# Learning From Imperfect Research Designs: Automating Causal Inference When Classic Assumptions Fail*

Kai R. D. Cooper
kaicoop@wharton.upenn.edu

Guilherme Duarte†
gjduarte@fas.harvard.edu

Luke Keele
luke.keele@uphs.upenn.edu

Dean Knox
dcknox@upenn.edu

Jonathan Mummolo
jmummolo@princeton.edu

June 11, 2026

Word Count: 9,557

**Abstract**

Social science has developed an expansive design-based toolkit for causal inference, but key assumptions often fail in real-world settings. Partial identification offers an alternative: researchers can learn as much as possible through sharp bounds while transparently acknowledging limitations of data and design. We propose methodological improvements to automated partial identification that make it viable for applied social-science research, including new approaches for quantifying uncertainty, adjusting for covariates, handling continuous variables, assessing the consequences of relaxing or falsifying assumptions, and interpreting why bounds are narrow or wide. We then replicate and extend published studies spanning several causal designs to show how these advances deepen our understanding of empirical robustness. In some applications, implausible assumptions directly drive key causal claims; in another, they are inconsistent with observed data. Among other results, we present updated findings on counterinsurgency violence, voter habit formation, and racial bias in policing.

*Keywords:* causal inference, partial identification, assumption testing, robustness

**AI Disclosure:** The authors used AI-assisted tools at various stages of manuscript preparation, including ChatGPT, OpenAI, GPT-5.5 Thinking; Claude Sonnet 4, Anthropic; and

*Kai Cooper is a Ph.D. student in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Guilherme Jardim Duarte is a postdoctoral fellow and incoming assistant professor in the Department of Government at Harvard University. Luke Keele is a Research Professor of Statistics in Surgery, Perleman School of Medicine, University of Pennsylvania. Dean Knox is an assistant professor in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Jonathan Mummolo is an associate professor of Politics and Public Affairs, Princeton University. We gratefully acknowledge financial support from AI for Business and the Analytics at Wharton Data Science and Business Analytics Fund. This research was made possible in part by grants from the Carnegie Corporation of New York and Arnold Ventures under Grant 22-06762. The statements made and views expressed are solely the responsibility of the authors. Authors listed in alphabetical order.

†Corresponding author.

Codex, OpenAI. These tools were used for copyediting, formatting suggestions, manuscript-submission checks, coding assistance, debugging, and assistance with figures and plots. All AI-assisted outputs, including text, code, analyses, and figures, were reviewed, verified, and edited by the authors, who remain responsible for the accuracy, validity, and final content of the manuscript.

# Contents


**1 Introduction** **1**

**2 Improving Automated Partial Identification with `autobounds`** **3**
2.1 Preliminaries . . . 3
2.2 An Improved Approach to Uncertainty Quantification . . . 7
2.3 New Theory and Methodology for Covariate Adjustment . . . 9
2.4 New Theory and Methodology for Continuous Outcomes . . . 11
2.5 New Diagnostic Tools and Sensitivity Analyses . . . 13
2.6 A Worked Example of Automated Partial Identification for Confounding . . . 14

**3 Testing Empirical Implications of Theorized Assumptions** **19**
3.1 Replication and Extension of Kocher et al. (2011) . . . 21

**4 Probing Sensitivity to Assumption Violations** **23**
4.1 Replication and Extension of Coppock and Green (2016) . . . 23

**5 Weighing Relative Importance of Assumptions** **31**
5.1 Replication and Extension of Knox et al. (2020) . . . 31

**6 Discussion and Conclusion** **36**

**A Detailed Example with `autobounds`** **44**
A.1 Simulated Data for Section 2.6 . . . 50
A.2 R Code for Section 2.6 . . . 52

**B An Application to Difference in Differences** **53**
B.1 Replication and Extension of Schubiger (2021) . . . 54
B.2 Code . . . 57

**C Statistical Uncertainty and Covariate Adjustment** **57**
C.1 Preliminaries . . . 57
C.2 Statistical Uncertainty . . . 59
C.2.1 Uncertainty Quantification without Covariates . . . 59
C.3 Simulation and Coverage . . . 63
C.4 Covariate Adjustment . . . 64
C.4.1 Validity of Covariate-averaged Bounds. . . . 69

**D Model specification for Kocher et al. (2011)** **74**

**E Code** **75**
E.1 Instrumental Variables . . . 75
E.2 Selection Bias . . . 77

# 1 Introduction

Social scientists now possess an expansive set of tools for causal inference. But applied work is messy: real-world scenarios often diverge from the ideal conditions under which standard approaches yield reliable estimates of causal quantities. Widely used identifying assumptions, like the "exclusion restriction" in instrumental-variables analysis, are often difficult to defend outside the tidy examples found in statistics textbooks (Bazzi and Clemens, 2013; Gallen, 2020). Applied researchers have long had limited options: ignore the problem and present unreliable results, narrow the focus to questions of lesser importance, or abandon projects altogether.

A recent advancement in automated partial identification, the `autobounds` algorithm developed in Duarte et al. (2024), offers an alternative. With this approach, users declare a causal quantity of interest (i.e. an estimand, like an average treatment effect), state assumptions, and provide data while acknowledging its limitations through explicit, tailored assumptions. Using a "branch-and-bound" optimization technique (Vigerske and Gleixner, 2018; Gamrath et al., 2020; Belotti et al., 2009), the algorithm then searches over possible data-generating processes (DGPs) and locates those consistent with stated assumptions and observed data. When complete, it outputs either *sharp bounds* on the estimand—best- and worst-case scenarios—or, if one exists, a point-identified solution. As Duarte et al. (2024) states, "This approach can accommodate scenarios involving any classic threat to inference, including but not limited to missing data, selection, measurement error, and noncompliance" (p. 1778). Importantly, `autobounds` also allows users to partially relax or abandon assumptions and re-compute bounds. However, the initial methodology had a number of limitations that impeded its use in applied work—not least, the fact that it was initially suitable only for settings with discrete data.

In this paper, Section 2 first reviews the fundamentals of partial identification in causal inference. We then extend `autobounds` in several ways. (1) We develop an uncertainty-

quantification procedure that sharply tightens the original's overly wide intervals. (2) We propose a practical, model-based approach for adjusting for background covariates, including continuous ones. (3) We extend the algorithm to bound continuous outcomes via increasingly fine-grained bins, akin to Riemann summation. (4) We provide new procedures for interpreting why conclusions are or are not informative: understanding empirical implications of theorized assumptions that are falsified by data; relaxing a wide range of assumptions using a new, general approach to sensitivity analyses; and probing best-/worst-case DGPs that are observationally equivalent and cannot be ruled out using available data.

Subsequent sections illustrate these techniques in applications covering several research designs and substantive subfields.[1] In Section 3, we replicate Kocher et al. (2011), which examines the effect of bombing campaigns in Vietnam. We use `autobounds` to reveal faulty theory: the algorithm reports that the IV exclusion restriction is incompatible with the observed data, despite the widespread perception that this assumption is "not testable" (Imbens and Angrist, 1994, p. 468). In Section 4, we reanalyze Coppock and Green (2016), a study of voter habit formation, using our new sensitivity analyses to gradually relax the exclusion restriction and allow for a likely violation in the electoral context studied. We show that when allowing for even a slight direct effect by the randomized get-out-the-vote encouragement, the study's core claim about habitual voting can no longer be supported. Section 5 then reexamines Knox et al. (2020), showing how our interpretative techniques can help researchers weigh the consequences of different identifying assumptions. We find that conclusions about racial bias in police use of force hinge more on an assumption about the nature of bias in officer's initial stopping decision, rather than another assumption about how force is distributed across different types of police-civilian encounters. Finally, in Appendix B, we extend Schubiger (2021) to demonstrate the robustness of difference-in-difference results on community mobilization against state violence.

We conclude in Section 6 by arguing that this new tool allows social scientists to flexibly

[1] All code to execute these replications appears in the Appendix.

confront the idiosyncracies of applied research without relying on the often-implausible assumptions that come bundled with off-the-shelf causal-inference designs. Instead, analysts can invoke only the assumptions plausible in their setting while acknowledging design and data limitations. This approach also addresses, at least in part, a longstanding critique of the credibility revolution: the concern that modern causal techniques lead to narrow or misguided lines of inquiry (Monroe, 2005; Deaton, 2010). With automated partial identification, researchers can more easily conduct *question-driven* research. The flexibility of this tool means the statistical quantity of interest—the research question, formally stated—need not be retroactively adjusted to suit a particular design or applied setting.[2] The goal of a study need not change to accommodate the method being used.

# 2 Improving Automated Partial Identification with `autobounds`

We first review the `autobounds` framework and introduce our methodological improvements: new approaches to statistical inference, covariate adjustment, continuous outcomes, and interpretability. These improvements are illustrated through applications presented in Sections 3–5). To guide readers in the use of the `autobounds` software package, the section concludes with a simulated example that bounds the average treatment effect in the familiar setting where researchers are concerned about omitted-variable bias in a selection-on-observables design.

## 2.1 Preliminaries

We first establish notation, review basic concepts in causal inference,[3] and demonstrate a simple partial identification problem using `autobounds`. To use our approach, the researcher must first specify a target quantity or causal *estimand* $\varphi$: a comparison of average counterfactual quantities. Estimands, which precisely define the scientific question, are distinct from

[2]For example, researchers can use this tool to easily bound the average treatment effect in an entire population, where the standard instrumental-variables approach could only target a *local* average treatment effect among compliers.

[3]See Keele (2015) for a more detailed review.

the *estimators*, or statistical methods that may be used to answer the question; they are also distinct from the *estimates*, or numeric values obtained when applying a particular estimator to available data. Though estimands are often left undefined in applied research (Lundberg et al., 2021), this step is indispensable: without the target quantity, one cannot tell whether a design functions as intended.

One way to define estimands is using the potential outcomes framework (Rubin, 1974). Potential outcomes are unit-level attributes representing the counterfactual result that would appear in the presence or absence of treatment; in contrast, the actual outcome depends on whether or not treatment was actually received. We denote treatment by $D$ and assume it is binary unless noted, though the approach generalizes to categorical or ordinal treatments. The potential outcomes are $Y(d)$, with $d$ representing possible treatment values, and the actual outcome $Y$ is a function of treatment assignment and potential outcomes such that $Y = Y(D) = D \cdot Y(d = 1) + (1 - D)Y(d = 0)$. In this framework, one possible estimand is the average treatment effect (ATE):

$$\text{ATE} = \mathbb{E}\left[Y(d = 1) - Y(d = 0)\right],$$

which is the difference in each unit's potential outcomes under treatment and control, averaged over the entire population of interest.

The central challenge in causal inference is that causal estimands contain elements that are fundamentally unobservable. For example, if unit $i$ received treatment $D_i = 1$, then the factual outcome $Y_i = Y_i(D_i) = Y_i(d = 1)$ is observed, but the potential outcome $Y_i(d = 0)$ in the counterfactual world where the unit was untreated is unobserved. Addressing this problem requires an *identification strategy*, a set of formal assumptions that allow for the estimation of unobservable quantities from observed data (Angrist and Pischke, 2010).

One tool that can be utilized for identification analysis is the language of causal graphs (Pearl, 1995a, 2009). Graphs are visual representations of causal theories. Their structure

determines whether a causal effect is nonparametrically identified—i.e., can be estimated without functional-form assumptions such as no-defiers or parallel-trends assumptions (formally defined in subsequent sections)—or whether these additional assumptions might be required. Causal graphs offer a compact and efficient approach to summarizing key information about research designs, and they form an essential part of our proposed methodology.

Most identification strategies are designed to achieve *point identification*, recovering a single, unique value for e.g. the causal effect of a treatment on an outcome. One alternative is partial identification (Manski, 1990, 1995), which instead seeks to recover the best- and worst-case scenarios—upper and lower bounds—that form an interval of possible values for that causal effect. These bounds are *sharp*, meaning they provably cannot be narrowed without further assumptions or data. With these bounds, analysts can simply decline to invoke assumptions that are indefensible—producing a wider range of possible solutions that may nevertheless be sufficient to answer substantive questions such as whether a particular effect is positive, negative, or indistinguishable from zero. Partial identification allows analysts to navigate the trade-off between plausibility of assumptions and informativeness of bounds by using a nested series of models that add assumptions one at a time, obtaining successively narrower ranges of possible answers on the quantity of interest if the assumptions are true. This approach clarifies the relationship between the strength of causal modeling assumptions and the amount of information about the quantity of interest that can be extracted from available data. (For examples in political science, see Mebane and Poast, 2013 or Keele and Minozzi, 2012.)

While partial identification precisely characterizes causal effects under incomplete information or questionable designs, deriving sharp bounds is often analytically intractable. To address this, Duarte et al. (2024) provides an easy-to-use algorithm, `autobounds`, to automatically compute sharp bounds for causal research questions. For details on how this algorithm functions, we refer readers to Duarte et al. (2024) and the "Problem Formulation" sections in

this paper's Appendix; here, we seek to convey high-level intuition.

At the heart of this procedure is the concept of principal stratification (Frangakis and Rubin, 2002), which characterizes units by their essential types based on how they respond counterfactually as model variables change. Perhaps the most well-known example of principal stratification appears in the instrumental variables approach outlined in Angrist et al. (1996). In this framework, an exogenous instrument, $Z$, is thought to encourage treatment, $D$, which in turn affects an outcome $Y$. Given a dichotomous instrument and treatment, Angrist et al. (1996) describes four principal strata: "always takers," units which would accept treatment regardless of the value of $Z$; "never takers," units which would never accept treatment; "compliers," units which accept treatment if encouraged by the instrument $Z$ but not otherwise; and "defiers," units which accept treatment in the absence of encouragement by $Z$, and reject treatment if encouraged.

While this classic setup identifies four principal strata based on how the treatment responds to the instrument, it is possible to represent any discrete causal model in terms of principal strata based on how every relevant variable in the system could possibly respond under various scenarios. As systems grow more complex, the number of strata grows explosively. However, so long as all variables in the system are discrete, the number of principal strata will be countable and finite, since there are only so many ways that a discrete variable can respond to other potentially manipulable discrete variables that are causally upstream.[4]

Building on this intuition, `autobounds` works by efficiently enumerating all the principal strata implied by a causal model. The software takes four inputs: (1) a causal estimand or target quantity, such as the ATE; (2) a causal theory, represented in a DAG; (3) any additional functional-form assumptions not captured in the DAG, e.g. "no defiers" or "parallel trends"; and (4) observed data distributions.[5] The software then expresses the causal estimand in terms

[4]This remains true even in the case where unobserved confounders are continuous or high-dimensional, because the class of estimands that we consider never involve manipulation of these unobserved confounders. Rather, the research questions that are typically asked involve holding these unobserved confounders fixed, only manipulating the discrete "main variables" such as treatment assignment.

[5]The user can also specify two tolerance parameters that control the amount of computation time, corre-

of the sizes of these principal strata. For example, the first stage of an instrumental-variables analysis—the increase in treatment uptake caused by encouragement, $\mathbb{E}[D(z=1) - D(z=0)]$—can be expressed as the size of the complier group minus the size of the defier group.[6] Next, it translates causal assumptions and observed data into constraints on the possible sizes of principal strata.[7] Duarte et al. (2024) prove that sharp bounds can be obtained for essentially any quantity of interest in any discrete causal graph by solving the resulting optimization problem—i.e., maximizing and minimizing the estimand, subject to constraints imposed by assumptions and data.

Originally, `autobounds` could only handle settings in which the modeled variables were discrete. In this paper, however, we extend the framework in two directions. First, we introduce a procedure for adjusting for continuous and potentially high-dimensional covariates; see Subsection 2.3. Second, we introduce a conservative discretization-based approach for approximating continuous-outcome problems; see Subsection 2.4. Together these extensions broaden `autobounds`' applicability while preserving its core logic.

We demonstrate the general `autobounds` approach in the following section with a simple coded example. For a step-by-step mathematical explanation of the optimization performed by `autobounds`, see Appendix A.

## 2.2 An Improved Approach to Uncertainty Quantification

Sharp bounds on a causal estimand $\varphi$, which we denote $[\underline{\varphi}, \overline{\varphi}]$, represent structural uncertainty that persists even with infinite data, owing to mismatches between the ideal and the sampled data (e.g., treatment-outcome confounding, nonrandomly missing outcomes). In applied work, however, researchers must also address *statistical uncertainty* arising from finite samples. That

sponding to the desired level of provable sharpness and width of bounds, beyond which further computation is deemed unnecessary. See discussion of $\epsilon^{\text{thresh.}}$ and $\theta^{\text{thresh.}}$ in Duarte et al. (2024).

[6] This is because $\mathbb{E}[D(z=1)] = \Pr(\text{always}) + \Pr(\text{complier})$ and $\mathbb{E}[D(z=0)] = \Pr(\text{always}) + \Pr(\text{defier})$.

[7] Crucially, it also uses automated graphical and computer-algebra techniques to simplify the problem by eliminating redundant information and strata that cannot possibly exist given the stated assumptions and observed data.

is, one must not only estimate the bounds $[\hat{\underline{\varphi}}, \hat{\overline{\varphi}}]$, but also construct confidence regions for them. The original KL-divergence approach in Duarte et al. (2024) was highly conservative, often yielding unusably wide regions. Here, we develop a new procedure based on *recentered subsampling*. When evaluating this new procedure using the same simulations as Duarte et al. (2024), we demonstrate that the new approach reduces the width of the statistical-uncertainty component by a factor of 3.8 to 7.1 times, leading to far greater statistical power.

Inference on bounds is challenging. First, one must clarify what the intervals are intended to cover. As Imbens and Manski (2004) note, one can construct confidence intervals either for the partially identified region or for the parameter of interest itself.[8] The 95% confidence bounds that we construct are of the form $[\hat{\underline{\varphi}} - \underline{C}, \hat{\overline{\varphi}} + \overline{C}]$, where positive confidence terms $\underline{C}$ and $\overline{C}$ widen the estimated bounds to the estimated 0.025 and 0.975 sampling quantiles of $\hat{\underline{\varphi}}$ and $\hat{\overline{\varphi}}$, respectively.[9] Second, standard frequentist asymptotic confidence intervals are generally invalid in this setting because bounds are typically defined as extremum functionals (maxima or minima) of multiple parameters. Such functionals are nonsmooth, and classical tools such as the delta method or the bootstrap may fail, often producing intervals that substantially undercover the true bounds (Andrews and Han, 2009; Bugni, 2010; Canay, 2010). To address this nonsmoothness, we employ subsampling, which is frequently used for related problems with moment inequalities (Chernozhukov et al., 2007; Romano and Shaikh, 2010). We approximate the statistic's distribution using smaller subsamples of the data. Unlike the bootstrap, which relies on smooth local linear approximations, subsampling does not require differentiability of the bound map and is therefore attractive in nonsmooth settings such as this one (Politis et al., 2001). This requires some additional notation. Given the initial sample of size $n$, let $[\hat{\underline{\varphi}}, \hat{\overline{\varphi}}] = [\hat{\underline{\varphi}}_n, \hat{\overline{\varphi}}_n]$ be the bound estimates computed from the full sample. To construct confidence intervals, we use *recentered subsampling*. Let $m$ denote the

[8] Typically, the latter are less conservative than the former for a fixed probability level $1-\alpha$. Intuitively, if the true parameter lies near the upper boundary of the identified set, the lower endpoint of the interval will cover it with probability strictly greater than $1-\alpha$, making such intervals conservative for the parameter.

[9] If those one-sided procedures are valid for the lower and upper endpoints, then the Bonferroni construction yields an interval for the partially identified region, and therefore also for the true parameter.

subsample size,[10] obtained by rounding $n^\gamma$ downward; following convention, we suggest a default of $\gamma = 2/3$, though `autobounds` checks various safeguards and allows alternate values. We draw $B$ subsamples without replacement from the original sample, each of size $m$. For each subsample $b \in \{1, \ldots, B\}$, we recompute the bound estimates, $\left[\hat{\underline{\varphi}}_m^{(b)}, \hat{\overline{\varphi}}_m^{(b)}\right]$. We then form the recentered statistics

$$\underline{T}^{(b)} = \sqrt{m}\left(\hat{\underline{\varphi}}_m^{(b)} - \hat{\underline{\varphi}}_n\right), \qquad \overline{T}^{(b)} = \sqrt{m}\left(\hat{\overline{\varphi}}_m^{(b)} - \hat{\overline{\varphi}}_n\right).$$

and denote their empirical quantile functions as $Q_{\underline{T}}$ and $Q_{\overline{T}}$. Then the reported confidence bounds for the identified set are

$$\left[\hat{\underline{\varphi}}_n - \frac{Q_{\underline{T}}(1-\alpha/2)}{\sqrt{n}},\ \hat{\overline{\varphi}}_n - \frac{Q_{\overline{T}}(1-\alpha/2)}{\sqrt{n}},\right]$$

In simulations (Appendix C.3), subsampling yields coverage comparable to oracle estimates, which use the true data-generating process and are infeasible in practice, which exploit knowledge of the true data-generating process and are infeasible in actual applied research. We view these simulations as evidence of practical usefulness, especially where nonsmoothness undermines the bootstrap, not as a coverage proof. A limitation of subsampling is that using smaller subsamples may degrade performance in very small samples ($n \leq 100$).

