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

August 20, 2025

Word Count: 16,845

Abstract

Social science has developed an expansive design-based toolkit for causal inference, but the assumptions that undergird standard approaches often fail in applied settings. In response, researchers often present unreliable results, narrow their questions post-hoc, or abandon projects altogether. We demonstrate an alternative approach—automated partial identification, i.e. best- and worst-case reasoning—that allows researchers to learn as much as possible in these imperfect settings, while transparently acknowledging limitations of data and design. Using the `autobounds` framework, analysts declare an estimand, state assumptions, and supply discrete data. The program then returns sharp bounds on the estimand, or a point-identified solution if one exists. By replicating five published studies across empirical political science subfields, we show how this approach allows researchers to accommodate threats to standard assumptions about instrumental variables, difference in differences, sample selection, and mediation in a flexible and principled manner.

Keywords: causal inference, partial identification, assumption testing, constrained optimization

*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. The statements made and views expressed are solely the responsibility of the authors. Authors listed in alphabetical order.

†Corresponding author.

Contents

1 Introduction **1**

2 Partial Identification with `autobounds` **5**
2.1 Preliminaries . . . 5
2.2 A Simple Example with Classic Confounding . . . 9

3 Applications **14**
3.1 Instrumental Variables . . . 15
3.1.1 Replication and Extension of Coppock and Green (2016) . . . 17
3.1.2 Replication and Extension of Kocher et al. (2011) . . . 20
3.2 Difference in Differences . . . 24
3.2.1 Replication and Extension of Schubiger (2021) . . . 25
3.3 Selection Bias . . . 28
3.3.1 Replication and Extension of Knox et al. (2020) . . . 31
3.4 Mediation Analysis . . . 33
3.4.1 Replication and Extension of Acharya et al. (2018) . . . 38

4 Discussion and Conclusion **41**

A Detailed Example with `autobounds` **51**
A.1 Simulated Data for Section 2.2 . . . 59

B Additional Sensitivity Analyses for Instrumental Variables **61**

C Practical Considerations: Statistical Uncertainty and Covariate Adjustment **63**
C.1 Preliminaries . . . 63
C.2 Statistical Uncertainty . . . 64
C.2.1 Uncertainty Quantification without Covariates . . . 65
C.2.2 Alternative Approaches to Uncertainty Quantification . . . 68
C.3 Covariate Adjustment . . . 68
C.3.1 Validity of Covariate-averaged Bounds. . . . 72

D Model specification for Kocher et al. (2011) **76**

E Problem Formulations **77**
E.1 Instrumental Variables . . . 77
E.2 Difference in Differences . . . 86
E.2.1 Problem Formulation . . . 86
E.3 Selection Bias . . . 91
E.3.1 Problem Formulation . . . 91
E.4 Mediation . . . 96
E.4.1 Formal Background . . . 96
E.4.2 On the Manipulation Exclusion . . . 98
E.4.3 Eliminated Effects . . . 99
E.4.4 Problem Formulation . . . 102
E.4.5 Inference in the Parallel Design . . . 105

1 Introduction

Social scientists now possess an expansive set of tools for causal inference, greatly improving researchers' ability to draw credible insights from observational data. But as all researchers know, applied work is messy: real-world scenarios often diverge from the ideal conditions under which standard causal-inference approaches return reliable estimates of causal quantities. For example, (i) instrumental variables approaches rely on the well-known "exclusion restriction" (Angrist et al., 1996), meaning no direct causal path exists between the instrument and outcome; (ii) difference-in-differences designs rely on a "parallel trends" assumption (Abadie et al., 2010; Xu, 2017) that in the absence of treatment, treatment and control groups would move in tandem over time; (iii) analyses of selected data (Heckman, 1979) rely on the assumption that selection is not driven by unobserved factors that affect outcomes (Elwert and Winship, 2014); and (iv) mediation designs rely on the "sequential ignorability" assumption (Imai et al., 2011) that unobserved factors do not affect any combination of treatment, mediator, and outcome. These assumptions, discussed in more detail below, are often difficult to defend outside the tidy examples that appear in statistics textbooks. Applied researchers have limited options when confronting thorny, real-world scenarios where there is reason to doubt key assumptions. Typically, researchers either ignore the problem and present unreliable results, narrow their focus to questions of lesser importance, or abandon projects altogether.

In this paper, we argue that an alternative approach, partial identification, represents a far more fruitful path for applied work when ideal conditions are unachievable. Rather than ignoring violations of key assumptions to make seemingly precise but unreliable claims, the goal of partial identification is to explicitly account for these violations and seek to *bound* the quantity of interest—that is, to report best- and worst-case values for the causal question, given the imperfect information at hand. In other words, partial identification acknowledges the difficulties of empirical research and asks, "Given all available information, what is the most that can be learned *in spite of these obstacles*?"

The concept of partial identification is not new: decades of research in statistics and econometrics (Robins, 1989; Manski, 1990a; Heckman and Vytlacil, 2001; Zhang and Rubin, 2003; Cai et al., 2008; Swanson et al., 2018; Gabriel et al., 2020; Molinari, 2020) have developed and refined techniques for bounding causal quantities in particular scenarios (Lee, 2009; Gabriel et al., 2020; Kennedy et al., 2019; Knox et al., 2020; Li and Pearl, 2021; Sjölander et al., 2014). However, depending on the problem, the process of deriving sharp bounds—the narrowest possible range of answers to a causal question—can be at best tedious, potentially involving dozens of pages of algebra, and at worst entirely intractable. Scholars have at various points attempted to automate this process in specific settings (Balke and Pearl, 1994, 1997) or derive bounds for several variants of a problem (Swanson et al., 2018; Gabriel et al., 2020). A general and computationally feasible solution was only recently presented in Duarte et al. (2023). With this algorithm, `autobounds`, users declare a causal quantity of interest (i.e. an estimand, such as an average treatment effect), state assumptions, and provide discrete data, i.e. data in which all variables take a finite and countable number of values—however incomplete or mismeasured. Using an efficient "branch-and-bound" optimization technique (Vigerske and Gleixner, 2018; Gamrath et al., 2020; Belotti et al., 2009), the algorithm then efficiently searches over possible data generating processes (DGPs) and locates ones which satisfy the constraints implied by the stated assumptions and observed data. When complete, it outputs either *sharp bounds* on the estimand or, if one exists, a point-identified solution. As Duarte et al. (2023) 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).

In short, this new tool allows social scientists to flexibly 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 selectively invoke only the assumptions that are plausible in their applied settings, acknowledging the design and data

limitations that they face. Researchers can use it to transparently narrow the range of possible answers—learning as much as possible about the research question—until more data is obtained or additional information comes to light that justifies new assumptions. This approach can do more than simply eliminate implausible assumptions; it can also partially relax assumptions, for example by stipulating that the treatment group may deviate from parallel trends by no more than some amount or that a monotonicity assumption may be violated in no more than some percentage of cases. This allows researchers to conduct sensitivity analyses on virtually any aspect of a study and to precisely characterize the empirical consequences of an assumption by quantifying the degree to which its relaxation leads to wider bounds, which in turn can help direct future research to areas where it can advance scientific progress the most. As we show below, this tool can also test every possible observable implication of a set of assumptions, in some cases flagging them as collectively *falsified* if observed data were inconsistent with the theorized DGP. In other words, this tool also automatically informs researchers when empirical implications of their theories are contradicted by observed information.

Perhaps most importantly, this approach allows for *question-driven* research. The flexibility of this tool means that the statistical quantity of interest—the research question, formally stated—need not be retroactively adjusted to suit a particular applied setting. 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. In other words, the goal of a study need not change to accommodate the method being used—a desirable feature as scholars seek to accumulate knowledge on research questions across independent studies.

To demonstrate the benefits of this approach, we replicate and extend published findings across empirical subfields of political science to show how standard assumptions can be relaxed in applied settings. These studies include the consequences of bombing campaigns in Vietnam

(Kocher et al., 2011); effectiveness of voter mobilization in the United States (Coppock and Green, 2016); determinants of counterinsurgency in the Peruvian Civil War (Schubiger, 2021); racial bias in policing (Knox et al., 2020); and the democratic peace theory (Tomz and Weeks, 2013; Acharya et al., 2018). These examples cover many classic approaches in the causal inference toolkit, including instrumental variables, difference-in-differences, and mediation. In some cases, we show that standard assumptions can be relaxed only slightly before the study becomes uninformative about the direction of the effect (i.e. bounds include zero). In other cases, we show that informative bounds can be recovered even after abandoning assumptions previously thought to be pivotal. Regardless of the precision of the bounds, we show in each case how automated partial identification allows researchers to confront difficult challenges in a principled manner, all while pursuing the same meaningful social question that motivated a study to begin with.

This paper proceeds as follows. We first review the fundamentals of partial identification in causal inference and discuss the mechanics of the `autobounds` algorithm at a high level to convey intuition and demonstrate its basic functions with a simulated example. Having established the basic approach, we then apply it to several common research designs and inferential obstacles, replicating and extending published work to show how the algorithm performs in practice. These exercises illustrate how `autobounds` allows analysts to relax or abandon key assumptions when their validity is suspect, investigate more meaningful estimands, and direct future research in efficient ways. We conclude with a discussion of the potential for automated partial identification to promote more credible, productive and transparent empirical social science.

2 Partial Identification with `autobounds`

2.1 Preliminaries

We first establish notation, review basic concepts in causal inference,[1] and demonstrate a simple partial identification problem using `autobounds`. To use our approach, the researcher must first specify a target quantity, or causal *estimand*: a comparison of average counterfactual quantities. These estimands, which precisely define the scientific question of interest, are distinct from 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 not specifically defined in applied social science research (Lundberg et al., 2021), this step is indispensable—without knowing the target quantity, it is impossible to tell whether a research 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 the actual treatment with D; for simplicity, we will consider binary treatments unless otherwise stated, though the proposed approach generalizes straightforwardly 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.

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

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). Once the graph is written down, it can be defended as a causal representation of a theory. Based on that structure, one can then determine 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, 1990b, 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* if they provably cannot be narrowed without additional assumptions or data, meaning that researchers have learned as much as possible from the available information. The advantage of sharp bounds is that analysts can simply choose not to invoke assumptions that are indefensible—producing a wider range of possible answers 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 thus offers a principled approach to characterizing causal effects in the presence of incomplete information or questionable research designs, deriving bounds and proving their sharpness can often be analytically intractable. To address this, Duarte et al. (2023) provides an easy-to-use algorithm, `autobounds`, to automatically compute sharp bounds given the specifics of a causal research question. For details on how this algorithm functions, we refer readers to Duarte et al. (2023) and the "Problem Formulation" sections in this paper's Appendix; here, we seek to convey high-level intuition for `autobounds`.

At the heart of this procedure is the concept of principal stratification (Frangakis and Rubin, 2002), which is the process of characterizing units in a study into their essential types based on how they would respond, counterfactually, when variables in the model take different values. 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 causal systems grow in complexity, the number of principal strata can grow 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.[2]

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.[3] The software then expresses the causal estimand in terms 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.[4] Next, it translates causal assumptions and observed data into constraints on the possible sizes of principal strata.[5] Duarte et al. (2023) 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, to obtain best- and worst-case scenarios for the quantity of interest.

Importantly, `autobounds` currently only works with analyses in which each variable is discrete. This is a limitation, but there are feasible workarounds in most applied cases. For

[2]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.

[3]The user can also specify two tolerance parameters that control the amount of computation time, corresponding 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. (2023).

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

[5]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.

example, continuous variables can be discretized (binned into a finite number of values, e.g. low, medium and high) for use with `autobounds`.[6] For many social science applications, this poses no serious problems, since theories are rarely detailed enough to credibly predict distinct outcomes for, e.g., voters who are 40 vs. 40 $\frac{1}{12}$ years old. And in practice, data is rarely truly continuous: in any real-world dataset, variables will always take on a countably finite number of values. However, if variables must remain many-valued for an application, it will make computation of the `autobounds` algorithm more difficult. We are developing new techniques, discussed in the Section 4, that will allow for the incorporation of continuous variables under certain modeling assumptions.

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.

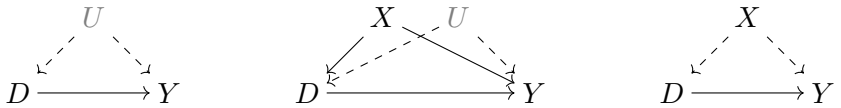
2.2 A Simple Example with Classic Confounding

To demonstrate the algorithm, we consider a simulated data set in which the analyst is interested in estimating the effect of obtaining a college education, D, on whether one turns out to vote, Y, but the relationship is confounded by a number of factors. 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.[7]

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,

[6] This coarsening is often preferable even in standard regression settings since it relaxes common assumptions. When treatments are continuous, for example, common monotonicity assumptions, as in the case of Instrumental Variables, become much more difficult to satisfy. Continuous variables also often impose functional form assumptions such as linearity. However, because it is nonparametric, `autobounds` makes no such functional form assumptions by default; they must be added as additional assumptions by the user.

[7] See Appendix 2.2 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."

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.

Researchers faced with this challenge are often confronted 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, scholars with substantive expertise in the area may find it difficult to argue that the observed parental SES X is the only factor that influences college education and voting, and they may feel uncomfortable making claims that are premised on this argument.

Those preferring this cautious approach are ill-served by the current state of applied causal methodology, which too often seeks to shoehorn research 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.[8]

```
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)
```

[8]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.

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.[9] It then translates this observed data distribution into implied constraints on the sizes of each principal stratum. See Appendix A, Equation (12) 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.[10]

```
# compute bounds, verify sharpness, and conduct statistical inference
confounding_problem.solve(ci=True, nsamples=1000)
```

The analyst finds that `autobounds` outputs an estimated interval of $[-0.165, 0.835]$. Estimated 95% confidence intervals on the bounds, (i.e. statistical uncertainty due to sampling error) are also reported as $[-0.183, 0.861$. Confidence bounds are computed using the methods described in using the method described in appendices C.2.1 and E.4.5. 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.[11]

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 treat-

[9] 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.

[10] 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).

[11] 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, 1990b). Perhaps surprisingly, this remains true when adjusting for observed confounders as well, absent further assumptions.

ment 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. This is directly grounded in existing research on voting behavior, which has identified the relative cost of voting as a key factor influencing turnout decisions (Li et al., 2018). Because individuals from high-SES families have greater resources available, the costs associated with voting—e.g. transportation, lost wages—represent a relatively lower share of those resources and thus are likely to be perceived as a lesser obstacle, on average. Moreover, high-SES parents are argued to foster politically rich home environments, further increasing political participation later in life (Verba et al., 2005).

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

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

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 $[0.101, 0.841]$ (95% CI $[0.035, 0.859]$). Thus, the analyst is able to estimate informative bounds and sign the effect, while keeping her research question the same, and avoiding untenable assumptions.

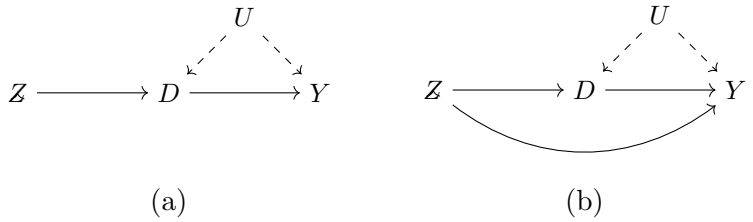
We emphasize that in practice, these restrictions should not be chosen arbitrarily; rather, they should be based on subject knowledge or evidence from prior work. An ideal workflow would state these assumptions once the causal estimand is defined, perhaps even in a pre-registration step. We provide these assumptions and the corresponding results merely to illustrate the flexibility of the algorithm in accommodating contextual information in a relatively simple example. In what follows, we show how `autobounds` can also be used to relax assumptions in published analyses that use more complex research designs, often while still returning informative results.

3 Applications

In this section, we replicate and extend several published papers to demonstrate how the assumptions of commonly used research designs can be relaxed or abandoned using automated partial identification while still allowing for useful inferences. All code to execute these replications appears in the Appendix.

We note that our intention is not to validate or invalidate any of these published studies. We replicate only select results from any given study, whereas each of the original papers also offers other evidence to support its claims. In addition, we sometimes make specification changes, such as discretizing variables, to implement `autobunds`. These exercises are meant to show how our approach can work in practice, with real data, in order to investigate the robustness of results when key assumptions are added, dropped, or partially relaxed.

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 D, which indicates a violation of exclusion restriction.



3.1 Instrumental Variables

When selection on observables cannot be credibly assumed—i.e., when treatment is not as-if randomly assigned, even conditional on observed covariates X—a popular identification strategy is instrumental variables analysis. With this approach, treatment D and outcome Y still share a common unobserved confounder, U, but analysts locate an instrumental variable, Z, that "encourages" treatment to occur as-if randomly, which allows for the identification of a local average treatment effect of D on Y (Angrist et al., 1996). This data-generating process (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 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 X, 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.[12] 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

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

LATE,[13] 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 (user-supplied assumptions). In the IV case, these observable implications are known as the *instrumental inequalities* (Pearl, 1995b; Bonet, 2001). For a valid instrument Z, discrete treatment D, and discrete outcome Y, the exclusion restriction implies that

$$\max_{d} \sum_{y} \Big[\max_{z} \Pr(Y = y, D = d | Z = z) \Big] \leq 1. \tag{2}$$

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

Finally, `autobounds` can 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. We demonstrate these features by replicating and extending an IV study in Section 3.1.1. In Section 3.1.2, we reexamine a second IV study to show how `autobounds` can alert researchers to faulty assumptions by testing whether their empirical implications are violated.

[13] 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 4 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: $\mathbb{E}[Y(d=1) - Y(d=0) | \text{D-complier}] = \big[\Pr(\text{Y-helped, D-complier}) - \Pr(\text{Y-hurt, D-complier})\big] / \Pr(\text{D-complier})$.

3.1.1 Replication and Extension of Coppock and Green (2016)

In a well-known get-out-the-vote (GOTV) experiment, Gerber et al. (2008) 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 one treatment arm, that their neighbors would be informed as to whether they voted. Subsequently, Coppock and Green (2016) extended this study using the instrumental variables framework to test whether voting in one election affects the chances of voting in subsequent elections. In this revised 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 $t+1$ onwards. In Figure 15, we show how this analysis can be replicated and extended in merely 10 lines of `autobounds` code.

Importantly, the assumptions of the IV design imply that habit formation is the only mechanism through which social pressure would affect subsequent turnout in election $t+1$. However, as Davenport et al. (2010) notes, another mechanism that could explain subsequent turnout is "social learning," whereby the content of the encouragement causes people to internalize the civic norm of voting (Bandura and Walters, 1977). If this mechanism were operating, it would constitute a violation of the exclusion restriction—a way in which the encouragement directly influences subsequent turnout that is not mediated by contemporaneous turnout—invalidating the IV design as a means to identify the local average treatment effect among compliers. This is particularly of concern due to 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 general election held just three months later.

