Potential Outcomes vs. Graphical Causality
Part I: Encoding Shape Constraints

Written by: D. Plecko, A. Maiti, E. Bareinboim

Background: *A longstanding debate between the proponents of the potential outcomes (PO) and the graphical models (GM) approach to causal inference concerns the identification of causal effects under shape constraints (such as monotonicity). Scholars in the PO framework have developed seminal results leveraging monotonicity constrains in practical applications, such as in identification of local average treatment effects (LATE). Various assertions have been made in the PO literature that the graphical approach may be inherently limited for encoding shape constraints. The GM approach has successfully provided complete algorithms for non-parametric identification of causal effects. Some special cases, such as linear or additive-noise models, are also well understood within this approach. However, incorporating shape constraints such as monotonicity, has so far eluded its scope. In this blogpost, we discuss a recent advance which allows one to encode monotonicity constraints in a graphical model, and leverage monotonicity for identification. We also provide background for readers to understand the core methodological questions at hand.*

1. Local Average Treatment Effects

A seminal result of Imbens and Angrist (1994) leverages monotonicity in the setting of instrumental variables. Consider the causal diagram in Fig. 1, where $Z \in \{0,1\}$ represents an invitation to a job training program, $X \in \{0,1\}$ represents training program participation, and Y represents job promotion. The local average treatment effect (LATE) is defined as:

$$\text{LATE} = \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X_{z_0} = 0, X_{z_1} = 1]. \tag{1}$$

Here, Y_x represents the potential outcome under the hypothetical intervention that sets $X = x$, possibly contrary to the fact (this potential outcome is sometimes also abbreviated by $Y(x)$). The LATE measures the causal effect of X on Y within the group of units who "comply" with the treatment Z (i.e., respond to the training program invitation), meaning that $X = 0$ when $Z = 0$ and $X = 1$ when $Z = 1$. This is encoded in the conditioning event $X_{z_0} = 0, X_{z_1} = 1$.

In the general non-parametric setting, the LATE cannot be identified (uniquely computed) from observational data. However, under the monotonicity assumption of $Z \to X$, that is if for all individuals $\mathbf{u}$[1] the setting $Z = z_1$ produces a greater or equal outcome in X than $Z = z_0$, written

$$X_{z_1}(\mathbf{u}) \geq X_{z_0}(\mathbf{u}) \tag{2}$$

for all assignments $\mathbf{u}$ of unobserved variables, then the quantity can in fact be identified from observational data (Imbens and Angrist, 1994).

We provide a sketch of the proof, to highlight the key steps of the derivation. Note that the term $\mathbb{E}[Y_x \mid X_{z_0} = 0, X_{z_1} = 1]$ can be written, in the case of binary Y, as:

$$\mathbb{E}[Y_x \mid X_{z_0} = 0, X_{z_1} = 1] = \frac{P(Y_x = 1, X_{z_0} = 0, X_{z_1} = 1)}{P(X_{z_0} = 0, X_{z_1} = 1)}. \tag{3}$$

[1] Here, $\mathbf{u}$ denotes a fixed set of exogenous (noise) variables, and corresponds to a single individual observed in the data.

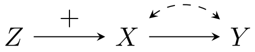


Figure 1: LATE IV setting.

The key implication of the monotonicity constraint in Eq. 2 is that the probability of the event $X_{z_0} = 1, X_{z_1} = 0$ equals 0, written as $P(X_{z_0} = 1, X_{z_1} = 0) = 0$. Note that we can write

$$P(X_{z_0} = 0, X_{z_1} = 1) = \underbrace{P(X_{z_0} = 0, X_{z_1} = 1) + P(X_{z_0} = 1, X_{z_1} = 1)}_{P(X_{z_1}=1) \text{ by law of tot. prob.}} - \underbrace{P(X_{z_0} = 1, X_{z_1} = 1)}_{P(X_{z_0}=1) \text{ by monotonicity}} \tag{4}$$