## 2.3 New Theory and Methodology for Covariate Adjustment

Many applications include background covariates for which researchers want to adjust but which are not the object of causal interest. In such settings, the target is typically a marginal estimand that averages over the covariate distribution, such as ATE $= \mathbb{E}_{\boldsymbol{X}}[\text{ATE}_{\boldsymbol{X}}]$. We therefore treat the covariates as background variables that come first in the causal ordering:

[10] Though $m$ will grow with $n$, we temporarily suppress this dependence in the main text.

they may cause the other modeled variables, but are not themselves caused by them.[11]

Under these assumptions, when the estimand is collapsible over the covariate distribution, the marginal lower and upper bounds are obtained by averaging the covariate-specific sharp bounds. Writing $\underline{\varphi}(\boldsymbol{x})$ and $\overline{\varphi}(\boldsymbol{x})$ for the sharp lower and upper bounds conditional on $\boldsymbol{X} = \boldsymbol{x}$, the corresponding marginal bounds are

$$\int \underline{\varphi}(\boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}) \qquad \text{and} \qquad \int \overline{\varphi}(\boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}).$$

For exact discrete stratification, the key point is that, given the assumptions above, averaging preserves sharpness rather than merely validity, so the covariate-adjusted bounds remain the sharp bounds for the marginal estimand. Appendix C.4 gives the formal argument. By contrast, the model-assisted binning procedure introduced below for continuous or high-dimensional covariates is an approximation to this exact stratified target.

When the number of unique covariate values is small, `autobounds` can compute these bounds separately within each level of $\boldsymbol{X}$ and then average them directly. The complication is that continuous or high-dimensional covariates make exact stratification infeasible. Our solution is a model-assisted covariate adjustment procedure. We first estimate the conditional observed-data law $P(\boldsymbol{V} \mid \boldsymbol{X} = \boldsymbol{x})$ using a multinomial model, treating this first-stage fit as a smoothing device rather than as a structural model of substantive interest. We then map each observation to its fitted conditional law, summarize those fitted laws with a low-dimensional score, partition the sample into a manageable number of bins using empirical quantiles of that score, and compute lower and upper bounds within each bin. The stratum-specific bounds are aggregated using the bins' empirical frequencies. In this sense, the method induces a matching-like subclassification on units with similar fitted conditional laws while keeping the

[11] This is not as strong as it may sound, because it is mainly an ordering assumption. The more restrictive content lies in the exclusion restrictions given by the absent arrows into the covariates; by contrast, drawing arrows out of the covariates is comparatively weak, because the presence of an arrow also covers the limiting case in which the corresponding effect is in fact zero.

optimization problem tractable.

To quantify uncertainty, we extend the recentered subsampling procedure described above. On each subsample, the model is refit, bins are rebuilt, and bounds are recomputed. This propagates uncertainty from the covariate model, the subclassification, and the bounding step in one procedure. We regard this approach as a practical generated-regressor procedure, not a fully nonparametric solution. The formal result in the appendix concerns the population covariate-averaged bounds; by contrast, we do not provide a general coverage theorem here for the full fitted-and-binned sampling procedure.

## 2.4 New Theory and Methodology for Continuous Outcomes

Continuous outcomes are difficult to handle in a nonparametric framework for at least two reasons. First, if outcome $Y$ is truly unbounded, then best-/worst-case reasoning becomes vacuous: if even a single observation has a missing value, then without additional restrictions, the sharp lower and upper bounds on many mean-based estimands will be $-\infty$ and $\infty$. And second, then there can be infinitely many types of units with very different potential responses, which undermines the finite principal-strata representation on which `autobounds` relies.[12]

To see the problem, let $F_d(y) = \Pr(Y(d) \leq y)$ denote the counterfactual cumulative distribution function of the potential outcome under treatment level $d$. For any fixed cutoff $y$, this quantity can in principle be bounded directly with `autobounds`, since it is just the probability of the event $\mathbf{1}\{Y(d) \leq y\} = 1$. The difficulty arises when evaluating the estimand as a functional of the outcome distribution. For example, when the estimand involves a mean such as $\mathbb{E}[Y(d)]$, bounding requires reasoning not only about the proportion of units above/below the cutoff $y$, but also weighting these events by the truncated means $\mathbb{E}[Y(d) \mid Y(d) \leq y]$ and $\mathbb{E}[Y(d) \mid Y(d) > y]$.

The bounds are then approximated by a conservative procedure. We first cut the contin-

[12]This challenge is exacerbated if the conditional-expectation function of $Y$ is allowed to be non-smooth.

uous outcome into an ordered discrete approximation. By default, the software partitions the empirical distribution of $Y$ into $n = 5$ quantile-based bins, though the number of bins can be changed by the user. Let these bins be indexed by $j = 1, \ldots, n$, and let $y_j^{\min}$ and $y_j^{\max}$ denote the observed lower and upper endpoints of bin $j$. We input the problem into autobounds as if the outcome were categorical with $n$ levels. Notice that this procedure does not directly give the conservative approximation. The key idea of the conservative method is to decompose a more complex estimand into simpler components that `autobounds` can bound directly. In the ATE case, this means separately bounding the treated and control potential-outcome means and then differencing them at the end. Let $\underline{\tau}_{dj}$ and $\overline{\tau}_{dj}$ denote the resulting lower and upper bounds for the event $\mathbf{1}\{Y(d) \geq j\}$ for treatment state $d \in \{0, 1\}$.

These threshold bounds are then aggregated conservatively to bound each potential-outcome mean. For a given treatment level $d$, the lower bound uses the lower endpoint differences of the bins,

$$\underline{\mu}_d = y_1^{\min} + \sum_{j=2}^{n} \left(y_j^{\min} - y_{j-1}^{\min}\right) \underline{\tau}_{dj},$$

while the upper bound uses the upper endpoint differences,

$$\overline{\mu}_d = y_1^{\max} + \sum_{j=2}^{n} \left(y_j^{\max} - y_{j-1}^{\max}\right) \overline{\tau}_{dj}.$$

The ATE bounds are then obtained by subtraction,

$$\underline{\mu}_1 - \overline{\mu}_0 \leq \mathbb{E}[Y(1) - Y(0)] \leq \overline{\mu}_1 - \underline{\mu}_0.$$

This is computationally attractive: autobounds replaces continuous outcomes with a discrete approximation. The resulting interval should therefore be interpreted as an approximation to the continuous-outcome estimand, rather than an exact bound for the original continuous problem. Two distinct sources of uncertainty are therefore present. First, there is statistical uncertainty from finite-sample estimation of the discretized bounds. Second, there

is approximation error from replacing the original continuous outcome with the chosen partition. The conservativeness established here concerns the coarsened representation induced by the partition: because the aggregation step uses lower bin endpoints for lower bounds and upper bin endpoints for upper bounds, the resulting interval is conservative relative to that coarsened problem. We do not provide a general bound here on the approximation error between the coarsened estimand and the original continuous-outcome, so the choice of bin count is substantively consequential rather than merely computational.

## 2.5 New Diagnostic Tools and Sensitivity Analyses

We also develop a diagnostic that improves interpretability. Beyond determining whether a given set of assumptions implies a narrow or wide identified interval, researchers may also want to understand *why* the data permit that much ambiguity. To address this, `autobounds` can now return the extremal DGPs—that is, two DGPs that are consistent with assumptions and observed data, corresponding to the lowest and highest values of the estimand that cannot be ruled out. Operationally, the analyst calls the solver with `return_dgps=True`, after which the software returns one candidate DGP for the lower bound and one for the upper bound. We demonstrate how this diagnostic is especially informative in the instrumental-variables replication of Coppock and Green (2016) in Section 4.

Sensitivity analysis offers another way to assess the consequences of assumptions. Rather than building separate sensitivity analyses per design, we treat assumption violations as a partial-identification problem: researchers begin with an initial causal model and then relax selected assumptions by allowing them to fail for at most a proportion $\theta$ of units. When $\theta = 0$, the original identifying assumptions are maintained; when $\theta = 1$, the assumption is effectively discarded. For each value of $\theta$, the method derives sharp bounds on the estimand of interest, allowing researchers to determine how much violation is needed before the original causal conclusion can no longer be signed. Theoretical details of this method are studied in

Duarte (2026). An application of this sensitivity framework is also demonstrated in Section 4.

## 2.6 A Worked Example of Automated Partial Identification for Confounding

To demonstrate the algorithm, we consider simulated data in which the analyst is interested in estimating the effect of a college education, $D$, on individual voter turnout, $Y$, but the relationship is confounded. Some of these may be observed confounders—$X$, such as an individual's parental socioeconomic status (SES), say, low ($X = 0$) or high ($X = 1$). Others may be unobserved—$U$, e.g. birth in a urban or rural region or interest in politics. The true data-generating process is represented in the DAG of Figure 1b.[13]

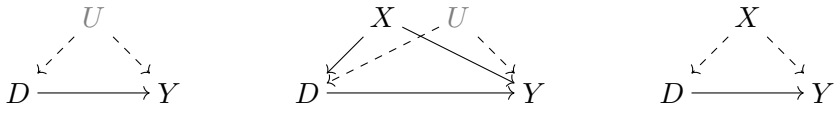
Suppose the analyst is interested in the ATE, $\mathbb{E}[Y(d = 1) - Y(d = 0)]$, equivalent to $\Pr(Y(d = 1) = 1) - \Pr(Y(d = 0) = 1)$ since $Y$ is binary. In our simulation, the ATE is 0.13, meaning that a college education increases the probability of voting by 13 percentage points. Attempting to adjust for observed $X$ by taking

$$\sum_x \Big[ \Pr(Y = 1 \mid D = 1, X = x) - \Pr(Y = 1 \mid D = 0, X = x) \Big] \Pr(X = x)$$

yields a biased estimate of 0.72, or 72 percentage points, due to the presence of additional unmeasured confounding by $U$.

This challenge typically presents researchers with an unpalatable choice. One option is to act as if the problem does not exist and report estimates for the effect of education under a "selection on observables" (SOO) assumption that rules out the existence of unobserved confounders, as in Figure 1c, perhaps supplemented with a post-hoc sensitivity analysis. However, it is arguably implausible that parental SES $X$ is the only factor influencing college education and voting.

[13] See Appendix A for simulation details.




(a) (b) (c)

Figure 1: Data generating processes under (a) completely unmeasured confounding, (b) partially measured confounding, and (c) perfectly measured confounding, sometimes referred to as "selection [into treatment] on observables."

Cautious researchers are ill-served by current methodology, which too often shoehorns projects into existing frameworks, such as the SOO paradigm, for the sake of reporting an estimate—regardless of whether the premises underlying that estimate are substantively defensible. Given current practices, it may seem that the only other options are to shift focus to a different but more feasible question other than the one that originally motivated the data collection—such as the descriptive correlation between education and voting—or to abandon the project entirely.

Our alternative approach using `autobounds` proceeds as follows. To begin, the analyst states her assumed causal model of the world, the DAG of Figure 1b, using the Python programming language (an implementation in the R language is in early development).

```
# load package
from autobounds import *
# define each arrow in fig 1b, flagging variable U as unobserved
confounding_model = DAG("D -> Y, X -> D, X -> Y, U -> D, U -> Y", unob = "U")
confounding_problem = causalProblem(confounding_model)
```

Running `causalProblem` creates an object that will eventually hold all information available to the analyst. Upon creation, this object holds the causal graph over treatment $D$, outcome $Y$, observed confounder $X$, and unobserved confounder $U$; unless otherwise specified, observed variables are treated as binary and unobserved variables are allowed to be continuous or high-dimensional. This allows `autobounds` to implicitly define the principal strata described in Section 2.1.

The researcher then defines her causal question—the estimand—in this case, the ATE.

```
# estimand is ATE of independent variable D on dependent variable Y
confounding_problem.set_ate(ind="D", dep="Y")
```

The `set_ate` method then defines the quantity of interest, in this case the ATE of the independent variable $D$, college education, on the dependent variable $Y$, voting. Using additional arguments, this shorthand method can also be used to specify estimands that are conditional ATEs; a more general alternative, `set_estimand`, offers a flexible syntax for defining other quantities of interest.[14]

```
import pandas
# load observed data with D, X and Y columns, one row per unit
confounding_data = pandas.read_csv("confounding_data.csv")
# inference argument preps autobounds to compute confidence intervals
confounding_problem.read_data(raw=confounding_data, inference=True)
```

Next, the analyst loads the raw data from a .csv file—in which each row represents one unit and columns are given for $X$, $D$, and $Y$—using the standard Python package for data manipulation, `pandas`. The `read_data` method automatically tabulates observable strata in the data, e.g. the proportion of individuals who are college-educated voters with high parental SES ($X = 1$, $D = 1$, and $Y = 1$), the proportion who are college-educated voters with low parental SES ($X = 0$, $D = 1$, and $Y = 1$), and so on. This provides information the algorithm can use to rule out possible values of the estimand.[15] It then translates this observed data distribution into implied constraints on the sizes of each principal stratum. See Appendix A, Equation (11) for a formal statement of this problem.

Finally, the following code produces sharp bounds on the ATE—the narrowest bounds possible absent further data or assumptions.[16]

```
# compute bounds, verify sharpness, and conduct statistical inference
confounding_problem.solve(ci=True, progress = True, nsamples = 2000)
```

[14]The `set_ate` method is given for ease of use; for this common use case, it provides a shortcut that is equivalent to `with respect_to(confounding_problem):` followed by either `set_estimand(E("Y(d=1)") - E("Y(d=0)"))` or `set_estimand(p("Y(d=1) = 1") - p("Y(d=0) = 1"))`. More generally, by writing `E` and `p` functions of potential outcomes, users can formulate arbitrarily complex quantities for use in estimands and assumptions. For additional documentation and worked examples, see appendices.

[15]An alternative is to provide a data frame with one row per combination of possible values for $X$, $D$, and $Y$ columns, with a third column `prob` containing the frequencies of each combination.

[16]The bounds are then computed with the `solve` method, which utilizes the SCIP Optimization Suite (Bolusani et al., 2024) via the `PySCIPOpt` interface (Maher et al., 2016).

The analyst finds that `autobounds` outputs an estimated interval of $[-0.16, 0.84]$. Estimated 95% confidence intervals on the bounds, (i.e. statistical uncertainty due to sampling error) are also reported as $[-0.18, 0.86]$. Confidence bounds are computed using the method described in Appendix C.2.1. These bounds cover the true ATE, but they are quite uninformative: the worst- and best-case scenarios are far apart, so the range of possible answers to the causal question is wide, and they cross zero, so the sign of the effect is unidentified.[17]

In this simple example, $X$ is low-dimensional—specifically, binary—which makes it straightforward to stratify by $X$ directly. However, when $X$ contains many levels or is high-dimensional, stratification becomes computationally or practically infeasible. In such cases, an alternative strategy is available: one can assume that observed covariates affect all variables in the model (i.e., they influence treatment, outcome, and their relationship), rather than assuming they affect none. This covariate adjustment approach allows for efficient computation even with many-valued or high-dimensional covariates. For technical details on this strategy, see Appendix C.4. To illustrate, suppose the analyst ran `autobounds` using this covariate-adjustment approach on the same confounding problem, without explicitly including $X$ as a node in the DAG. The code would look as follows:

```
from autobounds import respect_to
confounding_model_no_x = DAG('D -> Y, U -> D, U -> Y', unob='U')
confounding_problem_with_x = causalProblem(confounding_model_no_x)
with respect_to(confounding_problem_with_x):
    read_data(raw=confounding_data, covariates=["X"])
    set_ate(ind="D", dep="Y")
    solve(ci=True, nsamples = 2000, progress=True, workers=20)
```

This approach leverages observed covariates to help narrow bounds while maintaining computational tractability even with high-dimensional covariates. In settings where the covariate adjustment reduces to exact stratification on $X$, the resulting bounds coincide, up to sampling error due to subsampling, with those obtained by explicitly including $X$ in the DAG

[17] In fact, without further assumptions, it has been shown that the bounds on the ATE, given confounding between a binary treatment and a binary outcome, will always contain zero because they are always of width one (Robins, 1989; Manski, 1990). Perhaps surprisingly, this remains true when adjusting for observed confounders as well, absent further assumptions.

and averaging the stratum-specific sharp bounds. In the more general continuous- or high-dimensional-covariate case, however, the fitted score-binning procedure should be understood as an approximation to that exact stratified target. In this example, the bounds remain $[-0.16, 0.84]$ (95% CI $[-0.18, 0.86]$), which are still uninformative.

Not all may be lost, however. These bounds are wide because they are nonparametric: among other implications, this means that they allow for any possible way that observed confounders $X$ and unobserved confounders $U$ might interact with each other and with treatment $D$. However, expert knowledge might permit the analyst to add an assumption which indirectly limits the impact of this issue. The analyst may make the following observation: treatment is indeed confounded, but people who come from families with higher parental SES have a higher counterfactual propensity to vote, regardless of their education. In addition to the potential impact of financial constraints among their children on voting rates, high-SES parents are also argued to foster politically rich home environments, further increasing political participation later in life (Verba et al., 2005; Schafer et al., 2022).

Therefore, we might assume those with means would, on average, vote more often than those without, given any counterfactual level of education. Formally, the analyst assumes

$$\mathbb{E}[Y(d) \mid X = 1] \geq \mathbb{E}[Y(d) \mid X = 0] \quad \text{for each } d.$$

This notion, which was first proposed by Manski and Pepper (2000) in the instrumental-variable context, is known as a monotone response assumption across subgroups.

In `autobounds`, this assumption can be implemented as follows:

```
from autobounds import respect_to
# shorthand to facilitate repeated statements about this problem
with respect_to(confounding_problem):
    # define key quantities
    turnout_if_college_in_highSES   = E("Y(D=1)", cond="X=1")
    turnout_if_college_in_lowSES    = E("Y(D=1)", cond="X=0")
    turnout_if_nocollege_in_highSES = E("Y(D=0)", cond="X=1")
    turnout_if_nocollege_in_lowSES  = E("Y(D=0)", cond="X=0")
    # state the assumption of monotone response across subgroups
    add_assumption(
        turnout_if_college_in_highSES, ">=", turnout_if_college_in_lowSES
```

```
)
add_assumption(
    turnout_if_nocollege_in_highSES, ">=", turnout_if_nocollege_in_lowSES
)
# compute bounds, verify sharpness, and conduct statistical inference
solve(ci=True, nsamples=1000)
```

Under this assumption, the bounds become $[0.10, 0.84]$ (95% CI $[0.04, 0.86]$). Thus, the analyst is able to estimate informative bounds and sign the effect, while keeping her research question the same, and avoiding untenable assumptions. The same simulated confounding setup can also be analyzed in the continuous-outcome mode described in Section 2.4.

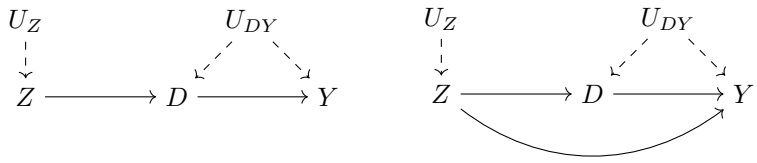
We emphasize that in practice, such restrictions should rest on subject knowledge or prior evidence, not be chosen arbitrarily. An ideal workflow would state these assumptions once the causal estimand is defined, perhaps even in a pre-registration step. We provide them merely to illustrate the algorithm's flexibility in a simple example. Next, we show how autobounds relaxes assumptions in published analyses with more complex designs, often while still returning informative results.

# 3 Testing Empirical Implications of Theorized Assumptions

In this section, we investigate the empirical consequences of violations of the exclusion restriction in an instrumental variables (IV) framework, an assumption often described as "not testable" (Imbens and Angrist, 1994, p. 468). IV studies assume that treatment $D$ and outcome $Y$ share a common unobserved confounder, $U_{DY}$, but that an instrumental variable, $Z$, "encourages" treatment to occur as-if randomly, allowing for the identification of a local average treatment effect of $D$ on $Y$ (Angrist et al., 1996). This DGP is represented in the left panel of Figure 2.

As shown in Angrist et al. (1996), the traditional instrumental variables strategy will recover the local average treatment effect among "compliers" (LATE)—units which respond to

Figure 2: **IV DAGs.** Panel (a) depicts the simple IV setting of Angrist et al. (1996). In this setting, if monotonicity of Z to D is assumed (no defiers), the local ATE on the outcome $Y$, among those that "comply" with encouragement, is point identified. Panel (b) represents the same DAG with an additional arrow from $Z$ to $Y$, which indicates a violation of exclusion restriction.




(a) (b)

the encouragement provided by the instrument as instructed—if several assumptions hold. These conditions include (i) ignorability of $Z$, satisfied if the instrument is as-if randomly assigned; (ii) a non-null effect of $Z$ on $D$, also known as "relevance"; (iii) an exclusion restriction, or the absence of a direct effect of $Z$ on $Y$; and (iv) monotonicity, or the absence of "defiers" that behave inversely to instructions.[18] As Balke and Pearl (1997) shows, even when monotonicity is not assumed, it is possible to calculate sharp bounds for the ATE using a linear-programming approach. `autobounds` generalizes this approach allowing for the calculation of sharp bounds not only for the ATE, but also for nonlinear quantities such as the LATE,[19] and indeed for essentially any estimand. Moreover, this `autobounds`-based estimator will produce valid results both with and without monotonicity assumptions. In cases where stronger assumptions make the estimand point identifiable (e.g., the LATE under monotonicity, Angrist et al., 1996), the interval of possible answers from `autobounds` collapses so that the best-case upper bound is exactly equal to the worst-case lower bound.

