Omitted Labels Induce Nontransitive Paradoxes in Causality

Bijan Mazaheri*　　　　　　　　　　　　　　　　　　　　BIJAN.H.MAZAHERI@DARTMOUTH.EDU
Dartmouth Engineering

Siddharth Jain†　　　　　　　　　　　　　　　　　　　　　　　　SID496@GMAIL.COM
California Institute of Technology

Matthew Cook　　　　　　　　　　　　　　　　　　　　　　　　　M.COOK@RUG.NL
University of Groningen

Jehoshua Bruck　　　　　　　　　　　　　　　　　BRUCK@PARADISE.CALTECH.EDU
California Institute of Technology

Editors: Biwei Huang and Mathias Drton

Abstract

We explore "omitted label contexts," in which training data is limited to a subset of the possible labels. This setting is standard among specialized human experts or specific focused studies. By studying Simpson's paradox, we observe that "correct" adjustments sometimes require non-exchangeable treatment and control groups. A generalization of Simpson's paradox leads us to study networks of conclusions drawn from different contexts, within which a paradox of nontransitivity arises. We prove that the space of possible nontransitive structures in these networks exactly corresponds to structures that form from aggregating ranked-choice votes.

1. Introduction

Knowledge is powered and limited by the data that drives it. When seeking to understand the relevance of a study, the most critical aspect is its data's *context*. Two common "data contexts" are (1) the population of participants and (2) interventions made on that population. Optimizing a context for utility might involve a census of the target population with perfectly focused interventions. The real world, however, weighs utility against feasibility; budget constraints limit a study's participants, and moral constraints limit its interventions.

Domain Adaptation　　A study may attempt to transfer conclusions from a sub-optimal *data context* to a *target* one. Most broadly, this topic is known as domain adaptation (DA). DA formally deals with transferring from a training probability distribution $p(\mathbf{V})$ to a target distribution $q(\mathbf{V})$. For a specific training task, we often separate the measured variables into $\mathbf{V} = (Y, \mathbf{X})$, where Y is the "label" and $\mathbf{X}$ are the "covariates."

DA is made possible through assumptions on the shared information between p and q. Two such assumptions are *covariate shift* (Shimodaira, 2000) and *label shift* (Schweikert et al., 2008). Another setting that can be considered to fall within DA is causal inference, which seeks to identify the effects of intervening on a variable (often a treatment) without actually performing that intervention (Pearl, 2009; Imbens and Rubin, 2015; Peters et al., 2017). From a DA perspective, our data-context is "observational" and our target context is "interventional," i.e. $q(\mathbf{V}) = p(\mathbf{V} \mid \mathrm{do}(T = t))$ for $T \in \mathbf{V}$.

---

* Work completed while a PhD student at California Institute of Technology.

† Now at Ripple Medical Inc., Irvine, CA. Work completed while a postdoc at California Institute of Technology.

The best way to learn causal effects is to perform interventions in a clinical "randomized controlled trial" (RCT). RCTs utilize a random partition into treatment and control groups, which ensures that no potentially relevant factors (like the severity of a disease) can affect the probability of receiving treatment. This "exchangeability" between treatment and control forms a key principle of the interventional distributions that causal inference aims to emulate. That is, an intervened variable's value (such as $\mathrm{do}(T = t)$) contains no information about that variable's usual causes.

A common technique within DA is to transform the data's distribution to the target context through reweighting. These weights make up for poorly represented portions of a data distribution by giving them more importance. For label shift, this corresponds to weighting data by the ratio of label probabilities ($w(\mathbf{x}, y) = q(y)/p(y)$). Similarly, covariate shift can be accounted for using a ratio of covariate probability densities ($w(\mathbf{x}, y) = q(\mathbf{x})/p(\mathbf{x})$). Causal inference performs similar reweightings to transform observational data into distributions with exchangeable treatment and control groups, e.g. "inverse probability weighting" (Imbens and Rubin, 2015; Cole and Hernán, 2008; Hernán MA, 2020). Covariate adjustment techniques like the "backdoor adjustment" (Pearl, 2009) and "G-computation" (Robins et al., 2009) perform adjustments at the distribution level that can nonetheless be thought of as reweighting data.

Omitted Label Contexts This paper will discuss a relatively new branch of the DA tree introduced by Mazaheri et al. (2021), which we will call "omitted label contexts." Such settings are limited to *only a subset* $\mathcal{Y}^* \subset \mathcal{Y}$ of the labels for $Y \in \mathcal{Y}$. For example, "dogs vs. cats" is omitted label context (omitting all other animals), but "dogs vs. non-dogs" is not. While the relative probabilities of classes within this subset are maintained, data from all other labels are unobserved. More precisely, $p(y_1^*)/p(y_2^*) = q(y_1^*)/q(y_2^*)$ for $y*_1, y*_2 \in \mathcal{Y}^*$, but $p(y') = 0$ if $y' \notin \mathcal{Y}^*$. Within the scope of this paper, we will restrict our focus to $|\mathcal{Y}^*| = 2$.

Omitted label contexts are motivated by a few real-life scenarios within medicine and epidemiology. Omitted label contexts are extremely common in the study of rare conditions. For example, a census genome sequencing of the US population would be an impractical and financially infeasible task. Instead, databases like TCGA (Tomczak et al., 2015) allow focused access to patients with specific (and often rare) cancers by omitting the "healthy label". In study designs, investigators may opt for an omitted label context (e.g. breast vs lung cancer) or induce further label shift by working with a uniform distribution on the labels of interest. In these cases, it is useful to determine a notion of consistency. Though less related to the issues discussed in this paper, "immortal time bias" (Suissa, 2008) can also be thought of as an omitted label context in which patients are excluded due to death before the end of the study.

Most broadly, omitted labels are a form of sampling bias — a topic that has been studied in detail within the causal inference literature (Correa et al., 2019). Bareinboim and Tian (2015) calls a causal effect "recoverable" if it can be computed in the presence of a selection mechanism. An important difficulty within omitted label contexts is that they are "irreversible." Zero-probability labels cannot be "weighted-up" to transform the distribution to that of the general population. With respect to covariate adjustments, this makes the causal effects unrecoverable.

1.1. Summary

This paper will begin by illustrating the potentially severe consequences of omitted label bias, particularly with respect to unrecoverable causal effects. We consider what happens when these effects are ignored, demonstrating that omitted labels can give rise to reversals in calculated treatment

effects. This paradox is driven by the non-exchangeability of "correct" causal inference weights in omitted label contexts, i.e., standard adjustments to exchangeability give us incorrect quantities.