$$= P(X_{z_1} = 1) - P(X_{z_0} = 1). \tag{5}$$

Using ignorability $X_z \perp\!\!\!\perp Z$ (or the rules of do-calculus in the graphical approach), we can see that $P(X_z = 1) = P(X = 1 \mid Z = z)$, which in combination with Eq. 5 identifies $P(X_{z_0} = 0, X_{z_1} = 1)$ from observational data. By applying the same re-writing trick as in Eq. 4, one can also obtain that

$$P(Y_{x_1} = 1, X_{z_0} = 0, X_{z_1} = 1) = P(Y_{x_1} = 1, X_{z_1} = 1) - P(Y_{x_1} = 1, X_{z_0} = 1) \tag{6}$$

$$= P(Y_{x_1} = 1, X_{z_1} = 1)\frac{P(Z = 1)}{P(Z = 1)} - P(Y_{x_1} = 1, X_{z_0} = 1)\frac{P(Z = 0)}{P(Z = 0)} \tag{7}$$

$$= \frac{P(Y_{x_1} = 1, X_{z_1} = 1, Z = 1)}{P(Z = 1)} - \frac{P(Y_{x_1} = 1, X_{z_0} = 1, Z = 0)}{P(Z = 0)} \tag{8}$$

$$= \frac{P(Y = 1, X = 1, Z = 1)}{P(Z = 1)} - \frac{P(Y = 1, X = 1, Z = 0)}{P(Z = 0)} \tag{9}$$

$$= P(Y = 1, X = 1 \mid Z = 1) - P(Y = 1, X = 1 \mid Z = 0). \tag{10}$$

Eq. 8 follows from $Y_x, X_z \perp\!\!\!\perp Z$, while Eq. 9 follows from the consistency axiom. The key step of the derivation is therefore in Eq. 6, in which the monotonicity constraint is leveraged.

The LATE approach has proven as very popular over the years. Interestingly, in his recent discussion of the potential outcomes vs. the graphical approach to causal inference, one of the authors of LATE, Imbens (2020) writes that "causal diagrams have difficulty coding shape restrictions such as monotonicity". While this may have been a valid criticism in the past, we next describe a recent advance which allows users to encode monotonicity constrains in a graphical model, and leverage them for effect identification.

2. A General Algorithmic Approach

The key insight of Imbens and Angrist (1994) for identifying the LATE can in fact be generalized. Let W be a variable, and $\mathbf{Y}$ a set of variables disjoint from W. Let $\mathbf{T}, \mathbf{S}$ be a partition of the parents of W, such that W is monotonic with respect to parents $\mathbf{T}$, meaning that

$$W_{\mathbf{t},\mathbf{s}}(u) \leq W_{\mathbf{t}',\mathbf{s}}(u) \quad \forall \mathbf{t} \leq \mathbf{t}', \mathbf{s}. \tag{11}$$

Let $\mathbf{Y}_*$ denote an arbitrary counterfactual of the variables $\mathbf{Y}$. If W is binary, we can introduce the following *monotonicity reduction* rules:

(1) **Simplification Rule:** If $\mathbf{t} \leq \mathbf{t}'$, then

(a) $P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 0) = P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 0)$

(b) $P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) = P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1)$

(2) **Difference Rule:** If $\mathbf{t} \leq \mathbf{t}'$, then

$$\begin{aligned} P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) &= P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 1) - P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1) \quad (12) \\ &= P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0) - P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 0). \quad (13) \end{aligned}$$

The rules can be comprehended intuitively. Suppose that $\mathbf{t}, \mathbf{t}'$ are two values of $\mathbf{T}$ such that $\mathbf{t} \leq \mathbf{t}'$. By definition of monotonicity, we have that

$$W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) \leq W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}). \quad (14)$$

