Prediction-Intervention Games and Invariant Sets

Linus Kühne
Seminar for Statistics and ETH AI Center
ETH Zurich
Zurich, Switzerland
linus.kuehne@ai.ethz.ch

Felix Schur
Seminar for Statistics
ETH Zurich
Zurich, Switzerland
felix.schur@stat.math.ethz.ch

Jonas Peters
Seminar for Statistics
ETH Zurich
Zurich, Switzerland
jonas.peters@stat.math.ethz.ch

Abstract

We consider the following two-player game: using observational data, the leader chooses a prediction function for a response variable Y from given covariates. The follower then reacts with an intervention on some covariates in the underlying structural causal model to maximize their own objective. The leader knows the intervention targets, but may have limited knowledge of the follower's objective. We call this setup a prediction-intervention game, a special case of a Stackelberg game. Finding an optimal strategy for the leader is generally difficult. To avoid severe performance loss, the leader may base their prediction on the causal parents of Y, or more generally on an invariant subset of covariates. We prove, for two common classes of follower objectives, that predictors based on the stable blanket, an invariant subset, are at least as good as those based on the causal parents. We further upper bound the leader's post-intervention risk by a worst-case risk over allowed interventions and strengthen existing distribution generalization results to analyze this bound: we give sufficient conditions under which stable-blanket predictors are worst-case optimal, and show by examples that these conditions cannot in general be dropped. Finally, we discuss practical strategies for settings with known and unknown graph, and test them on simulated and real-world data.

1 Introduction

We consider a two-player Stackelberg game for predicting a real-valued or binary response $Y \in \mathcal{Y}$ from covariates $X \in \mathcal{X}$: based on a training distribution (or data), a leader commits to a prediction function f, and a follower (an individual subject to the prediction) observes or queries f and intervenes on some covariates to optimize their own, possibly non-adversarial, objective. The test distribution thus emerges only after the prediction function has been deployed, and can depend on that function. In particular, training and test distributions need not coincide.

Example. Consider an insurer who prices policies with a function f predicting the probability that an individual will submit a claim. Suppose that f uses two covariates that are correlated with Y, but only one is causal: the car's number of safety features and the car's color. An individual i with access to f (such as through querying a website for a quote) may change their covariates from X_i to $\tilde{X}_i$ (yielding $\tilde{Y}_i$) to minimize $f(\tilde{X}_i)$ and obtain a smaller premium. Installing safety features changes both the predicted and true probability of filing a claim: $f(\tilde{X}_i) \neq f(X_i)$ and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

$\mathbb{P}(\tilde{Y}_i = 1) \neq \mathbb{P}(Y_i = 1)$. Choosing a car with a different color may instead yield $f(\tilde{X}_i) \neq f(X_i)$ without changing the probability of filing a claim: $\mathbb{P}(\tilde{Y}_i = 1) = \mathbb{P}(Y_i = 1)$. A similar phenomenon occurs in usage-based car insurance, where Lee et al. [36] show that temporarily monitoring drivers to price policies changes their behavior during that period, worsening prediction performance.

Prediction-intervention games. We formalize the problem as a causal version of a Stackelberg game [66], which we call a prediction-intervention game. The leader provides a prediction function f, and the follower reacts by choosing an intervention $e^*(f)$ in a structural causal model (SCM) that maximizes their objective: $e^*(f) \in \operatorname{argmax}_e V_e^{\text{foll}}(f)$. We model the allowed intervention targets as the children of an intervention node in the causal graph. Knowing the causal structure helps the leader design a prediction function that is robust under the follower's interventions, that is, find $\operatorname{argmin}_f R^{\text{lead}}_{e^*(f)}(f)$. In practice, the leader estimates R^{lead}_e from data and may have only partial knowledge about the follower's objective and possible interventions.

Related work. Strategic classification [5, 24] and its causal extensions [26, 41, 59] model followers changing their covariates at a cost to obtain favorable predictions. Unlike strategic classification, the causal extensions and our formulation model how these changes affect Y, using an SCM. In these causal extensions, the leader knows how the follower reacts to f, through a known cost function [26, 41] or by deploying predictors and observing the resulting data [59]. Our guarantees do not need this knowledge: they hold for all interventions that $\mathcal{G}_0$ and A allow for and for classes of follower objectives; the causal extensions provide no such guarantees. Performative prediction [46] studies the phenomenon that deployed models shift the data distribution, without specifying what causes the shift. Our game is an instance of this broader framework in which we specify how the follower causes the shift, namely by intervening in an SCM. In our game, the leader chooses f without observing the follower's response, so the retraining procedures analyzed in performative prediction cannot be used; our specification lets us derive guarantees before deployment. We adopt the leader-follower structure from Stackelberg games [66] but to the best of our knowledge, no methodology other than the one stated above is directly applicable to our setting. Our focus on invariant subsets connects prediction-intervention games to causal distribution generalization [9, 47, 51]; in particular, we analyze stable-blanket predictors [48], whose previous optimality guarantee concerns average risk among invariant subsets. We provide a tight bound, connecting our setting to worst-case distribution generalization [17, 34, 56], which specifies test distributions without a causal model. We also give conditions for the worst-case optimality of the stable-blanket predictor among all measurable predictors, over a family of distributions that, unlike in Rojas-Carulla et al. [51, Thm. 4], does not depend on the invariant subset. The stable-blanket predictor belongs to the family considered by existing methods [37, 63], which rely on conditional independencies in interventional distributions [61]. Their selected predictor need not be worst-case optimal in our setting. § B gives further details on existing work, e.g., on performative prediction, and lists additional work that we consider more distant than the above.

Contributions. (i) We formally introduce prediction-intervention games and show how the intervention view connects to modeling a group of self-optimizing individuals (§ 2). (ii) We prove that, for any fixed intervention, stable-blanket prediction is optimal among predictors that do not use a specific set of descendants of Y (namely, $\mathrm{DE}(Y)_{\text{forb}}$, see § 3.2), and show that, for two common classes of follower objectives, it weakly dominates parent-based prediction for the leader's game objective (§ 3). (iii) The worst-case risk over the possible interventions upper-bounds the leader's post-intervention risk and does not require knowledge of the follower's objective. This connects to distribution generalization and worst-case prediction. We strengthen existing results to give sufficient conditions under which predictors based on the stable blanket are worst-case optimal among all measurable predictors, and prove that in general these conditions cannot be dropped (§ 4). (iv) We propose learning methods that the leader can use when they do not know the causal graph (§ 5). (v) We test our methodology on simulated and real data (§ 6). The source code to reproduce all experiments is available under `github.com/LinusKuehne/prediction-intervention-games`.

2 Prediction-intervention games

The game. We assume that the leader's task is to perform prediction; in this paper, we consider both regression and binary classification. Let $Y \in \mathcal{Y}$ be a binary or continuous response with covariates $X = (X_1, \dots, X_d) \in \mathcal{X}$. We use $\mathcal{Y} := \{0, 1\}$ and $\mathcal{A} := [0, 1]$ for classification,

and $\mathcal{Y} \coloneqq \mathbb{R}$ and $\mathcal{A} \coloneqq \mathbb{R}$ for regression. The training data are generated by a structural causal model [SCM; 3, 44] $\mathbb{C}_0$ over (X, Y, E) with induced distribution $\mathbb{P}^0$ and DAG $\mathcal{G}_0$, in which E is a source node indexing environments. Each $e \in \mathcal{E}_0 \coloneqq \operatorname{supp}_{\mathbb{C}_0}(E)$ corresponds to a distribution $\mathbb{P}^0_e \coloneqq \mathbb{P}^0(\cdot \mid E = e)$ over (X, Y). The leader has access to one or more training distributions $\mathbb{P}^0_e$ (or i.i.d. samples thereof) for $e \in \mathcal{E}_{\mathrm{tr}} \subseteq \mathcal{E}_0$, and chooses a prediction function $f \colon \mathcal{X} \to \mathcal{A}$. We assume that $\mathcal{G}_0$ is known unless stated otherwise. After observing f, the follower reacts by strategically choosing an environment $e \in \mathcal{E}$ that maximizes a utility $V^{\mathrm{foll}}_e(f)$. Here, $\mathcal{E}$ is a set of environments indexing distributions $\mathbb{P}_e$ over (X, Y) that can occur after deployment of f; these environments correspond to intervention distributions and we even allow for $\mathcal{E} \supsetneq \mathcal{E}_0$ (below, we provide more details on the follower's interventions). After the follower has chosen $e \in \mathcal{E}$, the leader incurs risk $R^{\mathrm{lead}}_e(f)$. This defines a prediction-intervention game:[1]

$$f^* \in \operatorname*{argmin}_{f \colon \mathcal{X} \to \mathcal{A}} \sup_{e \in \mathcal{E}^*(f)} R^{\mathrm{lead}}_e(f), \qquad \mathcal{E}^*(f) \coloneqq \operatorname*{argmax}_{e \in \mathcal{E}} V^{\mathrm{foll}}_e(f). \tag{1}$$

