \mathbb{E}[Y(1-A)|X, Z=1] + Y^{a=1}(1 - A^{z=1}) - Y^{a=0}A^{z=0}\right\}|g=1]
\end{aligned}$$

Where the above equality follows from IV Assumptions 1-5 and the fact that $Y^{a=0}(1 - A^{z=1}) = Y^{z=1,a=0}(1 - A^{z=1}) = Y^{z=1,a=A^{z=1}}(1 - A^{z=1} = Y^{z=1}(1 - A^{z=1})$ and $Y^{a=1}A^{z=0} = Y^{z=0,a=1}A^{z=0} = Y^{z=0,a=A^{z=0}}A^{z=0} = Y^{z=0}A^{z=0}$.

By unconfoundedness of Z and tower law we obtain

$$\begin{aligned}
\text{“} &= \mathbb{E}[\{\mathbb{E}(YA|X, Z=1) - \mathbb{E}(Y(1-A)|X, Z=0)\} \\
&+ \left\{\mathbb{E}(Y^{a=1}|A^{z=1}=0)\mathbb{E}(1-A|X, Z=1) - \mathbb{E}(Y^{a=0}|A^{z=0}=1)\mathbb{E}(A|X, Z=0)\right\}|g=1]
\end{aligned}$$

Note that two terms $\mathbb{E}(Y^a|A^{z=a} = 1 - a, g = 1)$ for $a = 0, 1$ are not identified by the observed data. If we set them to their extreme values, 0 and 1, we show that $\beta(g)$ is bounded above by

$$\begin{aligned}
\beta_u(g) &= \mathbb{E}[\mathbb{E}(YA|X, Z=1) - \mathbb{E}(Y(1-A)|X, Z=0) + \mathbb{E}(1-A|X, Z=1)|g=1] \\
&\equiv \mathbb{E}[\mathbb{E}(YA + 1 - A|X, Z=1) - \mathbb{E}(Y(1-A)|X, Z=0)|g=1]
\end{aligned}$$

And is bounded below by

$$\begin{aligned}\beta_\ell(g) &= \mathbb{E}[\mathbb{E}(YA|X, Z=1) - \mathbb{E}(Y(1-A)|X, Z=0) + \mathbb{E}(A|X, Z=0)|g=1] \\ &\equiv \mathbb{E}[\mathbb{E}(YA|X, Z=1) - \mathbb{E}(Y(1-A)+A|X, Z=0)|g=1]\end{aligned}$$

□

Proof. Corollary 3.9, Bound length in identifiable subgroups

$$\begin{aligned}\ell(g) &:= \beta_u(g) - \beta_\ell(g) \\ &= \mathbb{E}[\mathbb{E}(YA+1-A|X, Z=1) - \mathbb{E}(Y(1-A)|X, Z=0)|g=1] \\ &- \mathbb{E}[\mathbb{E}(YA|X, Z=1) - \mathbb{E}(Y(1-A)+A|X, Z=0)|g=1] \\ &= \mathbb{E}[\mathbb{E}(1-A|X, Z=1) + \mathbb{E}(A|X, Z=0)|g=1] \\ &= \mathbb{E}[1 - (\mathbb{E}(A|X, Z=1) - \mathbb{E}(A|X, Z=0))|g=1] \\ &= \mathbb{E}[1-\gamma(\mathbf{X})|g=1]\end{aligned}$$

Where the last equality follows form the definition of the compliance score: $\gamma(X) = \mathbb{E}(A|X, Z=1) - \mathbb{E}(A|X, Z=0)$.

Under Assumptions 1-6, we have that $\gamma(\mathbf{X}) = \mathbb{P}(C=1|\mathbf{X})$ and $\mathbb{E}[\gamma] = \mathbb{P}(C=1)$. Therefore, $\mathbb{E}[1-\gamma(\mathbf{X})|g=1] = 1 - \mathbb{P}(C=1|g=1) = \mathbb{P}(C=0|g=1)$ by total probability. □

Result : Subgroup of size t with minimal bound length. Recall that $\ell(g) = \mathbb{E}[1-\gamma(\mathbf{X})|g=1]$ from Corollary 3.9.

$$\begin{aligned}\ell(g) &= 1 - \mathbb{E}[\gamma|g=1] \\ &= 1 - \frac{\mathbb{E}[\gamma g]}{\mathbb{E}[g]} \\ &= 1 + \frac{\mathcal{E}(g) - \mathbb{E}(\gamma+g)}{2\mathbb{E}[g]}\end{aligned}$$