Therefore, $W_{\mathbf{t}',\mathbf{s}} = 0$ implies that $W_{\mathbf{t},\mathbf{s}} = 0$ as well. Similarly, $W_{\mathbf{t},\mathbf{s}} = 1$ implies that $W_{\mathbf{t}',\mathbf{s}} = 1$. These two simple observations yield the simplification rule. This rule was used, for instance, in Eq. 4 of the LATE derivation to reduce the term $P(X_{z_0} = 1, X_{z_1} = 1)$ to $P(X_{z_0} = 1)$. Now, to derive the difference rule, we note that $P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 0) = 0$ by monotonicity. Thus, we can write

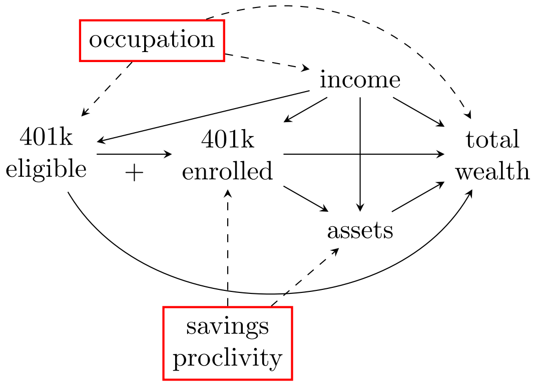
$$\begin{aligned} P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) &= P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) + P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) \quad (15) \\ &\quad - P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) \quad (16) \\ &= P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 1) - P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) \quad (17) \\ &= P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 1) - P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1), \quad (18) \end{aligned}$$

where the last line follows from the simplification rule. The difference rule was used in Eq. 6 of the LATE derivation.

Our recent work (Maiti et al., 2024) demonstrates that the monotonicity reduction rules discussed above can be used in combination with the standard machinery for counterfactual identification in the graphical approach (Shpitser and Pearl, 2007; Correa et al., 2021). In particular, the paper introduces a sound algorithm for identifying arbitrary counterfactual queries from combinations of observational and experimental data while leveraging monotonicity constraints, which are naturally encoded by labeling edges of the graph along which the effects are monotonic. We refer the reader to the paper for technical details, and next describe an example in which monotonicity can be used to extend the scope of identification results in the existing literature.

3. Example of Monotonicity Identification

We consider an example that may appeal to econometricians. We investigate the dataset studied in (Abadie, 2003), called 401k. Consider the graphical model given by:



.

Abadie (2003) studied the LATE of 401k enrollment (X) on net financial assets (M), with 401k eligibility (Z) as the instrumental variable. We consider an extended setting, in which there could possibly exist important confounding variables, including occupation type (U_1) and proclivity for savings (U_2). It is worth noting that our setting is more general than the setting considered in (Abadie, 2003), since focusing on variables Z, X, M as done in (Abadie, 2003) would not render Z a valid instrument (in other words, the set of assumption in the graph above seem to include more hidden confounding possibilities than considered in previous literature). We are interested in identifying the LATE of 401k enrollment on the total wealth (Y) in different income (W) groups. Note that eligibility (Z) cannot serve as an instrument in our setting. Our query of interest can be written as:

$$\text{LATE}(w) = \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X_{z_0} = 0, X_{z_1} = 1, W = w]. \tag{19}$$

Applying the classical non-parametric LATE estimator from Imbens and Angrist (1994), conditional on $W = w$, would yield

$$\text{classic-LATE}(w) = \frac{\mathbb{E}[Y \mid z_1, w] - \mathbb{E}[Y \mid z_0, w]}{\mathbb{E}[X \mid z_1, w] - \mathbb{E}[X \mid z_0, w]}. \tag{20}$$

This expression, however, leads to an incorrect result. Similarly, an attempt of conditioning on M before applying the classical LATE estimator, labeled cond(m)-LATE(w), given by

$$\text{cond}(m)\text{-LATE}(w) = \sum_m P(m \mid w) \frac{\mathbb{E}[Y \mid z_1, m, w] - \mathbb{E}[Y \mid z_0, m, w]}{\mathbb{E}[X \mid z_1, m, w] - \mathbb{E}[X \mid z_0, m, w]} \tag{21}$$

would also lead to an incorrect result. Therefore, remarkably, state-of-the-art methods in the econometrics literature cannot solve the problem at hand.

