Achievable distributional robustness when the robust risk is only partially identified

Julia Kostin
Department of Computer Science
ETH Zurich
jkostin@ethz.ch

Nicola Gnecco
Gatsby Computational Neuroscience Unit
University College London
nicola.gnecco@gmail.com

Fanny Yang
Department of Computer Science
ETH Zurich
fan.yang@inf.ethz.ch

Abstract

In safety-critical applications, machine learning models should generalize well under worst-case distribution shifts, that is, have a small robust risk. Invariance-based algorithms can provably take advantage of structural assumptions on the shifts when the training distributions are heterogeneous enough to identify the robust risk. However, in practice, such identifiability conditions are rarely satisfied – a scenario so far underexplored in the theoretical literature. In this paper, we aim to fill the gap and propose to study the more general setting of *partially identifiable robustness*. In particular, we define a new risk measure, the identifiable robust risk, and its corresponding (population) minimax quantity that is an algorithm-independent measure for the best achievable robustness under partial identifiability. We introduce these concepts broadly, and then study them within the framework of linear structural causal models for concreteness of the presentation. We use the introduced minimax quantity to show how previous approaches provably achieve suboptimal robustness in the partially identifiable case. We confirm our findings through empirical simulations and real-world experiments and demonstrate how the test error of existing robustness methods grows increasingly suboptimal as the proportion of previously unseen test directions increases.

1 Introduction

The success of machine learning methods typically relies on the assumption that the training and test data follow the same distribution. However, this assumption is often violated in practice. For instance, this can happen if the test data are collected at a later time or using a different measuring device. Without further assumptions on the test distribution, generalization under distribution shift is impossible. However, practitioners often have partial information about the set of possible "shifts" that may occur during test time, inducing a set of *feasible test distributions* that the model should generalize to. We refer to the resulting set as the *robustness set*. With $\mathcal{R}(\beta; \mathbb{P})$ denoting the population risk of a model parameterized by β for distribution $\mathbb{P}$, the robust risk can be written as

$$\mathcal{R}_{\mathrm{rob}}(\beta) := \sup_{\mathbb{P} \in \mathcal{P}_{\mathrm{rob}}(\theta^\star)} \mathcal{R}(\beta; \mathbb{P}), \tag{1}$$

where $\mathcal{P}_{\mathrm{rob}}(\theta^\star)$ corresponds to the robustness set that we assume to be fully characterized by some true parameter $\theta^\star$. In safety-critical applications, the goal is often to find a minimizer of the robust

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Comparison of various distributional robustness frameworks and what kind of assumptions their analysis can account for (with an incomplete list of examples for each framework).

Framework accounts for	bounded shifts	partial identifiability of causal parameters	partial identifiability of robustness set
DRO [7, 16, 53, 35, 47]	✓	–	✗
Infinite robustness [38, 18, 33, 42, 6, 2, 51, 57, 29, 1]	✗	✗	✗
Finite robustness [45, 26, 14, 28, 49]	✓	✓	✗
Partially id. robustness (this work)	✓	✓	✓

risk, i.e. a *robust prediction model* that shows the best performance on the *worst-case* distribution out of the robustness set.

A number of subfields in machine learning and optimization have addressed this problem. For example, in distributionally robust optimization (DRO) [7, 16], the parameter $\theta^\star$ may be the training distribution $\mathbb{P}$ and the robustness set the *neighborhood* of $\mathbb{P}$ in some probability distance metric [30, 35, 22, 15]. Relatedly, adversarial robustness [23, 32] studies the risk on worst-case transformations of examples drawn from some distribution $\mathbb{P}$ and can be seen as equivalent to distribution shift robustness [53]. DRO-type methods minimize the worst-case robustness against arbitrary distribution shifts in the neighborhood without structural assumptions. Although being assumption-agnostic can be viewed as a strength, it also has its caveat: even when available, prior knowledge about the structure of expected test shifts cannot be incorporated. In such cases, the robust model's prediction might be overly conservative, resulting in suboptimal performance when the test shifts are in fact more benign. [47].

In many practical scenarios, data from *heterogeneous sources* is available at training time – for example, data from different geographic locations or time ranges. Due to the lack of modeling assumptions, multiple environments in the DRO setting cannot, in general, be leveraged to achieve better robustness in a given robustness set – in those contexts, the presence of multiple environments is usually argued to enable robustness against a larger robustness set. Instead, domain experts can anticipate which aspects of the joint probability distribution of (X, Y) are more likely to shift. Such prior structural information can, for example, be incorporated through the framework of structural causal models (SCMs), via the approach of *causality-oriented robustness* [34, 11]. Importantly, the literature in this area has so far focused on settings when the desired robust objective $\mathcal{R}_{\mathrm{rob}}$ is identifiable, i.e. computable from training data. Traditional causal learning and invariance-based methods aim to fully identify some underlying causal parameter of the SCM for robustness against *all* (potentially infinite) interventions [38, 42, 6, 29]. However, the training data is often not heterogeneous enough to fully identify the causal parameter. Thus, another line of work [45, 50, 28] focuses on the scenario when the causal parameter is not necessarily identifiable, but the test shifts only occur in training directions, rendering the robust risk (1) identifiable. We provide an overview in Table 1 and an additional discussion of related work in Appendix A.

In practice, causality- and invariance-based methods often result in wrong representations of the data [27, 44] and end up performing similarly to empirical risk minimization (ERM) that ignores the multi-environment information [3, 24, 43]. Many possible explanations for this observation have been proposed in the literature. In our work, we focus on the *non-identifiability* failure scenario. In particular, we extend the discussion of invariance-based methods to include the partially identifiable setting, where not only the causal parameter, but the robust risk (1) is not determinable using training data either[1]. Specifically, we aim to discuss the following question:

What is the optimal worst-case performance any model can have for given structural relationships between test and training data and how do existing methods comparatively perform in such settings?

[1] Here, we mean partial identifiability of the robust risk, which is reminiscent of outputting uncertainty sets for a quantity of interest in the field of partial identification [54, 19].

When the robust risk is not identifiable from training data, we obtain a whole *set* of possible objectives that includes the true robust risk. In this case, we are interested in the best achievable robustness for *any algorithm* that we capture in a quantity called the *identifiable robust risk*:

$$\mathcal{R}_{\text{rob,ID}}(\beta) := \sup_{\substack{\text{possible} \\ \text{true model } \theta^\star}} \sup_{\mathbb{P} \in \mathcal{P}_{\text{rob}}(\theta^\star)} \mathcal{R}(\beta; \mathbb{P}). \tag{2}$$

Note that $\mathcal{R}_{\text{rob,ID}}(\beta)$ is well-defined even when the standard robust risk is not identifiable – it takes the supremum over the robust risks induced by all possible true model parameters $\theta^\star$ that are consistent with the given set of training data distributions. Furthermore, the minimal value of the identifiable robust risk corresponds to the optimal worst-case performance in the partially identifiable setting. Spiritually, this *minimax population quantity* is reminiscent of the algorithm-independent limits in classical statistical learning theory [58].[2] Even though our partial identifiability framework can be evaluated for arbitrary modeling assumptions on the distribution shift (such as covariate/label shift, DRO, etc.), we present it in a concrete setting for clarity of the exposition. Specifically, we discuss linear structural causal models (SCMs) with unobserved confounding (cf. Section 2), similar to the setting of IV (instrumental variables) and anchor regression [45, 46].

The identifiable robust risk (2) not only represents a notion of algorithm-independent optimality for any combination of training and test shifts. In the linear SCM setting in Section 2, we also show theoretically and empirically that the ranking and optimality of different robustness methods change drastically in identifiable vs. partially identifiable settings. The same can be observed in experiments on real-world data. Our experimental results strongly indicate that evaluation and benchmarking on partially identifiable settings are important for determining the effectiveness of robustness methods. Finally, while the identifiable robust predictor is only provably optimal for the linear SCM, experiments on real-world data in Section 3.3 suggest that our estimator may significantly improve upon other invariance-based methods in more realistic scenarios.

2 Setting

In this section, we first introduce the linear causal model setting and describe our structural assumptions on the training and test distributions. Then, we introduce our framework for distributional robustness that allows for partial identifiability and define the *identifiable robust risk*, the worst-case robust risk among the possible robust risks induced by the training distributions.

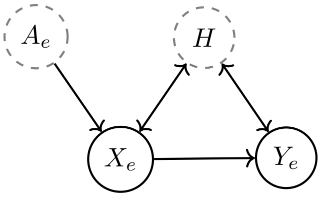
2.1 Data distribution and a model of additive environmental shifts

Data distribution. We are given multiple training environments indexed by $e \in \mathcal{E}_{\text{train}}$, where $\mathcal{E}_{\text{train}}$ is a countable environment index set. For each training environment e, we observe data $(X_e, Y_e) \sim \mathbb{P}_e^{X,Y}$ consisting of input covariates $X_e \in \mathbb{R}^d$ and the target variable $Y_e \in \mathbb{R}$, which are generated by the linear structural causal model (SCM) (3) and its corresponding causal graph, depicted in Figure 1. Throughout the paper, we assume that we observe the collection of training distributions $\mathcal{P}_{\theta^\star, \mathcal{E}_{\text{train}}} = \{\mathbb{P}_{\theta^\star, e}^{X,Y}\}_{e \in \mathcal{E}_{\text{train}}}$, omitting $\theta^\star$ when it is clear from the context. We discuss the corresponding finite sample setting in Appendix D.

We assume that the true unobserved parameters $\theta^\star := (\beta^\star, \Sigma^\star) \in \Theta$, where $\Theta \subset \mathbb{R}^{d+(d+1)(d+1)}$, are invariant across environments. The joint covariance $\Sigma^\star$ of the noise (η, ξ) can be written in block form as $\Sigma^\star = \begin{pmatrix} \Sigma_\eta^\star & \Sigma_{\eta,\xi}^\star \\ \Sigma_{\eta,\xi}^{\star\top} & (\sigma_\xi^\star)^2 \end{pmatrix}$. We allow the presence of latent confounders between X_e and Y_e, and hence the case where $\Sigma_{\eta,\xi}^\star \neq 0$. Note that the *confounded noise setting* is, in general, more challenging than the independent noise setting, since, given any number of environments, common estimators such as the linear regression estimator are biased away from the causal parameter $\beta^\star$.

Additive distribution shift. The distribution shift between environments is modeled by the (random) additive shift $A_e \in \mathbb{R}^d$ with mean $\mathbb{E}[A_e] = \mu_e$ and covariance matrix $\text{Cov}[A_e] = \Sigma_e$. In

[2] In particular, extending (2) to its finite-sample counterpart would introduce a more natural extension of the classical minimax risk statistical learning theory. In this work, we focus on identifiability aspects instead of statistical rates.



$$\begin{aligned} A_e &\sim \mathbb{P}_e^A; \\ (\eta, \xi) &\sim \mathcal{N}(0, \Sigma^\star); \\ X_e &= A_e + \eta; \\ Y_e &= \beta^{\star\top} X_e + \xi. \end{aligned} \tag{3}$$

Figure 1: (Left) Causal graph corresponding to the SCM in Equation (3). Observed variables (X_e, Y_e) are indicated by solid circles while unobserved variables, namely the additive shift A_e and confounders H, are shown in dashed circles. Note that here, bidirectional edges indicate that the relationship between two nodes can be in either direction.

general, the environment shifts A_e can be degenerate, i.e. the covariances Σ_e are not assumed to be full rank. For simplicity of presentation, we further assume that $\mathcal{E}_{\text{train}}$ contains a reference environment $e = 0$ satisfying $\mu_0 = 0$ and $\Sigma_0 = 0$. In Appendix B, we discuss how our results apply if this condition is not met. Our additive shift structure implies that the joint distribution $\mathbb{P}_e^{X,Y,A} = \mathbb{P}_e^A \times \mathbb{P}^{X,Y|A}$ changes in each environment. However, we do not allow for direct interventions on Y or the latent confounders, that is $\mathbb{P}^{X,Y|A}$ remains invariant. Note that our distribution shift setting is *more general than covariate shift*: due to unobserved confounding, the conditional distribution $\mathbb{P}_e^{Y|X}$ also varies across environments.

In summary, our model (3) describes a multi-environment setting where different training distributions vary by changing the distribution $\mathbb{P}_e^A$ of the random additive shifts A_e, but the causal model parameters $\theta^\star$ remain invariant across all training and test environments. It can model a variety of multi-environment settings in the related literature. For instance, choosing $|\mathcal{E}_{\text{train}}| = 1$ and $A \sim \mathcal{N}(0, M\Sigma_A M^\top)$, where $M \in \mathbb{R}^{d\times q}, \Sigma_A \in \mathbb{R}^{q\times q}$, yields the setup of continuous anchor regression [45][3]. Discrete anchor regression [45] corresponds to a discrete environment index set $\mathcal{E}_{\text{train}} = [m], m \in \mathbb{N}$, and deterministic mean shifts $A_e = \mu_e \in \mathbb{R}^d$. The more general additive shift setting in [49] corresponds to $\mathcal{E}_{\text{train}} = [m]$ and $A_e \sim \mathcal{N}(\mu_e, \Sigma_e)$. Note that in the above works, the environment index is modeled as a random variable $E \sim \mathbb{P}^E$ through which one assigns weights to different training environments. The results in this paper only depend on the support of $\mathbb{P}^E$, that is, whether an environment was seen or not. Thus, the population-level guarantees – the focus of this paper – are the same for any distributions with the same support on $\mathbb{P}^E$.

Structural assumptions on test distribution shift. During test time, we expect to observe data that follows a new, previously unseen distribution $\mathbb{P}_{\text{test}}^{X,Y}$. The test data are generated by the same SCM (3), but with a new additive shift $A_{\text{test}} \sim \mathbb{P}_{\text{test}}^A$ with corresponding finite mean μ_{test} and covariance Σ_{test}. Even though we do not observe $\mathbb{P}_{\text{test}}^{X,Y}$ during training, we do assume partial knowledge about the directions and sizes of possible distributions of the shift variable $\mathbb{P}_{\text{test}}^A$, that is

$$\mathbb{E}\left[A_{\text{test}} A_{\text{test}}{}^\top\right] = \Sigma_{\text{test}} + \mu_{\text{test}} \mu_{\text{test}}{}^\top \preceq M_{\text{test}}; \tag{4}$$

where $M_{\text{test}} \succeq 0$. If the test distribution of X is given (as in the *domain adaptation* setting), one can directly estimate the shift of the test environment and set $M_{\text{test}} := \mathbb{E}\left[A_{\text{test}} A_{\text{test}}{}^\top\right]$[4]. In the following, we consider the distributional robustness setting in which *partial knowledge* about test shifts is given in form of their *maximum strength* γ and *general direction* $\mathcal{M} \subseteq \mathbb{R}^d$. We can then formalize this partial knowledge by setting $M_{\text{test}} = \gamma \Pi_{\mathcal{M}}$, where $\gamma > 0$ and $\Pi_{\mathcal{M}}$ is an orthogonal projection onto the subspace $\mathcal{M}$.[5]

2.2 Classical distributional robustness

Given the test shift directions $\mathcal{M}$ and strength γ, our goal is to find an estimator using the training data that has a small risk over the entire set of shifted test distributions, called the *robustness set*, that

[3] Note that in the anchor regression setup, the environment shift A, called *anchor*, is also observed, thus the training data consist of (A, X, Y).

[4] Even though we do not observe A_{test}, the structural assumptions on A_{test} correspond to assumptions on $\mathbb{E}\left[X^{\text{test}} X^{\text{test}\top}\right] - \mathbb{E}\left[X_0 X_0^\top\right]$.