Which can be obtained by recalling the definition of classification error in Equation 2: $\mathcal{E}(g) = \mathbb{E}[\gamma(\mathbf{X})(1-g(\mathbf{X})) + (1-\gamma(\mathbf{X}))g(\mathbf{X})] = \mathbb{E}[\gamma - 2\gamma g + g]$. Algebra yields $-\mathbb{E}[\gamma g] = \mathcal{E}(g) - \mathbb{E}[\gamma+g] + \mathbb{E}[\gamma g]$, implying $-\mathbb{E}[\gamma g] = \frac{\mathcal{E}(g) - \mathbb{E}[\gamma+g]}{2}$.

Minimizing $\ell(g)$ is equivalent to minimizing $\mathcal{E}(g)$ when $\mathbb{E}(g) = t$ is fixed. For rules of a fixed size, we showed previously in Part 2 of Theorem 3.4 that the quantile-classifier, $h := \mathbb{1}[\gamma > F^{-1}(1-t)]$ is optimal with respect to classification error, yielding our desired result. □

Proof. Theorem 3.11: Asymptotic Results for Bound Estimators

Suppose we have two independent samples of size n, where the first sample is used to estimate the nuisances. We can generalize the result to the actual estimator target quantity which is computed using K-fold cross validation, where the nuisances are calculated on the $K-1$ folds and the estimator is evaluated on the K-th fold, then swaps, then averages.

The proposed estimator is as follows

$$\hat{\beta}_j(\hat{h}_q) = \mathbb{P}_n\left[\{\varphi_1(V_{j1};\hat{\eta}) - \varphi_0(V_{j0};\hat{\eta})\}\hat{h}_q(\mathbf{X};\hat{\eta})\right]/\hat{\mu}$$

Where $(\hat{\eta}, \hat{\nu}, \hat{\mu})$ were all constructed on the independent sample.

Let $\varphi_q = \{\varphi_1(V_{j1};\eta) - \varphi_0(V_{j0};\eta)\}h_q(X;\eta)$ and $\hat{\varphi}_q = \{\varphi_1(V_{j1};\hat{\eta}) - \varphi_0(V_{j0};\hat{\eta})\}\hat{h}_q(X;\hat{\eta})$ be the corresponding estimated version. Letting $\phi_\mu = \varphi_1(A;\eta) - \varphi(A;\eta)$ and $\hat{\phi}_\mu$ be the estimated version. Then

$$\hat{\beta}_j(\hat{h}_q) - \beta_j(h_q) = \frac{1}{\mathbb{P}_n\hat{\phi}_\mu}\left((\mathbb{P}_n\hat{\varphi}_q - \mathbb{P}\varphi_q) - \beta_j(h_q)(\mathbb{P}_n\hat{\phi}_\mu - \mathbb{P}\phi_\mu)\right) \qquad (\star)$$

Let's first focus on estimating instrument strength, $\Psi(P) := \mu \equiv \mathbb{E}[\mathbb{E}[\gamma|X]] \equiv \mathbb{E}[\mathbb{E}[A|Z=1, X] - \mathbb{E}[A|Z=0, X]]$. The efficient influence function of this functional is therefore $\phi_\mu = \varphi_1(A;\eta) - \varphi_0(A;\eta)$ where η is a function of the instrument propensity and treatment regression as shown in Equation 12.

Let $\hat{\phi}_\mu = \varphi_1(A;\hat{\eta}) - \varphi_0(A;\hat{\eta})$ be the estimated influence function of μ based on the estimated nuisances.

$$\hat{\mu} - \mu \equiv \mathbb{P}_n\hat{\phi}_\mu - \mathbb{P}\phi_\mu = \underbrace{(\mathbb{P}_n - \mathbb{P})\phi_\mu}_{\text{Term 1}} + \underbrace{(\mathbb{P}_n - \mathbb{P})(\hat{\phi}_\mu - \phi_\mu)}_{\text{Term 2}} + \underbrace{\mathbb{P}(\hat{\phi}_\mu - \phi_\mu)}_{\text{Term 3}}$$