However, armed with a general algorithm for identification under monotonicity constraints, we can in fact identify the query of interest. We provide a derivation for obtaining an expression, together with an intuition for what allows us to identify the effect. We show how to compute $P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1, W = w)$ for the given causal graph. Once we can compute this term, the expression can be identified with relative ease. We expand $P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1, W = w)$:

$$\sum_{z,m} P(Y_{x_1 mzw} = y, M_{x_1 w} = m, X_{z_0 w} = 0, X_{z_1 w} = 1, Z_w = z, W = w) \tag{22}$$

$$\overset{(i)}{=} \sum_{z,m} P(Y_{x_1 mzw} = y, Z_w = z, W = w) P(M_{x_1 zw} = m, X_{z_0 w} = 0, X_{z_1 w} = 1) \tag{23}$$

$$\overset{(ii)}{=} \sum_{z,m} P(Y_{x_1 mzw} = y, Z_w = z, W = w) \big[P(M_{x_1 w} = m, X_{z_1 w} = 1) - P(M_{x_1 w} = m, X_{z_0 w} = 1)\big] \tag{24}$$

$$\overset{(iii)}{=} \sum_{z,m} P(Y_{x_1 mzw} = y, Z_w = z, W = w) \big[P(m, x_1 \mid z_1, w) - P(m, x_1 \mid z_0, w)\big] \tag{25}$$

$$\overset{(iv)}{=} \sum_{z,m} P(y \mid w, z, x_1, m) P(w, z) \big[P(m, x_1 \mid w, z_1) - P(m, x_1 \mid w, z_0)\big] \tag{26}$$

In the first step of the expansion, we use the law of total probability and the consistency axiom. Then, we note that M_{xw}, X_{zw} depend on the unobserved confounder U_2, while Y_{xmzw}, Z_w, W depend on the confounder U_1. Written in the language of ignorability, we have that

$$Y_{xmzw}, Z_w, W \perp\!\!\!\perp M_{xw}, X_{zw}, \tag{27}$$

which allows us to write the step (i). Then, the term $P(M_{x_1 zw} = m, X_{z_0 w} = 0, X_{z_1 w} = 1)$ can be simplified using the monotonicity reduction rules (difference rule), giving step (ii). Now, since we

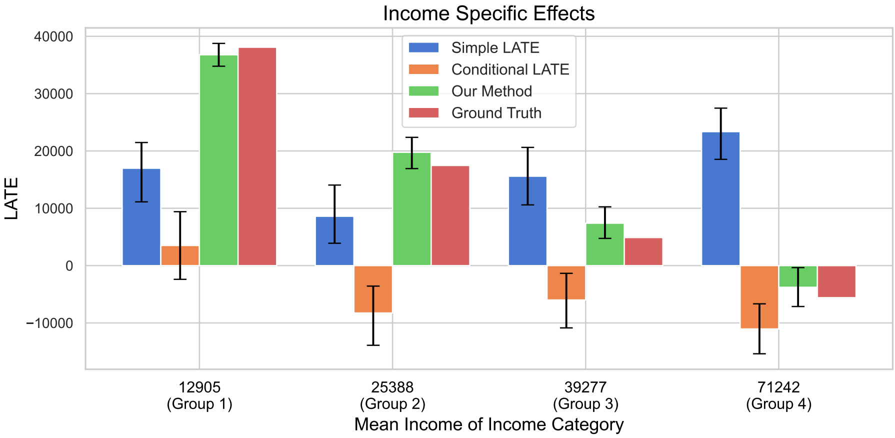


Figure 2: Experimental results on the 401k dataset studied in (Abadie, 2003).

