SYMMETRY-CONSTRAINED CAUSAL PARTIAL IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

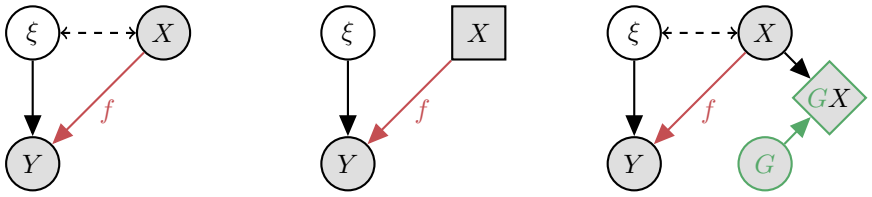
We present a novel framework for using knowledge of data symmetries to sharpen bounds in causal *partial identification (PI)*. The causal effect of the treatment X on outcome Y is generally not identifiable from observational data alone if their common causes, also known as confounders, are unobserved. 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. We show that knowledge of symmetries in data generation—formalized as invariance under transformation groups—provides additional constraints that tighten these bounds. We operationalize this insight through two approaches: (i) adding explicit invariance error constraints to existing PI methods, and (ii) applying symmetry-preserving *data augmentation (DA)* as a pre-processing step. Under a linear Gaussian model, we show that the later yields bounds that provably valid (containing the true causal effect), sharper (smaller identified sets), and more robust (lower worst-case error). The key mechanism being that randomized symmetry transformations introduce exogenous variation in X that cannot be attributed to confounding, thereby reducing ambiguity in the identified set. Experiments on synthetic and real data validate our approach. More broadly, our findings establish known data symmetries—ubiquitously employed in DA for variance reduction—can be repurposed as a principled tool for causal inference when point-identification is impossible.

1 INTRODUCTION

The problem of regression in machine learning aims to fit a model to observational (X, Y) data that predicts outcome Y from treatment X. Improving the generalization of such predictors to unlabeled samples of X often requires regularization techniques like *data augmentation (DA)* (Vapnik, 1998; Shorten & Khoshgoftaar, 2019; Lyle et al., 2020). However, such predictive models are generally not causal: the statistical relationship between X and Y may be driven by unobserved common causes, i.e. confounders, rather than the true influence of X on Y. The gold standard for eliminating confounding is direct intervention, i.e. explicit randomization of X during data generation (Peters et al., 2017; Pearl, 2009). Since these are often inaccessible, a common workaround is to correct for confounding via auxiliary variable (Zhang et al., 2023). However, these too may be insufficient to recover the causal effect (Kilbertus et al., 2020), or scarce in many applications (Akbar et al., 2025).

In such cases, identifying the true causal effect is not possible from observational data alone. *Partial identification (PI)* offers a principled alternative by computing bounds guaranteed to contain the true effect (Padh et al., 2023)—often sufficient for decision-making even without point-identification.. These bounds are obtained by solving an optimization problem whose constraints encode assumptions about how the data were generated. The informativeness of the bounds depends entirely on the strength and structure of these constraints.

This paper introduces known data symmetries—formalized as invariance under transformations of X—as a new source of constraints for PI. Such symmetries are ubiquitous in scientific and causal modeling, ranging from geometric stabilities in physical systems (Bronstein et al., 2021; Satorras et al., 2022) to semantic invariance in natural language (Veitch et al., 2021a) and permutation in-



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

Figure 1: Graphs of respective SEMs. (a) The original SEM from Eq. (1) with confounded (X, Y). (b) The original SEM post intervention on X. (c) The original SEM post DA application.

variance in exchangeable data (Veitch et al., 2021b). We demonstrate that enforcing this structural knowledge restricts the set of plausible causal hypotheses, effectively pruning the identified set. We operationalize this insight in two complementary ways:

1. **Explicit constraints via invariance error (Section 4.1)**: we propose bounding the invariance error under specified transformations as an explicit constraint for improving PI.
2. **Implicit constraints via DA (Section 4.2)**: we show the effectiveness of a simple symmetry-preserving DA based pre-processing step before running off-the-shelf PI solvers.

The first approach gives strict improvement in PI by construction, is amenable to modern Monte Carlo, gradient-based solvers, is not restricted to any specific hypothesis class and allows tunability to easily handle symmetry miss-specification. The later provides a cheap tool that can be composed with black-box PI solvers, offering guarantees under the linear Gaussian regime with well-specified symmetries. Our methods are simple to apply, compatible with existing solvers, and re-purpose the rather pervasive ML tool of data symmetries as a practical approach for strengthening causal conclusions when point-identification not possible.

2 Preliminaries

2.1 Statistical vs. causal inference

Consider random treatment X, outcome Y taking values in $\mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{Y} \subseteq \mathbb{R}$ respectively. The function $f \in \mathcal{H} \coloneqq \{h : \mathcal{X} \to \mathcal{Y}\}$ 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 $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, y_i)\}_{i=0}^n$ 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 $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$,

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

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 for causal effect identification

We can make X and the residual ξ uncorrelated via an *intervention* $\mathrm{do}(X \coloneqq X')$ 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 $\mathbb{P}_{X,Y}^{\mathrm{do}(X \coloneqq X')}$. 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)* $\mathbb{E}^{\mathrm{do}(X:=\boldsymbol{x})}[Y \mid X = \boldsymbol{x}]$ 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)

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

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

$$\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x}) = \Big[\min\nolimits_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} h^{\mathbb{Q}}_{\mathrm{adj}}(\boldsymbol{x}), \;\; \max\nolimits_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} h^{\mathbb{Q}}_{\mathrm{adj}}(\boldsymbol{x}) \Big]. \tag{4}$$

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 (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 $\mathcal{Q}_{\mathrm{pi}}$ 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 SYMMETRIES AND INVARIANCE

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,

$$R_{\mathrm{da+erm}}(h) \coloneqq \mathbb{E}[\ell(Y, h(GX))]. \tag{5}$$

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

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

Group $\mathcal{G}$ has a (unique) normalized Haar measure, $\mathbb{P}_G$ 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. (5) and subsequent theoretical results in that context. Nonetheless, increasing sample size via DA also bears on our work, a point we shall briefly discuss. Section 4.2 also makes the following assumption which characterizes many standard DA operations.

Assumption 1 (unbiased group action)**.** The group action $G \sim \mathbb{P}_G$ is identity-centered, meaning

$$\mathbb{E}[G\boldsymbol{x}] = \boldsymbol{x}, \qquad \forall\, \boldsymbol{x} \in \mathcal{X}.$$

Lemma 1 (added exogenous variation with DA)**.** *Under Assumption 1, G inflates the data variance,*

$$\boldsymbol{\Delta} \coloneqq \boldsymbol{\Sigma}_{GX} - \boldsymbol{\Sigma}_X \succcurlyeq \mathbf{0}, \qquad \textit{equality iff} \quad GX = X \quad \textit{a.s.}$$

Proof. See Appendix A.5 for the proof. □

3 CAUSAL IMPLICATIONS OF DATA SYMMETRIES

Crucially, the random group action G from Lemma 1 introduces additional *exogenous* variation in X that is independent of other system variables. Consequently, Akbar et al. (2025) showed that for a $\mathcal{G}$-invariant target function f, the transformation GX simulates a *soft* intervention on X—perturbing X to weaken the confounding association $X \leftrightarrow \xi$ while preserving the causal mechanism $X \to Y$. To formalize how this improves point estimation of f, Akbar et al. (2025) employs the following linear version of Eq. (1), which also serves as the basis for our subsequent PI analysis in Section 4.2.

Assumption 2 (a linear, Gaussian SEM)**.** The SEM Eq. (1) is centered, joint Gaussian with $\boldsymbol{f} \in \mathbb{R}^m$,

$$Y = \boldsymbol{f}^\top X + \xi.$$

In this setting, the causal estimation error (excess risk) under a squared loss takes the following form:

$$E^{\mathrm{do}(X)}(\boldsymbol{h}) = \left\|\boldsymbol{h} - \boldsymbol{f}\right\|^2_{\boldsymbol{\Sigma}_X}, \qquad E^{\mathrm{do}(\boldsymbol{x})}(\boldsymbol{h}) = \left(\boldsymbol{h}^\top \boldsymbol{x} - \boldsymbol{f}^\top \boldsymbol{x}\right)^2, \tag{6}$$

where we overload the notation $\mathrm{do}(\boldsymbol{x})$ (as opposed to $\mathrm{do}(X)$) to meant the *hard* intervention $\mathrm{do}(X \coloneqq \boldsymbol{x})$, i.e. fixing X to a constant $\boldsymbol{x}$ during data generation. Similar formulations have been used to measure causal error (Vankadara et al., 2022; Kania & Wit, 2023; Akbar et al., 2025) or quantify confounding strength (Janzing, 2019; Janzing & Schölkopf, 2018a;b). (Akbar et al., 2025) established the following result, which directly bears on our work:

Proposition 1 (estimation with DA (Akbar et al. (2025) lifted)). *For $\mathcal{G}$-inv. $\boldsymbol{f}$, Assumptions 1 and 2,*

$$0 \;\leq\; \frac{\kappa}{1+\kappa}\cdot \underbrace{\|\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\boldsymbol{h}_{\mathrm{erm}}-\boldsymbol{f})\|^2_{\boldsymbol{\Sigma}_X}}_{\text{estimation error within } \mathrm{range}(\boldsymbol{\Delta})} \;\leq\; E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{erm}}) - E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{da+erm}}),$$

$$\leq\; \|\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\boldsymbol{h}_{\mathrm{erm}}-\boldsymbol{f})\|^2_{\boldsymbol{\Sigma}_X}, \qquad \textit{eq. iff} \quad \overbrace{\boldsymbol{\Delta}\perp\boldsymbol{\Sigma}_{X,\xi}}^{\text{DA orthogonal to confounding}},$$

where $\kappa := \lambda^{+}_{\min}\big(\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Delta}\big) < \infty$ represents the lowest positive eigenvalue of the product $\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Delta}$.

