CAUSAL PARTIAL IDENTIFICATION WITH DATA AUGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We provide a first analysis for using knowledge of symmetries in data generation via *data augmentation (DA)* transformations for sharpening bounds on causal effects derived from observational data. The causal effect of the treatment X on outcome Y is generally not identifiable from observational data alone if their common causes, also known as confounders, are unobserved. Partial identification (PI) entails estimating bounds on such treatment effects by solving a constrained optimization problem that encodes different assumptions imposed on data generation. PI has use in many application domains where such bounds are sufficient to inform policy decisions, even if the treatment effect itself is not identifiable. To this end, we propose that the cheap and ubuquitous tool of DA, which is otherwise used for mitigating estimation variance, can also be repurposed for sharpening bounds in PI. This is especially useful when the data is complex (i.e., continuous, high-dimensional), as imposing additional constraints becomes expensive compared to a simple pre-processing via DA.

1 INTRODUCTION

One of the classical problems in machine learning is that of regression—predicting an outcome Y from a set of predictors X. The standard approach involves learning a model from i.i.d. samples to generalize to new, unseen data, often employing regularization techniques like data augmentation (DA) to improve performance Vapnik (1998); Shorten & Khoshgoftaar (2019); Lyle et al. (2020). However, we generally cannot interpret such predictive models as causal. The statistical relationship between X and Y may not reflect the true influence of X on Y, but could instead be driven by unobserved common causes, or *confounders*. The gold standard for eliminating such confounding bias is a direct *intervention* on X, where we explicitly assign its values during data generation. However, performing interventions are often impractical or prohibitively costly Peters et al. (2017); Pearl (2009).

When confounders are unobserved, identifying the true causal effect is notoriously difficult. While methods like instrumental variables (IVs) can simulate interventions under specific conditional independence assumptions, valid instruments are scarce in many high-dimensional applications like computer vision and natural language processing Singh et al. (2019); Zhang et al. (2023); Kilbertus et al. (2020). This scarcity has motivated a line of inquiry into repurposing standard machine learning tools for causal inference. For instance, regularization techniques like ℓ_1 and ℓ_2 have been studied not just for improving i.i.d. generalization, but also for reducing confounding bias in causal estimates Janzing (2019); Kania & Wit (2023); Vankadara et al. (2022).

This perspective has recently been extended to *data augmentation (DA)*, a ubiquitous regularization method Shorten & Khoshgoftaar (2019); Lyle et al. (2020). It has been demonstrated that when DA respects the symmetries of the outcome-generating process (i.e., is outcome-invariant), it can be framed as a "soft" intervention Akbar et al. (2025). This interventionist view of DA allows it to reduce confounding bias and improve the *point-estimation* of unidentifiable causal effects. However, in many real-world scenarios with significant uncertainty, the causal effect is fundamentally not point-identifiable. The goal then shifts from seeking a single best estimate to the more robust task of *partial identification* (PI)—deriving rigorous bounds that are guaranteed to contain the true causal effect. This raises a crucial question: if DA can improve point estimates of causal effects, can it also

be leveraged to sharpen the bounds derived through partial identification? We answer this question in the affirmative.

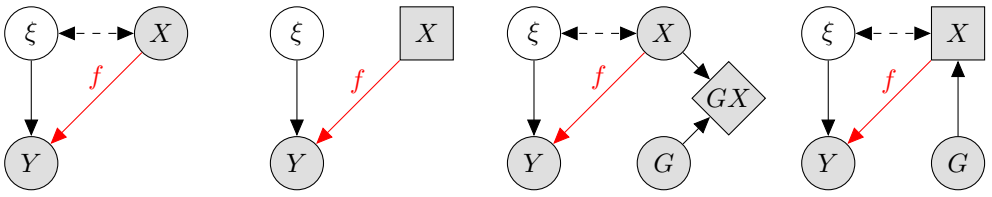
Our contribution. To this end, we provide a first analysis of DA for partial identification of causal effects. Specifically, we leverage the insight that outcome-invariant DA is equivalent to a transformation intervention on the treatment variable. We demonstrate, both theoretically and empirically, that this interventionist perspective allows DA to improve partial identification in three key ways:

- **Valid Bounds:** We first establish that applying outcome-invariant DA is a *valid* procedure. The resulting identified set of causal functions does not move farther from the true causal function, ensuring the integrity of the bounds (Theorem 1).
- **Sharper Bounds:** We show that DA strictly sharpens the bounds on the causal effect. It reduces the volume of the identified set of functions and tightens the point-wise identification intervals, provided the augmentation is not trivial for the treatment queries (Proposition 1).
- **More Robust Bounds:** Finally, we demonstrate that the identified set becomes more robust. DA reduces the worst-case causal excess risk, meaning that even the least accurate hypothesis within the identified set is improved, leading to more reliable and informative bounds for decision-making (Theorem 2).

Taken together, our results position outcome-invariant DA as a simple, largely model-agnostic pre-processing step that can be composed with many existing partial identification method to enhance performance.

2 Preliminaries

2.1 Statistical vs. Causal Inference



(a) Original SEM (b) $\mathrm{do}(X)$ (c) SEM post DA (d) $\mathrm{do}(X := GX)$

Figure 1: Graphs of respective SEMs. (a) The original SEM from Eq. (1) with confounded (X, Y). (b) Graph obtained via intervention on X in Eq. (1). (c) Graph for DA. (d) Graph for transformation intervention. Observational distributions of (GX, Y, ξ) in (c) and (X, Y, ξ) in (d) are identical.

Consider random treatment X, outcome Y taking values in $\mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{Y} \subseteq \mathbb{R}$ respectively. The function $f \in \mathcal{H} := \{h : \mathcal{X} \to \mathcal{Y}\}$ defines their causal relationship via a *structural equation model (SEM)*

$$Y = f(X) + \xi, \qquad \mathbb{E}[\xi] = 0. \tag{1}$$

We want to estimate f given a dataset $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=0}^n$ of n samples from the distribution $\mathbb{P}_{X,Y}$.

With the assumption $X \perp\!\!\!\perp \xi$, we have $\mathbb{E}[Y \,|\, X = \boldsymbol{x}] = f(\boldsymbol{x})$ in Eq. (1). *Statistical inference* entails identifying precisely the Bayes optimal predictor $\mathbb{E}[Y \,|\, X = \boldsymbol{x}]$ from $\mathcal{D}$ by minimizing an empirical version of the *statistical risk* over hypotheses $h \in \mathcal{H}$ for some proper, convex loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$,

$$R_{\mathrm{erm}}(h) := \mathbb{E}[\ell(Y, h(X))]. \tag{2}$$

Then, for a sufficiently rich hypothesis class, the minimizer h_{erm} gives an unbiased estimation of f.

However, the residual ξ in Eq. (1) may generally be correlated with X, i.e., $\mathbb{E}[\xi \,|\, X] \neq 0$, so that the conditional $\mathbb{E}[Y \,|\, X = \boldsymbol{x}]$ now gives a biased estimate of $f(\boldsymbol{x})$ (Pearl, 2009; Peters et al., 2017). This correlation arises due to unobserved common causes of X and Y, known as *confounders*. We say that X and Y are confounded and refer to the resulting bias as the *confounding bias* Pearl (2009). *Causal inference* entails adjusting for this bias to identify f, or at the very least account for it by finding bounds on f should identification not be possible. Both approaches are outlined below.

2.2 Intervention and adjustment for causal effect identification

We can make X and the residual ξ uncorrelated via an *intervention*[1] $\mathrm{do}(X \coloneqq X')$ that explicitly sets X to some independently sampled X' in Eq. (1) during data generation. The induced distribution, referred to as the *interventional distribution*, is represented by $\mathbb{P}^{\mathrm{do}(X:=X')}_{X,Y}$. We use the shorthand notation $\mathrm{do}(X)$ for an intervention where $X' \sim \mathbb{P}_X$, under which the objective from Eq. (2) now defines the *causal risk* (Kania & Wit, 2023; Vankadara et al., 2022; Janzing & Schölkopf, 2018b) as

$$R^{\mathrm{do}(X)}_{\mathrm{erm}}(h) \coloneqq \mathbb{E}^{\mathrm{do}(X)}[\ell(Y, h(X))]. \tag{3}$$

The target estimand of Eq. (3) is the *average treatment effect (ATE)* $\mathbb{E}^{\mathrm{do}(X:=\boldsymbol{x})}[Y \,|\, X = \boldsymbol{x}]$ which equals $f(\boldsymbol{x})$ for the SEM under consideration in Eq. (1). Minimizers of Eq. (3) therefore give an unbiased estimation of f. To better capture the *estimation error* for a candidate hypothesis $h \in \mathcal{H}$, we use the *causal excess risk* (Vankadara et al., 2022) by removing irreducible noise from Eq. (3) as

$$E^{\mathrm{do}(X)}(h) \coloneqq R^{\mathrm{do}(X)}_{\mathrm{erm}}(h) - R^{\mathrm{do}(X)}_{\mathrm{erm}}(f).$$

Since interventions are often inaccessible for computing the risk Eq. (3), estimating f usually relies on access to the full joint distribution $\mathbb{P}$ of (X, Y, ξ) via *back-door adjustment* (Xu & Gretton, 2022)

$$h^{\mathbb{P}}_{\mathrm{adj}}(\boldsymbol{x}) \coloneqq \mathbb{E}_{\xi}[\mathbb{E}[Y \,|\, X = \boldsymbol{x}, \xi]], \qquad (X, Y, \xi) \sim \mathbb{P}.$$