By assumption, $||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| = o_{\mathbb{P}}(1)$. When we inspect Term 2, we see

$$\begin{aligned}
(\hat{\phi}_\mu - \phi_\mu) &\leq ||(\hat{\phi}_\mu - \phi_\mu)|| \\
&= \left|\left| \frac{(2Z-1)(A-\hat{\lambda}_Z)}{\pi_Z \hat{\pi}_Z}(\pi_Z - \hat{\pi}_Z) + \frac{(2Z-1)}{\pi_Z}(\lambda_Z - \hat{\lambda}_Z) + (\hat{\gamma} - \gamma) \right|\right| \\
&= \left|\left| \frac{(2Z-1)(A-\hat{\lambda}_Z)}{\pi_Z \hat{\pi}_Z}(\pi_Z - \hat{\pi}_Z) + \frac{(2Z-1)}{\pi_Z}(\lambda_Z - \hat{\lambda}_Z) + ((\hat{\lambda}_1 - \lambda_1) - (\hat{\lambda}_0 - \lambda_0)) \right|\right| \\
&\lesssim ||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z||
\end{aligned}$$

Where the first equality invokes the form of the efficient influence function and the second invokes the definition of γ and $\hat{\gamma}$.

We assume that $||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| = o_{\mathbb{P}}(1)$. The following Lemma says that for $\hat{f}$ estimated from an outside sample,

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{||\hat{f} - f||}{\sqrt{n}} \right)$$

Letting $f =: 0$ and $\hat{f} =: ||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| = o_{\mathbb{P}}(1)$, yields

$$\begin{aligned}
(\mathbb{P}_n - \mathbb{P})(\hat{\phi}_\mu - \phi_\mu) &\lesssim (\mathbb{P}_n - \mathbb{P})\left(||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| \right) \\
&= O_{\mathbb{P}}\left(\frac{||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z||}{\sqrt{n}} \right) \\
&= o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

Let's inspect the Term 3: $\mathbb{P}(\hat{\phi}_\mu - \phi_\mu)$. Recall the form of the efficient influence function $\varphi_Z(A;\eta)$ that ϕ_μ is built from:

$$\begin{aligned}
\varphi_z(A;\eta) &= \frac{\mathbb{1}(Z=z)}{\pi_z(\mathbf{X})}(A - \mathbb{E}[A|\mathbf{X}, Z=z]) + \mathbb{E}[A|\mathbf{X}, Z=z] \\
&= \frac{\mathbb{1}(Z=z)}{\pi_z(\mathbf{X})}(A - \lambda_z) + \lambda_z
\end{aligned}$$

And $\hat{\phi}_\mu$ is built with plugin estimators for the nuisances. Then term three can be written as:

$$|\mathbb{P}(\hat{\phi}_\mu - \phi_\mu)| = |\mathbb{P}\left(\varphi_1(A;\hat{\eta}) - \varphi_0(A;\hat{\eta}) - (\varphi_1(A;\eta) - \varphi_0(A;\eta))\right)|$$

Focusing in on the φ_1 terms:

$$\begin{aligned}
|\mathbb{P}\left(\varphi_1(A;\hat{\eta}) - \varphi_1(A;\eta)\right)| &= \left| \mathbb{P}\left(\frac{\mathbb{1}(Z=z)}{\hat{\pi}_1(\mathbf{X})}(A - \hat{\lambda}_1) + \hat{\lambda}_1 - \left(\frac{\mathbb{1}(Z=z)}{\pi_1(\mathbf{X})}(A - \lambda_1) + \lambda_1 \right) \right) \right| \\
&= \mathbb{P}\left[-(\lambda_1 - \hat{\lambda}_1) + \left(\frac{\pi_1}{\hat{\pi}_1(\mathbf{X})}(\lambda_1 - \hat{\lambda}_1) - \frac{\pi_1}{\pi_1(\mathbf{X})}(\lambda_1 - \lambda_1) \right) \right] \qquad \text{(Tower Law)} \\
&= \mathbb{P}\left[\frac{-\hat{\pi}_1}{\hat{\pi}_1}(\lambda_1 - \hat{\lambda}_1) + \left(\frac{\pi_1}{\hat{\pi}_1(\mathbf{X})}(\lambda_1 - \hat{\lambda}_1) \right) \right] \\
&= \left| \mathbb{P}\left\{ \frac{\pi_1 - \hat{\pi}_1}{\hat{\pi}_1}(\lambda_1 - \hat{\lambda}_1) \right\} \right|
\end{aligned}$$