Proof. See Appendix A.3, cf. (Akbar et al., 2025, Thm. 3) for the proof. □

Essentially, for $\mathcal{G}$-invariant $\boldsymbol{f}$, DA *dominates* ERM on causal estimation—performing strictly better iff it aligns with the confounding within X, but never worse. Note that in Proposition 1, for the "large DA" regime, which we define as $\kappa \to \infty$, the lower-bound approaches the upper-bound, which is simply the sq.-norm of the *projection* $\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\cdot)$ of estimation bias $(\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f})$ onto $\mathrm{range}(\boldsymbol{\Delta})$. This confirms that identification is generally not possible in this setting. Therefore the principled approach is to undertake PI of $\boldsymbol{f}$ instead of the point-estimation approach by Akbar et al. (2025).

This motivates our current work, where we leverage knowledge of symmetries in data generation to improve partial identification and/ or sensitivity analysis of f, as discussed in the following section.

4 Symmetry-Constrained Partial Identification

Our objective is to leverage symmetry knowledge to restrict the identified sets $\mathcal{H}_{\mathrm{pi}}$ and $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$. We give two strategies to operationalize this: (i) integrating an explicit invariance error constraint into the optimization, and (ii) inducing implicit regularization through data augmentation pre-processing.

4.1 Explicit constraint with invariance error

We start off by considering the most obvious approach to incorporate symmetry knowledge into PI—add an explicit invariance error constraint to any baseline PI method defined by a constraint set $\mathcal{C}_{\mathrm{pi}}$,

$$E_{\mathrm{inv}}(h) := \mathbb{E}\Big[\,\|h(X) - h(GX)\|^2\,\Big],$$

$$\mathcal{Q}_{\mathrm{inv+pi}}(\mathbb{P}_{X,Y}) := \Big\{\; \mathbb{Q} \in \mathcal{C}_{\mathrm{pi}} \;\Big|\; \mathbb{Q}_{X,Y} = \mathbb{P}_{X,Y},\;\; E_{\mathrm{inv}}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big) \leq \epsilon \;\Big\}.$$

Remark 1 (sharper, robust bounds with invariance error). By construction, subset inclusion holds:

$$\mathcal{H}_{\mathrm{inv+pi}} \subseteq \mathcal{H}_{\mathrm{pi}}, \qquad\qquad \mathcal{H}_{\mathrm{inv+pi}}(\boldsymbol{x}) \subseteq \mathcal{H}_{\mathrm{pi}}(\boldsymbol{x}).$$

Consequently, this guarantees that the volume of the identified set and the corresponding worst-case excess risk does not increase. Note that due to this same set inclusion, the approximation error generally cannot decrease, and may even potentially increase if $\mathcal{C}_{\mathrm{pi}}$ does not contain the true distribution $\mathbb{P}$. For $\epsilon = 0$, these metrics are equal to "large DA" regime results in Sections 3 and 4.2.

Nevertheless, whenever the baseline PI constraints $\mathcal{C}_{\mathrm{pi}}$ are valid and $E_{\mathrm{inv}}(f) \leq \epsilon$ holds, $\mathcal{Q}_{\mathrm{inv+pi}}$ guarantees validity. Furthermore, the parameter ϵ enables sensitivity analysis, allowing us to inspect how the bounds evolve as we vary the assumed invariance error in our choice of transformations G. Of course we can similarly use other formulations for E_{inv}, such as ones stated in Yang et al. (2019), or restrict ourselves to a hypothesis class that follows our symmetry by design (Cohen & Welling, 2016). However, the later may be restrictive in terms of compatibility with standard PI methods.

While our experiments discuss settings where Eq. (4) for $\mathcal{H}_{\mathrm{inv+pi}}(\boldsymbol{x})$ can be solved via convex programming, we emphasize that this explicit constraint formulation is fully compatible with modern deep learning-based PI. Since the functional E_{inv} is differentiable and amenable to Monte Carlo evaluation, it can be readily incorporated as a regularizer in augmented Lagrangian and/or gradient-based solvers (Padh et al., 2023; Kilbertus et al., 2020; Meresht et al., 2022; Hu et al., 2021).

Despite this compatibility, incorporating the invariance error constraint still requires modifying the solver logic or objective function. This imposes an implementation burden and precludes the use of specialized or "black-box" PI software where the internal optimization is fixed. This limitation motivates our second approach—a simple data pre-processing strategy that implicitly impose symmetry

constraints by simply feeding augmented data into standard off-the-shelf PI methods. Furthermore, when modeling complex, high-dimensional data, enforcing non-convex invariance constraints during optimization is often more expensive and notoriously unstable (Schweisthal et al., 2025; Padh et al., 2023) as opposed to a simple data pre-processing step.

4.2 Implicit constraint with DA pre-processing

We draw inspiration from IV methods, where "strong" instruments—those inducing significant exogenous variation in X—are known to yield sharper identification bounds compared to weak instruments (Kilbertus et al., 2020; Padh et al., 2023). This motivates our central inquiry in this section:

Can the synthetic exogenous variation introduced by DA similarly sharpen PI?

As with Akbar et al. (2025), which we extend now to the PI setting, the fundamental mechanism for PI sharpening is Lemma 1. Our main insight into why DA aids PI is summarized as follows:

(i) **Statistical Efficiency:** Most straightforwardly, DA grows effective data size, quelling sampling variation and finite-sample errors; key sources of uncertainty in PI (Imbens & Manski, 2004).

(ii) **Sharper Bounds:** DA adds variation in X that is explicitly exogenous, and therefore cannot be attributed to confounding. This reduces ambiguity in PI, which leads to sharper bounds.

(iii) **Robust Bounds:** By perturbing spurious features, DA reduces confounding bias, centering and contracting the PI bounds around the true solution. This directly minimizes the worst-case error.

(iv) **Valid Bounds:** Crucially, $\mathcal{G}$-invariance of f guarantees valid bounds with DA if $\mathcal{C}_{\text{pi}}$ is valid.

We elaborate these via analysis of the linear model from Assumption 2. But first we explicitly define the composition $\mathcal{Q}_{\text{da+pi}}$ of DA and PI, as well as the specific PI model that we use for our analysis,

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

Assumption 3 (a bounded confounding sensitivity model)**.** Consider the following constraint set.

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

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

Assumption 3 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 necessarily restrict ourselves to it—under the linear Gaussian setting of Assumption 2, several families of PI and sensitivity models reduce to ellipsoidal constraints equivalent to the form:

Lemma 2 (characterizing the identified set in a linear, Gaussian case)**.** *Under Assumptions 2 and 3,*

$$\mathcal{H}_{\text{pi}} = \left\{ \boldsymbol{h} \;\middle|\; \|\boldsymbol{h} - \boldsymbol{h}_{\text{erm}}\|^2_{\boldsymbol{\Sigma}_X} \leq r(\boldsymbol{\Gamma})^2 \right\},$$

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

$$\mathcal{H}_{\text{pi}}(\boldsymbol{x}) = \left[\; \boldsymbol{h}_{\text{erm}}^\top \boldsymbol{x} - r(\boldsymbol{\Gamma}) \cdot \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}}, \;\; \boldsymbol{h}_{\text{erm}}^\top \boldsymbol{x} + r(\boldsymbol{\Gamma}) \cdot \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}} \; \right].$$

Proof. See Appendix A.5 for the proof. □

Our results thus carry broader implications for PI and sensitivity analysis, as we discuss in Section 7.

4.2.1 Better bounds with data augmentation

First and foremost, we investigate if the post-DA bounds are, in some way, better than the baseline PI bounds. That is, if this exercise is useful at all. We present two results to support this claim.

Proposition 2 (sharper bounds with DA)**.** *For Assumptions 1 to 3, Lebesgue measure (volume) $|\cdot|$,*

$$\frac{|\mathcal{H}_{\text{da+pi}}|}{|\mathcal{H}_{\text{pi}}|} = \sqrt{\frac{\det \boldsymbol{\Sigma}_X}{\det \boldsymbol{\Sigma}_{GX}}} < 1, \qquad \frac{|\mathcal{H}_{\text{da+pi}}(\boldsymbol{x})|}{|\mathcal{H}_{\text{pi}}(\boldsymbol{x})|} = \frac{\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}}{\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}}} \leq 1, \qquad \textit{equality iff} \qquad \boldsymbol{x} \perp \boldsymbol{\Delta}.$$

Proof. See Appendix A.4 for the proof. □

Proposition 2 states that the hypothesis set $\mathcal{H}_{\text{da+pi}}$ is strictly smaller than the baseline $\mathcal{H}_{\text{pi}}$. The same holds true for the intervals $\mathcal{H}_{\text{da+pi}}(\boldsymbol{x})$ vs. $\mathcal{H}_{\text{pi}}(\boldsymbol{x})$, unless the query point $\boldsymbol{x}$ is orthogonal to the variation induced by DA [1], in which case the size of the interval remains the same. Importantly, Proposition 2 shows that this increase in "sharpness" is a direct consequence of the added variation from of DA in Lemma 1. And because this variation is exogenous and independent of ξ, our hypothesis/assumption about the strength of confounding $\boldsymbol{\Gamma}$ in the system should remain the same. This combination of increased data variation, but same confounding assumptions is what reduces ambiguity in PI, resulting in the sharper bounds of Proposition 2. Lastly, in the "large DA" regime, the ellipsoid $\mathcal{H}_{\text{da+pi}}$ collapses onto $\text{null}(\boldsymbol{\Delta})$, and the interval width $|\mathcal{H}_{\text{da+pi}}(\boldsymbol{x})|$ becomes $\left\|\boldsymbol{\Pi}^{\perp}_{\boldsymbol{\Delta}}\boldsymbol{x}\right\|^2_{\boldsymbol{\Sigma}_X^{-1}}$. Meaning, the DA removes *all but* the uncertainty that it cannot "see" within its $\text{range}(\boldsymbol{\Delta})$.

Although smaller identified sets/ intervals are in general desirable, size alone may not be the most appropriate measure of "goodness" of the identified set. The next result is based on worst-case error.