Contributions The problems that arise within omitted label contexts are unsurprising — computing (what is known to be) an unrecoverable causal effect is impossible. However, within the details of these failures emerges a structure between the conclusions of multiple studies with *different* restrictions on their labels. Specifically, a "new" [1] paradox manifests within the combination of erroneous causal conclusions. We will link this paradox to a 200-year-old observation from social choice theory known as the "Condorcet paradox," which demonstrates how ranked-choice votes (i.e., a preference ordering of the candidates for each voter) can result in a cycle of aggregate preference (Nicolas et al., 1785). In a surprising connection between social choice theory, causality, and machine learning, we prove that these phenomena are one and the same.

The task of combining conclusions from different models is sometimes referred to as decision fusion (Castanedo, 2013). The usual reasons for studying decision fusion involve lack of access to data or inability to combine datasets (e.g., because of non-overlapping covariates). As advances in large models (such as LLMs) transition us from data-based to model-based vehicles of information, studying how to fuse decisions from multiple models will be of growing importance.

Outline Section 2 begins by explaining Simpson's paradox through an example that will be expanded on throughout the paper. Section 3 studies omitted label contexts and shows how and why they can exhibit Simpson's paradox in causality. Section 4 introduces Condorcet's paradox and the linear ordering polytope and shows their relationship to networks of conclusions from different omitted label contexts. Section 5 concludes the paper and discusses its implications.

Notation In general, we will use the capital Latin alphabet (i.e., X, Y, T) to denote random variables, with Y being the "label" or class we wish to predict or determine the causal effect on, X being the covariates, and T being the treatment. The lowercase Latin alphabet will denote assignments to these variables, e.g., $x^{(1)}$ means $X = x^{(1)}$. Vectors and sets of random variables will be in bold-face font, while other types of sets will use caligraphic font (e.g., $\mathcal{Y} = \{y^{(1)}, y^{(2)}, \ldots\}$). The following notation is used throughout the paper:

- $\Pr(\cdot)$ will be used for probability. $p(\cdot)$ and $q(\cdot)$ will also be used when discussing DA.
- $\mathbf{1}^\ell$ denotes an all 1 vector of size ℓ.
- $\triangle^\ell$ denotes the ℓ-dimensional simplex. That is, $\lambda \in \triangle^\ell$ iff $\lambda \in [0, 1]^\ell$ and $\mathbf{1}^\top \lambda = 1$.
- $\prec$ and $\preceq$ denote element-wise inequality. For example, we say $\mathbf{w} \preceq \mathbf{v}$ if $w_i \leq v_i \, \forall \, i \in [\ell]$.
- We will use $\mathrm{Co}(S)$ to denote the open convex hull of S, $\overline{\mathrm{Co}}(S)$ to denote the closed convex hull, and $\mathrm{Bo}(\cdot)$ to denote the boundary.

2. Simpson's Paradox

We will ease into our dissection of errors in causal quantities by discussing Simpson's paradox. While these observations are well-studied, we present them here in a form that is generalized to multiple labels in the following section. This discussion will rely on hypothetical observational data

---

1. The paradox is very similar to the one introduced in Mazaheri et al. (2021) but has not been observed for causality of covarite shift.

T	X	$y^{(0)}$	$y^{(1)}$
$t^{(0)}$	$x^{(0)}$	3	7
$t^{(0)}$	$x^{(1)}$	1	0
$t^{(1)}$	$x^{(0)}$	0	1
$t^{(1)}$	$x^{(1)}$	7	3

T	X	$y^{(0)}$	$y^{(1)}$
$t^{(0)}$	$x^{(0)}$	3	7
$t^{(0)}$	$x^{(1)}$	**10**	**0**
$t^{(1)}$	$x^{(0)}$	**0**	**10**
$t^{(1)}$	$x^{(1)}$	7	3

Table 1: The left table shows a specification of counts for Simpson's Paradox. On the right, the 2nd and 3rd rows have been weighted up so that both the treated ($t^{(1)}$) and untreated ($t^{(0)}$) groups have a (weighted) distribution on X that matches the marginal distribution, which is uniform (all rows must sum to 10).

T	X	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$
$t^{(0)}$	$x^{(0)}$	3	7	0
$t^{(0)}$	$x^{(1)}$	1	0	99
$t^{(1)}$	$x^{(0)}$	0	1	99
$t^{(1)}$	$x^{(1)}$	7	3	0

T	X	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$
$t^{(0)}$	$x^{(0)}$	**30**	**70**	**0**
$t^{(0)}$	$x^{(1)}$	1	0	99
$t^{(1)}$	$x^{(0)}$	0	1	99
$t^{(1)}$	$x^{(1)}$	**70**	**30**	**0**

Table 2: On the left we give an augmentation of Table 1 with a third column. Now, the 1st and 4th rows must be weighted up in order to balance the distribution of X (all rows must sum to 100).

on a treatment T and its outcome Y, given in Table 1 (left). To map these quantities to a real-world example, suppose X indicates the severity of a disease. Severe cases receive treatment at higher rates (and recover at lower rates) than mild ones. In aggregate, treated patients have lower rates of improvement than untreated patients.

The driver for Simpson's paradox is the difference in severity between those who did and did not receive treatment and the effect of this severity on patient outcomes. That is to say, the treatment and control groups are not exchangeable. To remedy this, we can reweight the rows of our table to exchangeability by emphasizing the severely ill patients who did not receive treatment and the mildly ill patients who did receive treatment. This is accomplished by reweighting datapoints (t, x, y) according to the inverse probability/propensity of receiving the treatment that they got, $w(t, x, y) = 1/\Pr(t \mid x)$, sometimes referred to as "Inverse Probability Weighting" (IPW) (Imbens and Rubin, 2015). In our example, IPW weights up the second and third rows by a factor of 10, as shown in Table 1 (right). When this reweighting is interpreted as a synthetic study on 40 participants (20 treated and 20 control, each with a 10 : 10 split on severity), the new apparent treatment effect is $13/20 - 7/20 = 30\%$.

An alternative perspective is that the causal effect of the treatment lies in the outcome changes *within each severity group*. By separately considering the severe and mild patients, we can average

T	X	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$
$t^{(1)}$	$x^{(0)}$	2	1	0
$t^{(1)}$	$x^{(1)}$	0	2	1
$t^{(1)}$	$x^{(2)}$	1	0	2
$t^{(0)}$	$x^{(0)}$	0	1	2
$t^{(0)}$	$x^{(1)}$	2	0	1
$t^{(0)}$	$x^{(2)}$	1	2	0

Table 3: A specification of counts that mimics Condorcet's paradox.

outcomes according to the marginal probability distribution of severity. Following this intuition, the "backdoor adjustment" (Pearl, 2009) calculates the probability distribution of $Y = y^{(i)}$ under an intervention of $T = t^{(j)}$:

$$\Pr(y^{(i)} \mid \text{do}(t^{(j)})) := \sum_x \Pr(x) \Pr(y^{(i)} \mid x, t^{(j)}).$$

The difference between the two possible interventions gives the "average treatment effect" (ATE),

$$\text{ATE} = \Pr(y^{(1)} \mid \text{do}(t^{(1)})) - \Pr(y^{(1)} \mid \text{do}(t^{(0)})) = \frac{1/1 + 3/10}{2} - \frac{7/10 + 0/3}{2} = .3.$$

Notice that the marginal probability distribution of X is uniform, corresponding to an equal weighting of the $x^{(0)}$ and $x^{(1)}$ rows in Table 1. In fact, both IPW and backdoor approaches result in the same weightings of the rows of the table because $\Pr(t, x)/\Pr(t \mid x) = \Pr(x)$.