The same calculations follow for the φ_0 terms, yielding

$$\begin{aligned}
|\mathbb{P}(\hat{\phi}_\mu - \phi_\mu)| &= \left| \mathbb{P}\left\{ \frac{\pi_1 - \hat{\pi}_1}{\hat{\pi}_1}(\lambda_1 - \hat{\lambda}_1) - \frac{\pi_0 - \hat{\pi}_0}{\hat{\pi}_0}(\lambda_0 - \hat{\lambda}_0) \right\} \right| \\
&= \left| \mathbb{P}\left\{ \frac{\pi_1 - \hat{\pi}_1}{\hat{\pi}_1}(\lambda_1 - \hat{\lambda}_1) - \frac{(1-\pi_1) - (1-\hat{\pi}_1)}{\hat{\pi}_0}(\lambda_0 - \hat{\lambda}_0) \right\} \right| \qquad \text{(Complement probs)} \\
&= \left| \mathbb{P}\left\{ \frac{\pi_1 - \hat{\pi}_1}{\hat{\pi}_1}(\lambda_1 - \hat{\lambda}_1) - \frac{\pi_1 - \hat{\pi}_1}{\hat{\pi}_0}(\lambda_0 - \hat{\lambda}_0) \right\} \right| \\
&\lesssim ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z|| \right)
\end{aligned}$$

Therefore under the assumption that $||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right) = o_{\mathbb{P}}(1)$ and $||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| = o_{\mathbb{P}}(1)$.

$$\hat{\mu} - \mu = (\mathbb{P}_n - \mathbb{P})\phi_\mu + o_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(n^{-1/2}) \quad \text{(WLLN)}$$

Now focusing on $\mathbb{P}_n\hat{\varphi}_q - \mathbb{P}\varphi_q$ term in the expression in $(\star)$. We have

$$\mathbb{P}_n\hat{\varphi}_q - \mathbb{P}\varphi_q = \underbrace{(\mathbb{P}_n - \mathbb{P})\varphi_q}_{\text{Term 1}} + \underbrace{(\mathbb{P}_n - \mathbb{P})(\hat{\varphi}_q - \varphi_q)}_{\text{Term 2}} + \underbrace{\mathbb{P}(\hat{\varphi}_q - \varphi_q)}_{\text{Term 3}}$$

Focusing on Term 2, $(\mathbb{P}_n - \mathbb{P})(\hat{\varphi}_q - \varphi_q)$. By the following Lemma, letting $f =: \varphi_q$ and $\hat{f} =: \hat{\varphi}_q$, yields

$$(\mathbb{P}_n - \mathbb{P})(\hat{\varphi}_q - \varphi_q) = O_{\mathbb{P}}\left(\frac{||\hat{\varphi}_q - \varphi_q||}{\sqrt{n}}\right)$$

And upper bounding $||\hat{\varphi}_q - \varphi_q||$;

$$\begin{aligned}
||\hat{\varphi}_q - \varphi_q|| &= ||\{\hat{\varphi}_1 - \hat{\varphi}_0\}\hat{h}_q - \{\varphi_1 - \varphi_0\}h_q|| \\
&\lesssim ||\{\hat{\varphi}_1 - \hat{\varphi}_0\} - \{\varphi_1 - \varphi_0\}|| + ||\hat{h}_q - h_q|| \\
&\lesssim ||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\nu}_z - \nu_z|| + \mathbb{P}(\hat{h}_q \neq h_q)
\end{aligned}$$

Thus, term 2 $= o_{\mathbb{P}}(n^{-1/2})$ if $||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\nu}_z - \nu_z|| + \mathbb{P}(\hat{h}_q \neq h_q) = o_{\mathbb{P}}(1)$.

Term 3 can be written as

$$\begin{aligned}
\mathbb{P}(\hat{\varphi}_q - \varphi_q) &= \mathbb{P}\left[(\hat{\varphi}_1 - \hat{\varphi}_0)\hat{h}_q - (\varphi_1 - \varphi_0)h_q\right] \\
&= \mathbb{P}\left(((\hat{\varphi}_1 - \hat{\varphi}_0) - (\varphi_1 - \varphi_0))\hat{h}_q + (\varphi_1 - \varphi_0)(\hat{h}_q - h_q)\right)
\end{aligned}$$