Theorem 1 (robust bounds with DA)**.** *For $\mathcal{G}$-inv. $\boldsymbol{f}$, Assumptions 1 to 3, $\kappa := \lambda_{\max}\big(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_{GX}^{-1}\big) \leq 1$,*

$$E^{\text{do}(X)}_{\text{worst}}(\mathcal{Q}_{\text{pi}}) \;=\; \Big(\|\boldsymbol{h}_{\text{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} + r(\boldsymbol{\Gamma})\Big)^2,$$
$$\overset{\text{(i), (ii)}}{\geq} \Big(\underbrace{\|\boldsymbol{h}_{\text{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X}}_{\text{lower estimation error}} + \underbrace{\sqrt{\kappa}\cdot r(\boldsymbol{\Gamma})}_{\text{sharper bounds}}\Big)^2 \overset{\text{(ii)}}{\geq} E^{\text{do}(X)}_{\text{worst}}(\mathcal{Q}_{\text{da+pi}}).$$

Equality iff (i) DA adds low variance $\kappa = 1$, and *(ii) DA orthogonal to confounding $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$. Also,*

$$\mathbb{E}_{\boldsymbol{x}}\Big[E^{\text{do}(\boldsymbol{x})}_{\text{worst}}(\mathcal{Q}_{\text{pi}})\Big] \;>\; \underbrace{\|\boldsymbol{h}_{\text{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}}_{\text{lower estimation error}} + \underbrace{\nu\cdot r(\boldsymbol{\Gamma})^2}_{\text{sharper bounds}} \;+\; s \;=\; \mathbb{E}_{\boldsymbol{x}}\Big[E^{\text{do}(\boldsymbol{x})}_{\text{worst}}(\mathcal{Q}_{\text{da+pi}})\Big],$$

where $\nu := \text{tr}\big(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_{GX}^{-1}\big) < \text{tr}\big(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_X^{-1}\big) = m$, queries $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ and some slack $s \geq 0$.

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

Theorem 1 shows that DA dominates PI on worst-case error through two mechanisms: (i) From Proposition 1, confounding aligned DA causes the PI centroid $\boldsymbol{h}_{\text{erm}}$ to drift closer to $\boldsymbol{f}$, bringing the bounds with it, thereby reducing worst-case error. (ii) Independently, from Proposition 2, the bounds themselves shrink, pushing the worst-case point closer still to $\boldsymbol{f}$. Given that the worst-case error bounds how bad the performance of any one hypothesis in the identified set may be, application of a DA pre-processing to PI therefore makes subsequent decision making more robust and reliable. Theorem 1 also gives a lower bound on this improvement via the factor κ, which in the "large DA" regime approaches 0 when $\boldsymbol{\Delta}$ has full span on $\mathbb{R}^m$, but is 1 otherwise. In our linear setting of Assumption 2, the former implies a trivial $\boldsymbol{f}$, and so the last inequality in Theorem 1 more clearly shows the independent, and strictly positive (average) effect ν of sharper bounds for individual queries $\boldsymbol{x} \sim \mathbb{P}_X$. Which in the "large DA" regime shrinks to $\nu \to \dim(\text{null}(\boldsymbol{\Delta})) =: k$, reducing by a factor $(m-k)/m < 1$, and directly improving (average) worst-case error for a random query $\boldsymbol{x}$.

4.2.2 Valid bounds with data augmentation

Finally, we discuss perhaps the most important property in PI—bound validity. We address this as:

Theorem 2 (valid bounds with DA)**.** *For any $\mathcal{G}$-invariant $\boldsymbol{f}$, it holds under Assumptions 1 to 3 that*

$$E^{\text{do}(X)}_{\text{approx}}(\mathcal{Q}_{\text{da+pi}}) \leq E^{\text{do}(X)}_{\text{approx}}(\mathcal{Q}_{\text{pi}}), \qquad \textit{equality iff} \qquad \mathbb{P} \in \mathcal{Q}_{\text{pi}}, \qquad \textit{or} \qquad \boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}.$$

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

Meaning the identified set $\mathcal{H}_{\text{da+pi}}$ is no farther from $\boldsymbol{f}$ compared to the original set $\mathcal{H}_{\text{pi}}$, and is strictly closer to $\boldsymbol{f}$ so long as the DA induced variation aligns with confounding. Of course it follows that when $\mathcal{Q}_{\text{pi}}$ contains the true joint distribution $\mathbb{P}$, then $\boldsymbol{f} \in \mathcal{H}_{\text{pi}}$ and so we should also have $\boldsymbol{f} \in \mathcal{H}_{\text{da+pi}}$. Instead of such a simple set inclusion criteria, we keep the more general approximation error framing of Theorem 2 because we also position DA as a tool for improved sensitivity analysis

[1] Intuitively, this would be like rotating an image $\boldsymbol{x}$ of a centered circle —the rotation leaves $\boldsymbol{x}$ unchanged.

where the constraint set may not necessarily be valid for some values of $\mathbf{\Gamma}$. Theorem 2 is then reassures that with $\mathcal{G}$-invariant $\boldsymbol{f}$, DA at the very least should not cause $\mathcal{H}_{\text{pi}}$ to drift away from $\boldsymbol{f}$.

Immediately following from Theorem 2, when $\mathbb{P} \in \mathcal{Q}_{\text{pi}}$, we also get $\boldsymbol{f}^\top \boldsymbol{x} \in \mathcal{H}_{\text{da+pi}}(\boldsymbol{x}), \mathcal{H}_{\text{pi}}(\boldsymbol{x})$. It is difficult, however, to show a similar result as Theorem 1 for the point-wise evaluation of $E^{\text{do}(\boldsymbol{x})}_{\text{approx}}(\mathcal{Q}_{\text{da+pi}})$ vs. $E^{\text{do}(\boldsymbol{x})}_{\text{approx}}(\mathcal{Q}_{\text{pi}})$ when $\mathbb{P} \notin \mathcal{Q}_{\text{pi}}$, as the approximation error non-trivially depends on the alignment of unknown confounding $\mathbf{\Sigma}_{X,\xi}$ with the query $\boldsymbol{x}$, and both can be arbitrary.

5 Related Work

PI and sensitivity analysis. We give an account of related PI and sensitivity analysis literature in Section 2.3. Our work is largely orthogonal but complementary to this: we introduce a new source of constraints—symmetry knowledge—that is compatible and composes with existing PI frameworks.

Symmetry and invariance in causal inference. Invariance is fundamental to causality: causal mechanisms yield predictions invariant to interventions (Peters et al., 2016). Methods enforce such invariances using auxiliary variables for identification (Peters et al., 2016; Heinze-Deml et al., 2018; Arjovsky et al., 2019; Dance & Bloem-Reddy, 2024; Kilbertus et al., 2020; Singh et al., 2019; Zhang et al., 2023) or robust prediction (Rothenhäusler et al., 2021; Krueger et al., 2021; Christiansen et al., 2022). Akbar et al. (2025) leverage symmetry knowledge for robust prediction; whereas we address the orthogonal, but more principled problem of PI when point identification is infeasible.

Counterfactual DA. The literature on counterfactual DA has been the main focus of causal analysis of DA. These methods achieve robust predictors by synthesizing counterfactual examples (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), but require restrictive assumptions: full SEMs (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 complete causal graphs (Pitis et al., 2022; Armengol Urpí et al., 2024). We, like Akbar et al. (2025), require a more accessible symmetry knowledge about the data.

6 Experiments

We validate our frameworks in finite samples. We fix the augmented sample size to match the original to show that bound sharpening stems from symmetry constraints rather than variance reduction.

6.1 Simulation experiment

We follow Akbar et al. (2025) to instantiate a simulation for the linear Gaussian SEM from Assumption 2. To do this, we first sample the SEM parameters—standard normal matrix $\boldsymbol{T} \in \mathbb{R}^{m\times m}$, and vectors $\boldsymbol{f}, \boldsymbol{e} \in \mathbb{R}^m$, keeping them fixed throughout the experiment. Then sample standard normal exogenous variables (U, N_X, N_Y) and pass them through the following model to get observable (X, Y) confounded by the unobserved U as:

$$X \coloneqq \boldsymbol{T}^\top U + 0.1 \cdot N_X, \quad Y \coloneqq \boldsymbol{f}^\top X + \boldsymbol{e}^\top U + 0.1 \cdot N_Y.$$

Next, we construct a DA G such that $\boldsymbol{f}$ respects $\mathcal{G}$-invariance. As with Akbar et al. (2025), we do this by first taking the SVD of $\boldsymbol{f}$,

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

The matrix $\boldsymbol{V}_0$ represents the orthonormal basis of $\text{null}(\boldsymbol{f})$. We take k of these rows to construct $\boldsymbol{A} \in \mathbb{R}^{k\times m}$ which defines G:

$$GX \coloneqq X + a \cdot \boldsymbol{A}^\top G, \qquad G \sim \mathcal{N}(\mathbf{0}_k, \boldsymbol{I}_k).$$

Therefore, by construction we have $\mathcal{G}$-invariance and therefore $\boldsymbol{f}^\top X = \boldsymbol{f}^\top GX$. Figure 2 provides an intuitive visualization

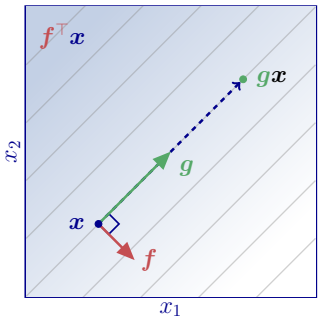


Figure 2: A cartoon visualization of the linear simulation setup. The augmentation generates $\boldsymbol{gx}$ by randomly translating $\boldsymbol{x}$ along the level-sets (contours) defined by the causal parameter $\boldsymbol{f}$, using additive noise sampled from the null-space of $\boldsymbol{f}$.