It is straightforward to gauge the sensitivity of results to violations of this assumption using `autobounds`. 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 (if this share is hypothesized to zero, then we are effectively re-imposing the exclusion restriction, but if it is greater than zero, we are relaxing it). Finally, we recompute bounds on the estimand when allowing for some violation of the assumption.[14] After setting up the causal diagram, loading data, and stating the monotonicity (no-defiers) assumption, the following lines of Figure 15 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 = edge_is_active("Z -> Y")
    # assume that the size of this group is limited
    add_assumption(p_excl_restr_violation, "<=", 0.01)
```

By varying this hypothesized maximum share of exclusion-restriction violators to many values between 0 and 1, we can produce a full sensitivity curve describing the empirical consequences of any degree of violation.

The bottom left panel of Figure 4 shows the results of this exercise when the complier group is estimated to represent 8% of the population (based on the first-stage estimate, which recovers the complier-group size under the monotonicity assumption). We view these results as indicating extreme sensitivity to the exclusion restriction: the LATE cannot be signed if the social-pressure encouragement mailed in Aug. 2006 directly influenced turnout in Nov. 2006 for more than 1% of units. However, we recognize that others may disagree, especially in other settings where 1% is an implausibly high rate for exclusion-restriction violations. (Substantive knowledge of the topic under study is crucial when interpreting the output of

[14] 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.

this algorithm.) Regardless, the flexibility of our approach allows for a precise discussion about the plausibility of a research design and the robustness of its conclusions, rather than the informal arguments typically used in discussions of identifying assumptions.

The `autobounds` algorithm also allows us to easily explore alternative estimands, such as the ATE. The top left panel of Figure 4 shows that with the observed compliance rate in the data, the ATE can never be signed. In the following three columns of the plot, we show how the sensitivity of results to exclusion-restriction violations changes as compliance rates grow higher, for both the LATE and the ATE, using naturalistic simulations that we generate by reweighting the original dataset. The results show that as compliance rates increase, bounds become more narrow, making it easier to sign the effect. This is especially apparent in the bottom row of the plot, which shows bounds on the LATE. Intuitively, this is because a fixed number of exclusion-restriction violators (say, 1% of the population) can "explain away" much of the change in outcomes when there are not many compliers and the reduced form is therefore small. As the number of compliers grows larger, the instrument will induce bigger shifts in the outcome, and the fraction of those shifts that can be explained away by the 1% of exclusion-restriction violators will be relatively smaller. In addition, we see that under high rates of compliance, the bounds on ATE and LATE are highly similar. This also makes sense, because under perfect compliance, the two quantities are identical.

These results show how IV results can be fragile in a way that current best practices do not adequately emphasize. Practitioners are typically advised to proceed with IV analysis if the first stage is *statistically* strong—according to a common rule of thumb, if the null hypothesis of no $Z \rightarrow D$ effect can be rejected with an F-statistic greater than 10. In the case of the model in Coppock and Green (2016) reanalyzed here, the first-stage F-statistic is 753, indicating an instrument that is statistically extremely strong, due to the large sample size. However, our results show that when compliance rates are relatively low, IV results can still be *causally* fragile, in the sense that they remain vulnerable to minor violations of the

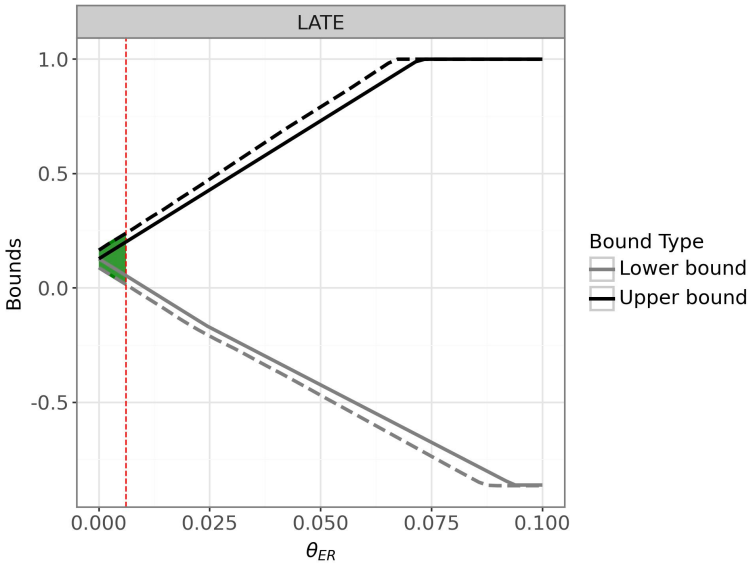


Figure 3: **Sharp Bounds on the LATE Under Increasing Violations of the Exclusion Restriction.** The plot below shows sharp bounds on the LATE of an initial GOTV intervention on turnout in a future election (solid lines are estimated bounds, dashed lines are 95% confidence intervals on those bounds). The plot shows that if 0% of units are assumed to exhbit 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.

exclusion restriction. Conversely, instruments which induce high rates of compliance can be more causally robust to violations of this key assumption.

3.1.2 Replication and Extension of Kocher et al. (2011)

Next, to illustrate how autobounds can alert users to faulty assumptions, we examine 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 most apparent challenge in identifying this effect, which the paper addresses with a number of analytic approaches, is that whether a hamlet is bombed is generally not random but rather is 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.

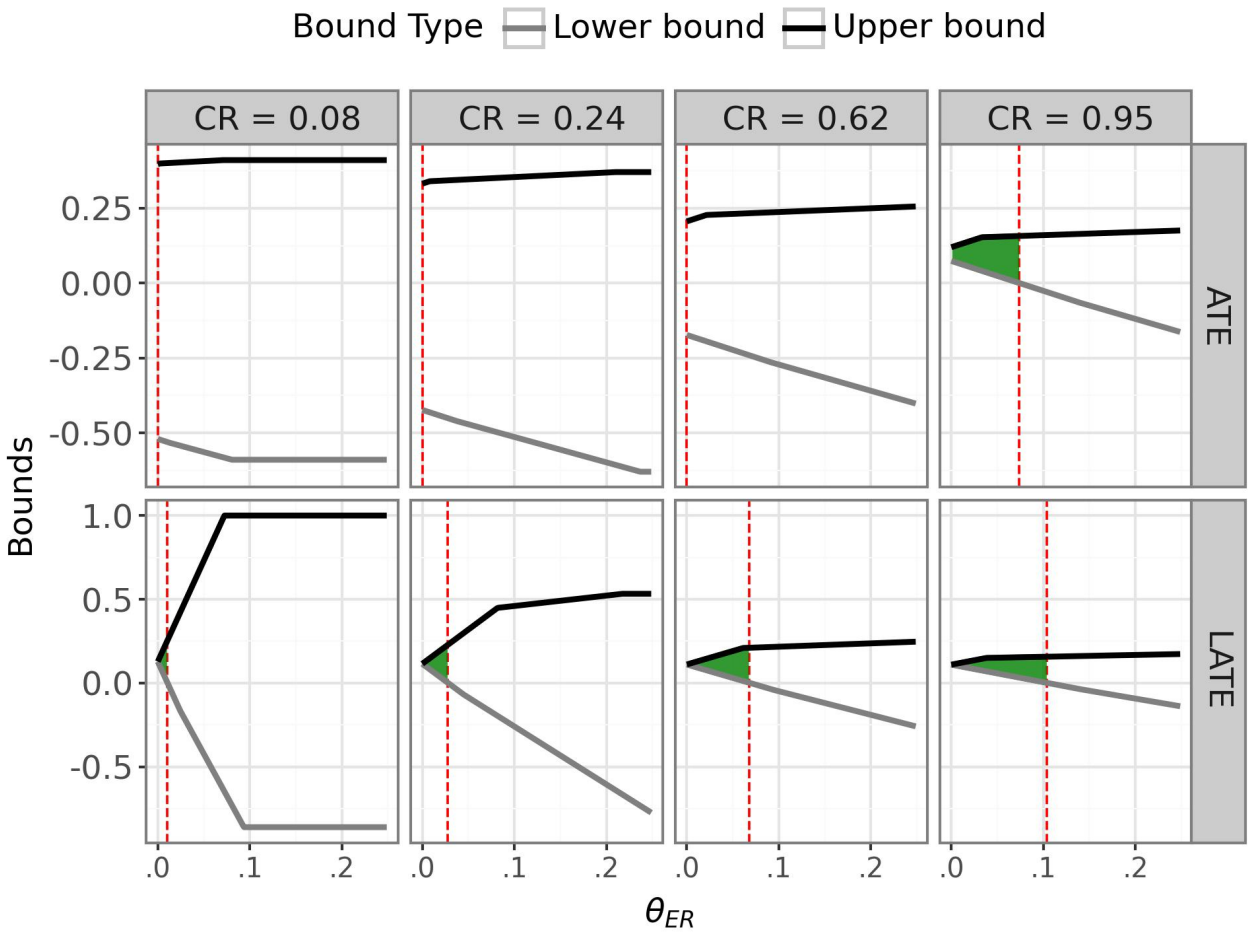


Figure 4: **Sharp Bounds on the ATE/LATE under Increasing Compliance Rates.** The green shaded areas highlight regions where the sign of the effect is identified. The compliance rate in the original data is 0.08, we synthetically reweighted the raw data to generate datasets with higher compliance rates.

To preview our findings, `autobounds` reveals that the instrumental-variables assumption set is *falsified* in this context. In other words, their empirical implications are not satisfied, suggesting that the variables used are not valid instruments.[15] We emphasize that we focus on one particular model specification from Kocher et al. (2011). It is therefore possible the overall conclusions of the paper still hold based on other evidence provided in Kocher et al. (2011). Further, we make a number of specification choices that are distinct from (though conceptually consistent with) the analysis in Kocher et al. (2011).[16]

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.[17] 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 conditions 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 will use `autobounds` to first test the observable implications of these assumptions—i.e.,

[15] 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.

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

[17] 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.

the instrumental inequalities given in Equation 2. 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 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.[18] 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). To probe the intuition behind this result, we manually examined groups of hamlets that exactly match on a coarsened version of the control covariates, $\tilde{\mathbf{X}}$ (Branson and Keele, 2020). For example, consider the 127 highly developed, high-population hamlets that are located in low-roughness terrain that is far from the border. Of these well-off villages, the vast majority were untouched in September 1969 ($D = 0$)—only two were subject to aerial bombing ($D = 1$). Thus, within this subgroup, there is virtually no variation in the causal factors that influence the outcome, $\tilde{\mathbf{X}}$ and D.

While there is large variation in the instrument—for example, 10 hamlets were under full government control, and 14 were under moderate insurgent control—the IV assumptions imply that this variation should be nearly independent of the outcome variable, for two reasons. First, by holding $\tilde{\mathbf{X}}$ fixed within these 127 hamlets, the instrument Z should not be confounded with the outcome Y. And second, the instrument Z should not have an effect on the outcome Y except through variation induced on the treatment D, which in this case is nearly nonexistent. This is the intuition behind the instrumental inequalities given in Equation 2—under a valid IV design, the potential outcomes $Y(d = 0)$ and $Y(d = 1)$ must be independent of Z.

[18] 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.

The observed data is clearly inconsistent with this notion: for example, no hamlets with $Z = 1$ (total government control in July 1969) or $Z = 4$ (moderate insurgent control in July 1969) were bombed, which under the IV assumptions suggests that the distribution of Y (control in December 1969) should be identical between the $Z = 1$ and $Z = 4$ groups. In fact, their distributions of Y do not overlap at all: the $Z = 1$ group stays almost entirely under full government control ($Y = 1$) with one hamlet slipping to moderate government control ($Y = 2$), while all hamlets in the $Z = 4$ group either end up in contested control ($Y = 3$) or remain in moderate insurgent control ($Y = 4$).[19] This suggests either that there is $Z - Y$ confounding or that there is a $Z \to Y$ direct effect that does not flow through D. When evaluating the null hypotheses suggested by a $\tilde{\mathbf{X}}$-conditional version of the instrumental inequalities from Equation 2, we reject the null at $p < 0.001$ in this particular group of 127 hamlets. When using Fisher's method for combining p-values across different groups of coarsened-exact-matched hamlets, we obtain an overall $p_{\text{Fisher}} < 0.001$ as well, decisively rejecting the overall null hypothesis that the Vietnam aerial bombing case is a valid IV design.[20] See Figure 16 in the appendix for the Python code for the `autobounds` implementation.

3.2 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

[19] In general, we only observe $Y(d = 0)$ among the subset of units that actually receive $D = 0$, and because $p(Y(d = 0)|D = 0)$ can differ from $p(Y(d = 0))$ due to D-Y confounding, the IV inequalities implicitly engage in best-/worst-case reasoning about the remaining units with $D = 1$. However, in this particular case, none of the well-off villages with $Z = 1$, $Z = 2$, and $Z = 4$ were bombed—i.e., $\hat{p}(D = 0|\tilde{\mathbf{X}} = \tilde{\mathbf{x}}, Z \in \{1, 2, 4\}) = 1$. Thus, in this subgroup, we are able to directly test whether $Y(d = 0) \perp\!\!\!\perp Z$ using a χ^2 test.

[20] Specifically, we identified groups of coarsened-exact-matched hamlets containing 100 units or more. Within these 16 levels of $\tilde{\mathbf{X}}$, we tested the null hypothesis that $Y(d = 0) \perp\!\!\!\perp Z$ as follows. First, among units with $D = 0$, we constructed a 5×5 contingency table of $Y(d = 0)$ and Z. Then, among units with $D = 1$—for which Z is known but $Y(d = 0)$ is not—we allocated hypothetical $Y(d = 0)$ values that, if true, would result in the lowest χ^2 test statistic. We conduct a χ^2 test of independence for this hypothetical contingency table, representing the most conservative test that is consistent with the observed information. Among the 18 resulting tests, half were significant at the conventional 95% confidence level: we obtained six p-values below 0.001, one more below 0.01, and an additional two below 0.05. Combining these independent tests by Fisher's method, which states that $-2\sum_{j=1}^{J} \log(p_j) \sim \chi^2(J)$ under the overall null hypothesis, yields an overall $p_{\text{Fisher}} < 0.001$.

Figure 5: **Difference-in-differences DAGs.** (a) Standard DiD model. (b) DiD with bracketing on covariates.

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 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.

In what follows, we use `autobounds` to demonstrate what can be learned from these data when the parallel trends assumption is relaxed.

3.2.1 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 5a 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 3.2.1 (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; Figure 17 shows how users input the assumed causal graph, data, and estimand. The parallel-trends assumption can then be stated as follows:

```
# 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.[21]

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 5b. 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

[21] 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.

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

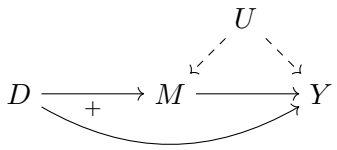
Under this assumption, we find that the bounds for the ATT are $[0.006, 0.046]$ (95% CI $[-0.013, 0.065]$), and the bounds for the ATE are $[-0.042, 0.955]$ (95% CI $[-0.044, 0.958]$). As this final example shows, replacing the relatively restrictive parallel trends assumption with the more lenient bracketing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

3.3 Selection Bias

Selection bias is often inherent to the data sources used by social scientists. A particularly difficult problem arises when treatment status determines whether units are observed or not, and a large literature has grappled with the statistical bias that can result from this form of post-treatment conditioning (Rosenbaum, 1984; Acharya et al., 2016; Nyhan et al., 2017; Blackwell, 2013).

In one recent example, Knox et al. (2020) examines the problem of estimating racial bias

Figure 6: **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.



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) 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 6. 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.[22]

The first assumption we reexamine is "mediator monotonicity," or the assumed absence of anti-white police stops:

Assumption 3.3.1 (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 3.3.2 (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"

[22] 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 ρ indicating racial bias in stopping at 0.32 in the replication below.

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 s 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 may allow racial bias 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. Moreover, it would be difficult to apply this bounding solution in other settings where post-treatment selection occurs, even within the topic of policing—norms and data availability vary tremendously across the roughly 18,000 state and local agencies in the U.S. In many situations, a more flexible approach is needed. We use `autobounds` to relax and abandon these assumptions, showing it is still possible to obtain informative results.

3.3.1 Replication and Extension of Knox et al. (2020)

While the original bounds in Knox et al. (2020) rely on the aforementioned assumptions, `autobounds` lets us to relax or abandon them, see Figure 18 for the relevant code (pages shorter than the proofs required in the original Knox et al. (2020) paper).[23] The far left result in Figure 7 shows analytic results based on prior work for the average treatment effect among the detained, $\text{ATE}_{M=1}$, among encounters involving Black and White civilians, with estimated bounds of $[0.106, 0.121]$ (95% CI $[0.105, 0.122]$).[24]. The second estimate shows

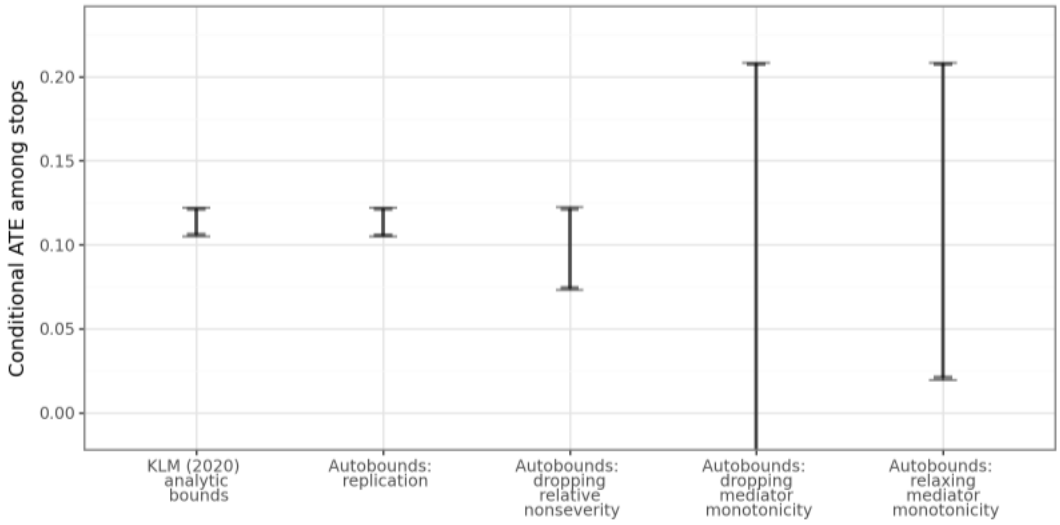
[23]For all numerical analyses, we assume that the probability of being stopped, $P(M = 1)$, is at least 0.02 to ensure numerical stability, as the denominator in the computation of the estimand $ATE_{M=1}$ must be greater than zero; when the denominator approaches zero too closely, numerical tolerance 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 `autobounds` is under development and is expected to be available soon.

[24]These results correspond to the Black-to-white comparison in Knox et al. (2020) under the "baseline specification" without covariate adjustment.

that when given the same information `autobounds` recovers identical bounds $[0.106, 0.121]$ (95% CI $[0.105, 0.122]$. In the third entry in the plot, we eliminate the assumption about the relative severity of racial-stop and always-stop encounters. Somewhat surprisingly, the resulting bounds widen only slightly to $[0.074, 0.121]$ (95% CI $[0.073, 0.122]$). This means analysts can still learn a great deal about the possible severity of racial bias in the use of force in scenarios where this assumption fails.

However, the following estimate in the plot shows that mediator monotonicity is much more consequential. When dropped, the bounds on the $\text{ATE}_{M=1}$ become $[-0.237, 0.207]$ (95% CI $[-0.239, 0.208]$). However, as the final estimate shows, the effect can be signed as positive by relaxing this assumption such that no more than 5% of stops of White civilians are discriminatory, producing $[0.020, 0.207]$ (95% CI $[0.019, 0.208]$). Under the plausible view that anti-white bias in detainment, while possible, is relatively rare in the U.S., this revised analysis shows the original results are robust to substantively large violations of the monotonicity assumption.

Figure 7: **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.



3.4 Mediation Analysis

In a mediation analysis, researchers seek to uncover how an intervention works. Specifically, the goal is to identify an indirect channel by which a treatment affects the outcome. Mediation analyses are used in many contexts including social and political sciences, psychology, epidemiology and health services research (David P. MacKinnon et al., 2007; Richiardi et al.,

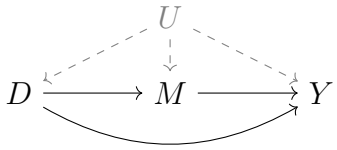
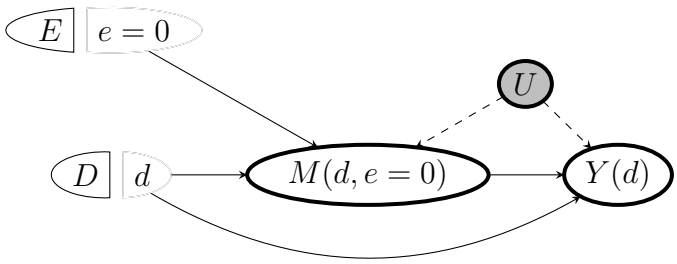


Figure 8: **Mediation data generating process.** The total effect of D on Y may be partially explained through M, the mediator. Obstacles to the validity of mediation designs are often driven by U, which may jointly confound treatment, mediator, and outcome (or any subset thereof). Additional observed confounders may exist but are not depicted. of treatment-outcome, treatment-mediator and mediator-outcome relationships.