2.3 Partial identification and sensitivity analysis

For unobserved noise ξ, identification of f is generally not possible from $\mathbb{P}_{X,Y}$ alone. Nevertheless, given assumptions on the data generating process in Eq. (1), we can do *partial identification (PI)* (Padh et al., 2023) of f by considering all the joint distributions $\mathbb{Q}$ consistent with said assumptions,

$$\mathcal{Q}_{\mathrm{pi}}\big(\mathbb{P}_{X,Y}\big) \coloneqq \Big\{ \mathbb{Q} \in \mathcal{C}_{\mathrm{pi}} \,\Big|\, \mathbb{Q}_{X,Y} = \mathbb{P}_{X,Y} \Big\},$$

where the constraint set $\mathcal{C}_{\mathrm{pi}}$ encodes our assumptions. If correctly specified, $\mathbb{P} \in \mathcal{C}_{\mathrm{pi}}$ and the following set $\mathcal{H}_{\mathrm{pi}}$ of candidate hypotheses contains the true solution f, or the interval $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ holds $f(\boldsymbol{x})$,

$$\mathcal{H}_{\mathrm{pi}} \coloneqq \Big\{ h^{\mathbb{Q}}_{\mathrm{adj}} \,\Big|\, \mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}} \Big\}, \qquad \mathcal{H}_{\mathrm{pi}}(\boldsymbol{x}) \coloneqq \Big\{ h^{\mathbb{Q}}_{\mathrm{adj}}(\boldsymbol{x}) \,\Big|\, \mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}} \Big\},$$

where $\mathcal{Q}_{\mathrm{pi}}$ is shorthand for $\mathcal{Q}_{\mathrm{pi}}\big(\mathbb{P}_{X,Y}\big)$. Computing the interval $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ at $\boldsymbol{x}$ is often more practical than characterizing the set $\mathcal{H}_{\mathrm{pi}}$, since it amounts to solving two constrained optimization problems as

$$\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x}) = \Big[\ \min\nolimits_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} h^{\mathbb{Q}}_{\mathrm{adj}}(\boldsymbol{x}),\ \ \max\nolimits_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} h^{\mathbb{Q}}_{\mathrm{adj}}(\boldsymbol{x})\ \Big].$$

In either case, we want the identified sets to (i) contain the true solution, (ii) be as small as possible.

The constraint set may also be parameterized as $\mathcal{C}_{\mathrm{pi}}(\boldsymbol{\Gamma})$ to conduct *sensitivity analyses* (Frauen et al., 2024) by varying parameters $\boldsymbol{\Gamma}$ to see how $\mathcal{H}_{\mathrm{pi}}$, $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ evolve as assumptions are relaxed/tightened.

Lastly, since we are now discussing hypothesis sets, we define the two appropriate evaluation metrics

$$E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}}) \coloneqq \min_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \qquad E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}) \coloneqq \max_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big).$$

The *approximation error* $E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}})$ measures how far the target f is from $\mathcal{H}_{\mathrm{pi}}$ (Brown & Ali, 2024), and the *worst-case excess risk* $E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}})$ upper bounds the performance of the identified set $\mathcal{H}_{\mathrm{pi}}$ relative to the target f. Similarly, $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ is evaluated using $E^{\mathrm{do}(\boldsymbol{x})}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}})$ and $E^{\mathrm{do}(\boldsymbol{x})}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}})$.

Choice of constraints. The nature and construction of $\mathcal{C}_{\mathrm{pi}}$ often depends on domain knowledge. Popular approaches involve bounding the spurious correlation between X, Y, including the sensitivity model by Rosenbaum (2002) which parameterizes the strength of unmeasured confounding through odds ratios, its generalization of the Marginal Sensitivity Model (MSM) by Tan (2006) that does the same using propensity scores and the partial R-squared approach by Cinelli & Hazlett

[1] *Transformation* interventions swap X with a transformation GX in Eq. (1) (Dance & Bloem-Reddy, 2024).

(2020) bounds the proportion of variance explained by unobserved confounders. More recently, Fan et al. (2024); Guo et al. (2022) formulated the PI problem as *robust optimization (RO)* over $\mathcal{Q}_{\text{pi}}$ constructed as a total variation ball around the observational distribution $\mathbb{P}_{X,Y}$, and Meresht et al. (2022) similarly uses Wasserstein constraints. An equivalent approach to modeling the confounding is to instead model the random function $f_\xi(\cdot) \coloneqq f(\cdot) + \xi$ itself, also known as the *response function* (Padh et al., 2023). Hu et al. (2021) modeled these using generative adversarial networks (GANs) to then match $\mathbb{P}_{X,Y}$ in distribution. Most of these methods can also leverage auxiliary variables in addition to X, Y for imposing constraints in the form of conditional independences to sharpen bounds. Of note is the instrumental variable (IV) based PI by Balke & Pearl (1997) for when ξ arbitrarily influences Y instead of the additive model in Eq. (1). Modern neural-network based variants for continuous, high-dimensional treatments and/or IVs are explored by Schweisthal et al. (2025); Kilbertus et al. (2020); Hu et al. (2021); Padh et al. (2023); Meresht et al. (2022); Gunsilius (2020).

2.4 Data augmentation

For finite samples, the technique of *data augmentation (DA)* is used to reduce estimation variance (Lyle et al., 2020; Chen et al., 2020) in statistical inference. This is achieved by applying random transformations $G \sim \mathbb{P}_G$ to the data, generating multiple transformed samples $(G\boldsymbol{x}_i, y_i)$ from each original sample $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$, thereby increasing variability in the data for statistical risk evaluation,

$$R_{\text{da+erm}}(h) \coloneqq \mathbb{E}[\ell(Y, h(GX))]. \tag{4}$$

In this work we restrict ourselves to DA with respect to which f is invariant (Lyle et al., 2020; Chen et al., 2020). The action of a group $\mathcal{G}$ is a mapping $\alpha : \mathcal{X} \times \mathcal{G} \to \mathcal{X}$ compatible with the group operation. Writing $\boldsymbol{g}\boldsymbol{x} \coloneqq \alpha(\boldsymbol{x}, \boldsymbol{g})$ as shorthand, we say that f is *invariant* under $\mathcal{G}$ (or $\mathcal{G}$*-invariant*) if

$$f(\boldsymbol{g}\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \forall\ (\boldsymbol{g}, \boldsymbol{x}) \in \mathcal{G} \times \mathcal{X}.$$

Less formally, we say that the map $\boldsymbol{g}\boldsymbol{x}$ is a valid *outcome-invariant* DA transformation. Let $\mathcal{G}$ have a (unique) normalized Haar measure and $\mathbb{P}_G$ be the corresponding distribution defined over it.

Of course one needs to have prior knowledge about the symmetries of f to construct such a DA. We argue that the popularity of this modeling assumption in the DA and invariance literature (Lyle et al., 2020; Chen et al., 2020) is precisely because such symmetries are already established in many application domains. For example, when classifying images of cats and dogs we already know that whatever the true labeling function may be, it would certainly be invariant to rotations on the images. G would then represent the random rotation angle, whereas $G\boldsymbol{x}$ would be the rotated image $\boldsymbol{x}$.

While DA is canonically used to mitigate finite-sample estimation variance, our focus is primarily on the infinite-sample setting, and we present Eq. (4) and subsequent theoretical results in that context. Nonetheless, increasing sample size via DA also bears on our work, a point we shall briefly discuss.

3 Intervention with Data Augmentation

We start by first framing data augmentation as a tool for causal inference. Assume we have access to a valid outcome-invariant DA $G \sim \mathbb{P}_G$ for the SEM in Eq. (1). Now, consider an intervention on the SEM where we substitute the treatment X with the transformation GX. With some abuse of notation, we shall represent this intervention by $\text{do}(X \coloneqq GX)$, the graph of which is shown in Fig. 1d. Comparing the DA mechanism in Fig. 1c and the intervention $\text{do}(X \coloneqq GX)$ in Fig. 1d, we note:

Observation 1 (DA as transformation intervention)**.** $\mathbb{P}_{GX,Y}$ and $\mathbb{P}^{\text{do}(X \coloneqq GX)}_{X,Y}$ are identical.

We can hence treat samples generated via DA as if they were instead generated from an intervention $\text{do}(X \coloneqq GX)$ on X. This now allows us to re-write the post-DA statistical risk from Eq. (4) as

$$R_{\text{da+erm}}(h) = R^{\text{do}(X \coloneqq GX)}_{\text{erm}}(h),$$

to emphasize that DA is equivalent to a transformation intervention on the treatment. As such, Akbar et al. (2025) showed that when DA targets spurious features of X, it mitigates confounding bias in point-estimation of f, even if full identification may not be possible. However, standard approach in such a non-identifiable setting is not to point-estimate f, but to undertake partial identification of f.

This motivates our current work, where we leverage intervention properties of outcome-invariant DA to improve partial identification and/or sensitivity analysis of f, as discussed in the next section.

4 Partial Identification with Data Augmentation

Our main insight for why data augmentation might help with partial identification is summarized as:

(i) Most straightforwardly, increasing sample size via DA can mitigate sampling variation and finite sample errors, a key source of uncertainty in PI (Imbens & Manski, 2004).

(ii) DA adds variation in X beyond Y, reducing ambiguity in PI, which leads to sharper bounds.

(iii) When spurious features of X are perturbed by the DA, it reduces sensitivity to confounding resulting in more robust and informative bounds in sensitivity analysis.

(iv) Most importantly, outcome-invariance of DA guarantees valid bounds whenever $\mathcal{C}_{\text{pi}}$ is valid.