[5] When more refined information on the test shifts is given by a general PSD matrix $M_{\text{test}} \in \mathbb{R}^{d\times d}$, we can replace it by the upper bound $M_{\text{test}} \preceq \lambda_{\max}(M_{\text{test}}) \Pi_{\text{range}(M_{\text{test}})}$ and apply our results on the upper bound.

we define as

$$\mathcal{P}_{\theta^\star}(\gamma\Pi_{\mathcal{M}}) := \{\mathbb{P}^{X,Y}_{\theta^\star,\text{test}} : \mathbb{E}\left[A_{\text{test}}{A_{\text{test}}}^\top\right] \preceq \gamma\Pi_{\mathcal{M}}\}. \tag{5}$$

For a given robustness set $\mathcal{P}_{\theta^\star}(\gamma\Pi_{\mathcal{M}})$, we define the *robust risk*

$$\mathcal{R}_{\text{rob}}(\beta;\theta^\star,\gamma\Pi_{\mathcal{M}}) := \sup_{\mathbb{P}\in\mathcal{P}_{\theta^\star}(\gamma\Pi_{\mathcal{M}})} \mathcal{R}(\beta;\mathbb{P}), \tag{6}$$

and the corresponding *robust predictor* $\beta^{rob} := \arg\min_{\beta\in\mathbb{R}^d} \mathcal{R}_{\text{rob}}(\beta;\theta^\star,\gamma\Pi_{\mathcal{M}})$, where $\mathcal{R}(\beta;\mathbb{P}) := \mathbb{E}_{\mathbb{P}}[(Y-\beta^\top X)^2]$ denotes the risk w.r.t. to the squared loss. The *robust risk* and the *robust predictor* can be explicitly computed as a function of the true model parameters $\theta^\star = (\beta^\star,\Sigma^\star)$ and the test shift bound $\gamma\Pi_{\mathcal{M}}$:

$$\beta^{rob}_{\theta^\star} = \arg\min_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathcal{P}_{\theta^\star}(\gamma\Pi_{\mathcal{M}})} \mathbb{E}_{\mathbb{P}}[(Y-\beta^\top X)^2] = \beta^\star + (\gamma\Pi_{\mathcal{M}} + \Sigma^\star_\eta)^{-1}\Sigma^\star_{\eta,\xi}. \tag{7}$$

In practice, neither the model parameters $\theta^\star$ nor the robust risk $\mathcal{R}_{\text{rob}}$ itself can generally be fully identified. Instead, in the next sections we show that the robust prediction model can usually only be *set-identified*, except for specific combinations of the training and test shifts.

2.3 Partially identifiable robustness framework

In this section, we formally introduce the identifiable robust risk and related notions that allow us to characterize model robustness in the case when the robust predictor cannot be identified. We start with the notion of *observational equivalence* [17]:

Definition 1 (Observational equivalence)**.** *We call model parameters $\theta_1 = (\beta_1,\Sigma_1)$ and $\theta_2 = (\beta_2,\Sigma_2)$* ***observationally equivalent*** *with respect to a set of shift distributions $\{\mathbb{P}^A_e : e\in\mathcal{E}_{\text{train}}\}$*[6] *if they induce the same set $\mathcal{P}_{\theta,\mathcal{E}_{\text{train}}}$ of training distributions over the observed variables (X_e,Y_e) as described in Section 2.1, i.e.*

$$\textit{For all } e\in\mathcal{E}_{\text{train}} : \mathbb{P}^{X,Y}_{\theta_1,e} = \mathbb{P}^{X,Y}_{\theta_2,e}.$$

By observing $\mathcal{P}_{\theta^\star,\mathcal{E}_{\text{train}}}$, we can identify the model parameters up to the ***observationally equivalent set*** *defined as*

$$\Theta_{\text{eq}} := \{\theta = (\beta,\Sigma)\in\Theta : \mathcal{P}_{\theta,\mathcal{E}_{\text{train}}} = \mathcal{P}_{\theta^\star,\mathcal{E}_{\text{train}}}\}.$$

When the observationally equivalent set is not a singleton, prior work only considers scenarios where the robustness set (5) and hence also the robust prediction model are still identifiable (see Equation (7)). This scenario is shown in Figure 2a and discussed again in Section 3.2. In general, however, the observation of multiple training environments $\mathcal{P}_{\theta^\star,\mathcal{E}_{\text{train}}}$ neither identifies the model parameters nor the robustness set or robust risk, which is the partially identified setting that we focus on. Instead, we can only compute a superset of the robustness set

$$\mathcal{P}_{\Theta_{\text{eq}}}(\gamma\Pi_{\mathcal{M}}) := \bigcup_{\theta\in\Theta_{\text{eq}}} \mathcal{P}_\theta(\gamma\Pi_{\mathcal{M}}) \supset \mathcal{P}_{\theta^\star}(\gamma\Pi_{\mathcal{M}})$$

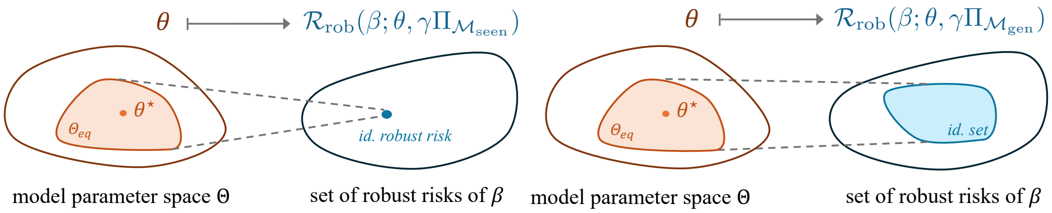
and correspondingly, a set of robust risks $\{\mathcal{R}_{\text{rob}}(\beta;\theta,\gamma\Pi_{\mathcal{M}}) : \theta\in\Theta_{\text{eq}}\}$ and robust predictors $\mathcal{B}^{rob}_{\Theta_{\text{eq}}} := \{\beta^{rob}_\theta : \theta\in\Theta_{\text{eq}}\}$. In this case, we would still like to achieve the "best-possible" robustness, which is intuitively test shift robustness for the "hardest-possible" parameters that could have induced the observed training distributions.

Definition 2 (Identifiable robust risk and the minimax quantity)**.** *Consider the data model (3). The identifiable robust risk is defined as*

$$\mathcal{R}_{\text{rob,ID}}(\beta;\Theta_{\text{eq}},\gamma\Pi_{\mathcal{M}}) := \sup_{\theta\in\Theta_{\text{eq}}} \mathcal{R}_{\text{rob}}(\beta;\theta,\gamma\Pi_{\mathcal{M}}). \tag{8}$$

We will denote its minimizer as $\beta^{\text{rob,ID}}$ and refer to it as the identifiable robust predictor. The optimal robustness on test shifts bounded by $\gamma\Pi_{\mathcal{M}}$ given training data $\mathcal{P}_{\theta^\star,\mathcal{E}_{\text{train}}}$ is described by the minimax quantity

$$\mathfrak{M}(\Theta_{\text{eq}},\gamma\Pi_{\mathcal{M}}) = \inf_{\beta\in\mathbb{R}^d} \mathcal{R}_{\text{rob,ID}}(\beta;\Theta_{\text{eq}},\gamma\Pi_{\mathcal{M}}). \tag{9}$$


θ
$\mathcal{R}_{\text{rob}}(\beta;\theta,\gamma\Pi_{\mathcal{M}_{\text{seen}}})$
θ_{eq}
$\theta^\star$
id. robust risk
model parameter space Θ
set of robust risks of β
θ
$\mathcal{R}_{\text{rob}}(\beta;\theta,\gamma\Pi_{\mathcal{M}_{\text{gen}}})$
θ_{eq}
$\theta^\star$
id. set
model parameter space Θ
set of robust risks of β

(a) Identifiable setting

(b) Partially identifiable setting

Figure 2: Relationship between identifiability of the model parameters and identifiability of the robust risk. (a) The classical scenario where the test shift directions $\mathcal{M}_{\text{seen}}$ are contained in the span of training shifts so that the robust risk and thus its minimizer are point-identified. (b) The more general scenario of this paper, where the shift directions during test time $\mathcal{M}_{\text{gen}}$ can contain new shift directions and the robust risk can only be set-identified.

The definition of the identifiable robust risk reflects the absence of knowledge of the model parameters in test shift directions that were not observed during training. In words, the most robust parameter choice $\beta^{\text{rob,ID}}$ is the one that minimizes the worst-case robust risk for parameters in the observationally equivalent set . In the next sections, we compute these quantities explicitly for the setting of Section 2. This will allow us to compare the best achievable robustness in the partially identified case with the guarantees of prior methods in this setting.

3 Main results for partially identified robustness

We now compute the identifiable robust risk (8) and derive a lower bound for the minimax quantity (9) in the additive shift setting. We then compare the identifiable robust risk of existing robustness methods and ordinary least squares (OLS) with the minimizer of the identifiable robust risk both theoretically and empirically.

3.1 Minimax robustness results for the SCM

The degree to which the model parameters $\theta^\star$ in the linear SCM setting (3) can be identified depends on the number of environments and the total rank of the additive shifts. This is well-studied, for instance, in the instrumental variable (IV) regression literature [4, 9]. In particular, in the setting of Section 2.1, the causal parameter $\beta^\star$ can *only* be identified along the mean and variance shifts of the covariates across the training data. Therefore, if not enough shift directions are observed, it is merely *set-identifiable*. In the following, we show how set-identifiability of the model parameters translates into set-identifiability of the robust prediction model (7). More formally, we denote by $\mathcal{S}$ the subspace consisting of all *additive shift directions seen during training*:

$$\mathcal{S} := \text{range}\left[\sum_{e\in\mathcal{E}_{\text{train}}} \left(\Sigma_e + \mu_e\mu_e^\top\right)\right]. \tag{10}$$

The definition of the space $\mathcal{S}$ induces the following orthogonal decompositions of the causal parameter and test shift directions $\mathcal{M}$:

$$\beta^\star = \beta^{\mathcal{S}} + \beta^{\mathcal{S}^\perp}, \quad \text{and} \quad \Pi_{\mathcal{M}} \preceq SS^\top + RR^\top, \tag{11}$$

where S and R are matrices with orthonormal columns such that range $S \subset \mathcal{S}$, range $R \subset \mathcal{S}^\perp$ and range S, range R are the smallest subspaces satisfying Equation (11)[7]. The matrix S corresponds to test shift directions along the model can be identified. Conversely, R corresponds to test shift directions, along which the model is non-identified. The vector $\beta^{\mathcal{S}}$ is the *identifiable part* of the causal parameter. It uniquely defines a set of *identified model parameters* that reads

$$\theta^{\mathcal{S}} := (\beta^{\mathcal{S}}, \Sigma_\eta^{\mathcal{S}}, \Sigma_{\eta,\xi}^{\mathcal{S}}, (\sigma_\xi^{\mathcal{S}})^2) = (\beta^{\mathcal{S}}, \Sigma_\eta^\star, \Sigma_{\eta,\xi}^\star + \Sigma_\eta^\star\beta^{\mathcal{S}^\perp}, (\sigma_\xi^\star)^2 + 2\langle\Sigma_{\eta,\xi}^\star, \beta^{\mathcal{S}^\perp}\rangle + \langle\beta^{\mathcal{S}^\perp}, \Sigma_\eta^\star\beta^{\mathcal{S}^\perp}\rangle)$$

[6] In general, the distribution of A_e is unknown, since A_e is unobserved. In our setting, $\mathbb{P}_e^A$ can be identified because of the reference environment. Otherwise, one proceeds with relative shifts as described in Appendix B.

[7] The choice of S and R is not unique, but the subspaces range S, range R, which matter for our results, are.

and can be computed from the training distributions. In the next proposition, we show that the model parameters and robust predictor can be identified up to a set around $\theta^{\mathcal{S}}$, which can be interpreted as the set's geometric center. From the characterization of this set, it directly follows that the robust predictor is only identifiable if the test shifts are in the direction of the training shifts, i.e. $\mathcal{M} \subset \mathcal{S}$.

Proposition 1 (Identifiability of model parameters and robust predictor)**.** *Suppose that the set of training and test distributions is generated according to Section 2.1, with some model parameter $\theta \in \Theta$. Then, it holds that*

(a) the model parameters generating the training distribution (3) can be identified up to the following observationally equivalent set :

$$\Theta_{\text{eq}} = \Theta \cap \{\beta^{\mathcal{S}} + \alpha, \Sigma_\eta^\star, \Sigma_{\eta,\xi}^{\mathcal{S}} - \Sigma_\eta^\star \alpha, (\sigma_\xi^{\mathcal{S}})^2 - 2\alpha^\top \Sigma_{\eta,\xi}^{\mathcal{S}} + \alpha^\top \Sigma_\eta \alpha \colon \alpha \in \mathcal{S}^\perp\} \ni \theta^\star; \quad (12)$$

(b) the robust predictor β^{rob} as defined in Equation (7) is identified up to the set

$$\mathcal{B}_\Theta^{rob} \cap \{\beta^{\mathcal{S}} + (\gamma \Pi_{\mathcal{M}} + \Sigma_\eta^\star)^{-1} \Sigma_{\eta,\xi}^{\mathcal{S}} + (\gamma \Pi_{\mathcal{M}} + \Sigma_\eta^\star)^{-1} \alpha \colon \ \alpha \in \text{range } R\} \ni \beta^{rob}. \quad (13)$$

The proof of Proposition 1 is provided in Appendix F.1. Proposition 1 implies two well-known settings: If we observe a rich enough set $\mathcal{P}_{\mathcal{E}_{\text{train}}}$ of training environments such that $\mathcal{S} = \mathbb{R}^d$, the model parameters are uniquely identified, corresponding to the setting of full-rank instruments [4]. However, even in the under-identified case $\mathcal{S} \neq \mathbb{R}^d$, if the test shift directions $\mathcal{M}$ are contained in the space $\mathcal{S}$ of training-time shifts, i.e. $R = 0$, the robust prediction model is identifiable from training data *regardless of the identifiability of the model parameters*. This is the setting considered e.g. in anchor regression [45] and discussed again in Section 3.2 and Appendix C.

So far, we have described how the identifiability of the robust prediction model depends on the structure of both the training environments (via the space $\mathcal{S}$) and the test environments (via $\mathcal{M}$). We now aim to compute the smallest achievable robust loss for the general partially identifiable setting, which allows for $R \neq 0$. In particular, we provide a lower bound on the *best-possible achievable distributional robustness* formalized by the minimax quantity (9). First observe that without further assumptions on the parameter space Θ, the observationally equivalent set is unbounded, and the identifiable robust risk (8) can be infinite. The following assumption allows us to provide a fine-grained analysis of robustness in a partially identified setting.

Assumption 3.1 (Boundedness of the causal parameter)**.** *There exists a constant $C > 0$ such that any causal parameter β generating the SCM (3) is norm-bounded by C, i.e. $\|\beta\|_2 \leq C$ and hence $\Theta = \mathcal{B}^d(C) \times \mathbb{R}^{(d+1)\times(d+1)}$.*

Furthermore, two key quantities that appear in the bounds are $S_{\text{tot}} = \mathbb{R}^d - \text{range } R$, the space of directions which are either identified or unperturbed during test time, and $C_{\text{ker}} = \sqrt{C^2 - \|\beta^{\mathcal{S}}\|^2}$, the maximum norm of the non-identified part of the causal parameter $\beta^\star$. Finally, recall that the reference distribution $\mathbb{P}_{\theta^\star,0}^{X,Y}$ is observed and hence identifiable.