Simpson's paradox has been the subject of a long list of works to which it would be impossible to do full justice. Pearl (2022) and Hernán et al. (2011) describe Simpson's paradox as "solved" by causal modeling. We will focus on one key takeaway: the choice of how to re-weight sub-cases (rows of our table) plays a key role in the conclusion of a study, sometimes reversing the apparent relationship (as in Simpson's Paradox).

An important observation is that there is a geometry to the way in which these errors occur. Notice that the reversal in Table 1 would be increased by further increasing the probability of rows 1 and 4, e.g., by changing the 3, 7 counts to 300, 700. This reweighting strengthens the dependence of T on X. While we will not dive further into the geometry of Simpson's paradox, its existence motivates the structures we will study within networks of contexts.

3. Omitted Label Contexts

Now that we understand the potential effects of reweighting distributions on covariates, we will move our focus to the study of omitted label contexts. As discussed in the introduction, omitted label contexts involve the removal of some labels while preserving the relative probabilities of the non-removed labels. This removal can shift the apparent distribution of any variable that is associated with Y, including both treatment T and covariates X.

3.1. Causality within Omitted Label Contexts

Consider a second hypothetical dataset that augments Table 1 with an additional column, shown in Table 2. When $y^{(2)}$ is excluded, the observed dataset is equivalent to Table 1, which we recall has an ATE of .3 on the outcome of $y^{(1)}$ (relative to $y^{(0)}$).

Although the full context has the exact same (uniform) marginal probability distribution on X, the inclusion of the $y^{(2)}$ column now requires that we weigh up the 1st and 4th rows (rather than the 2nd and 3rd rows). See the right side of Table 2 for the resulting (weighted) counts. When we use these weights to compute the ATE on $y^{(1)}$, we see a reversal of the treatment effect:

$$\text{ATE} = \frac{1/100 + 3/10}{2} - \frac{7/10 + 0/100}{2} = -.195. \tag{1}$$

This can be understood by graphically modeling the selection bias as in Bareinboim and Tian (2015), shown in Figure 1. Figure 1 (a) shows the graph describing a confounding variable X causing
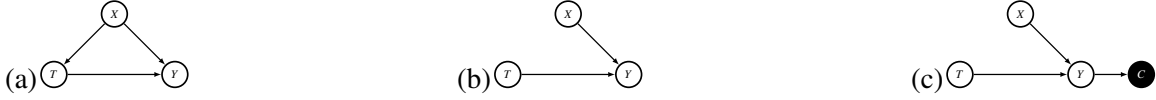
Figure 1: (a) A causal DAG depicting confounding from a common cause X. (b) The causal DAG that "severs" $X \rightarrow T$ by reweighting for exchangeability. (c) The causal DAG depicting the effect of an omitted label context C, which has been conditioned on.

both T and Y. The goal of IPW and the backdoor adjustment is to reweigh the distribution to fit the DAG in Figure 1 (b), i.e. the distribution of X is exchangeable in both $t^{(0)}$ and $t^{(1)}$ or equivalently $X \perp\!\!\!\perp T$. Figure 1 (c) shows the effect of restricting the labels of Y within a dataset context (such as with omitted label contexts), which involves conditioning on a child of Y. That is, let C be an indicator for Y's membership in the restricted set of labels and condition on $C = 1$ to induce omitted label bias. X and T are not d-separated[2] because conditioning on a variable that is causally downstream of both X and T can induce a spurious correlation between them.

This effect darkens the outlook for causal inference in omitted label contexts, particularly because the globally correct reweighting of Table 2 involves counts that do not appear exchangeable when we omit the $y^{(2)}$ column. We are no longer guided by the principle of exchangeability in our observed data. Finding the correct adjustment requires knowledge of the context-induced distribution $\Pr(X, T, Y \mid y \in \{y^{(i)}, y^{(i)}\})$. This distribution cannot be obtained without extending the study to all labels.

In response to this difficulty, we will study what can be learned when we have many *different* omitted label contexts. These contexts do not provide access to the full joint distribution that would be needed to correctly compute a backdoor adjustment. However, the aggregation of these contexts provides partial information that can be formalized.

4. Networks of Contexts

Before we discuss the structures within networks of omitted label contexts, we will introduce another paradox from social choice theory: the Condorcet Paradox (Nicolas et al., 1785). We will see that this paradox and its structure are deeply related to networks of conclusions drawn from omitted label contexts.

4.1. The Condorcet Paradox

The Condorcet paradox works as follows: three voters each have preferences $y^{(0)} \rightarrow y^{(1)} \rightarrow y^{(2)}$, $y^{(1)} \rightarrow y^{(2)} \rightarrow y^{(0)}$, and $y^{(2)} \rightarrow y^{(0)} \rightarrow y^{(1)}$, with $a \rightarrow b$ indicating a preference of a over b. These preferences are generated by rotating a starting ordering three times, meaning that $y^{(i)} \rightarrow y^{(i+1 \mod 2)}$) in two out of the three voters. The result is an aggregate cycle of preference $y^{(0)} \rightarrow y^{(1)} \rightarrow y^{(2)} \rightarrow y^{(0)}$ with frequencies of 2/3 voters for each edge.

This paradox can be generalized into what we will call an "aggregation of rankings" (AR) — a complete directed-graph[3] on the set of labels $\mathcal{Y}$ with weights on each $y^{(i)} \rightarrow y^{(j)}$ corresponding to the fraction of voters who prefer $y^{(i)}$ to $y^{(j)}$. AR structures are a convex combination of total

---

2. see Pearl (2009) for more information on d-separation.
3. These graphs are always complete, but we use graph terminology as in Mazaheri et al. (2021) in order to reference properties that are dependent on cycles.
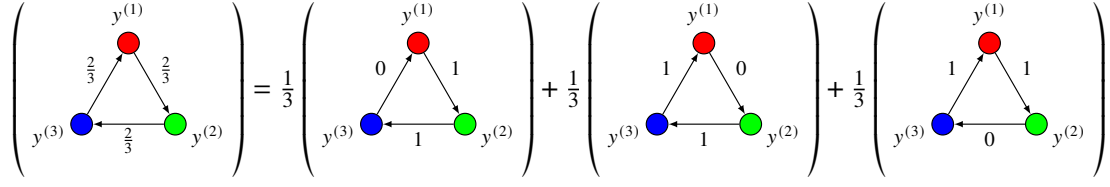
Figure 2: The Condorcet paradox as an aggregation of rankings.

orderings (i.e., graphs with edge weights of 0 or 1), with component weights corresponding to the fraction of voters carrying each total ordering. See Figure 2 for an illustration of this perspective for the Condorcet paradox. As a result, the space occupied by all possible AR structures is known as the "linear ordering polytope," which has been the subject of extensive study (Fishburn, 1992; Alon, 2002).