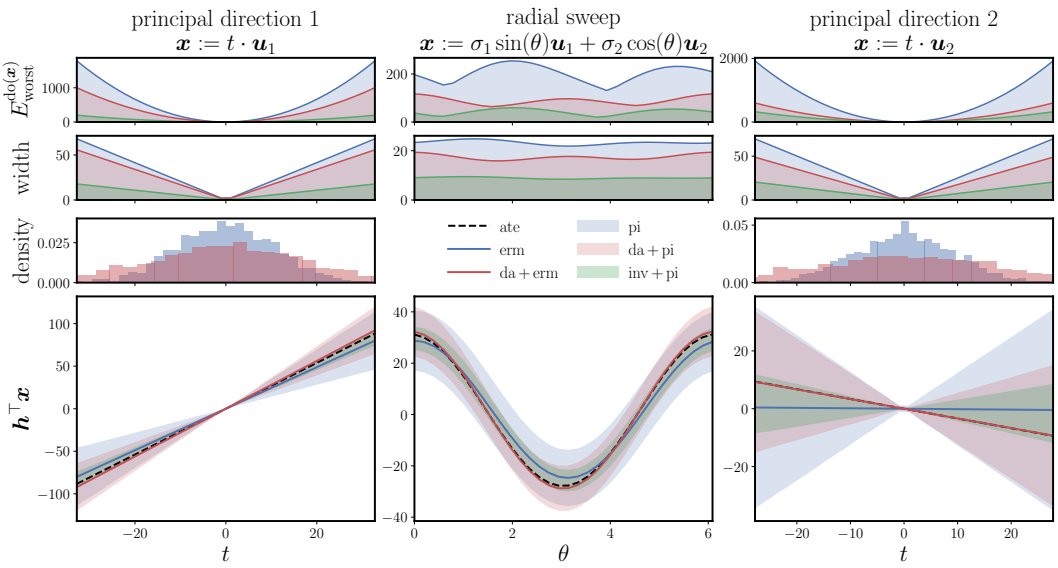


Figure 3: Data augmentation consistently sharpens partial identification bounds in a linear simulation. Across query points aligned with principal components (PC1, PC2) and a radial sweep, DA+PI (red) yields narrower intervals, lower worst-case excess risk (E_{worst}), and predictions closer to the true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

of such a transformation. While this construction technically utilizes the ground truth $\boldsymbol{f}$, we treat access to $\boldsymbol{A}$ as representing prior domain knowledge about the functional symmetries, noting that this information alone is insufficient to identify $\boldsymbol{f}$ due to the unobserved confounding $\boldsymbol{e}^\top U$.

Taking $m = 32$, $k = 31$, we generate $n = 2048$ samples for (X, Y, G) with the DA std parameter a specified as 4. For ERM we use a closed form linear OLS solution. And for PI we use the partial R-squared sensitivity model from Assumption 3 for a range of query points $\boldsymbol{x}_0, \boldsymbol{x}_1$ with the sensitivity parameter set as $\Gamma = \Gamma_0 = 2^9$, and $\epsilon = 2^{-3}$ for the invariance error constraint.

To visualize the results, we chose $\boldsymbol{x} := t \cdot \boldsymbol{u}_1$ and $\boldsymbol{x} := t \cdot \boldsymbol{u}_2$ where $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ are the first and second principal components of the data. We then sweep t over ± 3 standard-deviations, computing intervals $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x}), \mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})$ via convex programming (separately for the upper and lower bounds). The results are shown in Fig. 3 (left, right). Fig. 3 (center) also shows a radial sweep over $\theta \in [0, 2\pi]$ to generate queries $\boldsymbol{x} := \sigma_0 \cdot \sin(\theta) \cdot \boldsymbol{u}_0 + \sigma_1 \cdot \cos(\theta) \cdot \boldsymbol{u}_1$.

6.2 Optical Device Experiment

We utilize the benchmark dataset provided by Janzing & Schölkopf (2018b), consisting of 3×3 pixel images X displayed on a laptop screen which generate voltage readings Y across a photodiode. The system involves a physically instantiated hidden confounder U that controls the intensity of two LEDs; the first affects the webcam capturing X, while the second influences the photodiode measuring Y. We derive the ground-truth causal predictor $\boldsymbol{f}$ by regressing Y on the joint features $(\phi(X), U)$, where $\phi(X)$ denotes polynomial features of X. We select the polynomial degree $d \in \{1, \cdots, 5\}$ that best explains the data (degree 2 in our case) and subsequently remove the learned component corresponding to U 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 $\Gamma = \Gamma_0 = 10^2$ for the partial R-squared model from Assumption 3, and $\epsilon = 2^{-3}$ for the invariance error constraint 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.

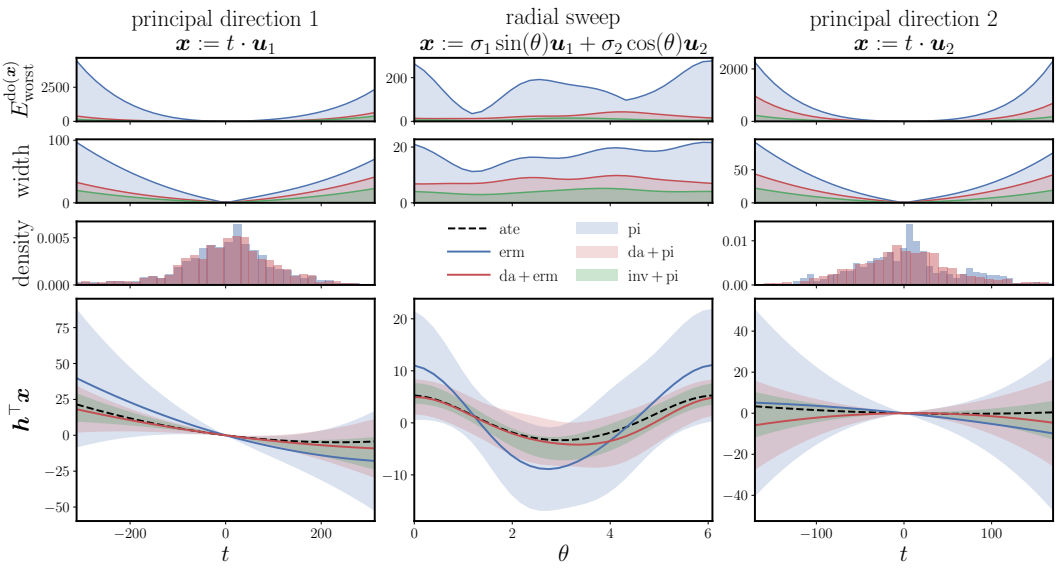


Figure 4: Our method sharpens causal bounds on the real-world Optical Device dataset. Even with complex, non-linear relationships, applying outcome-invariant DA (red) substantially narrows the partial identification bounds compared to the baseline (blue).

7 Limitations, Assumptions and Future Work

Symmetry knowledge. Our approach hinges on the untestable assumption that the target f is $\mathcal{G}$-invariant for the chosen symmetry transformation. While this does require prior knowledge, our framework also allows to handle symmetry miss-specification in the explicit invariance error constraint from Section 4.1 via the ϵ parameter. Additionally, we remind the readers that untestable assumptions are fundamental for making any causal conclusions from observational data with unobserved confounding (Pearl, 2009), as is the norm in partial identification. This also includes access to auxiliary variables since the conditional independences that they represent are also merely untestable assumptions. Furthermore, Akbar et al. (2025) argues that a symmetry-based knowledge assumption is actually quite practical given its 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 PI 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 our symmetry transformation 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 PI models can be reduced to the constraints in Assumption 3. These include, of course, the partial R-squared model, Rosenbaum (2002), 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 results here are more general than just the partial R-squared model.

8 Conclusion

We show that causal symmetries sharpen partial identification bounds by restricting the hypothesis space. We operationalize this via explicit invariance constraints and implicit data augmentation. Through construction and linear analysis, respectively, we prove these methods yield valid, strictly tighter, and more robust bounds. Empirically validated and broadly compatible, our framework establishes symmetry as a powerful resource within the tool-belt for causal partial identification.

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.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. arXiv:1907.02893.

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.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

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.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2022. doi: 10.1109/TPAMI.2021.3094760.

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.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR.

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.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

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.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REX). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

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, Peter Bühlmann, and Nicolai Meinshausen. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

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.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2022. URL https://arxiv.org/abs/2102.09844.

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.

Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208. Curran Associates, Inc., 2021a.

Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208. Curran Associates, Inc., 2021b.

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.

Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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—ROBUST BOUNDS WITH DA

Theorem 1 (robust bounds with DA). *For $\mathcal{G}$-inv. $\boldsymbol{f}$, Assumptions 1 to 3, $\kappa \coloneqq \lambda_{\max}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}\big) \leq 1$,*

$$E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}) = \left(\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} + r(\boldsymbol{\Gamma})\right)^2,$$
$$\overset{(i),\,(ii)}{\geq} \left(\underbrace{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X}}_{\text{lower estimation error}} + \underbrace{\sqrt{\kappa}\cdot r(\boldsymbol{\Gamma})}_{\text{sharper bounds}}\right)^2 \overset{(ii)}{\geq} E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{da+pi}}).$$

Equality iff (i) DA adds low variance $\kappa = 1$, and *(ii) DA orthogonal to confounding $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$. Also,*

$$\mathbb{E}_{\boldsymbol{x}}\Big[E^{\mathrm{do}(\boldsymbol{x})}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}})\Big] > \underbrace{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}}_{\text{lower estimation error}} + \underbrace{\nu \cdot r(\boldsymbol{\Gamma})^2}_{\text{sharper bounds}} + s = \mathbb{E}_{\boldsymbol{x}}\Big[E^{\mathrm{do}(\boldsymbol{x})}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{da+pi}})\Big],$$

where $\nu \coloneqq \mathrm{tr}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}\big) < \mathrm{tr}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_X^{-1}\big) = m$, queries $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ and some slack $s \geq 0$.

Proof. We show the two inequalities below in the respective sections.