The first half of the above term, $\mathbb{P}\left(((\hat{\varphi}_1 - \hat{\varphi}_0) - (\varphi_1 - \varphi_0))\hat{h}_q\right)$ can be written as, using identical steps to the instrument strength case:

$$\begin{aligned}
\mathbb{P}\left(((\hat{\varphi}_1 - \varphi_1) - (\hat{\varphi}_0 - \varphi_0))\hat{h}_q\right) &= \mathbb{P}[\hat{h}_q\{\hat{\pi}_1^{-1}(\pi_1 - \hat{\pi}_1)(\nu_1 - \hat{\nu}_1) + \hat{\pi}_0^{-1}(\pi_1 - \hat{\pi}_1)(\nu_0 - \hat{\nu}_0)\}] \\
&\lesssim ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\nu}_z - \nu_z||\right)
\end{aligned}$$

Thus, the first half of the above term be $o_{\mathbb{P}}(n^{-1/2})$ if $||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\nu}_z - \nu_z||\right) = o_{\mathbb{P}}(n^{-1/2})$. The second half of the expression

$$\begin{aligned}
\mathbb{P}\left((\varphi_1 - \varphi_0)(\hat{h}_q - h_q)\right) &= \mathbb{P}\left(\nu(\hat{h}_q - h_q)\right) \quad \text{(By definition of } \nu\text{)} \\
&\leq \mathbb{P}\left(\nu \,|\mathbb{1}(\hat{\gamma} > \hat{q}) - \mathbb{1}(\gamma > q)|\right) \quad \text{(Def quant-thresh classifier)}
\end{aligned}$$

Recall the following Lemma: for real valued $\hat{f}$ and f,

$$\left|\mathbb{1}(\hat{f} > 0) - \mathbb{1}(f > 0)\right| \leq \mathbb{1}(|f| \leq |\hat{f} - f|)$$

Let's apply this lemma to the previous expression

$$\begin{aligned}
\mathbb{P}\left(\nu \,|\mathbb{1}(\hat{\gamma} > \hat{q}) - \mathbb{1}(\gamma > q)|\right) &\leq \mathbb{P}\left(\nu \mathbb{1}(|\gamma - q| \leq |(\hat{\gamma} - \gamma) - (\hat{q} - q)|)\right) \\
&\lesssim \mathbb{P}\left(|\gamma - q| \leq |(\hat{\gamma} - \gamma) - (\hat{q} - q)|\right) \quad \text{(Bc } \nu \text{ bdd)} \\
&\lesssim \mathbb{P}\left(|\gamma - q| \leq |(\hat{\gamma} - \gamma)| + |(\hat{q} - q)|\right) \quad \text{(Triangle ineq)} \\
&\lesssim (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^\alpha \quad \text{(Margin Assumption)}
\end{aligned}$$

Thus, under the stated conditions, including $(||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^\alpha = o_{\mathbb{P}}(n^{-1/2})$

$$\mathbb{P}_n\hat{\varphi}_q - \mathbb{P}\varphi_q = (\mathbb{P}_n - \mathbb{P})\varphi_q + o_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(n^{-1/2}) \quad \text{(WLLN)}$$

Thus, both terms of $(\star)$ are $o_{\mathbb{P}}(n^{-1/2})$, yielding, under all the conditions stated above including when $R_{1n} := ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z|| + \max_z ||\hat{\nu}_z - \nu_z|| \right) = o_{\mathbb{P}}(n^{-1/2})$, $R_{2n} := ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z|| + \max_z ||\hat{\nu}_z - \nu_z|| \right) = o_{\mathbb{P}}(n^{-1/2})$

$$\hat{\beta}_j(\hat{h}_q) - \beta_j(h_q) = o_{\mathbb{P}}(n^{-1/2})$$

And

$$\sqrt{n}(\hat{\beta}_j(\hat{h}_q) - \beta_j(h_q)) \rightsquigarrow N(0, \mu^{-2} \cdot \text{Var}[(\varphi_1(V_{j,1}) - \varphi_0(V_{j,0}))h_q - \beta_j(h_q)\phi_\mu])$$

Which follows from the asymptotic variance being the square of the influence function for an asymptotically linear estimator. □

Proof. Result 3.12, Sharpness identification

Recall the definition of instrument sharpness.

$$\begin{aligned}
\psi &= \frac{\text{cov}(C, h_q)}{\text{var}(C)} \\
&= \frac{\mathbb{E}[Ch_q] - \mathbb{E}[C]\mathbb{E}[h_q]}{\mathbb{E}[C^2] - \mathbb{E}[C]^2} \\
&= \frac{\mathbb{E}[\mathbb{E}[Ch_q|X]] - \mu^2}{\mu - \mu^2} \quad (C \text{ binary, } h_q \text{ strength calibrated}) \\
&= \frac{h_q\gamma - \mu^2}{\mu - \mu^2} \quad (\text{Bc } \mathbb{E}[C|X] = \gamma(X))
\end{aligned}$$