Theorem 3.1. *Assume that the training and test data follow the data-generating mechanism in Section 2.1 with test time shifts decomposed as in Equation (11) for some semi-orthogonal matrices S, R with* range $S \subset \mathcal{S}$, range $R \subset \mathcal{S}^\perp$*. Further, let Assumption 3.1 hold with parameter C. The identifiable robust risk (8) is then given by*

$$\mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, \gamma \Pi_{\mathcal{M}}) = \gamma \mathbb{I}_{R \neq 0} (C_{\text{ker}} + \|R^\top \beta\|_2)^2 + \gamma \|S^\top (\beta^{\mathcal{S}} - \beta)\|_2^2 + \mathcal{R}(\beta; \mathbb{P}_{\theta^\star,0}^{X,Y}), \quad (14)$$

Further, we obtain the following lower bound for the minimax quantity as defined in Equation (9):

$$\mathfrak{M}(\Theta_{\text{eq}}, \gamma \Pi_{\mathcal{M}}) \begin{cases} = \gamma \mathbb{I}_{R \neq 0} C_{\text{ker}}^2 + \min_{R^\top \beta = 0} \mathcal{R}_{\text{rob}}(\beta; \theta^{\mathcal{S}}, \gamma S S^\top), & \textit{if } \gamma \geq \gamma_{\text{th}}; \\ \geq \gamma \mathbb{I}_{R \neq 0} C_{\text{ker}}^2 + \min_{\beta \in \mathbb{R}^d} \mathcal{R}_{\text{rob}}(\beta; \theta^{\mathcal{S}}, \gamma S S^\top), & \textit{else}, \end{cases}$$

where $\gamma_{\text{th}} = \frac{(\kappa(\Sigma_\eta^\star)+1)\|RR^\top \Sigma_{\eta,\xi}^{\mathcal{S}}\|}{C_{\text{ker}}}$. Moreover, if $R \neq 0$, for small shifts

$$\lim_{\gamma \to 0} \frac{\mathfrak{M}(\Theta_{\text{eq}}, \gamma \Pi_{\mathcal{M}})}{\gamma} = (C_{\text{ker}} + \|RR^\top {\Sigma_\eta^\star}^{-1} \Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2. \quad (15)$$

We prove Theorem 3.1 in Appendix F.2. In the case of no new test shifts, i.e., $R = 0$, is discussed in prior work [45, 49], as the strength γ of the shift grows, the identifiable robust risk saturates. On the other hand, if $R \neq 0$, i.e., the test shift contains new directions w.r.t. to the training data, the best achievable robustness $\mathfrak{M}(\Theta_{\text{eq}}, \gamma\Pi_{\mathcal{M}})$ grows linearly with γ, and thus no infinite robustness is possible. We highlight that the minimax quantity is attained by the *identifiable robust predictor*

$$\beta^{\text{rob,ID}} = \arg\min_{\beta \in \mathbb{R}^d} \mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, \gamma\Pi_{\mathcal{M}}),$$

and for $\gamma \geq \gamma_{\text{th}}$, the lower bound corresponds to its identifiable robust risk and thus is tight. Moreover, $\gamma \geq \gamma_{\text{th}}$, $\beta^{\text{rob,ID}}$ can be explicitly computed from the training distributions (cf. Appendix F.2) and is *orthogonal* to the space $\operatorname{range} R$ of non-identifiable test shift directions. In other words, for large shifts γ in non-identified directions, the optimal robust model would "abstain" from prediction in those directions. For smaller γ, $\beta^{\text{rob,ID}}$ gradually utilizes less information in the non-identified directions, thus interpolating between maximum predictive power (OLS) and robustness w.r.t. new directions (abstaining). Note that the model $\beta^{\text{rob,ID}}$ is a population quantity that is identifiable from the collection of training *distributions*. When only finite samples are available, we discuss in Appendix D how we can still estimate the minimax quantity by minimizing an empirical loss function (17) that can be computed from multi-environment data. Additionally, in Appendix D, we provide details on the computation of the empirical identifiable robust risk and the corresponding estimator.

3.2 Theoretical analysis of existing finite robustness methods

We now evaluate existing finite robustness methods in our partial identifiability framework and discuss in which scenarios they are far from the best achievable robustness. A spiritually similar systematic comparison of domain adaptation methods is presented in [12], however, in our setting, the robust risk is not identifiable from data. We impose a probability distribution on the environment variable $E \in \mathcal{E}_{\text{train}}$ s.t. $\mathbb{P}[E = e] = w_e$, which allows us to compare to the anchor regression framework and similar, where the environment weights are required to obtain an estimate of M_{test}. In our discussion, we focus on discrete anchor regression [45] and pooled OLS estimators[8]. In discrete anchor regression, for each environment e, we observe data (X_e, Y_e) following the SCM $X_e = \mu_e + \eta$; $Y_e = \beta^{\star\top} X_e + \xi$, where $\mu_e \in \mathbb{R}^d$ are mean shifts and the noise is distributed like in Equation (3). The discrete anchor regression estimator minimizes the following robust risk:

$$\beta_{\text{anchor}} = \arg\min_{\beta \in \mathbb{R}^d} \mathcal{R}_{\text{rob}}(\beta; \theta^\star, \gamma M_{\text{anchor}}),$$

where $M_{\text{anchor}} = \sum_{e \in \mathcal{E}_{\text{train}}} w_e \mu_e \mu_e^\top$. The pooled ordinary least squares (OLS) estimator β_{OLS} corresponds to β_{anchor} with $\gamma = 1$. We observe that the test shifts bounded by γM_{anchor} are fully contained in the space of identified directions $\mathcal{S}$, since $\mathcal{S} = \operatorname{range} \cup_{e \in \mathcal{E}_{\text{train}}} \mu_e \mu_e^\top = \operatorname{range} M_{\text{anchor}}$. Thus, according to Proposition 1, the robust risk and robust predictor β_{anchor} are identifiable for all $\gamma > 0$. We now evaluate the robustness performance of β_{anchor} and β_{OLS} with respect to the more general shifts bounded by $M_{\text{new}} := \gamma M_{\text{anchor}} + \gamma' R R^\top$, thus consisting of training-identified shifts M_{anchor} and a possibly smaller share of previously unseen shifts in $\operatorname{range} R \subset \mathcal{S}^\perp$. With respect to M_{new}, the robust risk is only partially identified, and identifiable robust risk (8) given by

$$\mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, M_{\text{new}}) = \gamma'(C_{\text{ker}} + \|R^\top \beta\|_2)^2 + \mathcal{R}_{\text{rob}}(\beta; \theta^\star, \gamma M_{\text{anchor}}).$$

We evaluate how the identifiable robust risks (14) of both previous methods depend on the strength γ' of the previously unseen shift:

$$\mathcal{R}_{\text{rob,ID}}(\beta_{\text{anchor}}; \Theta_{\text{eq}}, M_{\text{new}})/\gamma' = (C_{\text{ker}} + \|RR^\top(\Sigma_\eta^\star + \gamma M_{\text{anchor}})^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2 + o(\gamma');$$
$$\mathcal{R}_{\text{rob,ID}}(\beta_{\text{OLS}}; \Theta_{\text{eq}}, M_{\text{new}})/\gamma' = (C_{\text{ker}} + \|RR^\top(\Sigma_\eta^\star + M_{\text{anchor}})^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2 + o(\gamma').$$

In contrast, for the best achievable robustness in the anchor setting[9] it holds

$$\frac{\mathfrak{M}(\Theta_{\text{eq}}, M_{\text{new}})}{\gamma'} = C_{\text{ker}}^2 + o(\gamma'), \text{ if } \gamma' \geq \gamma_{\text{th}};$$
$$\lim_{\gamma' \to 0} \frac{\mathfrak{M}(\Theta_{\text{eq}}, M_{\text{new}})}{\gamma'} = (C_{\text{ker}} + \|RR^\top(\Sigma_\eta^\star + \gamma M_{\text{anchor}})^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2.$$

[8] In Appendix C, we show that analogous results hold for continuous anchor regression and the method of distributionally robust invariant gradients (DRIG) [49].

[9] Here, we only vary γ', whereas γ is fixed.

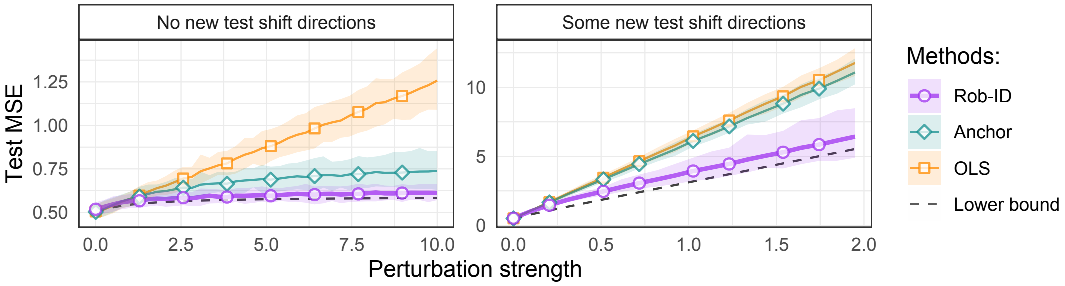


Figure 3: Test error under a partially unidentified distribution shift A_{test} of the baseline estimators $\beta_{\text{OLS}}, \beta_{\text{anchor}}$ (using the "correct" γ) for finite robustness and the identifiable robust predictor in (mean-shifted) multi-environment finite-sample experiments in the classical identified setting (left) and the partially identified robustness setting (right). The details of the experimental setting can be found in Appendix E.

Thus, the anchor regression estimator is optimal in the limit of small unseen shifts but significantly deviates from the best achievable robustness for smaller shifts. This is due to the fact that the term $\|RR^\top(\Sigma_\eta^\star + \gamma M_{\text{anchor}})^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|$ only goes to zero as $\gamma \to \infty$ (yielding the minimax risk) if M_{anchor} is full-rank, otherwise, it is strictly bounded from below as $\Sigma_\eta^\star$ is full-rank. Moreover, under some conditions on the covariance matrix (e.g., if $\Sigma_\eta^\star$ is block-diagonalizable w.r.t. $\mathcal{S}$ and $\mathcal{S}^\perp$), pooled OLS and the anchor estimator achieve the same rate in γ', showcasing how finite robustness methods can perform similarly to empirical risk minimization if the assumptions on the robustness set are not met. We provide additional comparisons in Appendix C.

3.3 Experimental results

In this section, we provide empirical evidence of our theoretical conclusions in Section 3.1 and Section 3.2. In particular, we compare the prediction performance of multiple existing robustness methods to the minimax lower bound, estimated by the identifiable robust predictor – including partially idenfitiable settings in which the test data contains shifts in *previously unseen* directions. We observe that both in a synthetic adversarial setting, empirical risk minimization and invariance-based robustness methods have significantly sub-optimal test loss in the partially identified setting, confirming our theoretical predictions in Section 3.2.Furthermore, we observe that even though the minimizer of identifiable robust risk is optimal only for the linear causal setting in Section 2.1, it surprisingly outperforms existing methods in a real-world experiment.

Experiments on synthetic Gaussian data We simulate Gaussian covariates according to Equation (3) with multiple environments differing by linearly independent randomly selected mean shifts. Given a fixed confounding model represented by a noise covariance Σ and fixed directions $\mathcal{S}$ of training mean shifts, we "evaluate" the identifiable robust risk by first picking the most adversarial $\beta^\star$ for fixed Σ and $\mathcal{S}$, and then computing its robust risk (6). We describe the full details of the data generation and loss evaluation in Appendix E. We consider two shift scenarios: in the first one, corresponding to the identifiable case in Figure 2a, the test environment is only perturbed by bounded shifts in training directions, as considered in prior work [45, 49]. In the second scenario, corresponding to the non-identifiable case Figure 2b, the test environment is perturbed by a mixture of training shifts and shifts in previously unobserved directions. We compute the empirical minimizers $\hat\beta_{\text{OLS}}, \hat\beta_{\text{anchor}}$ and $\hat\beta^{\text{rob,ID}}$ of the OLS, anchor regression and identifiable robust losses, respectively, and compare their test MSE (mean squared error) in Figure 3. In the first (identifiable) setting – Figure 3 (left) – the robust risk is asymptotically constant across γ for both robust methods, while the error for the vanilla ERM or OLS estimator increases linearly. In contrast, in the second, partially identified, setting – Figure 3 (right) – all estimators exhibit linearly increasing test errors; however the slopes of the anchor and OLS estimator are much steeper and lead to larger errors than the empirical minimizer of (14) that closely matches the analytic theoretical lower bound.

Real-world data experiments We consider the single-cell gene expression dataset from [41], which consists of single-cell observations over $d = 622$ genes collected from both observational and several interventional environments. Following [48], we only select the 28 genes that are active in the observational environment. For each gene $j = 1, \ldots, 28$, we generate a dataset D_j where $Y := X_j$

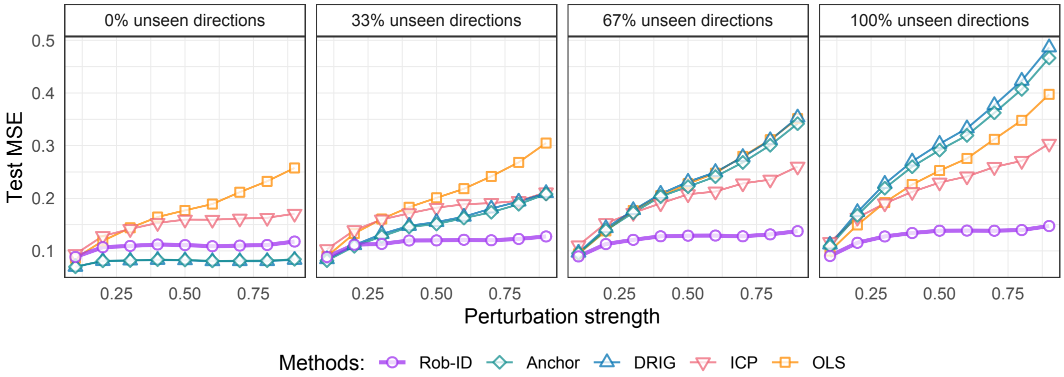


Figure 4: The figures show the performance of the *identifiable robust predictor* (Rob-ID) compared to other methods as a function of perturbation strength s (which is obtained by selecting the $s \times 100$ percent of data points closest to the observational mean). Different panels correspond to the proportion of unseen shift directions at test time. For each panel and perturbation strength s, each point represents an average over the 28 target genes and three experiments (i.e., training environments).

is the target variable and the covariates are the three genes most strongly correlated with Y. For each D_j, we perform three experiments – every experiment uses a different interventional environment besides the observational data as training data (an illustration of the data structure can be found in Figure 5). We then separately compute the mean-squared error on subsets of samples from all three interventional environments (including held-out samples from the interventional environment used for training). As a proxy for shift strength $\gamma > 0$, for each test environment, we pick the $s \times 100\%$ of data points closest to the observational mean. More details on this process can be found in Appendix E. We describe the computation of the identifiable robust estimator in Appendix D. In Figure 4, we show the test MSE of various OOD methods and the identifiable robust estimator as a function of s, presented in four different scenarios: no unseen shifts (left), some proportion of unseen shifts (middle panels) and 100% unseen shift directions (right). We compare the performance of anchor regression [45], invariant causal prediction (ICP) [38], Distributional Robustness via Invariant Gradients (DRIG) [49], and OLS with our estimated lower bound Rob-ID. We observe that the performance ranking of the robustness methods significantly varies with the proportion of new test shift directions. When no new shift information is present, anchor regression and DRIG are optimal. However, as soon as some unseen directions are present, their performance becomes inferior to OLS/ERM and the gap to the minimizer of the identifiable robust risk (in the setting in Section 2) grows with the proportion of unseen shifts. While the MSE of the previous invariant methods increases drastically with the strength of the test shift, the test loss of the identifiable robust predictor remains relatively stable.

4 Conclusion and future directions

This paper introduces the identifiable robust risk that is well-defined even in settings where the robust risk is not computable from training distributions. When the robustness set is identifiable (such as anchor regression-related methods [45, 49]), the identifiable robust risk reduces to the conventional robust risk. In this paper, we instantiate our general framework for linear structural causal models with additive shifts. We compute tight lower bounds for this setting and show how existing invariance-based methods are suboptimal. Further, we demonstrate how i) the benefits of invariance-based methods strongly decrease in the partially identifiable setting; and ii) this suboptimality increases with perturbation strength and proportion of previously unobserved test shifts.

The main limitation of our paper is its reliance on a linear causal setting to explicitly compute the observationally equivalent set and estimate the minimax quantity. However, we expect that the results and intuition developed in this paper can be extended to linear shifts in a lower-dimensional latent space via a suitable parametric or non-linear map [55, 10]. Important future directions include extending our results to more general causal graphs, non-linear relationships between covariates, non-additive shifts and the classification setting. Further, a potential use of our work is in the field of *active intervention selection* (e.g, [60, 21]). By computing the most adversarial model parameter for a given estimator, e.g., OLS, we can obtain an intervention which minimizes the identifiable robust risk of the estimator on the next unseen shift.

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

[3] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? A sample complexity perspective. In *International Conference on Learning Representations*, 2021.

[4] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[5] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

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

[7] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[8] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Roger J Bowden and Darrell A Turkington. *Instrumental variables*. Number 8. Cambridge university press, 1990.

[10] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2024.

[11] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[12] Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22(261):1–80, 2021.

[13] Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*, 2022.

[14] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021.

[15] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

[16] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[17] Jean-Marie Dufour and Cheng Hsiao. *Identification*, pages 65–77. Palgrave Macmillan UK, London, 2010.

[18] Jianqing Fan, Cong Fang, Yihong Gu, and Tong Zhang. Environment invariant linear least squares. *arXiv preprint arXiv:2303.03092*, 2023.

[19] Justin Frake, Anthony Gibbs, Brent Goldfarb, Takuya Hiraiwa, Evan Starr, and Shotaro Yamaguchi. From perfect to practical: Partial identification methods for causal inference in strategic management research. *Available at SSRN 4228655*, 2023.

[20] Charlie Frogner, Sebastian Claici, Edward Chien, and Justin Solomon. Incorporating unlabeled data into distributionally robust learning. *Journal of Machine Learning Research*, 22(56):1–46, 2021.

[21] Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475, 2020.

[22] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

[23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[24] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[25] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

[26] Martin Emil Jakobsen and Jonas Peters. Distributional robustness of k-class estimators and the pulse. *The Econometrics Journal*, 25(2):404–432, 2022.