The preferences of voters in the Condorcet paradox can be embedded into a table of frequencies, with each voter becoming a specific value for covariate X. Table 3 demonstrates this using the counts $2 > 1 > 0$ to induce high, medium, and low preference. Notice that the order of preferences for each x in the $t^{(1)}$ half of the table (first three rows) exactly correspond to the order of preferences given by the voters in the Condorcet paradox, starting with $y^{(0)} \to y^{(1)} \to y^{(2)}$ for $x^{(0)}$ and shifting the order for the other values of X.

The $t^{(0)}$ half of the table complements the $t^{(1)}$ half so that the counts for $(t^{(0)}, x^{(i)}, y^{(j)})$ and $(t^{(1)}, x^{(i)}, y^{(j)})$ always sum to three. As a result, restricting our table to any two columns still yields a uniform probability distribution on X, i.e. $\Pr(x^{(0)} \mid y \in \{y^{(i)}, y^{(j)}\}) = \Pr(x^{(1)} \mid y \in \{y^{(i)}, y^{(j)}\}) = \Pr(x^{(2)} \mid y \in \{y^{(i)}, y^{(j)}\})$. This is the distribution that a naive study would average over when applying a backdoor adjustment, meaning that

$$
\begin{aligned}
\Pr(y^{(0)} \mid \mathrm{do}(t^{(1)}), y \in \{y^{(0)}, y^{(1)}\}) &= \frac{2/3 + 0/2 + 1/1}{3} = 5/9 \\
\Pr(y^{(0)} \mid \mathrm{do}(t^{(0)}), y \in \{y^{(0)}, y^{(1)}\}) &= \frac{0/3 + 2/2 + 1/3}{3} = 4/9.
\end{aligned}
\tag{2}
$$

The calculations in Equation 2 conclude that the ATE on $y^{(0)}$ in the $y \in \{y^{(0)}, y^1\}$ context is $+1/9$. These calculations are the same for the ATE on $y^{(1)}$ for $y \in \{y^{(1)}, y^{(2)}\}$ and the ATE on $y^{(2)}$ for $y \in \{y^{(2)}, y^{(0)}\}$ due to the cyclic shifting of columns. Hence, the studies separately conclude that the treatment increases the relative frequency of all three labels, which is clearly impossible. The embedding of the Condorcet paradox into causal conclusions implies a correspondence between aggregations of rankings and backdoor adjustments (or any other case-based weighting).

Cardinal vs Ordinal Systems The Condorcet paradox is primarily driven by the loss of information in an ordinal system. That is, an ordering of $A \to B \to C$ cannot distinguish between the magnitude of the preferences $A \to B$ and $A \to C$. If voters were instead allowed to allocate multiple votes among the candidates (e.g. one voter gives 2 votes for A, 1 for B, and 0 for C), it would be clear that $A \to C$ is a stronger opinion than $A \to B$ (that is, $2 - 1 < 2 - 0$). Indeed, such "cardinal" voting systems do not give rise to paradoxes of transitivity (Conklin and Sutherland, 1923).

Consider mapping the paradox we have described here to a weighted DAG structure. The edge-weight on $y^{(i)} \to y^{(j)}$ corresponds to a do-intervention that has been computed on a context

7

that omits all labels other than $y^{(i)}, y^{(j)}$. Since these edge-weights are constructed using ordinal values (i.e., the probability distribution on the Y), we would not expect the Condorcet paradox to emerge. Nonetheless, we will show that these "aggregations of soft rankings" occupy the same linear ordering polytope as "aggregations of rankings."

4.2. Aggregations of Rankings and Soft Rankings

Table 3 used counts of $2, 1, 0$ to induce preference between labels in each row. Because this system is cardinal, the "preference" (the relative frequency of each label) can take on any value in $[0, 1]$. For this reason, we refer to the induced preferences in each row of our tables as a "soft ranking." We will now formally define both aggregations of rankings (ARs) and aggregations of soft rankings (ASRs).

Definition 1 (Ranking) *A ranking of $\mathcal{Y}$ is a function $A : \mathcal{Y} \times \mathcal{Y} \rightarrow \{0, 1\}$ generated by a total ordering. We use $A(y^{(i)}, y^{(j)}) = 1$ to denote preference $y^{(i)} \rightarrow y^{(j)}$ and $A(y^{(i)}, y^{(j)}) = 0$ for $y^{(i)} \leftarrow y^{(j)}$.*

Definition 2 (Aggregation of Rankings (AR)) *An aggregation of rankings is specified by a set of rankings $\mathbf{A}$ and a corresponding weight function $\alpha \in \triangle^{|\mathbf{A}|}$ (indexed by $A \in \mathbf{A}$).*

Definition 3 (Aggregate Preference) *An aggregation preference in an AR between $y^{(i)}, y^{(j)} \in \mathcal{Y}$ is defined to be*

$$R_{\mathbf{A}, \alpha}(y^{(i)}, y^{(j)}) := \sum_{A \in \mathbf{A}} \alpha_A A(y^{(i)}, y^{(j)}).$$

Corresponding to rankings, ARs, and aggregate preferences $R_{\mathbf{A}, \alpha}$, we will have soft rankings, ASRs, and aggregate probabilities $F_{\mathbf{B}, \beta}$.