□

Theorem 3.13, Relationship of sharpness to classification error and bound length. We start by showing the relationship to classification error. We'll show

$$\mathcal{E}(h) = 2\mu(1-\mu)(1-\psi(h)) + (1-2\mu)(\mathbb{E}(h) - \mu)$$

Which reproduces the result in the Theorem because $\mathbb{E}(h_q) = \mu$. First note that

$$\psi(h) = \frac{\text{cov}(C, h_q)}{\text{var}(C)} = \frac{\mathbb{E}((\gamma - \mu)h)}{\mu - \mu^2} = \frac{\mathbb{E}(\gamma h) - \mu\mathbb{E}(h)}{\mu(1-\mu)}$$

Next we show,

$$\begin{aligned}
2\mu(1-\mu)(1-\psi(h)) + (1-2\mu)(\mathbb{E}(h) - \mu) &= 2(\mu - \mu^2 - \mathbb{E}(\gamma h) + \mu\mathbb{E}(h))) + (\mathbb{E}(h) - \mu - 2\mu\mathbb{E}(h) + 2\mu^2) \\
&= \mu + \mathbb{E}(h) - 2\mathbb{E}(\gamma h) \\
&= \mathcal{E}(h)
\end{aligned}$$

from Result 3.3. When $h = h_q$, $\mathbb{E}(h_q) = \mu$, yielding the desired result.

Next we'll show the relationship with bound length. We'll show

$$\ell(h) = (1-\mu)\left(1 - \mu\frac{\psi(h)}{\mathbb{E}(h)}\right)$$

Which yields the desired result when $h = h_q$ implying $\mathbb{E}(h_q) = \mu$. Recognizing that

$$\begin{aligned}
(1-\mu)\left(1 - \frac{\mu}{\mathbb{E}(h)}\psi(h)\right) &= (1-\mu)\left(1 - \frac{\mathbb{E}(\gamma h) - \mu\mathbb{E}(h)}{(1-\mu)\mathbb{E}(h)}\right) \\
&= (1-\mu) - \frac{\mathbb{E}(\gamma h) - \mu\mathbb{E}(h)}{\mathbb{E}(h)} \\
&= 1 - \frac{\mathbb{E}(\gamma h)}{\mathbb{E}(h)} \\
&= 1 - \mathbb{E}(\gamma | h = 1) = \ell(h)
\end{aligned}$$

When $h = h_q$ implying $\mathbb{E}(h_q) = \mu$, we get the desired equality. □

Proof. Theorem 3.14, Asymptotic Results for Instrument Sharpness

We pursue a very similar approach to the proof for Theorem 3.11. Suppose we have independent samples of size n, where the first sample is used to estimate the nuisances to avoid empirical process restrictions. We can generalize the result to the actual estimator target quantity which is computed using K-fold cross fitting, where the nuisances are calculated on the $K-1$ folds and the estimator is evaluated on the K-th fold, then swaps, then averages.

Using logic identical to Theorem 3.11, we show that instrument strength has the following expansion under the assumption $||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| = o_{\mathbb{P}}(1)$:

$$\hat{\mu} - \mu = (\mathbb{P}_n - \mathbb{P})\phi_\mu + \mathcal{O}_{\mathbb{P}}\left(||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right)\right) + o_{\mathbb{P}}(n^{-1/2})$$

Now let $\phi_\xi = \phi_\mu(\mathbf{O}; \eta) h_q(\mathbf{X})$ be the influence function of $\xi = \mathbb{P}(\gamma h_q)$ and $\hat{\phi}_{\xi,-b} = \phi_\mu(\mathbf{O}; \hat{\eta}_{-b})\hat{h}_{q,-b}(\mathbf{X})$ be the estimated influence function leaving out the b-th fold of data. Let $\hat{\xi} = \mathbb{P}_n(\hat{\phi}_{\xi,-B})$ and $\xi = \mathbb{P}(\gamma h_q) = \mathbb{P}(\phi_\xi)$. Write $\hat{\xi} - \xi$ according to the four term expansion

$$\hat{\xi} - \xi = (\mathbb{P}_n - \mathbb{P})(\hat{\phi}_{\xi,-B} - \phi_\xi) + (\mathbb{P}_n - \mathbb{P})\phi_\xi + \mathbb{P}(\hat{h}_{q,-B}(\hat{\phi}_{\mu,-B} - \phi_\mu)) + \mathbb{P}(\phi_\mu(\hat{h}_{q,-B} - h_q))$$

We will study each of these terms separately.