We assume throughout that $\mathcal{E}^*(f) \neq \varnothing$ and that the follower can evaluate $V^{\mathrm{foll}}_e(f)$ for all $e \in \mathcal{E}$, for example by knowing $\mathbb{C}_0$ or by sampling those variables from each[2] $\mathbb{P}_e$ that are required to evaluate $V^{\mathrm{foll}}_e(f)$, possibly not containing Y. We detail R^{lead}, V^{foll}, and the construction of $\mathcal{E}$ below.

Notation. For a node j in a DAG $\mathcal{H}$, let $\mathrm{PA}_{\mathcal{H}}(j), \mathrm{CH}_{\mathcal{H}}(j), \mathrm{DE}_{\mathcal{H}}(j), \mathrm{ND}_{\mathcal{H}}(j)$ denote its parents, children, descendants, and non-descendants (all excluding j), respectively. We drop the subscript if it is clear from the context. We sometimes write $k \to j$ as a shorthand for an edge $X_k \to X_j$. For $S \subseteq \{1, \dots, d, Y\}$, we define $X_S \coloneqq (X_j)_{j \in S}$ with $X_Y \coloneqq Y$. Slightly abusing notation, S^C is always defined as $\{1, \dots, d\} \setminus S$. Expectations under $\mathbb{P}^0_e$ and $\mathbb{P}_e$ are denoted by $\mathbb{E}^0_e$ and $\mathbb{E}_e$.

The leader's evaluation loss. We assume that the leader evaluates a predictor $f \colon \mathcal{X} \to \mathcal{A}$ using a loss $\ell \colon \mathcal{Y} \times \mathcal{A} \to [0, \infty]$ strictly consistent for the mean functional [20]: for every distribution F on $\mathcal{Y}$ with finite mean $\mathbb{E}_F[Y]$ and for which the expectations below are finite, and for all $q \in \mathcal{A}$, $\mathbb{E}_F[\ell(Y, \mathbb{E}_F[Y])] \leq \mathbb{E}_F[\ell(Y, q)]$, with equality iff $q = \mathbb{E}_F[Y]$. Examples include the squared loss $\ell(y, q) = (y - q)^2$ for regression and, for binary classification, any strictly proper scoring rule such as the Brier score $\ell(y, q) = (y - q)^2$ or the log loss [21]. Strict consistency ensures that the risk $R^{\mathrm{lead}}_e(f) \coloneqq \mathbb{E}_e[\ell(Y, f(X))]$ is uniquely minimized by $f(x) = \mathbb{E}_e[Y \mid X = x]$ (thus, it plays a similar role to the orthogonality of the conditional expectation, but extends to other loss functions). We assume throughout that Y has finite mean under each $\mathbb{P}_e$ and that all risks in our results are finite.

Follower's objective. The follower's objective need not coincide with the leader's: while the leader evaluates f by $R^{\mathrm{lead}}_e(f)$ and prefers low risk, the follower selects $e \in \mathcal{E}$ to maximize a separate utility $V^{\mathrm{foll}}_e(f) \in \mathbb{R}$. For instance, the leader may use $R^{\mathrm{lead}}_e(f) = \mathbb{E}_e[(Y - f(X))^2]$, while the follower maximizes an objective such as $\mathbb{E}_e[Y - f(X)]$ or $\mathbb{E}_e[f(X)]$.

Follower's interventions. The follower reacts to f by strategically intervening on the children of E in $\mathbb{C}_0$, possibly adding new edges to $\mathcal{G}_0$. This yields a deployment SCM $\mathbb{C}$ over (X, Y, E) that extends $\mathbb{C}_0$, with induced distribution $\mathbb{P}$ and DAG $\mathcal{G}$. Each $e \in \mathcal{E} \coloneqq \operatorname{supp}_{\mathbb{C}}(E)$ yields $\mathbb{P}_e \coloneqq \mathbb{P}(\cdot \mid E = e)$ over (X, Y). The source node E indicates the potential intervention targets: an edge $E \to X_j$ in $\mathcal{G}_0$ means the follower may intervene on the mechanism of X_j. We assume no edge $E \to Y$. Compared to training, for each $j \in \mathrm{CH}_{\mathcal{G}_0}(E)$, the follower may let X_j depend on its parents $X_{\mathrm{PA}_{\mathcal{G}_0}(j)}$ in $\mathcal{G}_0$, which may include Y, and on additional variables from an action set $A_j \subseteq \{1, \dots, d\}$ of permitted new parents; we write $A \coloneqq (A_1, \dots, A_d)$. Thus, for $k \in A_j$, the new mechanism of X_j may use X_k as an input (so that $\mathcal{G}$ contains the additional edge $X_k \to X_j$) but does not alter the mechanism of X_k. Formally, $\mathbb{C}$ is obtained from $\mathbb{C}_0$ by replacing, for each $j \in \mathrm{CH}_{\mathcal{G}_0}(E)$, the structural assignment of X_j with a measurable function of $X_{\mathrm{PA}_{\mathcal{G}_0}(j) \cup S_j}$ for some $S_j \subseteq A_j$, and independent exogenous noise. Since the new assignment may be constant, hard interventions are included in this intervention model, as are randomized and soft interventions.[3] All

[1] Instead of the supremum, one could also average $R^{\mathrm{lead}}_e(f)$ over the follower's best responses $e \in \mathcal{E}^*(f)$.

[2] An individual follower may also condition on covariates that cannot be changed, such as a birth year; see § 2.1.

[3] The results in § 3 and § 4 concern population predictors that are conditional means of Y given a subset of the covariates. If these are continuous functions and the support of X under every $e \in \mathcal{E}$ lies in the union of the training supports, the training distributions identify them. When predictors are learned from data, interventions can move the covariates outside the training support, where learning requires extrapolation assumptions that we do not address.

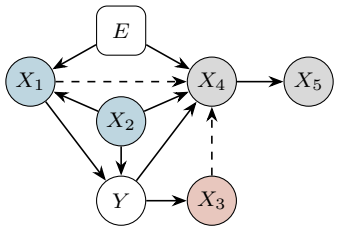


Figure 1: Running example. The solid edges form the training DAG $\mathcal{G}_0$, in which the children X_1 and X_4 of E are the intervention targets. During training, X_4 is a function of X_2, Y, and noise; the follower may replace this assignment by a function of X_1, X_2, X_3, Y, and new noise, which adds the dashed edges and yields the deployment DAG $\mathcal{G}$. Colors: parents $\mathrm{PA}_{\mathcal{G}}(Y)$ (blue), stable blanket $\mathrm{SB}_{\mathcal{G}}(Y)$ (blue and red), and forbidden descendants $\mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}}$ (gray), defined in § 3.2.

other structural assignments are unchanged. The deployment DAG $\mathcal{G}$ is the DAG of $\mathbb{C}$, which we assume to be acyclic. Choosing $S_j = \varnothing$ for all j together with the original assignments recovers $\mathbb{C}_0$, so $\mathcal{E}_{\mathrm{tr}} \subseteq \mathcal{E}_0 \subseteq \mathcal{E}$ and $\mathbb{P}_e = \mathbb{P}_e^0$ for $e \in \mathcal{E}_0$. Larger A_j allow for more flexible interventions. Fig. 1 shows our running example of $\mathcal{G}_0$ and $\mathcal{G}$, with $A_1 = \varnothing$ and $A_4 = \{1, 3\}$.