have $M_{xw}, X_{zw} \perp\!\!\!\perp W, Z$ as mentioned already, we can also write

$$P(M_{x_1w} = m, X_{z_1w} = 1) = P(M_{x_1w} = m, X_{z_1w} = 1 \mid Z = 1, W = w) \tag{28}$$
$$= P(M = m, X = 1 \mid Z = 1, W = w) \tag{29}$$

where the second line follows by consistency. Plugging this into the derivation gives step (iii). Since $Y_{xmzw}, Z_w, W \perp\!\!\!\perp M_{xw}, X_{zw}$, we have that

$$P(Y_{x_1mzw} = y, Z_w = z, W = w) = P(Y_{x_1mzw} = y, Z = z, W = w) \tag{30}$$
$$= P(Y_{x_1m} = y \mid Z = z, W = w)P(Z = z, W = w) \tag{31}$$
$$= P(Y_{x_1m} = y \mid M = m, X = 1, Z = z, W = w) \tag{32}$$
$$* P(Z = z, W = w)$$
$$= P(Y = y \mid M = m, X = 1, Z = z, W = w)P(Z = z, W = w), \tag{33}$$

where the second to last step follows from the ignorability statement $Y_{x_1m} \perp\!\!\!\perp X, M \mid Z, W$. This gives the final step (iv) of the above derivation. Putting everything together, we can obtain the LATE identification expression which in this case equals:

$$\text{LATE}(w) = \frac{\sum_y y \cdot Q_1(w, y)}{\sum_y Q_1(w, y)} - \frac{\sum_y y \cdot Q_0(w, y)}{\sum_y Q_0(w, y)}, \tag{34}$$

where the weights $Q_i(y)$ are given by

$$Q_i(w, y) = \sum_{m,z} P(y \mid w, z, m, x_i)P(z \mid w)\big[P(m, x_i \mid w, z_i) - P(m, x_i \mid w, z_{1-i})\big]. \tag{35}$$

An important question is how to understand the above derivation intuitively. For this, we make the following observations, which provide the intuition for why the LATE can be identified:

(O1) The total effect consists of the direct effect $X \to Y$ and the indirect effect $X \to M \to Y$,

(O2) $X \to M \to Y$ requires inferring the effect $X \to M$, and $M \to Y$,

(O3) For the $X \to M$ effect, Z is a valid instrument, assuming monotonicity, (O4) When considering X, M jointly, the pair Z, W in fact gives a valid back-door set for the effect of X, M on Y, reflected in the ignorability statement $Y_{xm} \perp\!\!\!\perp X, M \mid Z, W$. This allows the identification of the effect of a joint intervention of X, M on Y.

By combining the above observations, one can intuitively understand the steps of the LATE derivation in this instance.

We conclude with an empirical analysis performed on the dataset studied by (Abadie, 2003). We compare the estimators classic-LATE(w) from Eq. 20, cond-LATE(w) from Eq. 21, and the true identification expression for LATE in Eq. 34. We discretize income, net financial assets, and total wealth into groups corresponding to quartiles. Using such discretized data, we design a synthetic SCM M that matches the observational distribution of the original data. We then generate 30,000 data points from this model to obtain estimators LATE, classic-LATE, and cond-LATE. The results are shown Fig. 2, with bootstrap-derived 95% confidence intervals indicated, together with the ground truth indicated in red. Clearly, our method is able to infer the effects correctly, while the existing methods in the literature cannot perform this task.

References

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

Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. Advances in Neural Information Processing Systems, 34:6856–6867, 2021.

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Journal of Economic Literature, 58(4):1129–1179, 2020.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. Econometrica, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/2951620`.

Aurghya Maiti, Drago Plecko, and Elias Bareinboim. Counterfactual identification under monotonicity constraints. 2024.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. In Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, UAI'07, page 352–359, Arlington, Virginia, USA, 2007. AUAI Press. ISBN 0974903930.