SYMMETRY-INFORMED CAUSAL PARTIAL IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Partial identification (PI) entails estimating bounds on causal effects by encoding different assumptions on data generation as a constrained optimization problem. Such bounds can suffice to inform policy decisions even if the causal effect itself is not identifiable. Often vacuous in practice, practitioners seek to exhaustively encode domain knowledge as additional constraints to make the PI bounds more informative. We introduce known *data symmetries*—invariance of the causal effect under certain data transformations—as a new source of constraints to inform PI. We operationalize this as a *shape constraint* on the causal function, and via a *change of measure* against which PI is posed using simple data pre-processing. Both approaches are shown to sharpen bounds under two canonical PI models. This is shown both theoretically for the population case, and via experiments in the finite-sample case. More broadly, our framework establishes data symmetries as a natural, underutilized source of background knowledge for robust causal inference.

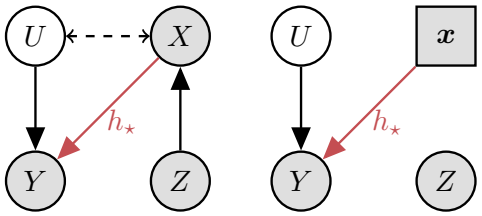
1 INTRODUCTION

Interventions are the gold standard for answering causal queries (Pearl, 2009; Peters et al., 2017). Absent interventions, observed associations may be due to hidden confounders rather than causal effects, so researchers resort to adjusting via auxiliary variables, which rest on strong, untestable assumptions that may simply fail (Zhang et al., 2023). Partial identification (PI) offers a principled relaxation to this problem. Rather than insisting on assumptions sufficient for point identification, PI returns *bounds* on the causal effect under a milder set of assumptions (Padh et al., 2023). These bounds are constructed by considering all possible data-generating processes (DGPs) compatible with both the observed data and the chosen assumptions. Consequently, they are only as informative as the strength of said assumptions, and can often be near-vacuous. To make PI actionable for decision-making, practitioners must exhaustively incorporate credible domain knowledge into the PI solver to restrict the set of compatible DGPs, thereby sharpening the bounds.

While existing PI families encode domain knowledge by bounding the strength of hidden confounding (Cinelli & Hazlett, 2020; Rosenbaum, 2002), positing a causal graph (Meresht et al., 2022), or imposing shape restrictions on the class of causal functions (Chetverikov et al., 2018), a natural and accessible source of prior knowledge remains underutilized: *known data symmetries*. Symmetries describe properties of a system that remain invariant under certain data transformations; for example, image content may not change under rotation/reflection/translation (Bronstein et al., 2021). Readily used in statistical regularization (Chen et al., 2020; Lyle et al., 2020), we propose re-purposing known symmetries as a natural shape restriction on the set of candidate causal functions to sharpen PI bounds.

Contributions. We give two operationalizations: an invariance shape constraint on the causal functions added to white-box PI solvers (§3), and a change-of-measure approach where off-the-shelf or black-box solvers are simply fed pre-processed data via symmetry transformations (§4). We license the latter by showing it realizes a soft intervention: preserving the causal mechanism and graph up to relabeling, while diluting confounding. Thus, validity is preserved on aforementioned PI families. We then develop observable conditions for bound sharpening, and a simple pad-and-intersect protocol to guard against misspecified symmetry under two canonical PI models: bounded confounding (Cinelli & Hazlett, 2020; Rosenbaum, 2002) and leaky instruments (Deaner, 2022; Cinelli & Hazlett, 2025). Finally, we validate both approaches on simulated and real data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Hidden confounding model (b) Intervention $\mathrm{do}(\boldsymbol{x})$

Figure 1: *(a)* Graph of Eq. (1) with IV Z and hidden noise U that confounds (X, Y). *(b)* An intervention $\mathrm{do}(X = \boldsymbol{x})$ removes this confounding.

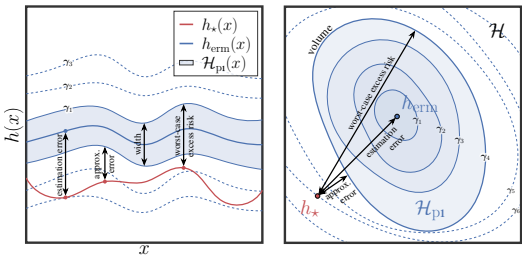


Figure 2: PI sets $\mathcal{H}_{\mathrm{pi}}(\boldsymbol{x})$ *(Left)* and $\mathcal{H}_{\mathrm{pi}}$ *(Right)* for various sensitivity parameters $\boldsymbol{\gamma}$ and different metrics of interest; estimation errors $E_{\boldsymbol{x}}, E$, approx. errors $E^-_{\boldsymbol{x}}, E^-$, worst-case errors $E^+_{\boldsymbol{x}}, E^+$.

2 Preliminaries

2.1 Statistical vs. causal inference

Given n samples $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, y_i, \boldsymbol{z}_i)\}_{i\in[n]}$ from the observational distribution p_{XYZ} of treatment $X \in \mathcal{X} \subseteq \mathbb{R}^{d_{\boldsymbol{x}}}$, outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and pre-treatment co-variates $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_{\boldsymbol{z}}}$, we aim to find the function $h_\star \in \mathcal{H} \coloneqq \{h : \mathcal{X} \to \mathbb{R}\}$ defining the causal effect of X on Y under the following DGP:

$$Y = h_\star(X) + U + \xi, \qquad \mathbb{E}[U] = \mathbb{E}[\xi] = 0, \qquad \xi \perp\!\!\!\perp (U, X, Z). \tag{1}$$

The unobserved noise (ξ, U) is assumed mean-centered w.l.o.g., with the intercept absorbed into $h_\star$. Independence of noise $U \perp\!\!\!\perp X$ implies $\mathbb{E}[Y \,|\, \boldsymbol{x}] = h_\star(\boldsymbol{x})$. *Statistical inference* entails identifying precisely the Bayes optimal predictor $\mathbb{E}[Y \,|\, \boldsymbol{x}]$ from $\mathcal{D}$ via *empirical risk minimization* (ERM): for a proper, convex loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, minimize over $h \in \mathcal{H}$ an empirical version of the *statistical risk*,

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

Then, for a sufficiently rich hypothesis class, the minimizer h_{erm} provides an unbiased estimate of $h_\star$.

However, independence of X and the unobserved noise U may not hold in general, in which case $\mathbb{E}[U \,|\, \boldsymbol{x}] \neq 0$ and the conditional $\mathbb{E}[Y \,|\, \boldsymbol{x}]$ now gives a biased estimate of $h_\star(\boldsymbol{x})$ (Pearl, 2009; Peters et al., 2017). This dependence arises due to unobserved common causes U of X and Y, known as *confounders*. We say that X and Y are confounded and refer to the resulting bias as *confounding bias* (Pearl, 2009). The observed pre-treatment co-variates Z may contain auxiliary information, e.g. proxies for such confounders, that can help to mitigate this bias. This encompasses *causal inference* which seeks to identify $h_\star$ under hidden confounding, or at the very least account for it by finding bounds on $h_\star$ should point-identification not be possible. Both approaches are outlined below.

2.2 Point identification via intervention and adjustment

We can make treatment X and the confounder U independent via an *intervention* $\mathrm{do}(X = \boldsymbol{x})$ that sets X to a fixed value $\boldsymbol{x}$ in Eq. (1) during data generation. Under such a treatment intervention, henceforth denoted as $\mathrm{do}(\boldsymbol{x})$ for brevity, outcome Y follows the *interventional distribution* $p_{Y|\mathrm{do}(\boldsymbol{x})}$. The loss from Eq. (2) evaluated under $p_{Y|\mathrm{do}(\boldsymbol{x})}$ now defines the *causal risk* R (Kania & Wit, 2023),

$$R_{\boldsymbol{x}}(y) \coloneqq \mathbb{E}[\ell(Y, y) \,|\, \mathrm{do}(\boldsymbol{x})], \qquad R(h) \coloneqq \mathbb{E}_{X\sim p_X}[R_X(h(X))]. \tag{3}$$

The target estimand of Eq. (3) is the *average treatment effect (ATE)* $\mathbb{E}[Y \,|\, \mathrm{do}(\boldsymbol{x})]$ which equals $h_\star(\boldsymbol{x})$ for the data generation in Eq. (1) (Xu & Gretton, 2022). Minimizers of Eq. (3) hence yield unbiased estimates of $h_\star(\boldsymbol{x})$. To capture the *estimation error* for a candidate hypothesis $h \in \mathcal{H}$, we use *causal excess risk* (Akbar et al., 2025; Chennuru Vankadara et al., 2022)

$$E(h) \coloneqq R(h) - R(h_\star),$$

which removes the irreducible noise from Eq. (3) and measures how far h is from $h_\star$, with $E(h_\star) = 0$.

As interventions are often inaccessible for computing the risk in Eq. (3), estimating $h_\star$ usually relies on access to the full joint distribution p of (U, X, Y, Z) via *back-door adjustment* $h^p_\star$ (Pearl, 2009),

$$h^p_\star(\boldsymbol{x}) \coloneqq \mathbb{E}[\mathbb{E}[Y \,|\, \boldsymbol{x}, Z, U]], \qquad (U, X, Y, Z) \sim p.$$

Alternatively, $h_\star$ is recovered via *instrumental variables (IVs)* Z satisfying (i) **un-confoundedness:** $Z \perp\!\!\!\perp U$, (ii) **exclusion:** $Y \perp\!\!\!\perp Z \mid X, U$ (Newey & Powell, 2003). Conditioning Eq. (1) on Z:

$$\mathbb{E}[Y - h_\star(X) \,|\, \boldsymbol{z}] = \mathbb{E}[U \,|\, \boldsymbol{z}] = 0. \tag{4}$$

An IV estimate h_{IV} solves Eq. (4) for $h_\star$, which now requires only observed p_{XYZ} (Sun et al., 2025).

2.3 Partial identification and sensitivity analysis

Uniqueness of the IV solution in Eq. (4) needs a *complete* Z, i.e. $\mathbb{E}[h(X) \,|\, Z] = 0$ iff $h(X) = 0$ (Deaner, 2022). Without completeness, $h_\star$ is unidentifiable (Ailer et al., 2024). Furthermore, the IV properties themselves may be violated (Watson et al., 2024; Penn et al., 2025; Conley et al., 2012; Cinelli & Hazlett, 2025). And since U is unobserved, $h_\star$ is generally not recoverable from p_{XYZ} alone. Nevertheless, milder assumptions on the model in Eq. (1) still allows *partial identification* (PI) of $h_\star$ by considering *all* the joint distributions q consistent with said assumptions (Padh et al., 2023),

$$\mathcal{Q}_{\text{PI}}(p_{XYZ}) \coloneqq \Big\{ q \in \mathcal{P}_{\text{PI}} \,:\, D_{\mathcal{F}}(p_V, q_V) = 0 \Big\}, \qquad D_{\mathcal{F}}(p_V, q_V) \coloneqq \sup_{f \in \mathcal{F}} \Big| \mathop{\mathbb{E}}_{p}[f(V)] - \mathop{\mathbb{E}}_{q}[f(V)] \Big|,$$

where some discrepancy metric $D_{\mathcal{F}}$, chosen here to be an *integral probability metric* (IPM) (Chau et al., 2025), matches q, p on observed variables $V \subseteq \{X, Y, Z\}$ over a set of witness functions $\mathcal{F} = \{f : \mathcal{V} \to \mathbb{R}\}$ and the constraint set $\mathcal{P}_{\text{PI}}$ encodes our assumptions on the DGP Eq. (1). If correctly specified, $p \in \mathcal{P}_{\text{PI}}$, and the following set $\mathcal{H}_{\text{PI}}$ of candidate hypotheses contains the true $h_\star$,

$$\mathcal{H}_{\text{PI}} \coloneqq \big\{ h^q_\star \,:\, q \in \mathcal{Q}_{\text{PI}} \big\}, \qquad\qquad \mathcal{H}_{\text{PI}}(\boldsymbol{x}) \coloneqq \big\{ h^q_\star(\boldsymbol{x}) \,:\, q \in \mathcal{Q}_{\text{PI}} \big\},$$

where $\mathcal{Q}_{\text{PI}}$ abbreviates $\mathcal{Q}_{\text{PI}}(p_{XYZ})$. Computing interval $\mathcal{H}_{\text{PI}}(\boldsymbol{x})$ at $\boldsymbol{x}$ is often more practical than characterizing set $\mathcal{H}_{\text{PI}}$, as for convex $\mathcal{Q}_{\text{PI}}$ it amounts to solving two constrained optimization problems

$$\mathcal{H}_{\text{PI}}(\boldsymbol{x}) = \big[\, \min\nolimits_{q \in \mathcal{Q}_{\text{PI}}} h^q_\star(\boldsymbol{x}), \quad \max\nolimits_{q \in \mathcal{Q}_{\text{PI}}} h^q_\star(\boldsymbol{x}) \,\big].$$

In either case, we want the identified sets to (i) contain the true solution, and (ii) be sharp, i.e. as small as possible. Of course these respectively depend on the validity and strength of assumptions that one is willing to make within $\mathcal{P}_{\text{PI}}$, with point-estimation now being a special case where $\mathcal{Q}_{\text{PI}}$ is singleton.

Hence, one may parameterize the constraints as $\mathcal{P}^{\boldsymbol{\gamma}}_{\text{PI}}$ to conduct *sensitivity analysis* by varying parameters $\boldsymbol{\gamma}$ to see how $\mathcal{H}^{\boldsymbol{\gamma}}_{\text{PI}}(\boldsymbol{x}), \mathcal{H}^{\boldsymbol{\gamma}}_{\text{PI}}$ evolve as assumptions are relaxed/tightened (Frauen et al., 2024).

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

$$E^-(\mathcal{H}_{\text{PI}}) \coloneqq \min_{h \in \mathcal{H}_{\text{PI}}} E(h), \qquad \overline{E}(\mathcal{H}_{\text{PI}}) \coloneqq \mathbb{E}_{h \sim \text{Unif}(\mathcal{H}_{\text{PI}})}[E(h)], \qquad E^+(\mathcal{H}_{\text{PI}}) \coloneqq \max_{h \in \mathcal{H}_{\text{PI}}} E(h).$$

The *approximation error* E^- measures how far the target $h_\star$ is from $\mathcal{H}_{\text{PI}}$ (Brown & Ali, 2024), enabling a notion of approximate validity: $\mathcal{H}_{\text{PI}}$ is ϵ-*approximately valid*, or just ϵ-*valid*, whenever $E^-(\mathcal{H}_{\text{PI}}) \leq \epsilon$ (Li et al., 2026). The *worst-case error* E^+ gives our primary notion of bound *informativeness* for decision making, upper bounding the excess risk of candidates within $\mathcal{H}_{\text{PI}}$ relative to $h_\star$. We additionally introduce the *average error* $\overline{E}$ (Gustafson, 2010) as an analytic surrogate for informativeness, approximating worst-case error up to a universal constant factor (see Lem. 4).

Similarly, the intervals $\mathcal{H}_{\text{PI}}(\boldsymbol{x})$ are evaluated using $E^-_{\boldsymbol{x}}(\mathcal{H}_{\text{PI}}(\boldsymbol{x}))$, $\overline{E}_{\boldsymbol{x}}(\mathcal{H}_{\text{PI}}(\boldsymbol{x}))$ and $E^+_{\boldsymbol{x}}(\mathcal{H}_{\text{PI}}(\boldsymbol{x}))$. We define these analogously to $E^-, \overline{E}$ and E^+ respectively, but with $R_{\boldsymbol{x}}$ as the primitive instead of R.

Choice of constraints. The construction of $\mathcal{P}_{\text{PI}}$ depends on domain knowledge. At the agnostic extreme, Manski's (1990) "no-assumption" bounds use only a bounded support of Y to constrain $\mathcal{H}_{\text{PI}}(\boldsymbol{x})$ for binary $\boldsymbol{x}$. Other approaches bound the strength of unmeasured confounding through odds ratios (Rosenbaum, 2002), propensity scores in the Marginal Sensitivity Model (MSM) (Tan, 2006), or the share of variance explained by hidden confounders in the partial r^2 model (Cinelli & Hazlett, 2020). Robust optimization (RO) formulations instead take $\mathcal{Q}_{\text{PI}}$ to be a total-variation (Bertsimas et al., 2023; Guo et al., 2022) or Wasserstein (Meresht et al., 2022) ball around p_{XYZ}. Alternatively, one may directly model the *response function* $H_\star(\cdot) \coloneqq h_\star(\cdot) + U + \xi$ (Padh et al., 2023), e.g., with GANs matched to p_{XYZ} (Hu et al., 2021). Many methods further sharpen bounds via conditional independences on auxiliary variables (Ailer et al., 2024; Zheng et al., 2025). A notable example is IV-based PI (Balke & Pearl, 1997), where U may affect Y non-additively, unlike Eq. (1). Neural

variants extend this to continuous, high-dimensional treatments and IVs (Kilbertus et al., 2020; Hu et al., 2021; Padh et al., 2023; Meresht et al., 2022; Gunsilius, 2020).

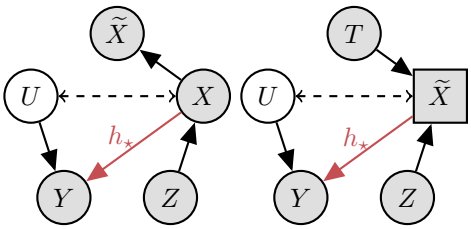
Perhaps the closest to our work are *shape constraints* (Chetverikov et al., 2018), which sharpen PI by restricting the class of structural or response functions $h_\star$ or $H_\star$. Examples include positivity (Leamer, 1981), monotonicity (Manski, 1997; Manski & Pepper, 2000), convexity (Freyberger & Horowitz, 2015), curvature (Melnychuk et al., 2023), smoothness (Kim et al., 2018), sparsity (Pfister & Peters, 2022), and other regularities (Meresht et al., 2022), each narrowing the set of compatible DGPs. In the same vein, we propose known invariances of $h_\star$ as a natural shape constraint.

2.4 Known data symmetries and invariance

Symmetries are ubiquitous in scientific modeling, describing structural properties that do not change under certain data transformations (Bronstein et al., 2021). Formalized as a map $\alpha : \mathcal{X} \times \mathcal{T} \to \mathcal{X}$ parameterized by $\boldsymbol{\tau} \in \mathcal{T}$ with $\boldsymbol{\tau}\boldsymbol{x} \coloneqq \alpha(\boldsymbol{x}, \boldsymbol{\tau})$ as shorthand, we say that $h_\star$ is *invariant* under $\mathcal{T}$, or $\mathcal{T}$*-invariant*, if

$$h_\star(\boldsymbol{\tau}\boldsymbol{x}) = h_\star(\boldsymbol{x}), \qquad \forall\, (\boldsymbol{x}, \boldsymbol{\tau}) \in \mathcal{X} \times \mathcal{T}.$$

This captures the notion that certain transformations preserve the causal relationship $X \to Y$. E.g., spatial translations in physics (Brandstetter et al., 2022), permutations of exchangeable units (Zaheer et al., 2017), semantics-preserving text transformations (Ribeiro et al., 2020), and reverse-complements in genomics (Weinstein & Blei, 2026). In image classification, for instance, a rotated image $\boldsymbol{\tau}\boldsymbol{x}$ of a cat $\boldsymbol{x}$ remains a cat (Cohen & Welling, 2016).



(a) Data augmentation (b) Soft-intervention

Figure 3: For $\mathcal{T}$-inv. $h_\star$, *(a)* DA on observed data induces an identical distribution to *(b)* a soft intervention $\mathrm{do}(X = \widetilde{X})$ on Eq. (1).

In practice, symmetries may hold only approximately, and $h_\star$ is $(\varepsilon, \mathcal{T})$*-inv.* if (Huang et al., 2023)

$$\sup_{\boldsymbol{x},\boldsymbol{\tau}} |h_\star(\boldsymbol{x}) - h_\star(\boldsymbol{\tau}\boldsymbol{x})| \leq \varepsilon.$$

Symmetry assumptions are well-established across many applications and routinely used to improve statistical estimation. For example, given some probability measure p_T on $\mathcal{T}$, *data augmentation* (DA) reduces estimation variance by generating multiple m augmentations $\{(\boldsymbol{\tau}_{i,j}\boldsymbol{x}_i,\ y_i)\}_{j\in[m]}, \boldsymbol{\tau}_{i,j} \sim p_T$ for each sample $(\boldsymbol{x}_i, y_i, \cdot) \in \mathcal{D}$, increasing data variability for statistical risk evaluation (Lyle et al., 2020; Chen et al., 2020; Savkin et al., 2020),

$$R_{\widetilde{\mathrm{erm}}}(h) \coloneqq \mathbb{E}\big[\ell\big(Y, h(\widetilde{X})\big)\big], \tag{5}$$

where we denote $\widetilde{X} \coloneqq TX$ as random perturbation of X under $T \sim p_T$, such as depicted in Fig. 3a.

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.

Despite their success in statistical inference, symmetry priors remain understudied in causal inference (Akbar et al., 2025). This gap is particularly consequential for PI, where bounds are often vacuous and practitioners seek to exhaustively impose constraints encoding domain knowledge to sharpen them. In the language of §2.3, $\mathcal{T}$-inv. is precisely a shape constraint on $h_\star$. Symmetry assumptions hence provide a natural and powerful addition to this toolkit, especially in domains where they are already established. In what follows, we provide a first analysis of their application to sharpen PI.

3 A Symmetry Shape Constraint

Following the shape constraint tradition in PI, the most direct way is to incorporate $\mathcal{T}$-inv. as a solver feasibility requirement. E.g., impose an *invariance error* constraint with ε as a sensitivity parameter:

$$E_{\mathrm{inv}}(h) \coloneqq \mathbb{E}_p\big[|h(X) - h(\widetilde{X})|^2\big], \qquad \mathcal{P}_{\mathrm{PI+inv}} \coloneqq \big\{\ q \in \mathcal{P}_{\mathrm{PI}}\ :\ E_{\mathrm{inv}}(h^q_\star) \leq \varepsilon^2\ \big\}.$$

yielding the *sharp* set $\mathcal{H}_{\mathrm{PI+inv}} = \mathcal{H}_{\mathrm{PI}} \cap \mathcal{H}^{\varepsilon}_{\mathrm{inv}}$, with $\mathcal{H}^{\varepsilon}_{\mathrm{inv}} \coloneqq \big\{h \in \mathcal{H} : E_{\mathrm{inv}}(h) \leq \varepsilon^2\big\}$ (Cor. 1, cf. §D). Other E_{inv} functionals may serve equally (Yang et al., 2019), as does restricting $\mathcal{H}$ to be invariant by design (Cohen & Welling, 2016), though the latter can sit awkwardly with standard PI solvers and may not admit sensitivity parameters.

While our experiments focus on settings where $\mathcal{H}_{\text{PI+INV}}(\boldsymbol{x})$ is computable via convex programming, this constraint formulation is fully compatible with modern deep learning-based PI. Since $E_{\text{INV}}(\cdot)$ is differentiable and amenable to Monte Carlo evaluation, it can be incorporated as a regularizer in augmented Lagrangian or gradient-based solvers (Padh et al., 2023; Meresht et al., 2022; Hu et al., 2021). Nevertheless, there are operational limitations: it requires white-box access to the solver's objective, which precludes the specialized or black-box software where the optimization is fixed. Furthermore, optimizing non-convex invariance constraints for complex, high-dimensional data can be expensive and notoriously unstable (Schweisthal et al., 2025; Padh et al., 2023). This motivates our second approach; an operationalization that touches only the data, not the solver.

4 A Change of Measure under Symmetry Transformations

DA as a change of measure. Where §3 constrains the feasible set, we instead change the measure $p_{XYZ} \to p_{\widetilde{X}YZ}$ under which PI is posed while leaving the solver untouched, composing with it as

$$\mathcal{Q}_{\widetilde{\text{PI}}}(p_{XYZ}) \coloneqq \mathcal{Q}_{\text{PI}}(p_{\widetilde{X}YZ}).$$

This change of measure reweights candidate evaluations, pruning previously feasible functions while, crucially, preserving validity across broad families of standard solvers. We now isolate the two mechanisms driving these effects: variance inflation (sharpening) and simulated intervention (validity).

Variance inflation. In IV methods, "strong" IVs that induce more exogenous variation in X yield sharper PI bounds than weak ones (Kilbertus et al., 2020; Padh et al., 2023). So we pose the question:

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

We answer yes. Fundamentally, PI uncertainty is an *attribution problem*; the solver must split observed (X, Y) covariation between $h_\star$ and U. Only variation in (X, Y) whose relation to U is *known*, e.g. an IV Z, can break the tie. DA adds exactly such variation since T is exogenous by construction and cannot be attributed to U. This reduces ambiguity in PI, giving sharper, more informative bounds. This effect is formalized in §4.1, here we give a notion of variance inflation itself.