Unknown graph. In § 5, we drop the assumption that the graph is known to the leader. In that case, we assume access to training data from several training environments $\mathcal{E}_{\mathrm{tr}} \subseteq \mathcal{E}_0$ with $|\mathcal{E}_{\mathrm{tr}}| > 1$. For each $e \in \mathcal{E}_{\mathrm{tr}}$, we observe n_e i.i.d. draws $(X_i, Y_i)_{i=1}^{n_e} \sim \mathbb{P}_e^0$.

As far as we know, the formulation (1) is novel and determining its value $\sup_{e \in \mathcal{E}^*(f^*)} R_e^{\mathrm{lead}}(f^*)$ is relevant. § A gives two real-world examples. Prediction-intervention games are special instances of bilevel optimization problems [69], which are known to be hard to solve in general [2]. In § 3 and § 4 we make first steps by connecting the problem to invariant sets and distribution generalization, respectively, and strengthening and applying results from these fields.

2.1 Individuals and population interventions

The follower of § 2 chooses mechanisms for a whole population; we now show that the framework still applies when considering individual followers who each choose values for the intervention targets given their own realization of the observed variables. In this section, all graph quantities are taken with respect to $\mathcal{G}_0$. An individual follower may observe a realization of X_O for some $O \subseteq \mathrm{ND}(\mathrm{CH}(E))$, the variables among $X_1, \dots, X_d, Y$ that are non-descendants of every target (in particular, $O \cap \mathrm{CH}(E) = \emptyset$), and is allowed to choose deterministic values for $\mathrm{CH}(E)$, taking into account the observation $X_O = x_O$. For simplicity, let us assume that for all $j \in \mathrm{CH}(E)$, we have $\mathrm{PA}(j) \setminus \{E\} \subseteq O \cup \mathrm{CH}(E)$, so the follower has access to all parents of j when setting j's value, and $O \setminus \mathrm{PA}(j) \subseteq \{1, \dots, d\}$, that is, Y is observed if and only if it is a parent of j. Let us further assume that the follower maximizes a utility of the form $\mathbb{E}_e[k(f, X, Y) \,|\, X_O = x_O]$. We now consider a solution $(c_j)_{j \in \mathrm{CH}(E)}$ maximizing the individual follower's objective

$$\mathbb{E}_{(c_j)_{j \in \mathrm{CH}(E)}}[k(f, X, Y) \,|\, X_O = x_O], \tag{2}$$

where the leftmost subscript indicates that the $X_{\mathrm{CH}(E)}$ are intervened on and set to c_j, $j \in \mathrm{CH}(E)$. The optimal constants may vary with x_O, so there are functions $(g_j)_{j \in \mathrm{CH}(E)}$ such that for all x_O, they maximize $\mathbb{E}_{(g_j)_{j \in \mathrm{CH}(E)}}[k(f, X, Y) \,|\, X_O = x_O]$; here the subscript indicates that the structural assignments of $X_{\mathrm{CH}(E)}$ are set to the g_j's. Thus, the same $(g_j)_{j \in \mathrm{CH}(E)}$ maximize $\mathbb{E}_{(g_j)_{j \in \mathrm{CH}(E)}}[k(f, X, Y)]$, that is, they solve the problem of a population follower with $A_j := O \setminus \mathrm{PA}(j)$.[4] Thus, taking the average over many individual followers optimizing (2) has, from the perspective of the leader and the game's value, the same effect as considering a single population follower in an augmented graph. In § C.1, we relax the assumption of a single O to different O_j for $j \in \mathrm{CH}(E)$, still satisfying $O_j \subseteq \mathrm{ND}(j)$. This relaxation is relevant for considering individual followers in Fig. 1 and the experiments in § 6.1 and § F.5.

3 Strategies based on invariant sets

A natural strategy is to restrict the predictor f to a subset of covariates $S \subseteq \{1, \dots, d\}$. We first discuss the common choice $S = \mathrm{PA}(Y)$, which is a special case of choosing f based on an invariant

[4] Since $\mathrm{PA}(j) \setminus \{E\} \subseteq O \cup \mathrm{CH}(E)$, any population strategy with these A_j sets $X_{\mathrm{CH}(E)}$ as a function of X_O and independent noise; given $X_O = x_O$, its objective is therefore an average over constant choices, which the g_j maximize.

set. Throughout this section, we work with the augmented SCM $\mathbb{C}$, DAG $\mathcal{G}$, and test family $\mathcal{E}$ from § 2. Quantities such as $\mathrm{PA}(Y)$ and $\mathrm{CH}(Y)$ (see § 2), and the sets $\mathrm{DE}(Y)_{\mathrm{forb}}$ and $\mathrm{SB}(Y)$ defined below, are computed in $\mathcal{G}$. § 3.3 relates these to the training-stage DAG $\mathcal{G}_0$. In the remainder of the paper, we write R_e for the leader's risk R_e^{lead}.

3.1 Strategies based on causal parents

Using predictors f based on the causal parents of Y is common in causal strategic prediction [68] and has the following attractive property: if the follower's utility depends on the intervention only through the prediction $f(X)$, then a parent-based predictor can be affected only by interventions changing the distribution of $X_{\mathrm{PA}(Y)}$. 'Gaming' in the sense of manipulating descendants of Y to change the prediction of f without inducing a change in Y is therefore impossible. Restricting f to $X_{\mathrm{PA}(Y)}$ can thus incentivize followers to genuinely improve their outcomes.

Predictors based on $\mathrm{PA}(Y)$ further satisfy a simple robustness guarantee. Since $Y \notin \mathrm{CH}(E)$, the conditional distribution $Y \mid X_{\mathrm{PA}(Y)}$ is the same in all $e \in \mathcal{E}$. We refer to such subsets as invariant (see Def. 1 below). Thus, $f_{\mathrm{PA}(Y)}(x) \coloneqq \mathbb{E}[Y \mid X_{\mathrm{PA}(Y)} = x_{\mathrm{PA}(Y)}]$ is environment-independent. Hence, if we restrict f to the parents, the follower can induce a covariate shift in $X_{\mathrm{PA}(Y)}$; the optimal regression function using $X_{\mathrm{PA}(Y)}$ does not change. In the special additive-noise case $Y = g(X_{\mathrm{PA}(Y)}) + \varepsilon$ with $\varepsilon \perp\!\!\!\perp X_{\mathrm{PA}(Y)}$ and $\mathbb{E}[\varepsilon] = 0$, the follower's interventions are inconsequential for parent-based prediction under the squared loss: for all $e \in \mathcal{E}$, $R_e(f_{\mathrm{PA}(Y)}) = \mathrm{Var}(\varepsilon)$. Without this additive-noise structure, the risk may still vary across environments through the marginal distribution of $X_{\mathrm{PA}(Y)}$, but the variation is only due to covariate shift under a fixed conditional $Y \mid X_{\mathrm{PA}(Y)}$.

The above arguments do not apply exclusively to $X_{\mathrm{PA}(Y)}$. Similar reasoning extends to other invariant subsets that contain children of Y that the follower cannot manipulate. We now define invariant sets and show that there is an invariant set, the stable blanket, that improves upon the parents.