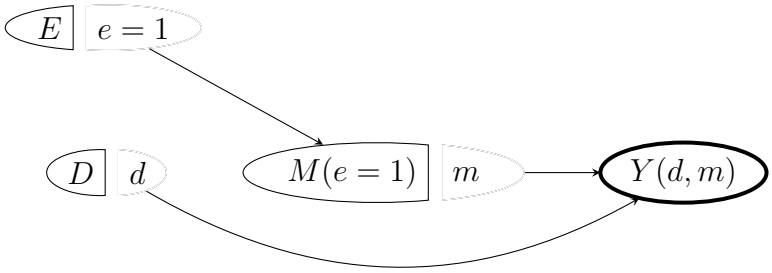
2013; VanderWeele, 2015). Figure 8 contains a standard DAG for a mediation analysis. For simplicity of exposition, we will work with the setting where variables on the main channel $D \to M \to Y$ are all binary. The ATE represents the *total effect* of D on Y, represented by $\mathbb{E}[Y(d=1) - Y(d=0)] = \mathbb{E}[Y(d=1, M(d=1)) - Y(d=0, M(d=0))]$.[25]. However, this estimand only tells us if D causes changes in Y, and is uninformative about the role of M, i.e. the causal mechanism. Mediation analysis aims to decompose this total effect of treatment into direct and indirect effects, which we will together refer to as *mediated effects*. The direct effect represents the influence of the treatment on the outcome that is not mediated by the mediator(s) of interest (i.e. $D \to Y$ path in Figure 8), while the indirect effect represents the influence that is transmitted through M, which represents the proposed causal mechanism. To demonstrate the application of `autobounds` to mediation analysis, we develop new partial identification results for the *parallel design*, introduced by Imai et al. (2011), which is a design-based approach that enables mediation analysis under more plausible identification assumptions. Importantly, the parallel design does not require the "sequential ignorability" assumption—which in effect states that (a potentially unmanipulated) mediator can be considered randomly assigned within each treatment group—and is considered one of the strongest identifying assumptions in mainstream applied causal inference methods. We focus on a recent application of the design in Acharya et al. (2018).

In this design, the researcher randomly assigns subjects to experimental arms, E, which are carried out in parallel. In the *natural arm*, $E = 0$, treatment is randomized but the mediator is not manipulated; thus, the mediator behaves naturally in response to the assigned treatment. In the *manipulated arm*, $E = 1$, the researcher intervenes upon both treatment and mediator. To better represent this design in causal graphical language, Figure 9 represents each experimental arm using its own Single World Intervention Graph (SWIG, Richardson

[25] In a mediation setting $Y(d) := Y(d, M(d))$ is known as the composition assumption (VanderWeele and Vansteelandt, 2009). Intuitively, by intervening on D without intervening on M, Y responds to D at $D = d$ and M equal to the natural level it would have taken with $D = d$, that is, $M(d)$.



(a) Natural arm



(b) Manipulated arm

Figure 9: Single world intervention graphs for the parallel design of survey-based experiments, in which assumptions 3.4.1 and 3.4.2 are explicit. (a) Natural arm. Here, only treatment is manipulated; the ATE is identified. (b) Experimental arm. Both D and M are randomized, and therefore we observe Y in levels of D and M. Here, the CDE is identified. Owing to the randomization of E, information from both arms is pooled to compute (bounds on) the mediated effects under the DGP in (a), where the mediator takes on its natural value in response to treatment.

and Robins, 2013a; Sarvet et al., 2020).[26]

This design involves two assumptions, which we examine in turn below. At a high level, these are assumptions that follow from a careful implementation of the design.

Assumption 3.4.1 (Manipulation exclusion)**.** $Y(d, m, e = 1) = Y(d, M(d), e = 0)$ *whenever* $M(d) = m$ *for any* $d \in \mathcal{S}(D)$ *and* $m \in \mathcal{S}(M)$.

In essence, we assume that participants are unaware of their manipulation. More precisely, the assumption states that if an individual's mediator organically takes on the value m in response to treatment d in the natural arm—meaning that $M(d, e = 0) = m$—then forcibly setting the mediator to the same m in the manipulated arm would not have any

[26]SWIGs generalize DAGs to include counterfactual variables via the 'node-splitting' operation which illustrates how factual variables transform under an intervention, but are otherwise functionally very similar. SWIGs have proven especially effective at visually representing the conditional independencies in causal models, in particular by capturing how these relationships change under different interventions (Sarvet et al., 2020), an important consideration here. For more details, consult Richardson and Robins (2013b).

downstream consequences. In other words, if the individual's response in the natural arm would be $Y(d, M(d, e = 0) = m, e = 0) = y$ in the natural arm, then their response in the manipulated arm would also be $Y(d, m, e = 1) = y$. This assumption is encoded in the causal graphs of Figure 9, because the experimental arm assignment E does not impact the outcome Y. If this assumption is violated, then observed differences across experimental arms could be an artifact of the experimental assignment, meaning that an analyst would not be able to convincingly learn about mediated effects. See Appendix E.4.2 for further discussion of this assumption.

Assumption 3.4.2 (Randomization)**.** *(a) The experimental arm is ignorable,* $\{Y(d, m), M(d'), D\} \perp\!\!\!\perp E$*; (b) treatment assignment is ignorable in the natural arm,* $\{Y(d, m), M(d')\} \perp\!\!\!\perp D \mid E = 0$*; and (c) treatment and mediator assignments are ignorable in the manipulated arm* $Y(d, m) \perp\!\!\!\perp \{D, M\} \mid E = 1$ *for any* $d, d' \in \mathcal{S}(D), m \in \mathcal{S}(M)$*.*

This multi-part assumption provides a basis for comparisons within and across arms. First, it states that the experimental arm is as good as randomly assigned. It further states that the treatment is as good as randomly assigned in both the natural and manipulated arms, and moreover, the mediator is also as good as randomly assigned in the manipulated arm. Acharya et al. (2018) argues that in survey experiments, information-based mediators can be effectively randomized: when M represents a piece of information that is not provided in the natural arm, so that respondents must indirectly infer from treatment, then researchers can intervene upon M by explicitly providing this information in the manipulated arm. This contrasts with scenarios where M represents ephemeral objects such as emotions or beliefs, which can only be primed and which must also be measured post-treatment—in these cases, treatment-induced confounding represents a major threat to validity in a mediation analysis (VanderWeele and Vansteelandt, 2009; Imai et al., 2010, for more discussion, see Appendix E.4.1).

We now discuss matters of identification. In a mediation analysis, the goal is to commonly to decompose the ATE into components that do and do not flow through the mediator.

The first component is typically represented in terms of the natural indirect effect (NIE). Conceptually, the NIE represents the average change in potential outcomes that would occur if the treatment were held fixed, but the mediator varied (i.e. took whatever value it would "naturally" take given treatment status for each unit), all else equal. In other words, it is one way to describe the share of total effect of the treatment on the outcome that can be attributed to the mediator.

The second component, representing the part of the ATE that does not flow through the mediator, is most commonly described in terms of the natural direct effect (NDE); however, some work has also focused on a related quantity, the controlled direct effect (CDE). The NDE characterizes the portion of the total effect that does not involve the mediator, again while allowing the mediator to take its "natural value" given treatment status. In contrast, the CDE is the direct effect of the treatment on the outcome holding the mediator at a specific value (regardless of whether the mediator would ever "naturally" take that value for a given unit). In other words, it is another conception of the portion of the total effect that does not involve the mediator. (See Appendix E.4.1 for formal definitions of these estimands.) Assumptions 3.4.1 and 3.4.2 identify the ATE and the CDE, but the NIE and NDE—which are more commonly of interest in mediation settings—remain unidentified.

While this may seem to be an important limitation of the parallel design, Acharya et al. (2018) seek to circumvent the issue by arguing that the *difference* between the ATE and CDE, which they call the eliminated effect (EE), is itself a meaningful quantity of interest for explaining causal mechanisms. As noted above, the approach in Acharya et al. (2018) does not allow for the calculation of the NIE or NDE.[27] Rather, the EE is a difficult-to-interpret combination of an indirect effect (which is of central importance in mediation analyses) and a causal interaction (which is typically of lesser interest). In this section, we show how `autobounds` can

[27]To see this, Acharya et al. (2018) show the following decomposition of the eliminated effect at level m into the natural indirect effect and the *reference interaction* (RI), $\text{EE}(m) := \text{NIE} + \text{RI}(m)$, where the RI is defined to be $\text{RI}(m) := \text{NDE} - \text{CDE}(m)$. This quantity provides a combined measure of direct mediation, NIE, and the impact of the interaction between the treatment and mediator on the direct effect (VanderWeele and Knol, 2014; VanderWeele, 2015). For a fuller discussion of this estimand, see the appendix E.4.3.

make the parallel design more informative. Rather than giving up on the standard quantity of interest in mediation analysis—the NIE—simply because it cannot be point identified, we instead use partial identification under weak assumptions to obtain informative bounds on it.

3.4.1 Replication and Extension of Acharya et al. (2018)

Here, we replicate and extend Study 2 in Acharya et al. (2018) which features an experiment from Tomz and Weeks (2013). The study investigates how Democratic Peace theory structures public opinion, and investigates whether support for a preemptive strike, Y, against a democratic nation ($D = 1$) as opposed to an autocratic nation ($D = 0$) is mediated by the belief that said nation is perceived to be a threat, M. We use `autobounds` to bound the NDE and NIE (for sample code, see Figure 19 in the appendix), which was not possible in the original analysis.

We do so by placing an additional assumption, grounded in the literature on public support for war, on the way that regime type interacts with the perception of threat to national security. At a high level, this assumption states that national security threats will diminish the role of other factors—in other words, democratic norms may matter, but they will matter less when national security is on the line.

Assumption 3.4.3 (Threat Dominates Regime Type for Most Respondents)**.**

$$\Pr\left(\begin{array}{l} Y_i(d=0,m=0) - Y_i(d=1,m=0) \\ \quad > \; Y_i(d=0,m=1) - Y_i(d=1,m=1) \end{array}\right) \quad \geq \quad 1-\gamma$$

allowing some fraction of respondents $\gamma \in [0,1]$ *to behave contrary to the theorized patterns.*

The upper term in the probability, $Y_i(d=0,m=0) - Y_i(d=1,m=0)$, represents the difference in respondent i's support for a preemptive strike against a nonthreatening autocracy, compared to a similarly nonthreatening democracy. When this term is positive, unit i is more supportive of unprovoked attacks on autocracies and less supportive of unprovoked attacks on nations that share democratic , when both nations pose no threat. The lower term, $Y_i(d=0,m=1) - Y_i(d=1,m=1)$, represents the difference in i's support for an attack on a *threatening* autocracy (compared to a similarly threatening democracy). The assumption states that for most people, the latter difference is diminished: when the nation in question poses a threat, the autocracy-democracy difference will shrink. This would be consistent with a model in which national-security threats loom largest and respondents have limits on their cognitive capacity or attention span for considering other criteria. The assumption does not require that this hold for *every* respondent; rather, it states that when drawing respondents from the population, *most* individuals (a fraction of $1-\gamma$) will think this way, while the remaining γ fraction may respond differently.

Armed with this assumption, we can parcel out the component mediation effect and interaction effect which comprise the eliminated effect. In particular, Figure 10 displays our

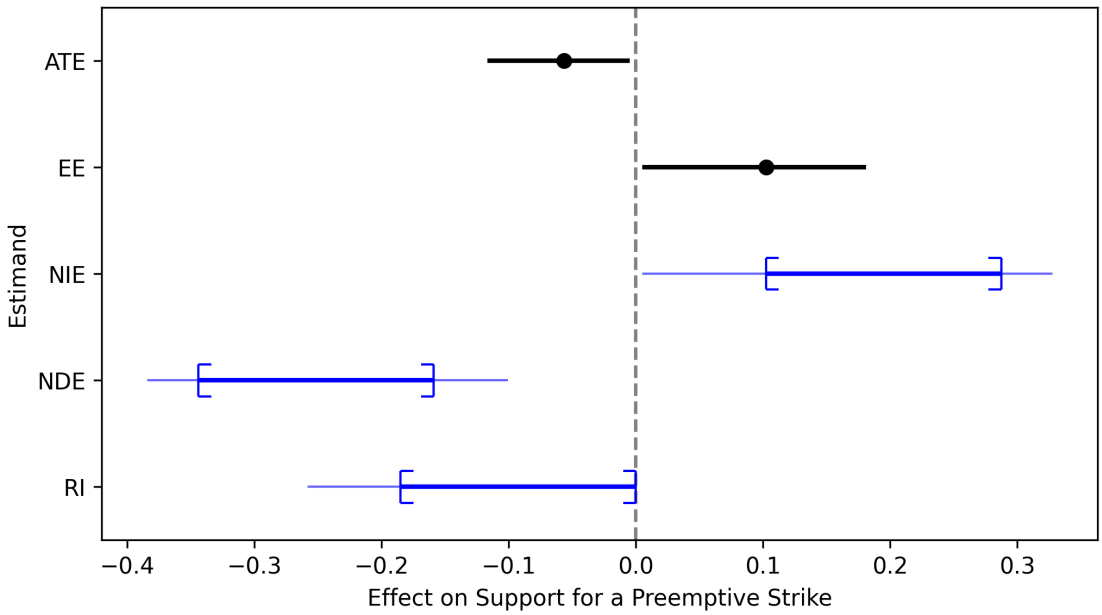


Figure 10: The effect of democracy on supporting a preemptive strike. **(Black)** Point estimates and 95% confidence intervals from Acharya et al. (2018). **(Blue)** Point estimates and 95% confidence bands of the bounds on the NIE, RI and NDE calculated using `autobounds` under assumption 3.4.3.

partial identification analysis of the NIE and NDE and RI (the latter two fix the mediator at the level corresponding to "threat") for $\gamma = 0$ and Figure 11 shows the NIE and NDE as γ varies within $[0, 0.5]$. (The sensitivity curve for the RI is not shown, as its sign is only identified at $\gamma = 0$.) Notably, the point estimates are informative in that we can sign the NIE for $\gamma \in [0, 0.05)$. For $\gamma = 0$ the RI is estimated to be negative, the NIE is positive and the NDE is strongly negative. The latter result is very intuitive: fixing perception of threat, the results suggest that Americans will be much less likely to support an attack on a democracy than an autocracy. The dashed lines of Figure 11 show 95% confidence bands around the bounds which reflect sampling uncertainty. Accounting for uncertainty, the region of sign identification is smaller for the NDE and vanishes for the NIE and RI, so we cannot draw strong inferences about these quantities.[28]

[28] We believe that for this study, the width of the bounds and confidence intervals are in part due to insufficiently varied data to optimally constrain the strata frequencies without further assumptions; see appendix E.4.5 for more discussion.

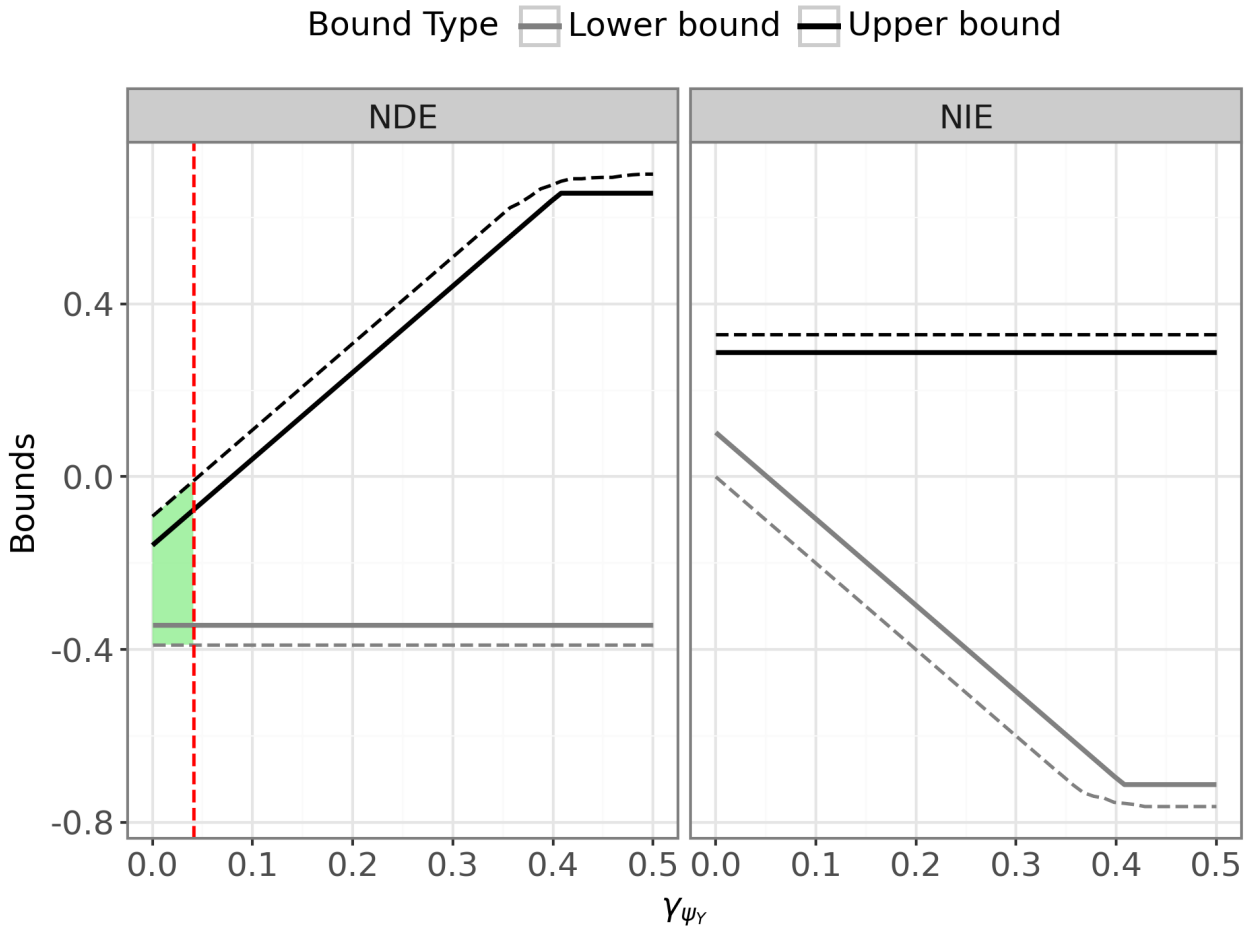


Figure 11: Mediation and interaction effects on support for democratic peace. Bounds on the NDE and NIE under assumption 3.4.3 in the replication of Acharya et al. (2018). The red vertical line indicates the point at which we lose sign identification at the 95% confidence level.

4 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 from data in a variety of scenarios. But as any applied researcher knows, these assumptions rarely hold perfectly in practice.

In this paper, we demonstrate how recent advances in automated partial identification can allow researchers to easily adapt several common causal designs to accommodate potential violations of assumptions while still extracting as much information from data as possible.

As our applications show, this approach affords numerous benefits, and addresses several obstacles that have prevented the widespread use of partial identification strategies in the past. 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 data generating process while easily recomputing sharp bounds on the solution. 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. Many components of causal research—namely, estimands and identifying assumptions—are crucial to replicating and extending prior work, but these key factors are often left vague or implicit 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. 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. With this knowledge, researchers can better

target their efforts, designing projects which interrogate the validity of more consequential assumptions, or seeking out data and scenarios that may 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 and fertile area of research. For instance, while we provide a method that includes covariates, incorporating more complex data structures—such as clustered standard errors—into the modeling 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.

We caution that our approach is not a panacea, and does not obviate the need for careful research design. Without sound theory and identification strategies that make assumptions plausible, the results produced by `autobunds` 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., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

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.

Acharya, A., M. Blackwell, and M. Sen (2018). Analyzing Causal Mechanisms in Survey Experiments. *Political Analysis 26*(4), 357–378.

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. (1999). Estimation when a parameter is on a boundary. *Econometrica 67*(6), 1341–1383.