Fit vs. evaluation geometries. Center Y w.l.o.g., then $\mathbb{E}[h_\star(X)] = 0$. So *estimation* of $h_\star$ lives only in the centered *shape space* $\mathcal{H}_X \coloneqq \{h \in \mathcal{H} \colon \mathbb{E}[h(X)] = 0\}$, which we equip with covariance forms

$$\langle f, g\rangle_X \coloneqq \operatorname{Cov}(f(X), g(X)), \qquad \|h\|_X^2 \coloneqq \operatorname{Var}(h(X)),$$

and $(\mathcal{H}_{\widetilde{X}}, \langle\cdot,\cdot\rangle_{\widetilde{X}})$ likewise under $p_{\widetilde{X}}$. Of course, *evaluation* is still reported in the ambient $L_2(X)$-norm; in particular, the causal excess risk under a sq.-loss $\ell(a, b) \coloneqq (a-b)^2$ decomposes orthogonally,

$$E(h) = \left\|h - h_\star\right\|_{L_2(X)}^2 = \underbrace{\|h - h_\star\|_X^2}_{\text{shape; lives on } \mathcal{H}_X} + \underbrace{\left(\mathbb{E}[h(X)] - \mathbb{E}[h_\star(X)]\right)^2}_{\text{level; lives off } \mathcal{H}_X}.$$

Fitting in $(\mathcal{H}_{\widetilde{X}}, \langle\cdot,\cdot\rangle_{\widetilde{X}})$ can change the shape. But $\widetilde{X}$-centered candidates are not X-centered, so a level penalty appears. This forces a separate treatment of covariance and mean, which we do under:

Assumption 1. Let $\mathcal{H} \subseteq L_2(X) \cap L_2(\widetilde{X})$ be closed under constant shifts $\mathcal{H} + \mathbb{R}1 = \mathcal{H}$. Optionally:[1]
(a) Realizability: Suppose $\mathcal{H}$ contains both $h_\star(X)$, $\mathbb{E}[Y|X]$, and post-DA $\mathbb{E}[h_\star(X)|\widetilde{X}]$, $\mathbb{E}[Y|\widetilde{X}]$.
(b) Absolute continuity: $p_X \ll p_{\widetilde{X}}$, i.e. the post-DA distribution covers the baseline distribution.
(c) Finite dim. Hilbert: $\mathcal{H}$ is linear with $\dim(\mathcal{H}_X) = k < \infty$, and $\langle\cdot,\cdot\rangle_X$ positive-definite over $\mathcal{H}_X$.

Variance vs. mean-shifts. We ask how much a solver-unit direction actually weighs under evaluation, on average. Sum the $L_2(X)$-norms over *any* $\widetilde{X}$-orthonormal basis $\{e_i\} \subset \mathcal{H}_{\widetilde{X}}$ (basis indep., cf. §G),

$$\sum\nolimits_i \|e_i\|_{L_2(X)}^2 \quad = \quad \sum\nolimits_i \|e_i\|_X^2 + \sum\nolimits_i \left(\mathbb{E}[e_i(X)] - \mathbb{E}[e_i(\widetilde{X})]\right)^2 \quad = \quad \operatorname{tr}(\mathcal{K}) + D_{\mathcal{B}}^2(p_X, p_{\widetilde{X}}).$$

[1] Any class with a free intercept satisfies the base clause. Clause (a) removes statistical approx. bias to isolate confounding bias, our focus. Clause (b) is essentially free under DA; sample Id $\in \mathcal{T}$, or mix-in p_{XYZ}. It lets the $\mathcal{H}_X$ Hilbert structure carry over to $\mathcal{H}_{\widetilde{X}}$. Clause (c) gives bounded evaluation $h(\boldsymbol{x})$, without which PI intervals $\mathcal{H}_{\text{PI}}(\boldsymbol{x})$ for continuous X are known to be vacuous (Kilbertus et al., 2020; Pearl, 1995; Bonet, 2001; Gunsilius, 2020). Crucially, finite dimension is only needed to compute set volume/avg.-error in closed form.

Here, $\mathrm{tr}(\mathcal{K}) := \sum_i \|e_i\|_X^2$ is the X-variance carried by k directions that are unit-sized under $\widetilde{X}$. And $D_{\mathcal{B}}^2$ is the squared *maximum mean discrepancy* (MMD) over a unit $\widetilde{X}$-ball $\mathcal{B} := \{h \in \mathcal{H}_{\widetilde{X}} : \|h\|_{\widetilde{X}} \leq 1\}$ (Gretton et al., 2012), collecting the e_i mean-shifts by Pythagoras. Read them as a lever and its price. Whenever $\mathrm{tr}(\mathcal{K})/k \leq 1$, a typical unit $\widetilde{X}$-direction carries less than unit X-variance; this is our notion of *variance inflation*, quantifying the added exogenous variation invoked above. We later connect it directly to bound sharpening in Prop. 2 with a price; the familiar $D_{\mathcal{B}}$ from *out-of-distribution* (OOD) generalization literature (Long et al., 2015), since we fit under $p_{\widetilde{X}}$ yet evaluate OOD on p_X. However, DA must first be licensed for PI at all, which we discuss next.

Simulated-intervention. We now give an observation that licenses DA for PI: for $\mathcal{T}$-inv. $h_\star$, DA realizes a *soft intervention* $\mathrm{do}(X = \widetilde{X})$ on Eq. (1) (Akbar et al., 2025; Ilse et al., 2021; Massidda et al., 2023; Jaber et al., 2020; Besserve & Schölkopf, 2022); the structural equation and causal graph are preserved up to relabeling $X \to \widetilde{X}$, as in Fig. 3. Meanwhile, confounding is diluted by the *data processing inequality* (DPI) $\mathrm{Var}(\mathbb{E}[U|X]) \geq \mathrm{Var}(\mathbb{E}[U|\widetilde{X}])$ on the chain $U \leftrightarrow X \to \widetilde{X}$. Crucially, this dilution is observable through the ratio $\rho := \tilde{\sigma}^2/\sigma^2$ of MMSE $\sigma^2 := \min_{h\in\mathcal{H}} R_{\mathrm{erm}}(h)$ and $\tilde{\sigma}^2 := \min_{h\in\mathcal{H}} R_{\widetilde{\mathrm{erm}}}(h)$, with $\rho \geq 1$ since a noisy $\widetilde{X}$ cannot predict Y better than X. Concretely:

Lemma 1 (soft intervention with DA)**.** *For $\mathcal{T}$-inv. $h_\star$, DA acts on Eq. (1) as a* soft-intervention*:*

(A) Structural equation: *Trivially, post-DA variables $(\xi, U, \widetilde{X}, Y, \widetilde{Z})$ with $\widetilde{Z} \in \{Z, (Z, T)\}$ follow:*

$$Y = h_\star(\widetilde{X}) + U + \xi, \qquad \mathbb{E}[U] = \mathbb{E}[\xi] = 0, \qquad \xi \perp\!\!\!\perp (U, \widetilde{X}, \widetilde{Z}).$$

(B) Markov transfer: *Let p be Markov w.r.t. a DAG $\mathcal{G}$ over $V := \{U, X, Y, Z_1, Z_2, \ldots\}$ where $\mathrm{NonDesc}_{\mathcal{G}}(X) = V \backslash \{X, Y\}$. Then, $p_{\widetilde{V}}$ is Markov w.r.t. $\widetilde{\mathcal{G}}$, where $\widetilde{V}, \widetilde{\mathcal{G}}$ relabel X to $\widetilde{X}$ in $V, \mathcal{G}$.*

(C) Attenuated confounding: *Under Asm. 1a, let biases $b := h_{\mathrm{erm}} - h_\star$ and $\tilde{b} := h_{\widetilde{\mathrm{erm}}} - h_\star$. Then,*

$$\|b\|_X^2 - \|\tilde{b}\|_{\widetilde{X}}^2 \;=\; \sigma^2(\rho - 1) \;\geq\; 0.$$

Hence, broad families of shape constraints (Chetverikov et al., 2018), graphical restrictions (Zhang & Bareinboim, 2021) or confounding bounds (Frauen et al., 2024) that are valid under p_{XYZ} continue to remain valid under $p_{\widetilde{X}YZ}$. Subsequent sections instantiate two such canonical PI models to study the specific identified set evaluation metrics we introduced in §2.3.

Fragility of point-estimation. To motivate the PI application of DA, we first understand its limitations in point-estimation. After all, Lem. 1 does beg the question why this should not work.

Proposition 1 (estimation with DA)**.** *Under Asm. 1a, $\mathcal{T}$-inv. $h_\star$ and* bias direction $e_{\tilde{b}} := \tilde{b} \Big/ \|\tilde{b}\|_{\widetilde{X}}$*,*

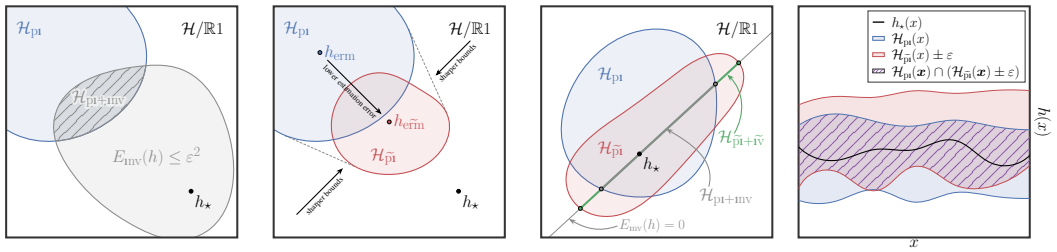
$$E(h_{\mathrm{erm}}) - E(h_{\widetilde{\mathrm{erm}}}) \;=\; \underbrace{\sigma^2(\rho - 1) \cdot \|e_{\tilde{b}}\|_{L_2(X)}^2}_{\textit{gain}} \;+\; \underbrace{(1 - \|e_{\tilde{b}}\|_{L_2(X)}^2) \cdot E(h_{\mathrm{erm}})}_{\textit{wild-card}}.$$

Proposition 1 reveals three fragilities: (i) **bias-direction dependence:** attenuation $\rho \geq 1$ alone does not guarantee improvement, which hinges on whether $\|e_{\tilde{b}}\|_{L_2(X)}^2 \leq 1$ for an unobserved bias direction, (ii) **misspecification:** if symmetry is misspecified, the gain term may itself become a loss, and (iii) **identifiability:** $h_\star$ remains unidentifiable regardless. All three defects are manageable under PI.

4.1 Partial identification with data augmentation

We show that DA yields three PI guarantees, each resolving one fragility above, as summarized in Table 1. **Validity** (Thm. 1) **:** Crucially, PI validity follows from $\mathcal{T}$-invariance alone, addressing (iii). **Sharpness / informativeness** (Prop. 2 and Thm. 2) **:** We show that PI sharpening is governed by *global* quantities like ρ, $\mathrm{tr}(\mathcal{K})$, $D_{\mathcal{B}}$, making sufficient conditions observable, unlike (i). **Robustness** (Thm. 3 and Rem. 1) **:** To make these guarantees robust to ε-misspecified symmetry, we deploy a simple ε-pad-and-intersect post-processing, resolving (ii).

Beyond these population effects, DA has the familiar statistical-efficiency benefit of growing effective sample size, mitigating sampling variation and finite-sample error, key sources of PI uncertainty (Imbens & Manski, 2004). To make these claims explicit, we instantiate the following PI model:



(a) $\mathcal{H}_{\text{PI+INV}}$. (b) $\mathcal{H}_{\widetilde{\text{PI}}}$. (c) $\mathcal{H}_{\widetilde{\text{PI+IV}}}$. (d) $\mathcal{H}_{\text{PI}}(\boldsymbol{x}) \cap (\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) \pm \varepsilon)$

Figure 4: Illustration of identified sets: *(a)* Invariance shape constraint prunes the baseline set (§3). *(b)* DA pre-processing yields a sharper (Prop. 2), better-centered (Prop. 1) set (§4). *(c)* T-as-IV enforces symmetry (§4.2, cf. §D). *(d)* ε-pad-and-intersect dominates under arbitrary DA (Rem. 1).

Assumption 2 (a bounded confounding sensitivity model)**.** For *confounding budgets* $\gamma_{\boldsymbol{x}}, \gamma_y$, let

$$\mathcal{P}_{\text{PI}} := \left\{ q : \begin{matrix} r^2_{U\sim X} \le \gamma_{\boldsymbol{x}}, \\ r^2_{U\sim Y|X} \le \gamma_y \end{matrix} \right\}, \qquad r^2_{A\sim B|C} := 1 - \frac{\mathbb{E}[\mathrm{Var}(A\,|\,B,C)]}{\mathbb{E}[\mathrm{Var}(A\,|\,C)]} \le 1,$$

where $r^2_{A\sim B|C}$ is the *partial-r^2*, i.e. fraction of variation in A explained by B after controlling for C. Further, make the IPM $D_{\mathcal{F}}$ test all quadratic functions $\mathcal{F} := \{f^2 : f \in \mathrm{span}(\mathcal{H}) + \mathbb{R}Y\}$ of (X, Y).

Assumption 2 adopts the widely used partial-r^2 model by Cinelli & Hazlett (2020). While we employ this model in our analysis, we do not necessarily restrict ourselves to it: several PI families in §2.3 impose equivalent constraints, reducing to ellipsoidal sets of the form

Lemma 2 (characterizing the identified set)**.** *Under Asm. 2 and 1a or 1c, $\mathcal{H}_{\text{PI}}$ is the covariance ball*

$$\mathcal{H}_{\text{PI}} = \left\{ h \in \mathcal{H}_X : \left\| h - h_{\text{erm}} \right\|^2_X \le \sigma^2 \gamma \right\},$$

within $\mathcal{H}_X$, centered at h_{erm} with radius $\sigma\sqrt{\gamma}$, and an effective confounding budget *of* $\gamma := \frac{\gamma_{\boldsymbol{x}}\gamma_y}{1-\gamma_{\boldsymbol{x}}}$.

The post-DA analysis now reduces to three questions about this ellipsoid: whether it covers $h_\star$, does its size contract, and if its center/radius tradeoff lowers avg./worst-case errors. We treat these in turn.

Validity. First and foremost, we establish post-DA validity. In terms of approximation error:

Theorem 1 (valid bounds with DA)**.** *Let $h_\star$ be $\mathcal{T}$-inv. under Asm. 1a and 2. Then, the post-DA set $\mathcal{H}_{\widetilde{\text{PI}}}$ is exactly valid whenever the baseline set $\mathcal{H}_{\text{PI}}$ is ϵ-valid with $\epsilon \le \sigma^2\left(\sqrt{\rho - 1 + \gamma\rho} - \sqrt{\gamma}\right)^2$.*

Consequently, $\mathcal{H}_{\widetilde{\text{PI}}}$ *repairs* the ϵ-validity of $\mathcal{H}_{\text{PI}}$ since the previously insufficient budget γ can now accommodate the diluted bias. Crucially, this follows from the simulated intervention in Lem. 1 alone, and does not rely on conditions over unobserved bias like Prop. 1. The only untestable requirement of $\mathcal{T}$-invariance is later relaxed to $(\varepsilon, \mathcal{T})$-invariance to allow for sensitivity analysis over the assumed ε-misspecification of symmetry.

Table 1: Point vs. *partial* identification under DA.

Condition	Identification metric	Safeguard
$\mathcal{T}$-inv. $h_\star \Longrightarrow$	*validity* (Thm. 1) ✓	ε-pad (Thm. 3)
$\|e_{\tilde{b}}\| \le 1 \Longrightarrow$	estimation error ↓ (Prop. 1)	✗
$\Omega \le 1 \implies$	*informativeness* ↑ (Thm. 2) ⇑ *sharpness* ↑ (Prop. 2)	intersect with baseline set (Rem. 1)

Having established validity, we now show DA can make bounds *better*, even under relaxed conditions.

Sharpness. While the centroid drift in Prop. 1 hinges on the bias direction, the identified set's *size* depends on *global* distribution shifts. This yields a milder, data-observable improvement condition:

Proposition 2 (sharper bounds with DA)**.** *Denoting $L_2(X)$-volume by $|\cdot|$, under Asms. 1b–1c and 2,*

$$\left(\frac{|\mathcal{H}_{\widetilde{\text{PI}}}|}{|\mathcal{H}_{\text{PI}}|}\right)^{2/k} \le \frac{\mathbb{E}\left[|\mathcal{H}_{\widetilde{\text{PI}}}(X)|^2\right]}{\mathbb{E}\left[|\mathcal{H}_{\text{PI}}(X)|^2\right]} = \underbrace{\rho}_{\text{over-budget}} \cdot \Big[\underbrace{\mathrm{tr}(\mathcal{K})}_{\text{lever}} + \underbrace{D^2_{\mathcal{B}}(p_X, p_{\tilde{X}})}_{\text{penalty}} \Big] \Big/ k =: \Omega.$$

Proposition 2 ties sharpening to competing factors. Noise inflation $\rho \geq 1$ and mean-shift $D_{\mathcal{B}}^2 \geq 0$ strictly expand the set. Thus, bounds tighten when inflated variance $\mathrm{tr}(\mathcal{K})/k \leq 1$ drives Ω small enough to satisfy $\Omega \leq 1$. Intuitively, Lem. 2 enforces a *variance cap* on allowable bias. DA can stretch admissible candidates beyond this fixed γ-budget, pruning them from the set. Notably, the soft intervention Lem. 1 keeps the true bias $\tilde{b}$ within budget even as other candidates are pruned. Finally, as the inherited γ over-budgets the now diluted confounding by ρ, one may re-calibrate to $\tilde{\gamma} \coloneqq \gamma/\rho$ if the solver exposes γ, keeping validity while reducing Prop. 2's term to $\Omega = (\mathrm{tr}(\mathcal{K}) + D_{\mathcal{B}}^2)/k$.

Informativeness. While smaller sets are desirable, we also gauge the "quality" of set candidates:

Theorem 2 (informative bounds with DA)**.** *Under Asms. 1b–1c and 2, the following holds for $d = k$,*

$$\overline{E}(\mathcal{H}_{\text{PI}}) - \overline{E}(\mathcal{H}_{\widetilde{\text{PI}}}) \quad = \quad \underbrace{E(h_{\text{erm}}) - E(h_{\widetilde{\text{erm}}})}_{\text{centroid drift (Prop. 1)}} + \frac{\sigma^2 \gamma k}{d+2} \cdot \underbrace{(1-\Omega)}_{\text{set size (Prop. 2)}} \quad =: \quad f(d),$$

And interval improvement averaged over queries: $\mathbb{E}_{X \sim p_X}\left[\overline{E}_X(\mathcal{H}_{\text{PI}}(X)) - \overline{E}_X(\mathcal{H}_{\widetilde{\text{PI}}}(X))\right] = f(1)$.

The decomposition isolates two mechanisms: the centroid drifts toward $h_\star$ and the set shrinks, as illustrated in Fig. 4b. Crucially, thresholds $\|e_{\tilde{b}}\|_{L_2(X)}^2 \leq 1$ and $\Omega \leq 1$ are *sufficient*, and not necessary for improvement. In general, DA "wobbles" the ellipsoid of Lem. 2, inflating some directions while deflating others. The intersection $\mathcal{H}_{\text{PI}}(\boldsymbol{x}) \cap \mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x})$ nonetheless guarantees no regression.

Robustness. DA encoding misspecified $\mathcal{T}$-inv. induces endogenous noise $\eta \coloneqq h_\star(X) - h_\star(\widetilde{X})$,

$$Y = h_\star(X) + U + \xi, \qquad \Longleftrightarrow \qquad Y = h_\star(\widetilde{X}) + U + \eta + \xi,$$

The post-DA intervals $\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x})$ may not account for this additional confounding. Therefore, to recover validity we apply ε-padding $[a, b] \pm \varepsilon \coloneqq [a - \varepsilon, b + \varepsilon]$.[2]Baseline intersection then absorbs any slack.

Theorem 3.A (robust bounds with DA)**.** *Under Asms. 1a–1b and 2, let valid baseline set $\mathcal{H}_{\text{PI}}$ cover an $(\varepsilon, \mathcal{T})$-invariant $h_\star$. Then, ε-padded post-DA intervals $\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) \pm \varepsilon$ also remain valid on p_X-a.e. $\boldsymbol{x}$.*

Remark 1 (dominant bounds with DA)**.** Under Asms. 1a–1b and 2, let $\mathcal{H}_{\text{PI}}$ cover $(\varepsilon, \mathcal{T})$-invariant $h_\star$. Then, $\mathcal{H}_{\text{PI}}(\boldsymbol{x}) \cap (\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) \pm \varepsilon)$ dominates $\mathcal{H}_{\text{PI}}(\boldsymbol{x})$ on sharpness and worst-case error for p_X-a.e. $\boldsymbol{x}$.

Theorem 3.A now enables DA based sensitivity analysis via the parameter ε; validity is preserved by ε-padding bounds, while baseline intersection maintains sharpness and informativeness. This is depicted in Fig. 4d where post-DA bounds may arbitrarily change, yet intersection guarantees dominance. It is worth contrasting this with §3; unlike $\mathcal{H}_{\text{PI+IV}}(\boldsymbol{x})$, $\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) \pm \varepsilon$ solves the entire ε-grid simultaneously.

4.2 Symmetry constraint with T-as-IV

The IV problem in Eq. (4) may be incomplete (Ailer et al., 2024), or even misspecified (Watson et al., 2024; Penn et al., 2025; Conley et al., 2012; Deaner, 2022; 2025; Cinelli & Hazlett, 2025). Consequently, $h_\star$ is unidentifiable. Nevertheless, even invalid instruments retain value for sharpening PI:

Assumption 3 (a leaky IV sensitivity model (Deaner, 2022))**.** For instrument violation budget $\gamma_{\boldsymbol{z}}$, let

$$\mathcal{P}_{\text{PI+IV}} \coloneqq \left\{ q \in \mathcal{P}_{\text{PI}} \,:\, \left\|\mathbb{E}_p[Y - h_\star^q(X) \,|\, Z]\right\|_Z^2 \leq \sigma^2 \gamma_{\boldsymbol{z}} \right\}.$$

Here, $\gamma_{\boldsymbol{z}}$ represents a bounded violation in un-confoundedness and/or exclusion of Z, while the baseline constraints $\mathcal{P}_{\text{PI}}$ may allow $\mathcal{P}_{\text{PI+IV}}$ to be sufficiently well-posed to yield non-vacuous bounds.

Can §4.1 be pushed further, to impose symmetry explicitly within an IV solver? Apparently so! Post-DA, the random transformation T satisfies the IV properties of un-confoundedness $T \perp\!\!\!\perp U$ and exclusion $Y \perp\!\!\!\perp T \,|\, \widetilde{X}, U$ by construction. But $\mathcal{T}$-invariance renders it incomplete, since $T \perp\!\!\!\perp X$ gives $\mathbb{E}[h_\star(\widetilde{X}) \,|\, T] = \mathbb{E}[h_\star(X) \,|\, T] = \mathbb{E}[h_\star(X)] = 0$. Because incomplete instruments still constrain PI, we plug T in as an IV $\widetilde{Z} \coloneqq (Z, T)$ to enforce symmetry within off-the-shelf or black-box IV solvers,

$$\mathcal{Q}_{\widetilde{\text{PI}}+\widetilde{\text{IV}}}(p_{XYZ}) \coloneqq \mathcal{Q}_{\text{PI+IV}}(p_{\widetilde{X}Y\widetilde{Z}}).$$

We defer readers to §C and §D for additional analysis under Asm. 3, including comparison with §3.

[2]For non-interval $\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x})$, read $\pm\varepsilon$ as the Minkowski sum $\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) \pm \varepsilon \coloneqq \mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) + [-\varepsilon, \varepsilon]$. Also, note that re-calibrating to *any* bigger budget $\tilde{\gamma} \geq \gamma$ may not suffice for validity since $\mathbb{E}[\eta] \neq 0$ is not represented by $\mathcal{H}_{\widetilde{X}}$.

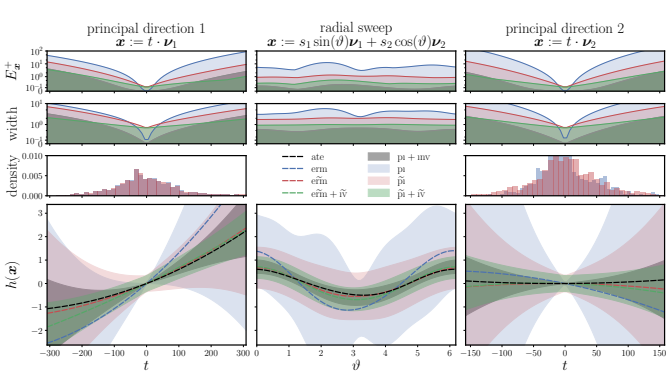


Figure 5: **Optical device data.** Across queries $\boldsymbol{x}$ along principal components $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2$ of $\mathrm{Var}(X)$ and a radial sweep, symmetries reduce interval width and worst-case error *(Top)* over baseline while maintaining validity *(Bottom)*.

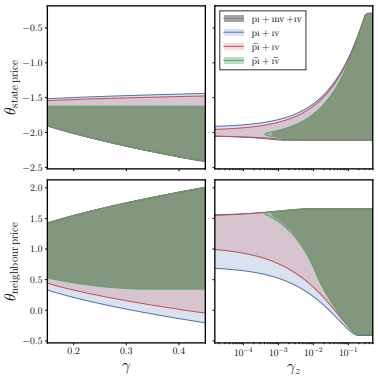


Figure 6: **US cigarette demand.** State *(Top)* and neighbour *(Bottom)* price elasticity across confounding γ *(Left)* and IV $\gamma_{\boldsymbol{z}}$ *(Right)* budgets.

5 Experiments

We validate our methods in finite samples. To isolate symmetry-driven sharpening from mere statistical efficiency, we fix the augmented dataset size to match the original. Other implementation details and supplementary ablations are discussed in §F. Our implementation is available at the following: https://anonymous.4open.science/r/causal-bounds-submission-0A1F

Optical device experiment. The dataset of Janzing & Schölkopf (2018) pairs 3×3 pixel images X shown on a screen driving voltage readings Y across a photo-diode. A physical latent U drives two LEDs, one affecting the webcam capturing X, the other the photo-diode. As image orientation does not affect irradiance, so we augment X by flips and $90°$ rotations. The true $h_\star$ is obtained by back-door adjustment on the recorded U. Figure 5 shows that symmetries shrink both interval width and worst-case error across queries, often by over an order of magnitude, while retaining validity.