[27] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

[28] Lucas Kook, Beate Sick, and Peter Bühlmann. Distributional anchor regression. *Statistics and Computing*, 32(3):39, 2022.

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

[30] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[31] Jiashuo Liu, Jiayun Wu, Bo Li, and Peng Cui. Distributionally robust optimization with data geometry. *Advances in Neural Information Processing Systems*, 35:33689–33701, 2022.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[33] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

[34] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

[35] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[36] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

[37] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[38] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[39] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. 2017.

[40] Lei S Qi, Matthew H Larson, Luke A Gilbert, Jennifer A Doudna, Jonathan S Weissman, Adam P Arkin, and Wendell A Lim. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

[41] Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022.

[42] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[43] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2657. PMLR, 2022.

[44] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

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

[46] Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, and Jonas Peters. Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning*, pages 18935–18958. PMLR, 2022.

[47] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[48] Christoph Schultheiss and Peter Bühlmann. Assessing the overall and partial causal well-specification of nonlinear additive noise models. *Journal of Machine Learning Research*, 25(159):1–41, 2024.

[49] Xinwei Shen, Peter Bühlmann, and Armeen Taeb. Causality-oriented robustness: exploiting general additive interventions. *arXiv preprint arXiv:2307.10299*, 2023.

[50] Wenqi Shi and Wenkai Xu. Learning nonlinear causal effect via kernel anchor regression. In *Uncertainty in Artificial Intelligence*, pages 1942–1952. PMLR, 2023.

[51] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.

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

[53] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[54] Elie Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.

[55] Nikolaj Thams, Michael Oberst, and David Sontag. Evaluating robustness to dataset shift via parametric robustness sets. *Advances in Neural Information Processing Systems*, 35:16877–16889, 2022.

[56] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

[57] Chuanlong Xie, Haotian Ye, Fei Chen, Yue Liu, Rui Sun, and Zhenguo Li. Risk variance penalization. *arXiv preprint arXiv:2006.07544*, 2020.

[58] Bin Yu. Assouad, Fano, and le Cam. In *Festschrift for Lucien Le Cam: Research papers in probability and statistics*, pages 423–435. Springer, 1997.

[59] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[60] Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10):1066–1075, 2023.

Appendix

The following sections provide deferred discussions, proofs and experimental details.

Table of contents

A Extended related work

To put our work into context, first, we discuss relevant distributional robustness literature organized according to structural assumptions on the desired robustness set. Second, we summarize existing views on partial identifiability in the causality and econometrics literature and how our findings connect to their perspective.

No structural assumptions on the shift. **DRO:** Distributionally robust optimization (DRO) tackles the problem of domain generalization when the robustness set is a ball around the training distribution w.r.t. some probability distance measure, e.g., Wasserstein distance [53, 35] or f-divergences [7, 16]. Considering all test distributions in a discrepancy ball can lead to overly conservative predictions, and therefore, alternatives have been proposed in, e.g., the Group DRO literature [47, 20, 31]. However, these methods cannot protect against perturbations larger than the ones seen during training time and do not provide a clear interpretation of the perturbations class [49].

Structural assumptions on the shift. Robustness from the lens of causality takes a step further, by assuming a structural causal model [37] generating the observed data (X, Y). **Infinite robustness methods:** The motivation of causal methods for robustness is that the causal function is worst-case optimal to predict the response under interventions of arbitrary direction and strength on the covariates [34, 11]. For this reason, causal models achieve what we call *infinite robustness*. Depending on the assumptions of the SCM, there are different ways to achieve infinite robustness. When there are no latent confounders, several works [38, 18, 33, 42, 2, 6, 51, 57, 29, 1] aim to identify the causal parents and achieve infinite robustness by exploiting the heterogeneity across training environments. In the presence of latent confounders, it is possible to achieve infinite robustness by identifying the causal function with, e.g., the instrumental variable method [5, 25, 52, 8, 36]. There are different limitations to *infinitely robust* methods. First, the identifiability conditions of the causal parents and/or causal function are often challenging to verify in practice. Second, ERM can outperform these methods when the interventions (read shifts) at test time are not arbitrarily strong or act directly on the response or latent variable [3, 24]. **Finite robustness methods:** In real data, shifts of arbitrary direction and strength in the covariates are unrealistic. Thus, different methods [45, 26, 28, 49, 14] trade off robustness against predictive power to achieve what we call *finite robustness*. The main idea of finite robustness methods is to learn a function that is as predictive as possible while protecting against shifts up to some strength in the directions that are observed during training time. These methods, however, only provide robustness guarantees that depend on the heterogeneity of the training data and do not offer insights into the limits of *algorithm-independent robustness* under shifts in new directions.

Partial identifiability: The problem of identification is at the center of the causal and econometric literature [39, 4]. It studies the conditions under which the (population) training distribution uniquely determines the causal parameters of the underlying SCM. Often, the training distribution only offers partial information about the causal parameters and, therefore, determines a set of observational equivalent parameters. This setting is known as *partial* or *set identification* and is used in causality and econometrics to learn intervals within which the true causal parameter lies [54]. In this work, we borrow the notion of partial identification to study the problem of distributional robustness when the robustness set itself is only partially identified.

B Extension to the general additive shift setting

We discuss how our setting changes when we relax the assumptions on the existence of the reference environment. We consider the data-generating process in Equation (3), where $\mathcal{E}_{\text{train}} = [m]$, $m \in \mathbb{N}$. If no environment e exists with $\mu_e = 0$ and $\Sigma_e = 0$, we first pick an arbitrary distribution $\mathbb{P}_{\text{ref}}^{X,Y}$ as the reference environment[10] . We denote $\Sigma'_\eta := \Sigma^\star_\eta + \Sigma_{\text{ref}}$.

First, we show we can express the space $\mathcal{S}$ of training additive shift directions defined in Equation (10) in the general case. We center all distributions by μ_{ref}, so that $\mathbb{E}[X_e] = \mu_e - \mu_{\text{ref}}$ for all $e \in \mathcal{E}_{\text{train}}$.

[10] In practice, it is useful to pick a distribution with the smallest covariance, i.e. $\operatorname{tr}\operatorname{Cov}(X_{\text{ref}}) \leq \operatorname{tr}\operatorname{Cov}(X_e)$ for all e.

With respect to the arbitrary reference environment, we now define

$$\tilde{\mathcal{S}} := \operatorname{range} \bigcup_{e \in \mathcal{E}_{\text{train}}} \left(\Sigma_e - \Sigma_{\text{ref}} + (\mu_e - \mu_{\text{ref}})(\mu_e - \mu_{\text{ref}})^\top \right) \subset \mathbb{R}^d.$$

We now consider test shifts with respect to the environment $\mathbb{P}_{\text{ref}}^{X,Y}$[11]. We define the test shift upper bound $\gamma \Pi_{\mathcal{M}}$. Again, we can decompose the upper bound as $\gamma \Pi_{\mathcal{M}} = \gamma S S^\top + \gamma R R^\top$, where $SS^\top$ and $RR^\top$ are orthogonal projections onto $\mathcal{S} \cap \mathcal{M}$ and $\mathcal{S}^\perp \cap \mathcal{M}$, respectively. Again, we can decompose the causal parameter $\beta^\star$ as $\beta^\star = \beta^{\mathcal{S}} + \beta^{\mathcal{S}^\perp}$. The projection $\beta^{\mathcal{S}}$ of the causal parameter onto the relative training shifts induces the following observationally equivalent parameters corresponding to the reference distribution:

$$\theta^{\mathcal{S}} := (\beta^{\mathcal{S}}, \Sigma'_\eta, \Sigma^{\mathcal{S}}_{\eta,\xi}, (\sigma^{\mathcal{S}}_\xi)^2) = (\beta^{\mathcal{S}}, \Sigma'_\eta, \Sigma^\star_{\eta,\xi} + \Sigma'_\eta \beta^{\mathcal{S}^\perp}, (\sigma^\star_\xi)^2 + 2\langle \Sigma^\star_{\eta,\xi}, \beta^{\mathcal{S}^\perp} \rangle + \langle \beta^{\mathcal{S}^\perp}, \Sigma'_\eta \beta^{\mathcal{S}^\perp} \rangle).$$

Again, $\theta^{\mathcal{S}}$ can be identified from the training distributions and is referred to as the *identified model parameters*. The following adapted version of Proposition 1 shows that assuming shifts on $\mathbb{P}_{\text{ref}}^{X,Y}$, the robust prediction model is only identifiable if the test shifts are in the direction of the relative training shifts:

Proposition 2 (Identifiability of reference distribution parameters and robust prediction model)**.** *Suppose that the set of training and test distributions is generated according to Equations* (3) *and* (4)*. Then, $\theta^{\mathcal{S}}$ is observationally equivalent to $\theta^\star$ and computable from training distributions. Furthermore, it holds that*

(a) the model parameters generating the reference distribution can be identified up to the following observationally equivalent set :

$$\Theta_{\text{eq}} = \{\beta^{\mathcal{S}} + \alpha, \Sigma'_\eta, \Sigma^{\mathcal{S}}_{\eta,\xi} - \Sigma'_\eta \alpha, (\sigma^{\mathcal{S}}_\xi)^2 - 2\alpha^\top \Sigma^{\mathcal{S}}_{\eta,\xi} + \alpha^\top \Sigma'_\eta \alpha \colon \alpha \in \mathcal{S}^\perp\} \ni \theta^\star$$

(b) the robust prediction model β^{rob} as defined in Equation (7) *is identified up to the set*

$$\beta^{\mathcal{S}} + (\gamma \Pi_{\mathcal{M}} + \Sigma'_\eta)^{-1} \Sigma^{\mathcal{S}}_{\eta,\xi} + \{(\gamma \Pi_{\mathcal{M}} + \Sigma'_\eta)^{-1} \alpha \colon \alpha \in \operatorname{range} R\} \ni \beta^{rob}$$

The proof is analogous to Appendix F.1. A version of Theorem 3.1 for perturbations on the reference environment follows accordingly.

C Comparison to finite robustness methods continued

C.1 Continuous anchor regression [45]

In the continuous anchor regression setting, during training we observe the distribution according to the SCM $X = MA + \eta$; $Y = \beta^{\star\top} X + \xi$, where $A \sim \mathcal{N}(0, \Sigma_A)$ is an observed q-dimensional anchor variable and $M \in \mathbb{R}^{d \times q}$ is a known matrix. Note that in this setting, we do not have a reference environment, but, since the anchor variable is observed, the distribution of the additive shift MA is known. The test shifts are assumed to be bounded by $M_{\text{test}} = \gamma M \Sigma_A M^\top$. Since $\operatorname{range} M_{\text{test}} \subset \mathcal{S} = \operatorname{range} M$, no new directions are observed during test time, in other words, $R = 0$. Thus, both the corresponding robust loss and the anchor regression estimator can be determined from training data. It holds that

$$\beta_{\text{anchor}} = \arg\min_{\beta \in \mathbb{R}^d} \mathcal{R}_{\text{rob}}(\beta; \theta^\star, \gamma M \Sigma_A M^\top).$$

Again, the pooled OLS estimator corresponds to β_{anchor} with $\gamma = 1$. Similar to the discrete anchor case, in case the test shifts are given by $M_{\text{new}} = \gamma M \Sigma_A M^\top + \gamma' R R^\top$, the identifiable robust risk (8) is given by

$$\mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, M_{\text{new}}) = \gamma' (C_{\text{ker}} + \|R^\top \beta\|_2)^2 + \mathcal{R}_{\text{rob}}(\beta; \theta^\star, \gamma M \Sigma_A M^\top)$$

[11] In other words, we require that the test distribution is a shifted version of the (arbitrarily) chosen reference distribution.

and for the best achievable robustness of the anchor estimator it holds

$$\mathcal{R}_{\text{rob,ID}}(\beta_{\text{anchor}}, \Theta_{\text{eq}}; M_{\text{new}})/\gamma' = (C_{\text{ker}} + \|RR^\top(\Sigma_\eta^\star + \gamma M \Sigma_A M^\top)^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2 + o(\gamma');$$
$$\lim_{\gamma'\to 0} \mathcal{R}_{\text{rob,ID}}(\beta_{\text{anchor}}, \Theta_{\text{eq}}; M_{\text{new}})/\gamma' = \lim_{\gamma'\to 0} \frac{\mathfrak{M}(\Theta_{\text{eq}}, M_{\text{new}})}{\gamma'}.$$

The above results follow by plugging M_{new} with $M := M_{\text{anchor}}$ into the proof of Theorem 3.1 in Appendix F.2.

C.2 Distributionally robust invariant gradients (DRIG) [49]

DRIG [49] uses the framework of Gaussian additive shifts $A_e \sim \mathcal{N}(\mu_e, \Sigma_e)$. For each environment e, we observe data (X_e, Y_e) distributed according to the SCM $X_e = A_e + \eta$; $Y_e = \beta^{\star\top} X_e + \xi$, where the noise is distributed like in Equation (3). DRIG consider more a more general intervention setting, additionally allowing additive shifts of Y and hidden confounders H. However, their identifiability results can only be shown for the case of interventions on X, and since identifiability of the causal parameter is a crucial part of our analysis, we only consider shifts on the covariates. DRIG assumes existence of a reference environment $e = 0$ with $\mu_0 = 0$ and for which it is required that the second moment of the reference environment is dominated by the second moment of the training mixture:

$$\Sigma_0 \preceq \sum_{e\in[m]} w_e(\Sigma_e + \mu_e\mu_e^\top).$$

This assumption allows [49] to derive the DRIG estimator which is robust against test shifts upper bounded by $M_{\text{DRIG}} := \gamma \sum_{e\in[m]} w_e(\Sigma_e - \Sigma_0 + \mu_e\mu_e^\top)$. The following lemma allows us to make further statements about M_{DRIG}:

Lemma C.1. *Let A and B be positive semidefinite matrices such that $B \preceq A$. Then it holds that* $\text{range } B \subset \text{range } A$.

Proof. It suffices to show that $\ker A \subset \ker B$. ($\ker A \subset \ker B$ implies that $\text{range } A = (\ker A)^\perp \subset (\ker B)^\perp = \text{range } B$.) Consider $x \in \ker A$, $x \neq 0$. Then it holds that $x^\top(A - B)x = x^\top Ax - x^\top Bx = 0 - x^\top Bx \geq 0$, from which it follows that $x^\top Bx = 0$ and thus $x \in \ker B$. □

Because of the assumption $\Sigma_0 \preceq \sum_{e\in[m]} w_e(\Sigma_e + \mu_e\mu_e^\top)$, by Lemma C.1 it follows that $\text{range } \Sigma_0 \subset \cup_{e\geq 1}\text{range } (\Sigma_e + \mu_e\mu_e^\top)$ and thus

$$\text{range } M_{\text{DRIG}} \subseteq \text{range } \left(\sum_{e\geq 1} w_e(\Sigma_e + \mu_e\mu_e^\top)\right).$$

Hence, the robustness directions achievable by DRIG in the "dominated reference environment" setting are the same as the ones under the assumption $\Sigma_0 = 0$.
Again, we observe that the test shifts bounded by γM_{DRIG} are fully contained in the space of identified directions $\mathcal{S}$. If the test shifts are instead bounded by $M_{\text{new}} := \gamma M_{\text{DRIG}} + \gamma' RR^\top$, including some unseen directions $\text{range } R \subset \mathcal{S}^\perp$, the robust risk in the DRIG setting is only partially identified. The identifiable robust risk (8) is given by

$$\mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, M_{\text{new}}) = \gamma'(C_{\text{ker}} + \|R^\top\beta\|_2)^2 + \mathcal{R}_{\text{rob}}(\beta; \theta^\star, \gamma M_{\text{DRIG}}),$$

and again, the DRIG estimator is optimal for infinitesimal shifts γ' and suboptimal for larger γ':

$$\mathcal{R}_{\text{rob,ID}}(\beta_{\text{DRIG}}; \Theta_{\text{eq}}, M_{\text{new}})/\gamma' = (C_{\text{ker}} + \|RR^\top(\Sigma_\eta^\star + \gamma M_{\text{DRIG}})^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2 + o(\gamma');$$
$$\frac{\mathfrak{M}(\Theta_{\text{eq}}, M_{\text{new}})}{\gamma'} = C_{\text{ker}}^2, \quad \text{if } \gamma' \geq \gamma_{\text{th}};$$
$$\lim_{\gamma'\to 0} \frac{\mathfrak{M}(\Theta_{\text{eq}}, M_{\text{new}})}{\gamma'} = (C_{\text{ker}} + \|RR^\top(\Sigma_\eta^\star + \gamma M_{\text{DRIG}})^{-1}\Sigma_{\eta,\xi}^{\mathcal{S}}\|)^2.$$

The above results follow by plugging M_{new} with $M := M_{\text{DRIG}}$ into the proof of Theorem 3.1 in Appendix F.2.