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.

Avin, C., I. Shpitser, and J. Pearl (2005). Identifiability of path-specific effects. In *IJCAI International Joint Conference on Artificial Intelligence*, pp. 357–363.

Balke, A. and J. Pearl (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier.

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.

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.

Branson, Z. and L. Keele (2020). Evaluating a key instrumental variable assumption using randomization tests. *American Journal of Epidemiology 189*(11), 1412–1420.

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

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

Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics 64*(3), 695–701.

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., S. Lee, and A. M. Rosen (2013). Intersection bounds: Estimation and inference. *Econometrica 81*(2), 667–737.

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.

David P. MacKinnon, Amanda J. Fairchild, and Matthew S. Fritz (2007). Mediation Analysis. *Annual Review of Psychology 58*(1), 593–614.

Duarte, G., N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser (2023). An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*.

Elwert, F. and C. Winship (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology 40*, 31.

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

Gabriel, E. E., M. C. Sachs, and A. Sjölander (2020). Causal bounds for outcome-dependent sampling in observational studies. *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2020.1832502.

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.

Heckman, J. and E. Vytlacil (2001). *Instrumental variables, selection models, and tight bounds on the average treatment effect*, pp. 1–15. Physica.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161.

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.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*(4), 765–789.

Imai, K., L. Keele, and T. Yamamoto (2010, February). Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science 25*(1), 51–71.

Ji, W., L. Lei, and A. Spector (2023). Model-agnostic covariate-assisted inference on partially identified causal effects. *arXiv preprint arXiv:2310.08115*.

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.

Kennedy, E. H., S. Harris, and L. J. Keele (2019). Survivor-complier effects in the presence of selection on treatment, with application to a study of prompt icu admission. *Journal of the American Statistical Association 114*(525), 93–104.

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., T. Yamamoto, M. A. Baum, and A. J. Berinsky (2019). Design, identification, and sensitivity analysis for patient preference trials. *Journal of the American Statistical Association 114*(528), 1532–1546.

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.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

Levis, A. W., M. Bonvini, Z. Zeng, L. Keele, and E. H. Kennedy (2023). Covariate-assisted bounds on causal effects with instrumental variables. *arXiv preprint arXiv:2301.12106*.

Li, A. and J. Pearl (2021). Bounds on causal effects and application to high dimensional data. arXiv preprint arXiv:2106.12121.

Li, Q., M. J. Pomante, and S. Schraufnagel (2018). Cost of voting in the american states. *Election Law Journal: Rules, Politics, and Policy 17*(3), 234–247.

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. (1990a). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Manski, C. F. (1990b). 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.

Manski, C. F. and J. V. Pepper (2009). More on monotone instrumental variables. *The Econometrics Journal 12*(suppl_1), S200–S216.

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.

Molinari, F. (2020). Microeconometrics with partial identification. arXiv:2004.11751.

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.

Pearl, J. (2012). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention science 13*, 426–436.

Pearl, J., J. Breese, and D. Koller (2001). Proceedings of the seventeenth conference on uncertainty in artificial intelligence.

Richardson, T. S., R. J. Evans, and J. M. Robins (2011). Transparent parameterizations of models for potential outcomes. *Bayesian statistics 9*, 569–610.

Richardson, T. S. and J. M. Robins (2013a). Single World Intervention Graphs: A Primer. In *Proceedings of the Second UAI Workshop on Causal Structure Learning*, Bellevue, WA. Accessed Feb 25, 2025.

Richardson, T. S. and J. M. Robins (2013b). Single World Intervention Graphs (SWIGs) : A Unification of the Counterfactual and Graphical Approaches to Causality. *Working Paper, Center for Stat. & Soc. Sci., U. Washington 128*(30).

Richiardi, L., R. Bellocco, and D. Zugna (2013). Mediation analysis in epidemiology: Methods, interpretation and bias. *International Journal of Epidemiology 42*(5), 1511–1519.

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., P. Green, N. Hjort, and S. Richardson (2003). Highly structured stochastic systems. *Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects. Oxford University Press, Oxford*.

Robins, J. M. and S. Greenland (1992, March). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*(2), 143–155.

Robins, J. M., T. S. Richardson, and I. Shpitser (2022). An interventionist approach to mediation analysis. pp. 713–764.

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.

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.

Sarvet, A. L., K. N. Wanis, M. J. Stensrud, and M. A. Hernán (2020, May). A graphical description of partial exchangeability. *Epidemiology 31*(3), 365–368.

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.

Shao, J. (1994). Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society 122*(4), 1251–1262.

Sjölander, A., W. Lee, H. Källberg, and Y. Pawitan (2014). Bounds on causal interactions for binary outcomes. *Biometrics 70*(3), 500–505.

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.

Swanson, S. A., M. A. Hernán, M. Miller, J. M. Robins, and T. S. Richardson (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association 113*(522), 933–947. DOI: 10.1080/01621459.2018.1434530.

Tchetgen Tchetgen, E. J. and T. VanderWeele (2014). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology 25*(2), 282–291.

Tomz, M. and J. L. P. Weeks (2013, November). Public Opinion and the Democratic Peace. *American Political Science Review 107*(4), 849–865.

VanderWeele, T. J. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction.

VanderWeele, T. J. and M. J. Knol (2014). A Tutorial on Interaction. *Epidemiologic Methods 3*(1), 33–72.

VanderWeele, T. J. and S. Vansteelandt (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface 2*(4), 457–468.

Vansteelandt, S. and R. M. Daniel (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.) 28*(2), 258.

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.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

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.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.

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 12. The DAG in this diagram is merely a graphical representation of the following structural causal model (Pearl, 2009)

$$\begin{aligned} D &= f_D(U) \\ Y &= f_Y(d, U), \end{aligned} \tag{3}$$

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

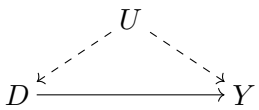


Figure 12: The data generating process of a simple case of unobserved confounding.

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

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

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

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

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 $D\text{-ctl} = \mathbf{1}\left\{f_D^U(\varnothing) = 0\right\}$. Figure 13(a) illustrates these functions pictorially. 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

$$f_Y^{(U=u)} : \{0, 1\} \to \{0, 1\}, \tag{5}$$

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 (5) 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* and display them pictorially in Figure 13(b). In more familiar terms, the first two response types are the units whose voting preferences are unaffected by their education: the "always voters," Y-av, who would vote regardless of their education and the "never voters," Y-nv, who would never vote, regardless of their education. The third type are the "helped," Y-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

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

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; Duarte et al., 2023).

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. Figure 13(c) represents this conclusion; for example, the region containing D-trt, Y-nv may be interpreted as the statement: there exists an $A \subseteq \mathcal{S}(U)$ such that $\Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{nv}) = \Pr(U \in A)$[29]. There

[29]For notational conciseness, we write $\Pr(V\text{-}\mathsf{type})$ as shorthand for $\Pr(V\text{-}\mathsf{type} = 1)$ throughout for any

are eight of the regions A, called *canonical partitions*, which can be thought of as eight point masses of the now discrete random variable U. For more details and examples, consult Balke and Pearl (1997) and Duarte et al. (2023).

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. (2023), which we restate below.

Proposition 1. *Suppose $\mathcal{G}$ is a canonical*[30] *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 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 $\mathcal{G}$*

$$\Pr\left(\bigcap_l C_l\right) = \sum_{\boldsymbol{U} \in \mathcal{S}(\boldsymbol{U})} \mathbf{1}\{\boldsymbol{U} \implies \{C_l : l\}\} \prod_k \Pr(U_k = u_k). \tag{6}$$

There are three key elements of this proposition we highlight here. Firstly, Equation 6 says that any factual[31] 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

variable V.

[30] 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. (2023) for an example.

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

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

$$\Pr(D = 1, Y = 0) = \sum_{u \in \mathcal{S}(U)} \mathbf{1}\{u \implies \{D = 1, Y = 0\}\} \Pr(U = u), \tag{7}$$

$$= \sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing) = 1, f_Y^{(U=u)}(d = 1) = 0 \right\} \Pr(U = u). \tag{8}$$

From this, we can now make the connection to principal strata clear. Notice that the right-hand side of Equation 8 contains only individuals who posses a Bachelor's degree ($D = 1$), and those who would not vote if they had one. Therefore, the voter can only have Y-type either Y-nv or Y-hu. Mathematically, this is a simple consequence of the equivalence between the the right-hand side of (8) and the following expression[32]

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

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

It therefore follows that

$$\Pr(D = 1, Y = 0) = \Pr(D\text{-trt}, Y\text{-nv}) + \Pr(D\text{-trt}, Y\text{-hu}). \tag{9}$$

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

[32] If X and W are propositions, then $\mathbf{1}\{X, W\} + \mathbf{1}\{X, W^c\} = \mathbf{1}\{X\}$ where W^c is the complement of W.

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

$$\begin{aligned} D &= 1 \\ Y(1) &= f_Y(d = 1, U) \end{aligned} \tag{10}$$

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 $(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{he}), (D\text{-}\mathsf{trt}, Y\text{-}\mathsf{av}), (D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{he}), (D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{av})$, which yields

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

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

$$\mathbb{E}[Y(1) - Y(0)] = \Pr(D\text{-}\mathsf{trt}, Y\text{-}\mathsf{he}) + \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{he}) - \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{hu}) - \Pr(D\text{-}\mathsf{ctl}, Y\text{-}\mathsf{hu}). \tag{11}$$

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*: $\Pr(D\text{-type}), \Pr(D\text{-type-}Y\text{-type}')$ for all stratum types, which are of course unknown. Secondly, define the objective: this is the ATE in equation (11). 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

$$\underset{\substack{\Pr(D\text{-type}),\\ \Pr(D\text{-type},\, Y\text{-type}')}}{\text{optimize}} \quad \Pr(D\text{-ctl}, Y\text{-he}) + \Pr(D\text{-trt}, Y\text{-he}) - \Pr(D\text{-ctl}, Y\text{-hu}) - \Pr(D\text{-trt}, Y\text{-hu}) \tag{12a}$$

$$\text{subject to} \quad 0 \leq \Pr(D\text{-type}) \leq 1,\ 0 \leq \Pr(D\text{-type}, Y\text{-type}') \leq 1,\ \forall\, \text{type}, \text{type}' \tag{12b}$$

$$\sum_{\text{type}} \Pr(D\text{-type}) = 1,\ \sum_{\text{type,type}'} \Pr(D\text{-type}, Y\text{-type}') = 1 \tag{12c}$$

$$\Pr(D=0, Y=0) = \Pr(D\text{-ctl}, Y\text{-he}) + \Pr(D\text{-ctl}, Y\text{-nv}) \tag{12d}$$

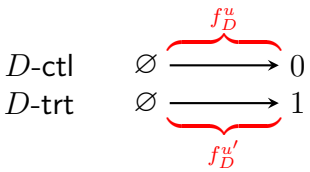
$$\Pr(D=0, Y=1) = \Pr(D\text{-ctl}, Y\text{-hu}) + \Pr(D\text{-ctl}, Y\text{-av}) \tag{12e}$$

$$\Pr(D=1, Y=0) = \Pr(D\text{-trt}, Y\text{-nv}) + \Pr(D\text{-trt}, Y\text{-hu}) \tag{12f}$$

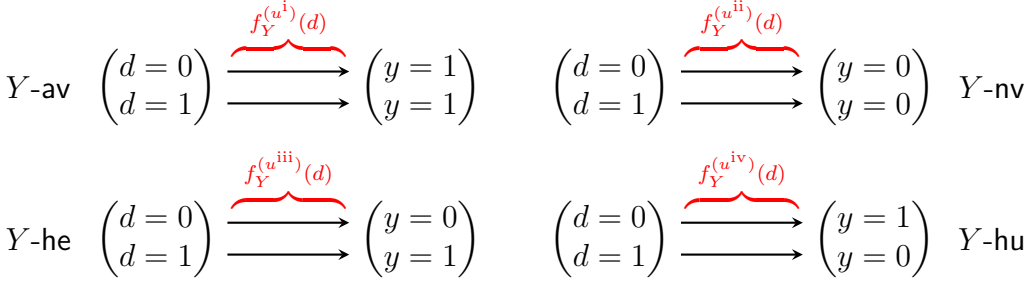
$$\Pr(D=1, Y=1) = \Pr(D\text{-trt}, Y\text{-he}) + \Pr(D\text{-trt}, Y\text{-av}) \tag{12g}$$

Lines (12b)-(12c) are the constraints contained in $\mathcal{C}_{\mathcal{P}}$ while lines (12d)-(12g) are the constraints contained in $\mathcal{C}_{\mathcal{E}}$. Intuitively, in problem (12), 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.

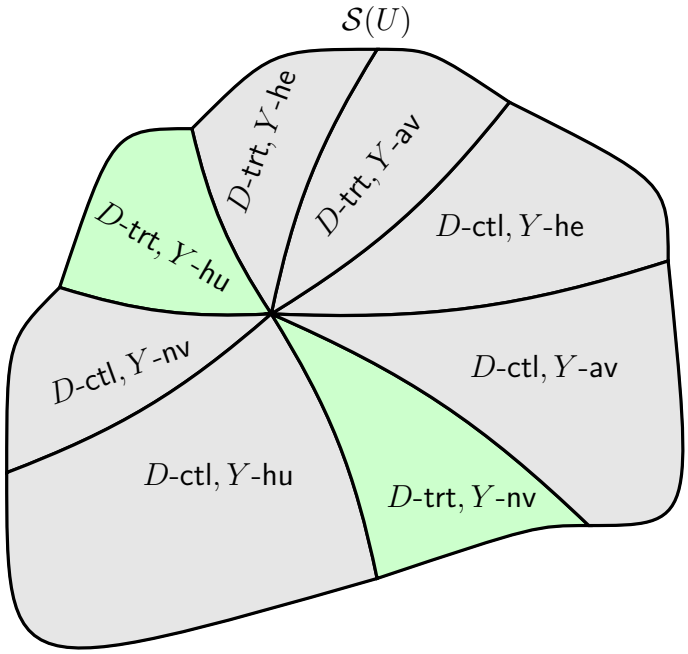
In Appendix E, we will repeat the polynomialization explained in detail in this section for each of the examples this paper discusses. For the sake of brevity, we will henceforth not convert every disturbance polynomialization into that over strata frequencies, however, to review examples of how to do this we recommend the reader revisit Appendices A and E.1.



(a) 2 possible education types



(b) 4 possible voter types



(c) 8 possible individual types

Figure 13: Schematic illustrating canonical partitions and principal stratification in the binary case. (a) Certain characteristics $u, u' \in \mathcal{S}(U)$ will determine if a unit obtains a bachelor's degree. (b) Potentially different characteristics $u^{\mathrm{i}}, u^{\mathrm{ii}}, u^{\mathrm{iii}}, u^{\mathrm{iv}} \in \mathcal{S}(U)$ determine how a unit's voting preferences respond to their education. (c) A diagram displaying how we can conceive $\mathcal{S}(U)$ as partitioned by eight non-overlapping regions corresponding to the combinations of types in (a) and (b); in this way, we show visually that our data are sufficiently described by these strata. For example, we can visualize $P(Y(1) = 1)$ as the probability of belonging to the green shaded region.

A.1 Simulated Data for Section 2.2

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 2.2, 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 pa_V is used to denote the observed parents of V; if 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 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 $y = y_0 y_1$, 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

[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.

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 2.2. 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 $d = d_0 d_1$, 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 $P(X_x) = 0.6^x(1-0.6)^{(1-x)}$),

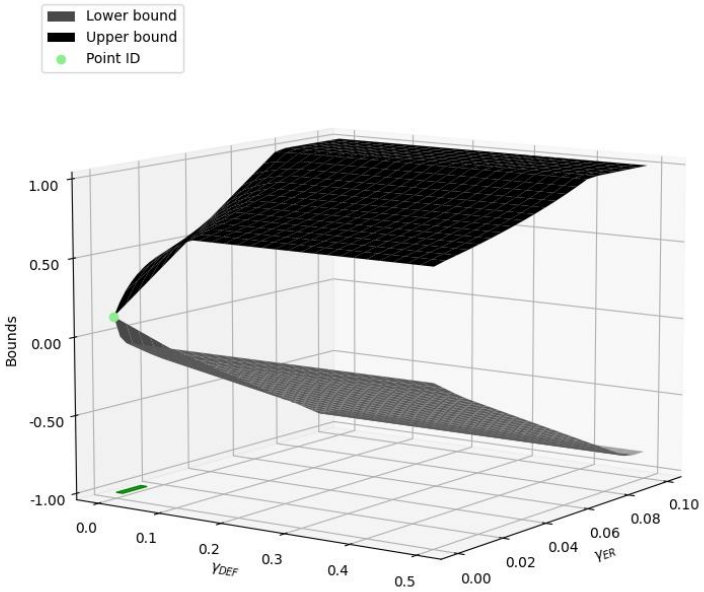
$$\Pr(D_{00}Y_{0000}) = 0.074576212961429,$$
$$\Pr(D_{01}Y_{0001}) = 0.0301702777607751,$$
$$\Pr(D_{01}Y_{0101}) = 0.186607173274612,$$
$$\Pr(D_{01}Y_{1101}) = 0.00798467737923486,$$
$$\Pr(D_{10}Y_{0010}) = 0.348297597134471,$$
$$\Pr(D_{10}Y_{1001}) = 0.134412818663854,$$
$$\Pr(D_{10}Y_{1010}) = 0.0480346194425266,$$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

and all other strata have probability zero.

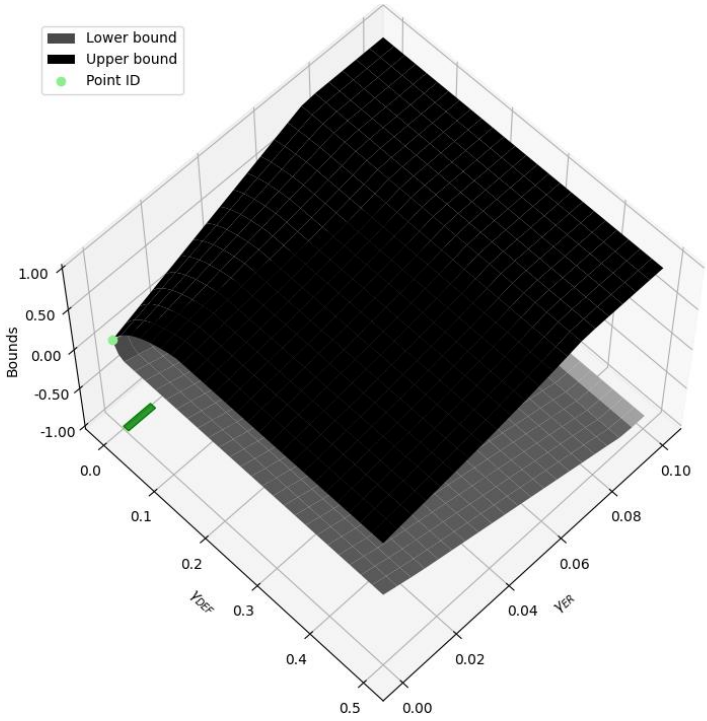
With this setup, using ideas discussed in the previous section, we compute the true ATE as the relevant combinations of these strata and then produce the observed law on which we apply `autobounds`. We also make sure that this DGP satisfies the weak monotonicity assumption. See `a_coded_example_simulation.py`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

B Additional Sensitivity Analyses for Instrumental Variables

We can simultaneously relax monotonicity and the exclusion restriction. Since we change several sensitivity parameters at once, we produce a surface of lower and upper bounds for the LATE. Figure 14 displays the results. This figure reveals several interesting points. First, remark that cross-sections of this figure along the line $\gamma_{\text{DEF}} = 0$ is Figure 3. Second, observe that the set of simultaneous violations $(\gamma_{\text{DEF}}, \gamma_{\text{ER}}) \in [0, 0.01] \times [0, 0.01]$ yield bounds on the LATE which do not include zero. This set is the cross of the green regions in Figure 14 and its existence demonstrates that violations of both assumptions simultaneously may still yield informative bounds on the LATE, however, in this study the tolerance for violations of assumptions is minimal. Lastly, the complex shape of the surface reveals a non-trivial interaction between the sensitivity parameters, the analysis of which is a subject of further study.