Note that Term 1 $(\mathbb{P}_n - \mathbb{P})(\hat{\phi}_{\xi,-B} - \phi_\xi) = \sum_{b=1}^K (\mathbb{P}_n - \mathbb{P})\left(\left(\hat{\phi}_{\xi,-b} - \phi_\xi\right)\mathbb{1}(B=b)\right)$. Noting that

$$\begin{aligned}
||\left(\hat{\phi}_{\xi,-b} - \phi_\xi\right)\mathbb{1}(B=b)|| &\leq ||\hat{\phi}_{\xi,-b} - \phi_\xi|| \lesssim ||\hat{\phi}_\xi - \phi_\xi|| \\
&= ||\hat{h}_q(\hat{\phi}_\mu - \phi_m u) + \phi_\mu(\hat{h}_q - h_q)|| \lesssim ||\hat{\phi}_\mu - \phi_m u|| + ||\hat{h}_q - h_q|| \\
&\lesssim ||\hat{\pi}_1 - \pi_1|| + \max_z ||\hat{\lambda}_z - \lambda_z|| + \mathbb{P}(\hat{h}_q \neq h_q)
\end{aligned}$$

Implying that the first term is $o_{\mathbb{P}}(n^{-1/2})$.

The second term is a linear term that we want to preserve.

The third term can be written as

$$\begin{aligned}
\mathbb{P}(\hat{h}_{q,-B}(\hat{\phi}_{\mu,-B} - \phi_\mu)) &= \sum_{b=1}^K \mathbb{P}(\hat{h}_{q,-b}(\hat{\phi}_{\mu,-b} - \phi_\mu))\mathbb{P}(B=b) \lesssim \mathbb{P}(\hat{h}_q(\hat{\phi}_\mu - \phi_\mu)) \\
&= \mathbb{P}\left(\hat{h}_q\left(\hat{\pi}_1^{-1}(\pi_1 - \hat{\pi}_1)(\lambda_1 - \hat{\lambda}_1) + \hat{\pi}_0^{-1}(\pi_1 - \hat{\pi}_1)(\lambda_0 - \hat{\lambda}_0)\right)\right) \\
&\lesssim ||\pi_1 - \hat{\pi}_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right)
\end{aligned}$$

The fourth term can be written as

$$\mathbb{P}(\phi_\mu(\hat{h}_{q,-B} - h_q)) = \mathbb{P}(\gamma(\hat{h}_{q,-B} - h_q)) = \mathbb{P}((\gamma - q)(\hat{h}_{q,-B} - h_q)) + q\mathbb{P}(\hat{h}_{q,-B} - h_q)$$

The first term on the far right hand side can be written as

$$\begin{aligned}
\mathbb{P}((\gamma - q)(\hat{h}_{q,-B} - h_q)) &= \sum_{b=1}^K \mathbb{P}((\gamma - q)(\hat{h}_{q,-b} - h_q))\mathbb{P}(B=b) \\
&\lesssim \mathbb{P}((\gamma - q)(\hat{h}_q - h_q)) \leq \mathbb{P}(|\gamma - q||\mathbb{1}(\hat{\gamma} > \hat{q}) - \mathbb{1}(\gamma > q)|) \\
&\leq \mathbb{P}(|\gamma - q|\mathbb{1}(|\gamma - q| \leq |\hat{\gamma} - \gamma| + |\hat{q} - q|)) \\
&\leq \mathbb{P}(|\hat{\gamma} - \gamma| + |\hat{q} - q|\mathbb{1}(|\gamma - q| \leq |\hat{\gamma} - \gamma| + |\hat{q} - q|)) \\
&\lesssim (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^{1+\alpha}
\end{aligned}$$

Where the third inequality follows from a Lemma stating $|\mathbb{1}(\hat{f} > 0) - \mathbb{1}(f > 0)| \leq \mathbb{1}(|f| \leq |\hat{f} - f|)$, the fourth inequality follows from the condition in the indicator, and the fifth follows from the margin assumption.