Besides calculating bounds for estimands, `autobounds` also tests all empirical implications of the theoretical model. In the IV case, these observable implications are known as the

[18]This result also assumes a stable unit treatment value assumption (SUTVA), which we employ throughout this paper.

[19]The ATE is a linear function of the principal strata sizes, as it is equal to $\mathbb{E}[Y(d=1)-Y(d=0)] = \Pr(\text{Y-helped}) - \Pr(\text{Y-hurt})$, where the group with outcomes "helped" by treatment is the group where $Y(d=1)=1$ and $Y(d=0)=0$; conversely, the group with outcomes "hurt" by treatment $Y(d=1)=0$ and $Y(d=0)=1$. (See Footnote 6 for additional discussion in a slightly different context.) In contrast, the LATE is a nonlinear function of the principal strata sizes because conditioning creates a fraction that cannot be eliminated: $\mathbb{E}[Y(d=1)-Y(d=0)|\text{D-complier}] = \big[\Pr(\text{Y-helped, D-complier}) - \Pr(\text{Y-hurt, D-complier})\big] / \Pr(\text{D-complier})$.

*instrumental inequalities* (Pearl, 1995b; Bonet, 2001). For a valid instrument $Z$, discrete treatment $D$, and discrete outcome $Y$, the exclusion restriction implies that

$$\max_{d} \sum_{y} \left[ \max_{z} \Pr(Y = y, D = d | Z = z) \right] \leq 1. \tag{1}$$

must hold. If data fails to satisfy it, and other assumptions are known to hold, then analysts may conclude there is a violation of the assumptions—e.g. a direct $Z \to Y$ effect, confounding between $Z$ and some other variable.

## 3.1 Replication and Extension of Kocher et al. (2011)

We now demonstrate how `autobounds` can alert researchers to *faulty* assumptions—assumptions which are logically inconsistent with observed data—by testing whether their empirical implications are violated. This example also incorporates our new method of covariate adjustment within the same framework.

We replicate Kocher et al. (2011), which seeks to estimate the effect of aerial bombing during the Vietnam War by the U.S.-backed Republic of Vietnam (RVN) of civilian hamlets on local control of hamlets by the Viet Cong. The paper concludes that bombing civilian targets increases the probability of insurgent control; in other words, the tactic backfires on the aggressor. The main identification challenge, which the paper addresses several ways, is that bombing is not random, but determined by military strategy. One approach used is an IV design in which prior insurgent control—a lagged outcome—is regarded as an encouragement for the treatment of aerial bombing.

To preview our findings, `autobounds` reveals the IV assumptions are *falsified* here. In other words, their empirical implications are not satisfied, suggesting that the variables used are not valid instruments.[20] We emphasize that we focus on one particular model specification from

[20] Consistent with this finding, the two-stage least squares results in Table 5 of Kocher et al. (2011) suggest that the binary treatment would have a 9-point effect on the outcome—an impossibility when the outcome is measured on a 5-point scale.

Kocher et al. (2011). So the paper's overall conclusions may still hold given other evidence in Kocher et al. (2011). Further, to adapt this problem to the `autobounds` framework, we make a number of specification choices that are distinct from (though conceptually consistent with) the analysis in Kocher et al. (2011).[21]

In one set of analyses, the study employs an IV design in which the instrument, $Z$, is insurgent control of a hamlet as measured in July 1969; the treatment, $D$, is an indicator of whether any bombs were dropped within a two-kilometer radius of the hamlet; and the outcome, $Y$, is control of the hamlet in December 1969. It further uses a number of control variables, $\mathbf{X}$, such as terrain roughness, population, and control of the hamlet at an intermediate point in time (September 1969). The validity of this strategy has a number of requirements, including that past control of a hamlet does not have a direct effect on future control (there is no $Z \to Y$ arrow). This is a questionable assumption, because control in period $t-1$ almost certainly influences control in $t$, e.g. by reinforcement of fortifications.[22] However, even if this were true, the design also requires that there are no unobserved common causes, such as the sympathies of hamlet residents, that jointly influence past and future control (i.e. there are no $Z \leftarrow U \to Y$ confounders). Kocher et al. (2011) argue that this condition is met: "there are no unobserved hamlet-specific variables that affected insurgent control in July, August, and December 1969, but *not* in September of that year as well," (p. 212, emphasis in original).

We first use `autobounds` to test these assumptions' observable implications—i.e., the instrumental inequalities given in Equation 1. We supply the assumed causal diagram—a *conditional* IV graph in which the 5-valued instrument $Z$ (prior insurgent control) influences the binary treatment $D$ (aerial bombing) which in turn influences $Y$ (future insurgent control), with background variables $\mathbf{X}$ that influence all of the above. Following the standard `autobounds` workflow, we then formally state what appears to be the implicit estimand of

[21] See Appendix D for details on our operationalization.

[22] One possible reason that the instrumental variable design might nonetheless remain valid is if hamlet control follows a Markov process: i.e., if control in $t-2$ (July 1969, $Z$) influences control in $t$ (December 1969, $Y$) only through control in $t-1$ (September 1969, in $X$). If this were true, then adjusting for hamlet control in September 1969 would be sufficient to block any direct effect.

Kocher et al. (2011), the LATE. We do not impose the assumption of monotonicity, so if `autobounds` detects a violation of assumptions, then it is either exclusion or randomization that must be violated.[23] Finally, we supply the data and start the computation.

Based on this information, `autobounds` reports that the IV assumptions are falsified: their observable implications are violated, i.e., the data distribution is inconsistent with the hypothesized IV data-generating process (Yang et al., 2014). The intuition is straightforward. Conditional on covariates, some strata exhibit substantial variation in the instrument $Z$ but almost no variation in treatment $D$, so under a valid IV design the distribution of outcomes should vary little with $Z$. Instead, outcomes differ sharply across values of the instrument, indicating either $Z$-$Y$ confounding or a direct effect of $Z$ on $Y$ that does not operate through $D$. A broader matched analysis yields the same conclusion; see Appendix D for details and Appendix Figure 9 for the `autobounds` implementation.

# 4 Probing Sensitivity to Assumption Violations

`autobounds` can also quantify the consequences if key assumptions, like monotonicity and the exclusion restriction, are violated to varying extents. In particular, we can use `autobounds` to recover sharp bounds on causal estimands after allowing for a given share of units to violate these assumptions.

## 4.1 Replication and Extension of Coppock and Green (2016)

To demonstrate these capabilities, we examine a well-known get-out-the-vote (GOTV) experiment, Gerber et al. (2008), which originally tested whether various forms of social pressure caused people to turn out in a 2006 primary election in Michigan. Specifically, voters were randomly informed that their turnout activity would be monitored by researchers and, in

[23] The absence of a monotonicity assumption means that the LATE will not be point-identified. In practice, the assumptions will be falsified regardless of what specific estimand is chosen here.

one treatment arm, that their neighbors would be informed as to whether they voted. Subsequently, Coppock and Green (2016) extended this study using the IV framework to test whether voting in one election affected the chances of voting in subsequent elections. In the extended setup, the initial social pressure intervention is treated as a binary instrument (encouragement, $Z$) to vote in the contemporaneous 2006 primary election (which is conceptualized as the treatment, $D$), but the outcome of interest is whether people vote in a subsequent general election later that year ($Y$). Substantively, this re-analysis sought to test whether voting is "habit forming," e.g. whether the experience of voting in the primary election at $t$ increases the probability of voting in subsequent elections from $t+1$ onward. Figure 10 in Appendix E.1 shows the core `autobounds` code for this example, including recovery of the bound-attaining DGPs.

Importantly, the assumptions of the IV design imply the only mechanism through which social pressure would affect subsequent turnout in election $t+1$ is by changing turnout in $t$, an action that then becomes habit forming. However, as Davenport et al. (2010) notes, the social-pressure encouragement could also cause people to internalize the civic norm of voting (Bandura and Walters, 1977)—thus directly influencing subsequent turnout in violation of the exclusion restriction—invalidating the IV design as a means to identify the local average treatment effect among compliers. This concern is heightened by the memorable nature of the "neighbors" social-pressure encouragement: delivered immediately before to the primary in August, it could potentially remain salient enough to directly affect some voting decisions in the November general election, held just three months later.

Our extension of `autobounds` offers a method for gauging the sensitivity of results to violations of this assumption. We start by describing a DAG of the form shown in Figure 2(b), allowing for a direct effect between the instrument (here, whether individuals received the "neighbors" treatment) and the outcome. Then, we add code which stipulates the hypothesized maximum share of units for whom the exclusion restriction may be violated. A share

of zero re-imposes the exclusion restriction; larger shares relax it. Finally, we recompute bounds on the estimand when allowing for some violation of the assumption.[24] After setting up the causal diagram, loading data, and stating the monotonicity (no-defiers) assumption, the following lines of Appendix Figure 10 show how to relax the potentially problematic exclusion-restriction assumption:

```
with respect_to(gotv_problem):
    # define group for which exclusion restriction is violated
    p_excl_restr_violation = p(is_active("Z -> Y"))
    # assume that the size of this group is limited
    add_assumption(p_excl_restr_violation, "<=", 0.01)
```

Varying this maximum share between 0 and 1 produces a full sensitivity curve for any degree of violation. Figure 3 shows that the LATE is point identified when the exclusion restriction is imposed exactly, but becomes unsigned once slightly more than 1% of units are allowed to violate it. In this sense, the design appears highly sensitive to exclusion-restriction violations despite its strong first stage.

The `autobounds` algorithm also allows us to easily explore alternative estimands, such as the ATE, and reason about its bounds. Under the baseline IV assumptions with monotonicity, the sharp ATE bounds in this application are $[-0.520, 0.397]$, so the sign of the effect is not identified. Let us understand why this is the case. In the IV setting studied here, any causal question could be answered for a given DGP by knowing how often $D(0) = d$, $D(1) = d'$, $Y(0) = y$, and $Y(1) = y'$ for all $d, d', y, y'$, (briefly, this is because the unobserved $U_{DY}$ can be equivalently represented by these latent response types, see Appendix A for more detail on this point). Thus, by reasoning about how these strata are constrained in the best- and worst-case DGPs (corresponding to the bounds of the partially identified region), the analyst may interrogate what kinds of possible worlds lead to best- and worst-case scenarios for their quantity of interest—here, the unconditional ATE. Because $Y$ is binary, the ATE is simply the difference between the share of voters who would vote in election $t+1$

[24] Specifically, we focus on the effect of the "neighbors" treatment on Nov. 2006 general election turnout reported in column 2 of Table 1 in Coppock and Green (2016). We note that the original paper contained evidence from a number of other IV models and regression discontinuity analyses to study this question.

if they voted in election $t$ (i.e. if they were treated) and the share who would vote in election $t+1$ if they did not vote in election $t$: the former group consists only of those who are "helped" (made more likely to vote in election $t+1$ by voting in election $t$) or always vote, while the latter consists only of always-voters or those "hurt" (made less likely to vote in election $t+1$ by voting in election $t$), so the always-voters cancel out and the ATE reduces to Pr(helped) − Pr(hurt). This contrast can be decomposed by unit response to encouragement, $\sum_{D\text{-type}\in\{\text{complier, defier, always-taker, never-taker}\}} \{\Pr(D\text{-type, helped}) - \Pr(D\text{-type, hurt})\}$, because (by random assignment) being helped or hurt by treatment does not constrain how a unit responds to encouragement.

We are now interested in how the strata comprising the ATE appear in the best- and worst-case DGPs. The `autobounds` software can return the sizes of principal strata in the DGPs associated with the best- and worst-case scenarios consistent with the data and assumptions.[25] There are many possible compositions of strata that could produce the bounds (subject to observed data constraints). `autobounds` by default tells us one of them, but it is possible to access more. We choose an exemplar here for exposition, as it permits us to demonstrate how to think more precisely about the origin of the bounds.[26] It is important to highlight that by studying the composition of the strata which attain the extremal bounds of the estimand we reveal sufficient but not necessary conditions for obtaining those bounds. An applied researcher may draw varied but all informative intuitions about the identification of her target quantity as a guide for further study, a discussion of which we will reach shortly.

Table 1 reports the results of this exercise. The results show that these DGPs are identical in terms of "compliers" with $D(1) > D(0)$ and, among this group, identical in the proportion "helped" or "hurt" by treatment—i.e., the subset of the compliers for whom voting in election

[25] Practically, this is done by evaluating the maximizers and minimizers of the bounding program. As the parameters of the program are the principal strata, it is possible to study the prevalence of each kind of voter-type in the DGPs which produce the bounds of the partially identified region. Please see the accompanying jupyter notebook on the IV analysis for a walkthrough.

[26] For instance, monotonicity is not required for the intuition discussed here to go through. We observed a solution consistent with an absence of defiers, and chose this solution for simplicity of exposition (which is equivalent to studying a problem which imposes monotonicity).

$t$ makes it more or less likely to vote in election $t+1$. This is because (assuming no defiers) (i) the known first stage fixes Pr(complier) $= 0.0678 + 0.0129 + 0.0023 = 0.083$, so that this quantity must be identical across the best- and worst-case DGPs; and (ii) the known reduced form, $\mathbb{E}[Y(z=1) - Y(z=0)] = \mathbb{E}[Y(D(1)) - Y(D(0))]$, fixes Pr(complier, helped) − Pr(complier, hurt) $= 0.0129 - 0.023 = 0.011$ as well. Thus, the best- and worst-case DGPs differ only among the remaining "always-" and "never-takers" with $D(0) = D(1) = 0$ or 1, respectively.

Among these encouragement-unaffected individuals, where the IV design is entirely uninformative, the best-case DGP represents the possibility that—unobservably to the analyst—57.9% would have been hurt (voting would cause them to not vote in the subsequent election) by a counterfactual intervention that forced them to be treated (e.g. to vote in the initial election).[27] Conversely, the worst-case DGP represents the alternative possibility that 42.1% of these encouragement-unaffected individuals would have been helped (voting would cause them to vote in the subsequent election) by such an intervention. Note that within each of the best- and worst-case DGPs, it cannot be that all encouragement-unaffected individuals are assigned to the helped or hurt categories, due to other constraints imposed by the observed data. However, between extremal DGPs there is an intuitive complementarity: subject to observed data constraints, the worst-case DGP minimizes the ATE by assigning 57.9% of the encouragement-unaffected mass to voters who would be hurt by treatment, while the best-case DGP maximizes the ATE by assigning the remaining 42.1% of that same mass to voters who would be helped by treatment. This imbalance in magnitude and direction also explains why the bounds permit more negative ATE values than positive ones.

These extremal DGPs yield substantive insights that could motivate new research. For example, at the upper bound of the ATE, we see that more than half of all voters would: (i) never be responsive to GOTV mailers and (ii) be dissuaded from voting in election $t+1$ if they voted in election $t$. This implies a massive block of individuals who are impervious

[27] (0.5192 + 0.0115) / (0.5192+0.2998+0.0865+0.0115) = 0.579.

Table 1: **Extremal DGPs producing the ATE bounds in the GOTV IV study.** $D$-type is the unit response to encouragement (i.e. complier, defier, always-taker, never-taker). $Y$-type is unit response to treatment (i.e. helped, hurt, always-vote, never-vote). Columns best- and worst-case prevalence report the mass of the joint principal strata that attain the best- (upper) and worst-case (lower) bounds for the ATE in the re-analysis of Coppock and Green (2016), respectively.

| $D$-type | $Y$-type | Worst-case prevalence | Best-case prevalence |
|---|---|---|---|
| never-taker | harmed by treatment | 0.5192 | 0.0000 |
| always-taker | always-voter | 0.2998 | 0.0000 |
| never-taker | never-voter | 0.0865 | 0.0000 |
| complier | always-voter | 0.0678 | 0.0678 |
| complier | helped by treatment | 0.0129 | 0.0129 |
| always-taker | harmed by treatment | 0.0115 | 0.0000 |
| complier | harmed by treatment | 0.0023 | 0.0023 |
| never-taker | always-voter | 0.0000 | 0.5192 |
| always-taker | helped by treatment | 0.0000 | 0.2998 |
| never-taker | helped by treatment | 0.0000 | 0.0865 |
| always-taker | never-voter | 0.0000 | 0.0115 |

to traditional campaign tactics and for whom the experience of voting would only depress levels of participation. Might other interventions besides these mailers be able to target these individuals? If so, the returns for participation could be immense. In the worst case, this same share of voters belongs to a stratum that would never be encouraged by a mailer, but would always vote nonetheless. Understanding this population could be invaluable to campaigns seeking to spend resources efficiently, avoiding households where persuasion is impossible. Separately, analysts with substantive knowledge in this domain might view these proportions as implausibly large, motivating new formal assumptions that can further narrow bounds on the estimand.

Figure 4 briefly illustrates the corresponding ATE point in synthetic reweightings of the data: when compliance is low, the ATE remains highly sensitive to exclusion-restriction violations, whereas higher compliance narrows the ATE bounds and brings them closer to the LATE.

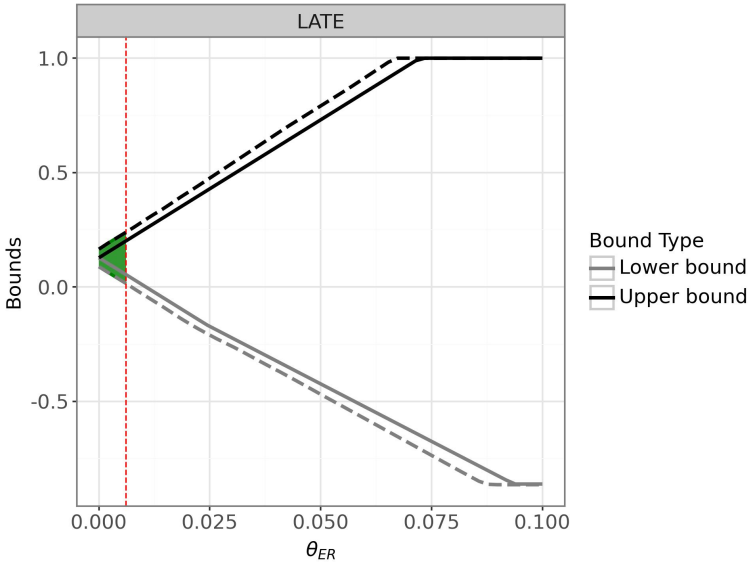



Figure 3: **Sharp Bounds on the LATE Under Increasing Violations of the Exclusion Restriction.** The plot below shows sharp bounds on the LATE of an initial GOTV intervention on turnout in a future election (solid lines are estimated bounds, dashed lines are 95% confidence intervals on those bounds). The $x$ axis displays $\theta_{ER}$, the hypothetical share of respondents violating the exclusion restriction.The plot shows that if 0% of units are assumed to exhibit a direct effect between instrument and outcome (i.e. if the exclusion restriction is assumed to hold perfectly), the LATE is point identified. However, assuming that more than 1% of units exhibit a direct effect—in violation of the exclusion restriction—the bounds on the LATE can no longer be signed.