We elaborate these via analysis of a simple, linear example. But first we explicitly define the composition of DA and PI as the set

$$\mathcal{Q}_{\text{da+pi}}\big(\mathbb{P}_{X,Y}\big) \coloneqq \mathcal{Q}_{\text{pi}}\big(\mathbb{P}_{GX,Y}\big) = \mathcal{Q}_{\text{pi}}\Big(\mathbb{P}_{X,Y}^{\text{do}(X:=GX)}\Big).$$

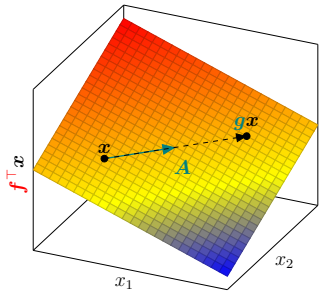


Figure 2: The ground truth function $\mathbf{f}$ in Example 1. The DA applied here corresponds to randomly translating the data samples along their level-set by adding random noise sampled from the null-space of $\mathbf{f}$.

Example 1 (a linear Gaussian DA example)**.** For $\sigma > 0$, non-zero $\boldsymbol{A}, \boldsymbol{T} \in \mathbb{R}^{*\times m}$ and $\boldsymbol{f}, \boldsymbol{\epsilon} \in \mathbb{R}^m$ in the following SEM, such that G, U, N_X, N_Y are conformable, centered Gaussian random vectors

$$X = \boldsymbol{T}^\top U + \sigma N_X, \qquad Y = \boldsymbol{f}^\top X + \boldsymbol{\epsilon}^\top U + \sigma N_Y, \qquad GX \coloneqq X + \boldsymbol{A}^\top G,$$

with $\text{range}\big(\boldsymbol{A}^\top\big) \subseteq \text{null}\big(\boldsymbol{f}^\top\big)$ so that GX is a valid outcome invariant DA transformation of X parameterized by G. Such a DA can be viewed as translating X along its level-set as shown in Fig. 2 and represents our prior knowledge about the symmetries of $\boldsymbol{f}$ for the purposes of this example. Finally, to recover the form Eq. (1), we simply set

$$\xi \coloneqq Y - \boldsymbol{f}^\top X.$$

Now, the task is to improve partial identification of $\boldsymbol{f}$ over the standard baseline using DA. For covariance $\boldsymbol{\Sigma}_X$ in Example 1, the causal excess risk used in our evaluation metrics takes the form

$$E^{\text{do}(X)}(\boldsymbol{h}) = \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \qquad E^{\text{do}(\boldsymbol{x})}(\boldsymbol{h}) = \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{x}\boldsymbol{x}^\top}. \tag{5}$$

Prior works also use similar formulations to measure causal estimation error (Vankadara et al., 2022; Kania & Wit, 2023; Akbar et al., 2025) or capture some notion of strength of confounding (Janzing, 2019; Janzing & Schölkopf, 2018a;b). The rest of this work uses the following constraint set for PI.

Assumption 1 (a bounded confounding sensitivity model)**.** Take the constraint set $\mathcal{C}_{\text{pi}}(\boldsymbol{\Gamma})$ such that

$$\mathcal{C}_{\text{pi}}(\boldsymbol{\Gamma}) \coloneqq \left\{ \mathbb{Q} = \mathcal{N}(\mathbf{0}, \cdot) \,\middle|\, \frac{\text{Var}(\mathbb{E}[\xi \,|\, X])}{\text{Var}(\xi)} \leq \Gamma, \quad \text{Var}(\xi) \leq \Gamma_0 \right\}, \qquad \boldsymbol{\Gamma} \coloneqq [\Gamma_0, \Gamma]^\top$$

where *confounding strength* $\Gamma \geq 0$ determines our assumption on the variation in ξ explained by X.

Assumption 1 adopts the widely used partial R-squared sensitivity model (Cinelli & Hazlett, 2019), itself a generalization of the classic Rosenbaum (2002). While we employ this model in our analyses, we do not restrict ourselves to it: under the linear Gaussian setting of Example 1, several other families of partial identification and sensitivity models yield equivalent constraints. Our results thus carry broader implications for PI and sensitivity analysis, as we further discuss in Section 7.

4.1 Valid bounds with data augmentation

First and foremost, we want to determine if the post-DA bounds are valid. This leads us to the result:

Theorem 1 (Valid bounds with DA)**.** *In Example 1, under Assumption 1, for some slack $s \geq 0$,*

$$E^{\text{do}(X)}_{\text{approx}}(\mathcal{Q}_{\text{da+pi}}) + s \leq E^{\text{do}(X)}_{\text{approx}}(\mathcal{Q}_{\text{pi}}), \quad \textit{equality iff} \quad \mathbb{P} \in \mathcal{Q}_{\text{pi}}, \quad \textit{or} \quad \mathbb{E}[GX \,|\, G] \perp_{\text{a.s.}} \mathbb{E}[X \,|\, \xi].$$

Proof. See Appendix A.1 for the proof. □

Which is to say that the identified set $\mathcal{H}_{\mathrm{da+pi}}$ is no farther from the true solution $\boldsymbol{f}$ compared to the original set $\mathcal{H}_{\mathrm{pi}}$, and is strictly closer to $\boldsymbol{f}$ so long as the DA perturbes spurious features of X, which is what the last equality-iff condition signifies. Of course it follows that when $\mathcal{Q}_{\mathrm{pi}}$ contains the true joint distribution $\mathbb{P}$, then $\boldsymbol{f} \in \mathcal{H}_{\mathrm{pi}}$ and so we should also have $\boldsymbol{f} \in \mathcal{H}_{\mathrm{da+pi}}$. Instead of such a simple set inclusion criteria, we keep the more general approximation error framing of Theorem 1 because we also position DA as a tool for improved sensitivity analysis where the constraint set may not necessarily be valid for some values of $\boldsymbol{\Gamma}$. Theorem 1 is then reassuring that under outcome invariance, DA at the very least should not cause $\mathcal{H}_{\mathrm{pi}}$ to drift away form the true solution.

It immediately follows from Theorem 1 that when $\mathbb{P} \in \mathcal{Q}_{\mathrm{pi}}$, we will also have $\boldsymbol{f}^\top \boldsymbol{x} \in \mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x}), \mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$. It is difficult, however, to show a similar general result as Theorem 1 for the point-wise evaluation of $E^{\mathrm{do}(\boldsymbol{x})}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{da+pi}})$ vs. $E^{\mathrm{do}(\boldsymbol{x})}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}})$ when $\mathbb{P} \notin \mathcal{Q}_{\mathrm{pi}}$. In this case the behavior of approximation error also depends on the alignment of unknown confounding parameters with our query vector $\boldsymbol{x}$, and both can be arbitrary. Nevertheless, we explore point-wise evaluation in our experiments.

4.2 Better bounds with data augmentation

Now that we have established that DA yields valid PI bounds, next we see if the resulting bounds can be, in some way, better than the baseline PI bounds. We present two results to support this claim.

Proposition 1 (Sharper bounds with DA)**.** *In Example 1, under Assumption 1, it holds that*

$$|\mathcal{H}_{\mathrm{da+pi}}| < |\mathcal{H}_{\mathrm{pi}}|, \qquad |\mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})| \leq |\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})|, \qquad \textit{equality iff} \quad \boldsymbol{x} \perp_{\mathrm{a.s.}} \mathbb{E}[GX \,|\, G]$$

where $|\cdot|$ denotes the Lebesgue measure (volume).

Proof. See Appendix A.3 for the proof. □

Proposition 1 states that the hypothesis set $\mathcal{H}_{\mathrm{da+pi}}$ is strictly smaller than the baseline $\mathcal{H}_{\mathrm{pi}}$. The same holds true for the intervals $\mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})$ vs. $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$, unless the query point $\boldsymbol{x}$ is orthogonal to the features that DA perturbs[2], in which case the size of the interval remains the same.

Although smaller identified sets/ intervals are in general desirable, size alone may not be the most appropriate measure of "goodness" of the identified set. We base the next result on worst-case excess risk.

Theorem 2 (Robust, informative bounds with DA)**.** *In Example 1, with Assumption 1, slack $s \geq 0$,*

$$E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{da+pi}}) + s \leq E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}), \qquad \textit{equality iff} \qquad \mathbb{E}[GX \,|\, G] \perp_{\mathrm{a.s.}} \mathbb{E}[X \,|\, \xi],$$

Proof. See Appendix A.2 for the proof. □

The Theorem 2 essentially states that when DA perturbes spurious features of the treatment, the identified set performs strictly better in terms of worst-case excess risk. Of course when spurious features are not perturbed, we are guaranteed to perform no worse than baseline PI. The worst case excess risk essentially bounds how bad the performance of any one hypothesis in the identified set may be, making our decision more robust and reliable.

5 Related Work

Our work connects two distinct but related lines of research: the vast literature on partial identification and sensitivity analysis for unobserved confounding, and the emerging field of causal data augmentation. We position our contribution as a simple, powerful tool that complements the former while offering a more practical alternative to the latter.