To pursue the form of the second term on the far right hand side, we start with the instrument strength.

$$\begin{aligned}
\hat{\mu} - \mu &= \mathbb{P}_n(\hat{h}_{q,-B}) - \mathbb{P}(h_q) \\
&= (\mathbb{P}_n - \mathbb{P})(\hat{h}_{q,-B} - h_q) + (\mathbb{P}_n - \mathbb{P})h_q + \mathbb{P}(\hat{h}_{q,-B} - h_q) \\
&= o_{\mathbb{P}}(n^{-1/2}) + (\mathbb{P}_n - \mathbb{P})h_q + \mathbb{P}(\hat{h}_{q,-B} - h_q)
\end{aligned}$$

Where the last line is the result of $||\hat{h}_q - h_q|| = o_{\mathbb{P}}(1)$ by $\mathbb{P}(\hat{h}_q \neq h_q) = o_{\mathbb{P}}(1)$. Rearranging the above display to reflect the second term on the far LHS yields

$$\begin{aligned} q\mathbb{P}(\hat{h}_{q,-B} - h_q) &\leq q(\hat{\mu} - \mu) + (\mathbb{P}_n - \mathbb{P})h_q + o_{\mathbb{P}}(n^{-1/2}) \\ &= q(\mathbb{P}_n - \mathbb{P})(\phi_\mu - h_q) + \mathcal{O}_{\mathbb{P}}\left(||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right)\right) + o_{\mathbb{P}}(n^{-1/2}) \end{aligned}$$

Putting this all together, we have

$$\hat{\mu} - \mu = (\mathbb{P}_n - \mathbb{P})\phi_\mu + \mathcal{O}_{\mathbb{P}}\left(||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right)\right) + o_{\mathbb{P}}(n^{-1/2})$$

$$\hat{\xi} - \xi = (\mathbb{P}_n - \mathbb{P})(\phi_\mu h_q + q(\phi_\mu - h_q)) + \mathcal{O}_{\mathbb{P}}\left(||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right)\right) + \mathcal{O}_{\mathbb{P}}\left((||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^{1+\alpha}\right)$$

Now we notice that we can write sharpness in terms of μ and ξ

$$\hat{\psi} - \psi = \frac{1}{\hat{\mu}(1-\hat{\mu})}(\hat{\xi} - \xi) + \frac{\xi(\hat{\mu} + \mu) - \xi - \mu\hat{\mu}}{\hat{\mu}(1-\hat{\mu})\mu(1-\mu)}(\hat{\mu} - \mu)$$

Now we can establish consistency by noting

$$\hat{\psi} - \psi = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + ||\hat{\pi}_1 - \pi_1|| \left(\max_z ||\hat{\lambda}_z - \lambda_z||\right) + (||\hat{\gamma} - \gamma||_\infty + |\hat{q} - q|)^{1+\alpha}\right)$$

To establish asymptotic normality under the conditions of terms above going to 0 at root-n rate, we have

$$\begin{aligned} \sqrt{n}\left(\begin{pmatrix}\hat{\mu}\\ \hat{\xi}\end{pmatrix} - \begin{pmatrix}\mu\\ \xi\end{pmatrix}\right) &= \sqrt{n}(\mathbb{P}_n - \mathbb{P})\begin{pmatrix}\phi_\mu \\ \phi_\mu h_q + q(\phi_\mu - h_q)\end{pmatrix} + o_{\mathbb{P}}(n^{-1/2}) \\ &\rightsquigarrow N\left(0, \text{Cov}\begin{pmatrix}\phi_\mu \\ \phi_\mu h_q + q(\phi_\mu - h_q)\end{pmatrix}\right) \end{aligned}$$

By the Delta method, we obtain

$$\hat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})\left(\frac{\phi_\mu h_q + q(\phi_\mu - h_q) - \xi}{(\mu - \mu^2)} + \frac{2\mu\xi - \xi - \mu^2}{(\mu - \mu^2)^2}(\phi_\mu - \mu)\right) + o_{\mathbb{P}}(n^{-1/2})$$

Thus, sharpness is an asymptotic linear estimator that has limiting distribution

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N\left(0, \text{Var}\left(\frac{\phi_\mu h_q + q(\phi_\mu - h_q) - \xi}{(\mu - \mu^2)} + \frac{2\mu\xi - \xi - \mu^2}{(\mu - \mu^2)^2}(\phi_\mu - \mu)\right)\right)$$

□