Population effect. Lemma 2, characterizes the identified sets $\mathcal{H}_{\mathrm{pi}}$, $\mathcal{H}_{\mathrm{da+pi}}$ as ellipsoids:

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

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

$$\begin{aligned}
E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}) &= \max\nolimits_{\mathbb{Q}\in\mathcal{Q}_{\mathrm{pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \\
&= \max\nolimits_{\boldsymbol{h}\in\mathcal{H}_{\mathrm{pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{pi}}.) \\
&= \max\nolimits_{\boldsymbol{h}\in\mathcal{H}_{\mathrm{pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&= \Big(\big\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\big\|_{\boldsymbol{\Sigma}_X} + r(\boldsymbol{\Gamma})\Big)^2, && \text{(Lemma 3)}
\end{aligned}$$

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

$$\begin{aligned}
&E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{da+pi}}) \\
&= \max\nolimits_{\mathbb{Q}\in\mathcal{Q}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}\Big(h^{\mathbb{Q}}_{\mathrm{adj}}\Big), \\
&= \max\nolimits_{\boldsymbol{h}\in\mathcal{H}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{da+pi}}.) \\
&= \max\nolimits_{\boldsymbol{h}\in\mathcal{H}_{\mathrm{da+pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&\overset{(\star)}{\leq} \boxed{\left(\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_X} + r(\boldsymbol{\Gamma}) \cdot \sqrt{\lambda_{\max}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}\big)}\right)^2}, \\
& && \text{(Lemma 3, = iff } (\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}) \parallel \boldsymbol{v}_{\max}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}\big).) \\
&\overset{(\dagger)}{\leq} \Big(\big\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\big\|_{\boldsymbol{\Sigma}_X} + r(\boldsymbol{\Gamma})\Big)^2, && \text{(Lemma 1, } \lambda_{\max}\big(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}\big) \leq 1.) \\
&\overset{(\ddagger)}{\leq} \Big(\big\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\big\|_{\boldsymbol{\Sigma}_X} + r(\boldsymbol{\Gamma})\Big)^2 = E^{\mathrm{do}(X)}_{\mathrm{worst}}(\mathcal{Q}_{\mathrm{pi}}). && \text{(Proposition 1, = iff } \boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}.)
\end{aligned}$$

Condition for equality. The bound involves three inequalities: ($\star$) the geometric bound from Lemma 3, ($\dagger$) the inflated variance implication from Lemma 1, and ($\ddagger$) the estimation bound from Proposition 1. Of these, ($\dagger$) is immediate, so we investigate ($\star$), ($\ddagger$) in isolation with $\lambda_{\max} = 1$. Now, assuming further that the condition $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$ is satisfied for ($\ddagger$), it follows from Proposition 1 that

$$\boldsymbol{h}_{\mathrm{da+erm}} = \boldsymbol{h}_{\mathrm{erm}} \qquad \Longleftrightarrow \qquad \boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}.$$

So we go ahead and substitute $\boldsymbol{h}_{\mathrm{da+erm}}$ with $\boldsymbol{h}_{\mathrm{erm}}$ in ($\star$). From Lemma 3, equality holds iff the bias vector $(\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f})$, now $(\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f})$, is a dominant eigenvector of $\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}$. Because $(\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}) = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,\xi}$ (OLS closed-form), and $\boldsymbol{\Sigma}_{GX} = \boldsymbol{\Sigma}_X + \boldsymbol{\Delta}$ (Lemma 1), we follow the steps of Propositions 1 and 2 to do a change of basis by jointly diagonalizing $\boldsymbol{\Sigma}_X, \boldsymbol{\Delta}$ (Lemma 7) to show

$$\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1} (\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}) = \boldsymbol{\Sigma}_X (\boldsymbol{\Sigma}_X + \boldsymbol{\Delta})^{-1} \big(\boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,\xi}\big) = (\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}), \quad \iff \quad \boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}.$$

The bias $(\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f})$ is an eigenvector of $\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_{GX}^{-1}$ with eigenvalue $\lambda_{\max} = 1$. Therefore, equality holds for both ($\star$) and ($\ddagger$) iff $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$ and the residual improvement is solely from radius contraction $\lambda_{\max} < 1$. Conditions for ($\star$), ($\dagger$), ($\ddagger$) form conditions (i), (ii) in the statement.

(Average) individual effect. Define $J(\mathcal{Q}) \coloneqq \mathbb{E}_{\boldsymbol{x}}\Big[E_{\mathrm{worst}}^{\mathrm{do}(\boldsymbol{x})}(\mathcal{Q}) \Big]$ for queries $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$. From Lemma 2, the worst-case risk at a query point $\boldsymbol{x}$ is (bias + radius) squared:

$$E_{\mathrm{worst}}^{\mathrm{do}(\boldsymbol{x})}(\mathcal{Q}_{\mathrm{pi}}) = \Big(\big|(\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f})^\top \boldsymbol{x}\big| + r(\boldsymbol{\Gamma}) \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}} \Big)^2.$$

Expanding the square, we decompose the total expected risk for DA into three terms:

$$J(\mathcal{Q}_{\mathrm{da+pi}}) = \underbrace{\mathbb{E}_{\boldsymbol{x}}\Big[\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{x}}^2 \Big]}_{\text{(a) estimation error}} + \underbrace{r(\boldsymbol{\Gamma})^2 \mathbb{E}_{\boldsymbol{x}}\Big[\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}^2 \Big]}_{\text{(b) average radius}},$$
$$+ \underbrace{2 r(\boldsymbol{\Gamma}) \mathbb{E}_{\boldsymbol{x}}\Big[\big|(\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f})^\top \boldsymbol{x}\big| \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}} \Big]}_{\text{(c) interaction of (a), (b)}}.$$

We analyze the reduction $J(\mathcal{Q}_{\mathrm{pi}}) - J(\mathcal{Q}_{\mathrm{da+pi}})$ term by term:

a) **Estimation error:** DA+ERM dominates ERM from Proposition 1, equality iff $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$

b) **Average radius:** DA+PI *strictly* dominates PI from Proposition 2, as $\boldsymbol{x} \not\perp \boldsymbol{\Delta}$ almost surely. Also, expand $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}^2$ into a trace term, and then use its cyclic permutation invariance gets

$$\mathbb{E}_{\boldsymbol{x}}\Big[\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}^2 \Big] = \mathrm{tr}\big(\boldsymbol{\Sigma}_{GX}^{-1} \boldsymbol{\Sigma}_X\big) < \mathrm{tr}\big(\boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_X\big) = m.$$

c) **Interaction term:** *Strictly* lower for any non-trivial DA $\boldsymbol{\Delta} \neq \boldsymbol{0}$ from Lemma 5.

Concluding that:

$$J(\mathcal{Q}_{\mathrm{da+pi}}) < J(\mathcal{Q}_{\mathrm{pi}}) \qquad \text{when} \qquad \boldsymbol{\Delta} \neq \boldsymbol{0}.$$

□

A.2 Proof of Theorem 2—Valid bounds with DA

Theorem 2 (valid bounds with DA). *For any $\mathcal{G}$-invariant $\boldsymbol{f}$, it holds under Assumptions 1 to 3 that*

$$E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{da+pi}}) \leq E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}}), \qquad \textit{equality iff} \qquad \mathbb{P} \in \mathcal{Q}_{\mathrm{pi}}, \qquad \textit{or} \qquad \boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}.$$

Proof. From Lemma 2, we can characterize the identified sets $\mathcal{H}_{\mathrm{pi}}$, $\mathcal{H}_{\mathrm{da+pi}}$ as ellipsoids of the form

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

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

$$\begin{aligned}
E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}}) &= \min_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{pi}}} E^{\mathrm{do}(X)}\left(h^{\mathbb{Q}}_{\mathrm{adj}}\right), \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{pi}}.) \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&= \left(\left\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\right\|_{\boldsymbol{\Sigma}_X} - r(\boldsymbol{\Gamma}) \right)^2, && \text{(Lemma 4)}
\end{aligned}$$

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

$$\begin{aligned}
&E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{da+pi}}) \\
&= \min_{\mathbb{Q} \in \mathcal{Q}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}\left(h^{\mathbb{Q}}_{\mathrm{adj}}\right), \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{da+pi}}} E^{\mathrm{do}(X)}(\boldsymbol{h}), && \text{(Re-parameterizing in terms of } \mathcal{H}_{\mathrm{da+pi}}.) \\
&= \min_{\boldsymbol{h} \in \mathcal{H}_{\mathrm{da+pi}}} \|\boldsymbol{h} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
&\overset{(\heartsuit)}{\leq} \left(1 - \frac{r(\boldsymbol{\Gamma})}{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|_{\boldsymbol{\Sigma}_{GX}}}\right)^2 \|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}, \\
& && \text{(Lemma 4, = iff } (\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}) \parallel \boldsymbol{v}\left(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}^{-1}_{GX}\right).) \\
&= \left(\left\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\right\|_{\boldsymbol{\Sigma}_{GX}} - r(\boldsymbol{\Gamma}) \right)^2 \frac{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_X}}{\|\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}\|^2_{\boldsymbol{\Sigma}_{GX}}}, \\
&\overset{(\spadesuit)}{\leq} \left(\left\|\boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f}\right\|_{\boldsymbol{\Sigma}_X} - r(\boldsymbol{\Gamma}) \right)^2 = E^{\mathrm{do}(X)}_{\mathrm{approx}}(\mathcal{Q}_{\mathrm{pi}}), && \text{(Similar to Proposition 1, = iff } \boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}.)
\end{aligned}$$

where the last inequality (‡) follows from a similar approach as used in Proposition 1 to show that

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

which holds with equality if and only if $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$. The case for $\mathbb{P} \in \mathcal{Q}_{\mathrm{pi}}$ is trivial from Lemma 4.

Condition for equality. The two types of inequalities that comprise the given approximation error bound are Proposition 1-type estimation bias related (♠), and the ellipsoidal geometry inequality (♡) from Lemma 4. We can proceed similar to the corresponding section in Theorem 1 to show that:

$$\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi} \qquad \Longleftrightarrow \qquad (\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f}) \parallel \boldsymbol{v}_{\max}\left(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}^{-1}_{GX}\right).$$

Since condition of (♡) requires alignment of $(\boldsymbol{h}_{\mathrm{da+erm}} - \boldsymbol{f})$ with *any* eigenvector $\boldsymbol{v}\left(\boldsymbol{\Sigma}_X \boldsymbol{\Sigma}^{-1}_{GX}\right)$, therefore $\boldsymbol{v}_{\max}$ suffices. Consequently, equality holds for both (♡) and (♠) iff $\boldsymbol{\Delta} \perp \boldsymbol{\Sigma}_{X,\xi}$. □

A.3 Proof of Proposition 1—Estimation with DA (Akbar et al. (2025) lifted)

Proposition 1 (estimation with DA (Akbar et al. (2025) lifted))**.** *For $\mathcal{G}$-inv. $\boldsymbol{f}$, Assumptions 1 and 2,*

$$0 \;\le\; \frac{\kappa}{1+\kappa}\cdot \underbrace{\|\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\boldsymbol{h}_{\mathrm{erm}}-\boldsymbol{f})\|^2_{\boldsymbol{\Sigma}_X}}_{\text{estimation error within } \mathrm{range}(\boldsymbol{\Delta})} \;\le\; E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{erm}}) - E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{da+erm}}),$$
$$\le\; \|\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\boldsymbol{h}_{\mathrm{erm}}-\boldsymbol{f})\|^2_{\boldsymbol{\Sigma}_X}, \qquad \textit{eq. iff} \quad \overbrace{\boldsymbol{\Delta}\perp\boldsymbol{\Sigma}_{X,\xi}}^{\substack{\textit{DA orthogonal}\\ \textit{to confounding}}},$$

where $\kappa := \lambda^+_{\min}\big(\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Delta}\big) < \infty$ represents the lowest positive eigenvalue of the product $\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Delta}$.