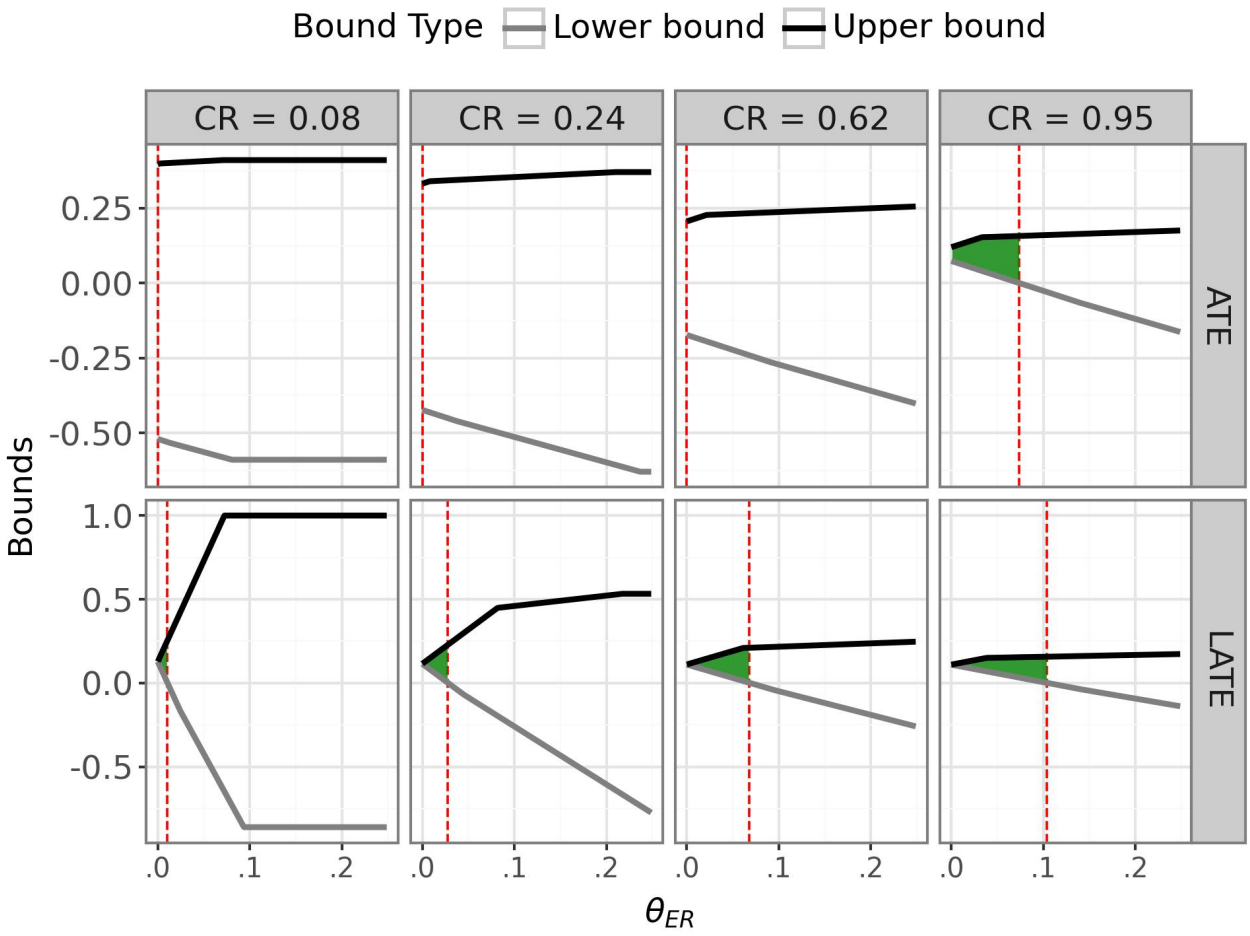



Figure 4: **Sharp Bounds on the ATE/LATE under Increasing Compliance Rates.** The green shaded areas highlight regions where the sign of the effect is identified. The compliance rate in the original data is 0.08, and the remaining panels use reweighted versions of the observed data with higher compliance rates.

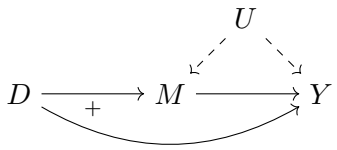
# 5 Weighing Relative Importance of Assumptions

What does it mean to say one assumption is "stronger" than another? One answer is that stronger assumptions are ones that, when violated, have relatively larger empirical implications on what we can learn from a study than weaker assumptions. Historically, determining the consequences of violating multiple assumptions required lengthy derivations. Fortunate, the modularity of `autobounds` allows such comparisons to be conducted in a straightforward manner. To illustrate, we revisit Knox et al. (2020) and Fryer (2019), which tested for potential racial discrimination in police use of force. Such analyses must address selection bias, unobserved confounding, and collider bias, requiring several assumptions to even partially identify effects. Below, we outline these assumptions and we leverage `autobounds` to re-evaluate their importance.

## 5.1 Replication and Extension of Knox et al. (2020)

Selection bias is sometimes inherent to the data sources used by social scientists. A hard problem arises when treatment status determines whether units are observed, and a large literature studies the resulting post-treatment conditioning bias (Rosenbaum, 1984; Acharya et al., 2016; Nyhan et al., 2017; Blackwell, 2013). In one example, Knox et al. (2020) examines the problem of estimating racial bias in the use of force by police using only data on encounters in which police choose to detain individuals. As the paper shows, because the race of civilians involved in police encounters (the treatment, $D$) very likely affects whether individuals are detained in the first place (an indicator for whether a person is stopped by police, $M$), then comparing the rates of force used against white and nonwhite civilians among the subset of encounters that involve a detainment leads to underestimates of racial bias in the use of force, absent implausible assumptions. While the original paper examined the use of various levels of discriminatory force against both Black and Hispanic civilians (compared to white civilians), for simplicity, this reanalysis focuses on a binary indicator for the use of any force at all ($Y$)

Figure 5: **Racial Discrimination in Police Use of Force.** $D$ represents the minority status of an encountered civilian. $M$ is the mediating decision of whether an officer chooses to detain the civilian. $Y$ is use of force.




and on the Black-white comparison only.

The source of the confounding that results from this form of sample selection can be seen in Figure 5. As the DAG shows, even if the analyst is able to adjust for all common causes of the treatment and outcome ($D$ and $Y$), and of the treatment and mediator ($D$ and $M$)—equivalent to rendering encounters comparable before the officer makes the decision to initiate a stop—conditioning on the mediator induces "collider bias" (Pearl, 2009), allowing common causes of stopping ($M$) and the use of force ($Y$) to confound comparisons. Theoretically, these unobserved confounders, represented collectively by $U$, could be factors never recorded in police administrative data such as the officer's mood at the time of the encounter.

To address this obstacle, Knox et al. (2020) analytically derives nonparametric sharp bounds on several causal estimands corresponding to racial bias. In general, these estimands consider the counterfactual substitution of a different individual of differing racial identity into an otherwise similar police-civilian encounter, and they compare the average counterfactual rates of force between two encounters involving two racial/ethnic groups of civilians. To sharply bound these estimands given available administrative data, Knox et al. (2020) appeals to four assumptions. We focus here on two in particular.[28]

The first assumption we reexamine is "mediator monotonicity," or the assumed absence of

[28]In addition to the assumptions that we relax and abandon in this reanalysis, Knox et al. (2020) imposes assumptions about the absence of unreported force and the ignorability of civilian race; see original paper for details. In addition, the bounding approach in Knox et al. (2020) allows analysts to specify the severity of bias in the initial decision to stop civilians to obtain sharp bounds for that scenario. The paper shows the severity of stopping bias can be lower bounded using a standard outcome test (Knowles et al., 2001), and estimates that at least 32% of stops of Black civilians would not have occurred had similarly situated white civilians been encountered in the New York City case. In keeping with this result, we specify the parameter $\rho$ indicating racial bias in stopping at 0.32 in the replication below.

anti-white police stops:

**Assumption 5.1.1** (Mediator Monotonicity)**.** $M(1) \geq M(0)$.

This assumption states that there are no encounters in which a stop would occur if a civilian was white, $M(d = 0) = 1$, but would not occur, counterfactually, if a civilian was nonwhite, $M(d = 1) = 0$. This could be violated if, for example, white civilians were more likely to be stopped, all else equal, when walking in majority Black neighborhoods, perhaps because they looked out of place.

Next, Knox et al. (2020) make an assumption about the average levels of force that would be applied, counterfactually, in two different types of encounters. The first are "always stops," or scenarios where officers would stop any civilian regardless of race, $M(d = 0) = M(d = 1) = 1$, e.g. armed robberies. The second are "racial stops," or scenarios in which officers would exercise discretion by stopping a minority civilian, $M(d = 1) = 1$ but would not stop a white civilian, $M(d = 0) = 1$—a pattern that might plausibly occur in, e.g., jaywalking incidents.

**Assumption 5.1.2** (Relative Non-severity of Racial Stops)**.**

$\mathbb{E}[Y(d, m) \mid M(d = 1) = 1, M(d = 0) = 1] \geq \mathbb{E}[Y(d, m) \mid M(d = 1) = 1, M(d = 0) = 0]$

This assumption holds if, for any race of civilians counterfactually inserted into the encounter, $d$, the average rate of force used during police encounters is larger in "always stop" encounters than in "racial stop" encounters where nonwhite civilians would be discriminatorily detained. The logic behind this assumption is that the former class of encounters are theorized to be serious incidents in progress, where officers have a duty to intervene by detaining the civilian and, it is assumed, a higher likelihood of using force; in contrast, the second class of encounters are discretionary scenarios in which officers have a choice about whether to intervene and where racial bias may therefore be more likely to influence the decision.

While Knox et al. (2020) argues that these assumptions are plausible in the empirical setting they examine (New York City in the 2000s, during which time the controversial "Stop,

Question and Frisk" tactic was prevalent (Mummolo, 2018)), others may question their validity. It is therefore worth investigating how consequential these assumptions are. Unlike the GOTV experiment of Section 4.1, it is a-priori unclear which assumptions are pivotal. There, voter behaviors which could violate monotonicity were uncontroversially thought to be unlikely to materialize; it was more plausible that the behavioral mechanism which may violate exclusion restriction—memory—could render conclusions fragile, as `autobounds` uncovered. However, in Knox et al. (2020) there exists an intersection of possible mechanisms which could produce the outcome.

Consider the ATE among the stopped, $\text{ATE}_{M=1}$. This quantity averages the effect over the always-stopped (e.g. bank robbers) and the racially stopped (e.g. jaywalkers), because anti-white stops (akin to defiers) are ruled out by mediator monotonicity. Assumption 5.1.2 compares the outcome across these two groups. Signing the effect is therefore a consequence of balancing two items: (i) the strength of the effect within the two groups, and (ii) the prevalence of each group. Mediator monotonicity rules out anti-white stops, which would attenuate the effect, but says nothing about the composition of the remaining units. On inspection, it would appear as if the latter assumption holds if officer taste for force is strongly motivated by the features of an encounter which removes discretion as opposed to one which does not. However, as neither of these assumptions are informative about the strength of that first stage, it is difficult to develop intuition for which assumption is more potent, and many other mechanisms could be consistent with a disparity between the two terms in Assumption 5.1.2, as noted by VanderWeele (2008).

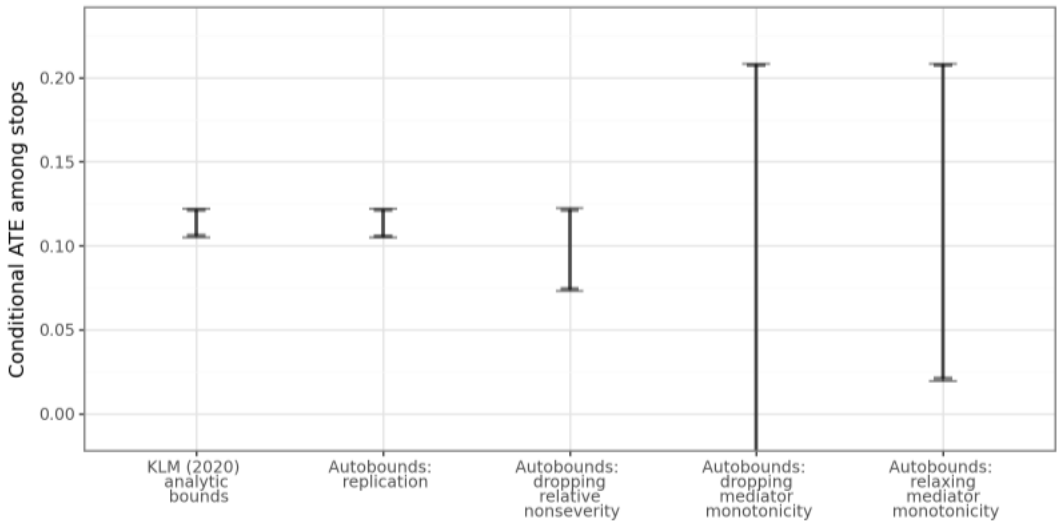
Rather than parsing the lengthy decomposition of the estimand by hand, `autobounds` lets us compare the empirical importance of these assumptions directly; see Figure 11 for the relevant code.[29] The far left result in Figure 6 shows the analytic baseline for the average

[29]For all numerical analyses, we assume that the probability of being stopped, $P(M = 1)$, is at least 0.02 to ensure numerical stability, as the denominator in the computation of the estimand $ATE_{M=1}$ must be greater than zero; when the denominator approaches zero too closely, numerical tolerance issues can lead to unreliable results. Additionally, the confidence intervals shown here were computed without clustering, which makes them substantially narrower than those reported in the original paper. Functionality to permit clustering in

treatment effect among the detained, $\text{ATE}_{M=1}$, among encounters involving Black and white civilians, with bounds of $[0.106, 0.121]$ (95% CI $[0.105, 0.122]$).[30] Given the same information, `autobounds` exactly recovers these bounds. Dropping the relative-severity assumption while retaining mediator monotonicity widens them only slightly, to $[0.074, 0.121]$ (95% CI $[0.073, 0.122]$).

By contrast, dropping mediator monotonicity yields $[-0.237, 0.207]$ (95% CI $[-0.239, 0.208]$), so the sign of the effect is no longer identified. In this application, mediator monotonicity is therefore more consequential than Assumption 5.1.2. The final estimate shows that the effect becomes positive again if this assumption is relaxed to allow no more than 5% of stops of white civilians to be discriminatory, producing $[0.020, 0.207]$ (95% CI $[0.019, 0.208]$).

Figure 6: **Sharp Bounds on Racial Bias in the Use of Force by Police Under Various Assumptions**. The figure displays sharp bounds on racial bias in the use of force ($\text{ATE}_{M=1}$) by police in New York City using data from Knox et al. (2020). `autobounds` replicates the analytic result computed in the original paper. Dropping an assumption about the relative severity of force across principal strata of encounters still allows anaylsts to bound racial bias as positive. Dropping the assumption of no anti-white bias in stopping leads to uninformative bounds, but relaxing this assumption to allow for no more than 5% of anti-white discriminatory stops results in positive bounds on the causal estimand.




`autobounds` is under development and is expected to be available soon.

[30] These results correspond to the Black-to-white comparison in the corrected version of Knox et al. (2020) under the "baseline specification" without covariate adjustment (Knox et al., 2026).

# 6 Discussion and Conclusion

Recent decades have seen the development of a raft of research designs for applied causal inference. These strategies have revolutionized social science by making explicit and precise the estimands and identifying assumptions under which causal relationships can be inferred. But these assumptions rarely hold perfectly in practice.

In this paper, we demonstrate how recent advances in automated partial identification allow researchers to easily adapt several common designs to accommodate potential violations of assumptions while still extracting as much information from data as possible. In particular, we provide several updates to the `autobounds` algorithm, including improved approaches to statistical inference and covariate adjustment, methods for handling continuous variables, sensitivity analyses, and diagnostics to aid interpretability. As our applications show, this approach offers several benefits and removes obstacles that have long impeded partial identification. First and foremost, this method is automated, and does not require the tedious and complex mathematical derivations that are currently necessary for bounding solutions in idiosyncratic settings. Second, the approach is fully flexible, allowing researchers not only to drop, but to partially relax, any assumption about the DGP while easily recomputing sharp bounds. This approach also addresses a frequent complaint of modern causal inference strategies—that they prompt a focus on narrow questions in order to satisfy the strictures of established research designs. With automated partial identification, researchers can hold their questions fixed even when assumptions fail and point identification is not possible, and still sharply bound the answer to the question that motivated their work to begin with. In extreme cases, this approach can also alert the researcher to inconsistencies between theory and data.

Further, we regard this approach as a boon for open science (Christensen et al., 2020). While the open science movement has traditionally emphasized making data and estimation transparent, our work focuses on a different aspect of the scientific workflow. Estimands and

identifying assumptions are crucial to replicating and extending prior work, yet are often left vague in applied work. To use our approach, all of these elements must be made explicit with precise definitions. Only when the target quantity and assumptions are transparently communicated can scientific debates yield meaningful progress.

Once applied, this technique can also reveal the most fruitful paths forward in a line of inquiry. Because our method precisely estimates the empirical implications of violations of assumptions, it can reveal which assumptions are most consequential. As we show above, in some cases, an apparently major violation may not alter a core conclusion. In others, even a small deviation from ideal conditions can overturn an inference. Researchers can then target their efforts toward the most consequential assumptions or seek data that obviate them.

While our proposed framework is broadly useful, several areas remain open for improvement. Although the identification problem is addressed, statistical inference for the resulting bounds is still an active area of research. For the no-covariate case, we argue for asymptotic pointwise validity of recentered subsampling under fixed-law stability conditions on the endpoint estimators; the covariate-adjusted uncertainty procedure, by contrast, should be understood as a practical extension rather than a general theorem for the full estimator. Incorporating more complex data structures—such as clustered standard errors—into the modeling also remains an open challenge. Because partial identification problems are often NP-hard, there are cases in which the computation may become intractable, even though this is not typically an issue in applied settings. One direction for improving tractability is to automate the derivation of symbolic solutions. This, too, remains an open research question.

Our approach is not a panacea and does not eliminate the need for careful design. Without sound theory and identification strategies that make assumptions plausible, the results produced by `autobounds` are unlikely to be informative. When used in conjunction with high-quality research designs, however, automated partial identification offers a powerful approach to learning from data under imperfect conditions.

# References


Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Acharya, A., M. Blackwell, and M. Sen (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *Biometrics 110*(3), 512–529.

Ahearn, C. E., J. E. Brand, and X. Zhou (2023). How, and for whom, does higher education increase voting? *Research in Higher Education 64*(4), 574–597.

Andrews, D. W. and P. Guggenberger (2009). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory 25*(3), 669–709.

Andrews, D. W. and P. Guggenberger (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory 26*(2), 426–468.

Andrews, D. W. and S. Han (2009). Invalidity of the bootstrap and the m out of n bootstrap for confidence interval endpoints defined by moment inequalities. *The Econometrics Journal 12*(suppl_1), S172–S199.

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

Angrist, J. D. and J̈.-S. Pischke (2010, Spring). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

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

Bandura, A. and R. H. Walters (1977). *Social learning theory*, Volume 1. Englewood cliffs Prentice Hall.

Bazzi, S. and M. A. Clemens (2013). Blunt instruments: Avoiding common pitfalls in identifying the causes of economic growth. *American Economic Journal: Macroeconomics 5*(2), 152–186.

Belotti, P., J. Lee, L. Liberti, F. Margot, and A. Wächter (2009). Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software 24*(4-5), 597–634.

Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science 57*(2), 504–520.

Bolusani, S., M. Besançon, K. Bestuzheva, A. Chmiela, J. Dionísio, T. Donkiewicz, J. van Doornmalen, L. Eifler, M. Ghannam, A. Gleixner, C. Graczyk, K. Halbig, I. Hedtke, A. Hoen, C. Hojny, R. van der Hulst, D. Kamp, T. Koch, K. Kofler, J. Lentz, J. Manns, G. Mexi, E. Mühmer, M. E. Pfetsch, F. Schlösser, F. Serrano, Y. Shinano, M. Turner, S. Vigerske, D. Weninger, and L. Xu (2024, February). The SCIP Optimization Suite 9.0. Technical report, Optimization Online.

Bonet, B. (2001). A calculus for causal relevance. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 40–47.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica 78*(2), 735–753.

Bugni, F. A. (2016). Comparison of inferential methods in partially identified models in terms of error in coverage probability. *Econometric Theory 32*(1), 187–242.

Burden, B. C. (2009). The dynamic effects of education on voter turnout. *Electoral studies 28*(4), 540–549.

Campbell, D. T. (2009). Prospective: Artifact and control1. *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*, 264.

Canay, I. A. (2010). El inference for partially identified models: Large deviations optimality and bootstrap validity. *Journal of Econometrics 156*(2), 408–425.

Chernozhukov, V., H. Hong, and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica 75*(5), 1243–1284.

Christensen, G., Z. Wang, E. Levy Paluck, N. Swanson, D. Birke, E. Miguel, and R. Littman (2020). Open science practices are on the rise: The state of social science (3s) survey.

Coppock, A. and D. P. Green (2016). Is voting habit forming? new evidence from experiments and regression discontinuities. *American Journal of Political Science 60*(4), 1044–1062.

Davenport, T. C., A. S. Gerber, D. P. Green, C. W. Larimer, C. D. Mann, and C. Panagopoulos (2010). The enduring effects of social pressure: Tracking campaign experiments over a series of elections. *Political Behavior 32*(3), 423–430.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature 48*(2), 424–455.

Duarte, G. (2026, May). A unified approach for assessing sensitivity to violations of causal assumptions. Working paper, Working paper.

Duarte, G., N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser (2024). An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association 119*(547), 1778–1793.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy 127*(3), 1210–1261.

Gallen, T. (2020). Broken instruments. *Available at SSRN 3671850*.

Gamrath, G., D. Anderson, K. Bestuzheva, W.-K. Chen, L. Eifler, M. Gasse, P. Gemander, A. Gleixner, L. Gottwald, K. Halbig, et al. (2020). The scip optimization suite 7.0.

Gerber, A. S., D. P. Green, and C. W. Larimer (2008, February). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review 102*(1), 33–48.

Glynn, A. N. and K. M. Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis 18*(1), 36–56.

Hasegawa, R. B., D. W. Webster, and D. S. Small (2019). Evaluating missouri's handgun purchaser law: a bracketing method for addressing concerns about history interacting with group. *Epidemiology 30*(3), 371–379.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

Huitfeldt, A., M. J. Stensrud, and E. Suzuki (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology 16*(1), 1.

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

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Keele, L. J. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis 23*(3), 313–335.

Keele, L. J. and W. Minozzi (2012). How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data. *Political Analysis 21*(2), 193–216.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–229.

Knox, D., W. Lowe, and J. Mummolo (2020). Administrative records mask racially biased policing. *American Political Science Review 114*, 619–637.

Knox, D., W. Lowe, and J. Mummolo (2026). Administrative records mask racially biased policing—corrigendum. *American Political Science Review 120*(1), 391–391.