US cigarette demand. We give an illustrative example of a textbook symmetry in a canonically causal setting; Scaling all prices and income by a common factor (e.g., inflation) should leave demand unchanged. On a panel of 49 US states (1970–2019) (Baltagi & Levin, 1986), consider Y the log taxed sales per capita of cigarettes and X the log state price, cheapest neighbour price, income and the consumer price index (CPI). We assume the latter two are exogenous, whereas the former two are confounded with (possibly leaky) instruments Z being the state's cigarette tax. Thus $h_\star(\cdot) := \boldsymbol{\theta}_\star^\top \log(\cdot)$ is invariant to translating X along $\mathbf{1}$, which we use as DA. Figure 6 shows PI bounds on the two endogenous price coefficients $\boldsymbol{\theta}_\star$, this symmetry makes the neighbour-price effect clearly positive unless the IV is heavily leaky, i.e., evidence of cross-border shopping.

do-MNIST experiment. Inspired by Arjovsky et al. (2019), we derive a semi-synthetic dataset from MNIST digits (LeCun et al., 1998), dubbed do-MNIST. A hidden U both colours the digit image X (red vs. blue) and corrupts its label Y, spuriously correlating colour with the label, while $h_\star(\boldsymbol{x})$ depends only on whether the digit is ≥ 5. We use the CopSens model of Zheng et al. (2025) with convolutional networks, and DA by small translations, rotations and hue jitter, none of which alter the digit. In Fig. 7, ERM tracks colour, and symmetries meaningfully sharpen PI around $h_\star$.

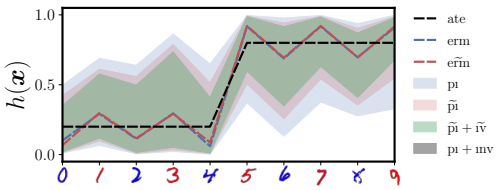


Figure 7: do-**MNIST data.** Bounds on 10 exemplar images with alternating colours. ERM follows colour; symmetries sharpen PI. Invariance shape constraint pi + inv is infeasible here (§F).

Conclusion. Researchers are advised to incorporate established domain symmetries as a standard component of their partial-identification toolkit. Integration is effortless with DA pre-processing. Further discussion, including limitations, broader applicability and future work is provided under §B.

AI USE STATEMENT

Generative AI tools were used to assist with LaTeX template adaptation, layout checking, and compilation verification for the ICLR submission. The scientific content and prose remain the authors' responsibility.

REPRODUCIBILITY STATEMENT

Implementation details, supplementary experiments, and complete proofs are provided in the appendices.

REFERENCES

Elisabeth Ailer, Niclas Dern, Jason Hartford, and Niki Kilbertus. Targeted sequential indirect experiment design. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 122029–122053. Curran Associates, Inc., 2024. doi: 10.52202/079017-3877. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dcc337bb2a4d25afefd9ab800721debb-Paper-Conference.pdf.

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, 2025. doi: 10.52202/085713-3290.

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

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.

Badi H Baltagi and Dan Levin. Estimating dynamic demand for cigarettes using panel data: The effects of bootlegging, taxation and advertising reconsidered. *The Review of Economics and Statistics*, 68(1):148–155, February 1986. doi: None. URL https://ideas.repec.org/a/tpr/restat/v68y1986i1p148-55.html.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems 32*. Curran Associates Inc., 2019.

Dimitris Bertsimas, Kosuke Imai, and Michael Lingzhi Li. Distributionally robust causal inference with observational data, 2023. URL https://arxiv.org/abs/2210.08326.

Michel Besserve and Bernhard Schölkopf. Learning soft interventions in complex equilibrium systems. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 170–180. PMLR, 2022.

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 2001. ISBN 1558608001.

Johannes Brandstetter, Max Welling, and Daniel E Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 2022.

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.

Siu Lun Chau, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

Leena 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, pp. 36627–36639. Curran Associates, Inc., 2022. doi: 10.52202/068431-2654. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ed7b8e1312f6ba8af6e4316dcd28bb3d-Paper-Conference.pdf.

Denis Chetverikov, Andres Santos, and Azeem M. Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10:pp. 31–63, 2018. ISSN 19411383, 19411391.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 2020. ISSN 1369-7412. doi: 10.1111/rssb.12348.

Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Biometrika*, 112(2), 2025. ISSN 1464-3510. doi: 10.1093/biomet/asaf004.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, 2016.

Timothy G. Conley, Christian B. Hansen, and Peter E. Rossi. Plausibly exogenous. *The Review of Economics and Statistics*, 94(1), 2012. ISSN 00346535, 15309142.

Hugh Dance and Benjamin Bloem-Reddy. Counterfactual cocycles: A framework for robust and coherent counterfactual transports, 2025. URL https://arxiv.org/abs/2405.13844.

Ben Deaner. Nonparametric instrumental variables estimation under misspecification, 2022. URL https://arxiv.org/abs/1901.01241.

Ben Deaner. The trade-off between flexibility and robustness in instrumental variables analysis. *American Economic Review*, 115(11):3975–98, November 2025. doi: 10.1257/aer.20231765. URL https://www.aeaweb.org/articles?id=10.1257/aer.20231765.

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.

Joachim Freyberger and Joel L. Horowitz. Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, 189(1), 2015. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2015.06.020.

AmirEmad Ghassami, Yunshu Zhang, Ilya Shpitser, and Eric Tchetgen Tchetgen. Partial identification of causal effects using proxy variables, 2026. URL https://arxiv.org/abs/2304.04374.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

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.

Paul Gustafson. Bayesian inference for partially identified models. *International Journal of Biostatistics*, 6(2), 2010. doi: 10.2202/1557-4679.1206.

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.

Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

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

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.

Wooyoung Kim, Koohyun Kwon, Soonwoo Kwon, and Sokbae Lee. The identification power of smoothness assumptions in models with counterfactual outcomes. *Quantitative Economics*, 9(2), 2018. doi: https://doi.org/10.3982/QE545.

Edward E. Leamer. Is it a demand curve, or is it a supply curve? partial identification through inequality constraints. *The Review of Economics and Statistics*, 63(3), 1981. ISSN 00346535, 15309142.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 1998. doi: 10.1109/5.726791.

Ang Li, Scott Mueller, Xin Shu, and Judea Pearl. ϵ-identifiability of causal quantities. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, 2015.

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.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990. ISSN 00028282.

Charles F. Manski. Monotone treatment response. *Econometrica*, 65(6):1311–1334, 1997. ISSN 00129682, 14680262.

Charles F. Manski and John V. Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010, 2000. doi: https://doi.org/10.1111/1468-0262.t01-1-00144a. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.t01-1-00144a.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. 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*. PMLR, 2023.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 7512–7523. PMLR, 2021.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5), 2003. ISSN 00129682, 14680262.

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. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995. ISBN 1558603859.

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

Jordan Penn, Lee M. Gunderson, Gecia Bravo-Hermsdorff, Ricardo Silva, and David Watson. BudgetIV: Optimal partial identification of causal effects with mostly invalid instruments. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2485–2493. PMLR, 03–05 May 2025.

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

Niklas Pfister and Jonas Peters. Identifiability of sparse causal effects using instrumental variables. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Association for Computational Linguistics (ACL)*, 2020.

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

Gian-Carlo Rota. Reynolds operators. In Richard Bellman (ed.), *Stochastic Processes in Mathematical Physics and Engineering*, pp. 70–83. American Mathematical Society, 1964. doi: 10.1090/psapm/016/0161140.

Artem Savkin, Thomas Lapotre, Kevin Strauss, Uzair Akbar, and Federico Tombari. Adversarial appearance learning in augmented Cityscapes for pedestrian recognition in autonomous driving. In *IEEE International Conference on Robotics and Automation*, pp. 3305–3311, 2020. doi: 10.1109/ICRA40945.2020.9197024.

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.

Haotian Sun, Antoine Moulin, Tongzheng Ren, Arthur Gretton, and Bo Dai. Spectral representation for causal estimation with hidden confounders. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, 2025.

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.

David Watson, Jordan Penn, Lee Gunderson, Gecia Bravo-Hermsdorff, Afsaneh Mastouri, and Ricardo Silva. Bounding causal effects with leaky instruments. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 3689–3710. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/watson24a.html.

Eli N. Weinstein and David M. Blei. Geometric causal models, 2026. URL https://arxiv.org/abs/2607.05153.

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.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

J. Zhang and E. Bareinboim. Non-parametric methods for partial identification of causal effects. Technical Report R-72, Causal Artificial Intelligence Lab, Columbia University, Feb 2021.

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

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *Journal of Machine Learning Research*, 26(36):1–60, 2025. URL http://jmlr.org/papers/v26/22-0372.html.

APPENDIX—SYMMETRY-INFORMED CAUSAL PARTIAL IDENTIFICATION

CONTENTS

A List of Symbols . . . 16
B Discussion . . . 17
C Misspecified DA under IV Constraints . . . 18
D Shape Constraint vs. Change of Measure . . . 19
E Implementation Details . . . 20
F Experiment Supplement . . . 21
F.1 Experiment Setup . . . 21
F.2 Ablations . . . 23
G Proofs . . . 26
G.1 Proof of Thm. 1—Valid bounds with DA . . . 27
G.2 Proof of Thm. 2—Informative bounds with DA . . . 28
G.3 Proof of Thms. 3.A and 3.B—Robust bounds with DA . . . 29
G.4 Proof of Prop. 1—Estimation with DA . . . 30
G.5 Proof of Prop. 2—Sharper bounds with DA . . . 31
G.6 Proof of Prop. 3—Shape constraint vs. change-of-measure . . . 32
G.7 Proof of Prop. 4—Sharp bounds with structural-function constraints . . . 33
G.8 Proof of Lem. 1—Soft intervention with DA . . . 34
G.9 Proof of Lem. 2—Characterizing the identified set . . . 35
G.10 Proof of Lem. 3—Shift decomposition . . . 38
G.11 Proof of Lem. 4—Average error as a worst-case surrogate . . . 39
G.12 Proof of Lem. 5—Reynolds projection with DA . . . 40
G.13 Proof of Lem. 6—Average Representer Norms . . . 41

A List of Symbols

Symbol	Description
Variables and Spaces	
$X, \mathcal{X}, \boldsymbol{x}$	Treatment, its support, and a realization
$Y, \mathcal{Y}, y$	Outcome, its support, and a realization
$Z, \mathcal{Z}, \boldsymbol{z}$	Pre-treatment co-variates / instruments (IVs), their support, and realization
U, ξ	Unobserved confounder, exogenous noise (mean-zero, $\xi \perp\!\!\!\perp (U, X, Z)$)
$\widetilde{X}, \widetilde{Z}$	Post-DA treatment $\widetilde{X} \coloneqq TX$ and instrument $\widetilde{Z} \coloneqq (Z, T)$
Hypotheses and Estimators	
$h, h_\star, \mathcal{H}$	Candidate hypothesis, causal target, hypothesis class
$\mathcal{H}_X, \mathcal{H}_{\widetilde{X}}, k$	X-, $\widetilde{X}$-centered $\mathcal{H}$ with covariance forms $\langle\cdot,\cdot\rangle_X, \langle\cdot,\cdot\rangle_{\widetilde{X}}$ and $k \coloneqq \dim(\mathcal{H}_X)$
$h_{\mathrm{erm}}, h_{\widetilde{\mathrm{erm}}}, b, \tilde{b}$	ERM estimators $h_\bullet \coloneqq \operatorname{argmin}_{h\in\mathcal{H}} R_\bullet(h)$, their respective biases $h_\bullet - h_\star$
$h_{\mathrm{IV}}, h_\star^p$	IV estimator, and back-door adjustment over joint p
$g_{\boldsymbol{x}}, \tilde{g}_{\boldsymbol{x}}$	Riesz representers for point-evaluation $h \mapsto h(\boldsymbol{x})$ in $\mathcal{H}_X$ and $\mathcal{H}_{\widetilde{X}}$
$\mathcal{H}_{\mathrm{inv}}, \mathcal{I}, \mathcal{\pi}$	$\mathcal{\pi}$ orthogonally projects onto ambient invariant space $\mathcal{I}$, and $\mathcal{H}_{\mathrm{inv}} \coloneqq \mathcal{H} \cap \mathcal{I}$
Symmetries, Interventions and Change of Measure	
$T, \mathcal{T}, \boldsymbol{\tau}, \alpha$	Random transformation, support, realization, and defining map $\boldsymbol{\tau x} = \alpha(\boldsymbol{x}, \boldsymbol{\tau})$
$\mathrm{do}(\boldsymbol{x}), \mathrm{do}(X = \widetilde{X})$	Hard intervention $\mathrm{do}(X = \boldsymbol{x})$, and DA-simulated soft intervention
ε, η	Invariance residual $\eta \coloneqq h_\star(X) - h_\star(\widetilde{X})$ uniformly bounded $\lvert\eta\rvert \le \varepsilon$ a.s.
$\mathcal{S}, \mathcal{K}, \delta^*\delta$	Shift operator on $\mathcal{H}_{\widetilde{X}}$, splits into shape/variance + rank-1 level/mean shifts
$\sigma^2, \tilde{\sigma}^2, \rho$	Baseline / post-DA MMSE, and the noise-inflation factor $\rho \coloneqq \tilde{\sigma}^2/\sigma^2 \ge 1$
Constraints and Partial Identification	
p, q	True observational joint, and a candidate joint distribution in $\mathcal{Q}_{\mathrm{PI}}$
$D_{\mathcal{F}}, \mathcal{F}$	IPM and its witness-function class defining $\mathcal{Q}_{\mathrm{PI}}$
$\gamma_{\boldsymbol{x}}, \gamma_y, \gamma$	Confounding budgets for X, Y; effective budget $\gamma \coloneqq \gamma_{\boldsymbol{x}}\gamma_y/(1-\gamma_{\boldsymbol{x}})$
$\gamma_{\boldsymbol{z}}, \gamma_{\boldsymbol{z}}^\star, \tilde{\gamma}_{\boldsymbol{z}}$	Assumed, true, and adjusted IV violation budgets
$\mathcal{P}_{\mathrm{PI}}, \mathcal{P}_{\mathrm{PI+INV}}, \mathcal{P}_{\mathrm{PI+IV}}$	PI constraint sets: baseline, augmented with invariance / IV constraints
$\mathcal{Q}_{\mathrm{PI}}, \mathcal{Q}_{\widetilde{\mathrm{PI}}}, \mathcal{Q}_{\widetilde{\mathrm{PI}}+\widetilde{\mathrm{IV}}}$	Candidate joints: baseline, post-DA, and DA-as-IV
$\mathcal{H}_{\mathrm{PI}}, \mathcal{H}_{\widetilde{\mathrm{PI}}}$	Identified set: baseline, post-DA
$\mathcal{H}_{\mathrm{PI}}(\boldsymbol{x}), \mathcal{H}_{\widetilde{\mathrm{PI}}}(\boldsymbol{x})$	Identified intervals at query $\boldsymbol{x}$: baseline vs. post-DA
Risks and Evaluation Metrics	
$\ell(y, y')$	Proper, convex loss function (e.g., squared error)
$R_{\mathrm{erm}}(h), R_{\widetilde{\mathrm{erm}}}(h)$	Observational risks, pre- and post-DA
$R_{\boldsymbol{x}}(y), R(h)$	Causal / interventional risks, point-wise and averaged
$E_{\boldsymbol{x}}(y), E(h)$	Causal excess risk/ estimation error $E(h) = R(h) - R(h_\star)$, pointwise via $R_{\boldsymbol{x}}$
$E^-(\mathcal{H}), \overline{E}(\mathcal{H}), E^+(\mathcal{H})$	Approximation (min), averaged, and worst-case (max) errors E over set $\mathcal{H}$
$E^-_{\boldsymbol{x}}(\mathcal{Y}), \overline{E}_{\boldsymbol{x}}(\mathcal{Y}), E^+_{\boldsymbol{x}}(\mathcal{Y})$	Point-wise versions of the above metrics via $E_{\boldsymbol{x}}$ over interval $\mathcal{Y}$
$E_{\mathrm{inv}}(h)$	Invariance error $\mathbb{E}_p\left[\lvert h(X) - h(\widetilde{X})\rvert^2\right]$

B Discussion

Limitations. Our proposed framework provides a practical hierarchy of symmetry constraints for PI: should an invariance shape constraint pose implementation challenges, practitioners can resort to the DA operationalization. If they suspect misspecification by some ε, they need only to ε-pad bounds. Which leads us to the main limitation of our work: while the soft intervention framing in Lem. 1 makes composability of DA with a broad class of disparate PI solvers plausible under well-specified symmetries, it is yet unclear whether this should also be the case under ε-misspecification. Having addressed this question under two canonical, arguably the most popular PI classes (Asm. 2 bounded confounding and Asm. 3 leaky IVs), further investigation of additional models is left as future work.

Broader applicability. Our framework is compatible with *all* PI models via the shape constraint operationalization. The DA approach also naturally extends to any PI or sensitivity model that imposes bounded confounding constraints equivalent to Asm. 2 (MSM (Tan, 2006), RO (Guo et al., 2022), Wasserstein (Meresht et al., 2022), total-variation (Padh et al., 2023), etc.). Under well-specified symmetries, DA also composes with other shape-constraints and auxiliary variable constraints other than IVs. E.g., Asm. 2 can admit conditioning on pre-treatment Z (Cinelli & Hazlett, 2020), which we omit only for clarity of exposition. With invariance in the response function $H_\star(\boldsymbol{x}) = H_\star(\boldsymbol{\tau x})$, the soft-intervention framing in Lem. 1 also extends DA to response function models (Padh et al., 2023), as well as IVs with latent U (Kilbertus et al., 2020), proxy variables (Ghassami et al., 2026; Mastouri et al., 2021; Sun et al., 2025) and others (Meresht et al., 2022; Zhang & Bareinboim, 2021).

Future work. We aim to generalize our change-of-measure framework to (i) general causal queries on $p_{Y\,|\,\mathrm{do}(\boldsymbol{x})}$ (Ailer et al., 2024), (ii) equi-variance in the response function $H_\star(\boldsymbol{\tau x}) = \beta(\boldsymbol{x}, \boldsymbol{\tau}) H_\star(\boldsymbol{x})$ under cocycle β (Dance & Bloem-Reddy, 2025), and (iii) latent U under a $\wp$-Wasserstein notion of equi-variance misspecification $\sup_{\boldsymbol{x},\boldsymbol{\tau}} W_\wp(p_{H_\star(\boldsymbol{\tau x})}, \beta(\boldsymbol{x}, \boldsymbol{\tau})_{\#} p_{H_\star(\boldsymbol{x})}) \leq \varepsilon$ (Chen et al., 2020).

C MISSPECIFIED DA UNDER IV CONSTRAINTS

For completeness, we explore the limits of the ε-padding approach from Thm. 3.A under the IV constraint in Asm. 3. Notice that $\mathcal{T}$-invariance violation manifests as violated IV exclusion / unconfoundedness. This violation may be accommodated by either DA reducing confounding via the DPI $\sigma^2 \leq \tilde{\sigma}^2$, or a sufficient slack in the baseline IV budget $\gamma_{\boldsymbol{z}}$.

Theorem 3.B (robust bounds with DA)**.** *Let a valid model following Asm. 3 yield intervals $\mathcal{H}_{\text{pi+iv}}(\boldsymbol{x})$ under the setting of Thm. 3.A. Then, $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}(\boldsymbol{x}) \pm \varepsilon$ is also valid for $\widetilde{Z} \in \{Z, (Z,T)\}$ on p_X-a.e. $\boldsymbol{x}$ if*

$$\varepsilon \leq \tilde{\sigma}\sqrt{\gamma_{\boldsymbol{z}}} - \sigma\sqrt{\gamma^\star_{\boldsymbol{z}}}, \qquad \text{where } \gamma^\star_{\boldsymbol{z}} \leq \gamma_{\boldsymbol{z}} \text{ is the true IV budget.}$$

The above inequality allows for some tolerance to ε-misspecified symmetry under DA so long as $\gamma_{\boldsymbol{z}}$ is positive. In practice this is often true even for perfect instruments to make the optimization problem more tractable. Importantly, practitioners can use the following observable diagnostic for Thm. 3.B:

$$\varepsilon \leq (\tilde{\sigma} - \sigma)\sqrt{\gamma_{\boldsymbol{z}}}.$$

Alternatively, if the sensitivity parameter $\gamma_{\boldsymbol{z}}$ is exposed, one can simply re-calibrate it to $\tilde{\gamma}_{\boldsymbol{z}}$ satisfying

$$\tilde{\gamma}_{\boldsymbol{z}}(\varepsilon) = \left(\varepsilon/\tilde{\sigma} + \sqrt{\gamma_{\boldsymbol{z}}/\rho}\right)^2.$$

Meaning the new budget $\tilde{\gamma}_{\boldsymbol{z}}$ should accommodate the IV leak as well as the misspecified symmetry. If one plausibly assumes symmetry violation is "orthogonal" to confounding (Rem. 2), this tightens to

$$\tilde{\gamma}_{\boldsymbol{z}}(\varepsilon) = \varepsilon^2/\tilde{\sigma}^2 + \gamma_{\boldsymbol{z}}/\rho.$$

D SHAPE CONSTRAINT VS. CHANGE OF MEASURE

In this section we aim to compare the two operationalizations of symmetry informed PI; via an invariance shape constraint in §3, and via change-of-measure through DA pre-processing in §4. Throughout, we assume no observable co-variates $Z = \varnothing$ for simplicity, and a mean-centered Y w.l.o.g. with the mean absorbed into the intercept of $h_\star$. Define the ambient $\mathcal{T}$-inv. subspace $\mathcal{I}$ as

$$\mathcal{I} \coloneqq \Big\{ h \in L_2^0(X) \cap L_2^0(\widetilde{X}) \,:\, h(X) = h(\widetilde{X}) \text{ a.s.} \Big\}, \qquad \mathcal{H}_{\text{inv}} \coloneqq \mathcal{H} \cap \mathcal{I}.$$

Notice that, by construction, all forms $\langle\cdot,\cdot\rangle_X$, $\langle\cdot,\cdot\rangle_{\widetilde{X}}$ $\langle\cdot,\cdot\rangle_{L_2(X)}$ and $\langle\cdot,\cdot\rangle_{L_2(\widetilde{X})}$ coincide on $\mathcal{I}$, $\mathcal{H}_{\text{inv}}$. Next, define the operator $\mathcal{E}_{A|B} : L_2^0(A) \to L_2^0(B)$ as conditional expectation on centered subspaces

$$(\mathcal{E}_{A|B}\, h)(\cdot) = \mathbb{E}[h(A) \,|\, B = \cdot].$$

Well defined because $\mathbb{E}[\mathbb{E}[h(A)\,|\,B]] = \mathbb{E}[h(A)] = 0$ by the tower property, and $\mathcal{E}_{A|B} h = 0$ iff $\mathbb{E}[h(A)\,|\,B] = 0$ a.s.. Of particular interest are $\mathcal{E}_{X|\widetilde{X}}$ and $\mathcal{E}_{\widetilde{X}|T}$, both of which we discuss below.

Now, under a $\mathcal{T}$-inv. $h_\star$, $\varepsilon = 0$ and the invariance shape constraint in §3 yields $\mathcal{H}_{\text{pi+inv}} = \mathcal{H}_{\text{pi}} \cap \mathcal{H}_{\text{inv}}$ (see Cor. 1). By contrast, DA pre-processing yields $\mathcal{H}_{\widetilde{\text{pi}}}$ and/or $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$. To make their comparison tractable in observables, we make two standard assumptions. First, we take $\mathcal{T}$ to be a compact group:

Assumption 4 (compact group action DA)**.** Let $\mathcal{T}$ be a compact group with action $\alpha : \mathcal{X} \times \mathcal{T} \to \mathcal{X}$. Furthermore, $\mathcal{T}$ is endowed with a (unique) normalized Haar measure p_T, i.e. p_T is uniform over $\mathcal{T}$.

Assumption 4 encompasses several standard DA transformations, including some used in our experiments such as rotations, translations, etc. Under Asm. 4, $p_{\widetilde{X}}$ itself becomes $\mathcal{T}$-invariant. Importantly, $\mathcal{E}_{X|\widetilde{X}}$ is now the Reynolds operator; an $L_2(X)$-orthogonal projector onto $\mathcal{I}$ (Rota, 1964). Under Asms. 1a–1b, the post-DA bias $\tilde{b}$ is then a projection of b onto $\mathcal{I}$ (Lem. 5). This now allows us to compare identification metrics entirely in observables, e.g. in Prop. 1,

$$\big\|e_{\tilde{b}}\big\|^2_{L_2(X)} = \frac{\|\tilde{b}\|^2_{L_2(X)}}{\|\tilde{b}\|^2_{\widetilde{X}}} = \frac{\|\tilde{b}\|^2_{\widetilde{X}}}{\|\tilde{b}\|^2_{\widetilde{X}}} = 1.$$

Second, since T-as-IV imposes Eq. (4) as a constraint, it filters functions with $\mathbb{E}[h(\widetilde{X})|T]$ a.s. constant. This holds for all $h \in \mathcal{H}_{\text{inv}}$, since $\mathbb{E}[h(\widetilde{X})|T] = \mathbb{E}[h(X)|T] = \mathbb{E}[h(X)] = 0$ a.s. following $T \perp\!\!\!\perp X$. Hence $\mathcal{H}_{\text{inv}} \subseteq \text{null}\big(\mathcal{E}_{\widetilde{X}|T}\big)$. We say that DA is *relevant*, or *complete*, when this holds with equality:

Assumption 5 (relevant / complete DA)**.** Let T be *relevant / complete*: $\text{null}\big(\mathcal{E}_{\widetilde{X}|T}\,\big|_{\mathcal{H}_{\widetilde{X}}}\big) = \mathcal{H}_{\text{inv}}$.