Proof. We start by first investigating the post-DA confounding vector $\mathbb{E}\big[(GX)\xi^\top\big] = \boldsymbol{\Sigma}_{GX,\xi}$ as

$$\begin{aligned}
\boldsymbol{\Sigma}_{GX,\xi} &= \mathbb{E}\big[(GX)\xi^\top\big] \\
&= \mathbb{E}_{X,\xi}\big[\mathbb{E}_G[GX \mid X,\xi]\xi^\top\big] && \text{(Law of total expectation.)}\\
&= \mathbb{E}_{X,\xi}\big[\mathbb{E}_G[GX \mid X]\xi^\top\big] && (G \text{ exogenous} \implies G \perp\!\!\!\perp \xi \mid X.)\\
&= \mathbb{E}_{X,\xi}\big[X\xi^\top\big] = \boldsymbol{\Sigma}_{X,\xi}. && (\text{As } \mathbb{E}[GX\,|\,X] = X \text{ from Assumption 1.})
\end{aligned}$$

Now define $\boldsymbol{c} := \boldsymbol{\Sigma}_{X,\xi} = \boldsymbol{\Sigma}_{GX,\xi}$ for brevity. The estimation error in Eq. (6) for the baseline ERM and DA+ERM is governed by the projection of confounding $\boldsymbol{c}$ onto the respective data manifolds as:

$$E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{erm}}) = \big\|\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c}\big\|^2_{\boldsymbol{\Sigma}_X} = \boldsymbol{c}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c},$$
$$E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{da+erm}}) = \big\|\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{c}\big\|^2_{\boldsymbol{\Sigma}_X} = \boldsymbol{c}^\top\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{c}$$

Using $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_{GX} - \boldsymbol{\Delta}$ and the Resolvent Identity $\boldsymbol{\Sigma}_X^{-1} - \boldsymbol{\Sigma}_{GX}^{-1} = \boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_X^{-1}$, we get:

$$\begin{aligned}
E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{da+erm}}) &= \boldsymbol{c}^\top\boldsymbol{\Sigma}_{GX}^{-1}(\boldsymbol{\Sigma}_{GX}-\boldsymbol{\Delta})\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{c}\\
&= \boldsymbol{c}^\top\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{c} - \boldsymbol{c}^\top\big(\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_{GX}^{-1}\big)\boldsymbol{c}\\
&= \big(\boldsymbol{c}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c} - \boldsymbol{c}^\top\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c}\big) - \boldsymbol{c}^\top\big(\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_{GX}^{-1}\big)\boldsymbol{c}\\
&= E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{erm}}) - \boldsymbol{c}^\top\big(\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_X^{-1}\big)\boldsymbol{c} - \boldsymbol{c}^\top\big(\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_{GX}^{-1}\big)\boldsymbol{c},\\
&= E^{\mathrm{do}(X)}(\boldsymbol{h}_{\mathrm{erm}}) \\
&\qquad - \overbrace{\boldsymbol{c}^\top\boldsymbol{\Sigma}_X^{-1}\big(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\big)\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c}}^{0 \le \text{first-order reduction}}\\
&\qquad\quad - \underbrace{\boldsymbol{c}^\top\boldsymbol{\Sigma}_X^{-1}\big(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}_{GX}^{-1}\boldsymbol{\Sigma}_X\big)\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c}}_{0 \le \text{second-order reduction}}.
\end{aligned}$$

Both reduction terms are quadratic forms of the PSD matrix $\boldsymbol{\Delta}$ and are therefore non-negative.

Define δ as their sum. Lemma 6 lower-bounds the first-order term, and by extension lower-bounds δ:

$$0 \;\le\; \frac{\kappa}{1+\kappa}\cdot\|\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\boldsymbol{h}_{\mathrm{erm}}-\boldsymbol{f})\|^2_{\boldsymbol{\Sigma}_X} \;\le\; \text{first order term} \overset{(\blacktriangle)}{\le} \delta.$$

Trace the same steps as Lemma 6 to bound δ from above via the simultaneous basis from Lemma 7 ($\boldsymbol{\Sigma}_X = \boldsymbol{S}^\top\boldsymbol{S}$, $\boldsymbol{\Delta} = \boldsymbol{S}^\top\boldsymbol{D}\boldsymbol{S}$). Taking $\boldsymbol{z} := \boldsymbol{S}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{c}$ and eigenvalues D_{ii} of $\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Delta}$, we can show

$$\delta \;=\; \sum_i z_i^2\cdot\Bigg(\overbrace{\underbrace{\frac{D_{ii}}{1+D_{ii}}}_{\text{1st order}} + \underbrace{\frac{D_{ii}}{(1+D_{ii})^2}}_{\text{2nd order}}}^{<1}\Bigg) \overset{(\triangledown)}{\le} \sum_{i:D_{ii}>0} 1\cdot z_i^2 \;=\; \|\boldsymbol{\Pi}_{\boldsymbol{\Delta}}(\boldsymbol{h}_{\mathrm{erm}}-\boldsymbol{f})\|^2_{\boldsymbol{\Sigma}_X}.$$

Condition for equality. Equality holds for ($\blacktriangle$) iff $\boldsymbol{\Delta}\perp\boldsymbol{\Sigma}_{X,\xi}$, as otherwise the second-order term is strictly positive. Equality also holds for ($\triangledown$) iff $\boldsymbol{\Delta}\perp\boldsymbol{\Sigma}_{X,\xi}$, because that entails $z_i = 0$ whenever $D_{ii} > 0$ so that the sums on both sides go to 0. □

A.4 Proof of Proposition 2—Sharper bounds with DA

Proposition 2 (sharper bounds with DA). *For Assumptions 1 to 3, Lebesgue measure (volume)* $|\cdot|$,

$$\frac{|\mathcal{H}_{\mathrm{da+pi}}|}{|\mathcal{H}_{\mathrm{pi}}|} = \sqrt{\frac{\det \boldsymbol{\Sigma}_X}{\det \boldsymbol{\Sigma}_{GX}}} < 1, \qquad \frac{|\mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})|}{|\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})|} = \frac{\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_{GX}^{-1}}}{\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}}} \leq 1, \qquad \textit{equality iff} \qquad \boldsymbol{x} \perp \boldsymbol{\Delta}.$$

Proof. We compare the geometric properties of the identified sets as characterized by Lemma 2.

Ellipsoid volume (global contraction). Given that the volume of a $\boldsymbol{\Sigma}$-ellipsoid $\propto (\det \boldsymbol{\Sigma})^{-1/2}$, it immediately follows from Lemmas 1 and 2, and the monotonicity of determinant for SPD matrices:

$$\begin{aligned} \boldsymbol{\Sigma}_X \preccurlyeq \boldsymbol{\Sigma}_{GX} &\implies \det(\boldsymbol{\Sigma}_X) < \det(\boldsymbol{\Sigma}_{GX}), \\ &\implies \det(\boldsymbol{\Sigma}_{GX})^{-1/2} < \det(\boldsymbol{\Sigma}_X)^{-1/2} \implies |\mathcal{H}_{\mathrm{da+pi}}| < |\mathcal{H}_{\mathrm{pi}}|. \end{aligned}$$

Interval width (point-wise contraction). From Lemma 2, the width of the interval $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ is simply $2r(\boldsymbol{\Gamma}) \cdot \|\boldsymbol{x}\|_{\boldsymbol{\Sigma}_X^{-1}}$. It then immediately follows from Lemma 1 and definition of the PSD order:

$$\begin{aligned} \boldsymbol{\Sigma}_X \preccurlyeq \boldsymbol{\Sigma}_{GX} &\implies \boldsymbol{\Sigma}_{GX}^{-1} \preccurlyeq \boldsymbol{\Sigma}_X^{-1}, \\ &\implies \boldsymbol{x}^\top \boldsymbol{\Sigma}_{GX}^{-1} \boldsymbol{x} \leq \boldsymbol{x}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{x} \implies |\mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})| \leq |\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})|. \end{aligned}$$

Condition for equality. The interval width is strictly smaller for $\mathcal{H}_{\mathrm{da+pi}}(\boldsymbol{x})$ compared to $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ unless the query point $\boldsymbol{x}$ lies in the null space of the difference $\boldsymbol{\Delta} \coloneqq \boldsymbol{\Sigma}_{GX} - \boldsymbol{\Sigma}_X$. From Lemma 7,

$$\boldsymbol{\Sigma}_{GX}^{-1} = (\boldsymbol{\Sigma}_X + \boldsymbol{\Delta})^{-1} = \left(\boldsymbol{S}^\top \boldsymbol{S} + \boldsymbol{S}^\top \boldsymbol{D} \boldsymbol{S}\right)^{-1} = \boldsymbol{S}^{-1} (\boldsymbol{I} + \boldsymbol{D})^{-1} \boldsymbol{S}^{-\top}.$$

When we analyze the ratio of squared norms using the basis $\boldsymbol{z} \coloneqq \boldsymbol{S}^{-\top} \boldsymbol{x}$, it simplifies to:

$$\frac{\|\boldsymbol{x}\|^2_{\boldsymbol{\Sigma}_{GX}^{-1}}}{\|\boldsymbol{x}\|^2_{\boldsymbol{\Sigma}_X^{-1}}} = \frac{\boldsymbol{z}^\top (\boldsymbol{I} + \boldsymbol{D})^{-1} \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}} = \frac{\sum_i z_i^2 (1 + D_{ii})^{-1}}{\sum_i z_i^2}.$$