Kocher, M. A., T. B. Pepinsky, and S. N. Kalyvas (2011). Aerial bombing and counterinsurgency in the vietnam war. *American Journal of Political Science 55*(2), 201–218.

Levis, A. W., M. Bonvini, Z. Zeng, L. Keele, and E. H. Kennedy (2025). Covariate-assisted bounds on causal effects with instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(5), 1508–1527.

Lundberg, I., R. Johnson, and B. M. Stewart (2021). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review 86*(3), 532–565.

Maher, S., M. Miltenberger, J. P. Pedroso, D. Rehfeldt, R. Schwarz, and F. Serrano (2016). PySCIPOpt: Mathematical programming in python with the SCIP optimization suite. In *Mathematical Software – ICMS 2016*, pp. 301–307. Springer International Publishing.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review Papers and Proceedings 80*(2), 319–323.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. F. and J. V. Pepper (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica 68*(4), 997–1010.

Mebane, W. R. and P. Poast (2013). Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis 22*(2), 169–182.

Monroe, K. R. (2005). *Perestroika!: The raucous rebellion in political science*. Yale University Press.

Mummolo, J. (2018). Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics 80*(1), 1–15.

Nyhan, B., C. Skovron, and R. Titiunik (2017). Differential registration bias in voter file data: A sensitivity analysis approach. *American Journal of Political Science 61*(3), 744–760.

Pearl, J. (1995a). Causal diagrams for empirical research. *Biometrika 82*(4), 669–710.

Pearl, J. (1995b). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435–443.

Pearl, J. (2009). *Causality*. New York: Cambridge University Press.

Politis, D. N., J. P. Romano, and M. Wolf (2001). On the asymptotic theory of subsampling. *Statistica Sinica*, 1105–1124.

Robins, J. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Romano, J. P. and A. M. Shaikh (2010). Inference for the identified set in partially identified econometric models. *Econometrica 78*(1), 169–211.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society 147*(5), 656–666.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 6*(5), 688–701.

Schafer, J., E. Cantoni, G. Bellettini, and C. Berti Ceroni (2022). Making unequal democracy work? the effects of income on voter turnout in northern italy. *American Journal of Political Science 66*(3), 745–761.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Schubiger, L. I. (2021). State violence and wartime civilian agency: Evidence from peru. *The Journal of Politics 83*(4), 1383–1398.

Sondheimer, R. M. and D. P. Green (2010). Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science 54*(1), 174–189.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters 78*(17), 2957–2962.

Verba, S., K. L. Schlozman, and N. Burns (2005). Family ties: Understanding the intergenerational transmission of political participation. In A. S. Zuckerman (Ed.), *Social logic of politics: Personal networks as contexts for political behavior*, pp. 95–116. Temple University Press.

Vigerske, S. and A. Gleixner (2018). Scip: Global optimization of mixed-integer nonlinear programs in a branch-and-cut framework. *Optimization Methods and Software 33*(3), 563–593.

Yang, F., J. R. Zubizarreta, D. S. Small, S. Lorch, and P. R. Rosenbaum (2014). Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician 68*(4), 253–263.

Ye, T., L. Keele, R. Hasegawa, and D. S. Small (2024). A negative correlation strategy for bracketing in difference-in-differences. *Journal of the American Statistical Association 119*(547), 2256–2268.

# Supporting Information

**A Detailed Example with `autobounds`** **44**

A.1 Simulated Data for Section 2.6 . . . . . 50

A.2 R Code for Section 2.6 . . . . . 52

**B An Application to Difference in Differences** **53**

B.1 Replication and Extension of Schubiger (2021) . . . . . 54

B.2 Code . . . . . 57

**C Statistical Uncertainty and Covariate Adjustment** **57**

C.1 Preliminaries . . . . . 57

C.2 Statistical Uncertainty . . . . . 59

C.2.1 Uncertainty Quantification without Covariates . . . . . 59

C.3 Simulation and Coverage . . . . . 63

C.4 Covariate Adjustment . . . . . 64

C.4.1 Validity of Covariate-averaged Bounds. . . . . 69

**D Model specification for Kocher et al. (2011)** **74**

**E Code** **75**

E.1 Instrumental Variables . . . . . 75

E.2 Selection Bias . . . . . 77

# A Detailed Example with `autobounds`

Consider a binary treatment, $D$, thought to cause a binary outcome, $Y$, where it is known that they share a set of unmeasured common causes, $U$. The analyst wishes to estimate the ATE, $\Pr(Y(1) = 1) - \Pr(Y(0) = 1)$. For instance, many studies have sought to estimate the causal effect of college education on voter turnout (e.g. Burden, 2009; Sondheimer and Green, 2010; Ahearn et al., 2023). Indeed, $U$ may contain factors such as income, home residence, age, parent education, some or all of which may be difficult or impossible to measure. Due to the presence of unobserved confounding, it is well known that this estimand is not identified and common means of estimating it, e.g. a difference in means, are biased. However, the ATE can be partially identified. That is, we can constrain its possible values. In the ensuing exposition, we will use the problem of estimating the effect of college education on turnout to illustrate how `autobounds` derives the bounds for the ATE.

The data generating process for this example is shown in Figure 1a. The DAG in this diagram is merely a graphical representation of the following causal model (Pearl, 2009)

$$D = f_D(U), \qquad Y \ = f_Y(d, U), \tag{2}$$

where $f_D$ and $f_Y$ are unspecified but *deterministic* functions of their (random) arguments. In words, we claim that obtaining an undergraduate degree is solely generated by unobserved factors $U$, e.g. income and geography, while voting is generated by both level of education and these unobserved factors for every subject under study.

Since all variables in the model are discrete, it is possible to enumerate all the possible ways in which $D$ and $Y$ are generated. To see this, note that since $D$ is binary, no matter what value $U$ takes, $f_D$ can only output the numbers 0 or 1. Therefore, for each $u \in \mathcal{S}(U)$—for example, a resident's age, income and home town—we can think of a function which labels whether or not that person would obtain an undergraduate degree

$$f_D^{(U=u)} : \varnothing \to \{0, 1\}, \tag{3}$$

a mapping we call the response function. The domain of this function is empty because $D$ is assumed to have no other causal parents than $U$. In other words, for each individual, once $U$ is fixed, $D$ will be determined. It follows then that this allows us to define two *disjoint* types of voter according to $U$: those who have an undergraduate degree and those who do not. Mathematically, this is equivalent to identifying which $u \in \mathcal{S}(U)$ activate treatment. That is we can define the random variable

$$D\text{-}\mathsf{trt} = \mathbf{1}\left\{f_D^U(\varnothing) = 1\right\} = \begin{cases} 1, & \text{for all } u \in \mathcal{S}(U) \text{ s.t. } f_D^U(\varnothing) = 1, \\ 0, & \text{otherwise} \end{cases}$$

which assigns treatment for any individual with characteristics $u$ that would determine their obtaining a college education. We can similarly define the control group as those with characteristics that lead them to avoid a college degree $D\text{-}\mathsf{ctl} = \mathbf{1}\left\{f_D^U(\varnothing) = 0\right\}$. We remark that these functions are random in $U$: only when $U$ is realized is your individual type determined; further, these groups are disjoint: a unit either has a bachelor's degree or they do not.

We can define a similar function for $Y$: upon fixing a voter's unobserved characteristics $u \in \mathcal{S}(U)$, only whether or not they obtain an undergraduate degree will determine if they vote, that is, the response function for $Y$ looks like

$$f_Y^{(U=u)} : \{0,1\} \to \{0,1\}, \tag{4}$$

which is a function of $D$ only (for each $u$). Now, since the input and output of $f_Y^{(U=u)}$ is binary, there are only four possible relationships the mapping in (4) could describe for any given $u$. These are $f_Y^{(U=u)}(d) = 1, f_Y^{(U=u)}(d) = 0, f_Y^{(U=u)}(d) = d, f_Y^{(U=u)}(d) = 1 - d$ for any $d \in \{0,1\}$. We call these functions *response types* In more familiar terms, the first two response types are the units whose voting preferences are unaffected by their education: the "always voters," $Y\text{-}\mathsf{av}$, who would vote regardless of their education and the "never voters," $Y\text{-}\mathsf{nv}$, who would never vote, regardless of their education. The third type are the "helped," $Y\text{-}\mathsf{he}$, individuals who would vote positively if they obtained a college degree, but would vote negatively without

one. The fourth type are the "hurt," $Y$-hu, who show the converse behavior to the helped. Any unit in the population many be described by one of these types, depending on their value of $U$, which is random. For example

$$\begin{aligned} Y\text{-he} &= \mathbf{1}\left\{f_Y^U(d=0)=0, f_Y^U(d=1)=1\right\} \\ &: \mathcal{S}(U) \to \{0,1\} \end{aligned}$$

which is a random function in $U$. The individual types with the indicators $D$-type and $Y$-type$'$ identify are called *principal strata* in the literature (Frangakis and Rubin, 2002).

At this juncture, it may seem that the exact nature of the support and even distribution of $U$ is important for describing probability distributions over the principal strata. However, the above exposition implies that this is not in fact the case. Intuitively, since $U$ is a common cause of a resident's education level and their decision to vote, the strata over $Y$ are not independent of those over $D$. These strata materialize jointly in the observed data. How many joint response types are there? For each two education types, there are four voting behaviors. Therefore, there are eight possible joint response types which describe all units in the population defined by the DGP in Figure 1. Informally, since $U$ dictates how $D$ and $Y$ are generated, it follows that by defining $U$ categorical with eight levels there is no loss of generality in the law governing the data generating process. Note that this conclusion is ignorant to the structure of $U$—it may be continuous, discrete, both or neither.

The previous exposition implies that we can represent any factual or counterfactual query in terms of the principal strata. In fact, this conclusion is exemplified in Proposition 2 of Duarte et al. (2024), which we restate below.

**Proposition 1.** *Suppose $\mathcal{G}$ is a canonical*[31] *DAG over a discrete causal model and define $\{C_l : l\}$ a set of counterfactual statements, indexed by $l$, in which $C_l = \{V_l(a_l) = v_l\}$ states that variable $V_l$ will take on value $v_l$ under manipulation(s) $a_l$. Let $\boldsymbol{U} = (U_1, \ldots, U_k)$ be the collection of all exogeneous variables in $\mathcal{G}$. Further, define $\mathbf{1}\{\boldsymbol{U} \implies \{C_l : l\}\}$ the indicator*

[31] A DAG is canonicalized by removing superfluous networks of exogeneous variables, distilling the DAG into its simplest form while losing no generality in the full data law. All DAGs can be canonicalized, and thus considering only this class of DAGs is unrestrictive. See section 3.1 of Duarte et al. (2024) for an example.

*function which takes on the value one if and only if the exogeneous realizations in* $\boldsymbol{U}$ *deterministically lead to* $C_l$ *being satisfied for every* $l$*. Then, under the structural equation model for* $\mathcal{G}$

$$\Pr\left(\bigcap_l C_l\right) = \sum_{\boldsymbol{U}\in\mathcal{S}(\boldsymbol{U})} \mathbf{1}\left\{\boldsymbol{U} \implies \{C_l : l\}\right\} \prod_k \Pr(U_k = u_k). \tag{5}$$

There are three key elements of this proposition we highlight here. Firstly, Equation 5 says that any factual[32] or counterfactual query may be equivalently expressed by first identifying which realizations of exogeneous variables in $\boldsymbol{U}$ generate that query, and second computing the likelihoods of hitting those realizations. Since all exogeneous variables are mutually independent, we may simply multiply the mass functions for each disturbance separately. Secondly, recalling that the previous exposition showed the equivalence between the exogenous disturbances and principal strata, we conclude that principal stratification is sufficient to describe any quantity we wish over a discrete causal model. Finally, the primary advantage of this result is that by connecting the disturbances to (counter)factual queries in this fashion, we may express any causal quantity we wish as a **polynomial** in the strata frequencies. This makes for ready use of modern optimization toolkits to quickly and accurately compute the bounds numerically.

Take, for instance, the probability that one has a Bachelor's degree but does not vote, $\Pr(D = 1, Y = 0)$. We may directly apply Proposition 1 and write

$$\Pr(D = 1, Y = 0) = \sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{u \implies \{D = 1, Y = 0\}\right\} \Pr(U = u), \tag{6}$$

$$= \sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_D^{(U=u)}(\varnothing) = 1, f_Y^{(U=u)}(d = 1) = 0\right\} \Pr(U = u). \tag{7}$$

It therefore follows that

$$\Pr(D = 1, Y = 0) = \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{nv}) + \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{hu}). \tag{8}$$

That is, this member of the electorate is the type of voter who possesses a Bachelor's degree,

[32] Factual queries correspond to the empty intervention, i.e. 'do nothing'.

and would not vote either (i) regardless of their education, or (ii) because of it. We cannot specify which of (i) or (ii) is true because we do not know how the subject would have voted had they not obtained a degree, i.e. another manifestation of the fundamental problem of causal inference (Holland, 1986).

We now return to the causal question of interest, the ATE. As we have just seen, an advantage of principal stratification is that we can turn queries about voting behavior into statements about the likelihood of being a certain type of voter. Focus on the first component of the ATE, $\mathbb{E}[Y(1)] = \Pr(Y(1) = 1)$ (because $Y$ is binary). Consider how an intervention on undergraduate education changes the structural causal model. We have

$$D = 1, \qquad Y(1) \;\; = f_Y(d = 1, U) \tag{9}$$

In this world, all units possess an undergraduate degree so there is no longer any randomness in the assignment of $D$. Subsequently, we generate the counterfactual $Y(1)$, which is the voting decision for a subject in a world where they are forced to obtain an undergraduate degree. To express the query $\Pr(Y(1) = 1)$ in terms of strata, observe that, if they vote in a world where a unit is forced to go to college, their $Y$-type can only be $Y$-he or $Y$-av. However, this logic assumes the unit is forced to obtain a college degree, i.e. where $U$ doesn't determine $D$; we do. The process of collecting observational data, though, does not provide us that luxury, since $U$ causes $D$ but is unknown. Therefore, this unit could be any of the four response types $(D\text{-trt}, Y\text{-he}), (D\text{-trt}, Y\text{-av}), (D\text{-ctl}, Y\text{-he}), (D\text{-ctl}, Y\text{-av})$, which yields

$$\Pr(Y(1) = 1) = \Pr(D\text{-trt}, Y\text{-he}) + \Pr(D\text{-trt}, Y\text{-av}) + \Pr(D\text{-ctl}, Y\text{-he}) + \Pr(D\text{-ctl}, Y\text{-av}).$$

By identical reasoning, it follows that $\Pr(Y(0) = 1) = \Pr(D\text{-trt}, Y\text{-av}) + \Pr(D\text{-trt}, Y\text{-av}) + \Pr(D\text{-ctl}, Y\text{-hu}) + \Pr(D\text{-ctl}, Y\text{-hu})$. Using these decompositions, the ATE may be expressed as:

$$\mathbb{E}[Y(1) - Y(0)] = \Pr(D\text{-trt}, Y\text{-he}) + \Pr(D\text{-ctl}, Y\text{-he}) - \Pr(D\text{-ctl}, Y\text{-hu}) - \Pr(D\text{-ctl}, Y\text{-hu}). \tag{10}$$

Even though this quantity is not point identified, it is always possible to derive bounds on its magnitude. The widest possible bounds are $[-1, 1]$, since $Y$ is binary. However, observational data may allow us to narrow these bounds significantly.

We now show how to form the polynomial program which `autobounds` uses to compute these bounds. To begin, we identify the parameters over which we optimize. These are the *strata frequencies*: $\Pr(D\text{-}\mathsf{type}), \Pr(D\text{-}\mathsf{type}\text{-}Y\text{-}\mathsf{type}')$ for all stratum types, which are of course unknown. Secondly, define the objective: this is the ATE in equation (10). Finally, identify constraints on the strata, these are: the laws of probability (strata frequencies must be contained in [0,1] and sum to unity), which we contain in the set $\mathcal{C}_{\mathcal{P}}$; and, the strata must combine to produce the observed data exactly, which we term the evidential constrains $\mathcal{C}_{\mathcal{E}}$. Thus, we aim to numerically solve

$$\underset{\substack{\Pr(D\text{-}\mathsf{type}),\\ \Pr(D\text{-}\mathsf{type},\, Y\text{-}\mathsf{type}')}}{\text{optimize}} \quad \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{he}) + \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{he}) - \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{hu}) - \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{hu}) \tag{11a}$$

$$\text{subject to} \quad 0 \leq \Pr(D\text{-}\mathsf{type}) \leq 1,\ 0 \leq \Pr(D\text{-}\mathsf{type}, Y\text{-}\mathsf{type}') \leq 1,\ \forall\, \mathsf{type}, \mathsf{type}' \tag{11b}$$

$$\sum_{\mathsf{type}} \Pr(D\text{-}\mathsf{type}) = 1,\ \sum_{\mathsf{type},\mathsf{type}'} \Pr(D\text{-}\mathsf{type}, Y\text{-}\mathsf{type}') = 1 \tag{11c}$$

$$\Pr(D = 0, Y = 0) = \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{he}) + \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{nv}) \tag{11d}$$

$$\Pr(D = 0, Y = 1) = \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{hu}) + \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{av}) \tag{11e}$$

$$\Pr(D = 1, Y = 0) = \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{nv}) + \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{hu}) \tag{11f}$$

$$\Pr(D = 1, Y = 1) = \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{he}) + \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{av}) \tag{11g}$$

Lines (11b)-(11c) are the constraints contained in $\mathcal{C}_{\mathcal{P}}$ while lines (11d)-(11g) are the constraints contained in $\mathcal{C}_{\mathcal{E}}$. Intuitively, in problem (11), the observed data constrain the possible compositions of voter types within the sample, from each of which we may compute an ATE. The output of the program is the best- and worst- case values of the ATE from those consistent with those compositions of voter types. The software `autobounds`, taking the DAG, estimand, observational data and any additional assumptions (e.g. monotonicity) as inputs, will build and solve this optimization program automatically. We note that because the data used to compute the probabilities in this optimization problem contain sampling error, these

bounds are estimated with statistical uncertainty; see Appendix C for details.

## A.1 Simulated Data for Section 2.6

Although `autobounds` is currently implemented in Python, we are actively developing an `R` version. The accompanying coded example is intended to show how the workflow from Section 2.6 can be translated into `R`, highlighting the interoperability and portability of the `autobounds` framework across languages while preserving the core logic of the analysis.

First some notation. In the previous section, we wrote $D$-type and $Y$-type to denote the principal strata for the treatment and outcome variables, as in those cases type had a simple, intuitive meaning for each possible type. This convenience is an artifact of the low cardinality of the problem, $D$ had no observed parents and the only observed parent of $Y$ was a binary variable, $D$. In the example we describe in section 2.6, the addition of the (even binary) observed parent $X$ as a common cause of $D$ and $Y$ increases the cardinality of the problem to the point that this group partisanship notational style becomes cumbersome: we would need a name for each stratum and some manner in which to remember them. Consequently, we adapt the notation to an algebraic style which we describe as follows: the object $V_v$ denotes the principal stratum of the factual variable $V$ with index $v$. The index $v$ has digits $(v_0 v_1 v_2 \dots v_{n-1})$, where each $v_i$ is the value that $V$ takes when its parents are set to the configuration associated with position $i$ in some fixed ordering of all possible parent value combinations (e.g. topological ordering). This implies that the index $v$ is a number written in base-$b$, where $b = |\mathcal{S}(V)|$ and the number of digits in $v$ is equal to $n = \prod_{W \in \mathsf{pa}_V} |S(W)|$, where $\mathsf{pa}_V$ is used to denote the observed parents of $V$; if $\mathsf{pa}_V$ is empty, we set $n = 1$. The reason why $v$ is represented in a base equal to the cardinality of $V$ is because each digit in $v$ represents the **value** that $V$ takes under a specific instantiation of its parents.[33]

This is best illustrated with an example. Consider the simple $D$-$Y$ confounding problem

[33]For this to hold consistently we assume all variables of interest take values in $\{0, 1, \dots, |S(V)|\}$, although this nomenclature can be generalized to variables that take on a finite number of nonconsecutive integer values. However, when the support of each variable $V$ is conveniently within the set $\{0, 1, \dots, |S(V)|\}$ notice that the index $v$, when read as a number in base-$b$, actually *counts* the number of strata which exist for variable $V$ alone.

as discussed in section A. Here, $D$ has no observed parents, therefore its strata are described by $D_d$ where $d$ is a number in base-2 ($b = |\mathcal{S}(D)| = 2$), of length unity ($n = 1$). As $D$ takes on values only 0 and 1, the strata for $D$ are $D_0$ and $D_1$. Now, $Y$ has one observed parent, $D$, which is binary. Therefore, $y$ is a binary number of length 2. Taking $y = y_0y_1$, we now specify an ordering for configurations of the parents of $Y$. Let's say $y_0$ represents the value of $Y$ when $D = 0$ and $y_1$ represents the value of $Y$ when $D = 1$. This ordering is mathematically arbitrary but we make this choice for readability. In this way, we have the equalities $Y_{00} = Y\text{-never}, Y_{01} = Y\text{-comply}, Y_{01} = Y\text{-defy}$ and $Y_{11} = Y\text{-always}$.