Definition 4 (Soft Rankings) *A soft ranking on $\mathcal{Y}$ is a function $B : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ generated by a categorical probability distribution on $\mathcal{Y}$, $\mathbf{p} \in \triangle^{|\mathcal{Y}|}$.*

$$B(y^{(i)}, y^{(j)}) := \frac{p_i}{p_i + p_j}.$$

Definition 5 (Aggregation of Soft Rankings (ASR)) *An aggregation of soft rankings is specified by a set of soft rankings $\mathbf{B}$ and a corresponding weight function $\beta \in \triangle^{|\mathbf{B}|}$ (indexed by $B \in \mathbf{B}$).*

Definition 6 (Aggregate Probability) *An aggregate probability in an ASR between $y^{(i)}, y^{(j)} \in \mathcal{Y}$ is*

$$F_{\mathbf{B}, \beta}(y^{(i)}, y^{(j)}) := \sum_{B \in \mathbf{B}} \beta_B B(y^{(i)}, y^{(j)}).$$

Observation 1 *Suppose the probability distribution for a covariate adjustment on $\mathbf{X}$ (e.g. X in our previous examples), $\Pr(\mathbf{X} \mid Y \in \{y^{(i)}, y^{(j)}\})$, is the same for all pairs of labels $\{y^{(i)}, y^{(j)}\}$. The treatment effects computed for each label pair then correspond to the difference between the aggregate probabilities in two ASRs with a $B \in \mathbf{B}$ for each assignment of $\mathbf{X} = \mathbf{x}$ and $\beta_B = \Pr(\mathbf{x} \mid Y \in \{y^{(i)}, y^{(j)}\})$.*

Observation 1 illustrates the connection between ASRs and the paradox of nontransitivity developed in the previous section — namely, that the backdoor adjustment is a weighted sum of conditional probabilities that can each be interpreted as a soft ranking. The remainder of this paper will be dedicated to showing that ARs and ASRs on the same cardinality $|\mathcal{Y}| = n$ can hold the exact same vectors of weights. This gives a correspondence between the notion of "consistency" when combining causal quantities in omitted label contexts and aggregations of rankings in the already well-studied linear ordering polytope.

To make our statement precise, we will denote $\mathcal{A}$ as the set of $\{0, 1\}^{n(n-1)}$ vectors associated with the output values of some A in a total ordering and $Co(\mathcal{A})$ as its convex hull. Note that $Co(\mathcal{A})$ is the space of edge-weights for ARs, i.e., vectors with entries corresponding to $R(y^{(i)}, y^{(j)})$. Similarly, denote $\mathcal{B}$ as the set of $[0, 1]^{n(n-1)}$ vectors generated by some categorical distribution and observe that its convex hull $Co(\mathcal{B})$ is the space of possible edge-weights for ASRs.

Theorem 7 *$Co(\mathcal{A})$ and $\overline{Co}(\mathcal{B})$ are the same.*

It is not difficult to see how soft rankings can be made "harder" by increasing the relative difference in probabilities. That is, replacing $2, 1, 0$ in Table 3 with $100, 1, 0$ more closely simulates an absolute preference. Showing that any set of aggregate probabilities from an ASR can be realized with aggregate preferences from an AR is less obvious. We will prove this direction by using the probability table in an ASR to directly construct a corresponding AR.

4.3. Probabilities can Emulate Preferences

We will begin with the simpler direction, given by Lemma 8.

Lemma 8 *$Co(\mathcal{A}) \subset \overline{Co}(\mathcal{B})$.*

To prove Lemma 8, we will first show that for every $A \in \mathcal{A}$, there exists a $B \in \mathcal{B}$ that is arbitrarily close to it. We will then make use of the following more general lemma, which we prove in Appendix A.

Lemma 9 *Consider a set of vectors $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_t\}$ and with $\mathbf{v}_i \in \mathbb{R}^m$ for all i. If we have $\tilde{V}$ such that for every $\varepsilon > 0$ and $\mathbf{v} \in V$, there exists $\tilde{\mathbf{v}} \in \tilde{V}$ such that $\|\tilde{\mathbf{v}} - \mathbf{v}\|_2 < \varepsilon$, then $Co(V) \subseteq \overline{Co}(\tilde{V})$.*

We now give the proof for Lemma 8.

Proof For a given A from a total ordering, we will show how to find a probability vector $\mathbf{p}$ that generates a B with values that are arbitrarily close to the $0, 1$ values of A. As already alluded to, this will involve blowing up the ratios of the probabilities in $\mathbf{p}$.

Let $y^{(0)} \rightarrow \ldots \rightarrow y^{(n-1)}$ be the ordering specified by A. Let the ith element of $\mathbf{p}$ be ε^i/z, where $z = \sum_{j=1}^{n} \varepsilon^j$ is a normalization factor so that $\mathbf{p}$ remains in the simplex. Notice that this assignment gives us

$$B(y^{(i)}, y^{(j)}) = \begin{cases} \frac{1}{1+\varepsilon^{j-i}} & \text{if } i < j \\ \frac{\varepsilon^{i-j}}{1+\varepsilon^{i-j}} & \text{if } i > j \end{cases}.$$

Setting $\varepsilon > 0$ arbitrarily close to 0 achieves

$$B(y^{(i)}, y^{(j)}) \longrightarrow \begin{cases} 1 & \text{if } i < j \\ 0 & \text{if } i > j \end{cases},$$

which means that $B(y^{(i)}, y^{(j)}) \rightarrow A(y^{(i)}, y^{(j)})$. Finally, we apply Lemma 9. ∎
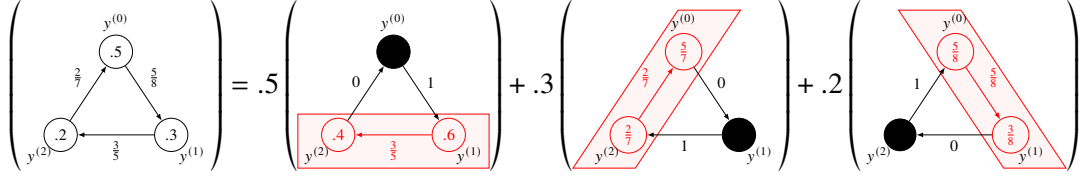
Figure 3: A demonstration of the inductive step in the proof for Lemma 10. The weights on the LHS are the aggregate probabilities $(y^{(i)}, y^{(j)})$ that we wish to generate, while the numbers within each vertex $y^{(i)}$ specify p_i. The weights of the graphs on the RHS are given by Equation 4, with adjusted (re-normalized) probabilities $p^{[k]}$ specified within the vertices. Three subgraphs are highlighted in red, which represent the smaller sets of labels which can be decomposed according to the inductive hypothesis.

4.4. Preferences can Emulate Probabilities

We will continue with the more difficult direction, given by Lemma 10.

Lemma 10 $Co(\mathcal{B}) \subset Co(\mathcal{A})$.

We will prove this direction by showing that every possible instance of B is in $Co(\mathcal{A})$. Convexity of $Co(\mathcal{A})$ will then complete the proof.