(a) Isometric view



(b) Plan view

Figure 14: **Multivariate sensitivity bounds on the LATE.** The green rectangular region is the collection of all $(\gamma_{\text{DEF}}, \gamma_{\text{ER}})$ where the sign of the LATE is identified.

C Practical Considerations: 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.[34] 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 $\varphi := \varphi(P_{\boldsymbol{V}(\mathcal{W})})$. 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 φ 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

[34] `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.

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 $\hat{P}_{\boldsymbol{V}}(d, y) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(D_i = d, Y_i = y)$.

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*.

Here, we develop a quasi-Bayesian technique for computing these confidence bounds that possesses asymptotic frequentist properties. Specifically, we utilize the *transparent reparameterization* proposed by Richardson et al. (2011); see also Knox et al. (2019). A model following this reparameterization distinguishes between causal parameters that are identified and unidentified, and it places priors only on the parameters which are identified. This approach

has two primary advantages. First, Richardson et al. (2011) shows that this approach can circumvent an undesirable property of standard Bayesian approaches: the fact that priors on unidentified parameters can induce strong posterior sensitivity to the specific prior chosen, because data are fundamentally unable to contradict these priors. Second, by utilizing the transparent parameterization, analysts can ensure that posterior inferences will respect the restrictions imposed by the assumed causal model. Even when off-the-shelf priors on the observed data law are used—i.e., priors that do not necessarily respect the restrictions implied by the causal assumptions—the procedure we propose can detect and reject posterior samples that are inconsistent with these assumptions. In general, implied constraints are not easy to obtain; the instrumental inequalities given in (2) are one of the few known examples. However, the `autobounds` paradigm ensures that they are accounted for because of the way that falsification testing is automatically handled (see Appendix A for details).

C.2.1 Uncertainty Quantification without Covariates

We now explain our quasi-Bayesian approach to uncertainty quantification in detail. Suppose the discrete variables in $\boldsymbol{V}$ can take on K possible combinations; in this case, $P_{\boldsymbol{V}}$ can be thought of as a K-valued categorical distribution. We will collect the parameters of $P_{\boldsymbol{V}}$ in the vector $\boldsymbol{\theta}_{\boldsymbol{V}}$, where the k-th element $\theta_{\boldsymbol{v}_k}$ represents the probability of the k-th combination, $P_{\boldsymbol{V}}(\boldsymbol{V} = \boldsymbol{v}_k)$. For instance, in the simple confounding example, there are $K = 4$ observable combinations of main variables: one parameter represents the probability $P_{\boldsymbol{V}}(D = 0, Y = 0)$, while another represents $P_{\boldsymbol{V}}(D = 1, Y = 1)$. Going forward, in a slight abuse of notation, we will write $\underline{\mathsf{A}}_\varphi(P_{\boldsymbol{V}})$ and $\underline{\mathsf{A}}_\varphi(\boldsymbol{\theta}_{\boldsymbol{V}})$ interchangeably, and similarly for the upper bound $\overline{\mathsf{A}}_\varphi$, as the observed data distributions $P_{\boldsymbol{V}}$ are in one-to-one correspondence with their parameter vectors $\boldsymbol{\theta}_{\boldsymbol{V}}$.

For the $\boldsymbol{\theta}_{\boldsymbol{V}}$ parameters of the observed data law $P_{\boldsymbol{V}}$, we impose an uninformative uniform prior, Dirichlet($\mathbf{1}_K$)—in other words, every possible combination of observed data values is given equal prior mass.[35] It can be seen that the exact prior is unimportant, as the posterior will be dominated by the likelihood as the number of samples grows large. Crucially, in

[35] $\mathbf{1}_K$ is the K-dimensional vector of all ones.

the transparent parameterization paradigm, we do not place priors on purely counterfactual quantities, which only appear in the full data law $P_{\boldsymbol{V}(\mathcal{W})}$ and not in the observed data law $P_{\boldsymbol{V}}$. In the confounding example, for instance, we do not place priors on the proportion of "never takers," $P_{\boldsymbol{V}(\mathcal{W})}(Y(d=0)=0, Y(d=1)=0)$, because the observed data could never contradict this prior due to the fundamental problem of causal inference. For the same reason, we do not conduct standard Bayesian inference with respect to these fundamentally unidentified parameters—e.g., by sampling or integrating over possible values for them—because their "posteriors" would merely be driven by the priors chosen. Rather, our approach is "quasi-Bayesian" in the sense that it conducts best- and worst-case reasoning over these unidentified parameters and is only truly Bayesian with respect to the identified parameters of the observed data law.

The dataset to be analyzed is a collection of samples $\mathbb{V} = \{\boldsymbol{V}_i\}_{i=1}^N$; this can be thought of as a matrix in which each row i contains $\boldsymbol{V}_i$, the vector of observed values for unit i, which is an i.i.d. draw from the true observed data law $P_{\boldsymbol{V}}$. As before, the categorical proportions in $P_{\boldsymbol{V}}$ are collected in the parameter vector $\boldsymbol{\theta}_{\boldsymbol{V}}$, on which we will conduct inference. By standard Bayesian Dirichlet-multinomial conjugacy results, a uniform prior $\boldsymbol{\theta}_{\boldsymbol{V}} \sim \text{Dirichlet}(\mathbf{1}_K)$ yields a posterior shaped by the prevalence of unique rows in the observed dataset, $(\boldsymbol{\theta}_{\boldsymbol{V}} \mid \mathbb{V}) \sim \text{Dirichlet}\left(\left[1 + \sum_{i=1}^N \mathbf{1}\{\boldsymbol{V}_i = \boldsymbol{v}_k\}\right]\right)$. For example, consider the all-binary confounding example where $\boldsymbol{V} = [D, Y]^\top$ and we wish to bound the ATE, $\varphi = \mathbb{E}[Y(d=1) - Y(d=0)]$. We possess a dataset $\mathbb{V} = \{[D_i, Y_i]^\top\}_{i=1}^N$ containing $N = n_{00} + n_{01} + n_{10} + n_{11}$ samples (rows) from P_{DY}, where n_{dy} is the number of times the row containing $[d, y] \in \{0,1\}^2$ appears in the dataset $\mathbb{V}$. We impose a uniform prior $\boldsymbol{\theta}_{DY} \sim \text{Dirichlet}([1,1,1,1]^\top)$ which yields the posterior $(\boldsymbol{\theta}_{DY} \mid \mathbb{V}) \sim \text{Dirichlet}([n_{00}+1, n_{01}+1, n_{10}+1, n_{11}+1]^\top)$.

For the quantity of interest φ, our goal is to sample from the posterior for its lower and upper bounds, respectively denoted $\underline{\varphi}$ and $\overline{\varphi}$. To preview the proposed method, each sample of the bounds will be constructed as follows: we will sample one possible observed data law that is consistent with the available data, $(\boldsymbol{\theta}^*_{\boldsymbol{V}} \mid \mathbb{V}) \sim \text{Dirichlet}\left(\left[1 + \sum_{i=1}^N \mathbf{1}\{\boldsymbol{V}_i = \boldsymbol{v}_k\}\right]\right)$, then transform it into one sample of the bounds under this sampled observed data law, the interval $\left[\underline{\varphi}^*, \overline{\varphi}^*\right] = \left[\underline{\mathsf{A}}_\varphi(\boldsymbol{\theta}^*_{\boldsymbol{V}}), \overline{\mathsf{A}}_\varphi(\boldsymbol{\theta}^*_{\boldsymbol{V}})\right]$, using `autobounds`.

This sampling procedure is justified because the posterior distribution over the bounds, conditional on the data, is given by

$$\begin{aligned} p\left(\underline{\varphi}, \overline{\varphi} \mid \mathbb{V}\right) &= \int p\left(\underline{\varphi}, \overline{\varphi} \mid \boldsymbol{\theta}_{\boldsymbol{V}}, \mathbb{V}\right)\, p(\boldsymbol{\theta}_{\boldsymbol{V}} \mid \mathbb{V})\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{V}} \\ &= \int p\left(\underline{\varphi}, \overline{\varphi} \mid \boldsymbol{\theta}_{\boldsymbol{V}}\right) p(\boldsymbol{\theta}_{\boldsymbol{V}} \mid \mathbb{V})\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{V}} \\ &= \int \delta\left(\underline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}_{\boldsymbol{V}}) - \underline{\varphi}\right) \delta\left(\overline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}_{\boldsymbol{V}}) - \overline{\varphi}\right)\, p(\boldsymbol{\theta}_{\boldsymbol{V}} \mid \mathbb{V})\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{V}}. \end{aligned} \tag{13}$$

Here, $\delta(x)$ is a Dirac delta distribution that places a unit point mass at $x = 0$. The first line follows from the law of total probability, marginalizing over uncertainty in the observed data law $\boldsymbol{\theta}_{\boldsymbol{V}}$. The second line utilizes the sufficiency of the category proportions $\boldsymbol{\theta}_{\boldsymbol{V}}$, a set of summary statistics that fully capture all information in the raw data $\mathbb{V}$. The third line then uses the fact that `autobounds` is a deterministic procedure: for any given $\boldsymbol{\theta}_{\boldsymbol{V}}$, the bounds $(\underline{\varphi}, \overline{\varphi})$ are fully determined via the functions $\underline{\mathsf{A}}_{\varphi}$ and $\overline{\mathsf{A}}_{\varphi}$. Consequently, the integral in (13) is simply the pushforward measure of the posterior $p(\boldsymbol{\theta}_{\boldsymbol{V}} \mid \mathbb{V})$ under the map

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

This structure enables straightforward Monte Carlo approximation of the integral in (13): for each posterior sample $s \in \{1, \ldots, S\}$ we draw $\boldsymbol{\theta}_{\boldsymbol{V}}^{(s)} \sim p(\boldsymbol{\theta}_{\boldsymbol{V}} \mid \mathbb{V})$, and for each sample compute the bounds

$$\underline{\varphi}^{(s)} = \underline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}_{\boldsymbol{V}}^{(s)}) \quad \text{and} \quad \overline{\varphi}^{(s)} = \overline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}_{\boldsymbol{V}}^{(s)}).$$

By the law of large numbers, the empirical distribution of these samples approximates the posterior over the bounds. In particular, the $\frac{\alpha}{2}$- and $(1-\frac{\alpha}{2})$-quantiles of the sampled $\{\underline{\varphi}^{(s)}\}_{s=1}^{S}$ and $\{\overline{\varphi}^{(s)}\}_{s=1}^{S}$, respectively, yield credible intervals that possess asymptotic frequentist guarantees on, e.g., coverage rates for the population interval.

C.2.2 Alternative Approaches to Uncertainty Quantification

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. Some prior work has sought to impose a "margin condition" that assumes the full data law being studied does not happen to lie near the points where the bounds are non-differentiable (Levis et al., 2023). This assumption is essentially one of convenience; it is difficult to empirically verify or justify from first principles. When it does not hold, previous analyses have shown that the estimated bounds $\underline{\mathsf{A}}_{\varphi}(\hat{P}_{\boldsymbol{V}})$ and $\overline{\mathsf{A}}_{\varphi}(\hat{P}_{\boldsymbol{V}})$ can underestimate (overestimate) the true bounds in finite samples of arbitrary size (Manski and Pepper, 2000, 2009). Finally, while resampling techniques could be used to approximate the bounds' joint asymptotic distribution, the aforementioned lack of regularity complicates standard methods to compute this approximation and precludes the direct use of methods like the nonparametric bootstrap (Shao, 1994; Andrews, 1999; Andrews and Han, 2009; Bugni, 2010; Canay, 2010). It is worth noting that there are several frequentist alternatives to performing inference over bounds (Chernozhukov et al., 2013; Kennedy et al., 2019; Ji et al., 2023).

C.3 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. (2023)—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. We also extend our quasi-Bayesian approach to this setting as well. 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 this quasi-Bayesian 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. At each distinct value of $\boldsymbol{x}$ observed in the data, we will construct a posterior over the conditional category proportions, $\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}$. We then follow the general quasi-Bayesian approach described in Appendix C.2.1 by repeating it within $\boldsymbol{x}$ values, yielding conditional bounds on the local quantities $\varphi_{\boldsymbol{x}}$. Finally, the overall posterior for bounds on the aggregate quantity of interest φ will be obtained by reaggregating the posteriors for these conditional bounds.

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

$$\theta_{\boldsymbol{v}_k|\boldsymbol{x}} = P_{\boldsymbol{V}\boldsymbol{X}}(\boldsymbol{V} = \boldsymbol{v}_k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_k^\top \phi(\boldsymbol{x}))}{\sum_{k'=1}^K \exp(\boldsymbol{\beta}_{k'}^\top \phi(\boldsymbol{x}))} = \text{softmax}(\boldsymbol{\beta}^\top \phi(\boldsymbol{x}))_k \tag{14}$$

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. There are a number of techniques for sampling from the posterior of these parameters, $p(\boldsymbol{\beta} \mid \mathbb{V}, \mathbb{X})$, which has no closed form; for simplicity, our implementation is based on the widely used Laplace approximation, though the proposed method easily accommodates other approaches such as Markov chain Monte Carlo sampling from the multinomial logistic regression posterior.

Recall that our target of inference is bounds on the marginal estimand $\varphi = \mathbb{E}_{\boldsymbol{X}}[\varphi_{\boldsymbol{X}}]$. We now illustrate how the previous construction yields a simple four-stage sampling algorithm from the posterior $p(\underline{\varphi}, \overline{\varphi} \mid \mathbb{V}, \mathbb{X})$. Each sample $s \in \{1, \ldots, S\}$ from the marginal bounds posterior, is constructed by first taking a sample $\boldsymbol{\beta}^{(s)}$ from the multinomial logistic regression posterior; second, applying the deterministic transformation in (14) to obtain $\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}}$; third, conducting the deterministic `autobounds` calculation yielding $\underline{\varphi}^{(s)}_{\boldsymbol{x}} = \underline{\mathsf{A}}_\varphi(\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}})$ and $\overline{\varphi}^{(s)}_{\boldsymbol{x}} = \overline{\mathsf{A}}_\varphi(\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}})$; and finally, averaging these bounds across the empirical distribution of $\boldsymbol{X}$, giving $\underline{\varphi}^{(s)} = \sum_{i=1}^N \underline{\varphi}^{(s)}_{\boldsymbol{x}_i}$ and $\overline{\varphi}^{(s)} = \sum_{i=1}^N \overline{\varphi}^{(s)}_{\boldsymbol{x}_i}$. This last step follows because φ is the expectation of $\varphi_{\boldsymbol{X}}$. This four-step procedure is justified because the posterior distribution over the marginal

bounds is

$$
\begin{aligned}
&p\left(\underline{\varphi}, \overline{\varphi} \mid \mathbb{V}, \mathbb{X}\right) \\
&= \int p\left(\underline{\varphi}, \overline{\varphi} \mid \boldsymbol{\beta}, \mathbb{V}, \mathbb{X}\right)\, p(\boldsymbol{\beta} \mid \mathbb{V}, \mathbb{X})\, \mathrm{d}\boldsymbol{\beta} \\
&= \int \delta\left(\underline{\varphi} - \int \underline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}) p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\right) \delta\left(\overline{\varphi} - \int \overline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}}) p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\right) p(\boldsymbol{\beta} \mid \mathbb{V}, \mathbb{X})\, \mathrm{d}\boldsymbol{\beta}
\end{aligned}
$$

where $\boldsymbol{\theta}_{\boldsymbol{V}|\boldsymbol{x}} = \mathrm{softmax}(\boldsymbol{\beta}^\top \phi(\boldsymbol{x}))$. The first equality follows by law of total probability, and the second follows from the fact that $\underline{\varphi}, \overline{\varphi}$ is a composition of deterministic functions of $\boldsymbol{\beta}$, as described above. We encapsulate this procedure in Algorithm 1.

Algorithm 1 Quasi-Bayesian Covariate Adjustment via Laplace Approximation

Input: Estimand φ, Data $\{\mathbb{V}_{\boldsymbol{x}_i}\}_{i=1}^N$, Resample count M, Confidence level α
Output: Partial identification interval $[\underline{\varphi}, \overline{\varphi}]$ with $(1-\alpha) \times 100\%$ coverage
Estimate posterior mode and covariance, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$, via Laplace's approximation of $p(\boldsymbol{\beta} \mid \mathbb{V}, \mathbb{X})$
for $s = 1$ **to** S **do**
  Sample $\boldsymbol{\beta}^{(s)} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$
  for $i = 1$ **to** N **do**
    Compute $\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}_i} \leftarrow \mathrm{softmax}(\boldsymbol{\beta}^{(s)\top} \phi(\boldsymbol{x}_i))$
    Assign $\underline{\varphi}^{(s)}_{\boldsymbol{x}_i} \leftarrow \underline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}_i})$, $\quad \overline{\varphi}^{(s)}_{\boldsymbol{x}_i} \leftarrow \overline{\mathsf{A}}_{\varphi}(\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}_i})$
  end
  $\underline{\varphi}^{(s)} \leftarrow \frac{1}{N} \sum_{i=1}^N \underline{\varphi}^{(s)}_{\boldsymbol{x}_i}$, $\quad \overline{\varphi}^{(s)} \leftarrow \frac{1}{N} \sum_{i=1}^N \overline{\varphi}^{(s)}_{\boldsymbol{x}_i}$
end
$\underline{\varphi} \leftarrow \mathrm{quantile}_{\alpha/2}(\{\underline{\varphi}^{(s)}\}_{s=1}^S)$ and $\overline{\varphi} \leftarrow \mathrm{quantile}_{1-\alpha/2}(\{\overline{\varphi}^{(s)}\}_{s=1}^S)$

First, approximate the posterior over the multinomial logistic regression parameters $\boldsymbol{\beta}$ using a standard Laplace approximation. Second, to account for uncertainty, draw S samples $\boldsymbol{\beta}^{(s)} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$ from this approximate posterior. For each sample and each observed covariate value $\boldsymbol{x}_i$, compute the conditional category probabilities $\boldsymbol{\theta}^{(s)}_{\boldsymbol{V}|\boldsymbol{x}_i}$ using the softmax transformation. Then, apply the `autobounds` procedure to these conditional distributions to obtain lower and upper bounds on the local estimand $\varphi_{\boldsymbol{x}_i}$. Aggregate these bounds across covariates by averaging over the observed values $\boldsymbol{x}_i$, i.e., using the empirical distribution of $\boldsymbol{X}$. Finally, return a $(1-\alpha)\%$ credible interval for the covariate-averaged partially identified estimand φ by computing the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated lower and upper bounds across all posterior samples. .

We offer two remarks on this procedure. First, in practice, we recommend using at least $S = 1000$ samples to ensure stable estimates of the bound distributions. Second, results are determined by the specific causal assumptions supplied to `autobounds` paradigm (see Section 2.2), and thus results are only guaranteed to be valid to the extent that those assumptions hold.

This combined approach provides several advantages over standard methods. The quasi-Bayesian procedure propagates estimation uncertainty in a principled manner, while the `autobounds` application ensures valid coverage even in partially identified settings. The final aggregated bounds appropriately account for observed heterogeneity and remain valid for the aggregate quantity of interest.

C.3.1 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.

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

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

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

$$P_{\boldsymbol{W}}(Y(d = 1) = 1) - P_{\boldsymbol{W}}(Y(d = 0) = 1) \tag{15}$$

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* φ *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, the second will illustrate the validity of these bounds upon averaging them. 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 φ under the `autobounds` procedure; in the ensuing exposition we focus only on the lower bound as proofs for the upper bound are analogous. Also, to complete the proof, we will rely on the graphical theory of Single World Intervention Graphs[37], or SWIGs, which combine DAGs and potential-outcomes into a single graphical framework, please see (Richardson and Robins, 2013b) for a complete description.

We begin with the following proposition.

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 SWIT (Richardson and Robins, 2013b) $\mathcal{G}(x)$ associated with $\mathcal{G}$, where the fixed node x is removed.[38] 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

[37] In their original paper, Richardson and Robins make a distinction between a SWIG and a SWIT, where a SWIG is a SWI(Template) instatiated at a specific numerical intervention.