Now, we apply this to our introductory example in section 2.6. Here, $X$ has no parents, so its index $x$ is a single binary digit. The binary treatment $D$ has one binary parent ($X$) and so $d$ is a binary number of length 2, while the binary outcome $Y$ has two binary parents, so its index $y$ is a number in base-2 of length 4. Next we impose an ordering on the indices. For $X$, this is trivial. Letting $d = d_0d_1$, take $d_i$ the value of $D$ when $X = i$. Writing $y = y_{00}y_{01}y_{10}y_{11}$, let $y_{ij}$ represent the value of $Y$ when $D = i$ and $X = j$.

To simulate data for the introductory example, we assume that $X \sim$ Bernoulli(0.6) (so $\Pr(X_x) = 0.6^x(1 - 0.6)^{(1-x)}$),

$$\Pr(D_{00}Y_{0000}) = 0.074576212961429 \qquad \Pr(D_{10}Y_{1001}) = 0.134412818663854$$
$$\Pr(D_{01}Y_{0001}) = 0.0301702777607751 \qquad \Pr(D_{10}Y_{1010}) = 0.0480346194425266$$
$$\Pr(D_{01}Y_{0101}) = 0.186607173274612 \qquad \Pr(D_{10}Y_{1111}) = 0.0940165222917155$$
$$\Pr(D_{01}Y_{1101}) = 0.00798467737923486 \qquad \Pr(D_{11}Y_{1100}) = 0.0345830964246296$$
$$\Pr(D_{10}Y_{0010}) = 0.348297597134471 \qquad \Pr(D_{11}Y_{1101}) = 0.041317004666752.$$

and all other strata have probability zero.

With this setup, using ideas discussed in the previous section, we compute the true ATE as the relevant combinations of these strata and then produce the observed law on which we apply `autobounds`. We also make sure that this DGP satisfies the weak monotonicity assumption. See `a_coded_example_simulation.py` for more details.

## A.2 R Code for Section 2.6

To illustrate this portability of appendix A.1 in practice, below we provide a compact R translation of the workflow used in Section 2.6. The example mirrors the confounding setup in the main text and shows how the same problem can be specified, solved, and then extended with covariate adjustment and a monotonicity restriction using the R interface.

```r
if (requireNamespace("autobounds", quietly = TRUE)) {
  library(autobounds)
} else {
  pkgload::load_all(".")
}
python_bin <- Sys.getenv(
  "AUTOBOUNDS_PYTHON",
  unset = "/home/xenakis/.pyenv/versions/3.13.5/bin/python"
)
autobounds_configure(python = python_bin)
set.seed(2138)
n <- 2500
X <- rbinom(n, 1, 0.55)
U <- rnorm(n)
D_latent <- -0.2 + 0.9 * X + 1.0 * U + rnorm(n)
D <- as.integer(D_latent > 0)
Y0_latent <- -0.5 + 0.6 * X + 0.9 * U + rnorm(n)
tau <- 0.35 + 0.10 * X
Y1_latent <- Y0_latent + tau
Y0 <- as.integer(Y0_latent > 0)
Y1 <- as.integer(Y1_latent > 0)
Y <- ifelse(D == 1, Y1, Y0)
confounding_data <- data.frame(X = X, D = D, Y = Y)
# Baseline confounding problem from Section 2
confounding_model <- dag(
  c("D -> Y", "X -> D", "X -> Y", "U -> D", "U -> Y"),
  unob = "U"
)
confounding_problem <- problem(confounding_model)
set_ate(confounding_problem, treatment = "D", outcome = "Y")
read_data(confounding_problem, data = confounding_data)
confounding_solution <- ab_solve(
  confounding_problem,
  theta = 0.005,
  verbose_result = FALSE
)
bounds(confounding_solution)
# Covariate adjustment via X rather than explicit stratification in the DAG
confounding_model_no_x <- dag(
  c("D -> Y", "U -> D", "U -> Y"),
  unob = "U"
)
confounding_problem_with_x <- problem(confounding_model_no_x)
read_data(confounding_problem_with_x, data = confounding_data, covariates = "X")
set_ate(confounding_problem_with_x, treatment = "D", outcome = "Y")
confounding_solution_with_x <- ab_solve(
  confounding_problem_with_x,
  theta = 0.005,
  verbose_result = FALSE
)
bounds(confounding_solution_with_x)
# Monotone response across observed subgroups
confounding_problem_monotone <- problem(confounding_model)
```

```
read_data(confounding_problem_monotone, data = confounding_data)
set_ate(confounding_problem_monotone, treatment = "D", outcome = "Y")
turnout_if_college_high <- E(confounding_problem_monotone, "Y(D=1)",
                             condition = "X=1")
turnout_if_college_low <- E(confounding_problem_monotone, "Y(D=1)",
                            condition = "X=0")
turnout_if_no_college_high <- E(confounding_problem_monotone, "Y(D=0)",
                            condition = "X=1")
turnout_if_no_college_low <- E(confounding_problem_monotone, "Y(D=0)",
                            condition = "X=0")
assume_geq(confounding_problem_monotone, turnout_if_college_high,
                               turnout_if_college_low)
assume_geq(confounding_problem_monotone, turnout_if_no_college_high,
                            turnout_if_no_college_low)
monotone_solution <- ab_solve(
  confounding_problem_monotone,
  theta = 0.005,
  verbose_result = FALSE
)
bounds(monotone_solution)
```

The same simulated confounding setup can also be extended to the continuous-outcome mode described in Section 2.1. In the notebook example, we replace the binary turnout measure with a continuous outcome on the unit interval and target the same marginal ATE, $\mathbb{E}[Y(d=1) - Y(d=0)]$. The true ATE in this DGP is 0.113. Using the exact observed-law construction based on sharp bounds for $\mathbb{E}[Y(1) \mid X = x]$ and $\mathbb{E}[Y(0) \mid X = x]$ within each covariate stratum and then averaging over $X$, the corresponding ATE bounds are $[-0.34, 0.66]$. The continuous-outcome routine in `autobounds` approximates this target by discretizing $Y$ into ordered bins, solving binary threshold subproblems, and aggregating the resulting bounds back to the ATE. In the example reported here, the 10-bin approximation yields $[-0.39, 0.71]$. Using more bins can further refine the approximation, and the procedure remains conservative throughout.

# B An Application to Difference in Differences

Difference-in-differences (DiD) is another widely used identification strategy designed to neutralize the influence of unobserved confounding. In the simplest case, the DiD strategy involves comparing the outcomes of two groups of observations (treatment and control) in two time periods (before and after some intervention is applied to the treatment group only). The average pre-post difference in outcomes is computed within both groups, and then the treatment

Figure 7: **Difference-in-differences DAGs.** (a) Standard DiD model. (b) DiD with bracketing on covariates.


difference is compared to the control difference.

This comparison can identify the average treatment effect among the treated (ATT), under the key identifying assumption of parallel trends—i.e., that in the absence of any intervention, the pre-post differences in both treatment and control groups would be equal, so that average outcomes would move in parallel. This assumption is generally regarded as not directly testable, since by definition the treatment group's counterfactual trend in the post period cannot be observed. Researchers typically examine pre-trends to see if outcomes were moved in parallel prior to the intervention, but such tests are not dispositive and are of limited use when extensive pre-treatment data is not available.

## B.1 Replication and Extension of Schubiger (2021)

We next replicate and extend Schubiger (2021), which relies in part on a DID strategy to estimate the effect of exposure to state violence on counterinsurgent mobilization in the Peruvian Civil War. As the study states, "The core challenge to answering this question lies in the fact that even though state violence was highly unpredictable during the counterinsurgency campaign of 1983–85, targeting did not occur at random, thus being potentially related to other important determinants of communities' propensity for counterinsurgent collective action," (p. 1388).

In this study, the units of analysis are *centro poblados*, "settlements of various sizes and types, such as villages and towns," some of which are targets of state violence. This analysis examines two time periods: 1983–1985, when human rights violations and other state violence were imposed on various towns and villages in response to a counterinsurgency, (the pre

period); and 1986–1988, during which time some localities employed "self-defense committees" which engaged in violent clashes with the state (the post period). The analysis examines two groups of localities: those that experienced state violence in the pre-period (the treatment group) and those that did not (the control group). The outcome is a binary indicator of "whether a given *centro poblado* was affected by violence against or perpetrated by self-defense committees in the period after the counterinsurgency campaign (1986–88)" (1390). However, as Schubiger (2021) notes, "As there is only one pretreatment period, pretreatment trends cannot be explored in detail..." (p. 1395).

We demonstrate how `autobounds` can be used to relax the parallel trends assumption by replicating and extending Schubiger (2021). The DAG of Figure 7a is one of several causal graphs that is consistent with the DiD design. As the graph shows, the treatment, $D$, the outcome in the pre period, $Y_{t=0}$, and the outcome in the post period, $Y_{t=1}$, are all caused by a common sent of unobserved confounders, $U$, which do not evolve in time. In other words, consistent with the traditional representation of DiD, the treatment and control groups differ in unobserved ways, and in turn, their levels of the outcome in all periods are not the same. However, a standard DiD analysis also imposes a functional restriction on the DAG, namely that the effect of $U$ on $Y_t$ does not evolve with $t$. In other words, absent treatment, the evolution of the outcome from pre- to post- treatment would be equal across groups. Mathematically, we assume that

**Assumption B.1.1** (Parallel Trends)**.** *The trend among the treated group,* $\mathbb{E}[Y_{t=1}(d=0) - Y_{t=0}(d=0) \mid D=1]$ *is equal to the trend among the control group,* $\mathbb{E}[Y_{t=1}(d=0) - Y_{t=0}(d=0) \mid D=0]$.

Translating this into `autobounds` is also straightforward:

```
# for clarity in code, we write Y_t0 as Ya and Y_t1 as Yb
with respect_to(peru_problem):
    trend_treated = p("Yb(D=0)=1", cond="D=1") - p("Ya=1", cond="D=1")
    trend_control = p("Yb(D=0)=1", cond="D=0") - p("Ya=1", cond="D=0")
    add_assumption(trend_treated, "==", trend_control)
```

Two estimands were analyzed. For the ATT, we obtained estimates in which the lower and upper bounds collapse to a point estimate of 0.047 (95% CI $[0.028, 0.069]$). This is exactly the

result obtained by Schubiger (2021). However, `autobounds` also allows us to calculate bounds for the ATE, and when we do, we are able to sign the effect as positive: $[0.001, 0.956]$ (95% CI $[0.000, 0.959]$). The parallel-trends assumption is crucial to these conclusions, as results without it are far less informative: for the ATE, bounds would have been $[-0.044, 0.956]$ (95% CI $[-0.048, 0.959]$), and for the ATT, $[-0.948, 0.052]$ (95% CI $[-0.961, 0.077]$). This also demonstrates a key feature of `autobounds`: the ability to easily obtain informative results about multiple estimands (i.e. multiple research questions) under multiple assumption sets.[34]

Finally, we relax the classic parallel-trends assumption using a type of bracketed trends (Campbell, 2009; Hasegawa et al., 2019; Ye et al., 2024), incorporating the background covariate $X$ depicted in the graph of Figure 7b. In contrast to the parallel-trends assumption, which states that background changes in the treated group's outcome are exactly balanced with with those in the control group, the bracketed-trends assumption say something weaker. Specifically, it states that these unobserved background changes in the treated group are sandwiched somewhere between the observed changes in groups which did not receive treatment. In other words, if one control group evolves faster than the treated group, and another evolves more slowly, then we can infer that the treated group's counterfactual would have fallen somewhere in between. To calculate the brackets, we use a single covariate: whether there was prior insurgent violence—i.e., whether the treatment affected that location in a previous period. Formally, if $\Delta(D = 1) \coloneqq \mathbb{E}[Y_{t=1}(d = 0) - Y_{t=0}(d = 0) \mid D = 1]$ is the treated group's trend, and $\Delta(D = 0, X = x) \coloneqq \mathbb{E}[Y_{t=1}(d = 0) - Y_{t=0}(d = 0) \mid D = 0, X = x]$ is the trend in the subset of the control group with $X = x$, then the bracketed trends assumption is as follows:

**Assumption B.1.2** (Bracketed Trends)**.** $\min\{\Delta(D = 0, X = 0), \Delta(D = 0, X = 1)\} \leq \Delta(D = 1) \leq \max\{\Delta(D = 0, X = 0), \Delta(D = 0, X = 1)\}$

Under this assumption, we find that the bounds for the ATT are $[0.006, 0.046]$ (95% CI $[-0.013, 0.065]$), and the bounds for the ATE are $[-0.042, 0.955]$ (95% CI $[-0.044, 0.958]$). As this final example shows, replacing the relatively restrictive parallel trends assumption with

[34] When assuming parallel trends, results for the ATE and ATT are identical regardless of which DAG is used; this is because the parallel trends assumption implies that the covariates $\mathbf{X}$ are irrelevant after accounting for their contribution to the pre-treatment outcome.

the more lenient bracketing trends assumption, analysts are still able to estimate the sign of the ATT as positive, though statistical significance is lost; the bounds for the ATE no longer indicate the direction of this effect, as the upper and lower bounds cross zero. This shows that in situations where parallel trends are thought to be violated, it is still possible to recover substantively informative results in a difference-in-differences setting using automated partial identification.

## B.2 Code

```
# DAG for this problem allows confounding of treatment D
# and pre/post outcomes Ya/Yb
peru_problem = causalProblem(
    DAG("U -> Ya, U -> D, U -> Yb, D -> Yb", unob="U")
)

# load data with D/Ya/Yb columns, one row per settlement
peru_data = pandas.read_csv("peru_data.csv")

# all statements below are w.r.t. this problem
with respect_to(peru_problem):
    # give data to autobounds, prepare to compute 95% CI
    read_data(peru_data, inference=True)
    # assume treated & control trends are parallel
    trend_treated = E("Yb(D=0)", cond="D=1") - E("Ya=1", cond="D=1")
    trend_control = E("Yb(D=0)", cond="D=0") - E("Ya=1", cond="D=0")
    add_assumption(trend_treated, "==", trend_control)
    # set estimand to be ATT: local ATE among treated (D=1)
    set_ate(ind="D", dep="Yb", cond="D=1")
    # calculate bounds
    peru_bounds = solve(ci=True, nsamples=1000)
```

# C Statistical Uncertainty and Covariate Adjustment

Next, we consider two practical considerations in applied research: (1) quantification of statistical uncertainty due to sampling error, and (2) adjustment for background covariates that are not of primary interest.

## C.1 Preliminaries

Before formalizing the proposed methods, we first introduce additional notation and key concepts. Let $P_{\boldsymbol{V}(\mathcal{W})}$ represent the *full data law*—that is, the full joint distribution over

all possible factual and counterfactual versions of variables $\boldsymbol{V}$ in response to every possible intervention $w \in \mathcal{W}$—in the population of interest. This distribution is generally unknowable, but it is a useful construct: all possible quantities, including observable factual quantities and unobservable counterfactual quantities of interest, are determined by the full data law.[35] We will use $\varphi(P_{\boldsymbol{V}(\mathcal{W})})$ to represent the estimand; this function essentially takes a full data law and reduces it down to one particular quantity of interest, such as the ATE. Where the full data law being discussed is clear from context, we will drop the argument and write $\varphi \coloneqq \varphi(P_{\boldsymbol{V}(\mathcal{W})})$. Let $P_{\boldsymbol{V}}$ denote the *observed data law*, a marginal of $P_{\boldsymbol{V}(\mathcal{W})}$ containing only the factual versions of variables in $\boldsymbol{V}$.

Next, let $\underline{\mathsf{A}}_\varphi$ and $\overline{\mathsf{A}}_\varphi$ represent the deterministic `autobounds` functions that respectively compute sharp lower and upper bounds on the estimand $\varphi$ by solving polynomial programs to global optimality when supplied some set of observed information. If analysts somehow possessed perfect information on the full data law $P_{\boldsymbol{V}(\mathcal{W})}$, then supplying this information to `autobounds` would point identify the quantity of interest, so that $\underline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}(\mathcal{W})}) = \overline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}(\mathcal{W})}) = \varphi(P_{\boldsymbol{V}(\mathcal{W})})$. If analysts possessed factual data on all units in the population, so that the observed law $P_{\boldsymbol{V}}$ is perfectly observed, then applying `autobounds` to this information would yield the *population bounds* $[\underline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}}), \overline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}})]$. In the more common case where analysts possess only a sample of units and must estimate the observed law, $\hat{P}_{\boldsymbol{V}}$, then applying `autobounds` to these inputs will yield the *estimated bounds* $[\underline{\mathsf{A}}_\varphi(\hat{P}_{\boldsymbol{V}}), \overline{\mathsf{A}}_\varphi(\hat{P}_{\boldsymbol{V}})]$ instead.

**A Running Example.** Consider a simple confounding scenario in which a binary treatment $D$ causes a binary outcome $Y$, with $D$ and $Y$ confounded by an unobserved $U$. The main variables are $\boldsymbol{V} = [D, Y]^\top$; the full data law $P_{\boldsymbol{V}(\mathcal{W})}$ is the distribution over $D$, $Y(d=0)$, and $Y(d=1)$; a common estimand is the ATE, $\varphi(P_{\boldsymbol{V}(\mathcal{W})}) = P_{\boldsymbol{V}(\mathcal{W})}(Y(d=1)=1) - P_{\boldsymbol{V}(\mathcal{W})}(Y(d=0)=1)$; and the observed data law $P_{\boldsymbol{V}}$ is the distribution over $D$ and $Y$ only. Given a sample of factual variables on $N$ units, the empirical analog of the observed data law is $\hat{P}_{\boldsymbol{V}}(d, y) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(D_i = d, Y_i = y)$.

[35] `autobounds` works by reasoning about possible full data laws that are consistent with available information. Specifically, it does so by reasoning about possible joint distributions of principal strata, which is one way to represent the full data law.

## C.2 Statistical Uncertainty

If data is available on the entire population of interest, `autobounds` derives sharp *population bounds*. These are ranges of possible answers that account for fundamental uncertainty, rather than statistical uncertainty: if key variables are confounded, even with an infinite number of observations, researchers may still only be able to recover a range of possible answers rather than uniquely identifying a single one. These population bounds can be calculated by supplying `autobounds` with perfectly measured information on the population distribution of the observed variables. When this distribution is measured with sampling error, computing bounds that ignore this error—i.e., treating the empirical distribution as if it were the population distribution, via the plug-in principle—`autobounds` produces *estimated bounds*. Finally, when directed to account for this sampling error in the observed quantities, `autobounds` will widen the estimated bounds to obtain *confidence bounds*.

As the sharp-bounding functions $\underline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})$ and $\overline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})$ are generally non-smooth functions of the observed data law, performing inference on estimators of these quantities by standard methods, e.g. delta method or bootstrap, is often intractable. This is because the max and min operators are typically nonsmooth, whereas the validity of asymptotic approximations relies on local linearization—such as von Mises or Taylor expansions—which in turn requires Hadamard differentiability. For this reason, a common approach in the literature is to replace max and min with smooth approximations, such as log-sum-exp transformations (Levis et al., 2025). These approximations, however, introduce tuning parameters and may provide a poor approximation to the underlying extremum functionals.

### C.2.1 Uncertainty Quantification without Covariates

The frequentist approach we employ here is to approximate the sampling distribution of each bound statistic using subsampling. Subsampling has been used more broadly in the econometrics literature as a tool for conducting inference in moment-inequality problems and other settings with nonsmooth extremum statistics; see, for example, Chernozhukov et al. (2007); Romano and Shaikh (2010); Andrews and Guggenberger (2009, 2010). Our implementation

is adapted to the present setting, where the lower and upper endpoints are computed numerically by `autobounds` rather than given by a closed-form moment-inequality statistic, and is particularly attractive in our setting because it does not rely on smooth local linear approximations. In contrast to bootstrap procedures that do rely on such approximations, subsampling only requires recomputing the statistic on smaller samples drawn from the observed data, and therefore naturally accommodates the extremum structure of sharp bounds (Politis et al., 2001). Throughout, the object of inference is the population identified set

$$\left[\underline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}}), \overline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})\right],$$

rather than only the plug-in estimate based on the empirical observed-data law.

The confidence interval we report is a pair of *confidence bounds* for the lower and upper endpoints of the population identified set. Let $\underline{\hat{\varphi}}_n \coloneqq \underline{\mathsf{A}}_{\varphi}(\hat{P}_{\boldsymbol{V}})$ and $\overline{\hat{\varphi}}_n \coloneqq \overline{\mathsf{A}}_{\varphi}(\hat{P}_{\boldsymbol{V}})$ denote the lower and upper estimated bounds computed from the full sample of size $n$. A two-sided confidence interval for the identified set can then be formed by constructing a one-sided lower confidence bound for $\underline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})$ and a one-sided upper confidence bound for $\overline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})$, using a Bonferroni correction to ensure simultaneous coverage. Intuitively, we widen the plug-in interval enough to account for the finite-sample variability of each endpoint separately.