D Empirical estimation of the identifiable robust predictor

In this section, we discuss how to compute the identifiable robust loss and its minimizer from finite-sample multi-environment training data. We first describe the finite-sample setting and provide a high-level algorithm. We then discuss some parts of the algorithm in more detail. Finally, we show that under certain assumptions, the empirical identifiable robust loss is consistent.

D.1 Computing the identifiable robust loss

Algorithm 1 Computation of the identifiable robust loss

```
Input: Multi-environment data $\mathcal{D} := \cup_{e \in \mathcal{E}_{\text{train}}} \mathcal{D}_e$, test shift strength $\gamma > 0$, test shift directions $M \in \mathbb{R}^{d \times d}$, causal parameter upper bound $C > 0$.
Step 1: Estimate the training shift directions $\hat{\mathcal{S}}(\mathcal{D})$, its orthogonal complement $\hat{\mathcal{S}}^\perp(\mathcal{D})$, and the identified causal parameter $\hat{\beta}^{\mathcal{S}}$.
Step 2: Estimate the identified and non-identified test shift directions $\hat{S}$, $\hat{R}$ and their projections $\hat{S}\hat{S}^\top$ and $\hat{R}\hat{R}^\top$.
Step 3: Estimate the norm $\hat{C}_{\text{ker}}$ of the non-identified causal parameter.
Step 4: Compute the identifiable robust loss function
    $$\mathcal{L}_n(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \hat{R}\hat{R}^\top) \leftarrow \underbrace{\mathcal{L}_{\text{ref}}(\beta; \mathcal{D}_0)}_{\text{reference loss}} + \underbrace{\mathcal{L}_{\text{inv}}(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \gamma)}_{\text{invariance penalty term}} + \underbrace{\mathcal{L}_{\text{id}}(\beta; \hat{C}_{\text{ker}}, \hat{R}\hat{R}^\top, \gamma)}_{\text{non-identifiability penalty term}}.$$
Return: identifiable robust predictor and the estimated minimax "hardness" of the problem:
    $$\hat{\beta}^{\text{rob,ID}} \leftarrow \arg\min_{\beta \in \mathbb{R}^d} \mathcal{L}_n(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \hat{R}\hat{R}^\top);$$
    $$\hat{\mathfrak{M}}(\mathcal{D}, \gamma, M) \leftarrow \min_{\beta \in \mathbb{R}^d} \mathcal{L}_n(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \hat{R}\hat{R}^\top).$$
```

Training data. We observe data from $m + 1$ training environments indexed by $E \in \mathcal{E}_{\text{train}} = \{0, ..., m\}$, where $E = 0$ represents the reference environment. We impose a discrete probability distribution $\mathbb{P}^E$ on the training environment $E \in \mathcal{E}_{\text{train}}$, resulting in the joint distribution $(X, Y, E) \sim \mathbb{P}^{X,Y|E} \times \mathbb{P}^E$. For each environment $E = e$, we observe the samples $\mathcal{D}_e := \{(X_{e,i}, Y_{e,i})\}_{i=1}^{n_e}$, where $(X_{e,i}, Y_{e,i})$ are independent copies of $(X_e, Y_e) \sim \mathbb{P}^{X,Y|E=e}$. Then, the resulting dataset is $\mathcal{D} := \cup_{e \in \mathcal{E}_{\text{train}}} \mathcal{D}_e$ with $n := n_0 + \cdots + n_m$. Furthermore, for each environment $E = e$, we define the weights $w_e := n_e/n$.

Computation of the identifiable robust loss. In Algorithm 1, we present a high-level scheme for computing the identifiable robust loss from multi-environment data, which consists of multiple steps. First, nuisance parameters related to the training and test shift directions are estimated, which we describe in more detail below. Afterwards, the three terms of the loss are computed: the (squared) loss $\mathcal{L}_{\text{ref}}(\beta; \mathcal{D}_0)$ on the reference environment is computed as

$$\mathcal{L}_{\text{ref}}(\beta; \mathcal{D}_0) = \sum_{i=1}^{n_0} (Y_{0,i} - \beta^\top X_{0,i})^2.$$

The invariance penalty term $\mathcal{L}_{\text{inv}}(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \gamma)$ (which increasingly aligns any estimator β in the direction of the estimated invariant causal predictor $\hat{\beta}^{\mathcal{S}}$ as $\gamma \to \infty$) can be computed as following in the linear SCM setting:

$$\mathcal{L}_{\text{inv}}(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \gamma) = \gamma \|\hat{S}\hat{S}^\top(\beta - \beta^{\mathcal{S}})\|_2^2.$$

Finally, the non-identifiability penalty term $\mathcal{L}_{\text{id}}(\beta; \hat{C}_{\text{ker}}, \hat{R}\hat{R}^\top, \gamma)$ can be computed as follows:

$$\mathcal{L}_{\text{id}}(\beta; \hat{C}_{\text{ker}}, \hat{R}\hat{R}^\top, \gamma) \leftarrow \gamma(C_{\text{ker}} + \|\hat{R}\hat{R}^\top \beta\|_2)^2.$$

The non-identifiability term, with increasing γ, penalizes any predictor β towards zero on the subspace R of non-identified test shift directions. In total, the identifiable robust loss (in the linear SCM setting) equals

$$\mathcal{L}_n(\beta; \hat{\beta}^{\mathcal{S}}, \hat{S}\hat{S}^\top, \hat{R}\hat{R}^\top) = \sum_{i=1}^{n_0}(Y_{0,i} - \beta^\top X_{0,i})^2 + \gamma\|\hat{S}\hat{S}^\top(\beta - \beta^{\mathcal{S}})\|_2^2 + \gamma(C_{\ker} + \|\hat{R}\hat{R}^\top\beta\|_2)^2,$$

where we suppress dependence on C and γ and only leave the dependence on the nuisance parameters.

Choice/Estimation of nuisance parameters. We now provide more details on the empirical estimation of the nuisance parameters $\hat{\mathcal{S}}$, $\hat{S}$, $\hat{R}$, and $\hat{\beta}^{\mathcal{S}}$.

- The **constant** C corresponds to the upper bound on the norm of the true causal parameter $\beta^\star$. Thus, the practitioner chooses C in advance to ensure that (with high probability) $\|\beta^\star\|_2 \leq C$.
- The **training shift directions** $\hat{\mathcal{S}}$ can be computed via

$$\hat{\mathcal{S}}(\mathcal{D}) = \operatorname{range}\sum_{e=1}^{m}(\operatorname{Cov}(X^e) - \operatorname{Cov}(X^0) + \mu_e\mu_e^\top - \mu_0\mu_0^\top), \tag{16}$$

  where for $e \in \mathcal{E}_{\text{train}}$, the matrix $\operatorname{Cov}(X^e)$ is the empirical covariance matrix estimated within the training environment $E = e$, and $\mu_e \in \mathbb{R}^d$ is the empirical mean of the covariates within the training environment $E = e$. Additionally, we compute the orthogonal complement $\hat{\mathcal{S}}^\perp(\mathcal{D})$ of the space $\hat{\mathcal{S}}(\mathcal{D})$[12].
- The **decomposition of the test shift directions** M into identified and non-identified shift directions (and their corresponding projection matrices) can be computed as follows. Let $\Pi_{\hat{\mathcal{S}}}$ and $\Pi_{\hat{\mathcal{S}}^\perp}$ denote the projection matrices on $\hat{\mathcal{S}}(\mathcal{D})$ and $\hat{\mathcal{S}}^\perp(\mathcal{D})$, respectively. Consider the singular value decompositions $\Pi_{\hat{\mathcal{S}}}M = U_{\hat{\mathcal{S}}}\Sigma_{\hat{\mathcal{S}}}V_{\hat{\mathcal{S}}}^\top$ and $\Pi_{\hat{\mathcal{S}}^\perp}M = U_{\hat{\mathcal{S}}^\perp}\Sigma_{\hat{\mathcal{S}}^\perp}V_{\hat{\mathcal{S}}^\perp}^\top$ Then, define

$$\hat{S} = U_{\hat{\mathcal{S}}}, \quad \hat{R} = U_{\hat{\mathcal{S}}^\perp}.$$

  The subspaces $\operatorname{range}(\Pi_{\hat{\mathcal{S}}}M)$ and $\operatorname{range}(\Pi_{\hat{\mathcal{S}}^\perp}M)$ are minimal subspaces contained in $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}}^\perp$, respectively, such that $\operatorname{range}(M) \subset \operatorname{range}(\Pi_{\hat{\mathcal{S}}}M) \oplus \operatorname{range}(\Pi_{\hat{\mathcal{S}}^\perp}M)$. The matrices $\hat{S}\hat{S}^\top$ and $\hat{R}\hat{R}^\top$ are their corresponding projection matrices.
- The **identified causal parameter** $\hat{\beta}^{\mathcal{S}}$ (approximately) equals the true causal parameter $\beta^\star$ on the space of training shift directions $\hat{\mathcal{S}}$. As conjectured in the anchor regression literature [45, 49, 26] (see, for example, the discussion right after Theorem 3.4 in [26] and Appendix H.3 therein) for $\gamma \to \infty$, the estimators $\beta_{\text{anchor}}^{\gamma}$ and $\beta_{\text{DRIG}}^{\gamma}$ converge to the causal parameter $\beta^\star$ on $\mathcal{S}$. Thus, the identified causal parameter can be estimated as

$$\hat{\beta}^{\mathcal{S}} := \Pi_{\hat{\mathcal{S}}}\beta_{\text{anchor}}^{\infty} \quad \text{or} \quad \hat{\beta}^{\mathcal{S}} := \Pi_{\hat{\mathcal{S}}}\beta_{\text{DRIG}}^{\infty}$$

  for the setting of mean or mean+variance shifts, respectively.

D.2 Consistency of the identifiable robust predictor

For any estimator $\beta \in \mathbb{R}^d$ and given the estimated nuisance parameters $\hat{\varphi} := (\hat{S}\hat{S}^\top, \hat{R}\hat{R}^\top, \hat{\beta}^{\mathcal{S}})$, we define the sample identifiable robust risk as

$$\mathcal{L}_n(\beta, \hat{\varphi}) := \frac{1}{n_0}\sum_{i \in \mathcal{D}_0}\left(Y_{0,i} - \beta^\top X_{0,i}\right)^2 + \gamma\|\hat{S}\hat{S}^\top(\hat{\beta}^{\mathcal{S}} - \beta)\|_2^2 + \gamma\left(\sqrt{C - \|\hat{\beta}^{\mathcal{S}}\|_2^2} + \|\hat{R}\hat{R}^\top\beta\|_2\right)^2. \tag{17}$$

[12] In general, $S(\mathcal{D})$ is a proper subspace of $\mathbb{R}^d$ and the RHS of (16) corresponds to a sum of low-rank second moments. This can be consistently estimated if, for instance, the rank of each shift is known (e.g. in the mean shift setting), or the covariances have a spiked structure, allowing to cut off small eigenvalues.

Correspondingly, we define the estimator of the identifiable robust predictor by

$$\hat{\beta}^{\mathrm{rob,ID}} := \arg\min_{\beta \in \mathcal{B}} \mathcal{L}_n(\beta, \hat{\varphi}), \tag{18}$$

where $\mathcal{B} \subseteq \mathbb{R}^d$ is some compact set whose interior contains the true parameter $\beta^{\mathrm{rob,ID}}$.

To show the consistency of (18), we first require consistency of the nuisance parameter estimators, which we state as an assumption.

Assumption D.1. *The estimated nuisance parameters $\hat{\varphi} := (\hat{S}\hat{S}^\top, \hat{R}\hat{R}^\top, \hat{\beta}^{\mathcal{S}})$ are consistent, that is, for $n \to \infty$,*

$$\|\hat{S}\hat{S}^\top - SS^\top\|_F \xrightarrow{\mathbb{P}} 0, \quad \|\hat{R}\hat{R}^\top - RR^\top\|_F \xrightarrow{\mathbb{P}} 0, \quad \hat{\beta}^{\mathcal{S}} \xrightarrow{\mathbb{P}} \beta^{\mathcal{S}} := \Pi_{\mathcal{S}}\beta^\star,$$

where for any matrix $A \in \mathbb{R}^{m\times q}$, $\|A\|_F = \sqrt{\mathrm{tr}\,(A^\top A)}$ denotes the Frobenius norm, and $SS^\top$, $RR^\top$ are the corresponding population projection matrices onto $\Pi_{\mathcal{S}}\mathcal{M}$, $\Pi_{\mathcal{S}^\perp}\mathcal{M}$ respectively.

Depending on the assumptions of the data-generating process, Assumption D.1 can be shown to hold. For example, in the anchor regression setting [45], the consistency of the projection matrices $\hat{S}\hat{S}^\top$, $\hat{R}\hat{R}^\top$, and $\Pi_{\hat{\mathcal{S}}}$ holds if the dimension of $\mathcal{S}$ is known (due to the mean shift structure). The proof relies on the Davis–Kahan theorem (see, for example, [59]) and the consistency of the covariance matrix estimator. Moreover, in the anchor regression setting, it is conjectured that the estimator $\beta^\infty_{\text{anchor}}$ converges to its population counterpart (as discussed right after Theorem 3.4 in [26] and Appendix H.3 therein) which implies that $\hat{\beta}^{\mathcal{S}} := \Pi_{\hat{\mathcal{S}}}\beta^\infty_{\text{anchor}}$ consistently estimates $\beta^{\mathcal{S}} = \Pi_{\mathcal{S}}\beta^\star$.

Under the assumption of the consistency of the nuisance parameter estimators, we can now show that (18) is a consistent estimator of the identifiable robust predictor.

Proposition 3. *Consider the estimator $\hat{\beta}^{\mathrm{rob,ID}}$ of the identifiable robust predictor defined in (18). Suppose the optimization problem is over a compact set $\mathcal{B} \subseteq \mathbb{R}^d$ whose interior contains the true minimizer $\beta^{\mathrm{rob,ID}}$. Moreover, suppose Assumption D.1 holds. Finally, assume that the covariance matrix $\mathbb{E}\,[X_0X_0^\top] \succ 0$ with bounded eigenvalues and $\mathbb{E}\,[Y_0^2] < \infty$. Then, $\hat{\beta}^{\mathrm{rob,ID}}$ is consistent, i.e., as $n, n_0 \to \infty$ it holds that*

$$\hat{\beta}^{\mathrm{rob,ID}} \xrightarrow{\mathbb{P}} \beta^{\mathrm{rob,ID}}.$$

D.3 Proof of Proposition 3

For ease of notation define $\beta_0 := \beta^{\mathrm{rob,ID}}$ and $\hat{\beta} := \hat{\beta}^{\mathrm{rob,ID}}$. For any parameter of interest $\beta \in \mathcal{B}$ and nuisance parameters $\varphi = (P_S, P_R, b)$, define the function

$$(x, y) \mapsto g_{\beta,\varphi}(x, y) := (y - \beta^\top x)^2 + \gamma\|P_S(b - \beta)\|_2^2 + \gamma\left(\sqrt{C - \|b\|_2^2} + \|P_R\beta\|_2\right)^2. \tag{19}$$

Using (19), the robust identifiable risk and its sample version defined in (17) can be written, respectively as

$$\mathcal{L}(\beta, \varphi) = \mathbb{E}\,[g_{\beta,\varphi}(X_0, Y_0)], \quad \mathcal{L}_n(\beta, \varphi) = \frac{1}{n_0}\sum_{i\in\mathcal{D}_0} g_{\beta,\varphi}(X_{0,i}, Y_{0,i}).$$

Our goal is to show that $\hat{\beta} \xrightarrow{\mathbb{P}} \beta_0$. First, we show that the minimum of the loss is well-separated.