Assumption 5 is akin to the standard injective $\mathcal{E}_{X|Z}$ assumption for relevant/complete IVs Z (Ailer et al., 2024; Sun et al., 2025). Similarly, Asm. 5 assumes $\mathcal{E}_{\widetilde{X}|T}$ injective on $\widetilde{X}$-orthogonal complement of $\mathcal{H}_{\text{inv}}$ within $\mathcal{H}_{\widetilde{X}}$. Meaning, the only directions whose data-averaged response is zero across augmentations are the invariant ones.

The inv. shape constraint and change-of-measure routes are now comparable under Asms. 4 and 5:

Proposition 3 (shape constraint vs. change-of-measure)**.** *Under Asms. 1a–1b, and Asms. 2 to 5 with $\mathcal{T}$-invariant $h_\star$ and $Z = \varnothing$. Set $\varepsilon = \gamma_{\boldsymbol{z}} = 0$, and suppose the baseline PI set is valid $h_\star \in \mathcal{H}_{\text{pi}}$. Then, $h_{\widetilde{\text{erm}}} = \Pi h_{\text{erm}} \in \mathcal{H}_{\text{inv}}$, where Π is the $L_2(X)$-orthogonal projection onto $\mathcal{I}$ (cf. §G). Further,*

$$\mathcal{H}_{\widetilde{\text{pi}}} = \Big\{ h \in \mathcal{H}_{\widetilde{X}} \,:\, \big\|h - h_{\widetilde{\text{erm}}}\big\|^2_{\widetilde{X}} \le \tilde{\sigma}^2 \gamma \Big\},$$

$$\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}} = \Big\{ h \in \mathcal{H}_{\text{inv}} \,:\, \big\|h - h_{\widetilde{\text{erm}}}\big\|^2_{\widetilde{X}} \le \tilde{\sigma}^2 \gamma \Big\},$$

$$\mathcal{H}_{\text{pi+inv}} = \Big\{ h \in \mathcal{H}_{\text{inv}} \,:\, \big\|h - h_{\widetilde{\text{erm}}}\big\|^2_{X} \le \sigma^2 (1 + \gamma - \rho) \Big\}.$$

All identified sets are balls centered about $h_{\widetilde{\text{erm}}}$. Both $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$ and $\mathcal{H}_{\text{pi+inv}}$ live on the slice $\mathcal{H}_{\text{inv}}$, the latter with a smaller radius $\sigma\sqrt{1+\gamma-\rho} \le \tilde{\sigma}\sqrt{\gamma}$ with equality iff $\rho = 1$. It trivially follows that,

$$h_\star \in \mathcal{H}_{\text{pi+inv}} \subseteq \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}} \subseteq \mathcal{H}_{\widetilde{\text{pi}}},$$

and the same dominance hierarchy holds for sharpness and worst-case errors over $\mathcal{H}_\bullet$ and $\mathcal{H}_\bullet(\boldsymbol{x})$.

Under Asm. 1c, closed form volume, width and avg. errors for the balls in Prop. 3 can now be derived in observables similar to Prop. 2 and Thm. 2. We leave it as an exercise for interested readers.

E Implementation Details

Partial identification. Write the finite-n sample version of the abstract PI program from §2.3 as:

$$\begin{aligned} &\min_q \,/\, \max_q \; h^q_\star(\boldsymbol{x}) && \text{// interval lower / upper bounds} \\ &\quad \text{s.t.} \quad D_{\mathcal{F}}(q_{XYZ},\, \widehat{p}_{XYZ}) \le \gamma_n, && \text{// observational consistency} \\ &\qquad\quad q \in \mathcal{P}_{\text{PI}}, \quad h^q_\star \in \mathcal{H}, \quad \widehat{E}_{\text{INV}}(h^q_\star) \le \varepsilon^2. && \text{// structural constraints} \end{aligned} \tag{P1}$$

Here, $\widehat{p}_{XYZ}$ is the empirical distribution of n observations. The IPM $D_{\mathcal{F}}$ matches q_{XYZ} to $\widehat{p}_{XYZ}$ within tolerance $\gamma_n \ge 0$ to account for sampling error. Further, $\widehat{E}_{\text{INV}}$ is a Monte Carlo approx. of E_{INV}:

$$\widehat{E}_{\text{INV}}(h) := \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \big| h(\boldsymbol{x}_i) - h(\boldsymbol{\tau}_{i,j}\, \boldsymbol{x}_i) \big|^2, \qquad (\boldsymbol{x}_i, \cdot) \in \mathcal{D}, \quad \boldsymbol{\tau}_{i,j} \sim p_T.$$

In practice, the partial-r^2 model (Asm. 2) directly imposes the Lem. 2 ball via $\widehat{R}_{\text{ERM}}(h) \le \widehat{\sigma}^2(1+\gamma)$ as

$$\widehat{R}_{\text{ERM}}(h) = \widehat{\sigma}^2 + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \big(h(\boldsymbol{x}_i) - \widehat{h}_{\text{ERM}}(\boldsymbol{x}_i) \big)^2}_{\le \widehat{\sigma}^2 \gamma \text{ by Lem. 2}}, \qquad \widehat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \widehat{h}_{\text{ERM}}(\boldsymbol{x}_i) \big)^2.$$

For leaky-IV, we implement a generalized method of moments (GMM) relaxation of the conditional moment in Asm. 3 (Akbar et al., 2025; Ailer et al., 2024). For IV test functions $\phi_{\boldsymbol{z}} : \mathcal{Z} \to \mathbb{R}^{d_{\phi_{\boldsymbol{z}}}}$, define

$$\widehat{R}_{\text{IV}}(h) := \left\| \frac{1}{n} \sum_{i=1}^{n} \big(y_i - h(\boldsymbol{x}_i) \big) \cdot \phi_{\boldsymbol{z}}(\boldsymbol{z}_i) \right\|^2_{\boldsymbol{W}}, \qquad \|\boldsymbol{v}\|^2_{\boldsymbol{W}} := \boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{v}.$$

We use $\phi_{\boldsymbol{z}}(\boldsymbol{z}) = \boldsymbol{z}$ and $\boldsymbol{W} = \widehat{\boldsymbol{\Sigma}}^{\dagger}_{\boldsymbol{z}}$ (Bennett et al., 2019), where $\dagger$ denotes the Moore–Penrose pseudoinverse. For point-estimation $\widehat{h}_{\text{IV}} := \operatorname{argmin}_{h \in \mathcal{H}} \widehat{R}_{\text{IV}}(h)$, whereas for PI, impose $\widehat{R}_{\text{IV}}(h) \le \widehat{\sigma}^2 \gamma_{\boldsymbol{z}}$.

We instantiate the hypothesis class by a parameterized family $\mathcal{H}_{\boldsymbol{\theta}} := \{\boldsymbol{\theta}^\top \phi(\cdot) : \boldsymbol{\theta} \in \mathcal{O}\}$ with some basis features $\phi : \mathcal{X} \to \mathbb{R}^k$, e.g. polynomial features for optical device, log transform in cigarette demand, and convolutional neural network in do-MNIST Kilbertus et al. (2020); Padh et al. (2023); Zheng et al. (2025). Program (P1) under Asms. 2 and 3 now reduces to:

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \mathcal{O}} \,/\, \max_{\boldsymbol{\theta} \in \mathcal{O}} \; h_{\boldsymbol{\theta}}(\boldsymbol{x}) \\ &\quad \text{s.t.} \quad \widehat{R}_{\text{ERM}}(h_{\boldsymbol{\theta}}) \le \widehat{\sigma}^2(1+\gamma), && \text{// bounded confounding (Lem. 2)} \\ &\qquad\quad \widehat{R}_{\text{IV}}(h_{\boldsymbol{\theta}}) \le \widehat{\sigma}^2 \gamma_{\boldsymbol{z}}, && \text{// leaky IV constraints (Asm. 3)} \\ &\qquad\quad \widehat{E}_{\text{INV}}(h_{\boldsymbol{\theta}}) \le \varepsilon^2, && \text{// invariance constraint (§3)} \\ &\qquad\quad \sum_i h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sum_i y_i, && \text{// center / mean-match} \end{aligned} \tag{P2}$$

where γ_n is folded into bootstrapped 95% *Imbens & Manski confidence intervals* (IM-CI) (2004) on the resulting intervals. Finally, Table 3 lists how programs (P1) or (P2) are invoked under different approaches of symmetry-informed PI. The simulation, optical device and cigarette demand experiments solve (P2) via QCQP, while do-MNIST uses SLSQP.

Change of measure. Take m samples $\boldsymbol{\tau}_{i,j} \sim p_T$ for each $i \in [n]$ with $n = |\mathcal{D}|$. Construct dataset $\widetilde{\mathcal{D}} := \{(\tilde{\boldsymbol{x}}_{i,j},\, y_i,\, \tilde{\boldsymbol{z}}_{i,j})\}_{i \in [n], j \in [m]}$ with $\tilde{\boldsymbol{x}}_{i,j} := \boldsymbol{\tau}_{i,j} \boldsymbol{x}_i$ and $\tilde{\boldsymbol{z}}_{i,j} := (\boldsymbol{z}_i, \boldsymbol{\tau}_{i,j})$.

Re-calibration. Post-DA, the budgets are set as per the diluted confounding (Lem. 1) and symmetry misspecification (Thm. 3.B) and swept in §F

$$\tilde{\gamma} = \gamma / \widehat{\rho}, \qquad\qquad \tilde{\gamma}_{\boldsymbol{z}}(\varepsilon) = \varepsilon^2 / \tilde{\sigma}^2 + \gamma_{\boldsymbol{z}} / \rho.$$

Table 3: Every object in the paper as one setting of the same program (P2); labels are those used in the legends of Figs. 5 and 6.

Dataset	Instruments	Point estimate	Sensitivity params. γ	$\gamma_{\boldsymbol{z}}$	ε	PI interval	Sensitivity params. γ	$\gamma_{\boldsymbol{z}}$	ε	Post-proc.
$\mathcal{D}$	Z	IV	∞	0	∞	PI + INV + IV	γ	$\gamma_{\boldsymbol{z}}$	ε	
$\mathcal{D}$	Z	ERM + IV	$\gamma_{\min}$	0	∞	PI + IV	γ	$\gamma_{\boldsymbol{z}}$	∞	
$\widetilde{\mathcal{D}}$	Z	ERM + IV	$\tilde{\gamma}_{\min}$	0	∞	PI + IV	$\tilde{\gamma}$	$\tilde{\gamma}_{\boldsymbol{z}}$	∞	$\pm\varepsilon$
$\widetilde{\mathcal{D}}$	$\widetilde{Z} := (Z, T)$	ERM + IV	$\tilde{\gamma}_{\min}$	0	∞	PI + IV	$\tilde{\gamma}$	$\tilde{\gamma}_{\boldsymbol{z}}$	∞	$\pm\varepsilon$

F Experiment Supplement

F.1 Experiment Setup

In addition to the three experiments described in §5, we also conduct a simple simulation experiment described below.

Simulation experiment. Fix $\boldsymbol{\theta}_\star \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and sample a $\boldsymbol{\Sigma}_{\boldsymbol{x}u}$ uniformly on a unit sphere. Generate

$$h_\star(X) := \boldsymbol{\theta}_\star^\top X, \qquad (X, U) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I} & \boldsymbol{\Sigma}_{\boldsymbol{x}u} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}u}^\top & 1 \end{bmatrix}\right), \qquad \xi \sim \mathcal{N}(0, 0.1^2),$$

and Y from Eq. (1). Next, we follow Akbar et al. (2025) to construct T such that $h_\star$ respects $\mathcal{T}$-inv. By selecting $d_{\boldsymbol{\tau}}$ basis of $\mathrm{null}(\boldsymbol{\theta}_\star)$, each with probability 0.8, as rows of $\boldsymbol{A} \in \mathbb{R}^{d_{\boldsymbol{\tau}} \times d_{\boldsymbol{x}}}$, we define:

$$\widetilde{X} = \alpha(X, T) = X + \boldsymbol{A}^\top T, \qquad T \sim \mathcal{N}(\mathbf{0}_{d_{\boldsymbol{\tau}}}, \boldsymbol{I}_{d_{\boldsymbol{\tau}}}).$$

By construction, $\widetilde{X}$ varies only in the null-space of $\boldsymbol{\theta}_\star$, ensuring $h_\star$ is $\mathcal{T}$-invariant. While this construction utilizes the true $\boldsymbol{\theta}_\star$, we treat $\boldsymbol{A}$ as illustrative of domain knowledge defining plausible data symmetries; crucially, this knowledge alone is insufficient to identify $\boldsymbol{\theta}_\star$ without observing U.

Taking dimensions $d_{\boldsymbol{x}} = 32$, $d_{\boldsymbol{\tau}} = 31$, we generate $n = 2048$ samples for (X, Y, T). The ERM baseline is the closed-form OLS solution. And for PI we use the partial-r^2 model (Asm. 2) with the effective confounding budget $\gamma = 1$, and invariance-error and IV validity budgets of $\varepsilon = \gamma_{\boldsymbol{z}} = 2^{-8}$.

Optical device. The benchmark dataset by Janzing & Schölkopf (2018) consists of 12 datasets of 3×3 pixel images X displayed on a laptop screen which generate voltage readings Y across a photo-diode. The system involves a physically instantiated latent confounder U controlling the intensity of two LEDs; the first affects the webcam capturing X, while the second influences the photo-diode measuring Y. We estimate the true target $h_\star(\boldsymbol{x}) = \boldsymbol{\theta}_\star^\top \phi(\boldsymbol{x})$ via back-door adjustment using joint features $(\phi(X), U)$, where $\phi(X)$ denotes polynomial features of X. We select the polynomial degree $\deg(h_\star) \in \{1, \cdots, 5\}$ that best explains the data ($\deg(h_\star) = 2$ for Fig. 5 on dataset #8). Our choice of DA on X composes random vertical/horizontal flips and $90°$ rotations. For a dataset of $n = 1000$ samples, we compute features $\phi(\widetilde{X})$ to be used with PI, setting effective budget $\gamma = 2^{-2}$ for the partial-r^2 model, and $\varepsilon = 2^{-1.5}, \gamma_{\boldsymbol{z}} = 2^{-6}$ for the inv. error and IV budgets respectively. Fig. 5 shows that symmetry information does improve PI.

In the ablations of Section F.2, all metrics are aggregated across all 12 datasets of optical device.

US Cigarette demand. Prices, sales and taxes are from the CDC's *Tax Burden on Tobacco*, state income from the BEA, and the CPI from FRED, for the 48 contiguous states and DC ($n = 2450$ state-years). The neighbor price is the lowest retail price among a state's land neighbors (Baltagi & Levin, 1986).

Ablation sweeps. Of course the true target is not available for this dataset, and Fig. 6 only illustrates sharpening effects without testing validity. We therefore generate a synthetic target $h_\star$ for ablations in Section F.2 via a two-stage least-squares fit with instruments Z subject to $\mathbf{1}^\top \boldsymbol{\theta} = 0$, thus satisfying both assumptions exactly. We still use $\varepsilon = 2^{-8}$ as a numerical tolerance. DA translates X along $\mathbf{1}$ with amplitude $2\,\mathrm{sd}(\mathbf{1}^\top X/2)$.

Fig. 6 sharpening. In Fig. 6, the γ sweep fixes $\gamma_{\boldsymbol{z}} = 0.0177$, and it should be noted that the corresponding result is illustrative of the sharpening effect of symmetry, validity contingent on budgets γ, γ_z and assumptions on endogeneity/endogeneity of X-components holding true. These of course cannot be validation from data, and rest of domain knowledge.

do-MNIST. With MNIST image D, we color it with C and generate labels Y as follows

$$U \sim \mathrm{Bern}(\tfrac{1}{2}), \quad C \coloneqq U \oplus \mathrm{Bern}(\eta), \quad \xi \sim \mathrm{Bern}(\beta),$$
$$X \coloneqq \mathrm{color}(D, C), \quad Y \coloneqq (1-\xi) \cdot \mathbf{1}[\mathrm{digit}(D) \geq 5] + \xi \cdot U,$$

with $\eta = 0.25$, $\beta = 0.4$, and $\mathrm{color}(\cdot, c)$ tinting the 14×14 digit red ($c = 1$) or blue ($c = 0$) with small jitter. The colour noise η keeps U non-deterministic given X. Then $h_\star(\boldsymbol{x}) = 0.2 + 0.6 \cdot \mathbf{1}[\mathrm{digit}(\boldsymbol{x}) \geq 5]$, whereas $\mathbb{E}[Y \mid X] \in \{0.1, 0.3, 0.7, 0.9\}$ leaks colour. ERM networks (same architecture as Arjovsky et al. (2019)) train on 1.2M draws from 40k images; CopSens (Zheng et al., 2025) (latent dimension 32) is fit on 60k draws from 10k disjoint images. A single $\gamma = 0.062$ is used for all methods, the smallest at which pi covers $h_\star$ on 99.5% of a third disjoint split. DA composes $\pm 20\%$ translations, $\pm 10^\circ$ rotations, and contrast, saturation and hue jitter; the symmetry is exact, and $\varepsilon = 0.04$ is finite-sample slack. 5% of augmented rows are left untransformed (Asm. 1b).

Infeasibility of shape-constraints. The pi + mv comes out infeasible for any reasonably small value of ε. The reason being that CopSens fixes all but the last layer of the ERM network. This is a standard approach in modern neural PI Padh et al. (2023); Ailer et al. (2024) to keep bounds non-vacuous. Of course one can always make the solver more flexible by unfreezing several penultimate layers at the expense of vacuity and compute. This substantiates our operational concerns about the shape constrant approach even further, making the case stronger for the change-of-measure approach.

F.2 Ablations

Sweeps report coverage of $h_\star$, interval width and worst-case error, over 8 seeds, against the budget ratio to its oracle value (e.g., $\gamma/\gamma^\star$).

Validity

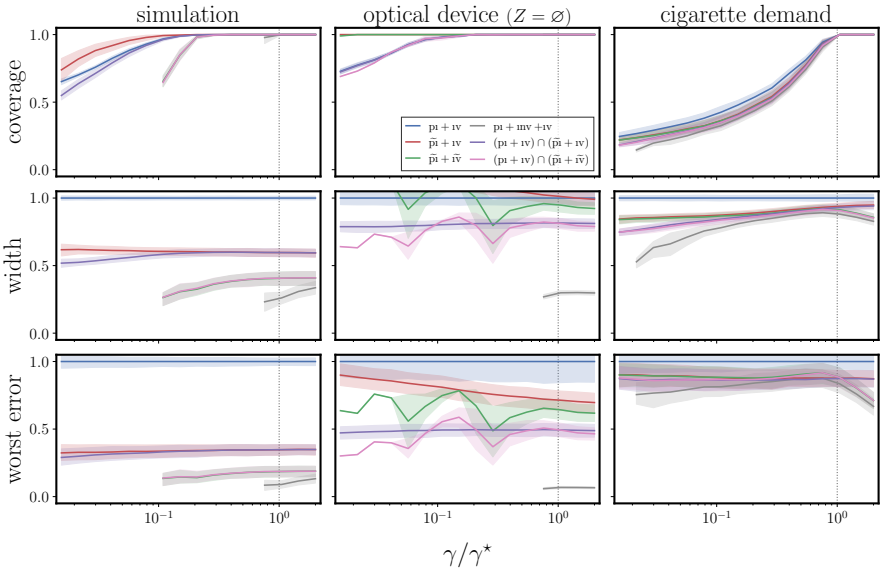


Figure 8: **Validity.** Coverage, interval width, and worst-case error as the confounding budget γ is varied across the simulation, optical device, and cigarette demand experiments.

Sharpness / informativeness

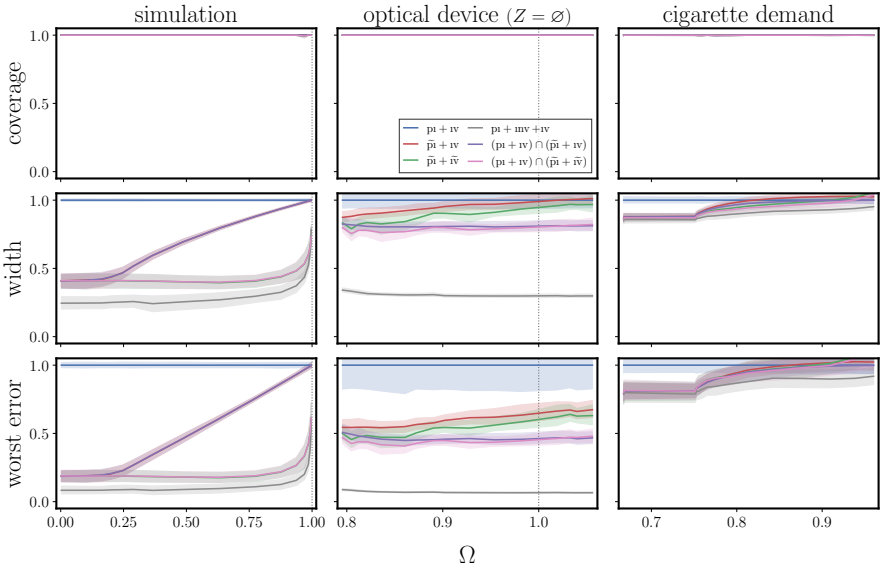


Figure 9: **Sharpness / informativeness.** Coverage, interval width, and worst-case error as the sharpening quantity Ω is varied across the simulation, optical device, and cigarette demand experiments.

ROBUSTNESS

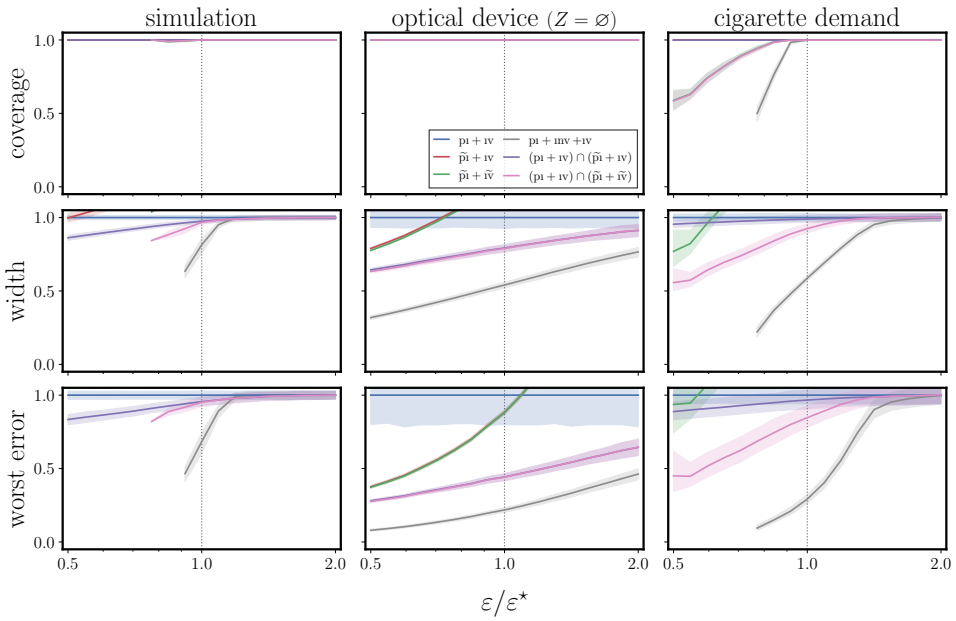


Figure 10: **Robustness.** Coverage, interval width, and worst-case error under symmetry misspecification $\varepsilon^\star$, with assumed misspecification ε varied across the simulation, optical device, and cigarette demand experiments.

RE-CALIBRATION

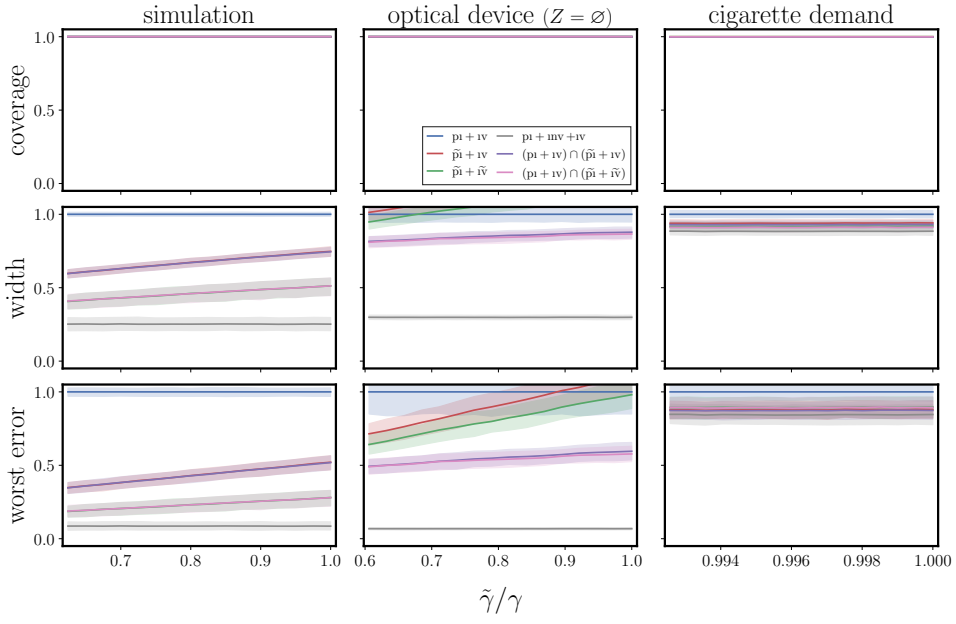


Figure 11: **Re-calibration.** Coverage, interval width, and worst-case error as the adjusted sensitivity budget $\widetilde{\gamma}$ is varied relative to the baseline budget γ across the simulation, optical device, and cigarette demand experiments.