3.2 Stable blankets are never worse than the causal parents

Definition 1 (Invariant subsets and predictors)**.** *A subset of covariates $S \subseteq \{1, \dots, d\}$ is called* invariant *if $Y \perp\!\!\!\perp E \mid X_S$ under $\mathbb{P}$, or, equivalently, if $Y \mid X_S$ is the same under all $\mathbb{P}_e$, $e \in \mathcal{E}$. For all invariant S, we define the predictor $f_S : \mathcal{X} \to \mathcal{A}$ as $f_S(x) \coloneqq \mathbb{E}[Y \mid X_S = x_S]$.*

The stable blanket of Y is an invariant subset capturing as much information about Y as possible while excluding irrelevant covariates. In this sense, it resembles the Markov blanket[5] $\mathrm{MB}(Y)$: $f_{\mathrm{MB}(Y)}$ minimizes the risk in a single environment, but $\mathrm{MB}(Y)$ may contain intervened children of Y and need not be invariant. To define the stable blanket, let $\mathrm{CH}_{\mathrm{int}}(Y) \coloneqq \mathrm{CH}(Y) \cap \mathrm{CH}(E)$ denote the intervened children of Y, and define

$$\mathrm{DE}(Y)_{\mathrm{forb}} \coloneqq \mathrm{CH}_{\mathrm{int}}(Y) \cup \mathrm{DE}(\mathrm{CH}_{\mathrm{int}}(Y)).$$

The reader may think of these nodes as 'forbidden descendants' of Y. The *stable blanket* [48, Def. 3.4] is a set satisfying $\mathrm{PA}(Y) \subseteq \mathrm{SB}(Y) \subseteq \mathrm{MB}(Y)$; it is defined as

$$\mathrm{SB}(Y) \coloneqq \mathrm{PA}(Y) \cup \big(\mathrm{CH}(Y) \setminus \mathrm{DE}(Y)_{\mathrm{forb}}\big) \cup \big(\mathrm{PA}\big(\mathrm{CH}(Y) \setminus \mathrm{DE}(Y)_{\mathrm{forb}}\big) \setminus \{Y\}\big). \tag{3}$$

The d-separation $E \perp_{\mathcal{G}} Y \mid \mathrm{SB}(Y)$ implies that $\mathrm{SB}(Y)$ is invariant by the global Markov property. In the deployment DAG $\mathcal{G}$ in Fig. 1, $\mathrm{DE}(Y)_{\mathrm{forb}} = \{4, 5\}$ (gray) and $\mathrm{SB}(Y) = \{1, 2, 3\}$ (blue and red).

In previous work, optimality of the stable blanket has been discussed only as an average optimality among invariant sets [48, Thm. 3.5]. We now prove a stronger result: among predictors that do not use information from $X_{\mathrm{DE}(Y)_{\mathrm{forb}}}$, $f_{\mathrm{SB}(Y)}$ minimizes the risk in all environments.

Lemma 2. *For all $\sigma(X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}})$-measurable $f \colon \mathcal{X} \to \mathcal{A}$ and all $e \in \mathcal{E}$, we have $R_e(f) \geq R_e(f_{\mathrm{SB}(Y)})$. In particular, for all $e \in \mathcal{E}$, $R_e(f_{\mathrm{PA}(Y)}) \geq R_e(f_{\mathrm{SB}(Y)})$.*

Proofs of all results are in § E. Lem. 2 implies a subset-level formulation, recovering [48, Thm. 3.5] under faithfulness; see § C.3. The leader's question is which f minimizes $\sup_{e \in \mathcal{E}^*(f)} R_e(f)$ in the

[5] In $\mathcal{G} \setminus \{E\}$ (with E and its incident edges removed), the *Markov blanket* (also called Markov boundary in other literature; we follow the terminology of Pfister et al. [48]) $\mathrm{MB}(Y) = \mathrm{PA}(Y) \cup \mathrm{CH}(Y) \cup (\mathrm{PA}(\mathrm{CH}(Y)) \setminus \{Y\})$ is the smallest covariate set S d-separating Y from all covariates outside S [43, § 3.3].

prediction-intervention game (1). Lem. 2 does not immediately answer this since $\mathcal{E}^*(f)$ depends on f. But it shows that for two common follower objectives, stable blankets beat parents.

Theorem 3 (Stable blanket vs. parents)**.** *If (i)* $\forall e \in \mathcal{E}\ V_e^{\mathrm{foll}}(f) = R_e(f)$ *or (ii)* $\forall e \in \mathcal{E}\ V_e^{\mathrm{foll}}(f) = \mathbb{E}_e[\,\alpha + \beta\, f(X) - c_e(X)\,]$ *for* $\alpha, \beta \in \mathbb{R}$ *and measurable* $c_e \colon \mathcal{X} \to \mathbb{R}$ *not depending on* f*, then*

$$\sup_{e \in \mathcal{E}^*(f_{\mathrm{SB}(Y)})} R_e(f_{\mathrm{SB}(Y)}) \leq \sup_{e \in \mathcal{E}^*(f_{\mathrm{PA}(Y)})} R_e(f_{\mathrm{PA}(Y)}).$$

Case (i) describes an adversarial follower maximizing the leader's risk. Case (ii) is common in strategic prediction [24]: the follower maximizes $f(X)$ while incurring a cost $c_e(X)$ for the shift. Thus, while the optimal f for the leader is unknown, $f_{\mathrm{SB}(Y)}$ weakly dominates $f_{\mathrm{PA}(Y)}$ for these two utility classes. For other utilities the ordering can reverse (Ex. 9).

3.3 Conditions for equality of stable blanket during training and deployment

The previous results use the DAG $\mathcal{G}$, which extends the training-time DAG $\mathcal{G}_0$ by including edges $X_k \to X_j$ for $j \in \mathrm{CH}(E)$ and $k \in A_j$ that encode interventions by followers. If the leader only knows $\mathcal{G}_0$, how can they compute $\mathrm{SB}(Y) = \mathrm{SB}_{\mathcal{G}}(Y)$? Would they need to anticipate which edges strategic followers add once f is deployed? The next lemma shows that this is not necessary: under weak assumptions on the A_j's (the sets of permitted new parents, see § 2) the stable blanket is the same in $\mathcal{G}_0$ and $\mathcal{G}$, so $\mathrm{SB}(Y)$ can be read off from the training-stage DAG $\mathcal{G}_0$ alone.

Lemma 4. *Assume that for all* $j \in \mathrm{CH}_{\mathcal{G}_0}(E)$*,* $A_j \subseteq \mathrm{DE}_{\mathcal{G}_0}(Y)^C_{\mathrm{forb}} \setminus \{Y, E\}$*, and that* $\mathcal{G}$ *is acyclic. Then* $\mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}} = \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$ *and* $\mathrm{SB}_{\mathcal{G}}(Y) = \mathrm{SB}_{\mathcal{G}_0}(Y)$*. In particular,* $Y \perp_{\mathcal{G}} E \,|\, \mathrm{SB}_{\mathcal{G}_0}(Y)$.

The set A_j, $j \in \mathrm{CH}_{\mathcal{G}_0}(E)$, restricts only the new parents that the follower may add to X_j in an intervention; the parents of X_j in $\mathcal{G}_0$, which may include Y, are allowed to remain parents in $\mathcal{G}$. Acyclicity of $\mathcal{G}$ is a timing condition: it rules out interventions that depend, directly or indirectly, on their own outputs. § C.8 discusses the assumptions of Lem. 4 further. In Fig. 1, for example, $A_1 = \varnothing$ and $A_4 = \{1, 3\}$ satisfy the assumption, since $\mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}} = \{4, 5\}$ and $\mathcal{G}$ is acyclic, so the dashed edges leave $\mathrm{SB}(Y)$ unchanged. Lem. 4 is the only result in this paper that uses this assumption; the other results in § 3 and § 4 hold for a fixed DAG $\mathcal{G}$.

§ 4 discusses the question whether $f_{\mathrm{SB}(Y)}$ is worst-case optimal among all measurable predictors and connects this question to standard distribution generalization. This worst-case setting corresponds to an adversarial follower, see Thm. 3 (i). Lem. 2, which holds in every environment, and Thm. 3 (ii), in which the follower is not adversarial, are not worst-case results.