Lemma D.1. *Suppose that $\mathbb{E}\,[X_0X_0^\top] \succ 0$. Then, for all $\delta > 0$, it holds that*

$$\inf\,\{\mathcal{L}(\beta, \varphi_0) \colon \|\beta - \beta_0\|_2 > \delta\} > \mathcal{L}(\beta_0, \varphi_0). \tag{20}$$

Fix $\delta > 0$. From the well-separation of the minimum from Lemma D.1, there exists $\varepsilon > 0$ such that

$$\left\{\|\hat{\beta} - \beta_0\|_2 > \delta\right\} \subseteq \left\{\mathcal{L}(\hat{\beta}, \varphi_0) - \mathcal{L}(\beta_0, \varphi_0) > \varepsilon\right\}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\beta}-\beta_0\|_2>\delta\right) &\leq \mathbb{P}\left(\mathcal{L}(\hat{\beta},\varphi_0)-\mathcal{L}(\beta_0,\varphi_0)>\varepsilon\right)\\
&= \mathbb{P}\Big(\mathcal{L}(\hat{\beta},\varphi_0)-\mathcal{L}_n(\hat{\beta},\varphi_0)+\mathcal{L}_n(\hat{\beta},\varphi_0)-\mathcal{L}_n(\hat{\beta},\hat{\varphi})\\
&\qquad +\mathcal{L}_n(\hat{\beta},\hat{\varphi})-\mathcal{L}_n(\beta_0,\hat{\varphi})+\mathcal{L}_n(\beta_0,\hat{\varphi})-\mathcal{L}(\beta_0,\varphi_0)>\varepsilon\Big)\\
&\leq \mathbb{P}\left(\mathcal{L}(\hat{\beta},\varphi_0)-\mathcal{L}_n(\hat{\beta},\varphi_0)>\varepsilon/4\right)+\mathbb{P}\left(\mathcal{L}_n(\hat{\beta},\varphi_0)-\mathcal{L}_n(\hat{\beta},\hat{\varphi})>\varepsilon/4\right) \qquad (21)\\
&\quad +\mathbb{P}\left(\mathcal{L}_n(\hat{\beta},\hat{\varphi})-\mathcal{L}_n(\beta_0,\hat{\varphi})>\varepsilon/4\right)+\mathbb{P}\left(\mathcal{L}_n(\beta_0,\hat{\varphi})-\mathcal{L}(\beta_0,\varphi_0)>\varepsilon/4\right). \qquad (22)
\end{aligned}$$

We now want to prove convergence the four terms in (21) and (22). For this, we use the following statements proved in Appendix D.4.

Lemma D.2. *Suppose $\mathcal{B}\subseteq\mathbb{R}^d$ is a compact set. Moreover, assume that the covariance matrix $\mathbb{E}\left[X_0X_0^\top\right]\succ 0$ with bounded eigenvalues and $\mathbb{E}\left[Y_0^2\right]<\infty$. Then, as $n,n_0\to\infty$ it holds that*

$$\sup_{\beta\in\mathcal{B}}|\mathcal{L}_n(\beta,\varphi_0)-\mathcal{L}(\beta,\varphi_0)|\xrightarrow{\mathbb{P}}0. \tag{23}$$

Lemma D.3. *As $n\to\infty$, it holds that*

$$\sup_{\beta\in\mathcal{B}}|\mathcal{L}_n(\beta,\hat{\varphi})-\mathcal{L}_n(\beta,\varphi_0)|\xrightarrow{\mathbb{P}}0. \tag{24}$$

The two terms in (21) converge to 0 by Lemma D.2 and Lemma D.3, respectively. The first term in (22) equals 0 since $\hat{\beta}$ minimizes $\beta\mapsto\mathcal{L}_n(\beta,\hat{\varphi})$. Finally, we observe that

$$\sup_{\beta\in\mathcal{B}}|\mathcal{L}_n(\beta,\hat{\varphi})-\mathcal{L}(\beta,\varphi_0)|\xrightarrow{\mathbb{P}}0, \tag{25}$$

since we have that

$$\sup_{\beta\in\mathcal{B}}|\mathcal{L}_n(\beta,\hat{\varphi})-\mathcal{L}(\beta,\varphi_0)|\leq\sup_{\beta\in\mathcal{B}}|\mathcal{L}_n(\beta,\hat{\varphi})-\mathcal{L}_n(\beta,\varphi_0)|+\sup_{\beta\in\mathcal{B}}|\mathcal{L}_n(\beta,\varphi_0)-\mathcal{L}(\beta,\varphi_0)|,$$

where the first term converges in probability by Lemma D.3, and the second term converges in probability by Lemma D.2. This implies that the second term in (22) converges to zero. Since $\delta>0$ was arbitrary, it follows that $\hat{\beta}\xrightarrow{\mathbb{P}}\beta_0$.

D.4 Proof of auxiliary lemmas

D.4.1 Proof of Lemma D.1

By definition,

$$\mathcal{L}(\beta,\varphi_0)=\mathbb{E}\left[(Y_0-\beta^\top X_0)^2\right]+\gamma\|SS^\top(\beta^{\mathcal{S}}-\beta)\|_2^2+\gamma\left(\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}+\|RR^\top\beta\|_2\right)^2.$$

Since $\mathbb{E}\left[X_0X_0^\top\right]\succ 0$, the first term is strongly convex in β. Moreover, the second and third terms are convex in β. Therefore, $\mathcal{L}(\beta,\varphi_0)$ is strongly convex in β. Since $\mathcal{L}(\beta,\varphi_0)$ is also continuous in β, it follows that there exists a unique global minimum. Let β_0 denote the global minimizer of $\mathcal{L}(\beta,\varphi_0)$. By the fact that $\mathcal{L}(\beta_0,\varphi_0)$ is a global minimum, and by definition of strong convexity, there exists a positive constant $m>0$ such that, for all $\beta\in\mathcal{B}$,

$$\mathcal{L}(\beta,\varphi_0)\geq\mathcal{L}(\beta_0,\varphi_0)+\frac{m}{2}\|\beta-\beta_0\|_2^2. \tag{26}$$

Fix $\delta>0$. Then, by (26), for all $\beta\in\mathcal{B}$ such that $\|\beta-\beta_0\|_2>\delta$ it holds that

$$\mathcal{L}(\beta,\varphi_0)\geq\mathcal{L}(\beta_0,\varphi_0)+\frac{m\delta^2}{2}>\mathcal{L}(\beta_0,\varphi_0).$$

Since the inequality holds for all $\beta\in\mathcal{B}$ such that $\|\beta-\beta_0\|_2>\delta$, we conclude that

$$\inf\{\mathcal{L}(\beta,\varphi_0)\colon\|\beta-\beta_0\|_2>\delta\}>\mathcal{L}(\beta_0,\varphi_0).$$

Since $\delta>0$ was arbitrary, the claim follows.

D.4.2 Proof of Lemma D.2

Recall that for any $\beta \in \mathcal{B}$

$$\mathcal{L}(\beta, \varphi_0) = \mathbb{E}\left[g_{\beta,\varphi_0}(X_0, Y_0)\right], \quad \mathcal{L}_n(\beta, \varphi_0) = \frac{1}{n_0} \sum_{i \in \mathcal{D}_0} g_{\beta,\varphi_0}(X_{0,i}, Y_{0,i}).$$

To show the result, we must establish that the class of functions $\{g_{\beta,\varphi_0} \colon \beta \in \mathcal{B}\}$ is Glivenko–Cantelli. From [56], a set of sufficient conditions for being a Glivenko–Cantelli class is that (i) $\mathcal{B}$ is compact, (ii) $\beta \mapsto g_{\beta,\varphi_0}(x, y)$ is continuous for every (x, y), and (iii) $\beta \mapsto g_{\beta,\varphi_0}$ is dominated by an integrable function. By assumption, (i) holds. Moreover, by (19), it follows that $\beta \mapsto g_{\beta,\varphi_0}$ is continuous for all (x, y) and thus (ii) holds. We now show that (iii) holds. Since $\mathcal{B}$ is compact we have that $\sup_{\beta\in\mathcal{B}}\|\beta\|_2 = C_1 < \infty$. For fixed $\gamma > 0$, and all (x, y), we have that

$$\begin{aligned} g_{\beta,\varphi_0}(x,y) &\le \sup_{\beta\in\mathcal{B}} |g_{\beta,\varphi_0}(x,y)| \\ &\le \sup_{\beta\in\mathcal{B}} (y - \beta^\top x)^2 + 2\gamma \|SS^\top\|_F^2 \left(\|\beta^{\mathcal{S}}\|_2^2 + \sup_{\beta\in\mathcal{B}} \|\beta\|_2^2 \right) \\ &\quad + \gamma \left(\sqrt{C - \|\beta^{\mathcal{S}}\|_2^2} + \|RR^\top\|_F \sup_{\beta\in\mathcal{B}} \|\beta\|_2 \right)^2 \\ &\le 2y^2 + 2C_1^2 \|x\|_2^2 + K =: G(x,y), \end{aligned} \tag{27}$$

where $K < \infty$ is a finite constant not depending on (x, y). Furthermore, we have that

$$\mathbb{E}\left[G(X_0, Y_0)\right] = 2\mathbb{E}\left[Y_0^2\right] + 2C_1^2 \operatorname{tr}\left(\mathbb{E}\left[X_0 X_0^\top\right]\right) + K < \infty, \tag{28}$$

since $\mathbb{E}\left[Y^2\right] < \infty$ and $\mathbb{E}\left[X_0 X_0^\top\right]$ has bounded eigenvalues by assumption. From (27) and (28), it follows that (iii) holds.

D.4.3 Proof of Lemma D.3

For fixed $\gamma > 0$, we have that

$$\frac{1}{\gamma} \sup_{\beta\in\mathcal{B}} |\mathcal{L}_n(\beta, \hat{\varphi}) - \mathcal{L}_n(\beta, \varphi_0)| \le \sup_{\beta\in\mathcal{B}} \left| \|\hat{S}\hat{S}^\top(\hat{\beta}^{\mathcal{S}} - \beta)\|_2^2 - \|SS^\top(\beta^{\mathcal{S}} - \beta)\|_2^2 \right| \tag{29}$$

$$+ \sup_{\beta\in\mathcal{B}} \left| \left(\sqrt{C - \|\hat{\beta}^{\mathcal{S}}\|_2^2} + \|\hat{R}\hat{R}^\top\beta\|_2 \right)^2 - \left(\sqrt{C - \|\beta^{\mathcal{S}}\|_2^2} + \|RR^\top\beta\|_2 \right)^2 \right| \tag{30}$$

We can upper bound (29) as follows,

$$\begin{aligned} &\sup_{\beta\in\mathcal{B}} \left| \|\hat{S}\hat{S}^\top(\hat{\beta}^{\mathcal{S}} - \beta)\|_2^2 - \|SS^\top(\beta^{\mathcal{S}} - \beta)\|_2^2 \right| \\ &= \sup_{\beta\in\mathcal{B}} \left| (\hat{\beta}^{\mathcal{S}} - \beta)^\top \hat{S}\hat{S}^\top (\hat{\beta}^{\mathcal{S}} - \beta) - (\beta^{\mathcal{S}} - \beta)^\top SS^\top (\beta^{\mathcal{S}} - \beta) \right| \\ &= \sup_{\beta\in\mathcal{B}} \Big| (\hat{\beta}^{\mathcal{S}} - \beta)^\top \hat{S}\hat{S}^\top (\hat{\beta}^{\mathcal{S}} - \beta^{\mathcal{S}}) + (\hat{\beta}^{\mathcal{S}} - \beta^{\mathcal{S}})^\top \hat{S}\hat{S}^\top (\beta^{\mathcal{S}} - \beta) \\ &\qquad + (\beta^{\mathcal{S}} - \beta)^\top (\hat{S}\hat{S}^\top - SS^\top)(\beta^{\mathcal{S}} - \beta) \Big| \\ &\le 2 \sup_{\beta\in\mathcal{B}} \|\hat{\beta}^{\mathcal{S}} - \beta\|_2 \, \|\hat{S}\hat{S}^\top\|_F \, \|\hat{\beta}^{\mathcal{S}} - \beta^{\mathcal{S}}\|_2 + \sup_{\beta\in\mathcal{B}} \|\beta^{\mathcal{S}} - \beta\|_2^2 \, \|\hat{S}\hat{S}^\top - SS^\top\|_F && (31) \\ &\le C_1 \|\hat{\beta}^{\mathcal{S}} - \beta^{\mathcal{S}}\|_2 + C_2 \|\hat{S}\hat{S}^\top - SS^\top\|_F \xrightarrow{\mathbb{P}} 0, && (32) \end{aligned}$$

where (31) follows from the Cauchy–Schwarz inequality and that $\|A\|_2 \le \|A\|_F$, the constants $C_1, C_2 < \infty$ in (32) follow from compactness of $\mathcal{B}$, and the convergence in probability follows from

Assumption D.1. Furthermore, we can upper bound (30) as follows,

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left|\left(\sqrt{C-\|\hat{\beta}^{\mathcal{S}}\|_2^2}+\|\hat{R}\hat{R}^\top\beta\|_2\right)^2-\left(\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}+\|RR^\top\beta\|_2\right)^2\right| \\
&= \sup_{\beta\in\mathcal{B}}\Big|C-\|\hat{\beta}^{\mathcal{S}}\|_2^2+\|\hat{R}\hat{R}^\top\beta\|_2^2+2\sqrt{C-\|\hat{\beta}^{\mathcal{S}}\|_2^2}\,\|\hat{R}\hat{R}^\top\beta\|_2 \\
&\qquad\qquad -C+\|\beta^{\mathcal{S}}\|_2^2-\|RR^\top\beta\|_2^2-2\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\,\|RR^\top\beta\|_2\Big| \\
&\le \sup_{\beta\in\mathcal{B}}\left|\|\hat{\beta}^{\mathcal{S}}\|_2^2-\|\beta^{\mathcal{S}}\|_2^2\right|+\sup_{\beta\in\mathcal{B}}\left|\beta^\top(\hat{R}\hat{R}^\top-RR^\top)\beta\right| \\
&\quad +2\sup_{\beta\in\mathcal{B}}\left|\sqrt{C-\|\hat{\beta}^{\mathcal{S}}\|_2^2}\,\|\hat{R}\hat{R}^\top\beta\|_2-\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\,\|RR^\top\beta\|_2\right| \\
&= (I)+(II)+(III).
\end{aligned}$$

By Assumption D.1, (I) converges in probability to zero. Regarding (II), we have

$$\sup_{\beta\in\mathcal{B}}\left|\beta^\top(\hat{R}\hat{R}^\top-RR^\top)\beta\right|\le\sup_{\beta\in\mathcal{B}}\|\beta\|_2^2\,\|\hat{R}\hat{R}^\top-RR^\top\|_F\xrightarrow{\mathbb{P}}0,$$

where the inequality follows from Cauchy–Schwarz and that $\|A\|_2\le\|A\|_F$, and the convergence in probability follows from Assumption D.1 along with the compactness of $\mathcal{B}$. It remains to upper bound (III). We have that

$$\begin{aligned}
\frac{(III)}{2} &\le \sup_{\beta\in\mathcal{B}}\left|\sqrt{C-\|\hat{\beta}^{\mathcal{S}}\|_2^2}\,\|\hat{R}\hat{R}^\top\beta\|_2-\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\,\|\hat{R}\hat{R}^\top\beta\|_2\right| \\
&\quad +\sup_{\beta\in\mathcal{B}}\left|\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\,\|\hat{R}\hat{R}^\top\beta\|_2-\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\,\|RR^\top\beta\|_2\right| \\
&\le \left(\sup_{\beta\in\mathcal{B}}\|\beta\|_2\,\|\hat{R}\hat{R}^\top\|_F\right)\left|\sqrt{C-\|\hat{\beta}^{\mathcal{S}}\|_2^2}-\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\right| \\
&\quad +\sup_{\beta\in\mathcal{B}}\left|\sqrt{\beta^\top\hat{R}\hat{R}^\top\beta}-\sqrt{\beta^\top RR^\top\beta}\right|\left(\sqrt{C-\|\beta^{\mathcal{S}}\|_2^2}\right) \\
&\le C_3\left|\|\beta^{\mathcal{S}}\|_2^2+\|\hat{\beta}^{\mathcal{S}}\|_2^2\right|^{1/2}+\sqrt{C}\sup_{\beta\in\mathcal{B}}\left|\beta^\top(\hat{R}\hat{R}^\top-RR^\top)\beta\right|^{1/2} &&(33)\\
&\le C_3\left|\|\beta^{\mathcal{S}}\|_2^2+\|\hat{\beta}^{\mathcal{S}}\|_2^2\right|^{1/2}+\sqrt{C}\left(\sup_{\beta\in\mathcal{B}}\|\beta\|_2^2\,\|\hat{R}\hat{R}^\top-RR^\top\|_F\right)^{1/2}\xrightarrow{\mathbb{P}}0. &&(34)
\end{aligned}$$

The inequality in (33) follows from the compactness of $\mathcal{B}$, the fact that $\hat{R}\hat{R}^\top$ has bounded eigenvalues, and that $|\sqrt{x}-\sqrt{y}|\le|x-y|^{1/2}$ for all $x,y\ge0$. The inequality in (34) follows from Cauchy–Schwarz and that $\|A\|_2\le\|A\|_F$. The convergence in probability follows form Assumption D.1 and the compactness of $\mathcal{B}$.