Partial Identification and Sensitivity Analysis. When causal effects are not point-identifiable due to unobserved confounders, partial identification (PI) is the standard framework for deriving

[2]Intuitively, this would be like rotating an image of a centered circle.

bounds on the true effect. The primary challenge in PI lies in specifying a credible set of assumptions, encoded in the constraint set $\mathcal{Q}_{\text{pi}}$, to define the space of possible data-generating processes. A rich body of work has proposed various methods for this. Classic approaches focus on sensitivity parameters that bound the influence of confounders, such as through odds ratios (Rosenbaum, 2002), propensity scores (Tan, 2006), or the proportion of unexplained variance (partial R-squared) (Cinelli & Hazlett, 2020). More recent work has framed PI as a robust optimization problem, constructing $\mathcal{Q}_{\text{pi}}$ as a total variation or Wasserstein ball around the observed data distribution (Fan et al., 2024; Guo et al., 2022; Meresht et al., 2022). An alternative is to directly model the *response function* itself, for instance using GANs (Hu et al., 2021; Padh et al., 2023). Many of these methods can be further refined by incorporating auxiliary variables, such as instrumental variables (IVs), to tighten the resulting bounds (Balke & Pearl, 1997; Schweisthal et al., 2025; Kilbertus et al., 2020; Gunsilius, 2020). Our work is orthogonal and complementary to all of these approaches. We do not propose a new method for constructing $\mathcal{Q}_{\text{pi}}$; rather, we introduce outcome-invariant DA as a simple data pre-processing step that can be applied in conjunction with many existing PI framework to sharpen its resulting bounds and improve robustness.

Counterfactual Data Augmentation. The causal analysis of DA has primarily been explored through the lens of *counterfactual data augmentation* (Ilse et al., 2021; Yuan et al., 2024; Feder et al., 2023; Pitis et al., 2022; Armengol Urpí et al., 2024; Mahajan et al., 2021; Aloui et al., 2023). These methods typically aim to improve a model's robustness to treatment interventions by synthesizing counterfactual training examples. However, they often rely on strong structural assumptions that limit their practical applicability, such as requiring access to the full structural equation model (SEM) (Yuan et al., 2024; Feder et al., 2023), specific auxiliary variables (Ilse et al., 2021; Feder et al., 2023; Mahajan et al., 2021; Aloui et al., 2023), or the complete causal graph (Pitis et al., 2022; Armengol Urpí et al., 2024). In contrast, our approach requires only the weaker, more practical assumption of outcome-invariance, which is often informed by prior domain knowledge about symmetries. A similar approach was undertaken by Akbar et al. (2025) for the task of point-estimation of causal effects even though identification itself if not possible with outcome-invariance alone. We therefore provide a more principled approach in this setting by instead focusing on partial-identification and sensitivity analysis.

6 Experiments

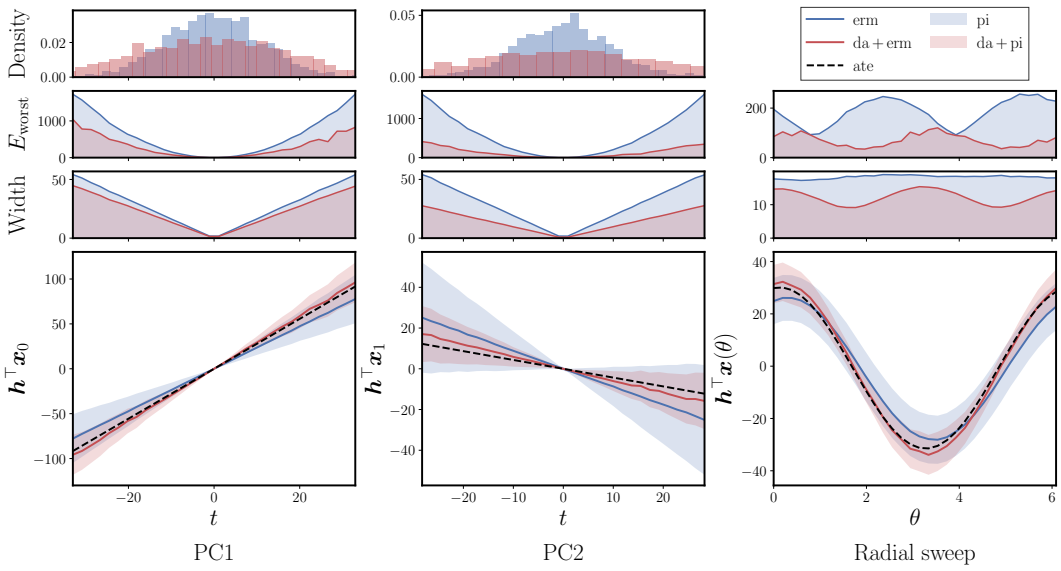


Figure 3: Data augmentation consistently sharpens partial identification bounds in a linear simulation. Across query points aligned with principal components (PC1, PC2) and a radial sweep, DA+PI (red) yields narrower intervals, lower worst-case excess risk (E_{worst}), and predictions closer to the true average treatment effect (ATE, black dash) compared to baseline PI (blue).

We began by presenting theoretical results in the infinite-sample setting to emphasize that simple pre-processing with data augmentation sharpens bounds in partial identification. In this section, we turn to the finite-sample regime and empirically evaluate the effectiveness of DA in PI. Importantly, we do not use DA for its conventional purpose of augmenting data to improve estimation variance, even though as mentioned Section 4 that is an important source of uncertainty in PI. This is because the use-case of DA for this is self-evident. Rather, we focus on the much more interesting setting where we fix the number of samples in the augmented dataset to match that of the original dataset throughout all experiments.

6.1 Simulation experiment

For the finite sample results of the linear Gaussian SEM from Example 1, by taking $m = 32$, $k = 31$ (dimension of G), $\sigma = 0.1$ and fixing $\boldsymbol{\tau}^\top = \mathbf{0}$, we sample a $\boldsymbol{f}, \boldsymbol{\epsilon}$ and $\boldsymbol{T} \in \mathbb{R}^{m \times m}$ from a standard normal distribution and then keep it fixed throughout the experiment. We construct a $\boldsymbol{A} \coloneqq \boldsymbol{V}_0$ with k rows as orthonormal basis of null($\boldsymbol{f}$), such that the SVD of $\boldsymbol{f}$ is

$$\boldsymbol{f} = [\boldsymbol{u} \quad \boldsymbol{U}_0] \begin{bmatrix} \sigma & \mathbf{0}_{1\times(m-1)} \\ \mathbf{0}_{(m-1)\times 1} & \mathbf{0}_{(m-1)\times(m-1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{V}_0^\top \end{bmatrix}.$$

Although this construction of $\boldsymbol{A}$ relies on direct knowledge of $\boldsymbol{f}$ (which is unavailable in practice), we include it here purely for illustrative purposes. We treat access to $\boldsymbol{A}$ as our prior knowledge about the symmetries of $\boldsymbol{f}$, noting that this information alone is insufficient to recover $\boldsymbol{f}$.

We then generate $n = 2048$ samples of (X, Y). For ERM we use a closed form linear OLS solution. And for PI we use the partial R-squared sensitivity model from Assumption 1 for a range of query points $\boldsymbol{x}_0, \boldsymbol{x}_1$ with the sensitivity parameter set as $\Gamma = \Gamma_0 = 20$.

To visualize the results, we chose $\boldsymbol{x}_0 \coloneqq t \cdot \boldsymbol{u}_0$ and $\boldsymbol{x}_1 \coloneqq t \cdot \boldsymbol{u}_1$ where $\boldsymbol{u}_0$, $\boldsymbol{u}_1$ are the first and second principal components of the data. We then sweep t over ± 3 standard-deviations, computing intervals $\mathcal{H}_{\text{pi}}(\boldsymbol{x}), \mathcal{H}_{\text{da+pi}}(\boldsymbol{x})$ via convex programming (separately for the upper and lower bounds). The results are shown in Fig. 3 (left, center). Fig. 3 (right) also shows a radial sweep over $\theta \in [0, 2\pi]$ to generate queries $\boldsymbol{x}(\theta) = \sigma_0 \cdot \sin(\theta) \cdot \boldsymbol{u}_0 + \sigma_1 \cdot \cos(\theta) \cdot \boldsymbol{u}_1$.

6.2 Optical device experiment

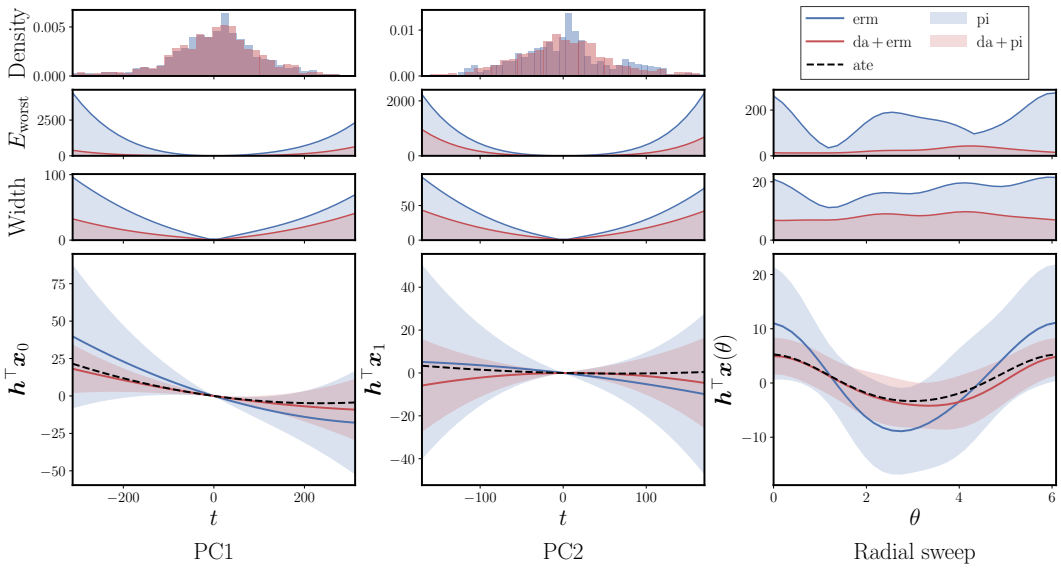


Figure 4: Our method sharpens causal bounds on the real-world Optical Device dataset. Even with complex, non-linear relationships, applying outcome-invariant DA (red) substantially narrows the partial identification bounds compared to the baseline (blue).