EFFICIENCY

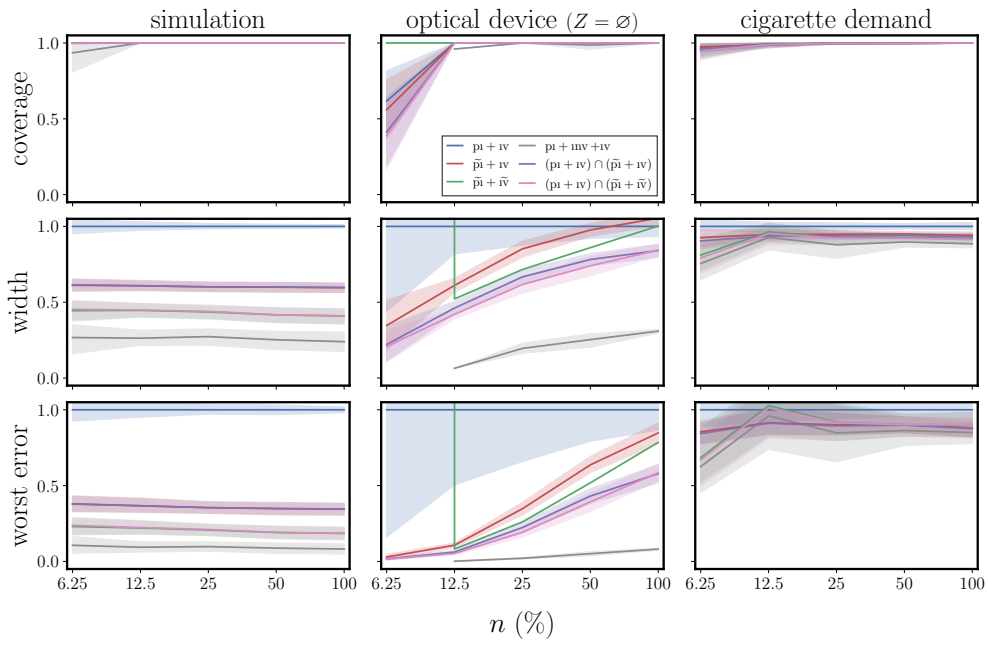


Figure 12: **Statistical efficiency in n.** Coverage, interval width, and worst-case error as the available sample size n is varied across the simulation, optical device, and cigarette demand experiments.

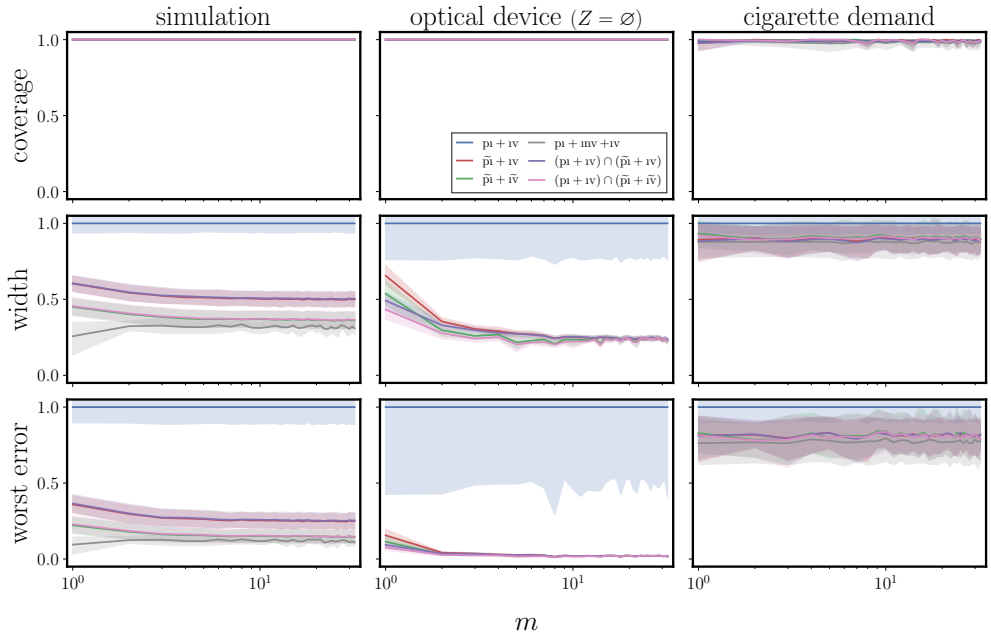


Figure 13: **Statistical efficiency in m.** Coverage, interval width, and worst-case error as the number of augmentation folds m is varied across the simulation, optical device, and cigarette demand experiments.

G PROOFS

TECHNICAL PRELIMINARIES

Finite dimension. For $V \in \{X, \widetilde{X}\}$, define the centering operator $\mathcal{C}_V \coloneqq h - \mathbb{E}[h(V)]$. Now, since $\mathcal{H} + \mathbb{R}1 = \mathcal{H}$ (Asm. 1), therefore $\mathcal{C}_V h \in \mathcal{H}_V$. These maps define a bijection between $\mathcal{H}_X$ and $\mathcal{H}_{\widetilde{X}}$,

$$\mathcal{H}_X \underset{\mathcal{C}_X}{\overset{\mathcal{C}_{\widetilde{X}}}{\rightleftarrows}} \mathcal{H}_{\widetilde{X}}$$

therefore, $\dim(\mathcal{H}_{\widetilde{X}}) = \dim(\mathcal{H}_X) = k$ (Asm. 1c).

Non-degeneracy. $p_X \ll p_{\widetilde{X}}$ (Asm. 1b) transfers Hilbert structure (Asm. 1c) to $\mathcal{H}_{\widetilde{X}}$. For $h \in \mathcal{H}_{\widetilde{X}}$,

$$\begin{aligned} \|h\|_{\widetilde{X}} = 0 \quad &\Longrightarrow \quad h(\widetilde{X}) = 0 \text{ a.s.} && \text{(0 is the only a.s. constant } h \in \mathcal{H}_{\widetilde{X}}.) \\ &\overset{\text{1b}}{\Longrightarrow} \quad h(X) = 0 \text{ a.s.} \quad \Longrightarrow \quad \|h\|_X = 0 \quad \overset{\text{1c}}{\Longrightarrow} \quad h \in \mathbb{R}1 \quad \overset{h \in \mathcal{H}_{\widetilde{X}}}{\Longrightarrow} \quad h = 0. \end{aligned}$$

Hence both $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_{\widetilde{X}}$ are positive-definite Hilbert norms on $\mathcal{H}_{\widetilde{X}}$, but so is $\langle \cdot, \cdot \rangle_{L_2(X)}$ because

$$\|h\|^2_{L_2(X)} = \|h\|^2_X + \mathbb{E}[h(X)]^2 \geq \|h\|^2_X.$$

Shift operator + decomposition. We give a basis-free form of the var.-/mean-shift identity in §4.

Under Asms. 1b–1c, Riesz gives a unique, positive-definite, self-adjoint *shift operator* $\mathcal{S} : \mathcal{H}_{\widetilde{X}} \to \mathcal{H}_{\widetilde{X}}$ that pulls the evaluation $L_2(X)$-inner product back into the solver's native $(\mathcal{H}_{\widetilde{X}}, \langle \cdot, \cdot \rangle_{\widetilde{X}})$-geometry,

$$\langle f, \mathcal{S} g \rangle_{\widetilde{X}} = \langle f, g \rangle_{L_2(X)}, \quad \forall f, g \in \mathcal{H}_{\widetilde{X}}.$$

$\mathcal{S}$ captures the fit vs. evaluation gap. And like $L_2(X)$-norm, its pull-back $\mathcal{S}$ also admits a shape / level split into operator $\mathcal{K} \colon \mathcal{H}_{\widetilde{X}} \to \mathcal{H}_{\widetilde{X}}$ and linear functional $\delta \colon \mathcal{H}_{\widetilde{X}} \to \mathbb{R}$ defined respectively as

$$\langle f, \mathcal{K} g \rangle_{\widetilde{X}} = \langle f, g \rangle_X, \qquad \delta h = \mathbb{E}[h(X)] - \mathbb{E}[h(\widetilde{X})], \qquad \forall f, g, h \in \mathcal{H}_{\widetilde{X}}.$$

Lemma 3 (shift decomposition)**.** *Under Asms. 1b–1c, $\mathcal{S} = \mathcal{K} + \delta^* \delta$, with $\mathcal{K} \succ 0$, $\delta^* \delta \succcurlyeq 0$ s.t. $\operatorname{rank}(\delta^* \delta) \leq 1$, and $\operatorname{tr}(\delta^* \delta) = D^2_{\mathcal{B}}(p_X, p_{\widetilde{X}})$ is MMD^2 over the unit ball $\mathcal{B} \coloneqq \{h \in \mathcal{H}_{\widetilde{X}} : \|h\|_{\widetilde{X}} \leq 1\}$.*

The *variance-shift operator* $\mathcal{K}$ governs how the estimation geometry changes between $(\mathcal{H}_X, \langle \cdot, \cdot \rangle_X)$ and $(\mathcal{H}_{\widetilde{X}}, \langle \cdot, \cdot \rangle_{\widetilde{X}})$. It differs from $\mathcal{S}$ precisely by the rank-1 *mean-shift operator* $\delta^* \delta$ with trace $D^2_{\mathcal{B}}$.

Invariant slice. Recall from §D,

$$\mathcal{I} \coloneqq \left\{ f \in L^0_2(X) \cap L^0_2(\widetilde{X}) \;\; : \;\; f(X) = f(\widetilde{X}) \text{ a.s.} \right\}.$$

Here $\mathcal{I}$ is a closed linear subspace, irrespective of any linearity assumed of $\mathcal{H}$. Now define

$$\mathcal{H}_{\text{inv}} \coloneqq \mathcal{H} \cap \mathcal{I}.$$

By construction, all four forms coincide on $\mathcal{I}$ and $\mathcal{H}_{\text{inv}}$,

$$\langle f, g \rangle_{L_2(X)} \;=\; \langle f, g \rangle_{L_2(\widetilde{X})} \;=\; \langle f, g \rangle_X \;=\; \langle f, g \rangle_{\widetilde{X}}, \qquad \forall\, f, g \in \mathcal{I}. \tag{6}$$

Covariance forms agree with L_2 as $\mathcal{I}$ is both X- and $\widetilde{X}$-centered. And X-, $\widetilde{X}$-forms agree by $\mathcal{T}$-inv.

Throughout, we denote $\mathcal{T}$ as the $L_2(X)$-orthogonal (equivalently, X-orthogonal) projection onto $\mathcal{I}$.

G.1 Proof of Thm. 1—Valid bounds with DA

Theorem 1 (valid bounds with DA)**.** *Let $h_\star$ be $\mathcal{T}$-inv. under Asm. 1a and 2. Then, the post-DA set $\mathcal{H}_{\widetilde{\text{PI}}}$ is exactly valid whenever the baseline set $\mathcal{H}_{\text{PI}}$ is ϵ-valid with $\epsilon \le \sigma^2\big(\sqrt{\rho-1+\gamma\rho}-\sqrt{\gamma}\big)^2$.*

Proof. We proceed as: (i) express baseline approx. error as point-to-ball distance ($*$), (ii) certify post-DA membership ($\dagger$) by comparing bias magnitudes ($\ddagger$) using Lem. 1.

First, write the Lem. 2 pre- and post-DA partially identified sets in their respective L_2 geometries as:

$$\mathcal{H}_{\text{PI}} = \Big\{h \in \mathcal{H}_X : \|h - h_{\text{erm}}\|^2_{L_2(X)} \le \gamma\sigma^2\Big\},$$
$$\mathcal{H}_{\widetilde{\text{PI}}} = \Big\{h \in \mathcal{H}_{\tilde{X}} : \|h - h_{\widetilde{\text{erm}}}\|^2_{L_2(\tilde{X})} \le \gamma\tilde{\sigma}^2\Big\}.$$

Note that target $h_\star$ lies in both centered classes: $\mathbb{E}[h_\star(X)] = \mathbb{E}[Y] = 0 \implies h_\star \in \mathcal{H}_X$, and by $\mathcal{T}$-inv., $\mathbb{E}[h_\star(\tilde{X})] = \mathbb{E}[h_\star(X)] = 0 \implies h_\star \in \mathcal{H}_{\tilde{X}}$. Define the pre- and post-DA bias magnitudes:

$$A \coloneqq \|h_{\text{erm}} - h_\star\|_{L_2(X)}, \qquad B \coloneqq \|h_{\widetilde{\text{erm}}} - h_\star\|_{L_2(\tilde{X})}.$$

For an $L_2(X)$-round ball, approx. error $E^-(\bullet) = \min_{h\in\bullet} \big\|h - h_\star\big\|^2_{L_2(X)}$ reduces to point-to-ball distance. Since $\mathcal{H}_{\text{PI}}$ is essentially $\mathcal{H}_X \cap L_2(X)$-ball, approx. error is lower-bounded by point-to-ball distance,

$$E^-(\mathcal{H}_{\text{PI}}) \ge \big(A - \sigma\sqrt{\gamma}\big)^2_+, \qquad h_\star \in \mathcal{H}_{\text{PI}} \iff A \le \sigma\sqrt{\gamma}. \tag{$*$}$$

As for $\mathcal{H}_{\widetilde{\text{PI}}}$, the following validity criterion gives $E^-(\mathcal{H}_{\widetilde{\text{PI}}}) = 0$ exactly,

$$h_\star \in \mathcal{H}_{\widetilde{\text{PI}}} \iff B \le \tilde{\sigma}\sqrt{\gamma}. \tag{$\dagger$}$$

For $C^2 \coloneqq \tilde{\sigma}^2 - \sigma^2$, Lem. 1 (Asm. 1a, $\mathcal{T}$-inv. $h_\star$) supplies the following two identities with $C^2 \ge 0$,

$$A^2 = B^2 + C^2, \qquad \tilde{\sigma}^2 = \sigma^2 + C^2. \tag{$\ddagger$}$$

Now suppose $\mathcal{H}_{\text{PI}}$ is ϵ-valid with

$$\epsilon \;\le\; \epsilon^+ \;\coloneqq\; \Big(\sqrt{C^2 + \gamma\tilde{\sigma}^2} - \sigma\sqrt{\gamma}\Big)^2 \;=\; \sigma^2\Big(\sqrt{\rho - 1 + \gamma\rho} - \sqrt{\gamma}\Big)^2,$$

the second equality substituting $C^2 + \gamma\tilde{\sigma}^2 = \sigma^2(\rho - 1 + \gamma\rho)$ from Eq. ($\ddagger$). Note $\epsilon^+ > 0$ iff $\rho > 1$, and

$$\sqrt{\epsilon^+} = \sqrt{C^2 + \gamma\tilde{\sigma}^2} - \sigma\sqrt{\gamma} \ge 0,$$

since the difference defining ϵ^+ is non-negative $C^2 + \gamma\tilde{\sigma}^2 - \gamma\sigma^2 = (1+\gamma)C^2 \ge 0$. From Eq. ($*$),

$$\begin{aligned}
&(A - \sigma\sqrt{\gamma})_+ \le \sqrt{\epsilon}, && (\epsilon\text{-validity.})\\
\implies\quad & A - \sigma\sqrt{\gamma} \le \sqrt{\epsilon^+}, && (x \le (x)_+ \text{ and } \sqrt{\epsilon} \le \sqrt{\epsilon^+}.)\\
\implies\quad & A - \sigma\sqrt{\gamma} \le \sqrt{C^2 + \gamma\tilde{\sigma}^2} - \sigma\sqrt{\gamma}, && (\text{Def. of } \epsilon^+, \text{ and } \sqrt{\epsilon^+} \ge 0.)\\
\implies\quad & A^2 \le C^2 + \gamma\tilde{\sigma}^2. && (\text{Add } \sigma\sqrt{\gamma}, \text{ then square.})\\
\implies\quad & B^2 = A^2 - C^2 \le \gamma\tilde{\sigma}^2. && (\text{From Eq. } (\ddagger).)
\end{aligned}$$

Because $B \le \tilde{\sigma}\sqrt{\gamma}$, thus by Eq. ($\dagger$) $h_\star \in \mathcal{H}_{\widetilde{\text{PI}}}$, and therefore

$$E^-(\mathcal{H}_{\widetilde{\text{PI}}}) = 0 \le E^-(\mathcal{H}_{\text{PI}}).$$

Finally, equality holds precisely when $E^-(\mathcal{H}_{\text{PI}}) = 0$, which by Eq. ($*$) is equivalent to $h_\star \in \mathcal{H}_{\text{PI}}$. □

G.2 Proof of Thm. 2—Informative bounds with DA

Theorem 2 (informative bounds with DA). *Under Asms. 1b–1c and 2, the following holds for $d = k$,*

$$\overline{E}(\mathcal{H}_{\text{PI}}) - \overline{E}(\mathcal{H}_{\widetilde{\text{PI}}}) \quad = \quad \underbrace{E(h_{\text{erm}}) - E(h_{\widetilde{\text{erm}}})}_{\text{centroid drift (Prop. 1)}} + \frac{\sigma^2\gamma k}{d+2} \cdot \underbrace{(1-\Omega)}_{\text{set size (Prop. 2)}} \quad =: \quad f(d),$$

And interval improvement averaged over queries: $\mathbb{E}_{X\sim p_X}\left[\overline{E}_X(\mathcal{H}_{\text{PI}}(X)) - \overline{E}_X(\mathcal{H}_{\widetilde{\text{PI}}}(X))\right] = f(1)$.

Proof. Gist of the proof: (i) reduce avg. error to second moment of a uniform law over a centered-ball, (ii) connect it to post-DA ball via operator $\mathcal{S}$, and (iii) repeat for intervals via representers and Lem. 6.

Setup. Write the Lem. 2 and Cor. 3 pre- and post-DA PI sets in their respective L_2 geometries:

$$\mathcal{H}_{\text{PI}} = \left\{h \in \mathcal{H}_X : \|h - h_{\text{erm}}\|^2_{L_2(X)} \le \sigma^2\gamma\right\}, \qquad \mathcal{H}_{\text{PI}}(\boldsymbol{x}) = \{h_{\text{erm}}(\boldsymbol{x})\} \pm \sigma\sqrt{\gamma}\|g_{\boldsymbol{x}}\|_{L_2(X)},$$
$$\mathcal{H}_{\widetilde{\text{PI}}} = \left\{h \in \mathcal{H}_{\widetilde{X}} : \|h - h_{\widetilde{\text{erm}}}\|^2_{L_2(\widetilde{X})} \le \tilde{\sigma}^2\gamma\right\}, \qquad \mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x}) = \{h_{\widetilde{\text{erm}}}(\boldsymbol{x})\} \pm \tilde{\sigma}\sqrt{\gamma}\|\tilde{g}_{\boldsymbol{x}}\|_{L_2(\widetilde{X})},$$

with $L_2(X)$-representers $g_{\boldsymbol{x}}, \tilde{g}_{\boldsymbol{x}}$ well defined on Hilbert $\mathcal{H}_X$ (Asm. 1c) and $\mathcal{H}_{\widetilde{X}}$ (Asm. 1b, cf. §G). Now, suppose u is uniform on a centered d-dim. ball of radius r w.r.t. *any* inner product $\langle\cdot,\cdot\rangle$, then

$$\mathbb{E}[u] = 0, \qquad\qquad \mathbb{E}[\langle u, \mathcal{A}u\rangle] = \frac{r^2}{d+2}\operatorname{tr}(\mathcal{A}),$$

for any self-adjoint operator $\mathcal{A}$ in $\langle\cdot,\cdot\rangle$. We show that avg. errors reduce to these identities up to offset.

Set version ($d = k$). Recall $\overline{E}(\cdot) = \mathbb{E}_{h\sim\text{Unif}(\cdot)}\left[\|h - h_\star\|^2_{L_2(X)}\right]$. Take the re-parameterizations

$$h = h_{\text{erm}} + u, \qquad u \sim \text{Unif}\{u \in \mathcal{H}_X : \|u\|_{L_2(X)} \le \sigma\sqrt{\gamma}\},$$
$$h = h_{\widetilde{\text{erm}}} + \tilde{u}, \qquad \tilde{u} \sim \text{Unif}\{u \in \mathcal{H}_{\widetilde{X}} : \|u\|_{L_2(\widetilde{X})} \le \tilde{\sigma}\sqrt{\gamma}\},$$

for $h \sim \text{Unif}(\mathcal{H}_{\text{PI}})$ and $h \sim \text{Unif}(\mathcal{H}_{\widetilde{\text{PI}}})$ respectively. Substituting back into avg. error $\overline{E}(\mathcal{H}_{\text{PI}})$ def.:

$$\begin{aligned}\overline{E}(\mathcal{H}_{\text{PI}}) &= \mathbb{E}_u\left[\|h_{\text{erm}} - h_\star + u\|^2_{L_2(X)}\right],\\ &= \|h_{\text{erm}} - h_\star\|^2_{L_2(X)} + \mathbb{E}_u\left[\|u\|^2_{L_2(X)}\right] = E(h_{\text{erm}}) + \frac{\sigma^2\gamma k}{k+2}, \qquad (\operatorname{tr}(\text{Id}) = k.)\end{aligned}$$

where $\mathbb{E}[u] = 0$ kills the cross-term. Likewise, for post-DA avg. error it follows from the def. of $\mathcal{S}$:

$$\begin{aligned}\overline{E}(\mathcal{H}_{\widetilde{\text{PI}}}) &= \mathbb{E}_{\tilde{u}}\left[\|h_{\widetilde{\text{erm}}} - h_\star + \tilde{u}\|^2_{L_2(X)}\right],\\ &= \|h_{\widetilde{\text{erm}}} - h_\star\|^2_{L_2(X)} + \mathbb{E}_{\tilde{u}}\left[\langle\tilde{u}, \mathcal{S}\tilde{u}\rangle_{L_2(\widetilde{X})}\right], \quad (\|h\|^2_{L_2(X)} = \langle h, \mathcal{S}h\rangle_{L_2(\widetilde{X})} \text{ for } h \in \mathcal{H}_{\widetilde{X}}.)\\ &= E(h_{\widetilde{\text{erm}}}) + \frac{\tilde{\sigma}^2\gamma}{k+2}\operatorname{tr}(\mathcal{S}).\end{aligned}$$

Again, cross-terms vanish because $\mathbb{E}[\tilde{u}] = 0$. Subtracting, and with $\rho = \tilde{\sigma}^2/\sigma^2$ gives the result $f(k)$.

Interval version ($d = 1$). A uniform draw from $[-w, w]$ has mean zero and variance $w^2/3$. Thus,

$$\overline{E}_{\boldsymbol{x}}(\mathcal{H}_{\text{PI}}(\boldsymbol{x})) = E_{\boldsymbol{x}}(h_{\text{erm}}(\boldsymbol{x})) + \frac{1}{3}\sigma^2\gamma\|g_{\boldsymbol{x}}\|^2_{L_2(X)},$$
$$\overline{E}_{\boldsymbol{x}}(\mathcal{H}_{\widetilde{\text{PI}}}(\boldsymbol{x})) = E_{\boldsymbol{x}}(h_{\widetilde{\text{erm}}}(\boldsymbol{x})) + \frac{1}{3}\tilde{\sigma}^2\gamma\|\tilde{g}_{\boldsymbol{x}}\|^2_{L_2(\widetilde{X})}.$$

Averaging over $X \sim p_X$ via Lem. 6 gives $\mathbb{E}_X\left[\|g_X\|^2_{L_2(X)}\right] = k$ and $\mathbb{E}_X\left[\|\tilde{g}_X\|^2_{L_2(\widetilde{X})}\right] = \operatorname{tr}(\mathcal{S})$. Substituting into pointwise approx. errors and subtracting gives the desired result $f(1)$. □

G.3 Proof of Thms. 3.A and 3.B—Robust bounds with DA

Theorem 3 (robust bounds with DA)**.** *Under Asms. 1a–1b, suppose valid PI models following Asms. 2 and 3 yield respective intervals $\mathcal{H}_{\text{pi}}(\boldsymbol{x})$ and $\mathcal{H}_{\text{pi+iv}}(\boldsymbol{x})$ covering an $(\varepsilon, \mathcal{T})$-invariant $h_\star$. Then,*

(A) ε-padded post-DA intervals $\mathcal{H}_{\widetilde{\text{pi}}}(\boldsymbol{x}) \pm \varepsilon$ continue to remain valid for all $\varepsilon \geq 0$ on p_X-a.e. $\boldsymbol{x}$.