[38] Richardson and Robins (2013b) do not advise removing fixed nodes from a SWIT because the map from a DAG to any of its SWITs is not injective in general. That is, a given SWIT does not imply a unique DAG as provenance. However, all relevant Markovian and graphical properties of the SWIT remain even without keeping the fixed nodes. As the `autobounds` paradigm requires the user to fix the DAG before doing any computation, this concern is alleviated.

$\mathbf{t}, \mathbf{t}'$ do not fix x. By Theorem 1 of Duarte et al. (2023), 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. (2023) 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. (2023) 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 validity of the bounds.

Proposition 3. *Suppose there exists an X satisfying the conditions of Proposition 2. 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 valid bounds for $\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}})$.

Proof. By Proposition 2, we know that $\left[\underline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x}), \overline{\mathsf{A}}_{\varphi^{\mathbf{t},\mathbf{t}'}}(P_{\boldsymbol{V}|x})\right]$ is an interval which contains $\varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|x})$. As P_X is positive everywhere, the integral operator with respect to P_X is monotonic in its arguments. Since $\int \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}|x}) \, \mathrm{d}P_X = \varphi^{\mathbf{t},\mathbf{t}'}(P_{\boldsymbol{W}})$ by assumption, the result follows. □

Proposition 3 tells that averaging the covariate specific bounds for each x produces valid bounds for the marginal estimand. To be clear, though, this result does not say that the averaged bounds are sharp; the technical details of such a result are beyond the scope of this paper and will be left to future work. However, as the covariate-specific bounds are sharp at each x, this is the best one can achieve in the meantime, and provides adequate justification for the quasi-Bayesian procedure presented above.

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.[39] Consistent with the original work,

[39] 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 $\tilde{X}$[40]. Throughout our analysis, hamlets are only compared if they exactly match on all coarsened $\tilde{X}$ values.

E Problem Formulations

E.1 Instrumental Variables

We will primarily examine the simple IV case of Figure 2(a). To simplify the problem, we transformed Kocher et al. (2011)'s measures of local control from a 5-point scale to binary variables, coded as indicators for being above or below the median. We will show how both the ATE and the LATE can be formulated as polynomial objective functions, using the simple IV case. The problem formulations for the more complex multi-instrument Figure 2(c–d) cases are similar; the approach generalizes straightforwardly to the multi-instrument setting, though we note that in these problems, the mere definition of a "complier" is more complex and researchers rarely state estimands in terms of local effects among the various complier groups. Therefore, in the replication of Kocher et al. (2011), we target only the ATE.

In the simple IV graph, the disturbance U_Z is responsible for determining only the response function for Z. Because the structural equation for Z has no inputs besides the disturbance U_Z, supplying a disturbance value will deterministically produce $Z = 0$ or $Z = 1$. For ease of reference, we will refer to these strata as Z-ctl or Z-enc. The disturbance U_{DY} determines

[40]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.

the response functions for both D and Y jointly. We will first define the possible response functions. For ease of reference, we use "V-always" and "V-never" (where V stands in for either D or Y) to respectively refer to response functions $f_V^{(U_{DY}=u_{DY}^{\text{i}})}(a) = 1$ and $f_V^{(U_{DY}=u_{DY}^{\text{ii}})}(a) = 0$, i.e. the individual types unaffected by treatment. Similarly, "V-comply" and "V-defy" respectively refer to response functions $f_V^{(U_{DY}=u_{DY}^{\text{iii}})}(a) = a$ and $f_V^{(U_{DY}=u_{DY}^{\text{iv}})}(a) = 1-a$, i.e. the individual types whose behavior is affected by treatment. For both D and Y, we find four individual types. Thus, U_{DY} can produce $4 \times 4 = 16$ possible joint response functions for D and Y: (D-never, Y-never), (D-never, Y-defy), (D-comply, Y-always), and so on. For a more detailed example of this procedure, please review Appendix A.

Through this principal stratification procedure, we can now parametrize our model by $2+16 = 18$ quantities, given by the strata probabilities $\Pr(Z\text{-type})$, $\Pr(D\text{-type}, Y\text{-type}')$ for the types outlined in the previous paragraph. We remark that, as Duarte et al. (2023) notes, when parametrizing the problem two parameters can be immediately eliminated due to axiomatic probability constraints. That is, $\Pr(Z\text{-ctl}) + \Pr(Z\text{-enc}) = 1$ and $\sum \Pr(D\text{-type}, Y\text{-type}') = 1$, meaning that one $\Pr(Z\text{-type})$ and one $\Pr(D\text{-type}, Y\text{-type})$ are surplus to requirements. For purposes of exposition, we will retain these superfluous parameters rather than eliminating them.

Case 2(a): Simple IV

Structural Eq.	Response Func.	Response Form
$Z = f_Z(U_Z)$	$f_Z^{(U_Z=u_Z)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(Z, U_{DY})$	$f_D^{(U_{DY}=u_{DY})}(z)$	$\{0,1\} \mapsto \{0,1\}$
$Y = f_Y(d, U_{DY})$	$f_Y^{(U_{DY}=u_{DY})}(d)$	$\{0,1\} \mapsto \{0,1\}$

Case 2(b): No exclusion restriction

Structural Eq.	Response Func.	Response Form
$Z = f_Z(U_Z)$	$f_Z^{(U_Z=u_Z)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(Z, U_{DY})$	$f_D^{(U_{DY}=u_{DY})}(z)$	$\{0,1\} \mapsto \{0,1\}$
$Y = f_Y(d, Z, U_{DY})$	$f_Y^{(U_{DY}=u_{DY})}(d,z)$	$\{0,1\}^2 \mapsto \{0,1\}$

Case 2(c): Kocher et al. (2011)

Structural Eq.	Response Func.	Response Form
$Z_7 = f_{Z_7}(U_{Z_7})$	$f_{Z_7}^{(U_{Z_7}=u_{Z_7})}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Z_8 = f_{Z_8}(Z_7, U_{Z_8})$	$f_{Z_8}^{(U_{Z_8}=u_{Z_8})}(z_7)$	$\{0,1\} \mapsto \{0,1\}$
$X_9 = f_{X_9}(Z_8, U_{X_9})$	$f_{X_9}^{(U_{X_9}=u_{X_9})}(z_8)$	$\{0,1\} \mapsto \{0,1\}$
$D_9 = f_{D_9}(Z_7, Z_8, U_{DY})$	$f_{D_9}^{(U_{DY}=u_{DY})}(z_7, z_8)$	$\{0,1\}^2 \mapsto \{0,1\}$
$Y_{12} = f_{Y_{12}}(X_9, D_9, U_{DY})$	$f_{Y_{12}}^{(U_{DY}=u_{DY})}(x_9, d_9)$	$\{0,1\}^2 \mapsto \{0,1\}$

Case 2(d): Kocher et al. (2011), relaxing exclusion restriction

Structural Eq.	Response Func.	Response Form
$Z_7 = f_{Z_7}(U_{Z_7})$	$f_{Z_7}^{(U_{Z_7}=u_{Z_7})}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Z_8 = f_{Z_8}(Z_7, U_{Z_8})$	$f_{Z_8}^{(U_{Z_8}=u_{Z_8})}(z_7)$	$\{0,1\} \mapsto \{0,1\}$
$X_9 = f_{X_9}(Z_8, U_{X_9})$	$f_{X_9}^{(U_{X_9}=u_{X_9})}(z_8)$	$\{0,1\} \mapsto \{0,1\}$
$D_9 = f_{D_9}(Z_7, Z_8, U_{DY})$	$f_{D_9}^{(U_{DY}=u_{DY})}(z_7, z_8)$	$\{0,1\}^2 \mapsto \{0,1\}$
$Y_{12} = f_{Y_{12}}(Z_8, X_9, D_9, U_{DY})$	$f_{Y_{12}}^{(U_{DY}=u_{DY})}(x_9, d_9)$	$\{0,1\}^3 \mapsto \{0,1\}$

Estimand Polynomialization

In the main text, we considered two estimands for this problem, the ATE, τ_{ATE}, and the LATE, τ_{LATE}. By Proposition 1 we have that

$$
\begin{aligned}
\tau_{\text{ATE}} &= \mathbb{E}[Y(d=1) - Y(d=0)] \\
&= \Pr(Y(d=1)=1) - \Pr(Y(d=0)=1) \\
&= \sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1}\left\{ f_Y^{(U_{DY}=u_{DY})}(d=1) = 1 \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY}) \\
&\quad - \sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1}\left\{ f_Y^{(U_{DY}=u_{DY})}(d=0) = 1 \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY})
\end{aligned}
$$

Notice that u_Z varies in only one term independently of all others in the sum, therefore we can factor it out. Since $\sum_{u_Z} \Pr(U_Z = u_Z) = 1$, we can drop it from the sum entirely. Interestingly, this is an example of a general phenomenon that if the downstream effect of the exogenous disturbance is blocked from the variables in the query by the intervention, then its parametrization contributes nothing to the polynomialization, see Proposition 3 in Duarte et al. (2023). Thus

$$\begin{aligned}\tau_{\text{ATE}} = &\sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1}\left\{f_Y^{(U_{DY}=u_{DY})}(d=1) = 1\right\} \cdot \Pr(U_{DY} = u_{DY}) \\ &- \sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1}\left\{f_Y^{(U_{DY}=u_{DY})}(d=0) = 1\right\} \cdot \Pr(U_{DY} = u_{DY})\end{aligned}$$

In terms of strata frequencies (recall that since D and Y are confounded their types manifest jointly), this is

$$\begin{aligned}&= \sum_{\text{type}} \Pr(D\text{-type}, Y\text{-comply}) + \Pr(D\text{-type}, Y\text{-always}) \\ &\quad - \sum_{\text{type}} \Pr(D\text{-type}, Y\text{-defy}) + \Pr(D\text{-type}, Y\text{-always}) \\ &= \sum_{\text{type}} \Pr(D\text{-type}, Y\text{-comply}) - \Pr(D\text{-type}, Y\text{-defy})\end{aligned}$$

which becomes our polynomial objective function.

The second estimand we consider, τ_{LATE}, is more complex because it involves conditional probabilities, which manifest as rational functions. We can address this issue with a sly trick,

however, which we demonstrate below.

$$
\begin{aligned}
\tau_{\text{LATE}} &= \mathbb{E}[Y(d=1) - Y(d=0) | D(z=1)=1, D(z=0)=0] \\
&= \frac{\Pr(Y(d=1)=1, D(z=1)=1, D(z=0)=0)}{\Pr(D(z=1)=1, D(z=0)=0)} \\
&\quad - \frac{\Pr(Y(d=0)=1, D(z=1)=1, D(z=0)=0)}{\Pr(D(z=1)=1, D(z=0)=0)} \\
&= \frac{\displaystyle\sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1}\left\{ \begin{array}{c} f_Y^{(U_{DY}=u_{DY})}(d=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY})}{\displaystyle\sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1}\left\{ \begin{array}{c} f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY})} \\
&\quad - \frac{\displaystyle\sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1}\left\{ \begin{array}{c} f_Y^{(U_{DY}=u_{DY})}(d=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY})}{\displaystyle\sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1}\left\{ \begin{array}{c} f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY})}
\end{aligned}
$$

Eliminating the blocked U_Z,

$$
\begin{aligned}
&= \frac{\displaystyle\sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1}\left\{ \begin{array}{c} f_Y^{(U_{DY}=u_{DY})}(d=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}{\displaystyle\sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1}\left\{ \begin{array}{c} f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})} \\
&\quad - \frac{\displaystyle\sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1}\left\{ \begin{array}{c} f_Y^{(U_{DY}=u_{DY})}(d=0)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}{\displaystyle\sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1}\left\{ \begin{array}{c} f_D^{(U_{DY}=u_{DY})}(z=1)=1, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}
\end{aligned}
$$

In terms of strata frequencies, this is

$$
\begin{aligned}
&= \frac{\Pr(D\text{-comply}, Y\text{-comply}) + \Pr(D\text{-comply}, Y\text{-always})}{\sum_{\text{type}} \Pr(D\text{-comply}, Y\text{-type})} \\
&\quad - \frac{\Pr(D\text{-comply}, Y\text{-defy}) + \Pr(D\text{-comply}, Y\text{-always})}{\sum_{\text{type}} \Pr(D\text{-comply}, Y\text{-type})} \\
&= \frac{\Pr(D\text{-comply}, Y\text{-comply}) - \Pr(D\text{-comply}, Y\text{-defy})}{\sum_{\text{type}} \Pr(D\text{-comply}, Y\text{-type})}
\end{aligned}
$$

Standard optimization software upon which `autobounds` relies does not permit rational objective functions. To circumvent this issue we simply define an auxiliary variable s equal to τ_{LATE}, where

$$
s = \frac{\Pr(D\text{-comply}, Y\text{-comply}) - \Pr(D\text{-comply}, Y\text{-defy})}{\sum_{\text{type}} \Pr(D\text{-comply}, Y\text{-type})},
$$

using which we can eliminate the fraction. Indeed, we specify the polynomial objective as: s, which is a valid polynomial in the expanded parameter space that includes s, while adding a single polynomial equality constraint that binds s to the target quantity (simply rearrange the terms in the identity $\tau_{\text{LATE}} = s$)

$$
\mathcal{C}_{\tau_{\text{LATE}}} = \left\{ \begin{array}{l}
\Pr(D\text{-comply}, Y\text{-comply}) - \Pr(D\text{-comply}, Y\text{-defy}) \\
\qquad -s \cdot \Pr(D\text{-comply}, Y\text{-never}) - s \cdot \Pr(D\text{-comply}, Y\text{-defy}) \\
\qquad\quad -s \cdot \Pr(D\text{-comply}, Y\text{-comply}) - s \cdot \Pr(D\text{-comply}, Y\text{-always}) \;=\; 0
\end{array} \right\}
$$

Constraint Polynomialization

The second stage of this process is to set the constraints which confine the possible values the estimand can take conditional on the given DGP and our assumptions. In particular, we will use the strata parameterization to polynomialize three types of information: (i) probability axioms; (ii) empirical evidence $\mathcal{E}$ that corresponds to observed quantities, and (iii) modeling assumptions $\mathcal{A}$.

The axiomatic constraints ("probabilities lie between zero and one and sum to unity") are

straightforwardly given by

$$\mathcal{C}_{\mathcal{P}} = \Big\{ \Pr(U_k = u_k) \geq 0 : u_k \in \mathcal{S}(U_k), k \in \{Z, DY\} \Big\} \\ \cap \Big\{ \textstyle\sum_{u_k \in \mathcal{S}(U_k)} \Pr(U_k = u_k) = 1 : k \in \{Z, DY\} \Big\}.$$

Equivalently,

$$\Pr(Z\text{-}\mathsf{type}) \geq 0 \quad \forall \quad \mathsf{type} \qquad\qquad \sum_{\mathsf{type}} \Pr(Z\text{-}\mathsf{type}) = 1$$

$$\Pr(D\text{-}\mathsf{type}, Y\text{-}\mathsf{type}') \geq 0 \quad \forall \quad \mathsf{type}, \mathsf{type}' \qquad\qquad \sum_{\mathsf{type},\mathsf{type}'} \Pr(D\text{-}\mathsf{type}, Y\text{-}\mathsf{type}') = 1$$

The first axiom translates into 18 inequality constraints, and the second into 2 equality constraints. We collect constraints arising from the laws of probability in $\mathcal{C}_{\mathcal{P}}$. As noted above, for simplicity of exposition, we do not exploit equality constraints to eliminate optimization variables and reduce the parameter space. However, it is important to note that doing so can speed computation dramatically.

Next, we polynomialize the empirical evidence, $\mathcal{E}$. Each piece of evidence is one of eight observed probabilities of the form $\Pr(Z = z, D = d, Y = y)$. Each piece of evidence can be polynomialized by

$$\begin{aligned} &\Pr(Z = z, D = d, Y = y) \\ &\quad = \sum_{\substack{u_Z \in \mathcal{S}(U_Z) \\ u_{DY} \in \mathcal{S}(U_{DY})}} \mathbf{1} \left\{ \begin{array}{c} f_Z^{(U_Z = u_Z)}(\varnothing) = z, \\ f_D^{(U_{DY} = u_{DY})}(z) = d, \\ f_Y^{(U_{DY} = u_{DY})}(d) = y \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{DY} = u_{DY}) \end{aligned}$$

which we collect in $\mathcal{C}_{\mathcal{E}}$ for all z, d, and y. For instance, in terms of strata frequencies, the cell $\Pr(Z = 0, D = 0, Y = 0)$ is given by

$$\Pr(Z\text{-}\mathsf{ctl}) \Big(\Pr(D\text{-}\mathsf{never}, Y\text{-}\mathsf{never}) + \Pr(D\text{-}\mathsf{never}, Y\text{-}\mathsf{comply}) + \\ \Pr(D\text{-}\mathsf{comply}, Y\text{-}\mathsf{never}) + \Pr(D\text{-}\mathsf{comply}, Y\text{-}\mathsf{comply}) \Big);$$

the remaining seven are similar.

Finally, we turn to the polynomialization of modeling assumptions $\mathcal{A}$. Formally, the monotonicity or "no defiers" assumption is that $\Pr(D(z=0)=1, D(z=1)=0)=0$, which polynomializes as

$$0 = \sum_{u_{DY} \in \mathcal{S}(U_{DY})} \mathbf{1} \left\{ \begin{array}{c} f_D^{(U_{DY}=u_{DY})}(z=1)=0, \\ f_D^{(U_{DY}=u_{DY})}(z=0)=1 \end{array} \right\} \cdot \Pr(U_{DY}=u_{DY})$$

In terms of strata frequencies, this constraint becomes

$$\mathcal{C}_{\mathcal{A}} = \left\{ \begin{array}{rcl} 0 & = & \Pr(D\text{-}\mathsf{defy}, Y\text{-}\mathsf{never}) + \Pr(D\text{-}\mathsf{defy}, Y\text{-}\mathsf{defy}) \\ & & + \Pr(D\text{-}\mathsf{defy}, Y\text{-}\mathsf{comply}) + \Pr(D\text{-}\mathsf{defy}, Y\text{-}\mathsf{always}) \end{array} \right\}.$$

We remark that, in conjunction with the first-axiom constraint (nonnegativity), the constrain in $\mathcal{A}$ implies

$$0 = \Pr(D\text{-}\mathsf{defy}, Y\text{-}\mathsf{type}) \quad \forall \quad \mathsf{type}$$

which is a literal statement of the no-defiers assumption on treatment response to encouragement, as expected.

To conclude, `autobounds` solves the problem

$$\underset{\mathsf{strata}}{\text{optimize}} \quad \tau_{\text{ATE}} \tag{16a}$$

$$\text{subject to} \quad \mathcal{C}_{\mathcal{P}} \cap \mathcal{C}_{\mathcal{E}} \cap \mathcal{C}_{\mathcal{A}} \tag{16b}$$

for the ATE, and the following for the LATE

$$\underset{\mathsf{strata}}{\text{optimize}} \quad s \tag{17a}$$

$$\text{subject to} \quad \mathcal{C}_{\tau_{\text{LATE}}} \cap \mathcal{C}_{\mathcal{P}} \cap \mathcal{C}_{\mathcal{E}} \cap \mathcal{C}_{\mathcal{A}}, \tag{17b}$$

which is expressed in Python code as follows

```
# DAG for this problem is a modified IV graph with potential
# direct effect from encouragement Z to outcome Y
gotv_problem = causalProblem(
    DAG("Z -> D, Z -> Y, D -> Y, U -> D, U -> Y", unob = "U")
)

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

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

    # give data to autobounds, prepare to compute 95% CI
    read_data(gotv_data, inference=True)

    # assume monotonicity: no "defiers" where plugging in
    # encouragement Z=z causes opposite treatment, D(z)=1-z
    p_defiers = p("D(Z=0)=1 & D(Z=1)=0")
    add_assumption(p_defiers, "==", 0)

    # restrict the scale of exclusion restriction violations
    p_excl_restr_violation = edge_is_active("Z -> Y")
    add_assumption(p_excl_restr_violation, "<=", 0.01)

    # set estimand to be LATE: local ATE of D on Y conditional
    # among "compliers" where encouragement Z=z causes same
    # treatment, D(z)=z
    set_ate(ind="D", dep="Y", cond="D(Z=0)=0 & D(Z=1)=1")

    gotv_bounds = solve(ci=True, nsamples=1000)
```

Figure 15: Code for the instrumental variable analysis of the Get-out-the-vote experiment.