4 Distribution generalization and an upper bound

§ 3 shows that no $\sigma(X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}})$-measurable predictor improves on $f_{\mathrm{SB}(Y)}$ in any environment, and that, for two natural classes of follower objectives V^{foll}, $f_{\mathrm{SB}(Y)}$ is a better choice than $f_{\mathrm{PA}(Y)}$ for the leader in the prediction-intervention game (1) (see Lem. 2 and Thm. 3). However, in general, the optimal leader strategy depends on V^{foll} (§ C.2), which may be unknown to the leader and may vary across followers. In this section, we instead analyze an upper bound that comes with two benefits: it holds for all V^{foll} and allows for easier optimization than the game objective. For all f, we have

$$\sup_{e \in \mathcal{E}^*(f)} R_e(f) \leq \sup_{e \in \mathcal{E}} R_e(f). \tag{4}$$

The bound is attained when the follower maximizes $V_e^{\mathrm{foll}}(f) = R_e(f)$, and can be loose when the objectives are different (see Rmk. 13). Minimizing the right-hand side is a distribution generalization (DG) problem. For the rest of the section, we focus on the worst-case problem

$$f^* \in \operatorname*{argmin}_{f \colon \mathcal{X} \to \mathcal{A}} \sup_{e \in \mathcal{E}} R_e(f). \tag{5}$$

This differs from the standard DG objective [1, 9, 51] in the uncertainty set: in our work, $\mathcal{E}$ is generated by strategic follower interventions after deployment, possibly including additional edges into intervened variables, which subsumes the standard DG problem. Still, the bound (4) makes prediction-intervention games concrete application scenarios for existing worst-case methods. In the following, we extend and strengthen existing DG results to make them applicable in prediction-intervention games. Our new results also hold under the augmentation construction in § 2 in that they apply to $(\mathbb{C}, \mathcal{G}, \mathcal{E})$ and to $(\mathbb{C}_0, \mathcal{G}_0, \mathcal{E}_0)$ alike. We formulate our results using the former notation.

4.1 Optimality of the stable blanket

By Lem. 2, we know that the stable blanket solves the worst-case problem (5) over predictors that do not use $X_{\mathrm{DE}(Y)_{\mathrm{forb}}}$. We now ask whether $f_{\mathrm{SB}(Y)}$ is worst-case optimal among all measurable predictors $f\colon \mathcal{X} \to \mathcal{A}$.[6] We will prove in § 4.2 that this is not always the case: $\mathrm{SB}(Y)$ excludes descendants of $\mathrm{CH}_{\mathrm{int}}(Y)$ but there are examples in which these covariates carry substantial information about Y and are only weakly affected by the intervention. We first introduce two assumptions under which worst-case optimality does hold (and which prevent such counterexamples).

Assumption A1 (Strong interventions on $\mathrm{DE}(E)$)**.** *For all $e \in \mathcal{E}$ there exists $e' \in \mathcal{E}$ such that (i) $X_{\mathrm{SB}(Y)}$ has the same marginal distribution under $\mathbb{P}_e$ and $\mathbb{P}_{e'}$, and (ii)*

$$Y \perp\!\!\!\perp X_{\mathrm{DE}(Y)_{\mathrm{forb}}} \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}} \text{ under } \mathbb{P}_{e'}.$$

Asm. A1 is inspired by the confounding-removing interventions of Christiansen et al. [9] (see also 55). Intuitively, (ii) asks for an environment e' in which $\mathbb{E}_{e'}[Y \mid X] = f_{\mathrm{SB}(Y)}(X)$. This is not always satisfied: it requires the environment to strongly influence descendants of intervened children that are not directly affected by interventions (see Ex. 7 for a setting violating this condition and § C.8 for a discussion of Asm. A1 and A2). A sufficient condition for Asm. A1 is that no such nodes exist and interventions on $\mathrm{CH}_{\mathrm{int}}(Y)$ are strong.

Assumption A2 (Strong interventions on $\mathrm{CH}_{\mathrm{int}}(Y)$ with graphical condition)**.** *In $\mathcal{G}$, we have*

$$\mathrm{CH}(Y) \cap \mathrm{DE}(\mathrm{CH}_{\mathrm{int}}(Y)) \subseteq \mathrm{CH}_{\mathrm{int}}(Y) \tag{6}$$

and for all $e \in \mathcal{E}$ there exists $e' \in \mathcal{E}$ for which the assignment of every X_j, $j \in \mathrm{CH}_{\mathrm{int}}(Y)$, is replaced by exogenous noise independent of all other noise variables, with all other assignments unchanged.

Either assumption ensures worst-case optimality of the stable blanket.

Theorem 5 (Worst-case optimality of $\mathrm{SB}(Y)$)**.** *If either Asm. A1 or A2 holds, then for all measurable $f\colon \mathcal{X} \to \mathcal{A}$, $f_{\mathrm{SB}(Y)}$ solves Problem* (5)*:*

$$\sup_{e\in\mathcal{E}} R_e(f) \geq \sup_{e\in\mathcal{E}} R_e(f_{\mathrm{SB}(Y)}). \tag{7}$$

Worst-case optimality of $f_{\mathrm{SB}(Y)}$ still holds if strong interventions are achieved only asymptotically, which can be relevant in some linear SCMs.

Proposition 6 (Asymptotic relaxation of Asm. A1)**.** *For all $e \in \mathcal{E}$, define the predictor $\mu_e(x) := \mathbb{E}_e[Y \mid X = x]$. If for all $e \in \mathcal{E}$ there exists a sequence $(e'_j)_{j\geq 1} \subseteq \mathcal{E}$ such that (i) $X_{\mathrm{SB}(Y)}$ has the same marginal distribution under $\mathbb{P}_e$ and $\mathbb{P}_{e'_j}$ for all j, and (ii) $R_{e'_j}(f_{\mathrm{SB}(Y)}) - R_{e'_j}(\mu_{e'_j}) \xrightarrow{j\to\infty} 0$, then* (7) *holds for all measurable $f\colon \mathcal{X} \to \mathcal{A}$.*

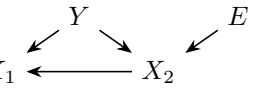
Thm. 5 is a function-level statement: under Asm. A1 or A2, no measurable predictor improves on $f_{\mathrm{SB}(Y)}$ in worst-case risk. Our methodology itself is subset-based: in strategic settings, the choice of subset shapes follower incentives, as in the example of § 1. Furthermore, subset-based predictors are interpretable and easy to audit, which matters in regulated settings such as lending [7, 11]. § C.4 contrasts the subset view with function-based invariance criteria such as IRM [1] and HSIC-X [54].

4.2 Counterexamples

We now show that, in general, the graphical condition (6) in Asm. A2 cannot be dropped.

Example 7. *Consider an SCM with independent noises $\varepsilon_1 \sim \mathrm{Ber}(p)$ for $p \in (0, 1/2)$ and ε_2:*

$$\begin{aligned} &Y \sim \mathrm{Ber}(1/2), \\ &X_2 := h(Y, E, \varepsilon_2) \in \{0, 1\}, \end{aligned} \qquad X_1 := \begin{cases} Y & \text{if } X_2 = 0, \\ Y \oplus \varepsilon_1 & \text{if } X_2 = 1, \end{cases}$$



where $\oplus$ is addition mod 2, and the function $h(\cdot, e, \cdot)$ may vary with e, so that different environments correspond to different mechanisms for X_2. Here, $\mathrm{SB}(Y) = \varnothing$, and the associated DAG violates Asm. A2. There exists a classifier $f\colon \mathcal{X} \to [0, 1]$ achieving a Brier score risk that is uniformly lower than the one of $f_{\mathrm{SB}(Y)} \equiv 1/2$: for all $e \in \mathcal{E}$, we have $R_e(f) \leq p(1-p) < \frac{1}{4} = R_e(f_{\mathrm{SB}(Y)})$ (§ E.6).