(B) ε-padded T-as-IV intervals $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}(\boldsymbol{x}) \pm \varepsilon$ are also valid for $\widetilde{Z} \in \{Z, (Z,T)\}$ on p_X-a.e. $\boldsymbol{x}$ if

$$\varepsilon \leq \tilde{\sigma}\sqrt{\gamma_{\boldsymbol{z}}} - \sigma\sqrt{\gamma^\star_{\boldsymbol{z}}}, \qquad \text{where } \gamma^\star_{\boldsymbol{z}} \leq \gamma_{\boldsymbol{z}} \text{ is the true IV budget.}$$

Proof. Sketch: (i) build a surrogate target $h_\dagger$ within ε of $h_\star$, (ii) certify $h_\dagger \in \mathcal{H}_{\widetilde{\text{pi}}}$ by DPI (Part A), and (iii) verify the IV moment via Minkowski (Part B); ε-pad to convert $h_\dagger$ membership into validity

Surrogate. Define the surrogate as the smoothed target $h_\dagger := \mathbb{E}[h_\star(X) \,|\, \widetilde{X} = \cdot] \in \mathcal{H}$ (Asm. 1a). Define $\eta := h_\star(X) - h_\star(\widetilde{X})$, with $|\eta| \leq \varepsilon$ a.s. by $(\varepsilon, \mathcal{T})$-invariance. Now, $h_\dagger$ obeys two identities:

$$\begin{aligned} h_\dagger(\widetilde{X}) &= h_{\widetilde{\text{erm}}}(\widetilde{X}) - \mathbb{E}[U \,|\, \widetilde{X}], && \text{(Take } \mathbb{E}[\cdot\,|\widetilde{X}] \text{ in Eq. (1).)} \\ &= h_\star(\widetilde{X}) + \mathbb{E}[\eta \,|\, \widetilde{X}], && \text{(Take } \mathbb{E}[\cdot\,|\widetilde{X}] \text{ in def. of } \eta.) \\ \implies \left\|h_\dagger - h_\star\right\|_{L_\infty(X)} &\overset{\text{Asm. 1b}}{\leq} \left\|h_\dagger - h_\star\right\|_{L_\infty(\widetilde{X})} \overset{\text{Jensen}}{\leq} \varepsilon. && (\P) \end{aligned}$$

Furthermore, by tower property $\mathbb{E}[h_\dagger(\widetilde{X})] = \mathbb{E}[h_\star(X)] = 0$, so surrogate is $\widetilde{X}$-centered $h_\dagger \in \mathcal{H}_{\widetilde{X}}$.

Part (A): $h_\dagger \in \mathcal{H}_{\widetilde{\text{pi}}}$. The surrogate's offset from the post-DA center is $h_{\widetilde{\text{erm}}} - h_\dagger = \mathbb{E}[U \,|\, \widetilde{X}]$, so

$$\begin{aligned} \|h_{\widetilde{\text{erm}}} - h_\dagger\|^2_{L_2(\widetilde{X})} &= \text{Var}\big(\mathbb{E}[U \,|\, \widetilde{X}]\big), && (\widetilde{X}\text{-centered } h_\dagger,\ h_{\widetilde{\text{erm}}},\ \Rightarrow \|\cdot\|^2_{L_2} = \text{Var}(\cdot).) \\ &\leq \text{Var}\big(\mathbb{E}[U \,|\, X]\big), && (\text{Contraction over Markov } U - X - \widetilde{X}.) \\ &\leq \gamma\sigma^2, && (\text{Baseline valid } h_\star \in \mathcal{H}_{\text{pi}} \text{ (Asm. 1a, 2).}) \\ &\leq \gamma\tilde{\sigma}^2. && (\text{DPI } \rho \geq 1.) \end{aligned}$$

Post-DA, Lem. 2 gives $\mathcal{H}_{\widetilde{\text{pi}}}$ as the $\mathcal{H}_{\widetilde{X}} \cap L_2(\widetilde{X})$-ball of sq.-radius $\tilde{\sigma}^2\gamma$ about $h_{\widetilde{\text{erm}}}$. So $h_\dagger \in \mathcal{H}$ (Asm. 1a) gives $h_\dagger \in \mathcal{H}_{\widetilde{\text{pi}}} \Rightarrow h_\dagger(\boldsymbol{x}) \in \mathcal{H}_{\widetilde{\text{pi}}}(\boldsymbol{x})$. Hence $h_\star(\boldsymbol{x}) \in \mathcal{H}_{\widetilde{\text{pi}}}(\boldsymbol{x}) \pm \varepsilon$ for p_X-a.e. $\boldsymbol{x}$ by Eq. (¶).

Part (B): $h_\dagger \in \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$. To verify membership, we need to test the surrogate's residual $Y - h_\dagger(\widetilde{X})$ against the IV moment constraint in Asm. 3. Note that this residual is mean-zero and reads as

$$Y - h_\dagger(\widetilde{X}) = \zeta + U + \xi, \qquad \zeta := \eta - \mathbb{E}[\eta \,|\, \widetilde{X}], \qquad \|\zeta\|^2_{L_2} = \mathbb{E}[\text{Var}(\eta \,|\, \widetilde{X})] \leq \varepsilon^2.$$

For any choice of IV $\widetilde{Z} \in \{Z, (Z,T)\}$, the conditional $\mathbb{E}[U + \xi \,|\, Z] = \mathbb{E}[U + \xi \,|\, \widetilde{Z}]$ is conserved since $T \perp\!\!\!\perp (\xi, U, Z)$. Writing the true budget $\gamma^\star_{\boldsymbol{z}} := \|\mathbb{E}[U + \xi \,|\, Z]\|^2_{L_2} \big/ \sigma^2 \leq \gamma_{\boldsymbol{z}}$ (Asm. 3), we have

$$\begin{aligned} \Big\|\mathbb{E}[Y - h_\dagger(\widetilde{X}) \mid \widetilde{Z}]\Big\|_{L_2} &\leq \Big\|\mathbb{E}[\zeta \mid \widetilde{Z}]\Big\|_{L_2} + \Big\|\mathbb{E}[U + \xi \mid \widetilde{Z}]\Big\|_{L_2}, && \text{(Minkowski.)} \\ &\leq \varepsilon + \sigma\sqrt{\gamma^\star_{\boldsymbol{z}}}, && (\text{Contraction; } T \perp\!\!\!\perp (\xi, U, Z).) \\ &\leq \tilde{\sigma}\sqrt{\gamma_{\boldsymbol{z}}}. && (\varepsilon \leq \tilde{\sigma}\sqrt{\gamma_{\boldsymbol{z}}} - \sigma\sqrt{\gamma^\star_{\boldsymbol{z}}}.) \end{aligned}$$

This is exactly the leaky-IV constraint of Asm. 3 under the augmented law, so together with Part (A), $h_\dagger \in \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$, and $h_\star(\boldsymbol{x}) \in \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}(\boldsymbol{x}) \pm \varepsilon$ for p_X-a.e. $\boldsymbol{x}$ uniformly over $\widetilde{Z} \in \{Z, (Z,T)\}$. □

Remark 2 (conservativeness of Thm. 3.B.)**.** Minkowski assumes the worst-case where the two residual components are perfectly aligned. Instead, one can define

$$\kappa := \Big\langle \mathbb{E}[\zeta|\widetilde{Z}],\, \mathbb{E}[U{+}\xi|\widetilde{Z}] \Big\rangle_{L_2} \Big/ \Big(\big\|\mathbb{E}[\zeta|\widetilde{Z}]\big\| \cdot \big\|\mathbb{E}[U{+}\xi|\widetilde{Z}]\big\| \Big) \in [-1, 1],$$

as an alignment sensitivity parameter. The sufficient condition for post-DA leaky-IV validity is now

$$\varepsilon^2 + 2\kappa\,\varepsilon\,\sigma\sqrt{\gamma^\star_z} + \sigma^2\gamma^\star_z \;\leq\; \tilde{\sigma}^2\gamma_z.$$

A conservative $\kappa = 1$ then recovers Thm. 3.B, while a plausible $\kappa = 0$ (misspecification uncorrelated to confounding) gives the Pythagorean budget $\varepsilon \leq \sqrt{\tilde{\sigma}^2\gamma_z - \sigma^2\gamma^\star_z}$, strictly dominating the smaller $\tilde{\sigma}\sqrt{\gamma_z} - \sigma\sqrt{\gamma^\star_z}$. The observable diagnostic $\varepsilon \leq (\tilde{\sigma} - \sigma)\sqrt{\gamma_z}$ of §C remains assumption-free.

G.4 Proof of Prop. 1—Estimation with DA

Proposition 1 (estimation with DA). *Under Asm. 1a, $\mathcal{T}$-inv. $h_\star$ and* bias direction $e_{\tilde{b}} \coloneqq \tilde{b} \big/ \|\tilde{b}\|_{\widetilde{X}}$,

$$E(h_{\mathrm{erm}}) - E(h_{\widetilde{\mathrm{erm}}}) \quad = \quad \underbrace{\sigma^2(\rho-1)\cdot\|e_{\tilde{b}}\|^2_{L_2(X)}}_{\textit{gain}} + \underbrace{(1-\|e_{\tilde{b}}\|^2_{L_2(X)})\cdot E(h_{\mathrm{erm}})}_{\textit{wild-card}}.$$

Proof. Under Asm. 1a, $h_{\mathrm{erm}} = \mathbb{E}[Y|X]$ and $h_{\widetilde{\mathrm{erm}}} = \mathbb{E}[Y|\widetilde{X}]$. Taking conditional means in Eq. (1),

$$b = \mathbb{E}[Y|X] - h_\star = \mathbb{E}[U|X], \qquad \tilde{b} = \mathbb{E}[Y|\widetilde{X}] - h_\star = \mathbb{E}[U|\widetilde{X}],$$

where the latter follows from $\mathcal{T}$-inv. of $h_\star$. Note that $b, \tilde{b} \in \mathcal{H}$ since $h_{\mathrm{erm}}, h_{\widetilde{\mathrm{erm}}}, h_\star \in \mathcal{H}$ (Asm. 1a). Further, $\mathbb{E}[U] = 0$ implies $b \in \mathcal{H}_X$ and $\tilde{b} \in \mathcal{H}_{\widetilde{X}}$. Now, re-write the error terms in respective biases:

$$E(h_{\mathrm{erm}}) = \|b\|^2_{L_2(X)}, \qquad E(h_{\widetilde{\mathrm{erm}}}) = \|\tilde{b}\|^2_{L_2(X)}.$$

On $\widetilde{X}$-centered functions $h \in L^0_2(\widetilde{X})$, define the shift ratio $s(h) \coloneqq \|h\|^2_{L_2(X)} \big/ \|h\|^2_{\widetilde{X}}$ ($s(0) \coloneqq 0$). Substitute into $E(h_{\widetilde{\mathrm{erm}}})$:

$$\begin{aligned}
E(h_{\widetilde{\mathrm{erm}}}) &= \|\tilde{b}\|^2_{L_2(X)}, \\
&= s(\tilde{b}) \cdot \|\tilde{b}\|^2_{\widetilde{X}}, && \text{(Multiply and divide by } \|\tilde{b}\|^2_{\widetilde{X}}.) \\
&= s(\tilde{b}) \cdot \|\tilde{b}\|^2_{L_2(\widetilde{X})}, && (\tilde{b} \in \mathcal{H}_{\widetilde{X}} \implies \|\tilde{b}\|^2_{L_2(\widetilde{X})} = \|\tilde{b}\|^2_{\widetilde{X}}.) \\
&= s(\tilde{b}) \cdot \Big(\|b\|^2_{L_2(X)} - (\tilde{\sigma}^2 - \sigma^2)\Big), && \text{(Lem. 1.)} \\
&= s(\tilde{b}) \cdot \big(E(h_{\mathrm{erm}}) - \sigma^2 \cdot (\rho - 1)\big), && \text{(Substitute } E(h_{\mathrm{erm}}) = \|b\|^2_{L_2(X)} \text{ and } \rho = \tilde{\sigma}^2/\sigma^2.)
\end{aligned}$$

Finally, subtract this from $E(h_{\mathrm{erm}})$ and re-arrange to get a general form of the proposition statement:

$$E(h_{\mathrm{erm}}) - E(h_{\widetilde{\mathrm{erm}}}) = \sigma^2 s(\tilde{b}) \cdot (\rho - 1) + E(h_{\mathrm{erm}}) \cdot (1 - s(\tilde{b})).$$

Operator version under Asms. 1b–1c. Recall from §G that Asms. 1b–1c gives $\|h\|^2_{L_2(X)} = \langle h, \mathcal{S}h\rangle_{\widetilde{X}}$ where operator $\mathcal{S}$ has a Rayleigh quotient $\mathcal{S}[h] \coloneqq \langle h, \mathcal{S}h\rangle_{\widetilde{X}} \big/ \|h\|^2_{\widetilde{X}}$ ($\mathcal{S}[0] \coloneqq 0$) and decomposition $\mathcal{S} = \mathcal{K} + \delta^*\delta$ by Lem. 3.

Trivially, substituting $s(\tilde{b}) = \mathcal{S}[\tilde{b}]$ into the claimed identity now gives an operator version where $\mathcal{S}[\tilde{b}] = \mathcal{K}[\tilde{b}] + (\delta^*\delta)[\tilde{b}]$ (Lem. 3 and Rayleigh quotient is linear in the operator). □

G.5 Proof of Prop. 2—Sharper bounds with DA

Proposition 2 (sharper bounds with DA). *Denoting $L_2(X)$-volume by $|\cdot|$, under Asms. 1b–1c and 2,*

$$\left(\frac{|\mathcal{H}_{\widetilde{\text{pi}}}|}{|\mathcal{H}_{\text{pi}}|}\right)^{2/k} \leq \frac{\mathbb{E}\big[|\mathcal{H}_{\widetilde{\text{pi}}}(X)|^2\big]}{\mathbb{E}\big[|\mathcal{H}_{\text{pi}}(X)|^2\big]} = \underbrace{\rho}_{\text{over-budget}} \cdot \Big[\underbrace{\text{tr}(\mathcal{K})}_{\text{lever}} + \underbrace{D^2_{\mathcal{B}}(p_X, p_{\widetilde{X}})}_{\text{penalty}}\Big] \Big/ k =: \Omega.$$

Proof. Sketching our steps: (i) **volume:** write the post-DA ball in an $L_2(X)$-orthonormal basis of $\mathcal{H}_{\widetilde{X}}$ and compare via AM-GM; (ii) **width:** write widths as representer norms and average via Lem. 6.

Setup. Write Lem. 2 and Cor. 3 pre- and post-DA identified sets in their respective L_2 geometries,

$$\mathcal{H}_{\text{pi}} = \Big\{h \in \mathcal{H}_X : \|h - h_{\text{erm}}\|^2_{L_2(X)} \leq \sigma^2\gamma\Big\}, \qquad \mathcal{H}_{\text{pi}}(\boldsymbol{x}) = \{h_{\text{erm}}(\boldsymbol{x})\} \pm \sigma\sqrt{\gamma}\|g_{\boldsymbol{x}}\|_{L_2(X)},$$
$$\mathcal{H}_{\widetilde{\text{pi}}} = \Big\{h \in \mathcal{H}_{\widetilde{X}} : \|h - h_{\widetilde{\text{erm}}}\|^2_{L_2(\widetilde{X})} \leq \tilde{\sigma}^2\gamma\Big\}, \qquad \mathcal{H}_{\widetilde{\text{pi}}}(\boldsymbol{x}) = \{h_{\widetilde{\text{erm}}}(\boldsymbol{x})\} \pm \tilde{\sigma}\sqrt{\gamma}\|\tilde{g}_{\boldsymbol{x}}\|_{L_2(\widetilde{X})}.$$

We measure both sets in ambient $L_2(X)$-volume and -width by bridging with $L_2(\widetilde{X})$ using operator $\mathcal{S}$.

Volume. Define the $L_2(X)$-volume of a set inside a k-dim. subspace as the Lebesgue volume of its coordinates in *any* $L_2(X)$-orthonormal basis of that subspace; in particular $V_k := \pi^{k/2}/\Gamma(k/2+1)$, the Euclidean unit-ball volume, is the $L_2(X)$-volume of a k-dim. unit ball in any $L_2(X)$-orthonormal basis. The baseline set is an $L_2(X)$-round ball of radius $\sigma\sqrt{\gamma}$ within $\mathcal{H}_X$, so directly

$$|\mathcal{H}_{\text{pi}}| = V_k(\sigma\sqrt{\gamma})^k.$$

For the post-DA set, take an $L_2(\widetilde{X})$-orthonormal eigenbasis $\{\psi_i\}_{i=1}^k$ of $\mathcal{S}$ on $\mathcal{H}_{\widetilde{X}}$ with eigenvalues $\lambda_i > 0$ (Asms. 1b–1c), and rescale $e_i := \psi_i/\sqrt{\lambda_i}$. First, notice $\{e_i\} \subset \mathcal{H}_{\widetilde{X}}$ is $L_2(X)$-orthonormal:

$$\langle e_i, e_j\rangle_{L_2(X)} = \langle e_i, \mathcal{S}e_j\rangle_{L_2(\widetilde{X})}, \qquad \text{(Def. of } \mathcal{S}.)$$
$$= \frac{\lambda_j}{\sqrt{\lambda_i\lambda_j}}\langle\psi_i, \psi_j\rangle_{L_2(\widetilde{X})} = \text{Id}_{ij}. \qquad \text{(Def. of } e_i\text{, and } \mathcal{S}\psi_j = \lambda_j\psi_j.)$$

Therefore we can measure $L_2(X)$-volume in the coordinates $h - h_{\widetilde{\text{erm}}} = \sum_i v_i e_i$. Second, we have:

$$\langle e_i, e_j\rangle_{L_2(\widetilde{X})} = \frac{1}{\sqrt{\lambda_i\lambda_j}}\langle\psi_i, \psi_j\rangle_{L_2(\widetilde{X})} = \frac{1}{\lambda_i}\text{Id}_{ij}. \qquad (L_2(\widetilde{X})\text{-orthonormality of } \psi_i.)$$

Using this identity and linearity of inner-product, substitute $h - h_{\widetilde{\text{erm}}} = \sum_i v_i e_i$ into the def. of $\mathcal{H}_{\widetilde{\text{pi}}}$:

$$\|h - h_{\widetilde{\text{erm}}}\|^2_{L_2(\widetilde{X})} = \left\langle \sum_i v_i e_i, \sum_j v_j e_j \right\rangle_{L_2(\widetilde{X})} = \sum_i\sum_j v_i v_j \langle e_i, e_j\rangle_{L_2(\widetilde{X})} = \boxed{\sum_{i=1}^k \frac{v_i^2}{\lambda_i} \leq \tilde{\sigma}^2\gamma.}$$

Out pops an ellipsoid with semi-axes $\tilde{\sigma}\sqrt{\gamma}\cdot\sqrt{\lambda_i}$. Now, taking its volume and subsequently the ratio:

$$|\mathcal{H}_{\widetilde{\text{pi}}}| = V_k(\tilde{\sigma}\sqrt{\gamma})^k \prod_i \sqrt{\lambda_i} = V_k(\tilde{\sigma}\sqrt{\gamma})^k (\det\mathcal{S})^{1/2},$$

$$\implies \frac{|\mathcal{H}_{\widetilde{\text{pi}}}|}{|\mathcal{H}_{\text{pi}}|} = \rho^{\frac{k}{2}}(\det\mathcal{S})^{\frac{1}{2}} = \rho^{\frac{k}{2}}\Big(\prod_i \lambda_i\Big)^{\frac{1}{2}} \overset{\text{AM–GM}}{\leq} \rho^{\frac{k}{2}}\Big(\frac{1}{k}\sum_i \lambda_i\Big)^{\frac{k}{2}} = \Big(\rho\cdot\frac{\text{tr}(\mathcal{S})}{k}\Big)^{\frac{k}{2}},$$

with equality iff eigenvalues λ_i coincide, i.e. iff $\mathcal{S} = \lambda\cdot\text{Id}$ for some $\lambda > 0$.

Width. Cor. 3 gives half-widths at a query $\boldsymbol{x}$ as $\sigma\sqrt{\gamma}\|g_{\boldsymbol{x}}\|_{L_2(X)}$ and $\tilde{\sigma}\sqrt{\gamma}\|\tilde{g}_{\boldsymbol{x}}\|_{L_2(\widetilde{X})}$. Averaging their squares over $X \sim p_X$ using Lem. 6,

$$\mathbb{E}_X\Big[|\mathcal{H}_{\text{pi}}(X)|^2\Big] = 4\sigma^2\gamma\,\mathbb{E}_X\Big[\|g_X\|^2_{L_2(X)}\Big] = 4\sigma^2\gamma\,k,$$
$$\mathbb{E}_X\Big[|\mathcal{H}_{\widetilde{\text{pi}}}(X)|^2\Big] = 4\tilde{\sigma}^2\gamma\,\mathbb{E}_X\Big[\|\tilde{g}_X\|^2_{L_2(\widetilde{X})}\Big] = 4\tilde{\sigma}^2\gamma\ \text{tr}(\mathcal{S}).$$

Splitting $\text{tr}(\mathcal{S}) = \text{tr}(\mathcal{K}) + \text{tr}(\delta^*\delta)$ (Lem. 3), their ratio gives the claimed identity. □

G.6 Proof of Prop. 3—Shape constraint vs. change-of-measure

Proposition 3 (shape constraint vs. change-of-measure)**.** *Under Asms. 1a–1b, and Asms. 2 to 5 with $\mathcal{T}$-invariant $h_\star$ and $Z = \varnothing$. Set $\varepsilon = \gamma_{\boldsymbol{z}} = 0$, and suppose the baseline PI set is valid $h_\star \in \mathcal{H}_{\text{pi}}$. Then, $h_{\widetilde{\text{erm}}} = \Pi h_{\text{erm}} \in \mathcal{H}_{\text{inv}}$, where Π is the $L_2(X)$-orthogonal projection onto $\mathcal{I}$ (cf. §G). Further,*

$$\mathcal{H}_{\widetilde{\text{pi}}} = \Big\{h \in \mathcal{H}_{\widetilde{X}} \;:\; \big\|h - h_{\widetilde{\text{erm}}}\big\|_{\widetilde{X}}^2 \le \tilde{\sigma}^2\gamma\Big\},$$
$$\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}} = \Big\{h \in \mathcal{H}_{\text{inv}} \colon \big\|h - h_{\widetilde{\text{erm}}}\big\|_{\widetilde{X}}^2 \le \tilde{\sigma}^2\gamma\Big\},$$
$$\mathcal{H}_{\text{pi}+\text{inv}} = \Big\{h \in \mathcal{H}_{\text{inv}} \colon \big\|h - h_{\widetilde{\text{erm}}}\big\|_{X}^2 \le \sigma^2(1+\gamma-\rho)\Big\}.$$

All identified sets are balls centered about $h_{\widetilde{\text{erm}}}$. Both $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$ and $\mathcal{H}_{\text{pi}+\text{inv}}$ live on the slice $\mathcal{H}_{\text{inv}}$, the latter with a smaller radius $\sigma\sqrt{1+\gamma-\rho} \le \tilde{\sigma}\sqrt{\gamma}$ with equality iff $\rho = 1$. It trivially follows that,

$$h_\star \in \mathcal{H}_{\text{pi}+\text{inv}} \subseteq \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}} \subseteq \mathcal{H}_{\widetilde{\text{pi}}},$$

and the same dominance hierarchy holds for sharpness and worst-case errors over $\mathcal{H}_\bullet$ and $\mathcal{H}_\bullet(\boldsymbol{x})$.

Proof. Fix the common center and geometry, characterize the three sets as balls, then compare radii:

Shared center. Under Asms. 1a–1b and 4, Lem. 5 gives: $h_{\widetilde{\text{erm}}} = \Pi h_{\text{erm}} \in \mathcal{H}_{\text{inv}}$.

The post-DA set. $\mathcal{H}_{\widetilde{\text{pi}}} = \Big\{h \in \mathcal{H}_{\widetilde{X}} \colon \|h - h_{\widetilde{\text{erm}}}\|_{\widetilde{X}}^2 \le \tilde{\sigma}^2\gamma\Big\}$ immediate from Lem. 2 under $p_{\widetilde{X}Y}$.