We also present the code for the problem in section 3.1.2 in Figure 16.

```
# 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):

    # give data to autobounds, prepare to do
    # covariate adjustment and compute 95% CI
    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 16: Code for the reanalysis of Kocher et al. (2011) with instrumental variables.

E.2 Difference in Differences

E.2.1 Problem Formulation

Difference-in-differences is a strategy that involves at least three variables: pre-treatment outcome Y_0, treatment D, and post-treatment outcome Y_1. Our estimand is the average treatment effect on the treated (ATT)[41] in the post period:

$$\mathbb{E}[Y_1(d=1)|D=1] - \mathbb{E}[Y_1(d=0)|D=1]. \tag{18}$$

For simplicity we will first work with the minimal model in Figure 5(a), in which the initial outcome Y_0 is confounded with both treatment D and the subsequent outcome Y_1, but

[41] By consistency, $E[Y_1(d=1)|D=1] = E[Y_1|D=1]$, so the first element of the estimand is immediately identifiable. However, the second is not.

does not have a direct causal effect. More complex causal models in which the pre-treatment outcome has various effects on subsequent variables, such as those depicted in Figure 5(b), are straightforward to accommodate with `autobounds`. We will further demonstrate how relaxations of the DID framework that use bracketing trends based on some auxiliary variable—a framework that includes results such as monotone trends (discussed below; Hasegawa et al., 2019; Ye et al., 2024)—are easily accommodated by `autobounds`. An example DGP for this approach is given in Figure 5(c). In this subsection, for brevity, we will employ Proposition 1 to create the polynomial expressions but we will not simplify them in terms of strata frequencies. For a fully simplified example of the polynomial procedure please refer back to the IV section above. Finally, the code which implements this problem in Python is shown in Figure 17.

Case 5(a): Simple DID

Structural Eq.	Response Func.	Response Form
$Y_0 = f_{Y_0}(U)$	$f_{Y_0}^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(U)$	$f_D^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Y_1 = f_{Y_1}(D,U)$	$f_{Y_1}^{(U=u)}(d)$	$\{0,1\} \mapsto \{0,1\}$

Case 5(b): DID with Effects of Pre-treatment Outcome

Structural Eq.	Response Func.	Response Form
$Y_0 = f_{Y_0}(U)$	$f_{Y_0}^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(Y_0,U)$	$f_D^{(U=u)}(y_0)$	$\{0,1\} \mapsto \{0,1\}$
$Y_1 = f_{Y_1}(Y_0,D,U)$	$f_{Y_1}^{(U=u)}(y_0,d)$	$\{0,1\}^2 \mapsto \{0,1\}$

Case 5(c): Bracketing Trends in X

Structural Eq.	Response Func.	Response Form
$X = f_X(U)$	$f_X^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Y_0 = f_{Y_0}(U,X)$	$f_{Y_0}^{(U=u)}(x)$	$\{0,1\} \mapsto \{0,1\}$
$D = f_D(Y_0,U)$	$f_D^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Y_1 = f_{Y_1}(X,D,U)$	$f_{Y_1}^{(U=u)}(x,d)$	$\{0,1\}^2 \mapsto \{0,1\}$

Estimand Polynomialization

We polynomialize the estimand

$$\begin{aligned}
\tau_{\text{ATT}} &= \mathbb{E}[Y_1(d=1)|D=1] - \mathbb{E}[Y_1(d=0)|D=1] \\
&= \frac{\Pr(Y_1(d=1)=1, D=1)}{\Pr(D=1)} - \frac{\Pr(Y_1(d=0)=1, D=1)}{\Pr(D=1)} \\
&= \frac{\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_{Y_1}^{(U=u)}(d=1)=1, f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u)}{\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u)} \\
&\quad - \frac{\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_{Y_1}^{(U=u)}(d=0)=1, f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u)}{\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u)}
\end{aligned}$$

but we note that this is a rational expression and thus cannot be inserted directly into the optimization routine. To handle this, we reuse the auxiliary variable trick from the LATE setting in IV, and define a new parameter s such that $s = \tau_{\text{ATT}}$. Rearranging the above expression we have the estimand constraint

$$\begin{aligned}
C_{\tau_{\text{ATT}}} = \Big\{ & s\cdot\textstyle\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u) \\
&= \textstyle\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_{Y_1}^{(U=u)}(d=1)=1, f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u) \\
&\quad - \textstyle\sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{f_{Y_1}^{(U=u)}(d=0)=1, f_D^{(U=u)}(\varnothing)=1\right\}\cdot\Pr(U=u)\Big\}.
\end{aligned}$$

Constraint Polynomialization

First, constraints arising from the axioms of probability are now standard

$$\mathcal{C}_{\mathcal{P}} = \Big\{\Pr(U=u)\geq 0 : u\in\mathcal{S}(U)\Big\} \cap \Big\{\textstyle\sum_{u\in\mathcal{S}(U)}\Pr(U=u)=1\Big\}.$$

Next, constraints arising from empirical evidence in the simple DID case of Figure 5(a) are

$$\mathcal{C}_{\mathcal{E}} = \left\{\Pr(Y_0=y_0, D=d, Y_1=y_1) = \sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{\begin{array}{c} f_{Y_0}^{(U=u)}(\varnothing)=y_0, \\ f_D^{(U=u)}(\varnothing)=d, \\ f_{Y_1}^{(U=u)}(d)=y_1 \end{array}\right\}\Pr(U=u)\right\}.$$

We omit the extensions to Figure 5(b–c) cases, which are largely identical.

Finally, we consider the assumption set the research might employ to narrow the bounds, $\mathcal{C}_{\mathcal{A}}$. We disucss two here. The first is the ubiquitous identification strategy for difference-in-difference analyses is the use of the parallel trends assumption, which states that on average, the post-treatment potential outcome under control, $Y_1(d = 0)$, will differ from the pre-treatment outcome, Y_0, by exactly the same amount for both treated and control groups. That is,

$$\mathbb{E}\left[Y_1(d = 0) - Y_0 | D = 1\right] = \mathbb{E}\left[Y_1(d = 0) - Y_0 | D = 0\right].$$

This is straightofrward to polynomialize. Indeed, parallel trends translates to

$$\begin{aligned}
0 = &\Pr(Y_1(d = 0) = 1 | D = 1) - \Pr(Y_0 = 1 | D = 1) \\
&- \Pr(Y_1(d = 0) = 1 | D = 0) + \Pr(Y_0 = 1 | D = 0) \\
= &\frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_1}^{(U=u)}(d = 0) = 1, f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U = u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U = u)} \\
&- \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_0}^{(U=u)}(\varnothing) = 1, f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U = u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U = u)} \\
&+ \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_1}^{(U=u)}(d = 0) = 1, f_D^{(U=u)}(\varnothing) = 0 \right\} \cdot \Pr(U = u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing) = 0 \right\} \cdot \Pr(U = u)} \\
&- \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_0}^{(U=u)}(\varnothing) = 1, f_D^{(U=u)}(\varnothing) = 0 \right\} \cdot \Pr(U = u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing) = 0 \right\} \cdot \Pr(U = u)}
\end{aligned}$$

which creates a polynomial-fractional constraint with differing denominators, in which the fraction can be cleared by further algebraic manipulation (for each fraction one auxiliary variable, say s_0 and s_1, can be defined and manipulated as previously seen for rational functions; we would need a further constraint that $s_1 - s_0 = 0$.)

An alternative strategy is to divide the control group into two groups on some variable, $X = 0$ and $X = 1$, creating two observed trends in their outcomes. The treatment group's

trend is then assumed to be *bracketed* by the trends in the two control groups (Hasegawa et al., 2019; Ye et al., 2024). That is,

$$\mathbb{E}[Y_1(d=0) - Y_0 | D=0, X=0] \;\leq\; \mathbb{E}[Y_1(d=0) - Y_0 | D=1] \;\leq\; \mathbb{E}[Y_1(d=0) - Y_0 | D=0, X=1]$$

The monotone trends approach of Hasegawa et al. (2019) is a special case of the bracketing approach in which $X = \mathbf{1}\{Y_0 \geq \theta\}$ for some threshold θ.[42] In this case, the key bracketing assumption may be justified if the outcome follows a "rich-get-richer" pattern, such as an exponential growth model.

These inequalities,

$$\mathbb{E}\left[Y_1(D=0) - Y_0 | D=0, X=x\right] \bigstar \mathbb{E}\left[Y_1(D=0) - Y_0 | D=1\right],$$

(where $\star$ is $\leq$ for $x=0$ and $\geq$ for $x=1$) then translate to

$$\begin{aligned}
0 \bigstar\; & \Pr[Y_1(D=0, X=x)=1 | D=1] - \Pr[Y_0=1 | D=1] \\
& - \Pr[Y_1(D=0, X=x)=1 | D=0, X=x] + \Pr[Y_0=1 | D=0, X=x] \\
\bigstar\; & \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_1}^{(U=u)}(d=0,x)=1, f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)} \\
& - \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_0}^{(U=u)}(\varnothing)=1, f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)} \\
& + \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_1}^{(U=u)}(d=0,x)=1, f_D^{(U=u)}(\varnothing)=0, f_X^{(U=u)}(\varnothing)=x \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing)=0, f_X^{(U=u)}(\varnothing)=x \right\} \cdot \Pr(U=u)} \\
& - \frac{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_{Y_0}^{(U=u)}(\varnothing)=1, f_D^{(U=u)}(\varnothing)=0, f_X^{(U=u)}(\varnothing)=x \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ f_D^{(U=u)}(\varnothing)=0, f_X^{(U=u)}(\varnothing)=x \right\} \cdot \Pr(U=u)}
\end{aligned}$$

where methods we have already illustrated to address rational functions would apply to instantiate these constraints in practice. The code for instantiating and solving this problem in

[42] We note that this approach behaves poorly with binary data, as X then perfectly separates units with $Y_0 = 0$ and $Y_0 = 1$ for $\theta \in (0, 1)$.

```
# 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)
```

Figure 17: Code for the analysis of the Peruvian Civil War problem with DiD.

`autobounds` is given in Figure 17.

E.3 Selection Bias

In this subsection we outline the formulation of the polynomial program for the case of selection bias studied in section 3.3. For brevity, we make some omissions of standard steps in the procedure which are explained fully in section E.1. Please see Figure 18 for the code to implement the selection bias problem in `autobounds`.

E.3.1 Problem Formulation

Structural Eq.	Response Func.	Response Form
$D = f_D(U_D)$	$f_D^{(U_D=u_D)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$M = f_M(D, U_{MY})$	$f_M^{(U_{MY}=u_{MY})}(d)$	$\{0,1\} \mapsto \{0,1\}$
$Y = f_Y(d, M, U_{MY})$	$f_Y^{(U_{MY}=u_{MY})}(d,m)$	$\{0,1\}^2 \mapsto \{0,1\}$

Estimand Polynomialization

The primary objective function is the ATE among detained individuals,

$$\begin{aligned}\text{ATE}_{M=1} &= \mathbb{E}[Y(d=1) - Y(d=0)|M=1] \\ &= \mathbb{E}[Y(d=1, M(d=1)) - Y(d=0, M(d=0))|M=1].\end{aligned}$$

Consider the term $\mathbb{E}[Y(d, M(d)) \mid M=1]$ for any d. We may write this as

$$\begin{aligned}\mathbb{E}[Y(d, M(d)) \mid M=1] &= \Pr(Y(d, M(d)=1 \mid M=1) \\ &= \sum_{d'} \Pr(Y(d, M(d)) = 1 \mid D=d', M(d')=1) \Pr(D=d' \mid M=1) \\ &= \sum_{d',m} \Big\{ \Pr(Y(d, M(d)) = 1 \mid D=d', M(d')=1, M(1-d')=m) \\ &\qquad \Pr(M(1-d')=m \mid D=d', M(d')=1) \Pr(D=d' \mid M=1) \Big\} \\ &= \sum_{d',m} \Big\{ \frac{\Pr(Y(d, M(d)) = 1, D=d', M(d')=1, M(1-d')=m)}{\Pr(D=d', M(d')=1)} \\ &\qquad \frac{\Pr(M(1-d')=m, D=d', M(d')=1)}{\Pr(D=d', M(d')=1)} \frac{\Pr(M(d')=1, D=d')}{\Pr(M=1)} \Big\} \\ &= \sum_{d',m} \Big\{ \frac{\Pr(Y(d, M(d)) = 1, D=d', M(d')=1, M(1-d')=m)}{\Pr(D=d', M(d')=1)} \\ &\qquad \frac{\Pr(M(1-d')=m, D=d', M(d')=1)}{\Pr(M=1)} \Big\}.\end{aligned}$$

For example, when $d=1$, the above reduces to

$$\begin{aligned}&\frac{\Pr(Y(d=1, m=1)=1, D=0, M(d=)=1, M(d=1)=1)}{\Pr(D=0, M(d=0)=1)} \frac{\Pr(M(d=1)=1, D=0, M(d=0)=1)}{\Pr(M=1)} \\ +&\frac{\Pr(Y(d=1, m=1)=1, D=1, M(d=1)=1, M(d=0)=0)}{\Pr(D=1, M(d=1)=1)} \frac{\Pr(M(d=0)=0, D=1, M(d=1)=1)}{\Pr(M=1)} \\ +&\frac{\Pr(Y(d=1, m=1)=1, D=1, M(d=1)=1, M(d=0)=1)}{\Pr(D=1, M(d=1)=1)} \frac{\Pr(M(d=0)=1, D=1, M(d=1)=1)}{\Pr(M=1)}\end{aligned}$$

since the $Y(d', m)$ term is eliminated when $d'=0$ and $m=0$. It is now straightforward to polynomialize these terms; we omit the detailed calculations for brevity. The $d=0$ case is

handled similarly.

Constraint Polynomialization

Probability constraints in $\mathcal{C}_{\mathcal{P}}$ are standard. The observed data consists only of $\Pr(D = d, Y = y|M = 1)$. Therefore, constraints in $\mathcal{C}_{\mathcal{E}}$ take the form for each d and y

$$\Pr(D = d, Y = y|M = 1) = \frac{\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbf{1}\left\{ \begin{array}{l} f_Y^{(U_{MY}=u_{MY})}(d, m = 1) = y, \\ f_M^{(U_{MY}=u_{MY})}(d) = 1, \\ f_D^{(U_D=u_D)}(\varnothing) = d \end{array} \right\} \cdot \Pr(U_{MY} = u_{MY})}{\sum_{\substack{d' \in \mathcal{S}(D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbf{1}\left\{ f_M^{(U_{MY}=u_{MY})}(d') = 1 \right\} \cdot \Pr(U_{MY} = u_{MY})}$$

Next we compile the constraints from expert assumptions, $\mathcal{C}_{\mathcal{A}}$.

First, assumption 3.3.1 (mediator monotonicity) which states that $\Pr[M(d = 1) = 0, M(d = 0) = 1] = 0$, or

$$0 = \sum_{u_{MY} \in \mathcal{S}(U_{MY})} \mathbf{1}\left\{ \begin{array}{l} f_M^{(U_{MY}=u_{MY})}(d = 1) = 0, \\ f_M^{(U_{MY}=u_{MY})}(d = 0) = 1 \end{array} \right\} \cdot \Pr(U_{MY} = u_{MY})$$

for each d.

Second, assumption 3.3.2 (relative non-severity of racial stops) which is given by $\Pr[Y(d, m) = 1|D = d', M(1) = 1, M(0) = 1] \geq \Pr[Y(d, m) = 1|D = d', M(1) = 1, M(0) = 0]$. We translate

this as

$$0 \geq \frac{\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbf{1}\left\{ \begin{array}{l} f_Y^{(U_{MY}=u_{MY})}(d,m) = 1, \\ f_D^{(U_D=u_D)}(\varnothing) = d', \\ f_M^{(U_{MY}=u_{MY})}(d=1) = 1, \\ f_M^{(U_{MY}=u_{MY})}(d=0) = 0 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}{\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbf{1}\left\{ \begin{array}{l} f_D^{(U_D=u_D)}(\varnothing) = d', \\ f_M^{(U_{MY}=u_{MY})}(d=1) = 1, \\ f_M^{(U_{MY}=u_{MY})}(d=0) = 0 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}$$

$$- \frac{\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbf{1}\left\{ \begin{array}{l} f_Y^{(U_{MY}=u_{MY})}(d,m) = 1, \\ f_D^{(U_D=u_D)}(\varnothing) = d', \\ f_M^{(U_{MY}=u_{MY})}(d=1) = 1, \\ f_M^{(U_{MY}=u_{MY})}(d=0) = 1 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}{\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbf{1}\left\{ \begin{array}{l} f_D^{(U_D=u_D)}(\varnothing) = d' \\ f_M^{(U_{MY}=u_{MY})}(d=1) = 1, \\ f_M^{(U_{MY}=u_{MY})}(d=0) = 1 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})}$$

where the complaint is duplicated for each (d, d', m) triple.

```
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 18: Code for the study of bias in police use of force.

E.4 Mediation

E.4.1 Formal Background

Here we provide a formal description of the estimands involved in standard mediation designs and the technical challenges for their identification. For the sake of simplicity, we assume the treatment is binary and all other variables are discrete. In the pursuit of quantifying causal mechanisms, scholars have aimed to identify three key quantities: the controlled direct effect (CDE), the natural indirect effect (NIE) and the natural direct effect (NDE), these are respectively

$$\begin{aligned}
\text{CDE}(m) &:= \quad \mathbb{E}[Y(d=1,m) - Y(d=0,m)], \\
\text{NIE}(d) &:= \quad \mathbb{E}[Y(d, M(d=1)) - Y(d, M(d=0))], \\
\text{NDE}(d) &:= \quad \mathbb{E}[Y(d=1, M(d)) - Y(d=0, M(d))].
\end{aligned}$$

We summarize these estimands in turn. Firstly, the CDE measures the effect of the treatment on the outcome when the mediator is held constant at a specific value. In effect, this quantity allows researchers to isolate the direct effect of the treatment, unaffected by changes in the mediator, that is, the pathway $D \to M \to Y$ is blocked by intervening on M (Pearl et al., 2001; Robins et al., 2003)—it is a measure of the interaction of the treatment and the mediator on the outcome. Secondly, the NIE quantifies the portion of the treatment effect that is explained by the mediator. To see this, note that NIE(d) is non-zero whenever a non-trivial causal effect of the treatment on the mediator, that is $M(1) \neq M(0)$, causes a change in the outcome. Holding treatment level fixed, this is exactly the pathway $D \to M \to Y$ in Figure 8. This concept is of great theoretical interest for understanding the role of the mediator in transmitting the treatment effect, as the mediator is not manipulated but naturally experienced. This intuition elucidates the contrast between the NIE and the CDE: it is generally understood that the NIE better explains causal mechanisms because the mediator is permitted to vary naturally in this estimand, however, if one were to aiming to understand interventional policy, the CDE is preferred, since the mediator is fixed by the researcher (Pearl, 2012; Acharya et al., 2016). Thirdly, the NDE captures the portion of the treatment effect that is not mediated by the mediator, representing the direct pathway ($D \to Y$) from the treatment to the outcome, this

readily understood by the noticing that M is held fixed in both counterfactual terms.[43] This concept is important for understanding the direct influence of the treatment, separate from its indirect effects (Robins and Greenland, 1992). Finally, notice that the global ATE is related to the mediated effects through

$$\text{ATE} = \text{NIE}(1) + \text{NDE}(0).$$

That is, the total effect of treatment can be additively decomposed into its direct and indirect effects with respect to a given mediator, M; this decomposition makes clear why the mediated effects are the ideal target for explaining causal relationships between two variables.