Since $\boldsymbol{D}$ is non-negative, the term $(1 + D_{ii})^{-1} < 1$ whenever $D_{ii} > 0$. Therefore, the ratio is strictly less than 1 unless $\boldsymbol{z}$ is supported only on indices where $D_{ii} = 0$. This requires $\boldsymbol{z}^\top \boldsymbol{D} \boldsymbol{z} = 0$, which transforms back to the condition that $\boldsymbol{x}$ must lie in the null-space of $\boldsymbol{\Delta}$ (i.e., $\boldsymbol{x} \perp \boldsymbol{\Delta}$). □

A.5 MISCELLANEOUS SUPPORTING LEMMAS

Lemma 1 (added exogenous variation with DA)**.** *Under Assumption 1, G inflates the data variance,*

$$\boldsymbol{\Delta} \coloneqq \boldsymbol{\Sigma}_{GX} - \boldsymbol{\Sigma}_X \succcurlyeq \mathbf{0}, \qquad \textit{equality iff} \quad GX = X \quad \textit{a.s.}$$

Proof. Represent $Z \coloneqq GX$. Now, by applying the Law of Total Covariance conditioning on X,

$$\boldsymbol{\Sigma}_{GX} = \mathbb{E}[\mathrm{Cov}(GX \,|\, X)] + \mathrm{Cov}(\mathbb{E}[GX \,|\, X]). \tag{7}$$

By Assumption 1 (unbiased group action) we have $\mathbb{E}[GX \,|\, X] = X$, and the second term reduces to

$$\mathrm{Cov}(\mathbb{E}[GX \,|\, X]) = \mathrm{Cov}(X) = \boldsymbol{\Sigma}_X. \tag{8}$$

The first term represents the exogenous variation injected by the group action. Let $\boldsymbol{\Delta} = \mathbb{E}[\mathrm{Cov}(GX \,|\, X)]$. Since covariance matrices are PSD by definition, we have $\boldsymbol{\Delta} \succcurlyeq \mathbf{0}$.

Condition for equality. The inequality holds with equality ($\boldsymbol{\Sigma}_{GX} = \boldsymbol{\Sigma}_X$) iff the injected noise matrix $\boldsymbol{\Delta} = \mathbf{0}$. Since $\mathrm{Cov}(GX \,|\, X) \succcurlyeq \mathbf{0}$ almost surely, its expectation is zero if and only if $\mathrm{Cov}(GX \,|\, X) = \mathbf{0}$ almost surely. This implies GX is a deterministic function of X. Given the unbiased assumption $\mathbb{E}[GX \,|\, X] = X$, this forces $GX = X$ almost surely (i.e., G acts as identity over support of X). Therefore, for any non-trivial augmentation, the inequality $\boldsymbol{\Delta} \succcurlyeq \mathbf{0}$ is strict. □

Lemma 2 (characterizing the identified set in a linear, Gaussian case)**.** *Under Assumptions 2 and 3,*

$$\mathcal{H}_{\mathrm{pi}} = \left\{ \boldsymbol{h} \;\middle|\; \|\boldsymbol{h} - \boldsymbol{h}_{\mathrm{erm}}\|_{\boldsymbol{\Sigma}_X}^2 \le r(\boldsymbol{\Gamma})^2 \right\},$$

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

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

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 $\boldsymbol{b} := \boldsymbol{h}_{\mathrm{erm}} - \boldsymbol{f} = \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{X\xi}$. By the partial-$R^2$ constraint in Assumption 3

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

we have

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

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

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

Thus the identified set for $\boldsymbol{f}$ is the stated ellipsoid with radius $r(\boldsymbol{\Gamma})^2 = \Gamma_0 \Gamma$. 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 $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$. □

Lemma 3 (upper bound on distance of a point to farthest point on ellipsoid)**.** *Take ellipsoid* $\mathcal{O} \subset \mathbb{R}^n$

$$\mathcal{O} = \left\{ \boldsymbol{x} \,\middle|\, (\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{\Sigma}_0 (\boldsymbol{x} - \boldsymbol{x}_0) \leq r_0^2 \right\},$$

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*

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

This distance is upper bounded as follows, with $\boldsymbol{v}_{\max}$ *as the eigenvector corresponding to* $\lambda_{\max}$*.*

$$D_{\boldsymbol{\Sigma}}^{\max}(\boldsymbol{y}, \mathcal{O}) \leq \|\boldsymbol{y} - \boldsymbol{x}_0\|_{\boldsymbol{\Sigma}} + r_0 \cdot \sqrt{\lambda_{\max}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right)},$$
$$\text{equality iff} \qquad \boldsymbol{y} - \boldsymbol{x}_0 \parallel \boldsymbol{v}_{\max}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right).$$

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 $r_0 \cdot \sqrt{\cdot \lambda_{\max}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right)}$. The result follows.

Condition for equality. The triangle inequality holds with equality iff $(\boldsymbol{y} - \boldsymbol{x}_0)$ and $(\boldsymbol{x} - \boldsymbol{x}_0)$ are collinear. The second term is maximized when $(\boldsymbol{x} - \boldsymbol{x}_0)$ aligns with the dominant eigenvector $\boldsymbol{v}_{\max}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right)$ (the generalized principal axis). Therefore, the total bound is tight iff $(\boldsymbol{y} - \boldsymbol{x}_0)$ is itself an eigenvector corresponding to $\lambda_{\max}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1})$, i.e. $(\boldsymbol{y} - \boldsymbol{x}_0) \parallel \boldsymbol{v}_{\max}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right)$. □

Lemma 4 (upper bound on distance of a point to an ellipsoid)**.** *Take the following ellipsoid* $\mathcal{O} \subset \mathbb{R}^n$

$$\mathcal{O} = \left\{ \boldsymbol{x} \,\middle|\, (\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{\Sigma}_0 (\boldsymbol{x} - \boldsymbol{x}_0) \le r_0^2 \right\},$$

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*

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

This distance is upper bounded by the following closed-form, with $\boldsymbol{v}(\cdot)$ *as* any *arbitrary eigenvector.*

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

$$\textit{equality iff} \quad \boldsymbol{y} \in \mathcal{O}, \quad \textit{or} \quad \boldsymbol{y} - \boldsymbol{x}_0 \parallel \boldsymbol{v}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right).$$

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

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

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

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

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

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

Condition for equality. The condition for $\boldsymbol{y} \in \mathcal{O}$ case is trivial. For $\boldsymbol{y} \notin \mathcal{O}$, the minimum distance from $\boldsymbol{y}$ to the ellipsoid occurs at the boundary intersection of the ray $\boldsymbol{x}(r)$ iff the gradient of the *objective* $\boldsymbol{\Sigma}(\boldsymbol{y} - \boldsymbol{x})$ is parallel to the gradient of the *constraint* $\boldsymbol{\Sigma}_0(\boldsymbol{x} - \boldsymbol{x}_0)$ at the intersection point. Since $(\boldsymbol{x} - \boldsymbol{x}_0)$ is proportional to $(\boldsymbol{y} - \boldsymbol{x}_0)$ along the ray, this optimality condition requires:

$$\boldsymbol{\Sigma}(\boldsymbol{y} - \boldsymbol{x}_0) \;\propto\; \boldsymbol{\Sigma}_0(\boldsymbol{y} - \boldsymbol{x}_0) \qquad \Longleftrightarrow \qquad (\boldsymbol{y} - \boldsymbol{x}_0) \;\propto\; \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0(\boldsymbol{y} - \boldsymbol{x}_0).$$

Thus, the ray bound is exact if and only if $(\boldsymbol{y} - \boldsymbol{x}_0)$ is an (*any*) eigenvector of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0$. □

Lemma 5 (centroid-radius interaction bound via coupling). *For $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$, consider two constant vectors $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^m$ (representing **centroid displacements**) and two symmetric positive definite matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \succ \mathbf{0}$ (representing respective **radius metrics**). Define the interaction integral:*

$$J(\boldsymbol{b}, \boldsymbol{\Sigma}) := \mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{b}^\top \boldsymbol{x}| \cdot \sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}} \right].$$

If $(\boldsymbol{b}_2, \boldsymbol{\Sigma}_2)$ has a strictly shorter "whitened" centroid and a strictly narrower radius (PSD-wise), i.e.,

1. ***Centroid Contraction:*** $\|\boldsymbol{b}_2\|_{\boldsymbol{\Sigma}_0} < \|\boldsymbol{b}_1\|_{\boldsymbol{\Sigma}_0}$,
2. ***Radius Contraction:*** $\boldsymbol{\Sigma}_2 \prec \boldsymbol{\Sigma}_1$,

then the interaction term strictly decreases:

$$J(\boldsymbol{b}_2, \boldsymbol{\Sigma}_2) < J(\boldsymbol{b}_1, \boldsymbol{\Sigma}_1).$$

Proof. To evaluate the integral, we transform it into spherically symmetric coordinates (whitening).

Whitening. We can express the data vector $\boldsymbol{x}$ as a linear transformation of a standard normal vector $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ such that $\boldsymbol{x} = \boldsymbol{\Sigma}_0^{1/2} \boldsymbol{z}$. Substituting this into the centroid and radius terms:

$$\text{Centroid:} \quad |\boldsymbol{b}^\top \boldsymbol{x}| = |\boldsymbol{b}^\top \boldsymbol{\Sigma}_0^{1/2} \boldsymbol{z}| = |(\boldsymbol{\Sigma}_0^{1/2} \boldsymbol{b})^\top \boldsymbol{z}| = |\tilde{\boldsymbol{b}}^\top \boldsymbol{z}|,$$
$$\text{Radius:} \quad \sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}} = \sqrt{\boldsymbol{z}^\top \boldsymbol{\Sigma}_0^{1/2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_0^{1/2} \boldsymbol{z}} = \sqrt{\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{z}},$$

where $\tilde{\boldsymbol{b}} := \boldsymbol{\Sigma}_0^{1/2} \boldsymbol{b}$ is the whitened centroid, and $\tilde{\boldsymbol{\Sigma}} := \boldsymbol{\Sigma}_0^{1/2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_0^{1/2}$ is the whitened radius metric.