The dataset from Janzing & Schölkopf (2018b) consists of 3×3 pixel images X displayed on a laptop screen that cause voltage readings Y across a photo-diode. A hidden confounder U controls two LEDs; one affects the webcam capturing X, the other affects the photo-diode measuring Y. The ground-truth predictor $\boldsymbol{f}$ is computed by first regressing Y on $(\phi(X), U)$, where $\phi(X)$ are polynomial features of X with degree $d \in \{1, \cdots, 5\}$ that best explains the data (degree 2 in our case). The component corresponding to U is then removed to recover $\boldsymbol{f}$. Our choice of DA on X includes additive Gaussian noise $G \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_X/10)$, random vertical/horizontal flips and 90^0 rotations for DA. We then compute the features $\phi(GX)$ to be used with PI, setting $\Gamma = \Gamma_0 = 10$ for the partial R-squared model from Assumption 1 on a datasets of $n = 1000$ samples. Figure 4 shows that DA+PI sharpens bounds over the PI baseline. The visualization approach in the same as in Section 6.1.

7 Limitations, assumptions and future work

Outcome invariance. Our approach hinges on the untestable assumption that the chosen data augmentations are outcome-invariant. While this requires prior knowledge, our framework has a benign failure mode: a valid outcome-invariant DA is guaranteed not to worsen the baseline PI bounds, even if it fails to target spurious features. Additionally, we would like to remind the readers that untestable domain knowledge is fundamentally unavoidable for making any causal conclusion from observational data Pearl (2009), as is the norm in partial identificaiton. This also includes access to auxiliary variables since the conditional independences that they represent are also merely untestable assumptions. Furthermore, we argue that our assumptions on DA are actually quite practical given that a symmetry-based DA model has precedence in the DA and invariance literature (Chen et al., 2020; Lyle et al., 2020; Shao et al., 2022; Fawzi & Frossard, 2015; Dubois et al., 2021; Petrache & Trivedi, 2023; Montasser et al., 2024; Romero & Lohit, 2022; Zhu et al., 2021; Wong et al., 2016).

Additional covariates. Many works in partial identification and sensitivity analysis leverage access to additional auxiliary variables, such as instrumental variables (IVs) and observable confounders or back-doors Kilbertus et al. (2020); Padh et al. (2023). Even though we do not explicitly model these to keep our analysis simple and tractable, we argue that out DA framing is still compatible with them–for example, applying DA on X does not invalidate an IV that enters into X.

Additional partial identification approaches. Many sensitivity and partial identificaiton models can be reduced to the constraints in Assumption 1. These include, of course, the partial R-squared model, Rosenbaum, MSM (under a mild bounded marginal ratio assumption), as well as DRO, Wasserstein, total-variation approaches. While a rigorous analysis is left for future work, it is important to specify that our resents here are more general than just the partial R-squared model.

8 Conclusion

In this work, we investigate the use of data augmentation (DA) for sharpening bounds in the partial identification (PI) of causal effects from observational data with unobserved confounding. While prior work has shown that outcome-invariant DA can mitigate confounding bias in point-estimation, full identification is often not possible. Here, we extend this line of reasoning to the more general setting of PI. We establish that outcome-invariant DA can be framed as a transformation intervention on the treatment variable. This perspective allows us to leverage DA not just for mitigating estimation variance, but as a tool to implicitly impose additional valid constraints on the PI problem.

Our theoretical analysis in a linear Gaussian setting demonstrates that composing PI with DA yields bounds that are provably valid, sharper in terms of interval width and hypothesis set volume, and more robust by offering lower worst-case excess risk. This provides a computationally inexpensive and readily available method for practitioners to strengthen causal inferences, especially in settings with complex, high-dimensional data where specifying other constraints can be difficult. Ultimately, this work solidifies the role of data augmentation as a versatile tool for causal inference, extending its utility from reducing bias in causal effect estimation to improving the precision and reliability of bounds when effects are only partially identified.

Ethics Statement

The authors have read and adhered to the ICLR Code of Ethics. This work is primarily theoretical and methodological, focusing on the mathematical foundations of using data augmentation for partial identification. The experimental validation relies on a synthetic dataset generated for illustrative purposes and a standard, publicly available benchmark dataset (Optical Device). No human subjects were involved in this research, no new data was collected, and therefore, no Institutional Review Board (IRB) approval was required. The goal of this research is to improve the rigor and reliability of causal inference from observational data, which can lead to more robust and fair decision-making in various applications. We do not foresee any direct negative ethical implications or societal consequences stemming from this work.

Reproducibility Statement

We are committed to ensuring the reproducibility of our work. All theoretical claims made in this paper are supported by detailed, step-by-step proofs, which can be found in the Appendix. The experimental setup for both the simulation study and the real-data experiment is described in Section 6. The complete source code to reproduce all experiments, figures, and results is included as supplementary material with this submission. The code is commented and contains all necessary implementation details, including hyperparameter settings and the specific data generation process for the simulation.

References

Uzair Akbar, Niki Kilbertus, Hao Shen, Krikamol Muandet, and Bo Dai. An analysis of causal effect estimation using outcome invariant data augmentation. In *Advances in Neural Information Processing Systems*, volume 38. NeurIPS, 2025.

Ahmed Aloui, Juncheng Dong, Cat P. Le, and Vahid Tarokh. Counterfactual data augmentation with contrastive learning, 2023. arXiv:2311.03630.

Núria Armengol Urpí, Marco Bagatella, Marin Vlastelica, and Georg Martius. Causal action influence aware counterfactual data augmentation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 1709–1729. PMLR, 2024.

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

Gavin Brown and Riccardo Ali. Bias/variance is not the same as approximation/estimation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=4TnFbv16hK.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 12 2019. ISSN 1369-7412. doi: 10.1111/rssb.12348. URL https://doi.org/10.1111/rssb.12348.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 2020.

Hugh Dance and Benjamin Bloem-Reddy. Causal inference with cocycles, 2024. arXiv:2405.13844.

Y. Dubois et al. Lossy compression for lossless prediction. In *NeurIPS*, 2021.

Qiuling Fan, Minghao Li, and Zhengling Wang. Distributionally robust optimization approaches for causal inference with unobserved confounding. *Journal of Machine Learning Research*, 25(87): 1–45, 2024.

A. Fawzi and P. Frossard. Manitest: Are classifiers really invariant? In *BMVC*, 2015.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. Data augmentations for improved (large) language model generalization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70638–70653. Curran Associates, Inc., 2023.

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ikX6D1oM1c.

Florian Gunsilius. A path-sampling method to partially identify causal effects in instrumental variable models, 2020. URL https://arxiv.org/abs/1910.09502.

Wenshuo Guo, Mingzhang Yin, Yixin Wang, and Michael Jordan. Partial identification with noisy covariates: A robust optimization approach. In *First Conference on Causal Learning and Reasoning*, 2022. URL https://openreview.net/forum?id=-NVBxy0TdU.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. A generative adversarial framework for bounding confounded causal effects. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12104–12112, May 2021. doi: 10.1609/aaai.v35i13.17437. URL https://ojs.aaai.org/index.php/AAAI/article/view/17437.

Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, pp. 4555–4562. PMLR, 2021.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3598769.

Dominik Janzing. Causal regularization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dominik Janzing and Bernhard Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2245–2253. PMLR, 2018a.

Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018b.

Lucas Kania and Ernst Wit. Causal regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees, 2023. arXiv:2205.01593.

Niki Kilbertus, Matt J. Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20108–20119, 2020.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks, 2020. arXiv:2005.00178.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7313–7324. PMLR, 2021.

Vahid Balazadeh Meresht, Vasilis Syrgkanis, and Rahul G Krishnan. Partial identification of treatment effects with implicit generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=8cUGfg-zUnh.

O. Montasser et al. Transformation-invariant learning and theoretical guarantees for ood generalization. In *NeurIPS*, 2024.

Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, and Niki Kilbertus. Stochastic causal programming for bounding treatment effects. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 142–176. PMLR, 11–14 Apr 2023. URL https://proceedings.mlr.press/v213/padh23a.html.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

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

M. Petrache and S. Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *NeurIPS*, 2023.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. MoCoDA: Model-based counterfactual data augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18143–18156, 2022.

D. Romero and S. Lohit. Learning partial equivariances from data. In *NeurIPS*, 2022.

Paul R Rosenbaum. *Observational studies*. Springer, 2002.

Jonas Schweisthal, Dennis Frauen, Maresa Schröder, Konstantin Hess, Niki Kilbertus, and Stefan Feuerriegel. Learning representations of instruments for partial identification of treatment effects. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=pgrJPhsk2w.

H. Shao et al. A theory of pac learnability under transformation invariances. In *NeurIPS*, 2022.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60, 2019.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. doi: 10.1198/016214506000000023. URL https://doi.org/10.1198/016214506000000023.

Chennuru Vankadara, Luca Rendsburg, Ulrike von Luxburg, and Debarghya Ghoshdastidar. Interpolation and regularization for causal learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Vladimir Naumovich Vapnik. *Statistical learning theory*. Wiley, New York, 1998.

S. Wong et al. Understanding data augmentation for classification: When to warp? In *DICTA*, 2016.

Liyuan Xu and Arthur Gretton. A neural mean embedding approach for back-door and front-door adjustment, 2022. arXiv:2210.06610.