E Details on finite-sample experiments

In this section, we provide more details of the data generation for our synthetic finite-sample experiments as well as data processing for the real-world data experiments.

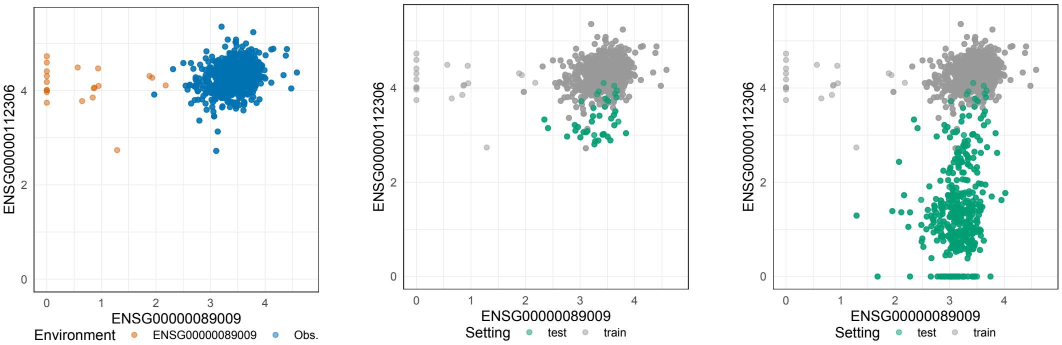
E.1 Synthetic experiments

For the synthetic experiments, we generate a random SCM which satisfies our assumptions. For $d=15$, we randomly sample the joint covariance $\Sigma^\star$ of (X,Y), fixing its total variance and the eigenvalues. We consider 7 environments including the reference environment, and for each environment except the reference, we randomly generate mean shifts of fixed norm. Since we have 6 non-zero random Gaussian mean shifts, it holds a.s. that $\dim \mathcal{S} = 6$. We then randomly generate an "initial guess" for $\beta^\star \in \mathbb{R}^d$ of fixed norm $C = 10$. Now, with respect to the space $\mathcal{S}$ of the identifiable directions induced by the mean shifts, we choose the most "adversarial" causal parameter $\beta^\star_{\text{adv}}$ which is equal to $\beta^\star$ on $\mathcal{S}$, but on $\mathcal{S}^\perp$ has the opposite direction of the noise OLS estimator ${\Sigma^\star_\eta}^{-1}\Sigma^\star_{\eta,\xi}$. We ensure that $\|\beta^\star_{\text{adv}}\|_2 = C$. Note that under the observed shifts, $\beta^\star$ and $\beta^\star_{\text{adv}}$ are observationally equivalent. We complete $\beta^\star_{\text{adv}}$ to the set θ_{adv} of observationally equivalent model parameters and generate the multi-environment training data according to θ_{adv} and the collection of mean shifts.

For Figure 3 (left), we define the test shift upper bound as $M_{\text{anchor}} = \gamma SS^\top$, where S is taken to be a two-dimensional subspace of $\mathcal{S}$. We vary γ from 0 to 10, and for each γ, we compute the oracle anchor regression estimator by minimizing the discrete anchor regression loss. Additionally, we compute the pooled OLS estimator and the identifiable robust predictor $\beta^{\text{rob,ID}}$ as described in Appendix D. Finally, we generate test data with a Gaussian additive shift $A_{\text{test}} \sim \mathcal{N}(0, M_{\text{anchor}})$. We evaluate the loss of β_{OLS}, β_{anchor} and $\beta^{\text{rob,ID}}$ on this test environment and include the population lower bound.

For Figure 3 (right), we define the test shift upper bound as $M_{\text{new}} = \gamma SS^\top + \gamma' RR^\top$, where R is a 2-dimensional subspace of the space $\mathcal{S}^\perp$ and we set $\gamma' = 0.05\gamma$ to showcase the effect of small unseen shifts compared to large identified shifts. We vary γ from 0 to 40 and for each γ, compute the oracle anchor regression estimator by minimizing the discrete anchor regression loss. Additionally, we compute the pooled OLS estimator and the identifiable robust predictor $\beta^{\text{rob,ID}}$ as described in Appendix D, for which we use knowledge of spaces S and R and prior knowledge of M_{new}. Finally, we generate test data with a Gaussian additive shift $A_{\text{test}} \sim \mathcal{N}(0, M_{\text{new}})$. We evaluate the loss of β_{OLS}, β_{anchor} and $\beta^{\text{rob,ID}}$ on this test environment, plot the resulting test losses for different estimators and include the population lower bound.

E.2 Real-world data experiments



(a) (b) (c)

Figure 5: The figures illustrate the structure of the (a) training-time shifts and (b-c) test-time shifts for different perturbation strengths on the example of two covariates. Panel (a) shows the training data containing two environments–observational (blue) and shifted (orange) corresponding to the knockout of the gene ENSG00000089009. Panels (b) and (c) show the training data in grey and test data from a previously unseen environment (green). Panel (b) depicts the top 10% test data points closest to the training support (perturbation strength = 0.1). Panel (c) illustrates the full test data (perturbation strength = 1.0).

We consider the K562 dataset from [41] and perform the preprocessing as done in [13]. The resulting dataset consists of $n = 162,751$ single-cell observations over $d = 622$ genes collected from observational and several interventional environments. The interventional environments arise by knocking down a single gene at a time using the CRISPR interference method [40]. Following [48], we select only always-active genes in the observational setting, resulting in a smaller dataset of 28 genes. For each gene $j = 1, \ldots, 28$, we set $Y := X_j$ as the target variable and select the three genes $X_{k_1}, \ldots, X_{k_3}$ most strongly correlated with Y (using Lasso), resulting in a dataset with columns $Y, X_{k_1}, \ldots, X_{k_3}$. Given this dataset, we construct the training and test environments as follows. Let $\mathcal{O}$ denote the 10,691 observations collected from the observational environment, and let $\mathcal{I}_i$ denote the observations collected from the interventional environment where the gene k_i was knocked down. We will denote by $\mathcal{I}_{i,s}$ the s-th quantile of datapoints in $\mathcal{I}_i$ w.r.t. to the expression value of the gene k_i. These are the $s \times 100\%$ of datapoints with the *weakest* shift compared to the observational mean of the gene k_i, and thus the parameter $s \in [0, 1]$ is a proxy for the *strength* of the shift. Furthermore, denote by $\mathcal{I}^*_{i,s}$ a random sample of $\mathcal{I}_{i,s}$ of a certain size. For each $i \in \{1, 2, 3\}$, we train the methods on $\mathcal{D}^{\text{train}}_i := \mathcal{O} \cup \mathcal{I}^*_{i,1}$, with $|\mathcal{I}^*_{i,1}| = 20$. An illustration of the training data $\mathcal{D}^{\text{train}}_i$ is shown in panel (a) of Figure 5. For each shift strength $s \in \{0.1, \ldots, 1\}$ we evaluate the models on the test samples from the three interventional environments. An example of the test data for different shift strengths s and a previously unseen direction is shown in Figure 5(b-c). Figure 4 shows the test MSE performance as a function of perturbation strength. We compare our method Rob-ID, defined as the minimizer of the empirical identifiable robust risk (17), with anchor regression [45], invariant causal prediction (ICP) [38], Distributional Robustness via Invariant Gradients (DRIG) [49], and OLS (corresponding to vanilla ERM). We use the following parameters for Rob-ID: $\gamma = 50$, $C_{\text{ker}} = 1.0$, and $M = \text{Id}$. For anchor regression and DRIG, we select $\gamma = 50$. For ICP, we set the significance level for the invariance tests to $\alpha = 0.05$.

These numerical experiments are computationally light and can be run in ≈ 5 minutes on a personal laptop.[13]

F Proofs

F.1 Proof of Proposition 1

For every environment $e \in \mathcal{E}_{\text{train}}$, we observe the first moments $\mathbb{E}\left(X_e\right)$ and $\mathbb{E}\left(Y_e\right)$, and second moments $\mathbb{E}\left(X_e X_e^\top\right)$, $\mathbb{E}\left(Y_e^2\right)$ and $\mathbb{E}\left(X_e Y_e\right)$. Since it holds by assumption that $\mu_0 = 0$ and $\Sigma_0 = 0$, we have that $\mathbb{E}\left(X_0 X_0^\top\right) = \Sigma^\star_\eta$, and so we can identify $\Sigma^\star_\eta$ uniquely. Furthermore, it holds that

$$\mathbb{E}\left(X_0 Y_0\right) = \Sigma^\star_\eta \beta^\star + \Sigma^\star_{\eta,\xi}, \tag{35}$$

$$\mathbb{E}\left(X_e Y_e\right) = (\Sigma_e + \mu_e \mu_e^\top + \Sigma^\star_\eta)\beta^\star + \Sigma^\star_{\eta,\xi}. \tag{36}$$

By taking the difference between Equation (36) and Equation (35), we can identify $(\Sigma_e + \mu_e \mu_e^\top)\beta^\star$. Thus, the causal parameter $\beta^\star$ is identifiable on the subspace $\mathcal{S}$ defined in Equation (10) and is not identifiable on its orthogonal complement $\mathcal{S}^\perp$. Thus, for any for vector $\alpha \in \mathcal{S}^\perp$, the vector $\beta = \beta^\star + \alpha$ is consistent with the data-generating process. It remains to compute the covariance parameters induced by an arbitrary $\tilde{\beta} := \beta^\star + \alpha$, for $\alpha \in \mathcal{S}^\perp$. For every environment $e \in \mathcal{E}_{\text{train}}$, the second mixed moment between X_e and Y_e has to satisfy the following equality

$$\mathbb{E}\left(X_e Y_e\right) = (\Sigma_e + \mu_e \mu_e^\top + \Sigma^\star_\eta)\beta^\star + \Sigma^\star_{\eta,\xi} = (\Sigma_e + \mu_e \mu_e^\top + \Sigma^\star_\eta)\tilde{\beta} + \tilde{\Sigma}_{\eta,\xi},$$

from which it follows that $\tilde{\Sigma}_{\eta,\xi} := \Sigma^\star_{\eta,\xi} - \Sigma^\star_\eta \alpha$. By computing $\mathbb{E}\left(Y_e^2\right)$ and inserting $\tilde{\beta} = \beta^\star + \alpha$ and $\tilde{\Sigma}_{\eta,\xi}$, we similarly obtain

$$\tilde{\sigma}_\xi^2 := (\sigma^\star_\xi)^2 - 2\alpha^\top \Sigma^\star_{\eta,\xi} + \alpha^\top \Sigma^\star_\eta \alpha.$$

Thus, we obtain the following set of observationally equivalent model parameters consistent with $\mathcal{P}_{\theta^\star, \mathcal{E}_{\text{train}}}$:

$$\Theta_{\text{eq}} = \{\beta^\star + \alpha, \Sigma^\star_\eta, \Sigma^\star_{\eta,\xi} - \Sigma^\star_\eta \alpha, (\sigma^\star_\xi)^2 - 2\alpha^\top \Sigma^\star_{\eta,\xi} + \alpha^\top \Sigma^\star_\eta \alpha \colon \alpha \in \mathcal{S}^\perp\}.$$

Since the observationally equivalent set is identifiable from the training distribution, but model parameters $\beta^\star$, $\Sigma^\star_{\eta,\xi}$, $(\sigma^\star_\xi)^2$ are not, it is helpful to re-express the observationally equivalent set through identifiable quantities. For this, we note that the "identified causal predictor" $\beta^{\mathcal{S}} = \beta^\star - \beta^{\mathcal{S}^\perp}$ induces an observationally equivalent model given by

$$\theta^{\mathcal{S}} := (\beta^{\mathcal{S}}, \Sigma^{\mathcal{S}}_\eta, \Sigma^{\mathcal{S}}_{\eta,\xi}, (\sigma^{\mathcal{S}}_\xi)^2) = (\beta^{\mathcal{S}}, \Sigma^\star_\eta, \Sigma^\star_{\eta,\xi} + \Sigma^\star_\eta \beta^{\mathcal{S}^\perp}, (\sigma^\star_\xi)^2 + 2\langle \Sigma^\star_{\eta,\xi}, \beta^{\mathcal{S}^\perp} \rangle + \langle \beta^{\mathcal{S}^\perp}, \Sigma^\star_\eta \beta^{\mathcal{S}^\perp} \rangle).$$

From this reparameterization, we infer the final form of the observationally equivalent set :

$$\Theta_{\text{eq}} = \{\beta^{\mathcal{S}} + \alpha, \Sigma'_\eta, \Sigma^{\mathcal{S}}_{\eta,\xi} - \Sigma'_\eta \alpha, (\sigma^{\mathcal{S}}_\xi)^2 - 2\alpha^\top \Sigma^{\mathcal{S}}_{\eta,\xi} + \alpha^\top \Sigma'_\eta \alpha \colon \alpha \in \mathcal{S}^\perp\} \ni \theta^\star$$

[13] We use a 2020 13-inch MacBook Pro with a 1.4 GHz Quad-Core Intel Core i5 processor, 8 GB of RAM, and Intel Iris Plus Graphics 645 with 1536 MB of graphics memory.

Therefore, Equation (12) follows. To find the robust predictor β^{rob}, we write down the robust loss with respect to M_{test} and any θ_α from the observationally equivalent set :

$$\begin{aligned}\mathcal{R}_{\text{rob}}(\beta; \theta_\alpha, M_{\text{test}}) &= (\beta^{\mathcal{S}} + \alpha - \beta)^\top (M_{\text{test}} + \Sigma^\star_\eta)(\beta^{\mathcal{S}} + \alpha - \beta) \\ &\quad + 2(\beta^{\mathcal{S}} + \alpha - \beta)^\top (\Sigma^\star_{\eta,\xi} - \Sigma^\star_\eta \alpha) + (\sigma^{\mathcal{S}}_\xi)^2 - 2\alpha^\top \Sigma^{\mathcal{S}}_{\eta,\xi} + \alpha^\top \Sigma^\star_\eta \alpha.\end{aligned}$$

inserting $\alpha \in \mathcal{S}^\perp$ and rearranging, Equation (13) follows.

F.2 Proof of Theorem 3.1

We structure the proof as follows: first, we quantify the non-identifiability of the robust risk by explicitly computing its supremum over the observationally equivalent set of the observationally equivalent model parameters (referred to as the identifiable robust risk). Second, we derive a lower bound for the identifiable robust risk by considering two cases depending on how a predictor $\overline{\beta}$ interacts with the possible test shifts M_{test}. In this proof, we use more general notation, with the test shifts bounded by a PSD matrix $M_{\text{test}} \preceq \gamma M + \gamma' RR^\top$, which $\text{range} M \subset \mathcal{S}$ and $\text{range} R \subset \mathcal{S}^\perp$. The statement of the theorem follows by setting $\gamma = \gamma'$. However, we believe that the more refined statement is useful, e.g., when one expects strong shifts in training directions and only weak "new" shifts.