To do so, we use *recentered subsampling*. Let $m = m_n$ denote the subsample size, with $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. In our implementation we use the rule

$$m_n = \lfloor n^{\gamma} \rfloor,$$

with default choice $\gamma = 2/3$, a commonly used practical calibration in the partially identified-models literature and one that performs well in comparative studies of coverage error for subsampling procedures (Bugni, 2016), together with finite-sample safeguards that prevent the subsample from becoming too small. We then draw $B$ subsamples without replacement from the original dataset, each of size $m$. For each subsample $b \in \{1, \ldots, B\}$, we recompute

the bound estimators

$$\underline{\hat{\varphi}}_m^{(b)} \coloneqq \underline{\mathsf{A}}_\varphi\left(\hat{P}_{\boldsymbol{V},m}^{(b)}\right) \qquad \text{and} \qquad \hat{\overline{\varphi}}_m^{(b)} \coloneqq \overline{\mathsf{A}}_\varphi\left(\hat{P}_{\boldsymbol{V},m}^{(b)}\right),$$

where $\hat{P}_{\boldsymbol{V},m}^{(b)}$ is the empirical observed-data law computed from the $b$th subsample.

The basic idea is to use the empirical distribution of the recentered endpoint statistics

$$T_L^{(b)} = \sqrt{m}\left(\underline{\hat{\varphi}}_m^{(b)} - \underline{\hat{\varphi}}_n\right), \tag{12}$$

$$T_U^{(b)} = \sqrt{m}\left(\hat{\overline{\varphi}}_m^{(b)} - \hat{\overline{\varphi}}_n\right), \tag{13}$$

to approximate the unknown sampling distribution of the full-sample bound estimators. Recentring at the full-sample estimates is important: it aligns the subsample statistics with the distribution we wish to approximate and avoids introducing additional bias from the fact that the identified set itself is not centered at zero.

Let $\hat{q}_{L,\alpha}$ denote the empirical $\alpha$-quantile of $\{T_L^{(b)}\}_{b=1}^B$, and let $\hat{q}_{U,\alpha}$ denote the empirical $\alpha$-quantile of $\{T_U^{(b)}\}_{b=1}^B$. The resulting $(1-\alpha)$ confidence interval for the identified set is

$$\left[\underline{\hat{\varphi}}_n - \frac{\hat{q}_{L,1-\alpha/2}}{\sqrt{n}},\ \hat{\overline{\varphi}}_n - \frac{\hat{q}_{U,\alpha/2}}{\sqrt{n}}\right]. \tag{14}$$

The lower endpoint uses the upper tail of the subsampling distribution of $T_L^{(b)}$ because we require a one-sided lower confidence bound for $\underline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}})$. Symmetrically, the upper endpoint uses the lower tail of the subsampling distribution of $T_U^{(b)}$ because we require a one-sided upper confidence bound for $\overline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}})$. If the one-sided endpoint procedures are valid, then Bonferroni's inequality implies that this interval is conservative for the entire partially identified region, and therefore also for the true scalar estimand whenever that estimand lies inside the region.

This procedure has several practical advantages in the present setting. First, it is agnostic to which constraint or moment happens to bind the lower or upper optimization problem in the population. This is critical because finite-sample perturbations can change the active constraint set, generating nonregular behavior exactly where standard bootstrap or delta-method

approximations are unreliable. Second, the procedure is modular: each subsample simply reruns the same `autobounds` program on a smaller empirical distribution, so no symbolic derivative calculations or closed-form characterization of the bound functionals are required. Third, because the method only requires repeated evaluation of the numerical optimization routine, it extends immediately to the broad class of discrete causal models handled by `autobounds`.

One caveat, emphasized by Andrews and Guggenberger (2009, 2010), is that subsampling and related plug-in asymptotic procedures need not be uniformly valid when the distribution of the statistic changes discontinuously across nearby data-generating processes. Their negative results are driven by drifting sequences under which the parameter effectively depends on $n$, so that the limiting experiment itself changes with the sequence and the nominal asymptotic approximation may fail. We view that concern as relevant here as well: if the identity of the binding extrema changes along such local sequences of observed-data laws, the law of the bound estimator may exhibit the same kind of discontinuous behavior. Accordingly, we do not claim uniform validity over all sequences of data-generating processes. The more modest claim we rely on is asymptotic pointwise validity: for a fixed population law, provided the lower- and upper-endpoint estimators admit stable pointwise asymptotic distributions and the recentered subsampling distribution consistently estimates those laws, the resulting one-sided endpoint procedures are asymptotically valid. In other words, the main concern is failure of uniform validity in neighborhoods of nonregular points, not failure of pointwise validity at a fixed, sufficiently regular data-generating process.

There is, however, a finite-sample tradeoff in choosing $m$. If $m$ is too close to $n$, the method begins to resemble the bootstrap and may fail to adequately capture nonregularity. If $m$ is too small, each subsample becomes too noisy, which can noticeably widen the resulting confidence interval. In our simulations, the default choice $\gamma = 2/3$ performs well across a range of designs, consistent with the comparative evidence in Bugni (2016), though very small samples can still be challenging. In short, the subsampling procedure is designed to prioritize validity and robustness in the presence of nonsmooth bound estimators, even at the cost of some conservativeness.

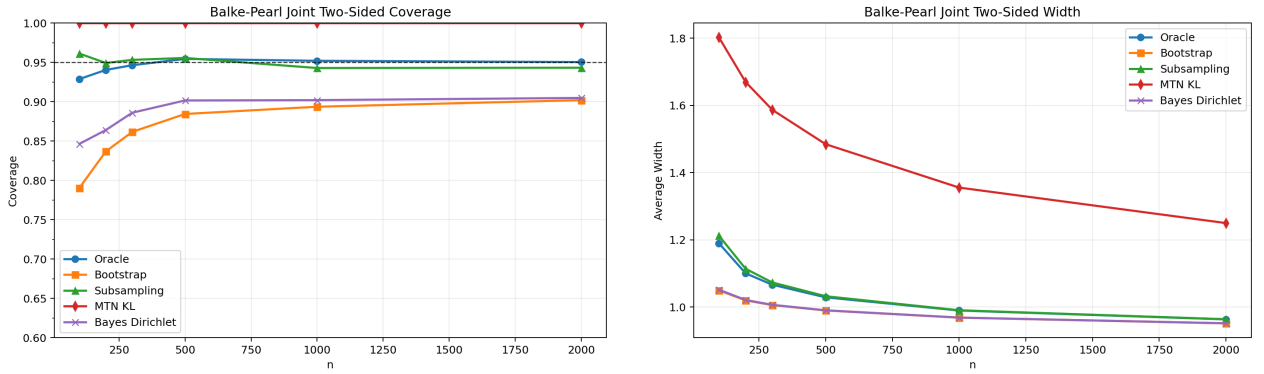
## C.3 Simulation and Coverage

Before turning to covariate adjustment, we compare the finite-sample behavior of several uncertainty-quantification procedures using the same IV simulation that was studied in Duarte et al. (2024). In this simulation, the population bounds for the ATE are $[-0.33, 0.56]$. However, the original KL-based method proposed in Duarte et al. (2024) is clearly too conservative—an issue that was acknowledged as a major limitation in that work. With $n = 1000$, it yields extremely wide confidence bounds that have an average width of 1.35. In other words, the portion of the confidence region that accounts for statistical uncertainty add a width of 0.46, above and beyond the width of the population bounds (0.89) which would remain even with infinite samples. In contrast, our newly proposed subsampling approach achieves nominal coverage rates (94.3%, compared to 100% coverage by the original KL-based method) with dramatically narrower confidence regions. On average, the confidence bounds have a width of 0.99, the portion of this accounting for statistical uncertainty contributes a width of 0.1.

Figure 8 reports the joint coverage and average joint interval width for the original KL-divergence method and the new recentered-subsampling procedure proposed here. We also evaluate a hypothetical "oracle" method that is infeasible in practice because it requires information about the true DGP that an analyst would not normally have. We find that the proposed recentered-subsampling procedure achieves performance that is competitive with this infeasible oracle procedure.

For reference, we also evaluate the performance of two alternative methods: the ordinary bootstrap, and a new quasi-Bayesian procedure. In this IV design example, the candidate maximizers and minimizers are only weakly separated, so small sampling fluctuations can easily switch which formula is active. This is the setting where bootstrap-based Gaussian approximations fail in practice. Because the bound map is nonsmooth at such near-ties, the bootstrap tends to behave as if the active extrema were stable when in fact they are not, and the resulting confidence intervals are too narrow. That is exactly what the simulations show: bootstrap coverage falls well below nominal levels, even though its average width remains

relatively small. The Dirichlet posterior method also fails to deliver strong frequentist coverage in this design, which is not surprising because it is a Bayesian procedure and there is no general frequentist guarantee once the Bernstein–von Mises conditions break down for these nonsmooth functionals (Van der Vaart, 2000).




(a) Joint coverage under weak separation.

(b) Average joint interval width under weak separation.

Figure 8: Finite-sample performance in a nearly tied IV design.

## C.4 Covariate Adjustment

Next, we turn to the scenario in which analysts possess some background covariates that are not of primary interest, for which they would like to adjust. When these background covariates are continuous, the theory developed in Duarte et al. (2024)—which is developed for fully discrete settings—requires some extension before it can be used. Here, we develop a model-based technique, utilizing a generalized linear model, for conducting this covariate adjustment. Because this approach relies on modeling assumptions that are unlikely to hold exactly, we recommend that the results be regarded as an *approximate* bias correction step, much like augmentation of potentially misspecified inverse propensity weighted estimators (Robins et al., 1994; Scharfstein et al., 1999; Glynn and Quinn, 2010) or linear corrections with inexact matching estimators (Abadie and Imbens, 2011).

We now extend the subsampling approach to uncertainty quantification to the common scenario where researchers wish to adjust for covariates to eliminate potential sources of confounding. We will suppose that these covariates are nuisances, in the sense that researchers

would like to estimate aggregate quantities that average over all covariate values, such as the ATE, rather than comparing conditional effects at different covariate values. We also develop an algorithm to perform statistical inference over these covariate-averaged bounds.

To fix the setting, suppose that as before, we possess dataset $[\mathbb{V}, \mathbb{X}] = \{[\boldsymbol{V}_i, \boldsymbol{X}_i]\}_{i=1}^N$ where each row $i$ is a sample from a now-expanded observed data law, $P_{\boldsymbol{VX}}$. Beyond the original $\boldsymbol{V}$, the discrete main variables previously discussed, we now allow for additional background covariate(s) $\boldsymbol{X}$, which are allowed to be continuous. We will suppose that within each value of $\boldsymbol{x}$ there exists a conditional estimand $\varphi_{\boldsymbol{x}}$, such as the conditional $\text{ATE}_{\boldsymbol{x}} = \mathbb{E}[Y(d=1) - Y(d=0)|\boldsymbol{x}]$. Our overall quantity of interest is $\varphi = \mathbb{E}_{\boldsymbol{X}}[\varphi_{\boldsymbol{X}}]$, such as the unconditional $\text{ATE} = \mathbb{E}[Y(d=1) - Y(d=0)] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[Y(d=1) - Y(d=0)|\boldsymbol{X}]]$.

We will conduct inference by reasoning about the covariate-conditional distributions over the main variables, or equivalently, the parameters of these conditional distributions $\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}$. Here, each $\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}$ is a parameter vector representing the categorical proportions of $\boldsymbol{V}$ conditional on the covariates taking on values $\boldsymbol{x}$. At a high level, the proposed method proceeds as follows. We first estimate the conditional category proportions $\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}$ using a multinomial model for $P(\boldsymbol{V} \mid \boldsymbol{X} = \boldsymbol{x})$. We then use those fitted conditional laws to form a manageable number of covariate strata, compute lower and upper bounds within each stratum, aggregate those stratum-specific bounds using the empirical stratum frequencies, and quantify sampling uncertainty with the same recentered subsampling logic developed in Appendix C.2.1. In this sense, the multinomial model serves as a smoothing device that induces a matching-like subclassification on units with similar fitted conditional observed-data laws. This subclassification step is essential computationally: solving a separate `autobounds` problem for every observed row would be infeasible in large datasets and would make repeated subsampling prohibitively expensive.

When the number of unique $\boldsymbol{x}$ values is small, the method of Appendix C.2.1 can be applied without modification within each level of the covariates. In this case, covariate adjustment can be handled nonparametrically. To accommodate scenarios where $\boldsymbol{X}$ is continuous or high-dimensional, however, we develop a model-based extension that allows researchers to approximately adjust for covariates by estimating the $\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}$ through multinomial logistic re-

gression. This places the following parametric structure on the main-variable proportions:

$$\theta_{\boldsymbol{v}_k|\boldsymbol{x}} = P_{\boldsymbol{V}\boldsymbol{X}}(\boldsymbol{V} = \boldsymbol{v}_k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_k^\top \phi(\boldsymbol{x}))}{\sum_{k'=1}^{K} \exp(\boldsymbol{\beta}_{k'}^\top \phi(\boldsymbol{x}))} = \text{softmax}(\boldsymbol{\beta}^\top \phi(\boldsymbol{x}))_k \tag{15}$$

where $\boldsymbol{v}_k$ is one possible combination of observed main-variable values, such as $D = 0$ and $Y = 0$ in the confounding example, and $\phi(\boldsymbol{x})$ is a basis expansion function that maps the covariates $\boldsymbol{x}$ to a set of features that may include indicator variables, nonlinear transformations, and interactions. Here, $\boldsymbol{\beta}_k$ is a vector of regression parameters indicating how the covariates $\boldsymbol{x}$ (and their expanded basis functions) translate into greater or lesser probabilities of observing a specific combination of main variables $\boldsymbol{v}_k$, and the summation in the denominator ensures that the $\boldsymbol{x}$-conditional category proportions sum to unity. We will use $\boldsymbol{\beta} \coloneqq [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$ to denote the joined vectors of regression coefficients. Rather than placing a posterior on these parameters, our implementation treats $\hat{\boldsymbol{\beta}}_n$ as an estimated nuisance object and propagates its uncertainty through recentered subsampling. To keep the procedure computationally feasible, we do not solve a separate `autobounds` problem for every observed $\boldsymbol{x}_i$. Instead, after obtaining fitted conditional laws, we coarsen them into a finite collection of strata. Concretely, let $s(\hat{\boldsymbol{\theta}}_{\boldsymbol{V}|\boldsymbol{x}})$ denote a one-dimensional summary score of the fitted conditional law; in practice, this can be chosen as a fitted probability or other low-dimensional summary of substantive interest. We then partition the sample into $G$ bins using empirical quantiles of $s(\hat{\boldsymbol{\theta}}_{\boldsymbol{V}|\boldsymbol{x}})$, and within each bin $g$ form the average fitted conditional law

$$\bar{\boldsymbol{\theta}}_{g,n} \coloneqq \frac{1}{n_g} \sum_{i:\, g_n(\boldsymbol{x}_i)=g} \hat{\boldsymbol{\theta}}_{\boldsymbol{V}|\boldsymbol{x}_i,n},$$

where $g_n(\boldsymbol{x}_i) \in \{1, \ldots, G\}$ is the bin assignment and $n_g$ is the number of observations in bin $g$. Recall that our target of inference is bounds on the marginal estimand $\varphi = \mathbb{E}_{\boldsymbol{X}}[\varphi_{\boldsymbol{X}}]$. On the full sample, we estimate $\hat{\boldsymbol{\beta}}_n$, map it through (15) to obtain fitted conditional proportions $\hat{\boldsymbol{\theta}}_{\boldsymbol{V}|\boldsymbol{x}_i,n}$ at each observed covariate value, use these fitted laws to construct the $G$ bins described

above, and then compute the binned covariate-adjusted bound estimators

$$\hat{\underline{\varphi}}_n = \sum_{g=1}^{G} \frac{n_g}{n} \underline{\mathsf{A}}_{\varphi}(\bar{\boldsymbol{\theta}}_{g,n}), \qquad \hat{\overline{\varphi}}_n = \sum_{g=1}^{G} \frac{n_g}{n} \overline{\mathsf{A}}_{\varphi}(\bar{\boldsymbol{\theta}}_{g,n}).$$

To quantify uncertainty, let $m = \lfloor n^{\gamma} \rfloor$ with $\gamma \in (0,1)$ and draw many subsamples of size $m$ without replacement. For each subsample $j$, we refit the multinomial logit on the subsample, recompute the fitted conditional proportions for observations in that subsample, rebuild the quantile bins from the subsample-specific score distribution, recompute the within-bin average fitted laws and the corresponding aggregated bounds, and obtain subsample analogues $\hat{\underline{\varphi}}_m^{*(j)}$ and $\hat{\overline{\varphi}}_m^{*(j)}$. We then form the recentered statistics

$$T_{m,\ell}^{*(j)} = \sqrt{m}\left(\hat{\underline{\varphi}}_m^{*(j)} - \hat{\underline{\varphi}}_n\right), \qquad T_{m,u}^{*(j)} = \sqrt{m}\left(\hat{\overline{\varphi}}_m^{*(j)} - \hat{\overline{\varphi}}_n\right),$$

and use their empirical quantiles to construct confidence bounds exactly as in the no-covariate case. That is, if $q_{\ell,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of $\{T_{m,\ell}^{*(j)}\}$ and $q_{u,\alpha/2}$ is the $\alpha/2$ quantile of $\{T_{m,u}^{*(j)}\}$, then the resulting interval is

$$\left[\hat{\underline{\varphi}}_n - \frac{q_{\ell,1-\alpha/2}}{\sqrt{n}},\ \hat{\overline{\varphi}}_n - \frac{q_{u,\alpha/2}}{\sqrt{n}}\right].$$

This procedure treats the covariate model as a smoothing device for estimating the conditional observed-data law, while the finite binning step converts the fitted laws into a tractable number of subclassification cells on which `autobounds` can be run. Final uncertainty quantification is driven by subsampling rather than by a parametric posterior approximation. We summarize the implementation in Algorithm 1.

Two clarifications are worth making. First, this is a model-assisted procedure: the target remains the population-average estimand $\varphi = \mathbb{E}_{\boldsymbol{X}}[\varphi_{\boldsymbol{X}}]$, but in finite samples we approximate the outer expectation by averaging over a finite partition induced by the fitted conditional laws, both on the full sample and within each subsample. Second, the first-stage multinomial fit creates a generated-regressor problem, so the role of subsampling is to propagate uncertainty

from the entire estimator, including estimation of $\hat{\boldsymbol{\beta}}_n$, formation of the bins, and the final `autobounds` step. We do not claim a general theorem here covering arbitrary misspecification or severe sparsity in the covariate distribution; rather, we view the procedure as a practical default for settings in which the multinomial model provides a stable approximation to the conditional observed-data law and the induced bins contain enough observations to support reliable optimization. For the same reason, we prefer subsampling to the ordinary bootstrap: even though the nuisance model is smooth, the final bound estimator remains a nonsmooth extremum-type functional, so the bootstrap may inherit the same nonregularity concerns discussed above.

**Algorithm 1** Covariate Adjustment via Recentered Subsampling

**Input:** Estimand $\varphi$, Data $\{(\boldsymbol{V}_i, \boldsymbol{X}_i)\}_{i=1}^n$, Subsample size $m$, Subsample count $M$, Confidence level $\alpha$

**Output:** Partial identification interval estimate $[\underline{\varphi}, \overline{\varphi}]$ with nominal level $(1-\alpha)\times 100\%$

Estimate $\hat{\boldsymbol{\beta}}_n$ on the full sample
**for** $i = 1$ **to** $n$ **do**
  Compute fitted probabilities and a scalar score for unit $i$
**end**
Partition $\{\hat{s}_{i,n}\}_{i=1}^n$ into $G$ bins using empirical quantiles
**for** $g = 1$ **to** $G$ **do**
  Compute the bin-level average fitted law
  Run `autobounds` to obtain lower and upper bounds for bin $g$
**end**
Aggregate the bin-specific bounds using bin frequencies
**for** $j = 1$ **to** $M$ **do**
  Draw a subsample of size $m$ without replacement
  Estimate $\hat{\boldsymbol{\beta}}_m^{*(j)}$ on that subsample
  **for** *each observation* $i$ *in subsample* $j$ **do**
    Compute fitted probabilities and a scalar score for unit $i$
  **end**
  Partition $\{\hat{s}_{i,m}^{*(j)}\}$ into $G$ bins using subsample empirical quantiles
  **for** $g = 1$ **to** $G$ **do**
    Compute the bin-level average fitted law
    Run `autobounds` to obtain lower and upper bounds for bin $g$
  **end**
  Aggregate the bin-specific bounds using subsample bin frequencies
  Form the recentered lower and upper statistics
**end**
Compute $q_{\ell,1-\alpha/2}$ and $q_{u,\alpha/2}$ from the empirical distributions of $\{T_{m,\ell}^{*(j)}\}$ and $\{T_{m,u}^{*(j)}\}$
Return the recentered confidence interval

We offer three remarks on this procedure. First, as in the no-covariate case, the choice

of $m$ trades off fidelity to the full-sample problem against the additional noise of very small subsamples; our implementation therefore uses the default rule $m = \lfloor n^{2/3} \rfloor$. Second, the resulting interval inherits robustness to the nonsmoothness of the bound estimators from the subsampling step, while the multinomial model enters only through estimation of the conditional observed-data law and the induced subclassification. Third, results are still determined by the causal assumptions supplied to `autobounds` (see Section 2.6) together with the quality of the working model for $P(\boldsymbol{V} \mid \boldsymbol{X})$, the choice of score $s(\cdot)$, and the number of bins $G$, so this adjustment should be interpreted as an approximate bias-correction device rather than a fully nonparametric procedure.