Jianhao Yuan, Francesco Pinto, Adam Davies, and Philip Torr. Not just pretty pictures: Toward interventional data augmentation using text-to-image generators. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=b89JtZj9gm.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1), 2023.

S. Zhu et al. Understanding the generalization benefit of model invariance from a data perspective. In *NeurIPS*, 2021.

A PROOFS

A.1 PROOF OF THEOREM 1 – VALID BOUNDS WITH DA

Theorem 1 (Valid bounds with DA)**.** *In Example 1, under Assumption 1, for some slack $s \geq 0$,*

$$E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{da+pi}}) + s \leq E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}}), \quad \textit{equality iff} \quad \mathbb{P} \in \mathcal{Q}_{\mathrm{pi}}, \quad \textit{or} \quad \mathbb{E}[GX \,|\, G] \perp_{\mathrm{a.s.}} \mathbb{E}[X \,|\, \xi].$$

Proof. From Lemma 5, we can characterize the identified sets $\mathcal{H}_{\mathrm{pi}}$, $\mathcal{H}_{\mathrm{da+pi}}$ as ellipsoids of the form

$$\mathcal{H}_{\mathrm{pi}} = \Big\{ \boldsymbol{h} \,\Big|\, \|\boldsymbol{h} - \boldsymbol{h}_{\mathrm{erm}}\|^2_{\boldsymbol{\Sigma}_X} \leq r(\boldsymbol{\Gamma}) \Big\}, \qquad \mathcal{H}_{\mathrm{da+pi}} = \Big\{ \boldsymbol{h} \,\Big|\, \|\boldsymbol{h} - \boldsymbol{h}_{\mathrm{da+erm}}\|^2_{\boldsymbol{\Sigma}_{GX}} \leq r(\boldsymbol{\Gamma}) \Big\}.$$

First consider $\mathbb{P} \notin \mathcal{Q}_{\mathrm{pi}}$. Now, from the definition of approximation error in Section 2.3 it follows

$$\begin{aligned}
E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}}) &= \min_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{pi}}.) \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&= \Big(\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} - \sqrt{r(\boldsymbol{\Gamma})} \Big)^2, && \text{(Lemma 1)}
\end{aligned}$$

where $r(\boldsymbol{\Gamma})$ is some constant entirely determined by $\boldsymbol{\Gamma}$. Now, we do a similar exercise with $\mathcal{Q}_{\mathrm{da+pi}}$,

$$\begin{aligned}
E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{da+pi}}) &= \min_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{da+pi}}.) \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{da+pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&= \left(1 - \frac{\sqrt{r(\boldsymbol{\Gamma})}}{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_{GX}}} \right)^2 \|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X} - s, && \text{(Lemma 1, } \exists s \geq 0.) \\
&= \Big(\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_{GX}} - \sqrt{r(\boldsymbol{\Gamma})} \Big)^2 \frac{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}}{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_{GX}}} - s, \\
&\leq \Big(\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} - \sqrt{r(\boldsymbol{\Gamma})} \Big)^2 - s,
\end{aligned}$$

where the last inequality follows from a similar approach as used in Proposition 2 and Lemma 4

$$\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X} \leq \|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_{GX}} \leq \|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X},$$

and equality holds iff $\mathbb{E}[GX \,|\, G] \perp_{\mathrm{a.s.}} \mathbb{E}[X \,|\, \xi]$. The case for $\mathbb{P} \in \mathcal{Q}_{\mathrm{pi}}$ is trivial from Lemma 1. □

A.2 Proof of Theorem 2 – Robust, informative bounds with DA

Theorem 2 (Robust, informative bounds with DA)**.** *In Example 1, with Assumption 1, slack $s \geq 0$,*

$$E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{da+pi}}) + s \leq E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}), \qquad \textit{equality iff} \qquad \mathbb{E}[GX \,|\, G] \perp_{\mathrm{a.s.}} \mathbb{E}[X \,|\, \xi],$$

Proof. From Lemma 5, we can characterize the identified sets $\mathcal{H}_{\mathrm{pi}}$, $\mathcal{H}_{\mathrm{da+pi}}$ as ellipsoids of the form

$$\mathcal{H}_{\mathrm{pi}} = \Big\{ \boldsymbol{h} \,\Big|\, \|\boldsymbol{h} - \boldsymbol{h}_{\mathrm{erm}}\|^2_{\boldsymbol{\Sigma}_X} \leq r(\boldsymbol{\Gamma}) \Big\}, \qquad \mathcal{H}_{\mathrm{da+pi}} = \Big\{ \boldsymbol{h} \,\Big|\, \|\boldsymbol{h} - \boldsymbol{h}_{\mathrm{da+erm}}\|^2_{\boldsymbol{\Sigma}_{GX}} \leq r(\boldsymbol{\Gamma}) \Big\}.$$

Now, from the definition of worst-case excess error in Section 2.3 it follows

$$\begin{aligned}
E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}) &= \max\nolimits_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \\
&= \max\nolimits_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{pi}}.) \\
&= \max\nolimits_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&= \Big(\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} + \sqrt{r(\boldsymbol{\Gamma})} \Big)^2, && \text{(Lemma 2)}
\end{aligned}$$

where $r(\boldsymbol{\Gamma})$ is some constant entirely determined by $\boldsymbol{\Gamma}$. Now, we do a similar exercise with $\mathcal{Q}_{\mathrm{da+pi}}$,

$$\begin{aligned}
&\Rightarrow E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{da+pi}}) \\
&= \max\nolimits_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \\
&= \max\nolimits_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{da+pi}}.) \\
&= \max\nolimits_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{da+pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&= \left(\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} + \sqrt{r(\boldsymbol{\Gamma}) \cdot \lambda_{\max}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}^{-1}_{GX}\big)} \right)^2 - s', && \text{(Lemma 2, } \exists s' \geq 0.) \\
&= \Big(\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} + \sqrt{r(\boldsymbol{\Gamma})} \Big)^2 - s, && \text{(} \exists s \geq s' \text{, since } \lambda_{\max}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}^{-1}_{GX}\big) \leq 1.) \\
&\leq \Big(\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} - \sqrt{r(\boldsymbol{\Gamma})} \Big)^2 - s,
\end{aligned}$$

where the last inequality follows from a similar approach as used in Proposition 2 and Lemma 4

$$\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X} \leq \|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X},$$

and equality holds iff $\mathbb{E}[GX \,|\, G] \perp_{\mathrm{a.s.}} \mathbb{E}[X \,|\, \xi]$.

□

A.3 Proof of Proposition 1 – Sharper bounds with DA

Proposition 1 (Sharper bounds with DA)**.** *In Example 1, under Assumption 1, it holds that*

$$|\mathcal{H}_{\mathrm{da+pi}}| < |\mathcal{H}_{\mathrm{pi}}|, \qquad |\mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})| \leq |\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})|, \qquad \textit{equality iff} \quad \boldsymbol{x} \perp_{\mathrm{a.s.}} \mathbb{E}[GX \,|\, G]$$

where $|\cdot|$ denotes the Lebesgue measure (volume).

Proof. We provide a sketch for a simple proof.

Ellipsoid volume. The result for $\mathcal{H}_{\mathrm{pi}}$ and $\mathcal{H}_{\mathrm{da+pi}}$ is straightforward by noting that these are ellipsoids from Lemma 5. Since the ellipsoids have the same radius but different shape parameters $\boldsymbol{\Sigma}_X$ (for) and $\boldsymbol{\Sigma}_{GX}$ (for), it follows that $\mathcal{H}_{\mathrm{pi}}$ has bigger volume than $\mathcal{H}_{\mathrm{da+pi}}$ since $\boldsymbol{\Sigma}_X \leq \boldsymbol{\Sigma}_{GX} = \boldsymbol{\Sigma}_X + \boldsymbol{A}^\top \boldsymbol{\Sigma}_G \boldsymbol{A}$.

Interval width. We again point to Lemma 5 to note that the width of $(\mathcal{H})_{\mathrm{pi}}\boldsymbol{x}$ and $(\mathcal{H})_{\mathrm{da+pi}}\boldsymbol{x}$ is only determined by (a scalar multiple of) $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}}$ and $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}$ respectively. Again, since $\boldsymbol{\Sigma}_X \leq \boldsymbol{\Sigma}_{GX} = \boldsymbol{\Sigma}_X + \boldsymbol{A}^\top \boldsymbol{\Sigma}_G \boldsymbol{A}$, we have that $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}$ is strictly smaller than $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}}$ unless $\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{\Sigma}_G \boldsymbol{A} = 0$, which from Lemma 4 is equivalent to saying that

$$\boldsymbol{x} \perp_{\mathrm{a.s.}} \mathbb{E}[GX \,|\, G].$$

□

A.4 Proposition — Causal estimation with DA

Proposition 2 (causal estimation with DA+ERM)**.** *For the SEM in Example 1, the following holds:*

$$E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{da+erm}}) \leq E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{erm}}), \qquad \textit{equality iff} \qquad \mathbb{E}[GX\,|\,G] \perp_{\mathrm{a.s.}} \mathbb{E}[X\,|\,\xi] \qquad \textit{a.s.}$$