The T-as-IV set. For $\widetilde{Z} = T$, $\gamma_{\boldsymbol{z}} = 0$, the IV constraint of Cor. 2 under $p_{\widetilde{X}Y\widetilde{Z}}$ filters $h \in \mathcal{H}_{\widetilde{X}}$ via

$$\begin{aligned}
&\operatorname{Var}(\mathbb{E}[Y - h(\widetilde{X})\,|\,T]) = 0 \iff \operatorname{Var}(\mathbb{E}[Y] - \mathbb{E}[h(\widetilde{X})\,|\,T]) = 0 && (T \perp\!\!\!\perp Y \text{ by construction.})\\
\iff\quad &\operatorname{Var}(\mathbb{E}[h(\widetilde{X})\,|\,T]) = 0 \iff \mathcal{E}_{\widetilde{X}|T} h = 0 && (\text{Def. of } \mathcal{E}_{\widetilde{X}|T}.)\\
&\iff h \in \operatorname{null}\big(\mathcal{E}_{\widetilde{X}|T}\big|_{\mathcal{H}_{\widetilde{X}}}\big) = \mathcal{H}_{\text{inv}}, && (\text{Asm. 5.})\\
&\implies \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}} = \mathcal{H}_{\widetilde{\text{pi}}} \cap \mathcal{H}_{\text{inv}}. && (\text{Cor. 2 under } p_{\widetilde{X}Y\widetilde{Z}}.)
\end{aligned}$$

The inv. constraint set. By Cor. 1, $\mathcal{H}_{\text{pi}+\text{inv}} = \mathcal{H}_{\text{pi}} \cap \mathcal{H}_{\text{inv}}$, with $\mathcal{H}_{\text{pi}}$ the X-ball of radius $\sigma\sqrt{\gamma}$ about h_{erm} (Lem. 2). For invariant $h \in \mathcal{H}_{\text{inv}}$, using Pythagoras along Π followed by $h = \Pi h$,

$$\begin{aligned}
\|h - h_{\text{erm}}\|_X^2 &= \|\Pi(h - h_{\text{erm}})\|_X^2 + \|(\mathrm{Id} - \Pi)(h - h_{\text{erm}})\|_X^2,\\
&= \|h - \Pi h_{\text{erm}}\|_X^2 + \|(\mathrm{Id} - \Pi)h_{\text{erm}}\|_X^2. && (\triangle)
\end{aligned}$$

Writing pre- and post-DA biases $b := h_{\text{erm}} - h_\star$ and $\tilde{b} := h_{\widetilde{\text{erm}}} - h_\star$, the second term reduces to:

$$\begin{aligned}
\|(\mathrm{Id} - \Pi)h_{\text{erm}}\|_X^2 &= \|(\mathrm{Id} - \Pi)(h_\star + b)\|_X^2 = \|(\mathrm{Id} - \Pi)b\|_X^2, && (\text{Def. of bias } b, \mathcal{T}\text{-inv. } h_\star.)\\
&= \|b - \tilde{b}\|_X^2 = \|b\|_X^2 - \|\tilde{b}\|_X^2, && (\text{Lem. 5 } \tilde{b} = \Pi b, \text{Pythagoras.})\\
&= \|b\|_X^2 - \|\tilde{b}\|_{\widetilde{X}}^2 = \tilde{\sigma}^2 - \sigma^2. && (\tilde{b} \in \mathcal{I} \Rightarrow \text{Eq. (6), Lem. 1.})
\end{aligned}$$

Substituting into ($\triangle$), together with $\Pi h_{\text{erm}} = h_{\widetilde{\text{erm}}}$ and baseline ball constraint $\|h - h_{\text{erm}}\|_X^2 \le \sigma^2\gamma$,

$$\implies \|h - h_{\widetilde{\text{erm}}}\|_X^2 \;\le\; \sigma^2\gamma - (\tilde{\sigma}^2 - \sigma^2) \;=\; \sigma^2(1 + \gamma - \rho).$$

Inclusion and dominance hierarchy. Both T-as-IV and inv.-constrained sets $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$, $\mathcal{H}_{\text{pi}+\text{inv}}$ are balls in $\mathcal{H}_{\text{inv}}$ about the common center $h_{\widetilde{\text{erm}}}$. Since $h - h_{\widetilde{\text{erm}}} \in \mathcal{I}$ for invariant h, their norms are measured in the common geometry of Eq. (6), with squared radii obeying

$$\tilde{\sigma}^2\gamma - \sigma^2(1 + \gamma - \rho) = \sigma^2(1+\gamma)(\rho - 1) \;\ge\; 0,$$

where $\rho \ge 1$ by the DPI (Lem. 1). Hence $\mathcal{H}_{\text{pi}+\text{inv}} \subseteq \mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}}$, while $\mathcal{H}_{\widetilde{\text{pi}}+\widetilde{\text{iv}}} \subseteq \mathcal{H}_{\widetilde{\text{pi}}}$ is immediate.

Finally, set inclusion directly implies a dominance hierarchy in sharpness and worst-case errors. □

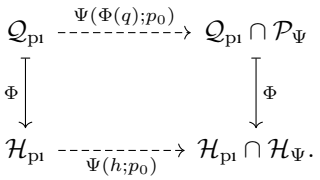


Figure 14: A picture of Prop. 4: Filtering baseline models $q \in \mathcal{Q}_{\text{PI}}$ by the predicate $\Psi(\Phi(q); p_0)$ and then mapping through Φ yields the same hypothesis set as mapping $\mathcal{H}_{\text{PI}} = \Phi(\mathcal{Q}_{\text{PI}})$ first and then filtering by the corresponding function-level predicate $\Psi(h; p_0)$. Dashed arrows denote set restriction.

G.7 Proof of Prop. 4—Sharp bounds with structural-function constraints

Proposition 4 (sharp bounds with structural-function constraints)**.** *Fix some distribution p_0. Define*

$$\Phi\colon \mathcal{P}_{\text{PI}} \to \mathcal{H},\ q \mapsto h^q_\star.$$

For a set of baseline PI candidate distributions $\mathcal{Q}_{\text{PI}} \subseteq \mathcal{P}_{\text{PI}}$, their candidate hypotheses are given by

$$\mathcal{H}_{\text{PI}} \coloneqq \Phi(\mathcal{Q}_{\text{PI}}).$$

Let $\Psi(h; p_0)$ be a predicate depending only on functions $h \in \mathcal{H}$ and the fixed law p_0. Define the sets

$$\mathcal{H}_\Psi \coloneqq \{h \in \mathcal{H}\colon \Psi(h; p_0)\}, \qquad \mathcal{P}_\Psi \coloneqq \{q \in \mathcal{P}_{\text{PI}}\colon \Psi(\Phi(q); p_0)\}.$$

Then $\mathcal{P}_\Psi = \Phi^{-1}(\mathcal{H}_\Psi)$ is the pre-image of $\mathcal{H}_\Psi$ under Φ, and the constrained identified set is sharp*:*

$$\Phi(\mathcal{Q}_{\text{PI}} \cap \mathcal{P}_\Psi) = \mathcal{H}_{\text{PI}} \cap \mathcal{H}_\Psi.$$

The same identity holds for any finite collection of such predicates. Fig. 14 states the prop. in a picture.

Proof. Re-write the claimed identity by substituting $\mathcal{H}_{\text{PI}} = \Phi(\cdot)$, $\mathcal{P}_\Psi = \Phi^{-1}(\cdot)$, $Q = \mathcal{Q}_{\text{PI}}$, $H = \mathcal{H}_\Psi$,

$$\Phi\big(Q \cap \Phi^{-1}(H)\big) = \Phi(Q) \cap H.$$

Necessity ($\subseteq$). Trivially, for any $h \in \Phi\big(Q \cap \Phi^{-1}(H)\big)$, we have $h = \Phi(q)$ for some candidate $q \in Q \cap \Phi^{-1}(H)$. Directly implying that $h \in \Phi(Q)$ and $h \in H$. Hence, the LHS implies the RHS.

Sufficiency ($\supseteq$). Let $h \in \Phi(Q) \cap H$. Then there exists $q \in Q$ such that $\Phi(q) = h$, which is in H. Since $\Phi(q) \in H$, then $q \in \Phi^{-1}(H)$. Thus, $q \in Q \cap \Phi^{-1}(H)$, giving $h = \Phi(q) \in \Phi\big(Q \cap \Phi^{-1}(H)\big)$. □

Corollary 1 (sharp bounds with inv. constraints)**.** *Define hypotheses via the inv.-error functional E_{INV}*

$$\mathcal{H}^{\varepsilon}_{\text{INV}} \coloneqq \big\{h \in \mathcal{H} : E_{\text{INV}}(h) \le \varepsilon^2\big\}.$$

Then, the invariance shape constraint PI model from §3 is sharp*, meaning $\mathcal{H}_{\text{PI+INV}} = \mathcal{H}_{\text{PI}} \cap \mathcal{H}^{\varepsilon}_{\text{INV}}$.*

Proof. Immediate from Prop. 4 using $p_0 = p_{X\tilde{X}}$ and $\Psi(h; p_0) = \mathbf{1}\big\{E_{\text{INV}}(h) \le \varepsilon^2\big\}$. □

Corollary 2 (sharp bounds with IV constraints)**.** *Let σ_p^2 be the ERM noise floor under true p. Define*

$$\mathcal{H}_{\text{IV}} \coloneqq \Big\{h \in \mathcal{H} : \mathrm{Var}_p\Big(\mathbb{E}_p[Y - h(X) \,|\, Z]\Big) \le \sigma_p^2 \gamma_{\boldsymbol{z}}\Big\},$$

as the leaky-IV hypotheses. Then, the PI model of Asm. 3 is sharp*, meaning $\mathcal{H}_{\text{PI+IV}} = \mathcal{H}_{\text{PI}} \cap \mathcal{H}_{\text{IV}}$.*

Proof. Immediate from Prop. 4 using $p_0 = p_{XYZ}$ and the predicate

$$\Psi(h; p_0) = \mathbf{1}\Big\{\mathrm{Var}_{p_0}\Big(\mathbb{E}_{p_0}[Y - h(X) \,|\, Z]\Big) \le \sigma_{p_0}^2 \gamma_{\boldsymbol{z}}\Big\}.$$ □

G.8 Proof of Lem. 1—Soft intervention with DA

Lemma 1 (soft intervention with DA). *For $\mathcal{T}$-inv. $h_\star$, DA acts on Eq. (1) as a* soft-intervention*:*

(A) Structural equation: *Trivially, post-DA variables $(\xi, U, \widetilde{X}, Y, \widetilde{Z})$ with $\widetilde{Z} \in \{Z, (Z,T)\}$ follow:*

$$Y = h_\star(\widetilde{X}) + U + \xi, \qquad \mathbb{E}[U] = \mathbb{E}[\xi] = 0, \qquad \xi \perp\!\!\!\perp (U, \widetilde{X}, \widetilde{Z}).$$

(B) Markov transfer: *Let p be Markov w.r.t. a DAG $\mathcal{G}$ over $V := \{U, X, Y, Z_1, Z_2, \ldots\}$ where $\mathrm{NonDesc}_{\mathcal{G}}(X) = V \backslash \{X, Y\}$. Then, $p_{\widetilde{V}}$ is Markov w.r.t. $\widetilde{\mathcal{G}}$, where $\widetilde{V}, \widetilde{\mathcal{G}}$ relabel X to $\widetilde{X}$ in $V, \mathcal{G}$.*

(C) Attenuated confounding: *Under Asm. 1a, let biases $b := h_{\mathrm{erm}} - h_\star$ and $\tilde{b} := h_{\widetilde{\mathrm{erm}}} - h_\star$. Then,*

$$\|b\|_X^2 - \|\tilde{b}\|_{\widetilde{X}}^2 \;=\; \sigma^2(\rho - 1) \;\geq\; 0.$$

Proof. Part *(a)* is immediate from def. of $\mathcal{T}$-inv. and $T \perp\!\!\!\perp (\xi, U, X, Y, Z)$. We prove parts *(b)–(c)*.

Part (b): Markov transfer. Write $W := V \backslash \{X, Y\}$, so $\widetilde{V} = W \cup \{\widetilde{X}, Y\}$. Now, disintegrate $p_{\widetilde{V}}$,

$$p_{W\widetilde{X}Y} = p_W \otimes p_{\widetilde{X}|W} \otimes p_{Y|\widetilde{X}W},$$

and reduce each kernel to its $\widetilde{\mathcal{G}}$-kernel,

$$p_W = \bigotimes\nolimits_{V \in W} p_{V|\mathrm{Pa}_{\mathcal{G}}(V)}, \qquad (W \text{ ancestral in } \mathcal{G}, \text{ so its } \mathcal{G}\text{-marginal factorizes.})$$

$$p_{\widetilde{X}|W}(\,\cdot\,|\,\boldsymbol{w}) \qquad (\text{Integrate out } \boldsymbol{\tau}.)$$

$$= \int \alpha(\cdot, \boldsymbol{\tau})_{\#}\, p_{X|W}(\,\cdot\,|\,\boldsymbol{w})\; p_T(\mathrm{d}\boldsymbol{\tau}), \qquad (\text{Use } T \perp\!\!\!\perp (W, X) \text{ and } \widetilde{X} = \alpha(X, \boldsymbol{\tau}) \text{ on } \{T = \boldsymbol{\tau}\}.)$$

$$= \int \alpha(\cdot, \boldsymbol{\tau})_{\#}\, p_{X|\mathrm{Pa}_{\mathcal{G}}(X)}\big(\,\cdot\,\big|\,\mathrm{pa}_{\mathcal{G}}(\boldsymbol{x})\big)\; p_T(\mathrm{d}\boldsymbol{\tau}), \qquad (W \text{ non-desc., so } X \perp\!\!\!\perp W \backslash \mathrm{Pa}_{\mathcal{G}}(X) \,\big|\, \mathrm{Pa}_{\mathcal{G}}(X).)$$

$$=: p_{\widetilde{X}|\mathrm{Pa}_{\mathcal{G}}(X)}\big(\,\cdot\,\big|\,\mathrm{pa}_{\mathcal{G}}(\boldsymbol{x})\big),$$

$$p_{Y|\widetilde{X}W}(\,\cdot\,|\,\boldsymbol{x}, \boldsymbol{w}) \qquad (\text{Part } \textit{(a)}\text{: } Y = h_\star(\widetilde{X}) + U + \xi \text{ with exogenous } \xi \perp\!\!\!\perp (W, \widetilde{X}).)$$

$$= p_\xi(\,\cdot\, - h_\star(\boldsymbol{x}) - u) =: p_{Y|\widetilde{X}U}(\,\cdot\,|\,\boldsymbol{x}, u).$$

Substitute back to get the desired factorization along $\widetilde{\mathcal{G}}$, i.e. all parent sets of $\mathcal{G}$ with X relabeled $\widetilde{X}$:

$$p_{W\widetilde{X}Y} = \bigotimes\nolimits_{V \in W} p_{V|\mathrm{Pa}_{\mathcal{G}}(V)} \;\otimes\; p_{\widetilde{X}|\mathrm{Pa}_{\mathcal{G}}(X)} \;\otimes\; p_{Y|\widetilde{X}U}.$$

Part (c): attenuated confounding. Under Asm. 1a, $h_{\mathrm{erm}} = \mathbb{E}[Y|X]$ and $h_{\widetilde{\mathrm{erm}}} = \mathbb{E}[Y|\widetilde{X}]$. Taking conditional means in Eq. (1),

$$b = \mathbb{E}[Y|X] - h_\star = \mathbb{E}[U|X], \qquad \tilde{b} = \mathbb{E}[Y|\widetilde{X}] - h_\star = \mathbb{E}[U|\widetilde{X}],$$

where the latter follows from $\mathcal{T}$-inv. of $h_\star$.

The identity: Since $\xi \perp\!\!\!\perp (U, X)$, the noise floor σ^2 decomposes as

$$\begin{aligned} \sigma^2 &= \mathbb{E}[\mathrm{Var}(Y|X)], && (\text{Asm. 1a.}) \\ &= \mathrm{Var}(\xi) + \mathbb{E}[\mathrm{Var}(U|X)], && (\xi \perp\!\!\!\perp (U, X).) \\ &= \mathrm{Var}(\xi) + \mathrm{Var}(U) - \|b\|_{L_2(X)}^2, && (\text{Law of total variance.}) \end{aligned}$$

and identical steps under $\widetilde{X}$ ($\mathcal{T}$-inv. at the first step) gives $\tilde{\sigma}^2 = \mathrm{Var}(\xi) + \mathrm{Var}(U) - \|\tilde{b}\|_{L_2(\widetilde{X})}^2$. Subtracting the two noise floors gives $\|b\|_{L_2(X)}^2 - \|\tilde{b}\|_{L_2(\widetilde{X})}^2 = \tilde{\sigma}^2 - \sigma^2$.

Non-negativity: As $\widetilde{X} = \alpha(X, T)$ with $T \perp\!\!\!\perp (U, X)$, the chain $U - X - \widetilde{X}$ is Markov, hence:

$$\tilde{b} = \mathbb{E}[U \,|\, \widetilde{X}] = \mathbb{E}\Big[\mathbb{E}[U|X]\,\Big|\,\widetilde{X}\Big] = \mathbb{E}[b(X) \,|\, \widetilde{X}] \qquad (\text{Tower rule.})$$

Because conditional expectation is an L_2-contraction, we have:

$$\|\tilde{b}\|_{L_2(\widetilde{X})}^2 = \mathrm{Var}(\mathbb{E}[U \,|\, \widetilde{X}]) \leq \mathrm{Var}(\mathbb{E}[U \,|\, X]) = \|b\|_{L_2(X)}^2,$$

i.e. $\tilde{\sigma}^2 - \sigma^2 = \|b\|_{L_2(X)}^2 - \|\tilde{b}\|_{L_2(\widetilde{X})}^2 \geq 0$. □

G.9 Proof of Lem. 2—Characterizing the identified set[3]

Lemma 2 (characterizing the identified set)**.** *Under Asm. 2 and 1a or 1c, $\mathcal{H}_{\text{PI}}$ is the covariance ball*

$$\mathcal{H}_{\text{PI}} = \Big\{ h \in \mathcal{H}_X : \big\|h - h_{\text{erm}}\big\|_X^2 \le \sigma^2 \gamma \Big\},$$

within $\mathcal{H}_X$, centered at h_{erm} with radius $\sigma\sqrt{\gamma}$, and an effective confounding budget *of $\gamma := \frac{\gamma_{\boldsymbol{x}}\gamma_y}{1-\gamma_{\boldsymbol{x}}}$.*

Proof. With $\mathcal{B}$ the ball in the statement, the proof has three parts. (i) **Anchor:** all $q \in \mathcal{Q}_{\text{PI}}$ share the center h_{erm} and radius σ. (ii) **Necessity** ($\mathcal{H}_{\text{PI}} \subseteq \mathcal{B}$)**:** $h^q_\star \in \mathcal{B}$ for every q by bounding on budgets $\gamma_{\boldsymbol{x}}, \gamma_y$; (iii) **Sufficiency** ($\mathcal{H}_{\text{PI}} \supseteq \mathcal{B}$)**:** Any $h \in \mathcal{B}$ is achieved by some Gaussian certificate $q^h \in \mathcal{Q}_{\text{PI}}$.

Anchor. Denote $\mathcal{V} := \operatorname{span}(\mathcal{H})$. The IPM test class $\mathcal{F} = \{f^2 : f \in \mathcal{V} + \mathbb{R}Y\}$ in Asm. 2 matches p_{XY} and q_{XY} up to second moments of $(v(X), Y)$ for all $v \in \mathcal{V}$. Hence, all q induce a common risk

$$R_{\text{erm}}(h) = \mathbb{E}[Y^2] - 2\,\mathbb{E}[h(X)Y] + \mathbb{E}[h(X)^2], \qquad h \in \mathcal{H},$$

with common minimizer h_{erm} and minimum σ^2. Moreover,

$$\mathbb{E}_p[v(X) \cdot (Y - h_{\text{erm}}(X))] = 0 \qquad \forall v \in \mathcal{V}. \tag{♣}$$

Under Asm. 1a, this holds because $h_{\text{erm}} = \mathbb{E}_p[Y \,|\, X]$; under Asm. 1c, $\mathcal{V} = \mathcal{H}$ and $h_{\text{erm}} = \mathcal{T}_{\mathcal{H}} Y$.

For $v = 1 \in \mathcal{V}$, Eq. (♣) gives $\mathbb{E}[h_{\text{erm}}(X)] = \mathbb{E}[Y]$, confirming that $h_{\text{erm}} \in \mathcal{H}_X$ under mean-zero Y.

Necessity ($\mathcal{H}_{\text{PI}} \subseteq \mathcal{B}$)**.** Write Eq. (1) under some fixed $q \in \mathcal{Q}_{\text{PI}}$

$$Y = h^q_\star(X) + U_q + \xi_q, \qquad\qquad \xi_q \perp\!\!\!\perp (U_q, X, Z),$$

omitting the sub-/superscript q where unambiguous. Define

$$b^q := h_{\text{erm}} - h^q_\star \in \mathcal{H} - \mathcal{H} \subseteq \mathcal{V}, \qquad\qquad r_q := \mathbb{E}_q[Y \,|\, X] - h_{\text{erm}} \perp_{L_2(X)} \mathcal{V},$$

where $r_q \perp_{L_2(X)} \mathcal{V}$ holds because for every $v \in \mathcal{V}$, tower rule, moment matching, and Eq. (♣) give

$$\begin{aligned}\mathbb{E}_q[v(X) r_q(X)] &= \mathbb{E}_q[v(X)Y] - \mathbb{E}_q[v(X) h_{\text{erm}}(X)] \\ &= \mathbb{E}_p[v(X)Y] - \mathbb{E}_p[v(X) h_{\text{erm}}(X)] = 0.\end{aligned}$$

Conditioning the structural on X now yields,

$$\mathbb{E}_q[U \,|\, X] = \mathbb{E}_q[Y \,|\, X] - h^q_\star = \underbrace{r_q}_{\perp \mathcal{V}} + \underbrace{b^q}_{\in \mathcal{V}}. \tag{♠}$$

with r_q, b^q both X-centered. Now define:

$$a := \mathbb{E}_q[\operatorname{Var}_q(U \,|\, X)], \qquad\qquad c := \operatorname{Var}_q\big(\mathbb{E}_q[U \,|\, X]\big).$$

To show that $h^q_\star \in \mathcal{B}$, we need to show $\|b^q\|_X^2 \le \sigma^2\gamma$. This constitutes a chain of three inequalities,

$$\|b\|_X^2 \overset{(**)}{\le} c \overset{(††)}{\le} \frac{\gamma_{\boldsymbol{x}}}{1-\gamma_{\boldsymbol{x}}} a \overset{(‡‡)}{\le} \frac{\gamma_{\boldsymbol{x}}\gamma_y}{1-\gamma_{\boldsymbol{x}}}\sigma^2 \le \sigma^2\gamma.$$

Here, (∗∗) holds by Pythagoras on Eq. (♠). And (††) follows from the X-budget

$$r^2_{U\sim X} := 1 - \frac{a}{\sigma_U^2} \le \gamma_{\boldsymbol{x}} \quad\Longrightarrow\quad \sigma_U^2 \le \frac{a}{1-\gamma_{\boldsymbol{x}}} \quad\Longrightarrow\quad c \overset{\text{total-var.}}{=} \sigma_U^2 - a \le \frac{\gamma_{\boldsymbol{x}}}{1-\gamma_{\boldsymbol{x}}} a.$$

To establish (‡‡), we need to upper-bound a. The Y-budget gives,

$$r^2_{U\sim Y|X} := 1 - \mathbb{E}[\operatorname{Var}(U \,|\, X, Y)]/\,a \;\le\; \gamma_y, \quad\Longrightarrow\quad (1-\gamma_y)\,a \;\le\; \mathbb{E}[\operatorname{Var}(U \,|\, X, Y)]. \tag{◇}$$

[3] By convention, set $\gamma := +\infty$ when $\gamma_{\boldsymbol{x}} = 1$ in Asm. 2, and $r^2 := 0$ whenever its denominator vanishes.

Now, condition on $X = \boldsymbol{x}$ and write $s(\boldsymbol{x}) \coloneqq \mathrm{Var}(U \,|\, X = \boldsymbol{x})$, so that $\mathbb{E}[s(X)] = a$. Note that $\mathbb{E}[\mathrm{Var}(U \,|\, Y, X) \,|\, X = \boldsymbol{x}]$ is the MMSE of predicting U from Y, so a *linear* MMSE upper-bounds it:

$$\mathbb{E}[\mathrm{Var}(U \,|\, Y, X) \,|\, X = \boldsymbol{x}] \leq \frac{s(\boldsymbol{x})\,\sigma_\xi^2}{s(\boldsymbol{x}) + \sigma_\xi^2}.$$

The RHS is concave in $s(\boldsymbol{x})$. For $a > 0$, Jensen and Eq. (♢) give (conclusion immediate for $a = 0$),

$$(1 - \gamma_y)\, a \leq \frac{a\,\sigma_\xi^2}{a + \sigma_\xi^2}, \qquad \Longrightarrow \qquad a \leq \gamma_y\,(a + \sigma_\xi^2).$$

This, combined with the below inequality gives $a \leq \gamma_y\,\sigma^2$, finally establishing $\|b\|_X^2 \leq \sigma^2\gamma$ via (‡‡).

$$a + \sigma_\xi^2 \overset{\text{total-var.}}{=} \mathbb{E}[\mathrm{Var}(Y \,|\, X)] \overset{\mathbb{E}[Y|X]\text{ optimality}}{\leq} \min_{h \in \mathcal{H}} \mathbb{E}\Big[\big(Y - h(X)\big)^2\Big] \quad =: \quad \sigma^2.$$

Sufficiency ($\mathcal{H}_{\text{PI}} \supseteq \mathcal{B}$). Fix $h \in \mathcal{B}$, then define

$$b^h \coloneqq h_{\text{erm}} - h, \qquad e^2 \coloneqq \big\|b^h\big\|_X^2 \leq \sigma^2\gamma.$$

Suppose first that $\gamma_{\boldsymbol{x}} > 0$ and $e > 0$. Define,

$$\sigma_W^2 \coloneqq \frac{1 - \gamma_{\boldsymbol{x}}}{\gamma_{\boldsymbol{x}}} e^2, \qquad \sigma_\xi^2 \coloneqq \sigma^2 - \sigma_W^2,$$

Since $e^2 \leq \sigma^2\gamma$,

$$\Longrightarrow\ \sigma_W^2 \leq \frac{1 - \gamma_{\boldsymbol{x}}}{\gamma_{\boldsymbol{x}}}\sigma^2\gamma = \gamma_y\sigma^2 \leq \sigma^2, \tag{♡}$$

so $\sigma_\xi^2 \geq 0$. Sample independently

$$(X, Z) \sim p_{XZ}, \qquad W_h \sim \mathcal{N}(0, \sigma_W^2), \qquad \xi_h \sim \mathcal{N}(0, \sigma_\xi^2),$$

and define

$$U_h \coloneqq b^h(X) + W_h, \qquad Y = h(X) + U_h + \xi_h = h_{\text{erm}}(X) + W_h + \xi_h.$$

This satisfies the structural equation with $h_\star^{q^h} = h$; U_h and ξ_h are centered, and $\xi_h \perp\!\!\!\perp (U_h, X, Z)$.

It remains to verify $q^h \in \mathcal{Q}_{\text{PI}}$.

(a) Moment matching. For every $v \in \mathcal{V}$, $q_X^h = p_X$ and $Y - h_{\text{erm}}(X) = W_h + \xi_h$ is centered, independent of X, and has variance σ^2. Thus

$$\mathbb{E}_{q^h}[v(X)Y] = \mathbb{E}_p[v(X)h_{\text{erm}}(X)] = \mathbb{E}_p[v(X)Y],$$

by Eq. (♣), while

$$\mathbb{E}_{q^h}[v(X)^2] = \mathbb{E}_p[v(X)^2], \qquad \mathbb{E}_{q^h}[Y^2] = \mathbb{E}_p[h_{\text{erm}}(X)^2] + \sigma^2 = \mathbb{E}_p[Y^2],$$

the latter by Pythagoras under p. First moments match similarly. Hence $D_{\mathcal{F}}(p_{XY}, q_{XY}^h) = 0$.