[6] The answer does not follow directly from existing work: e.g., Rojas-Carulla et al. [51, Thm. 4] (or, more precisely, Prop. 15, the extension to our setting) prove that for all invariant S, f_S is a worst-case optimal predictor over distributions sharing the conditional $Y \mid X_S$; this result is insufficient, as, for us, the considered distributions do not change with S.

In the regression setting with squared loss, for every DAG in which (6) fails, there exists an SCM with that DAG for which $f_{\mathrm{SB}(Y)}$ is not worst-case optimal (proved formally as Prop. 14).

5 Learning subset-based predictors

When the causal graph is known, the stable blanket $\mathrm{SB}(Y)$ is identified and $f_{\mathrm{SB}(Y)}$ can be estimated by pooling data across training environments. In practice, the causal graph is often unknown. In the population case, under faithfulness, every subset invariant on $\mathcal{E}$ is contained in $\mathrm{DE}(Y)^C_{\mathrm{forb}}$ (§ C.3), so by Lem. 2, $\mathrm{SB}(Y)$ is a most predictive invariant subset. With finite samples from $\mathcal{E}_{\mathrm{tr}}$, we therefore propose to select subsets $S \subseteq \{1, \dots, d\}$ that are invariant on $\mathcal{E}_{\mathrm{tr}}$ and predictive of Y, and estimate f_S by pooling data across $\mathcal{E}_{\mathrm{tr}}$. Since invariance on $\mathcal{E}_{\mathrm{tr}}$ need not extend to $\mathcal{E}$, this does not guarantee recovering $\mathrm{SB}(Y)$ (even for infinite data), so the methods below can be regarded as finding proxies for $f_{\mathrm{SB}(Y)}$. We build on existing tests and methods. Throughout, let $\hat{f}_S : \mathcal{X} \to \mathcal{A}$ estimate $f_S(x) = \mathbb{E}[Y \mid X_S = x_S]$, trained on data pooled across $\mathcal{E}_{\mathrm{tr}}$; any base model (such as random forests or neural networks) can be used. There are various ways of testing invariance, i.e., $H_{0,S} \colon Y \perp\!\!\!\perp E \mid X_S$, which we detail in § D.1. We take the resulting p-value as an invariance score $s_{\mathrm{inv}}(S)$, yielding $\hat{\mathcal{I}}_{\mathcal{E}_{\mathrm{tr}}} \coloneqq \{S \mid s_{\mathrm{inv}}(S) \geq \alpha_{\mathrm{inv}}\}$ for a tuning parameter $\alpha_{\mathrm{inv}} \in (0, 1)$. To reduce the exponential cost of computing s_{inv} for all 2^d subsets, one can perform a screening step (e.g., ℓ_1-penalization; see § D.2). We denote by $s_{\mathrm{pred}}(S)$ a prediction score of $\hat{f}_S$, such as the out-of-sample negative MSE for regression or the out-of-sample negative binary cross-entropy for classification on the training environments.

Invariant most-predictive subsets. Given $\hat{\mathcal{I}}_{\mathcal{E}_{\mathrm{tr}}}$, one option is to predict using the invariant subset with the highest prediction score, that is, $\hat{S}_{\max} \coloneqq \mathrm{argmax}_{S \in \hat{\mathcal{I}}_{\mathcal{E}_{\mathrm{tr}}}} s_{\mathrm{pred}}(S)$ [51]. We call this approach IMP (invariant most-predictive) [19].

Stabilized regression and classification. Since IMP is based on a single subset, errors in the invariance test may have a large impact on IMP's performance. Stabilized regression [48] addresses this by using an ensemble $\hat{f}^{\mathrm{SR}}(x) \coloneqq \sum_{S \subseteq \{1,\dots,d\}} \hat{\omega}_S \hat{f}_S(x)$, where the non-negative weights sum to one. Non-zero weights are assigned only to subsets that are both invariant and sufficiently predictive, with the predictiveness threshold calibrated by a bootstrap. To the best of our knowledge, we are the first to port this idea to classification by replacing the base regressors, the prediction score (cross-entropy instead of MSE), and the invariance tests with their classification counterparts, yielding stabilized classification (SC). Details on SC are given in § D.2.

6 Experiments

6.1 Synthetic prediction-intervention games

We validate our theoretical results on synthetic data. For this we consider two data-generating processes: a linear-Gaussian SCM and a nonlinear SCM. Both have the same induced training DAG $\mathcal{G}_0$ in Fig. 1. We compare three predictors (either in a population or finite-sample version): based on $\mathrm{PA}(Y)$, $\mathrm{SB}(Y)$, or all variables. The follower is represented by a neural intervention mechanism with intervention bound b and is trained either to maximize the leader's MSE or to minimize the prediction value (for details see § F.2). All results are averaged over 10 independent runs, with shaded regions indicating 95% confidence intervals for the mean. Further details are provided in § F.1.

We first consider the linear-Gaussian SCM (Fig. 2, left). In this setting the relevant conditional expectations are available in closed form, so the predictors use population regression quantities rather than estimates from finite training data. This isolates the effect of the follower's intervention from statistical estimation errors. Both $\mathrm{PA}(Y)$ and $\mathrm{SB}(Y)$ yield constant deployment MSE, as predicted for $\mathrm{PA}(Y)$ in § 3.1; for any invariant subset S, in particular $\mathrm{SB}(Y)$, the deployment MSE of f_S in e equals $\mathbb{E}_e[\mathrm{Var}(Y \mid X_S)]$, and $\mathrm{Var}(Y \mid X_S)$ is constant in linear-Gaussian SCMs. $\mathrm{SB}(Y)$ improves over $\mathrm{PA}(Y)$ because the non-intervened child X_3 carries additional stable information about Y. By contrast, the all-variable predictor is unstable: it performs well for weak interventions, but its MSE increases sharply as the intervention bound grows, since it relies on the manipulable variable X_4.

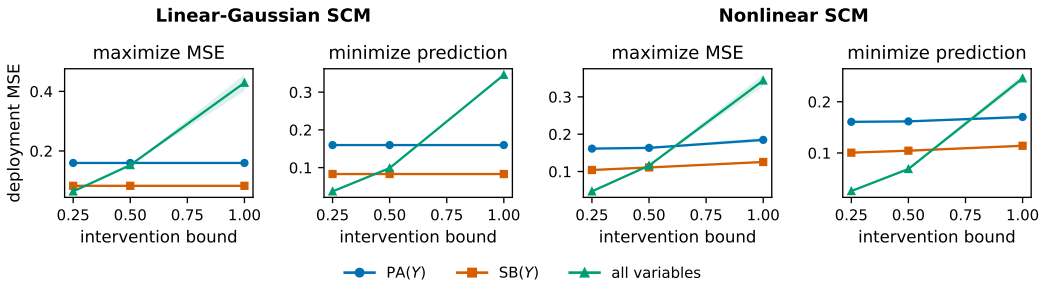


Figure 2: Synthetic prediction-intervention games. Left two panels: linear-Gaussian SCM with population regression predictors. Right two panels: nonlinear SCM with predictors estimated from finite data. The stable blanket is uniformly better than the parent set and both are substantially more stable than the all-variable predictor. In the nonlinear SCM, the small increase for $\mathrm{PA}(Y)$ is due to regression error alone; that for $\mathrm{SB}(Y)$ is mostly due to variation in population risk (shown in § F.3).

The nonlinear SCM shows the same qualitative ordering (Fig. 2, right). The stable blanket and the parent set remain more stable than the all-variable predictor, and the stable blanket again improves over the parents. Unlike in the linear-Gaussian experiment, however, the stable-blanket and parent predictors are not perfectly flat. This is expected: the predictors are now learned from finite training data, and for the stable blanket the population MSE $\mathbb{E}_e[\mathrm{Var}(Y \mid X_{\mathrm{SB}(Y)})]$ is no longer constant, since $\mathrm{Var}(Y \mid X_{\mathrm{SB}(Y)})$ is not constant in $X_{\mathrm{SB}(Y)}$ in the nonlinear SCM. The follower can exploit finite-sample regression errors by shifting the test distribution toward regions in which the fitted predictor differs from the population regression function. § F.3 investigates this effect directly: as the leader's training sample size grows, the MSE curves for $\mathrm{PA}(Y)$ and $\mathrm{SB}(Y)$ become more stable, while the all-variable predictor remains unstable.