Let $\mathbf{A}^{(i)} \subset \mathbf{A}$ denote the set of rankings for which $y^{(i)}$ is the "first choice." Equivalently, $\mathbf{A}^{(i)}$ is defined such that we have $A(y^{(i)}, y^{(j)}) = 1$ for all $j \neq i$ and $A \in \mathbf{A}^{(i)}$. We extend this notation to multiple indices, with $\mathbf{A}^{(ij)}$ encoding $y^{(i)}$ as first choice and $y^{(j)}$ as second choice (i.e., $y^{(j)}$ is preferred over all labels other than $y^{(i)}$). If only one ranking satisfies the restriction, then we will use an unbolded $A^{(ij)}$ to emphasize that we are no longer dealing with a set.

Proof We will induct on the number of labels n. The inductive hypothesis is that any $B \in \mathcal{B}$ generated by a categorical distribution $\mathbf{p} \in \triangle^n$ over n labels can be expressed as an AR $\mathbf{A}, \alpha$. This can easily be shown for the base case of $n = 2$ by assigning $\alpha_{A^{(01)}} = p_0$ and $\alpha_{A^{(10)}} = p_1$.

Now, assuming the inductive hypothesis to be correct for all B on n labels, we will show how to construct an AR for a B on $n + 1$ labels. First, expand $R_{\mathbf{A}, \alpha}$, which we have not yet specified, into aggregate rankings on $\mathbf{A}^{(0)}, \ldots, \mathbf{A}^{(n)}$,

$$R_{\mathbf{A}, \alpha} = \sum_{k=0}^{n} p_k R_{\mathbf{A}^{(k)}, \alpha^{(k)}}. \tag{3}$$

Now, choose some label $y^{(k)}$ and construct a new $\mathbf{p}^{[k]} \in \triangle^n$ that will represent the "conditional" probability table for all rankings with $y^{(k)}$ as their first choice. This is done by setting all $p_i^{[k]} = p_i/(1 - p_k)$, for $i \neq k$. This new $\mathbf{p}^{[k]}$ implies a $B^{[k]}(y^{(i)}, y^{(j)})$ that matches $B(y^{(i)}, y^{(j)})$ for all $i, j \neq k$. $B^{[k]}$ also satisfies the requirements for the inductive hypothesis when leaving out $y^{(k)}$ from the list of labels, thereby reducing the length of the list of labels. We can therefore assume there is a set of rankings $\mathbf{A}^{[k]}$ and corresponding $\alpha^{[k]}$ that forms an AR for which $R_{\mathbf{A}^{[k]}, \alpha^{[k]}} = B^{[k]}$.

Each $A^{[k]} \in \mathbf{A}^{[k]}$ can now be augmented with a first-choice preference of $y^{(k)}$ to generate the set $\mathbf{A}^{(k)}$ with corresponding $\alpha^{(k)} = \alpha^{[k]}$. Using this assignment, we have that

$$R_{\mathbf{A}^{(k)}, \alpha^{(k)}}(y^{(i)}, y^{(j)}) = \begin{cases} B(y^{(i)}, y^{(j)}) & \text{if } i, j \neq k \\ 1 & \text{if } i = k \\ 0 & \text{if } j = k \end{cases} \tag{4}$$

Applying Equation 4 to Equation 3 verifies that

$$\begin{aligned} R_{\mathbf{A}, \alpha}(y^{(i)}, y^{(j)}) = p_i + \sum_{k \neq i, j} p_k B(y^{(i)}, y^{(j)}) &= (p_i + p_j) B(y^{(i)}, y^{(j)}) + (1 - p_i - p_j) B(y^{(i)}, y^{(j)}) \\ &= B(y^{(i)}, y^{(j)}). \end{aligned}$$

Hence, our construction of $\mathbf{p}_i^{[k]}$ is valid. This completes the inductive proof. As stated earlier, convexity of $\mathrm{Co}(\mathcal{A})$ gives the desired result. ∎

Figure 3 illustrates an example inductive step in the proof for Lemma 10.

5. Discussion

This paper studies dataset contexts and how they affect the conclusions we take from them. Specifically, we have explored how the traditional principles of causal inference break down in omitted label contexts through the manifestation of two well-known paradoxes. These paradoxes warn practitioners about causal inference within a label-biased setting. In addition, the structures of these paradoxes (as described by Theorem 7) provide insight into their nature and define a notion of consistency between partial studies.

The study of omitted label contexts may have applications in handling other forms of label bias. In the study of rare diseases, it may be unfeasible to avoid scaling up the proportion of affected individuals. This is a form of label bias and comes with all of the caveats discussed in the preceding sections. An intriguing alternative is to instead combine the label of interest with another similarly rare disease. In a dataset of two rare diseases, one can maintain the relative probabilities of the labels while keeping their portion of the dataset nontrivial. In order to extend the results of this type of study to the broader population, one can imagine computing multiple ratios of increasingly more common labels in order to compute the context-induced distributions needed to perform adjustments. Such a combination requires a deep understanding of the structure in networks of contexts.

While the linear ordering polytope has a few counterintuitive properties, such as the breakdown of preference-transitivity, it also has limitations. For example, the Condorcet paradox actually represents the "maximum amount of nontransitivity" that is possible in a cycle of three choices. It is not possible for a cycle of 80% preferences to exist, nor is a cycle of 100%, 50%, 70% possible. In fact, for any cycle of length ℓ, the aggregate preferences along that cycle must sum to between 1 and $\ell - 1$, a property discussed in Fishburn (1992) as the "triangle inequality" and Mazaheri et al. (2021) as the "curl condition."

These limitations can be harnessed to provide bounds on unmeasured aggregate preferences, or "missing edge weights" as discussed in Mazaheri et al. (2021). That is, if we are missing a study on two labels, we can provide bounds on that study's likely outcome using the outcomes of related studies. Furthermore, we can use these properties to detect inconsistencies within sets of studies.

As larger machine learning models become more costly to train and comprehensive data drifts towards private datasets, many practitioners are choosing to re-purpose pre-trained "checkpoints" to new tasks. Of course, these checkpoints all come from data-contexts. Hence, we see decision fusion from different contexts as an essential upcoming field to launch data science into a new age.

Acknowledgments

References

Noga Alon. Voting paradoxes and digraphs realizations. *Advances in Applied Mathematics*, 29(1): 126–135, 2002.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Federico Castanedo. A review of data fusion techniques. *The scientific world journal*, 2013.

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

Edmund S Conklin and John W Sutherland. A comparison of the scale of values method with the order-of-merit method. *Journal of Experimental Psychology*, 6(1):44, 1923.

Juan D Correa, Jin Tian, and Elias Bareinboim. Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2744–2751, 2019.

Peter C Fishburn. Induced binary probabilities and the linear ordering polytope: A status report. *Mathematical Social Sciences*, 23(1):67–80, 1992.

Branko Grünbaum, Victor Klee, Micha A Perles, and Geoffrey Colin Shephard. *Convex polytopes*, volume 16. Springer, 1967.