In an experimental setting, if one can randomize D and M, the CDE is clearly identified. However, the NIE and NDE are far more difficult to capture, even with randomization. This is because there is no experiment in which one can design a manipulation to generate $Y(d, M(d'))$ where $d \neq d'$. We need stricter assumptions than those guaranteed by design. Most commonly, mediation designs produce valid inferences under four assumptions which the treatment, mediator and outcome must satisfy; these are: (i) no treatment-outcome confounding; (ii) no treatment-mediator confounding; (iii) no (pre-treatment) mediator-outcome confounding, and (iv) the absence of alternative mediators which confound the mediator and outcome relationship. If one can believe that the treatment was successfully randomized in the sample or one possesses a sufficiently rich set of pre-treatment covariates, (i) and (ii) are satisfied. Assumptions (iii) and (iv) are the most pernicious, particularly because they can be easily violated even in an experimental setting. Indeed, in the social sciences, the mediator is often a belief or emotion whose manipulation might be randomized but whose measurement is post-treatment. Due to the ephemeral nature of these objects, it makes difficult both their accurate measurement, and the prevention of alternative explanations for (in)direct effects induced by the treatment. Even if these alternate mediators are measured, identification remains a challenge (Avin et al., 2005; Robins et al., 2022; Tchetgen Tchetgen and VanderWeele, 2014; Vansteelandt and Daniel, 2017). In the mediation literature, assumptions

[43] Note that the ability to compute this quantity does not exclude the possibility of other mediators lying on the direct path $D \to Y$.

(i)-(iv) are together referred to as *sequential ignorability*. This is because these assumptions are equivalent to believing that *following* randomization of treatment, the mediator is effectively randomly assigned within each treatment group (VanderWeele and Vansteelandt, 2009; Imai et al., 2010).

E.4.2 On the Manipulation Exclusion

The parallel design aims to permit the researcher to identify causal mechanisms in a milder setting than that prescribed by sequential ignorability (Imai et al., 2010, 2011). However, one assumption of this design is quite opaque in its precise description, and so we offer a softer explanation of its importance. To do so, we first specify the causal model governing the parallel design

$$E = f_E(U_E), \quad M = f_M(D, E, U_{MY}),$$
$$D = f_D(U_D), \quad Y = f_M(D, E, M, U_{MY}).$$

When conducting the parallel design, we generate models for the natural arm

$$\begin{aligned} &E = 0, \quad M(d, e = 0) = f_M(d, e = 0, U_{MY}), \\ &D = d, \quad Y(d, M(d, e = 0), e = 0) = f_Y(d, M(d, e = 0), e = 0, U_{MY}), \end{aligned} \tag{19}$$

and the manipulated arm

$$\begin{aligned} &E = 1, \quad M = m, \\ &D = d, \quad Y(d, m, e = 1) = f_Y(d, m, e = 1, U_{MY}). \end{aligned} \tag{20}$$

The manipulation exclusion assumption implies that whenever $M(d, e = 0)$ in the natural model takes on the value m to which M is set in the manipulated model, $Y(d, M(d, e = 0), e = 0) = Y(d, m, e = 0) = Y(d, m, e = 1)$. That is, $Y(d, m, e)$ is constant in e. As the name suggests, this is a type of exclusion restriction, the like of which was already presented in section 3.1 on instrumental variables. This assumption suppresses the **direct** causal effect of E on Y; the experimental manipulation may only affect Y through its effect on M. Permitting E to only indirectly affect Y makes clear why assumption 3.4.1 is effectively equivalent to

ensuring that participants are unaware of their being manipulated. Indeed, this assumption encodes that the participant's decision is the same whether they infer that $M = m$ naturally or are explicitly told as such. For example, in the study by Tomz and Weeks (2013), this assumption might be violated if the mediator manipulation in some manner suggested which countries were democracies or autocracies. In this way, participants may then erroneously infer that the study itself is about the threat that these specifc nations pose to the US or Britain and therefore put increased weight on the threat cue in this arm but would ignore this context in arms where this information is obscured. Clearly, this assumption is crucial, since if it were violated, groups in the natural and manipulated arms are not comparable: the hidden mechanism by which the participants interpret the mediator is known to be different between arms in such a case.

With this assumption in hand, the natural model becomes

$$\begin{aligned} E = 0, \quad & M(d, e = 0) = f_M(d, e = 0, U_{MY}) \\ D = d, \quad & Y(d, M(d, e = 0)) = f_Y(d, M(d, e = 0), U_{MY}), \end{aligned} \tag{21}$$

while the manipulated model is

$$\begin{aligned} E = 1, \quad & M = m \\ D = d, \quad & Y(d, m) = f_Y(d, m, U_{MY}), \end{aligned} \tag{22}$$

and thus we need only deal with counterfactuals of the form $Y(d, m)$.

E.4.3 Eliminated Effects

As discussed in the main text, Acharya et al. (2018) advocates the use of the eliminated effect to quantify the effects of causal mechanisms. This is largely due to the fact that the designs permits the identification of its two components, the ATE and the CDE. Indeed, the eliminated effect is defined as

$$\text{EE}(m) = \text{ATE} - \text{CDE}(m). \tag{23}$$

Clearly, by randomization of D in the natural arm the first term is identified, and the randomization of D and M in the manipulated arm identifies the second term. To explain why a researcher might be interested in this estimand, the authors provide the following decomposition of the eliminated effect

$$\text{EE}(m) = \text{NIE}(1) + \text{RI}(m). \tag{24}$$

where $\text{RI}(m) = \text{NDE}(0) - \text{CDE}(m)$. This result follows immediately by noting that ATE = NIE(1) + NDE(0) as discussed in the previous section. The presence of mediated effects masked beneath the EE is now obvious, however, we give some intuition as to why one might find this estimand useful.

We will ground this exposition by the Tomz and Weeks (2013) setting. The NIE quantifies the portion of the effect of regime type on support for a preemptive strike that is explained by changes in how threatening the country is perceived to be. It isolates the indirect pathway where the regime type influences the perception of threat, and this perception, in turn, influences the level of support for a preemptive strike. The reference interaction is the difference between the direct effect of treatment with the mediator at its natural value under an autocratic regime and the direct effect of the treatment with the mediator fixed at level m. Noting that the NDE is a weighted average of the CDEs over the natural level of the mediator[44], the RI explains the amount of variation in the CDE due to natural variation in the mediator (at some baseline m). It is the part of the total effect due to interaction alone (Acharya et al., 2018). To illustrate this, return to the setting of Tomz and Weeks (2013): if we set $m = 0$, so the hypothetical country in question is known to be nonthreatening, the RI is the difference between the direct effect of democracy on support for attack under *inferred* threat and the same direct effect if the country is *known* to be nonthreatening. This difference is thus the part of the direct effect explained by how much the treatment and mediator manifest in tandem: indeed, if there is no such interaction, then the NDE is equal to the CDE so the RI vanishes, however, if $\text{RI}(0) < 0$ (say), then we can conclude that a decrease in support for a

[44] $\text{NDE}(d) = \sum_{m \in \mathcal{S}(M)} \mathbb{E}[Y(1,m) - Y(0,m) \mid M(d) = m] \Pr(M(d) = m)$

preemptive strike on a democratic nation is smaller under inferred threat, than under known threat.

Eliminated effects are easy targets since they are identified under mild assumptions. However, the question which most frequently drives scientific inquiry is the NIE, which is masked by this estimand. Acharya et al. (2018) argue that a large EE is indicative of a strong mediating factor, however, the exact manner in which it operates remains unknown. For example, the replication study finds a large and positive EE which suggests that support for an attack on a democracy is significantly mediated by the belief that the country is a threat. However, it is unclear how one might reconcile this result with the original finding in Tomz and Weeks (2013), which reports a negative indirect effect of democracy on support for a strike. Indeed, Acharya et al. (2018) proposes two reasons for this discrepancy, both of which stem from speculation on the magnitude and sign of the NIE or RI. We summarise these explanantions here. One explanation is that the reference interaction could be large and positive. This implies that, relative to a direct effect under a fixed threatening country baseline (which is negative), the direct effect of democracy (as opposed to autocracy) on support for an attack under inferred threat is lower in magnitude (that is, less negative). This seems unlikely, since it would suggest that the participants assumed a threat more significant than the one provided to them, which could occur if the participants were somehow primed to believe that the hypothetical country in question directly threatens America, but no such prime was administered. Acharya et al. alternatively suggest that due to the original study having employed a standard mediation analysis assuming sequential ignorability, their estimate is biased, and so the indirect effect could be nonnegative.

In the main text, we resolve this discrepancy through `autobounds`, which shows that under the mild sub-interaction assumption, the second explanation prevails. In particular, by directly targeting the desired mediated effects, our partial identification strategy yields informative bounds on the question of interest.

E.4.4 Problem Formulation

We build the polynomial program for the mediation problem as for all other problem types. For a more detailed explanation of how to build these polynomials see the IV example in Appendix E.1. The code for this example is in Figure 19. The DGP in which we wish to do inference is given by Figure 9(a). Explicitly, we have

Structural Eq.	Response Func.	Response Form
$E = f_E(U_E)$	$f_E^{(U_E=u_E)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(U_D)$	$f_D^{(U_D=u_D)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$M = f_M(D, E, U_{MY})$	$f_M^{(U=u)}(d, e)$	$\{0,1\}^2 \mapsto \{0,1\}$
$Y = f_Y(d, M, U_{MY})$	$f_Y^{(U=u)}(d, m)$	$\{0,1\}^2 \mapsto \{0,1\}$

Estimand Polynomialization

Here, we will show the problem formulation for the NIE, the NDE is similar. Indeed the objective function of the problem is

$$\text{NIE}(1) = \mathbb{E}[Y(d=1, M(d=1)) - Y(d=1, M(d=0))],$$

We now write this estimand in terms of the disturbances. The first term is

$$\begin{aligned}\Pr(Y(d=1)=1) &= \sum_{m \in \mathcal{S}(M)} \Pr(Y(d=1, M(d=1))=1, M(d=1)=m) \\ &= \sum_{m \in \mathcal{S}(M)} \sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ \begin{array}{c} f_M^{(U=u)}(1,m)=1, \\ f_Y^{(U=u)}(1)=m \end{array} \right\} \Pr(U=u).\end{aligned}$$

while the second is

$$\begin{aligned}\Pr(Y(1, M(0))=1) &= \sum_{m \in \mathcal{S}(M)} \Pr(Y(1, M(0))=1, M(0)=m) \\ &= \sum_{m \in \mathcal{S}(M)} \sum_{u \in \mathcal{S}(U)} \mathbf{1}\left\{ \begin{array}{c} f_M^{(U=u)}(1,m)=1, \\ f_Y^{(U=u)}(0)=m \end{array} \right\} \Pr(U=u).\end{aligned}$$

Constraint Polynomialization

The constraints, $\mathcal{C}_{\mathcal{P}}$, due to probability axioms are standard. Consider next the observed data, $\mathcal{C}_{\mathcal{E}}$. The observed data takes the form $\Pr(Y = y, D = d \mid E = 0)$ and $\Pr(Y = y, D = d, M = 1 \mid E = 1)$. To constrain our strata frequencies by these quantities observe that by randomization of E and assumption 3.4.2

$$\Pr(Y(d) = y) = \Pr(Y = y \mid D = d, E = 0),$$

therefore, expanding the LHS

$$\Pr(Y = y \mid D = d, E = 0) = \sum_{m \in \mathcal{S}(M)} \sum_{u_D \in \mathcal{S}(U_D)} \mathbf{1} \left\{ \begin{array}{c} f_M^{(U=u)}(d) = m \\ f_Y^{(U=u)}(d, m) = y \end{array} \right\} \Pr(U = u).$$

Similarly, by randomization of E and assumption 3.4.2 again

$$\Pr(Y(d, m) = y) = \Pr(Y = y \mid D = d, M = m, E = 1),$$

and so by expanding the LHS

$$\Pr(Y = y \mid D = d, M = 1, E = 1) = \sum_{u \in \mathcal{S}(U)} \mathbf{1} \left\{ f_Y^{(U=u)}(d, m = 1) = y \right\} \Pr(U = u).$$

Finally, we create the assumption set $\mathcal{C}_{\mathcal{A}}$. We add the sub-interaction assumption ψ_Y as a constraint. Since

$$\Pr(Y(1,0) = y_{10}, Y(0,0) = y_{00}, Y(1,1) = y_{11}, Y(0,1) = y_{01}) = \sum_{u \in \mathcal{S}(U)} \mathbf{1} \left\{ \begin{array}{l} f_Y^{(U=u)}(1,0) = y_{10}, \\ f_Y^{(U=u)}(0,0) = y_{00}, \\ f_Y^{(U=u)}(1,1) = y_{11}, \\ f_Y^{(U=u)}(0,1) = y_{01}, \end{array} \right\} \Pr(U = u),$$

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

# load data with D/M/Y columns, one row per respondent
dempeace_data = pandas.read_csv("dempeace_data.csv")

# data_do: a function which creates counterfactual probs from expt data, see notebooks
dempeace_doD = data_do(dempeace_data[dempeace_data["E"]==0], ["D"])
dempeace_doDM = data_do(dempeace_data[dempeace_data["E"]==1], ["D", "M"])

with respect_to(med_problem):
    # add experimental data as constraints
    for i in range(len(dempeace_doD)): # loop through rows of experimental summary
    D_val = dempeace_doD.loc[i, "D"] # identify D val
    Y_val = dempeace_doD.loc[i, "Y"] # identfiy Y val
    prob_val = dempeace_doD.loc[i, "prob"] # empirical probability

    add_assumption(
        p(f"Y(D={D_val})={Y_val}"), "==", prob_val
    )

    for i in range(len(dempeace_doDM)):
        D_val = dempeace_doDM.loc[i, "D"]
        M_val = dempeace_doDM.loc[i, "M"]
        Y_val = dempeace_doDM.loc[i, "Y"]
        prob_val = dempeace_doDM.loc[i, "prob"]

        add_assumption(
            p(f"Y(D={D_val},M={M_val})={Y_val}"), "==", prob_val
        )

    # gamma is the proportion of people whose direct effect would be
    # larger when the threat is diminished,
    # expect to be small or even 0, see main text and notebooks
    add_subinteraction(strength=gamma)

    # NIE
    nie_plus = E("Y(D=1)")
    nie_minus = p("Y(D=1,M=1)=1 & M(D=0)=1") + p("Y(D=1,M=0)=1 & M(D=0)=0")
    nie = nie_plus - nie_minus
    set_estimand(nie)
    dempeace_bounds = solve(ci=True)
```

Figure 19: Code for the mediation analysis of the Democratic Peace survey experiment.

it follows that

$$\psi_Y = \sum_{\substack{y_{10},y_{00},y_{11},y_{01}\in\mathcal{S}(Y):\\ y_{10}-y_{00}>y_{11}-y_{01}}} \sum_{u\in\mathcal{S}(U)} \mathbf{1}\left\{\begin{array}{l} f_Y^{(U=u)}(1,0)=y_{10},\\ f_Y^{(U=u)}(0,0)=y_{00},\\ f_Y^{(U=u)}(1,1)=y_{11},\\ f_Y^{(U=u)}(0,1)=y_{01},\end{array}\right\}\Pr(U=u)$$

We then add the constraint $\psi_Y \leq \gamma_{\psi_Y}$ to the polynomial program for user-selected γ_{ψ_Y}.

E.4.5 Inference in the Parallel Design

In appendix C.2, we illustrate how to compute inference on the bounds $(\underline{\varphi}, \overline{\varphi})$ of estimand φ when provided with the observed data law $P_{\boldsymbol{V}}$ parameterized by $\boldsymbol{\theta}_{\boldsymbol{V}}$. The procedure adopts a quasi-Bayesian framework, whereby we impose a Dirichlet prior on $\boldsymbol{\theta}_{\boldsymbol{V}}$, obtain its posterior using the observed data to produce the likelihood, and then by sampling from the posterior and applying `autobounds`, we obtain a distribution over the upper and lower bounds of the estimand, whence we can compute confidence intervals.

Parallel designs, as explained in section 3.4, may not provide the analyst with empirical samples of $P_{\boldsymbol{V}}$. In the case of Acharya et al. (2018), the analyst would only obtain samples from conditional margins of $P_{\boldsymbol{V}}$, in particular, (i) $P_{\boldsymbol{V}}(D = d, Y = y \mid E = 0)$ from the natural arm, and (ii) $P_{\boldsymbol{V}}(D = d, Y = y \mid M = 1, E = 1)$ from the manipulated arm. The natural arm only produces (i) because the authors chose not to measure M in this arm, while the manipulated arm we only obtain (ii) because the authors only set $M = 1$ in their survey roll-out. If one is interested only in EE(1), the authors' implementation is acceptable, since these data identify that quantity. However, if one wishes to target mediated effects, two challenges arise. First, the absence of data on any marginal or conditional distribution of M makes it imprudent to place priors on $P_{\boldsymbol{V}}(d, m, y)$: neither this distribution nor any parametrization of it is fully identified by the collected data, meaning that large regions of the posterior would be driven entirely by the prior chosen. In order to use the transparent parametrization of Richardson et al. (2011), we must adapt the procedure in section C.2. Secondly, many cells of $P_{\boldsymbol{V}}$ are missing in this design, making it difficult to obtain informative bounds on the estimand without additional assumptions. Consequently, estimated and confidence bounds are wider than those that would be obtained if complete samples of $P_{\boldsymbol{V}}$ were available. In the survey setting of Acharya et al. (2018), it would be inexpensive to modify the natural-arm design to measure M and the mediator-arm design to randomize M, making complete samples from $P_{\boldsymbol{V}}$ available. In future work, we recommend that analysts collect data on all variables and conditions in each arm to avoid this problem.

As discussed above, $\boldsymbol{\theta}_{\boldsymbol{V}}$ is not available for this design, we do not supply it as an argument

to `autobounds`. Rather, by using the following facts from Appendix E.4.4,

$$P_{\boldsymbol{W}}(Y(d) = y) = P_{\boldsymbol{V}}(Y = y \mid D = d, E = 0) \tag{25}$$

and

$$P_{\boldsymbol{W}}(Y(d, m = 1) = y) = P_{\boldsymbol{V}}(Y = y \mid D = d, M = 1, E = 1), \tag{26}$$

we can manually constrain the strata which combine to produce φ via the equations above. Now, let $\boldsymbol{\theta}_{\text{nat}}$ and $\boldsymbol{\theta}_{\text{man}}$ contain the four parameters on the RHS of (25) and the four parameters on the RHS of (26) respectively.

Since the bounds on the estimand (NIE or NDE) φ are computed as $[\underline{\mathsf{A}}(\boldsymbol{\theta}_{\text{nat}}, \boldsymbol{\theta}_{\text{man}}), \overline{\mathsf{A}}(\boldsymbol{\theta}_{\text{nat}}, \boldsymbol{\theta}_{\text{man}})]$, we simply apply the quasi-Bayesian framework outlined in appendix C.2 applied to inputs $\boldsymbol{\theta}_{\text{nat}}$ and $\boldsymbol{\theta}_{\text{man}}$. That is: (1) place a uniform Dirichlet prior on $\boldsymbol{\theta}_{\text{nat}}$ and a separate uniform Dirichlet prior on $\boldsymbol{\theta}_{\text{man}}$; (2) sample from their posteriors; (3) for each pair of samples $(\boldsymbol{\theta}^*_{\text{nat}}, \boldsymbol{\theta}^*_{\text{man}})$ from their posteriors, compute the posterior bounds sample $(\underline{\varphi}^*, \overline{\varphi}^*)$; and, (4), obtain $100(1-\alpha)\%$ confidence intervals by taking quantiles of the resulting distribution over the bounds. We are justified in treating the two parameters independently because of the design. Randomization on E partitions the sample into two independent groups, each governed by a distinct causal model (see Figure 9). In particular, in the $E = 1$ arm, the mediator M is set by the researcher, severing backdoor paths through unmeasured confounders U. This effectively removes U from the outcome model in that arm, preventing any dependence between $\boldsymbol{\theta}_{\text{nat}}$ and $\boldsymbol{\theta}_{\text{man}}$. See Figure 11 for an illustration of the results.