Finally, we also allow the follower's intervention on X_4 to depend on X_1 and X_3 in addition to its usual inputs (captured by the dashed edges in the deployment DAG $\mathcal{G}$ in Fig. 1). The qualitative results do not change; see § F. § F.5 presents a similar experiment with individual followers as in § C.1.

6.2 Prediction-intervention games on Causal Chambers

We complement the synthetic experiments of § 6.1 with a prediction-intervention game on real data from the Causal Chambers[7] light tunnel (Fig. 3 left) of Gamella et al. [18].

Setup. The chamber takes a light source setting $\texttt{RGB} \coloneqq (\texttt{red}, \texttt{green}, \texttt{blue}) \in \mathbb{R}^3$ and returns infrared and visible-light measurements ($\texttt{ir_j}$, $\texttt{vis_j}$) at positions $\texttt{j} \in \{1, 2, 3\}$ along the chamber. We aim to predict $Y \in \{0, 1\}$ (constructed as a function of $\texttt{ir_1}$) and observe the system under different environments. The leader's loss is the MSE; the follower's objective is to minimize the population mean prediction $V_e^{\mathrm{foll}}(\hat{f}) = -\mathbb{E}_e[\hat{f}(X)]$, as considered in Thm. 3. Given $\hat{f}$, the follower can query the Causal Chambers API to receive 500 observations of X under different interventions on $\texttt{ir_3}$ and $\texttt{vis_3}$. We consider a fixed collection of 49 candidate environments the follower can choose from. We restrict their choice with an intervention bound $b \in [0, 1]$: only the $\max\{1, \mathrm{round}(49b)\}$ environments closest to a reference environment e_{ref} are available. See § G.

Inaccurate causal graph. Gamella et al. [18] provide partial causal-graph information for the light tunnel. They note that absent edges do not necessarily exclude the existence of causal effects or hidden confounding. Ignoring this note and adding the edges from Y and E to $(\texttt{ir_3}, \texttt{vis_3})$, one might be tempted to consider the graph in Fig. 3 (middle) as correct. However, there is empirical evidence that this is not the correct graph for our experiment: conditional independence tests on training data reject[8] some of the implied independencies between sensors. We hypothesize that this is due to a hidden confounder (such as background lighting) acting on the sensors. Adding it results

[7] The Causal Chambers are devices controlled through an API and collect measurements from a physical system for which we have partial causal-graph information.

[8] Specifically, $\texttt{ir_i} \perp\!\!\!\perp \texttt{ir_j} \mid \texttt{RGB}$ and $\texttt{vis_i} \perp\!\!\!\perp \texttt{vis_j} \mid \texttt{RGB}$ for $\texttt{i} \neq \texttt{j} \in \{1, 2, 3\}$ are rejected in the majority of training environments at level 0.01; see § G.3.

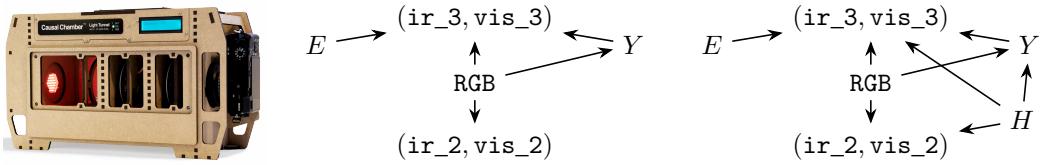


Figure 3: Left: light tunnel (adapted from Gamella et al. [18], licensed under CC BY 4.0). Middle: inaccurate graph, inferred from the partial causal-graph information by Gamella et al. [18]; the follower is allowed to intervene on $\texttt{ir_3}, \texttt{vis_3}$; the stable blanket of Y is $\{\texttt{RGB}\}$. Right: potentially correct graph with a hypothesized hidden confounder H pointing to all $\texttt{ir_j}$ and $\texttt{vis_j}$; the stable blanket of Y is $\{\texttt{RGB}, \texttt{ir_2}, \texttt{vis_2}\}$. Both graphs omit $\texttt{ir_1}$, which is only used to define Y.

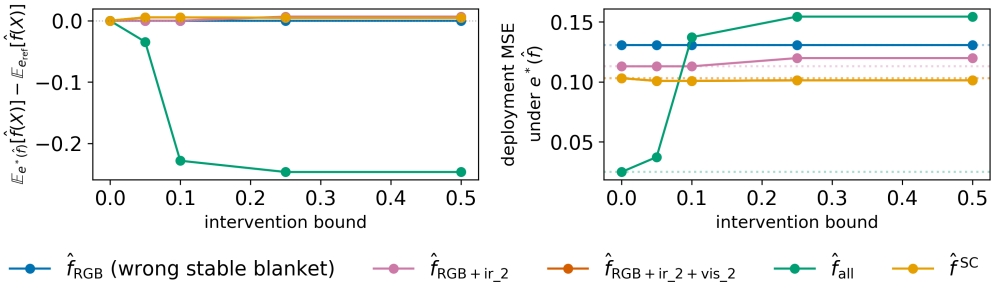


Figure 4: Prediction-intervention game on Causal Chambers data (§ 6.2). Left: follower-induced shift in the population mean prediction relative to a reference environment e_{ref}. Right: MSE of the leader's deployed predictor in e^* (chosen by the follower). The follower has a strong influence on their objective only for $\hat{f}_{\text{all}}$. The proposed $\hat{f}^{\text{SC}}$ (SC for stabilized classification) achieves the lowest MSE for $b \geq 0.1$, tied with $\hat{f}_{\texttt{RGB},\texttt{ir_2},\texttt{vis_2}}$ (curves overlap); $\hat{f}^{\text{SC}}$ does not require graph knowledge.

in Fig. 3 (right). Under this extended graph with hidden variables the stable blanket of Y can be defined analogously [67] and equals $\{\texttt{RGB}, \texttt{ir_2}, \texttt{vis_2}\}$; § C.7 provides more details.

Results. Fig. 4 shows the degree to which the follower can improve their objective (left), and the leader's MSE under the environment chosen by the follower (right), both computed on 500 new evaluation observations. Only under $\hat{f}_{\text{all}}$ (using all variables) can the follower improve their objective (left), and for large intervention bounds, this predictor performs worst for the leader. $\hat{f}_{\texttt{RGB}}$ is the stable-blanket predictor under the inaccurate graph and has a higher MSE than the larger subsets that include $\texttt{ir_2}$ and $\texttt{vis_2}$ on the reference environment. This is consistent with the hidden-confounding hypothesis that the graph is incorrect. The SC predictor $\hat{f}^{\text{SC}}$, which is purely based on data from three training environments and not the graph, selects $\{\texttt{RGB}, \texttt{ir_2}, \texttt{vis_2}\}$ (the stable blanket under the hidden-confounder graph) as the unique subset passing both the invariance and predictiveness filters, and therefore matches the performance of $\hat{f}_{\texttt{RGB},\texttt{ir_2},\texttt{vis_2}}$.

7 Discussion and future work

We have introduced prediction-intervention games, connected them to invariant sets and worst-case distribution generalization, and proved optimality results for strategies based on the stable blanket. We proposed practical methods that the leader may use for settings with known and unknown causal graph, and have successfully tested them on data. Many of our results concern the worst-case problem (5), such as Thm. 3 (i), Thm. 5, Prop. 6, and the examples in § 4.2. Exceptions are Lem. 2, which holds in every environment, and Thm. 3 (ii), in which the follower is not adversarial. Our results do not characterize the optimal leader strategy in general. One limitation of the worst-case optimality results in § 4.1 is that they rely on strong-intervention assumptions. Analyzing optimal strategies under intervention-constrained followers is an interesting direction for future research. We hypothesize that solving prediction-intervention games is in general a hard problem [e.g., 10]. Finally, in practice, the leader uses a finite-sample estimate $\hat{f}_S$ rather than the population predictor f_S, and the follower can shift covariates to high-error regions (§ F.3); a finite-sample analysis is left for future work.