Miguel A Hernán, David Clayton, and Niels Keiding. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3):780–785, 03 2011. ISSN 0300-5771. doi: 10.1093/ije/dyr041. URL https://doi.org/10.1093/ije/dyr041.

Robins JM Hernán MA. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Bijan Mazaheri, Siddharth Jain, and Jehoshua Bruck. Synthesizing New Expertise via Collaboration. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2447–2452, 2021. doi: 10.1109/ISIT45174.2021.9517822.

Jean Antoine Nicolas et al. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix. Par m. le marquis de Condorcet,...* de l'Imprimerie Royale, 1785.

Judea Pearl. *Causality.* Cambridge university press, 2009.

Judea Pearl. Comment: understanding simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl*, pages 399–412. 2022.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

JM Robins, MA Hernán, G Fitzmaurice, M Davidian, G Verbeke, and G Molenberghs. Longitudinal data analysis. *Handbooks of modern statistical methods*, pages 553–599, 2009.

Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, and Bernhard Schölkopf. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. *Advances in neural information processing systems*, 21, 2008.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Samy Suissa. Immortal time bias in pharmacoepidemiology. *American journal of epidemiology*, 167 (4):492–499, 2008.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

Appendix A. Proof of Lemma 9

Convex hulls of finite sets in $\mathbb{R}^\ell$ are *convex* polytopes, which can be expressed as an intersection of h halfspaces indexed by f with $\{\mathbf{x} : \mathbf{a}^{(\mathbf{f})^\top} \mathbf{x} x < b^{(f)}\}$ (Grünbaum et al., 1967). Vectors $\mathbf{a}^{(\mathbf{f})^\top}$ can be combined as row-vectors of a matrix, A, so that any convex polytope can be expressed as

$$\{x : A\mathbf{x} \prec \mathbf{b}\} = \left\{\mathbf{x} : \begin{pmatrix} (\mathbf{a}^{(\mathbf{1})})^\top \\ \vdots \\ (\mathbf{a}^{(\mathbf{h})})^\top \end{pmatrix} \mathbf{x} \prec \begin{pmatrix} b^{(1)} \\ \vdots \\ b^{(h)} \end{pmatrix} \right\}. \tag{5}$$

For convenience, the vectors $\mathbf{a}^{(\mathbf{f})}, \tilde{\mathbf{a}}^{(\mathbf{f})}$ are assumed to be unit vectors throughout.

The idea behind the proof will be to analyze the movement of the boundaries of the polytope defined by $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ and corresponding polytope defined by the "perturbed points" $\tilde{\mathcal{V}} = \{\tilde{\mathbf{v}}_1, \ldots \tilde{\mathbf{v}}_m\}$. The key is to show that a point that is far enough from the boundary of $\mathrm{Co}(\mathcal{V}$ will also be within $\mathrm{Co}(\tilde{\mathcal{V}})$, given by Lemma 11. This required distance from the boundary will be relative to the amount by which the perturbed points have moved. As we make the perturbation arbitrarily small (i.e. $\varepsilon \to 0$, all points in the interior of the polytope will be included.

Lemma 11 *Let*

$$Co(\mathcal{V}) = \{\mathbf{x} : A\mathbf{x} \prec \mathbf{b}\}$$
$$Co(\tilde{\mathcal{V}}) = \{\mathbf{x} : \tilde{A}\mathbf{x} \prec \tilde{\mathbf{b}}\}$$

as given by Equation 5. If $A\mathbf{x} \prec \mathbf{b} - \varepsilon \mathbf{1}^\ell$ and $\|\mathbf{v_i} - \tilde{\mathbf{v_i}}\|_2 < \varepsilon \ \forall i$, then $\tilde{A}\mathbf{x} \prec \tilde{\mathbf{b}}$.

To prove Lemma 11, we will need to show that the boundaries of the polytopes do not move too much. We will do this using Lemma 13, which bounds how far $\mathrm{Bo}(\mathrm{Co}(V))$ can be from $\mathrm{Bo}(\mathrm{Co}(\tilde{V}))$ along a single "face."

Definition 12 *Choose $f \in [h]$. Define:*

$$W^{(f)} = \{\mathbf{w} : (\mathbf{a^{(f)}})^\top w = b^{(f)}, \mathbf{w} \in V\}$$
$$\tilde{W}^{(f)} = \{\tilde{\mathbf{v_i}} : \mathbf{v_i} \in W^{(f)}\}$$

We restrict the size of $\left|W^{(f)}\right| = \ell$, which is the number of points needed to define a halfspace in $\mathbb{R}^\ell$. This can be done by allowing for multiple identical $\mathbf{a_f}, b_f$ combinations corresponding to all size ℓ subsets of the v_i along the boundary.

Note that $Co(W^{(f)})$ describes a "face" of the polytope $Co(\mathcal{V})$ indexed by f which is perpendicular to $\mathbf{a^{(f)}}$. $Co(\tilde{W}^{(f)})$ describes the perturbed face.

Lemma 13 *Choose $f, g \in [h]$ arbitrarily and let $W^{(f)} = \{\mathbf{w_1^{(f)}}, \ldots, \mathbf{w_\ell^{(f)}}\}$ and $\tilde{W}^{(f)} = \{\tilde{\mathbf{w_1^{(f)}}}, \ldots, \tilde{\mathbf{w_\ell^{(f)}}}\}$. For every $\mathbf{m^{(f)}} \in \overline{Co}(W^{(f)})$, we have $(\tilde{\mathbf{a}}^{(g)})^\top \mathbf{m^{(f)}} < \tilde{b}^{(g)} + \varepsilon$.*

Proof Because $m \in \overline{\mathrm{Co}}(W^{(f)})$, there is some $\lambda \in \triangle_\ell$ with

$$\mathbf{m^{(f)}} = \sum_{i=1}^{\ell} \lambda_i \mathbf{w_i^{(f)}} \in \overline{\mathrm{Co}}(W^{(f)}) \tag{6}$$

Consider also

$$\tilde{\mathbf{m}}^{(\mathbf{f})} = \sum_{i=1}^{\ell} \lambda_i \tilde{\mathbf{w_i^{(f)}}} \in \overline{\mathrm{Co}}(\tilde{W}^{(f)}) \tag{7}$$

Note that the norm of the difference between these two vectors is bounded:

$$\left\|\mathbf{m^{(f)}} - \tilde{\mathbf{m}}^{(\mathbf{f})}\right\|_2 = \left\|\sum_{i=1}^{\ell} \lambda_i (\mathbf{w_i^{(f)}} - \tilde{\mathbf{w_i^{(f)}}})\right\|_2$$
$$\leq \sum_{i=1}^{\ell} \lambda_i \underbrace{\left\|\mathbf{w_i^{(f)}} - \tilde{\mathbf{w_i^{(f)}}}\right\|_2}_{<\varepsilon} < \varepsilon \tag{8}$$