Proof. We have

$$\begin{aligned}
&\Rightarrow \|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} \\
&= \left\| \mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1} \mathbb{E}\big[(GX)Y^\top\big] - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, \\
&= \left\| \mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1} \mathbb{E}\Big[(GX)\big(\boldsymbol{f}^\top X + \xi\big)^\top\Big] - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Structural eq. of } Y\text{.)} \\
&= \left\| \mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1} \mathbb{E}\Big[(GX)\big(\boldsymbol{f}^\top (GX) + \xi\big)^\top\Big] - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Using } \mathcal{G}\text{-invariance of } \boldsymbol{f}\text{.)} \\
&= \left\| \left(\boldsymbol{f} + \mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1} \mathbb{E}\big[(GX)\xi^\top\big] \right) - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, \\
&= \left\| \mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1} \mathbb{E}\big[(GX)\xi^\top\big] \right\|_{\boldsymbol{\Sigma}_X}, \\
&= \left\| \mathbb{E}\Big[\Big(X + \tilde{G}\Big)\Big(X + \tilde{G}\Big)^\top\Big]^{-1} \mathbb{E}\Big[\Big(X + \tilde{G}\Big)\xi^\top\Big] \right\|_{\boldsymbol{\Sigma}_X}, \\
& && \text{(Where } \tilde{G} \coloneqq \mathbb{E}[GX\,|\,G] = \gamma \cdot \boldsymbol{A}^\top G\text{.)} \\
&= \left\| \Big(\mathbb{E}\big[XX^\top\big] + \mathbb{E}\Big[\tilde{G}\tilde{G}^\top\Big]\Big)^{-1} \mathbb{E}\big[X\xi^\top\big] \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Using } \tilde{G} \perp\!\!\!\perp X, \xi\text{.)} \\
&= \left\| \big(\boldsymbol{S}^\top \boldsymbol{S} + \boldsymbol{S}^\top \boldsymbol{D} \boldsymbol{S}\big)^{-1} \mathbb{E}\big[X\xi^\top\big] \right\|_{\boldsymbol{S}^\top \boldsymbol{S}}, && \text{(Lemma 3.)} \\
&= \left\| \boldsymbol{S}^{-1} (\mathbf{I}_m + \boldsymbol{D})^{-1} \boldsymbol{S}^{-\top} \mathbb{E}\big[X\xi^\top\big] \right\|_{\boldsymbol{S}^\top \boldsymbol{S}}, && (\boldsymbol{S}, \boldsymbol{S}^\top \text{ invertible.)} \\
&= \left\| \boldsymbol{S}\boldsymbol{S}^{-1} (\mathbf{I}_m + \boldsymbol{D})^{-1} \boldsymbol{S}^{-\top} \mathbb{E}\big[X\xi^\top\big] \right\|, && \text{(Switch to } \ell_2 \text{ norm.)} \\
&= \left\| (\mathbf{I}_m + \boldsymbol{D})^{-1} \boldsymbol{S}^{-\top} \mathbb{E}\big[X\xi^\top\big] \right\|, \\
&\leq \left\| \boldsymbol{S}^{-\top} \mathbb{E}\big[X\xi^\top\big] \right\|, && (6) \\
&= \left\| \boldsymbol{S}\boldsymbol{S}^{-1} \boldsymbol{S}^{-\top} \mathbb{E}\big[X\xi^\top\big] \right\|, && \text{(Substitute in } \mathbf{I}_m = \boldsymbol{S}\boldsymbol{S}^{-1}\text{.)} \\
&= \left\| \boldsymbol{S}^{-1} \boldsymbol{S}^{-\top} \mathbb{E}\big[X\xi^\top\big] \right\|_{\boldsymbol{S}^\top \boldsymbol{S}}, && \text{(Back to weighted norm.)} \\
&= \left\| \mathbb{E}\big[XX^\top\big]^{-1} \mathbb{E}\big[X\xi^\top\big] \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Substitute in } \boldsymbol{\Sigma}_X \coloneqq \mathbb{E}\big[XX^\top\big] = \boldsymbol{S}^\top \boldsymbol{S}\text{.)} \\
&= \left\| \boldsymbol{f} + \mathbb{E}\big[XX^\top\big]^{-1} \mathbb{E}\big[X\xi^\top\big] - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Add and subtract } \boldsymbol{f}\text{.)} \\
&= \left\| \mathbb{E}\big[XX^\top\big]^{-1} \big(\mathbb{E}\big[XX^\top\big]\boldsymbol{f} + \mathbb{E}\big[X\xi^\top\big]\big) - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Use } \mathbf{I}_m = \mathbb{E}\big[XX^\top\big]^{-1}\mathbb{E}\big[XX^\top\big]\text{.)} \\
&= \left\| \mathbb{E}\big[XX^\top\big]^{-1} \mathbb{E}\Big[X\big(\boldsymbol{f}^\top X + \xi\big)^\top\Big] - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Linearity of expectation.)} \\
&= \left\| \mathbb{E}\big[XX^\top\big]^{-1} \mathbb{E}\big[XY^\top\big] - \boldsymbol{f} \right\|_{\boldsymbol{\Sigma}_X}, && \text{(Structural eq. of } Y\text{.)} \\
&= \|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X}, && \text{(ERM closed form solution.)}
\end{aligned}$$

where inequality Eq. (6) holds because $\boldsymbol{D}$ is non-negative diagonal. Furthermore, inequality Eq. (6) only holds with equality iff $\boldsymbol{S}^{-\top}\mathbb{E}\big[X\xi^\top\big]$ is in the kernel of $\boldsymbol{D}$. Or equivalently, iff $\mathbb{E}\big[X\xi^\top\big]$ is in the kernel of $\boldsymbol{S}^\top \boldsymbol{D}\boldsymbol{S} = \boldsymbol{\Sigma}_{\tilde{G}}$, which from Lemma 4 is true iff $\mathbb{E}[GX\,|\,G] \perp \mathbb{E}[X\,|\,\xi]$ a.s. □

A.5 MISCELLANEOUS SUPPORTING LEMMAS

Lemma 1 (Upper bound on distance of a point to an ellipsoid)**.** *Take the following ellipsoid* $\mathcal{O} \subset \mathbb{R}^n$

$$\mathcal{O} = \Big\{ \boldsymbol{x} \,\Big|\, (\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{\Sigma}_0 (\boldsymbol{x} - \boldsymbol{x}_0) \leq r_0 \Big\},$$

with radius r_0*, centered at* $\boldsymbol{x}_0$ *and shape defined by the SPD matrix* $\boldsymbol{\Sigma}_0 \succ 0$*. For some arbitrary point* $\boldsymbol{y} \in \mathbb{R}^n$*, denote its distance from* $\mathcal{O}$ *as weighted by an SPD* $\boldsymbol{\Sigma} \succ 0$ *with the following notation*

$$D^{\min}_{\boldsymbol{\Sigma}}(\boldsymbol{y}, \mathcal{O}) \coloneqq \min_{\boldsymbol{x} \in \mathcal{O}} \|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}.$$

This distance can be upper bounded with the following closed-form, and equality holds for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$

$$D^{\min}_{\boldsymbol{\Sigma}}(\boldsymbol{y}, \mathcal{O}) \leq \begin{cases} 0, & \boldsymbol{y} \in \mathcal{O}, \\ \left(1 - \dfrac{\sqrt{r_0}}{\|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}_0}}\right) \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}}, & \boldsymbol{y} \notin \mathcal{O}. \end{cases}$$

Proof. The result for $\boldsymbol{y} \in \mathcal{O}$ case is immediate. To show the bound for $\boldsymbol{y} \notin \mathcal{O}$, consider the ray

$$\boldsymbol{x}(r) := \boldsymbol{x}_0 + r \cdot (\boldsymbol{y} - \boldsymbol{x}_0), \quad r \in [0, 1],$$

going from the ellipsoid center $\boldsymbol{x}_0$ through $\boldsymbol{y}$. This ray intersects with the ellipsoid boundary at

$$r^* = \frac{\sqrt{r_0}}{\|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}_0}} \in (0, 1),$$

due to $\mathcal{O}$ being a sphere under a $\boldsymbol{\Sigma}_0$ weighted norm. The point $\boldsymbol{x}^* \coloneqq \boldsymbol{x}(r^*)$ lies on the boundary.

$$\Rightarrow \boldsymbol{y} - \boldsymbol{x}^* = (1 - r^*) \cdot (\boldsymbol{y} - \boldsymbol{x}_0).$$

Since the closest point along an arbitrary ray is never closer than the true minimum, we have

$$\begin{aligned} D^{\min}_{\boldsymbol{\Sigma}}(\boldsymbol{y}, \mathcal{O}) &= \min_{\boldsymbol{x} \in \mathcal{O}} \|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}, \\ &\leq \|\boldsymbol{y} - \boldsymbol{x}^*\|_{\boldsymbol{\Sigma}}, \\ &= (1 - r^*) \cdot \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}}, \\ &= \left(1 - \frac{\sqrt{r_0}}{\|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}_0}}\right) \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}}. \end{aligned}$$

□

Lemma 2 (Upper bound on distance of a point to farthest point on ellipsoid)**.** *Take ellipsoid* $\mathcal{O} \subset \mathbb{R}^n$

$$\mathcal{O} = \Big\{ \boldsymbol{x} \,\Big|\, (\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{\Sigma}_0 (\boldsymbol{x} - \boldsymbol{x}_0) \leq r_0 \Big\},$$

with radius r_0*, centered at* $\boldsymbol{x}_0$ *and shape defined by the SPD matrix* $\boldsymbol{\Sigma}_0 \succ 0$*. For some arbitrary point* $\boldsymbol{y} \in \mathbb{R}^n$*, denote its distance from the farthest point on* $\mathcal{O}$ *as weighted by an SPD* $\boldsymbol{\Sigma} \succ 0$ *with*

$$D^{\max}_{\boldsymbol{\Sigma}}(\boldsymbol{y}, \mathcal{O}) \coloneqq \max_{\boldsymbol{x} \in \mathcal{O}} \|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}.$$