This approach provides a practical way to adjust for continuous or high-dimensional covariates while retaining the same inferential logic used elsewhere in the paper. A Bayesian or quasi-Bayesian treatment of the covariate model may also be workable, but we do not pursue it here.

### C.4.1 Validity of Covariate-averaged Bounds.

We now provide formal reasoning for the validity of the averaging of the bounds presented in the previous section. We will first demonstrate the validity and sharpness of the local bounds, i.e. at each covariate level $\boldsymbol{x}$. We then present a proof of validity for the bounds upon averaging.

To begin, we introduce some new notation. Let $\mathcal{G}$ be a directed acyclic graph over the variables $\boldsymbol{W} = \boldsymbol{V} \cup \boldsymbol{X} \cup \boldsymbol{U}$ where $\boldsymbol{V}$ are observed discrete variables, $\boldsymbol{U}$ are unobserved, and $\boldsymbol{X}$ are measured covariates which may be discrete or continuous. Let $P_{\boldsymbol{W}}$ be a distribution over $\boldsymbol{W}$ compatible[36] with $\mathcal{G}$, and define any distribution $P_{\mathbf{S}}$, $\mathbf{S} \subset \boldsymbol{W}$, as a margin of this $P_{\boldsymbol{W}}$, that is, $P_{\mathbf{S}} = \int P_{\boldsymbol{W}} \, \mathrm{d}P_{\boldsymbol{W} \setminus \mathbf{S}}$. Also, write $P_{\mathbf{S}|v}$ to be the analogous conditional margin of $P_{\boldsymbol{W}|v}$, where $P_{\boldsymbol{W}|v}$ is the conditional distribution of $\boldsymbol{W} \setminus \{V\}$ on the event $V = v$. We remark that $P_{\boldsymbol{V}(\mathcal{W})}$ as defined earlier is equal to $P_{\boldsymbol{W}}$ under the NPSEM (Pearl, 2009) we assume here. We have chosen to adjust notation slightly for the ensuing exposition as emphasis on the unobserved variables $\boldsymbol{U}$ will become important.

[36] *Compatibility* means that $P_{\boldsymbol{W}}$ obeys the Markov factorization over $\boldsymbol{W}$ implied by $\mathcal{G}$.

Next we describe the kinds of questions a user of `autobounds` might ask. We begin at the most granular level, with a *causal query*.

**Definition 1** (Causal Query)**.** *A causal query is a functional $\psi^{\mathbf{t}}(P_{\boldsymbol{W}})$ such that*

$$\psi^{\mathbf{t}}(P_{\boldsymbol{W}}) \coloneqq P_{\boldsymbol{W}}\left(\bigcap_{i=1}^{L} E_i(\mathbf{t}_i)\right)$$

*where $E_i(\mathbf{t}_i)$ is a counterfactual statement over variables in $\boldsymbol{V} \cup \boldsymbol{X}$, under a series of fixed interventions $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_L\}$ on elements of $\boldsymbol{V}$.*

Revisiting the running example, where $\boldsymbol{V} = [D, Y]$, $\psi^{\mathbf{t}}(P_{\mathbf{W}}) = P_{\mathbf{W}}(Y(d = 1) = 1)$ has $\mathbf{t}_1 = (1)$ with $E_1(\mathbf{t}_1) = \{Y(d = 1) = 1\}$.

**Definition 2** (Causal Estimand)**.** *A causal estimand is a functional*

$$\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}}) = g(\psi^{\mathbf{t}}(P_{\boldsymbol{W}}), \psi^{\mathbf{t}'}(P_{\boldsymbol{W}}))$$

*where $g$ is a measureable function.*

In the simple counfounding case the ATE

$$P_{\boldsymbol{W}}(Y(d = 1) = 1) - P_{\boldsymbol{W}}(Y(d = 0) = 1) \tag{16}$$

is a causal estimand with $g(a, b) = a - b$ and the specifications of $\psi^{\mathbf{t}}, \psi^{\mathbf{t}'}$ are readily apparent.

The following definition characterises all estimands for which our sharpness result will hold.

**Definition 3** (Causal Collapsibility; (Huitfeldt et al., 2019))**.** *We say that $\varphi$ is causally collapsible with respect to $P_X$ if*

$$\mathbb{E}_{P_X}[\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|X})] = \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}}).$$

In words, Definition 3 calls any causal estimand *causally collapsible with respect to $P_X$* if we can recover the marginal estimand by averaging the conditional estimand quantity over the

covariate distribution. With respect to this definition, examples of causal estimands which are collapsible with respect to $P_X$ include the ATE and the CDE; a non-example is the ATT, since the weights needed to average the causal estimand do not in general equal $P_X$. The notion of collapsibility can be easily generalised to weights which differ from $P_X$, however, such estimands are beyond the scope of the result we present below.

We are now ready to state the two main results in this section. The first describes the sharpness of the covariate-conditional bounds, while the second shows how these sharp conditional bounds can be aggregated into sharp marginal bounds under exact finite stratification. Adapting notation from previously, write $\underline{\mathsf{A}}_{\varphi}$, and $\overline{\mathsf{A}}_{\varphi}$, where for any evidence $P_{\boldsymbol{V}}$, $\underline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})$ and $\overline{\mathsf{A}}_{\varphi}(P_{\boldsymbol{V}})$ are the lower bound and upper bounds on estimand $\varphi$ under the `autobounds` procedure; in the ensuing exposition we focus only on the lower bound as proofs for the upper bound are analogous.

**Proposition 2.** *Let $\mathcal{G}$ be a canonical DAG over variables $\boldsymbol{W} = \boldsymbol{V} \cup \boldsymbol{X} \cup \boldsymbol{U}$, as in the preamble. Suppose there exists a variable $X \in \boldsymbol{X}$ such that*

1. *$X$ has no parents except its exogeneous disturbance, $U_X$;*
2. *$X$ is a parent of all variables $V \in \boldsymbol{V} \setminus \{X\}$;*
3. *$U_X$ has no children other than $X$.*

*Then, the quantity $\underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x})$ is a **sharp lower bound** for $\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|x})$.*

*Proof.* For any $x \in \mathcal{S}(X)$ fixed, consider the DAG $\mathcal{G}(x)$ associated with $\mathcal{G}$, where the fixed node $x$ is removed. Now $\mathcal{G}(x)$ is a canonical DAG over the variables $\boldsymbol{W}(x) \coloneqq \{V(x) : V \in \boldsymbol{V}\} \cup \boldsymbol{U} \cup \boldsymbol{X}$. Write $\boldsymbol{V}(x) \coloneqq \{V(x) : V \in \boldsymbol{V}\}$. Take any causal estimand $\varphi^{(\mathbf{t},x),(\mathbf{t}',x)}(P_{\boldsymbol{W}})$ comparing queries under interventions $(\mathbf{t}, x)$ and $(\mathbf{t}', x)$ where $\mathbf{t}, \mathbf{t}'$ do not fix $x$. By Theorem 1 of Duarte et al. (2024), it follows that $\underline{\mathsf{A}}_{\varphi^{(\mathbf{t},x),(\mathbf{t}',x)}}(P_{\boldsymbol{V}(x)})$ is a sharp lower bound on $\varphi^{(\mathbf{t},x),(\mathbf{t}',x)}(P_{\boldsymbol{W}(x)})$. Specifically, by Proposition 1 and Proposition 3 of Duarte et al. (2024) we have that $\underline{\mathsf{A}}_{\varphi^{(\mathbf{t},x),(\mathbf{t}',x)}}(P_{\boldsymbol{V}(x)})$ solves

$$\underset{P_{\boldsymbol{U}\setminus\{U_X\}}}{\arg\min}\, h(P_{\boldsymbol{U}\setminus\{U_X\}}) \quad \text{subject to } h(P_{\boldsymbol{U}\setminus\{U_X\}}) \in \mathcal{F}(P_{\boldsymbol{V}(x)}) \tag{P}$$

where $h(P_{\boldsymbol{U}\setminus\{U_X\}})$ is a polynomial in the probabilities $\Pr(U = u)$ for each $U \in \boldsymbol{U} \setminus \{U_X\}$, such that $h(P_{\boldsymbol{U}\setminus\{U_X\}}) = \varphi^{(\mathbf{t},x),(\mathbf{t}',x)}$ and $\mathcal{F}(P_{\boldsymbol{V}(x)})$ is the feasible set defined by the axiomatic and evidential constraints defined by $P_{\boldsymbol{V}(x)}$, see Appendix A. Note that Proposition 3 of Duarte et al. (2024) is used to exclude parameters $\Pr(U_X = u_X)$ from the program.

Now, notice that problem (P) is in terms of counterfactual distributions and therefore appears a-priori incomputable. However, by the assumptions of Proposition 2, it follows that

$$P_{\boldsymbol{V}(x)} = P_{\boldsymbol{V}|x}, \quad \varphi^{(\mathbf{t},x),(\mathbf{t}',x)}(P_{\boldsymbol{W}(x)}) = \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|x})$$

where the both equalities are due to the fact that $\{V(x) : V \in \boldsymbol{V} \setminus \{X\}\} \perp\!\!\!\perp X$ by the assumptions of this proposition and a standard application of consistency. Thus, we can cast problem (P) into an equivalent problem whereby each evidential constraint $\Pr(\boldsymbol{V}(x) = \mathbf{v})$ is replaced by $\Pr(\boldsymbol{V} = \mathbf{v} \mid X = x)$ and we retain the same objective function. Mathematically, this implies

$$\underline{\mathsf{A}}_{\varphi^{(\mathbf{t},x),(\mathbf{t}',x)}}(P_{\boldsymbol{V}(x)}) = \underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x})$$

which proves the result. □

In practice, Proposition 2 permits the researcher using `autobounds` to ignore $X$ when instantiating the DAG in their Python program, and simply provide the conditional distribution $\Pr(\boldsymbol{V} \setminus \{X\} = \mathbf{v} \mid X = x)$ as data. As the construction in the proof shows, one can do this because $X$ is disconnected from relevant nodes in $\mathcal{G}(x)$.

We are now ready to state the final result concerning the marginal bounds obtained by averaging over covariates.

**Proposition 3.** *Suppose there exists an $X$ satisfying the conditions of Proposition 2, and suppose further that $X$ has finite support $\mathcal{S}(X) = \{x_1, \ldots, x_K\}$. Let $\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}})$ be a causally collapsible estimand with respect to $P_X$. Then, the bounds*

$$l = \int \underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x})\ \mathrm{d}P_X, \quad u = \int \overline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x})\ \mathrm{d}P_X$$

*are sharp bounds for* $\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}})$.

*Proof.* We first show validity. By Proposition 2, for every $x$,

$$\underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x}) \leq \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|x}) \leq \overline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x}).$$

As the integral operator with respect to $P_X$ is monotone and $\varphi$ is causally collapsible with respect to $P_X$,

$$l \leq \int \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|x}) \, \mathrm{d}P_X = \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}}) \leq u.$$

Thus the averaged bounds are valid.

It remains to show attainability. We prove sharpness of the lower bound; the argument for the upper bound is analogous. Because $\mathcal{S}(X)$ is finite, it suffices to build a single feasible global model by combining one bound-attaining conditional optimizer for each stratum $x_k$. By Proposition 2, for each fixed $x_k$ there exists a conditional structural completion compatible with $\mathcal{G}$ and the observed law $P_{\boldsymbol{V}|x_k}$ whose conditional estimand attains the sharp lower bound

$$\varphi^{\mathbf{t},\mathbf{t}'}(P^*_{\boldsymbol{W}|x_k}) = \underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x_k}).$$

Write $P^*_{\boldsymbol{U}\setminus\{U_X\}|x_k}$ for one such bound-attaining latent law, and let $f^{(k)}$ denote the corresponding structural equations for the observed variables in stratum $x_k$.

We now paste these finitely many conditional optimizers into a single global model. Let $U_X$ generate $X$ according to $P_X$, as in the statement of the proposition. For each support point $x_k$, attach a latent block $\tilde{U}_k$ with law $P^*_{\boldsymbol{U}\setminus\{U_X\}|x_k}$, and take the finite collection $(\tilde{U}_1, \ldots, \tilde{U}_K, U_X)$ to be jointly independent. Define the structural equations for the observed variables so that when $X = x_k$, the variables in $\boldsymbol{V}$ are generated using the response functions $f^{(k)}$ associated with the optimizer $P^*_{\boldsymbol{U}\setminus\{U_X\}|x_k}$. Because $X$ is a parent of every variable in $\boldsymbol{V} \setminus \{X\}$, these piecewise structural equations are admissible in the canonical DAG. Because $X$ has no non-exogenous parents and $U_X$ has no children other than $X$, selecting the block $\tilde{U}_k$ when $X = x_k$ introduces no cross-stratum restriction beyond the shared marginal law $P_X$.

By construction, the resulting global model reproduces the observed law $P_{\boldsymbol{V}|x_k}$ at each

support point and attains the conditional lower bound in every stratum. Therefore,

$$\varphi^{\mathbf{t},\mathbf{t}'}(P^*_{\boldsymbol{W}}) = \sum_{k=1}^{K} \varphi^{\mathbf{t},\mathbf{t}'}(P^*_{\boldsymbol{W}|x_k}) \Pr(X = x_k) = \sum_{k=1}^{K} \underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x_k}) \Pr(X = x_k) = l,$$

where the first equality uses causal collapsibility for finite-support $X$. Hence the lower bound $l$ is attained by a feasible global model, so it is sharp. The upper bound is handled symmetrically. □

Proposition 3 shows that when conditioning on $X = x$ is exact and the covariate has finite support, averaging the covariate-specific sharp bounds produces the sharp marginal bounds for the collapsible estimand. This result applies to exact covariate stratification. Extending the same argument to more general support conditions would require additional measurable-selection and pasting arguments, which we do not pursue here. The model-assisted procedure used in the main text for continuous or high-dimensional covariates adds an additional approximation step by replacing exact conditioning with fitted laws and score-based bins; that approximation is computationally useful, but its sharpness is not implied by Proposition 3.

# D Model specification for Kocher et al. (2011)

We use insurgent control measured in July 1969 as the sole instrument (a 5-level ordinal variable ranging from full government control, $Z = 1$, to full insurgent control, $Z = 5$). The IV analyses in Kocher et al. (2011) also utilize insurgent control in August as a secondary instrument. We omit the second instrument for simplicity as replications show it does not meaningfully change results. Using insurgent control in July only is consistent with many other analyses presented by Kocher et al. (2011)—including all results in Tables 1–3 and 7 of that work—which similarly do not consider insurgent control in August. Following numerous analyses in the paper, we operationalize $D$ as a binary measure of whether any bombs were dropped within a two-kilometer radius of the hamlet.[37] Consistent with the original work,

[37] See e.g. Table 1, Model 5C, and Tables 6–7; however, note that Kocher et al. (2011) also present results that use the count of bombs.

our outcome $Y$ is a 5-level ordinal variable representing insurgent control of a hamlet in December 1969. Finally, Kocher et al. (2011) adjust for a number of additional covariates, $X$, including hamlet development level, population, distance to an international border, and terrain roughness. In general, we emphasize that Kocher et al. (2011) present a variety of specifications; we conduct an analysis that is consistent with their overall approach, but we cannot rule out the possibility that our findings are due to these implementation differences.

We cut all control variables at the median or terciles, as appropriate, to obtain the coarsened versions $\tilde{X}$[38]. Throughout our analysis, hamlets are only compared if they exactly match on all coarsened $\tilde{X}$ values.

# E Code

## E.1 Instrumental Variables

```
# DAG for this problem is the IV graph
vietnam_problem = causalProblem(
    DAG("Z -> D, D -> Y, U -> D, U -> Y", unob="U"),
    number_values={"Z" : 5, "D" : 2, "Y" : 2}
)
# read in data with Z/D/Y columns + other covariates
vietnam_data = pandas.read_csv("vietnam_data.csv")
# create binarized outcome by cutting at threshold
# and converting resulting booleans to 0/1 integers
thresh = 3  # repeated for every threshold in 2-5
vietnam_data = vietnam_data.assign(
    Y = (vietnam_data.Y_raw >= thresh).astype(int)
)
covariates = ['control_sep', 'development', 'log_dist_to_border',
                    'terrain_roughness', 'log_population']
# all statements below are w.r.t. this problem
with respect_to(vietnam_problem):
    read_data(
        vietnam_data, covariates = covariates, inference = True)
    # look for min/max ATE values (estimated bounds),
    # searching over all DGPs that are consistent
    # with assumptions & obs data
    vietnam_bounds = solve()
    # result: infeasible (can't find any such DGPs)
```

Figure 9: Code for the reanalysis of Kocher et al. (2011) with instrumental variables.

[38]Hamlet development level, population, and distance to an international border are binned into terciles. Terrain roughness is dichotomized by cutting at the median, as more than half of all hamlets have the lowest possible terrain roughness score of zero.

```
# load data with Z/D/Y columns, one row per voting-age adult
gotv_data = pandas.read_csv("gotv_data.csv")

# baseline IV graph
model = DAG("Z -> D, D -> Y, U -> D, U -> Y", unob="U")

def decode_extreme_point(values):
    rows = []
    for name, mass in values.items():
        if name == "objvar" or abs(mass) < 1e-10:
            continue
        d_type, y_type = name.split(".")
        d_bits = d_type.replace("D", "")
        y_bits = y_type.replace("Y", "")
        y0, y1 = int(y_bits[0]), int(y_bits[1])
        rows.append({
            "latent_type": name,
            "mass": mass,
            "treatment_stratum": {
                "00": "never-taker", "01": "complier",
                "10": "defier", "11": "always-taker"
            }[d_bits],
            "outcome_stratum": {
                "00": "never-voter", "01": "helped by treatment",
                "10": "harmed by treatment", "11": "always-voter"
            }[y_bits],
            "individual_ate": y1 - y0
        })
    return pd.DataFrame(rows).sort_values("mass", ascending=False)

problem_ate = causalProblem(model)
with respect_to(problem_ate):
    read_data(gotv_data)
    set_ate(ind="D", dep="Y")
    bounds_ate = solve(return_dgps=True)

print(bounds_ate["point lb dual"], bounds_ate["point ub dual"])
for label in ["lower", "upper"]:
    extreme_df = decode_extreme_point(bounds_ate["dgps"][label]["values"])
    print(extreme_df[[
        "latent_type", "mass", "treatment_stratum",
        "outcome_stratum", "individual_ate"
    ]].round({"mass": 4}))
```

Figure 10: Code for the instrumental variable analysis of the Get-out-the-vote experiment, including recovery of the extremal DGPs that attain the ATE bounds.

## E.2 Selection Bias

```
police_problem = causalProblem(
    DAG("D -> M, D -> Y, M -> Y, U -> M, U -> Y", unob="U")
)

# load data with D/Y columns, one row per police-civ encounter
# (only recorded if stop is made, so M=1 for all rows)
police_data = pandas.read_csv("police_data.csv")

# all statements below are w.r.t. this problem
with respect_to(police_problem):

    # give data to autobounds and
    read_data(police_data, cond="M=1")

    # assume no force would be used if stop was not made
    force_used_if_stop_not_made = p("Y(M=0)=1")
    add_assumption(force_used_if_stop_not_made, "==", 0)

    # assume no anti-white bias in stopping (encounters where
    # (white civs would be stopped but minority civs would not)
    anti_white_stop = p("M(D=0)=1 & M(D=1)=0")
    add_assumption(anti_white_stop, "==", 0.0)

    # assume that potential force is lower on average
    # in "racial stops" (discretionary, stop made iff civ is minority)
    # than in "always stops" (mandatory, stop made for any civ race)
    racial_stop = "M(D=0)=0 & M(D=1)=1"  # M(d)=1 if and only if d=1
    always_stop = "M(D=0)=1 & M(D=1)=1"  # M(d)=1 for all races d
    # state assumption when white civs are placed in these scenarios
    avg_force_if_white_among_racial_stops = E("Y(D=0, M=1)", cond=racial_stop)
    avg_force_if_white_among_always_stops = E("Y(D=0, M=1)", cond=always_stop)
    add_assumption(
        average_force_if_white_among_racial_stops,
        "<=",
        average_force_if_white_among_always_stops
    )
    # state assumption when minority civs are placed in these scenarios
    avg_force_if_minority_among_racial_stops = E("Y(D=1, M=1)", cond=racial_stop)
    avg_force_if_minority_among_always_stops = E("Y(D=1, M=1)", cond=always_stop)
    add_assumption(
        average_force_if_minority_among_racial_stops,
        "<=",
        average_force_if_minority_among_always_stops
    )

    # calculations based on Gelman, Fagan, Kiss (2007) data implies
    # that among stops of black civs, 32% are "racial stops" that
    # would not have occurred if white civs were placed in same
    # scenarios (remaining 68% are "always stops")
    p_racial_stops_among_all_minority_stops = p("M(D=0)=0", cond = "D=1 & M=1")
    add_assumption(p_racial_stops_among_all_minority_stops, "==", 0.32)

    # set estimand to be ATE conditional on the subset
    # of police-civ encounters where a stop was made
    set_ate("D", "Y", cond="M=1")

    # calculate bounds
    police_bounds = solve(ci=True, nsamples=1000)
```

Figure 11: Code for the study of bias in police use of force.