14

Also note that because $\tilde{\mathbf{m}}^{(\mathbf{f})} \in \overline{\mathrm{Co}}(\tilde{W}^{(f)}) \subseteq \overline{\mathrm{Co}}(\tilde{\mathcal{V}})$, we have that $(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \tilde{\mathbf{m}}^{(\mathbf{f})} \leq \tilde{b}^{(g)}$. Now, a simple application of Cauchy-Schwartz gives:

$$
\begin{aligned}
(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{m}^{(\mathbf{f})} &= (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top (\tilde{\mathbf{m}}^{(\mathbf{f})} + (\mathbf{m}^{(\mathbf{f})} - \tilde{\mathbf{m}}^{(\mathbf{f})})) \\
&= \underbrace{(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \tilde{\mathbf{m}}^{(\mathbf{f})}}_{\leq \tilde{b}^{(g)}} + (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top (\mathbf{m}^{(\mathbf{f})} - \tilde{\mathbf{m}}^{(\mathbf{f})}) \\
&\leq \tilde{b}^{(g)} + \left\| \tilde{\mathbf{a}}^{(\mathbf{g})} \right\|_2 \left\| \mathbf{m}^{(\mathbf{f})} - \tilde{\mathbf{m}}^{(\mathbf{f})} \right\|_2 \\
&< \tilde{b}^{(g)} + \varepsilon
\end{aligned}
\tag{9}
$$

$\blacksquare$

With this, we are now ready to prove Lemma 11.

Proof Choose an arbitrary face $g \in [h]$. Recall we have $\mathbf{x} \in \mathrm{Co}(V)$ with $(\mathbf{a}^{(\mathbf{g})})^\top \mathbf{x} < b - \varepsilon$ and we wish to show $(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{x} < \tilde{b}^{(g)}$.

Let $\mathbf{m}_{\mathbf{x}}^{(\mathbf{f})}$ be the result of extending $\tilde{\mathbf{a}}^{(\mathbf{g})}$ from $\mathbf{x}$ to $\mathrm{Bo}(V)$. This must hit some face with $(\mathbf{a}^{(\mathbf{f})})^\top \mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} = b^{(f)}$, so $\mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} \in \mathrm{Co}(W^{(f)})$. That is, find β such that

$$
\mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} = \beta \tilde{\mathbf{a}}^{(\mathbf{g})} + \mathbf{x} \in \mathrm{Co}(W^{(f)})
\tag{10}
$$

First, lets bound β. Notice that because $\mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} \in \mathrm{Co}(W^{(f)})$, we have

$$
\begin{aligned}
(\mathbf{a}^{(\mathbf{f})})^\top \mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} &= (\mathbf{a}^{(\mathbf{f})})^\top \left(\sum_{i=1}^{\ell} \lambda_i \mathbf{w}_{\mathbf{i}}^{(\mathbf{f})} \right) \\
&= \sum_{i=1}^{\ell} \lambda_i (\mathbf{a}^{(\mathbf{f})})^\top \mathbf{w}_{\mathbf{i}}^{(\mathbf{f})} = b^{(f)}
\end{aligned}
\tag{11}
$$

So, we have

$$
b^{(f)} = (\mathbf{a}^{(\mathbf{f})})^\top \mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} = \beta \underbrace{(\mathbf{a}^{(\mathbf{f})})^\top \tilde{\mathbf{a}}^{(\mathbf{g})}}_{\leq 1} + \underbrace{(\mathbf{a}^{(\mathbf{f})})^\top \mathbf{x}}_{<b^{(f)} - \varepsilon} \Rightarrow \varepsilon < \beta
\tag{12}
$$

Now, apply Lemma 13

$$
\begin{aligned}
(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{m}_{\mathbf{x}}^{(\mathbf{f})} &< \tilde{b}^{(g)} + \varepsilon \\
(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{x} + (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \tilde{\mathbf{a}}^{(\mathbf{g})} \beta &< \tilde{b}^{(g)} + \varepsilon \\
(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{x} &< \tilde{b}^{(g)}.
\end{aligned}
\tag{13}
$$

Face $g \in [h]$ was chosen arbitrarily, so this holds for all half-spaces in the convex polytope. Hence, we have $A\mathbf{x} \prec \mathbf{b}$. $\blacksquare$

Appendix B. Counts that Follow the Correct DAG

The paradox presented in Table 3 used counts yielding a distribution that does not precisely follow the given DAG. We will now show how to construct a similar set of counts with the statistics needed to imply the causal structure.

T	X	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$
$t^{(1)}$	$x^{(1)}$	$2\alpha_1$	α_1	0
$t^{(1)}$	$x^{(2)}$	0	$2\beta_1$	β_1
$t^{(1)}$	$x^{(3)}$	γ_1	0	$2\gamma_1$
$t^{(0)}$	$x^{(1)}$	0	α_2	$2\alpha_2$
$t^{(0)}$	$x^{(2)}$	$2\beta_2$	0	β_2
$t^{(0)}$	$x^{(3)}$	γ_2	$2\gamma_2$	0

Table 4: A more general specification of counts for our paradox.

Consider Table 4. The structure is copied from Table 3. If $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, and $\gamma_1 = \gamma_2$, then we notice that the relative probabilities of $\Pr(x)$ are given by the αs, βs, and γs. If these coefficients are all equal, then we have every row considered with equal weight, as in the main paper.

While setting all of the Greek coefficients to 1 provides a nice intuition for how the paradox emerges, it does not give a distribution that obeys the requirements of the given DAG. In order for our distribution to (1) factorize according to the DAG and (2) be faithful to the DAG, we must have the following properties:

1. $T \not\perp\!\!\!\perp Y$

2. $X \not\perp\!\!\!\perp Y$

3. $T \not\perp\!\!\!\perp X$

4. $T \not\perp\!\!\!\perp Y \mid X$

5. $X \not\perp\!\!\!\perp Y \mid T$

6. $T \not\perp\!\!\!\perp X \mid Y$

With all of the Greek coefficients set to 1, we notice that conditions 5,6 are met. The domain expertise setting effectively conditions on Y by restricting its values. When restricted to two columns, we also meet condition 4.

The remaining conditions (as well as condition 4 in the broader case) can be met by varying the Greek coefficients. A Jupyter notebook available here allows one to explore different settings to the Greek coefficients to achieve this paradox. One example is $\alpha_1 = \beta_1 = \gamma_1 = 1$ and $\alpha_2 = 1.1$, $\beta_2 = 1.2$, $\gamma_2 = 1.3$. The code returns whether the 6 independence conditions hold. The three ATEs for each omitted label context are then printed. All code was run on a Macbook Air with an M1 processor.