Rotational symmetry (coupling). The expectation is now over the standard normal variable $\boldsymbol{z}$:

$$J = \mathbb{E}_{\boldsymbol{z}}\left[|\tilde{\boldsymbol{b}}^\top \boldsymbol{z}| \cdot \sqrt{\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{z}} \right].$$

Since the distribution of $\boldsymbol{z}$ is spherically symmetric (invariant to rotations), the distribution of the dot product $\tilde{\boldsymbol{b}}^\top \boldsymbol{z}$ depends only on the length of $\tilde{\boldsymbol{b}}$. We can conceptually rotate the coordinate system for each scenario such that $\tilde{\boldsymbol{b}}$ aligns with the first basis vector $\boldsymbol{e}_1$. In this rotated frame, $|\tilde{\boldsymbol{b}}^\top \boldsymbol{z}| = \|\tilde{\boldsymbol{b}}\| \cdot |z_1|$. Crucially, note that $\|\tilde{\boldsymbol{b}}\| = \|\boldsymbol{\Sigma}_0^{1/2} \boldsymbol{b}\|_2 = \sqrt{\boldsymbol{b}^\top \boldsymbol{\Sigma}_0 \boldsymbol{b}} = \|\boldsymbol{b}\|_{\boldsymbol{\Sigma}_0}$. Thus:

$$J(\boldsymbol{b}, \boldsymbol{\Sigma}) = \|\boldsymbol{b}\|_{\boldsymbol{\Sigma}_0} \cdot \mathbb{E}_{\boldsymbol{z}}\left[|z_1| \cdot \sqrt{\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{z}} \right].$$

Comparison. We now compare $J_1 = J(\boldsymbol{b}_1, \boldsymbol{\Sigma}_1)$ and $J_2 = J(\boldsymbol{b}_2, \boldsymbol{\Sigma}_2)$.

$$\begin{aligned}
J_2 &= \|\boldsymbol{b}_2\|_{\boldsymbol{\Sigma}_0} \cdot \mathbb{E}_{\boldsymbol{z}}\left[|z_1| \cdot \sqrt{\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}}_2 \boldsymbol{z}} \right] \\
&< \|\boldsymbol{b}_1\|_{\boldsymbol{\Sigma}_0} \cdot \mathbb{E}_{\boldsymbol{z}}\left[|z_1| \cdot \sqrt{\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}}_2 \boldsymbol{z}} \right] && \text{(by centroid contraction)} \\
&< \|\boldsymbol{b}_1\|_{\boldsymbol{\Sigma}_0} \cdot \mathbb{E}_{\boldsymbol{z}}\left[|z_1| \cdot \sqrt{\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}}_1 \boldsymbol{z}} \right] && \text{(by radius contraction)} \\
&= J_1.
\end{aligned}$$

The second inequality holds because $\boldsymbol{\Sigma}_2 \prec \boldsymbol{\Sigma}_1$ implies $\tilde{\boldsymbol{\Sigma}}_2 \prec \tilde{\boldsymbol{\Sigma}}_1$, so $\boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}}_2 \boldsymbol{z} < \boldsymbol{z}^\top \tilde{\boldsymbol{\Sigma}}_1 \boldsymbol{z}$ for all $\boldsymbol{z} \neq \mathbf{0}$. Since $|z_1|$ is non-negative and not always zero, the expectation strictly decreases. □

Lemma 6 (sandwich bounds for SPD-PSD weighted norms). *For $n \times n$ matrices $\boldsymbol{A} \succ \mathbf{0}$, $\boldsymbol{B} \succcurlyeq \mathbf{0}$, denote the pseudo-inverse as $\boldsymbol{B}^\dagger$, and $\boldsymbol{\Pi}_{\boldsymbol{B}} \coloneqq \boldsymbol{B}^\dagger \boldsymbol{B}$ projects onto* $\mathrm{range}(\boldsymbol{B})$. *Then, for any $\boldsymbol{x} \in \mathbb{R}^n$,*

$$\underbrace{\frac{\kappa}{1+\kappa}}_{\text{shrinkage factor } \leq 1} \cdot \|\boldsymbol{\Pi}_{\boldsymbol{B}} \boldsymbol{x}\|_{\boldsymbol{A}}^2 \quad \leq \quad \boldsymbol{x}^\top \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\,\boldsymbol{x} \quad \leq \quad \|\boldsymbol{\Pi}_{\boldsymbol{B}} \boldsymbol{x}\|_{\boldsymbol{A}}^2,$$

for bounded minimum positive eigenvalue $\kappa \coloneqq \lambda_{\min}^{+}\big(\boldsymbol{A}^{-1}\boldsymbol{B}\big) < \infty$. Equality holds for upper bound iff $\boldsymbol{x} \perp \boldsymbol{B}$, and lower bound iff $\boldsymbol{\Pi}_{\boldsymbol{B}}\boldsymbol{x}$ is entirely in eigen-space of $\boldsymbol{A}^{-1}\boldsymbol{B}$ corresponding to κ.

Proof. From Lemma 7, we have $\boldsymbol{A} = \boldsymbol{S}^\top \boldsymbol{S}$ and $\boldsymbol{B} = \boldsymbol{S}^\top \boldsymbol{D}\boldsymbol{S}$ for invertible $\boldsymbol{S}$ and diagonal $\boldsymbol{D} \succcurlyeq \mathbf{0}$. Note that $\boldsymbol{D}$ are eigenvalues of $\boldsymbol{A}^{-1}\boldsymbol{B}$ by cyclic permutation invariance (i.e., $\lambda(\boldsymbol{A}\boldsymbol{B}) = \lambda(\boldsymbol{B}\boldsymbol{A})$).

Define the change of basis $\boldsymbol{z} \coloneqq \boldsymbol{S}\boldsymbol{x}$. Then $\boldsymbol{x} = \boldsymbol{S}^{-1}\boldsymbol{z}$, and

$$\begin{aligned}
\boldsymbol{x}^\top \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\,\boldsymbol{x} &= \boldsymbol{x}^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{S} + \boldsymbol{S}^\top \boldsymbol{D}\boldsymbol{S})^{-1}\boldsymbol{S}^\top \boldsymbol{D}\boldsymbol{S}\,\boldsymbol{x} \\
&= \boldsymbol{x}^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{S}^\top(\boldsymbol{I}+\boldsymbol{D})\boldsymbol{S})^{-1}\boldsymbol{S}^\top \boldsymbol{D}\boldsymbol{S}\boldsymbol{x} \\
&= \boldsymbol{x}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{S}^{-1}(\boldsymbol{I}+\boldsymbol{D})^{-1}\boldsymbol{S}^{-\top}\boldsymbol{S}^\top \boldsymbol{D}\boldsymbol{S}\boldsymbol{x} \\
&= \boldsymbol{x}^\top \boldsymbol{S}^\top (\boldsymbol{I}+\boldsymbol{D})^{-1}\boldsymbol{D}\boldsymbol{S}\boldsymbol{x} \\
&= \boldsymbol{z}^\top (\boldsymbol{I}+\boldsymbol{D})^{-1}\boldsymbol{D}\,\boldsymbol{z} \\
&= \sum_i \frac{D_{ii}}{1+D_{ii}} z_i^2.
\end{aligned}$$

Similarly, for the projected norm, noting that $\boldsymbol{\Pi}_{\boldsymbol{B}} = \boldsymbol{S}^{-1}\boldsymbol{D}^\dagger \boldsymbol{D}\boldsymbol{S}$ and $\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \|\boldsymbol{S}\boldsymbol{x}\|_2^2$:

$$\|\boldsymbol{\Pi}_{\boldsymbol{B}}\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \big\|\boldsymbol{S}\big(\boldsymbol{S}^{-1}\boldsymbol{D}^\dagger \boldsymbol{D}\boldsymbol{S}\big)\boldsymbol{x}\big\|_2^2 = \big\|\boldsymbol{D}^\dagger \boldsymbol{D}\boldsymbol{z}\big\|_2^2 = \sum_{i:D_{ii}>0} z_i^2.$$

Upper bound. Since $\frac{D_{ii}}{1+D_{ii}} < 1$ for all $D_{ii} > 0$, the following inequality is strict for any $z_i \neq 0$.

$$\boldsymbol{x}^\top \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\,\boldsymbol{x} = \sum_i \frac{D_{ii}}{1+D_{ii}} z_i^2 \qquad \leq \qquad \sum_i z_i^2 = \|\boldsymbol{\Pi}_{\boldsymbol{B}}\boldsymbol{x}\|_{\boldsymbol{A}}^2.$$

And equality holds iff $z_i = 0$ for *all* active indices, which implies $\boldsymbol{\Pi}_{\boldsymbol{B}}\boldsymbol{x} = \mathbf{0}$ (i.e., $\boldsymbol{x} \perp \boldsymbol{B}$).

Lower Bound. The function $f(d) = \frac{d}{1+d}$ is monotonically increasing for $d \geq 0$. Restricting our attention to the support of the vector (indices where $D_{ii} > 0$), we define $\kappa = \min\{D_{ii} : D_{ii} > 0\}$. It follows that for every active index, $\frac{D_{ii}}{1+D_{ii}} \geq \frac{\kappa}{1+\kappa}$. Summing over the support:

$$\sum_{i:D_{ii}>0} \frac{D_{ii}}{1+D_{ii}} z_i^2 \quad \geq \quad \sum_{i:D_{ii}>0} \frac{\kappa}{1+\kappa} z_i^2 \quad = \quad \frac{\kappa}{1+\kappa}\|\boldsymbol{\Pi}_{\boldsymbol{B}}\boldsymbol{x}\|_{\boldsymbol{A}}^2.$$

For the inequality to become an equality, we require $\frac{D_{ii}}{1+D_{ii}} = \frac{\kappa}{1+\kappa}$ for every index i where $z_i \neq 0$. This implies $D_{ii} = \kappa$ for all contributing dimensions. Geometrically, this means the vector $\boldsymbol{x}$ (after projection) must align only with the directions associated with the minimum eigenvalue κ. □

Lemma 7 (SPD, PSD joint denationalization via congruence Akbar et al. (2025))**.** *For any $n \times n$ matrices $\boldsymbol{A} \succ \mathbf{0}$, $\boldsymbol{B} \succcurlyeq \mathbf{0}$, there exists an $n \times n$ invertible $\boldsymbol{S}$ and non-negative diagonal $\boldsymbol{D}$ such that*

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

Proof. See (Akbar et al., 2025, Lem. 2), cf. (Horn & Johnson, 1985, Thm. 7.6.4, p. 465). □

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.