(b) The $\gamma_{\boldsymbol{x}}$ budget: Since $W_h \perp\!\!\!\perp X$,

$$a = \sigma_W^2, \qquad \sigma_U^2 = e^2 + \sigma_W^2 = e^2 \big/ \gamma_{\boldsymbol{x}},$$

and therefore

$$r_{U\sim X}^2 = 1 - \frac{\sigma_W^2}{e^2/\,\gamma_{\boldsymbol{x}}} = 1 - (1 - \gamma_{\boldsymbol{x}}) = \gamma_{\boldsymbol{x}} \leq \gamma_{\boldsymbol{x}}.$$

(c) The γ_y budget: Conditional on X, observing Y is equivalent to observing $W_h + \xi_h$. Therefore

$$\mathrm{Var}_q(U \mid X = \boldsymbol{x}, Y) = \mathrm{Var}(W \mid W + \xi) = \frac{\sigma_W^2 \cdot \sigma_\xi^2}{\sigma_W^2 + \sigma_\xi^2} = \frac{\sigma_W^2\,(\sigma^2 - \sigma_W^2)}{\sigma^2},$$

whence

$$r_{U\sim Y|X}^2 = 1 - \frac{\sigma_W^2\,(\sigma^2 - \sigma_W^2)/\sigma^2}{\sigma_W^2} = \frac{\sigma_W^2}{\sigma^2} \leq \gamma_y,$$

using $\sigma_W^2 \leq \gamma_y\sigma^2$ from Eq. (♡). Hence $q^h \in \mathcal{Q}_{\text{PI}}$.

Edge cases. If $e = 0$ (equivalently $h = h_{\mathrm{erm}}$ in $\mathcal{H}_X$; this covers $\gamma_{\boldsymbol{x}} = 0$ or $\gamma_y = 0$, where $\gamma = 0$ and the ball degenerates to $\{h_{\mathrm{erm}}\}$), take $U_h \coloneqq 0$ and $\xi_h \sim \mathcal{N}(0, \sigma^2)$. Moment matching *(a)* holds as before, and both budgets *(b)–(c)* hold with $r^2 = 0$.

Hence, every $h \in \mathcal{B}$ is attained by a Gaussian certificate (note, p itself is *not* assumed Gaussian). □

Corollary 3 (characterizing the identified interval)**.** *Under Asm. 1c and Asm. 2, for* $\boldsymbol{x} \in \mathrm{supp}(p_X)$*,*

$$\mathcal{H}_{\mathrm{PI}}(\boldsymbol{x}) = \{h_{\mathrm{erm}}(\boldsymbol{x})\} \pm \sigma\sqrt{\gamma}\,\|g_{\boldsymbol{x}}\|_{L_2(X)},$$

where $g_{\boldsymbol{x}} \in \mathcal{H}_X$ *is the* $L_2(X)$*-representer of* $h \mapsto h(\boldsymbol{x})$*, well defined by the non-degenerate* $\langle\cdot,\cdot\rangle_X$*.*

Proof. For any $h \in \mathcal{H}_{\mathrm{PI}}$ as given by Lem. 2, evaluation at query point $\boldsymbol{x} \in \mathrm{supp}(p_X)$ follows

$$\begin{aligned} |h(\boldsymbol{x}) - h_{\mathrm{erm}}(\boldsymbol{x})| &= \left|\langle h - h_{\mathrm{erm}},\, g_{\boldsymbol{x}}\rangle_{L_2(X)}\right|, && \text{(Reproducing identity.)} \\ &\le \|h - h_{\mathrm{erm}}\|_{L_2(X)} \cdot \|g_{\boldsymbol{x}}\|_{L_2(X)}, && \text{(Cauchy–Schwarz.)} \\ &\le \sigma\sqrt{\gamma}\,\|g_{\boldsymbol{x}}\|_{L_2(X)}, && \text{(From Lem. 2.)} \end{aligned}$$

with equality at $h - h_{\mathrm{erm}} = \pm\sigma\sqrt{\gamma}\, g_{\boldsymbol{x}}/\|g_{\boldsymbol{x}}\|_{L_2(X)} \in \mathcal{H}$, membership by linearity of $\mathcal{H}$ (Asm. 1c). These two extremal functions are attained since they lie in the ball of Lem. 2 by construction.

Finally, since $\mathcal{H}_{\mathrm{PI}}$ is convex (a ball within linear $\mathcal{H}$ under Asm. 1c) and evaluation $h \mapsto h(\boldsymbol{x})$ is linear, the image $\mathcal{H}_{\mathrm{PI}}(\boldsymbol{x})$ is an interval, and therefore coincides with the full interval above. □

G.10 Proof of Lem. 3—Shift decomposition

Lemma 3 (shift decomposition). *Under Asms. 1b–1c, $\mathcal{S} = \mathcal{K} + \delta^*\delta$, with $\mathcal{K} \succ 0$, $\delta^*\delta \succcurlyeq 0$ s.t.* $\operatorname{rank}(\delta^*\delta) \le 1$, *and* $\operatorname{tr}(\delta^*\delta) = D^2_{\mathcal{B}}(p_X, p_{\widetilde{X}})$ *is MMD^2 over the unit ball $\mathcal{B} := \{h \in \mathcal{H}_{\widetilde{X}} : \|h\|_{\widetilde{X}} \le 1\}$.*

Proof. We make three moves: (i) decompose the evaluation form as covariance-plus-mean, (ii) read off the two operators it splits into via Riesz, and (iii) identify the rank and trace of the mean part.

Evaluation form identity. Take $f, g \in \mathcal{H}_{\widetilde{X}}$. Scoring them in $L_2(X)$ gives

$$\begin{aligned}\langle f, g\rangle_{L_2(X)} &= \mathbb{E}\big[f(X)\, g(X)\big],\\ &= \operatorname{Cov}(f(X), g(X)) + \mathbb{E}[f(X)]\,\mathbb{E}[g(X)],\\ &= \langle f, g\rangle_X + (\delta f)\,(\delta g).\end{aligned}$$

Operator decomposition via Riesz. By Asm. 1c, $(\mathcal{H}_X, \langle\cdot,\cdot\rangle_X)$ is Hilbert. By Asm. 1b (cf. §G),

$$(\mathcal{H}_{\widetilde{X}}, \langle\cdot,\cdot\rangle_{\widetilde{X}})$$

is also Hilbert, so every symmetric bilinear form on it is represented by a unique self-adjoint operator. Additionally, the forms $\langle\cdot,\cdot\rangle_{L_2(X)}$ and $\langle\cdot,\cdot\rangle_X$ are also both positive-definite on $\mathcal{H}_{\widetilde{X}}$ (§G).

Riesz gives unique self-adjoint, positive-definite $\mathcal{S}, \mathcal{K} : \mathcal{H}_{\widetilde{X}} \to \mathcal{H}_{\widetilde{X}}$ satisfying

$$\langle f, \mathcal{S}g\rangle_{\widetilde{X}} = \langle f, g\rangle_{L_2(X)}, \qquad \langle f, \mathcal{K}g\rangle_{\widetilde{X}} = \langle f, g\rangle_X.$$

Likewise, $\delta : \mathcal{H}_{\widetilde{X}} \to \mathbb{R}$ being a bounded linear functional has a unique representer $\delta_\sharp \in \mathcal{H}_{\widetilde{X}}$ s.t.

$$\delta h = \langle h, \delta_\sharp\rangle_{\widetilde{X}}, \qquad \forall\, h \in \mathcal{H}_{\widetilde{X}}.$$

Therefore the adjoint $\delta^* c = c\delta_\sharp$, and

$$\delta^*\delta = \delta_\sharp \otimes_{\widetilde{X}} \delta_\sharp, \qquad \langle f, (\delta^*\delta)g\rangle_{\widetilde{X}} = (\delta f)(\delta g).$$

Combining these identities, for every $f, g \in \mathcal{H}_{\widetilde{X}}$,

$$\langle f, \mathcal{S}g\rangle_{\widetilde{X}} = \langle f, \big(\mathcal{K} + \delta^*\delta\big)g\rangle_{\widetilde{X}}.$$

Non-degeneracy of $\langle\cdot,\cdot\rangle_{\widetilde{X}}$ on $\mathcal{H}_{\widetilde{X}}$ therefore gives

$$\mathcal{S} = \mathcal{K} + \delta^*\delta.$$

Rank and trace. Finally, the cross-product $\delta^*\delta = \delta_\sharp \otimes_{\widetilde{X}} \delta_\sharp \succcurlyeq 0$ has image contained in $\operatorname{span}(\delta_\sharp)$, and its only possible nonzero eigenvalue is $\|\delta_\sharp\|^2_{\widetilde{X}}$. Hence,

$$\operatorname{rank}(\delta^*\delta) \le 1, \qquad \operatorname{tr}(\delta^*\delta) = \|\delta_\sharp\|^2_{\widetilde{X}}.$$

Moreover,

$$D_{\mathcal{B}}(p_X, p_{\widetilde{X}}) = \sup_{h\in\mathcal{B}} |\delta h| = \sup_{\|h\|_{\widetilde{X}}\le 1} |\langle h, \delta_\sharp\rangle_{\widetilde{X}}| = \|\delta_\sharp\|_{\widetilde{X}}.$$

by Cauchy–Schwarz, attained at $h = \delta_\sharp / \|\delta_\sharp\|_{\widetilde{X}}$. Therefore,

$$\operatorname{tr}(\delta^*\delta) = D^2_{\mathcal{B}}(p_X, p_{\widetilde{X}}).$$

□

G.11 Proof of Lem. 4—Average error as a worst-case surrogate

Lemma 4 (avg. error as worst-case surrogate)**.** *Let $\mathcal{H}(\boldsymbol{x}) \subset \mathbb{R}$ be a bounded interval at query $\boldsymbol{x}$. Under a squared-loss $\ell(a,b) \coloneqq (a-b)^2$, the following sandwich is tight in multiplicative constant 4:*

$$\overline{E}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x})) \;\le\; E^{+}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x})) \;\le\; 4 \cdot \overline{E}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x})).$$

Furthermore, under Asm. 1c and 2 the following inequality holds for the k-dim. ball $\mathcal{H}$ from Lem. 2:

$$\overline{E}(\mathcal{H}) \;\le\; E^{+}(\mathcal{H}) \;\le\; (k+3) \cdot \overline{E}(\mathcal{H}).$$

Proof. The lower bound is trivial (max. dominates mean). For upper bounds: (i) write closed forms of both interval metrics in center offset and half-width, (ii) compute a maximizer for their ratio, and (iii) derive a common bound for k-dim. ellipsoids.

Closed forms. Write $\mathcal{H}(\boldsymbol{x}) = [c-w, c+w]$ with interval center c and half-width $w \ge 0$. Let $d \coloneqq |c - h_\star(\boldsymbol{x})|$ be the center's offset from the truth. Under sq.-loss, the pointwise excess risk of a candidate y is $E_{\boldsymbol{x}}(y) = (y - h_\star(\boldsymbol{x}))^2$, so worst-case is attained at the endpoint farther from $h_\star(\boldsymbol{x})$,

$$E^{+}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x})) = \max_{|u|\le w}(d+u)^2 = (d+w)^2,$$

while a uniform draw decomposes into offset plus uniform second moment,

$$\overline{E}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x})) = \mathbb{E}_{u\sim\mathrm{Unif}[-w,w]}[(d+u)^2] = d^2 + \frac{1}{3}w^2. \qquad (\mathbb{E}[u]=0 \text{ kills cross term.})$$

Upper bound. Take $w > 0$, as otherwise the metrics are equal. Write their ratio in $t \coloneqq d/w \ge 0$,

$$\frac{E^{+}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x}))}{\overline{E}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x}))} = \frac{(t+1)^2}{t^2+1/3} =: f(t), \quad \Longrightarrow \quad f'(t) = \frac{2(t+1)(1/3-t)}{(t^2+1/3)^2},$$

so f increases on $t \in [0, 1/3]$ and decreases thereafter, with maximum

$$f(1/3) = \frac{(4/3)^2}{(1/3)^2+1/3} = \frac{16/9}{4/9} = 4.$$

Hence, $E^{+}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x})) \le 4 \cdot \overline{E}_{\boldsymbol{x}}(\mathcal{H}(\boldsymbol{x}))$, with equality exactly at the offset $d = w/3$.

k-dim. ellipsoids. Write an $L_2(X)$-ellipsoid $\mathcal{H} = c + \mathcal{A}\mathcal{B}_2^k$, with center c and k-dim. ball $\mathcal{B}_2^k$. Let

$$d \coloneqq \|c - h_\star\|_{L^2(X)}, \qquad v \coloneqq \frac{\mathrm{tr}(\mathcal{A}^*\mathcal{A})}{k+2}, \qquad R \coloneqq \sup_{\|u\|_2\le 1} \|\mathcal{A}u\|_{L^2(X)}.$$

with v the average sq.-radius of a centered $L_2(X)$-ellipsoid. Hence, for $U \sim \mathrm{Unif}(\mathcal{B}_2^k)$,

$$\overline{E}(\mathcal{H}) = \mathbb{E}\Big[\|c + \mathcal{A}U - h_\star\|^2_{L^2(X)}\Big] = d^2 + v.$$

Therefore,

$$\begin{aligned}
E^{+}(\mathcal{H}) &\le (d+R)^2, \\
&= d^2 + R^2 + 2dR, \\
&\le d^2 + R^2 + \frac{d^2R^2}{v} + v, && (\text{Cauchy-Schwarz } 2ab \le a^2+b^2 \text{ on } a = \tfrac{dR}{\sqrt{v}}, b = \sqrt{v}.) \\
&= \left(1 + \frac{R^2}{v}\right)(d^2+v), \\
&\le (1+(k+2))(d^2+v), && (\text{Pythagoras } R^2 \le \mathrm{tr}(\mathcal{A}^*\mathcal{A}) = (k+2)v.) \\
&= (k+3)\cdot\overline{E}(\mathcal{H}).
\end{aligned}$$

The bound is sharp: let one semi-axis be 1, and remaining $k-1$ semi-axes tend to 0. Place $h_\star$ on the longest-axis line at distance v/R from the center. The ratio $E^{+}(\mathcal{H}) \Big/ \overline{E}(\mathcal{H})$ then tends to $k+3$. □

G.12 Proof of Lem. 5—Reynolds projection with DA

Lemma 5 (Reynolds projection with DA). *Under Asms. 1a–1b, and 4, post-DA ERM $h_{\widetilde{\text{erm}}}$ satisfies*

$$h_{\widetilde{\text{erm}}} = \mathcal{T}h_{\text{erm}} \in \mathcal{H}_{\text{inv}},$$

where $\mathcal{T}$ is the $L_2(X)$-orthogonal proj. onto the ambient inv. subspace $\mathcal{I}$ with $\mathcal{H}_{\text{inv}} = \mathcal{H} \cap \mathcal{I}$ (cf. §D). Furthermore, when the target $h_\star$ is $\mathcal{T}$-invariant, the post-DA bias $\tilde{b} \coloneqq h_{\widetilde{\text{erm}}} - h_\star$ similarly follows

$$\tilde{b} = \mathcal{T}b.$$

Proof. We structure the proof in two parts: (i) the conditional expectation over augmentation acts as a proj. onto $\mathcal{I}$; (ii) both the post-DA ERM and bias naturally follow this conditional expectation.

Reynolds projection with DA. Take the operator $(\mathcal{R}h)(\cdot) \coloneqq \mathbb{E}[h(X) \mid \widetilde{X} = \cdot]$ on $h \in L_2^0(X)$, defined $p_{\widetilde{X}}$-a.e., hence p_X-a.e. (Asm. 1b). $\mathcal{R}$ coincides with projection $\mathcal{T}$ under Asm. 4, $\mathcal{R}h = \mathcal{T}h$. To prove this, we need to show that it is invariant $\mathcal{R}h \in \mathcal{I}$, and its residual $h - \mathcal{R}h$ is orthogonal to $\mathcal{I}$.

(a) Invariance: Since group $\mathcal{T}$ is compact, the Haar measure p_T is left-invariant, i.e. for any $\boldsymbol{\tau} \in \mathcal{T}$,

$$\begin{aligned}
& \boldsymbol{\tau}T \overset{d}{=} T, && \text{(Def. of left-inv. } p_T\text{, follows from Asm. 4.)} \\
\Rightarrow \quad & (\boldsymbol{\tau}T, X) \overset{d}{=} (T, X), && \text{(Follows from } T \perp\!\!\!\perp X\text{.)} \\
\Rightarrow \quad & (\boldsymbol{\tau}\widetilde{X}, h(X)) \overset{d}{=} (\widetilde{X}, h(X)), && \text{(Apply } (\boldsymbol{\tau}, \boldsymbol{x}) \mapsto (\alpha(\boldsymbol{x}, \boldsymbol{\tau}), h(\boldsymbol{x})) \text{ on both sides.)} \\
\Rightarrow \quad & \mathbb{E}[h(X) \mid \boldsymbol{\tau}\widetilde{X}] = \mathbb{E}[h(X) \mid \widetilde{X}], && \text{(Conditional expectation on both sides.)} \\
\Rightarrow \quad & (\mathcal{R}h)(\boldsymbol{\tau}^{-1}\widetilde{X}) = (\mathcal{R}h)(\widetilde{X}). && \text{(Re-parameterize to } \mathcal{R}h\text{. } \mathcal{T} \text{ group} \Rightarrow \boldsymbol{\tau}^{-1} \text{ exists.)}
\end{aligned}$$

Because $\boldsymbol{\tau}$ is arbitrary, it follows that $(\mathcal{R}h)(X) = (\mathcal{R}h)(\widetilde{X})$, where the left-hand evaluation under p_X is again licensed by Asm. 1b. Also, $\mathbb{E}[(\mathcal{R}h)(\widetilde{X})] = \mathbb{E}[h(X)] = 0$, hence $\mathcal{R}h \in \mathcal{I}$. Immediately, if

$$\mathcal{R}h \in \mathcal{H} \quad \Longrightarrow \quad \mathcal{R}h \in \mathcal{H}_{\text{inv}}.$$

(b) Orthogonality: The residual $h - \mathcal{R}h$ is $L_2(X)$-orthogonal to every invariant test function $g \in \mathcal{I}$:

$$\begin{aligned}
\langle \mathcal{R}h, g \rangle_{L_2(X)} \coloneqq \mathbb{E}[(\mathcal{R}h)(X) \cdot g(X)] &= \mathbb{E}[(\mathcal{R}h)(\widetilde{X}) \cdot g(\widetilde{X})], && \text{(Both } \mathcal{R}h \text{ and } g \text{ are } \mathcal{T}\text{-inv.)} \\
&= \mathbb{E}[\mathbb{E}[h(X) \,|\, \widetilde{X}] \cdot g(\widetilde{X})], && \text{(Def. of } \mathcal{R}\text{.)} \\
&= \mathbb{E}[\mathbb{E}[h(X) \cdot g(\widetilde{X}) \,|\, \widetilde{X}]], && (g(\widetilde{X}) \text{ is } \widetilde{X}\text{-measurable.)} \\
&= \mathbb{E}[h(X) \cdot g(\widetilde{X})], && \text{(Tower property.)} \\
&= \mathbb{E}[h(X) \cdot g(X)] = \langle h, g \rangle_{L_2(X)}. && (\mathcal{T}\text{-inv. of } g\text{.)}
\end{aligned}$$

Subtracting the two sides gives $\langle h - \mathcal{R}h, g \rangle_{L_2(X)} = 0$ for all $g \in \mathcal{I}$. Together with part *(a)*, $\mathcal{R}h \in \mathcal{I}$ and $h - \mathcal{R}h \perp_{L_2(X)} \mathcal{I}$, therefore for all $h \in L_2^0(X)$,

$$\mathcal{R}h = \mathcal{T}h.$$

Application to ERM and Bias. Under Asm. 1a, $h_{\text{erm}}(X) = \mathbb{E}[Y \,|\, X]$ and $h_{\widetilde{\text{erm}}}(\widetilde{X}) = \mathbb{E}[Y \,|\, \widetilde{X}]$. Since Y is centered, all of $h_\star, h_{\text{erm}}, h_{\widetilde{\text{erm}}}$ (hence $b, \tilde{b}$) are centered. Further, $T \perp\!\!\!\perp (X, Y, U, \xi)$ by construction, so $Y - X - \widetilde{X}$ is a Markov chain. Via the tower rule,

$$\begin{aligned}
h_{\widetilde{\text{erm}}}(\widetilde{X}) = \mathbb{E}[Y \,|\, \widetilde{X}] &= \mathbb{E}\Big[\mathbb{E}[Y \,|\, X, \widetilde{X}] \,\Big|\, \widetilde{X}\Big], \\
&= \mathbb{E}\Big[\mathbb{E}[Y \,|\, X] \,\Big|\, \widetilde{X}\Big] = \mathbb{E}[h_{\text{erm}}(X) \,|\, \widetilde{X}] = (\mathcal{R}h_{\text{erm}})(\widetilde{X}).
\end{aligned}$$

Note that $h_{\widetilde{\text{erm}}}$ and $\mathcal{R}h_{\text{erm}}$ are $\mathcal{T}$-inv., hence they must also be identically equal in $L_2(X)$ (Asm. 1b), giving $h_{\widetilde{\text{erm}}} = \mathcal{R}h_{\text{erm}} = \mathcal{T}h_{\text{erm}}$ by part *(b)*. And since $h_{\widetilde{\text{erm}}} \in \mathcal{H}_X$ by Asm. 1a, part *(a)* places it in the class: $h_{\widetilde{\text{erm}}} \in \mathcal{H}_{\text{inv}}$. Finally, for a $\mathcal{T}$-inv $h_\star \in \mathcal{I}$, using $\mathcal{T}h_\star = h_\star$ and linearity of $\mathcal{T}$ on $L_2^0(X)$,

$$\mathcal{T}b \quad = \quad \mathcal{T}h_{\text{erm}} - \mathcal{T}h_\star \quad = \quad h_{\widetilde{\text{erm}}} - h_\star \quad = \quad \tilde{b}. \qquad \square$$

G.13 Proof of Lem. 6—Average Representer Norms

Lemma 6 (avg. representer norms)**.** *Under Asms. 1b–1c, evaluation $h \mapsto h(\boldsymbol{x})$ admits the $L_2(X)$-representer $g_{\boldsymbol{x}} \in \mathcal{H}_X$ and the $L_2(\widetilde{X})$-representer $\tilde{g}_{\boldsymbol{x}} \in \mathcal{H}_{\widetilde{X}}$. Averaged over queries $X \sim p_X$ gives*

$$\mathbb{E}_X\Big[\,\|g_X\|^2_{L_2(X)}\Big] = k, \qquad\qquad \mathbb{E}_X\Big[\,\|\tilde{g}_X\|^2_{L_2(\widetilde{X})}\Big] = \operatorname{tr}(\mathcal{S}).$$

Proof. Two steps give the result: (i) **expand** each representer in an orthonormal basis of its own centered class; (ii) **average** over $X \sim p_X$, reading each summand as a quadratic form of the shift operator.

Expansion. Fix an $L_2(\widetilde{X})$-orthonormal basis $\{\psi_i\}_{i=1}^k$ of the class $\mathcal{H}_{\widetilde{X}}$ (Asms. 1b–1c, cf. §G). Then,

$$\begin{aligned}
\tilde{g}_{\boldsymbol{x}} &= \sum_{i=1}^k \langle \tilde{g}_{\boldsymbol{x}}, \psi_i\rangle_{L_2(\widetilde{X})} \cdot \psi_i, && \text{(Parseval in } \langle\cdot,\cdot\rangle_{L_2(\widetilde{X})}.)\\
&= \sum_{i=1}^k \psi_i(\boldsymbol{x}) \cdot \psi_i, && \text{(Reproducing identity } \langle \psi_i, \tilde{g}_{\boldsymbol{x}}\rangle_{L_2(\widetilde{X})} = \psi_i(\boldsymbol{x}).)\\
\Longrightarrow \quad \|\tilde{g}_{\boldsymbol{x}}\|^2_{L_2(\widetilde{X})} &= \sum_{i=1}^k \psi_i(\boldsymbol{x})^2. && \text{(Orthonormality of } \psi_i.)
\end{aligned}$$

Averaging. Take expectations over $X \sim p_X$ in each summand,

$$\mathbb{E}_X\Big[\,\psi_i(X)^2\Big] = \|\psi_i\|^2_{L_2(X)} = \langle \psi_i, \mathcal{S}\psi_i\rangle_{L_2(\widetilde{X})}, \qquad \text{(Def. of } \mathcal{S}.)$$

and sum over the basis to obtain a trace,

$$\mathbb{E}_X\Big[\,\|\tilde{g}_X\|^2_{L_2(\widetilde{X})}\Big] = \sum_{i=1}^k \langle \psi_i, \mathcal{S}\psi_i\rangle_{L_2(\widetilde{X})} = \operatorname{tr}(\mathcal{S}). \qquad \text{(Trace in any orthonormal basis.)}$$

Repeating the two steps in an $L_2(X)$-orthonormal basis $\{\psi_i\}_{i=1}^k$ of $\mathcal{H}_X$ replaces $\mathcal{S}$ by Id throughout. The averaged summand is then $\|\psi_i\|^2_{L_2(X)} = 1$, now resulting in $\mathbb{E}_X\Big[\,\|g_X\|^2_{L_2(X)}\Big] = \operatorname{tr}(\mathrm{Id}) = k$. □