Computation of the identifiable robust risk. For any model-generating parameter $\theta = (\beta, \Sigma)$ it holds that the robust risk of the model Equation (3) under test shifts $M_{\text{test}} \succeq 0$ is given by

$$\mathcal{R}_{\text{rob}}(\overline{\beta}; \theta, M_{\text{test}}) = (\beta - \overline{\beta})^\top (M_{\text{test}} + \Sigma^\star_\eta)(\beta - \overline{\beta}) + 2(\beta - \overline{\beta})^\top \Sigma_{\eta,\xi} + (\sigma_\xi)^2.$$

We recall that the observationally equivalent set of model parameters after observing the multi-environment training data Equation (3) is given by

$$\Theta_{\text{eq}} = \{\beta^{\mathcal{S}} + \alpha, \Sigma^\star_\eta, \Sigma^{\mathcal{S}}_{\eta,\xi} - \Sigma^\star_\eta \alpha, (\sigma^{\mathcal{S}}_\xi)^2 - 2\alpha^\top \Sigma^{\mathcal{S}}_{\eta,\xi} + \alpha^\top \Sigma_\eta \alpha : \alpha \in \mathcal{S}^\perp\}, \tag{37}$$

where $\mathcal{S}$ is the span of identified directions defined in Equation (10). Moreover, we recall that by Assumption 3.1, for any causal parameter β it should hold that $\|\beta\|_2 = \|\beta^{\mathcal{S}} + \alpha\|_2 \leq C$, which translates into the following constraint for the parameter α:

$$\|\alpha\|_2 \leq \sqrt{C^2 - \|\beta^{\mathcal{S}}\|_2^2} =: C_{\text{ker}}.$$

Inserting Equation (37) in Equation (8), we obtain

$$\mathcal{R}_{\text{rob,ID}}(\overline{\beta}; \Theta_{\text{eq}}, M_{\text{test}}) = \sup_{\substack{\alpha \in \mathcal{S}^\perp, \\ \|\alpha\|_2 \leq C_{\text{ker}}}} \mathcal{R}_{\text{rob}}(\overline{\beta}; \theta_\alpha, M_{\text{test}}),$$

where θ_α is a short notation for $(\beta^{\mathcal{S}} + \alpha, \Sigma^\star_\eta, \Sigma^{\mathcal{S}}_{\eta,\xi} - \Sigma^\star_\eta \alpha, (\sigma^{\mathcal{S}}_\xi)^2 - 2\alpha^\top \Sigma^{\mathcal{S}}_{\eta,\xi} + \alpha^\top \Sigma^\star_\eta \alpha)$. We now compute the supremum explicitly in case M_{test} has the form $M_{\text{test}} = \gamma M + \gamma' RR^\top$, where M is a PSD matrix with $\text{range } M \subseteq \mathcal{S}$ and R is a semi-orthogonal matrix with $\text{range } R \subseteq \mathcal{S}^\perp$. Note that this assumption includes both the setting of Theorem 3.1 and the setting of finite robustness methods in Section 3.2. For any $\alpha \in \mathcal{S}^\perp$, we write down the robust loss as

$$\begin{aligned}\mathcal{R}_{\text{rob}}(\overline{\beta}; \theta_\alpha, M_{\text{test}}) &= (\beta^{\mathcal{S}} - \overline{\beta})^\top (M_{\text{test}} + \Sigma^\star_\eta)(\beta^{\mathcal{S}} - \overline{\beta}) + 2(\beta^{\mathcal{S}} - \overline{\beta})^\top \Sigma^{\mathcal{S}}_{\eta,\xi} + (\sigma^{\mathcal{S}}_\xi)^2 \\ &\quad + \alpha^\top M_{\text{test}} \alpha + 2\alpha^\top M_{\text{test}} (\beta^{\mathcal{S}} - \overline{\beta}) \\ &= \mathcal{R}_{\text{rob}}(\overline{\beta}; \theta^{\mathcal{S}}, M_{\text{test}}) + \alpha^\top M_{\text{test}} \alpha + 2\alpha^\top M_{\text{test}} (\beta^{\mathcal{S}} - \overline{\beta}).\end{aligned}$$

The first term is the robust loss of $\overline{\beta}$ under test shift M_{test} and the identified model-generating parameter $\theta^{\mathcal{S}}$, thus it does not depend on α. By the structure of M_{test}, we obtain that

$$f(\alpha) := \alpha^\top M_{\text{test}} \alpha + 2\alpha^\top M_{\text{test}} (\beta^{\mathcal{S}} - \overline{\beta}) = \gamma' \alpha^\top RR^\top \alpha - \alpha^\top RR^\top \overline{\beta}.$$

If $R = 0$, i.e., the test shifts consist only of the identified directions, we have $f(\alpha) = 0$, independently of α, and thus

$$\mathcal{R}_{\text{rob,ID}}(\overline{\beta}; \Theta_{\text{eq}}, M_{\text{test}}) = \mathcal{R}_{\text{rob}}(\overline{\beta}; \theta^{\mathcal{S}}, M_{\text{test}}).$$

This implies the first statement of the theorem.

We now consider the case where $R \neq 0$, i.e., $RR^\top$ is a non-degenerate projection. Our goal is to maximize $f(\alpha)$ subject to constraints $\alpha \in \mathcal{S}^\perp$, $\|\alpha\|_2 \leq C_{\text{ker}}$. Let $\tilde{R}$ be an orthonormal extension of R such that $\text{range}\,(R|\tilde{R}) = \mathcal{S}^\perp$. Then, we can parameterize $\alpha \in \mathcal{S}^\perp$ as $\alpha = (R|\tilde{R})(\frac{w}{\tilde{w}})$ and the corresponding Lagrangian reads

$$\begin{aligned}\mathcal{L}(\alpha, \lambda) &= \gamma' \alpha^\top RR^\top \alpha - \alpha^\top RR^\top \overline{\beta} + \lambda(C_{\text{ker}}^2 - \|\alpha\|_2^2) \\ &= \gamma' \|w\|_2^2 - w^\top R^\top \overline{\beta} + \lambda(C_{\text{ker}}^2 - \|(w, \tilde{w})\|_2^2).\end{aligned}$$

Differentiating with respect to $w, \tilde{w}$ yields

$$\begin{aligned}w &= \frac{\gamma'}{\gamma' - \lambda} R^\top \overline{\beta}; \\ \tilde{w} &= 0.\end{aligned}$$

After differentiating w.r.t. λ, we obtain $\frac{\gamma'}{\gamma'-\lambda} = \pm \frac{C_{\text{ker}}}{\|R^\top \overline{\beta}\|_2}$. By inserting in the objective function and comparing, we obtain the **value of the identifiable robust risk**:

$$\mathcal{R}_{\text{rob,ID}}(\overline{\beta}; \Theta_{\text{eq}}, M_{\text{test}}) = \gamma' C_{\text{ker}}^2 + 2\gamma' \|R^\top \overline{\beta}\|_2 + \mathcal{R}_{\text{rob}}(\overline{\beta}; \theta^{\mathcal{S}}, M_{\text{test}}). \tag{38}$$

Putting together the cases $R = 0$ and $R \neq 0$, and plugging in $M_{\text{test}} = \gamma SS^\top + \gamma' RR^\top$, we obtain

$$\begin{aligned}\mathcal{R}_{\text{rob,ID}}(\overline{\beta}; \Theta_{\text{eq}}, M_{\text{test}}) &= \gamma' \mathbb{I}_{R \neq 0}(C_{\text{ker}} + \|R^\top \overline{\beta}\|_2)^2 + \mathcal{R}_{\text{rob}}(\overline{\beta}; \theta^{\mathcal{S}}, \gamma M) \\ &= \gamma' \mathbb{I}_{R \neq 0}(C_{\text{ker}} + \|R^\top \overline{\beta}\|_2)^2 + \gamma \|S^\top(\beta^{\mathcal{S}} - \overline{\beta})\|_2^2 + \mathcal{R}_0(\beta^{\mathcal{S}}, \overline{\beta}),\end{aligned}$$

where $\mathcal{R}_{\text{rob}}(\overline{\beta}; \theta^{\mathcal{S}}, \gamma M)$ is the robust risk of the estimator $\overline{\beta}$ w.r.t. the "identified" test shift γM and the identified model parameter $\theta^{\mathcal{S}}$, whereas $\mathcal{R}_0(\theta^{\mathcal{S}}, \overline{\beta})$ is the risk of $\overline{\beta}$ on the reference environment $e = 0$.

Derivation of the lower bound for the identifiable robust risk. Now that we have explicitly computed the identifiable robust risk, we devote ourselves to the computation of the lower bound for its best possible value

$$\inf_{\overline{\beta} \in \mathbb{R}^d} \mathcal{R}_{\text{rob,ID}}(\overline{\beta}; \Theta_{\text{eq}}, M_{\text{test}}).$$

In this part, we will only consider the case $R \neq 0$, since the case $R = 0$ corresponds to the (discrete) anchor regression-like setting, where both the robust risk and its minimizer are uniquely identifiable, and computable from training data. We will distinguish between two cases.

Case 1: $\|R^\top \overline{\beta}\|_2 = 0$. In this case, $\overline{\beta}$ is fully located in the orthogonal complement of R, which consists of $\mathcal{S}$ and $\tilde{R}$. We will denote (the basis of) this subspace by $S_{\text{tot}} = \mathcal{S} \oplus \tilde{R}$. Thus, S_{tot} is the "total" stable subspace consisting of identified directions in $\mathcal{S}$ and non-identified, but unperturbed directions $\tilde{R}$. We will parameterize $\overline{\beta}$ as $\overline{\beta} = S_{\text{tot}} w$. Thus, we are looking to solve the optimization problem

$$\beta^{\text{rob,ID}} = \arg\min_{w} (\beta^{\mathcal{S}} - S_{\text{tot}} w)^\top (\gamma SS^\top + \Sigma_\eta^\star)(\beta^{\mathcal{S}} - S_{\text{tot}} w) + 2(\beta^{\mathcal{S}} - S_{\text{tot}} w)^\top \Sigma_{\eta,\xi}^{\mathcal{S}} + (\sigma_\xi^{\mathcal{S}})^2.$$

Setting the gradient to zero yields the *asymptotic identifiable robust estimator*

$$\begin{aligned}\beta^{\text{rob,ID}} &= \beta^{\mathcal{S}} + S_{\text{tot}}[S_{\text{tot}}^\top(\gamma SS^\top + \Sigma_\eta)S_{\text{tot}}]^{-1} S_{\text{tot}}^\top \Sigma_{\eta,\xi}^{\mathcal{S}} \\ &= \beta^{\mathcal{S}} + S_{\text{tot}}[\gamma \text{Id}_{S_{\text{tot}}} + S_{\text{tot}}^\top \Sigma_\eta S_{\text{tot}}]^{-1} S_{\text{tot}}^\top \Sigma_{\eta,\xi}^{\mathcal{S}},\end{aligned} \tag{39}$$

which corresponds to the loss value of

$$\mathcal{R}_{\text{rob,ID}}(\beta^{\text{rob,ID}}; \Theta_{\text{eq}}, M_{\text{test}}) = \gamma' C_{\text{ker}}^2 + (\sigma_\xi^{\mathcal{S}})^2 - 2{\Sigma_{\eta,\xi}^{\mathcal{S}}}^\top S_{\text{tot}}[S_{\text{tot}}^\top(\gamma SS^\top + \Sigma_\eta)S_{\text{tot}}]^{-1} S_{\text{tot}}^\top \Sigma_{\eta,\xi}^{\mathcal{S}}.$$

Case 2:$\|R^\top \overline{\beta}\|_2 \neq 0$. Since for $\|R^\top \overline{\beta}\|_2 \neq 0$, the objective function is differentiable, we compute its gradient to be

$$\begin{aligned}\nabla \mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, M_{\text{test}}) &= 2\gamma' RR^\top \beta / \|RR^\top \beta\| + 2\gamma' RR^\top \beta + \nabla \mathcal{R}_{\text{rob}}(\beta; \theta^{\mathcal{S}}, \gamma M) \\ &= 2\gamma' RR^\top \beta / \|RR^\top \beta\| + 2\gamma' RR^\top \beta + 2(\Sigma_\eta^\star + \gamma M)(\beta - \beta^{\mathcal{S}}) - 2\Sigma_{\eta,\xi}^{\mathcal{S}}.\end{aligned}$$

This equation is, in general, not solvable w.r.t. β in closed form. Instead, we provide the limit of the optimal value of the function when the strength of the unseen shifts is small, i.e. $\gamma' \to 0$. We know that for $\gamma' = 0$, the minimizer of the identifiable robust risk is given by the anchor estimator

$$\beta_{\text{anchor}} = \beta^{\mathcal{S}} + (\Sigma_\eta^\star + \gamma M)^{-1} \Sigma_{\eta,\xi}^{\mathcal{S}}.$$

Thus, we lower bound the term $2\gamma' C_{\text{ker}} \|R^\top \beta\|$ by the scalar product $2\gamma' C_{\text{ker}} \langle R^\top \beta,\ R^\top \beta_{\text{anchor}} \rangle / \|\beta_{\text{anchor}}\|$ and expect it to be tight for small γ'. After inserting this lower bound in Equation (38) we obtain the minimizer of the lower bound of form

$$\beta_{LB} = \beta^{\mathcal{S}} + (\Sigma_\eta^\star + \gamma M + \gamma' RR^\top)^{-1} (\Sigma_{\eta,\xi}^{\mathcal{S}} - \gamma' C_{\text{ker}} RR^\top (\Sigma_\eta^\star + \gamma M)^{-1} \Sigma_{\eta,\xi}^{\mathcal{S}}).$$

We can now lower bound $\|RR^\top \beta_{LB}\|$ as

$$\|RR^\top \beta_{LB}\| \geq \|RR^\top (\Sigma_\eta^\star + \gamma M)^{-1} \Sigma_{\eta,\xi}^\star\| - o(\gamma'),$$

from which the rate for small γ' follows. If we set $\gamma = \gamma'$ and $M = SS^\top$, the claim (15) of the theorem follows. For Section 3.2, the lower bound directly implies optimality of the identifiable robust risk of the anchor estimator when the strength of the unseen shifts γ' is small. Additionally. if $\gamma = 0$, i.e. only unseen test shifts occur, we conclude that the OLS and anchor estimators have the same rates.

Lower bound γ_{th} for γ'. Finally, we want to derive a lower bound on the shift strength γ' such that for all $\gamma' \geq \gamma_{\text{th}}$ Case 1 of our proof is valid, i.e. it holds that $\beta^{\text{rob,ID}}$ is given by the closed form "abstaining" estimator (39). For this, we find γ_{th} such that for all $\gamma' \geq \gamma_{\text{th}}$ zero is contained in the subdifferential of $\mathcal{R}_{\text{rob,ID}}(\beta^{\text{rob,ID}}; \Theta_{\text{eq}}, M_{\text{test}})$ at $\beta^{\text{rob,ID}}$. Then the KKT conditions are met, and $\beta^{\text{rob,ID}}$ is the unique minimizer of the identifiable robust risk due to strong convexity of the objective. We compute the subdifferential to be

$$S = \gamma' C_{\text{ker}} \{RR^\top \beta : \|\beta\|_2 \leq 1\} + \nabla \mathcal{R}_{\text{rob}}(\beta^{\text{rob,ID}}; \theta^{\mathcal{S}}, \gamma M).$$

Since $\beta^{\text{rob,ID}}$ is the minimizer of $\mathcal{R}_{\text{rob}}(\beta; \theta^{\mathcal{S}}, \gamma M)$ under the constraint $R^\top \beta = 0$, the gradient is zero in $R^\perp$ and it remains to show that

$$\|RR^\top \nabla \mathcal{R}_{\text{rob}}(\beta^{\text{rob,ID}}; \theta^{\mathcal{S}}, \gamma M)\| \leq \gamma' C_{\text{ker}},$$

or

$$\gamma' \geq \|RR^\top \nabla \mathcal{R}_{\text{rob}}(\beta^{\text{rob,ID}}; \theta^{\mathcal{S}}, \gamma M)\| / C_{\text{ker}}.$$

Via an upper bound on the projected gradient, we derive the stricter condition

$$\gamma' \geq \frac{\|RR^\top \Sigma_{\eta,\xi}^{\mathcal{S}}\| (1 + \kappa(\Sigma_\eta^\star))}{C_{\text{ker}}},$$

where $\kappa(\Sigma_\eta^\star)$ is the condition number of the covariance matrix.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state our contributions relative to prior work in the abstract, in Section 1, and in Appendix A.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the abstract and in Section 1, we highlight the setting that we consider. We explicitly describe the assumptions in Section 2 and summarize the limitations in Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix F contains proofs of all results appearing in the main paper. Appendix B, Appendix C, and Appendix D are self-contained and contain derivations and proof of the results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix E provides all the necessary information to reproduce the experimental results presented in Section 3.2. We provide details on empirical estimation of the proposed loss function in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While we do not provide the code, the paper provides all necessary information on reproducing the experiment in Appendices D and E.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all details to understand the experimental results in Section 3.2 and Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   In the numerical experiment, shown in Figure 3, we provide the average test MSE and its 5% and 95%-quantiles over 100 repetitions for each method.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The numerical experiment described in Section 3.2 is computationally very light and can be run on a personal laptop in a few minutes. We describe this in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and confirm that our work conforms to it in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Even if our work addresses the theoretical limits of distributional robustness, we mention in the abstract and in Section 1 that the topic of distributional robustness is central to safety-critical applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The work develops a theoretical framework and considers synthetic experiments. Therefore, explicit safeguards do not seem applicable at this stage.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In the numerical experiment in Section 3.3, we cite the existing work that we compare to our framework and the dataset used. In running the numerical experiment, we reimplemented all the methods (including existing ones) for ease of comparison.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: At this stage, the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.