This distance can be upper bounded with the following closed-form, and equality holds for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$*,*

$$D^{\max}_{\boldsymbol{\Sigma}}(\boldsymbol{y}, \mathcal{O}) \leq \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}} + \sqrt{r_0 \cdot \lambda_{\max}\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\big)}$$

Proof. By triangle inequality

$$\|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{\Sigma}} \leq \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}.$$

Now, simply maximizing both sides over $\boldsymbol{x} \in \mathcal{O}$,

$$\max_{\boldsymbol{x} \in \mathcal{O}} \|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{\Sigma}} \leq \max_{\boldsymbol{x} \in \mathcal{O}} (\|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}) = \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}} + \max_{\boldsymbol{x} \in \mathcal{O}} \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}.$$

The last term $\max_{\boldsymbol{x} \in \mathcal{O}} \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}$ is simply the radius of the ellipsoid in the $\boldsymbol{\Sigma}$–norm, which is equal to $\sqrt{r_0 \cdot \lambda_{\max}\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\big)}$ where $\lambda_{\max}\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\big)$ is the maximum eigenvalue of the product $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}$. Hence, the result follows for $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$. For the case $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, the ellipsoid just becomes a sphere and equality holds. □

Lemma 3 (SPD and PSD simultaneous denationalization via congruence)**.** *For any $n \times n$ matrices $\mathbf{A} \succ \mathbf{0}$, $\mathbf{B} \succcurlyeq \mathbf{0}$, there exists an invertible $\mathbf{S} \in \mathbb{R}^{n\times n}$ and non-negative diagonal $\mathbf{D} \in \mathbb{R}^{n\times n}$ such that*

$$\mathbf{A} = \mathbf{S}^\top \mathbf{S}, \qquad \mathbf{B} = \mathbf{S}^\top \mathbf{D}\mathbf{S}.$$

Proof. This is similar to Theorem 7.6.4 in (Horn & Johnson, 1985, p. 465) for two SPD matrices. We proceed similarly; Since $\mathbf{A}$ is SPD, it admits a unique SPD square root $\mathbf{A}^{1/2}$. Define

$$\mathbf{C} := \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2},$$

which is SPD. By the spectral theorem, there exists an orthogonal matrix $\mathbf{U}$ such that

$$\mathbf{C} = \mathbf{U}^\top \mathbf{D}\mathbf{U},$$

where $\mathbf{D}$ is diagonal with non-negative entries (the eigenvalues of $\mathbf{C}$). Set

$$\mathbf{S} := \mathbf{U}\mathbf{A}^{1/2}.$$

Then

$$\mathbf{S}^\top \mathbf{S} = \mathbf{A}^{1/2}\mathbf{U}^\top \mathbf{U}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{I}\mathbf{A}^{1/2} = \mathbf{A},$$

and

$$\mathbf{S}^\top \mathbf{D}\mathbf{S} = \mathbf{A}^{1/2}\mathbf{U}^\top \mathbf{D}\mathbf{U}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \mathbf{B}.$$

Since $\mathbf{A}^{1/2}$ and $\mathbf{U}$ are invertible, $\mathbf{S}$ is invertible, completing the proof. □

Lemma 4 (Gaussian conditional orthogonality lemma)**.** *Let $X, Y, Z \in \mathbb{R}^n$ be zero-mean jointly Gaussian random vectors with covariance matrices $\mathbf{\Sigma}_X = \mathbb{E}[XX^\top]$, $\mathbf{\Sigma}_Z = \mathbb{E}[ZZ^\top]$, and cross-covariance $\mathbf{\Sigma}_{Y,Z} = \mathbb{E}[YZ^\top]$. Define the conditional expectation*

$$\mathbb{E}[Y \mid Z] \coloneqq \Big(\mathbb{E}\big[ZZ^\top\big]^{-1}\mathbb{E}\big[ZY^\top\big]\Big)^\top Z = \mathbf{\Sigma}_{Y,Z}\mathbf{\Sigma}_Z^{-1}Z.$$

Then the following are equivalent:

$$X \perp \mathbb{E}[Y \mid Z] = 0 \quad \textit{a.s.} \qquad \Longleftrightarrow \qquad \mathbf{\Sigma}_X\mathbf{\Sigma}_{Y,Z} = \mathbf{0}.$$

Proof. Since X, Y, Z are jointly Gaussian, $\mathbb{E}[Y \mid Z] = \boldsymbol{M}Z$ with $\boldsymbol{M} \coloneqq \mathbf{\Sigma}_{Y,Z}\mathbf{\Sigma}_Z^{-1}$. The scalar random variable

$$S \coloneqq X^\top \mathbb{E}[Y \mid Z] = X^\top \boldsymbol{M}Z$$

is Gaussian with mean zero. Hence,

$$S = 0 \quad \text{a.s.} \qquad \Longleftrightarrow \qquad \mathrm{Var}(S) = 0.$$

Compute the variance:

$$\mathrm{Var}(S) = \mathbb{E}\big[S^2\big] = \mathbb{E}\big[(X^\top \boldsymbol{M}Z)^2\big] = \mathbb{E}\big[Z^\top \boldsymbol{M}^\top XX^\top \boldsymbol{M}Z\big].$$

Using independence and zero-mean assumptions,

$$\mathrm{Var}(S) = \mathrm{tr}\big(\boldsymbol{M}^\top \mathbf{\Sigma}_X \boldsymbol{M}\mathbf{\Sigma}_Z\big).$$

Since covariance matrices are positive semidefinite, $\mathrm{Var}(S) = 0$ iff

$$\mathbf{\Sigma}_X^{1/2}\boldsymbol{M}\mathbf{\Sigma}_Z^{1/2} = \mathbf{0} \implies \mathbf{\Sigma}_X\boldsymbol{M}\mathbf{\Sigma}_Z = \mathbf{0}.$$

Substituting $\boldsymbol{M} = \mathbf{\Sigma}_{Y,Z}\mathbf{\Sigma}_Z^{-1}$ gives

$$\mathbf{\Sigma}_X\mathbf{\Sigma}_{Y,Z} = \mathbf{0},$$

completing the proof. □

Lemma 5 (Characterizing the identified set in a linear, Gaussian case)**.** *In Example 1, Assumption 1,*

$$\mathcal{H}_{\mathrm{pi}} = \Big\{ \boldsymbol{h} \,\Big|\, (\boldsymbol{h} - \boldsymbol{h}_{\mathrm{erm}})^\top \boldsymbol{\Sigma}_X (\boldsymbol{h} - \boldsymbol{h}_{\mathrm{erm}}) \le r(\boldsymbol{\Gamma}) \Big\},$$

where the ellipsoid radius $r(\boldsymbol{\Gamma}) \ge 0$ depends on the choice of constraint parameters. Furthermore,

$$\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x}) = \Big[\boldsymbol{h}_{\mathrm{erm}}^\top \boldsymbol{x} - \sqrt{r(\boldsymbol{\Gamma})} \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}},\ \boldsymbol{h}_{\mathrm{erm}}^\top \boldsymbol{x} + \sqrt{r(\boldsymbol{\Gamma})} \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}} \Big].$$

Proof. Compute the population covariance

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \boldsymbol{f}^\top X + \xi) = \boldsymbol{\Sigma}_X \boldsymbol{f} + \boldsymbol{\Sigma}_{X,\xi},$$

so the (naïve) ERM estimand satisfies

$$\boldsymbol{h}_{\mathrm{erm}} = \boldsymbol{\Sigma}_X^{-1} \mathrm{Cov}(X, Y) = \boldsymbol{f} + \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{X,\xi}.$$

Let $b := \boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f} = \Sigma_{XX}^{-1} \Sigma_{X\xi}$. By the partial-$R^2$ constraint in Assumption 1

$$R^2_{\xi|X} = \frac{\boldsymbol{\Sigma}_{X,\xi}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,\xi}}{\sigma_\xi^2} \le \Gamma,$$

we have

$$\boldsymbol{\Sigma}_{X\xi}^\top \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{X\xi} \le \sigma_\xi^2 \Gamma.$$

Substituting $\boldsymbol{\Sigma}_{X\xi} = \boldsymbol{\Sigma}_{XX} b = \boldsymbol{\Sigma}_{XX} (\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f})$ yields

$$(\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f})^\top \boldsymbol{\Sigma}_{XX} (\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}) \le \sigma_\xi^2 \Gamma,$$

which is equivalent to

$$(\boldsymbol{f} - \boldsymbol{h}_{\mathrm{erm}})^\top \boldsymbol{\Sigma}_{XX} (\boldsymbol{f} - \boldsymbol{h}_{\mathrm{erm}}) \le \sigma_\xi^2 \Gamma \le \Gamma_0 \Gamma.$$

Thus the identified set for $\boldsymbol{f}$ is the stated ellipsoid with radius $r(\boldsymbol{\Gamma}) = \Gamma_0 \Gamma$. The centred Gaussian assumption guarantees the linear projection interpretation used above is exact.

Lastly, since the identified set is an ellipsoid, maximizing/minimizing a linear functional $\boldsymbol{f}^\top \boldsymbol{x}$ is just moving along its principal axis in the direction of $\boldsymbol{x}$, giving us the bounds for $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$.

□

USE OF LARGE LANGUAGE MODELS

A large language model (LLM) was utilized as a writing assistant to help refine the prose, improve clarity, and ensure a consistent narrative tone during the preparation of this manuscript. The human authors directed this process, take full responsibility for the final content, and are solely responsible for all scientific contributions of this work.