Acknowledgments and Disclosure of Funding

We thank Niklas Pfister and Sorawit Saengkyongam for helpful discussions, and Juan Gamella for the help with the Causal Chambers data generation. During this work, LK was supported by the ETH AI Center through an ETH AI Center doctoral fellowship. We used Claude Opus 4.6, Fable 5, and Fable 5.1 to polish the language and proofread, to check and tighten proofs, to assist the literature search, to draft parts of the appendix (such as § C.7) from our notes, and to implement simulations and figures. The authors take full responsibility for the content.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv e-prints (1907.02893)*, 2019.

[2] Jérôme Bolte, Quốc-Tùng Lê, Edouard Pauwels, and Samuel Vaiter. Geometric and computational hardness of bilevel programming. *Mathematical Programming*, 215(1–2):539–574, 2026.

[3] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

[4] Kenneth P. Brevoort, Robert B. Avery, and Glenn B. Canner. Credit where none is due? Authorized-user account status and piggybacking credit. *Journal of Consumer Affairs*, 47 (3):518–547, 2013.

[5] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 547–555, 2011.

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

[7] Jiahao Chen. Fair lending needs explainable models for responsible recommendation. *arXiv e-prints (1809.04684)*, 2018.

[8] Yatong Chen, Jialu Wang, and Yang Liu. Learning to incentivize improvements from strategic agents. *Transactions on Machine Learning Research*, 2023.

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

[10] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 82–90, 2006.

[11] Consumer Financial Protection Bureau. Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms. `https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/`, 2022. Accessed: April 20, 2026.

[12] Consumer Financial Protection Bureau. Regulation B (12 CFR Part 1002), § 1002.6 Rules concerning evaluation of applications, paragraph (b)(6)(i). `https://www.consumerfinance.gov/rules-policy/regulations/1002/6/`, 2026. Accessed: October 13, 2026.

[13] Matt Cutts. Google does not use the keywords meta tag in web ranking. `https://developers.google.com/search/blog/2009/09/google-does-not-use-keywords-meta-tag`, 2009. Accessed: June 12, 2026.

[14] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[15] Emiliano Díaz Salas-Porras, Kenza Tazi, Ashwin Braude, Daniel Okoh, Kara D. Lamb, Duncan Watson-Parris, Paula Harder, and Nis Meinert. Identifying the causes of pyrocumulonimbus (PyroCb). *arXiv e-prints (2211.08883)*, 2022.

[16] Maayan Ehrenberg, Roy Ganz, and Nir Rosenfeld. Adversaries with incentives: A strategic alternative to adversarial robustness. In *International Conference on Learning Representations*, 2025.

[17] Francesco Freni, Anya Fries, Linus Kühne, Markus Reichstein, and Jonas Peters. Maximum risk minimization with random forests. *arXiv e-prints (2512.10445)*, 2025.

[18] Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025. See updated causal graph under `https://cchamber-box.s3.eu-central-2.amazonaws.com/gt_graph_lt_mk2_standard.pdf`. Accessed: May 6, 2026.

[19] Nicola Gnecco, Jonas Peters, Sebastian Engelke, and Niklas Pfister. Boosted control functions: Distribution generalization and invariance in confounded models. *Journal of Machine Learning Research*, 27(46):1–57, 2026.

[20] Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[21] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[22] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. Reasoning about causality in games. *Artificial Intelligence*, 320:103919, 2023.

[23] Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *Statistical Science*, 40(3):417–436, 2025.

[24] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016.

[25] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

[26] Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13233–13253, 2023.

[27] Xinyan Hu, Meena Jagadeesan, Michael I. Jordan, and Jacob Steinhardt. Incentivizing high-quality content in online recommender systems. In *Proceedings of the 7th Symposium on Foundations of Responsible Computing*, pages 15:1–15:18, 2026.

[28] Olav Kallenberg. *Foundations of Modern Probability*. Springer International Publishing, Cham, 3rd edition, 2021.

[29] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 4069–4077, 2021.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[31] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

[32] Lucas Kook. *tramicp: Model-Based Causal Feature Selection for General Response Types*, 2025. R package version 0.1-0.

[33] Lucas Kook, Sorawit Saengkyongam, Anton Rask Lundborg, Torsten Hothorn, and Jonas Peters. Model-based causal feature selection for general response types. *Journal of the American Statistical Association*, 120(550):1090–1101, 2025.

[34] Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

[35] Bogdan Kulynych. Causal prediction can induce performative stability. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

[36] Ho Cheung Brian Lee, Xinxin Li, and Siyuan Liu. The role of monitoring effect in risk classification: Evidence from telematics adoption. *Management Science*, 71(12):10122–10143, 2025.

[37] Kenneth Lee, Md Musfiqur Rahman, and Murat Kocaoglu. Finding invariant predictors efficiently via causal structure. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 1196–1206, 2023.

[38] 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. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

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

[40] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. In *Advances in Neural Information Processing Systems*, pages 31171–31185, 2022.

[41] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6917–6926, 2020.

[42] Manuj Mishra, James Fox, and Michael Wooldridge. Characterising interventions in causal games. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, pages 2560–2572, 2024.

[43] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1988.

[44] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[46] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7599–7609, 2020.

[47] 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.

[48] Niklas Pfister, Evan G. Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.

[49] Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[50] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

[51] 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.

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

[53] 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.

[54] Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, and Jonas Peters. Exploiting independent instruments: Identification and distribution generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18935–18958, 2022.

[55] Sorawit Saengkyongam, Nikolaj Thams, Jonas Peters, and Niklas Pfister. Invariant policy learning: A causal perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8606–8620, 2023.

[56] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[57] Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022.

[58] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

[59] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8676–8686, 2020.

[60] Xinwei Shen, Peter Bühlmann, and Armeen Taeb. Causality-oriented robustness: Exploiting general noise interventions. *Journal of the American Statistical Association*, 121(553):704–715, 2026.

[61] Ilya Shpitser and Judea Pearl. Dormant independence. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 1081–1087, 2008.

[62] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[63] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127, 2019.

[64] Adarsh Subbaswamy, Bryant Chen, and Suchi Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1):64–89, 2022.

[65] Thomas Sadanand Verma. Graphical aspects of causal models. Technical Report R-191, University of California at Los Angeles, Cognitive Systems Laboratory, 1993.

[66] Heinrich von Stackelberg. *Marktform und Gleichgewicht*. Julius Springer, Wien und Berlin, 1934.

[67] Hanqing Xiang. Stable blanket with hidden variables and cycles. *arXiv e-prints (2605.01856)*, 2026.

[68] Tom Yan, Shantanu Gupta, and Zachary Lipton. Discovering optimal scoring mechanisms in causal strategic prediction. *arXiv e-prints (2302.06804)*, 2023.

[69] Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.

Appendix contents

A Real-world examples **15**

B Extended related work **15**

C Additional theoretical results **18**

D Stabilized regression and classification **26**

E Proofs **27**

F Additional details for the synthetic experiments **34**

G Additional details for the Causal Chambers experiments **36**

A Real-world examples

Prediction-intervention games arise in US consumer credit scoring, where the score developer (such as FICO or VantageScore) takes the role of the leader and credit applicants are followers. Credit scores predict future repayment from covariates derived from an applicant's credit record. One covariate summarizes authorized-user accounts: revolving accounts (typically credit cards) that the applicant may use without being liable for the charges. Regulation B, which implements the 1974 US Equal Credit Opportunity Act, requires lenders to consider accounts an applicant shares with their spouse [12]. Since credit records do not reveal which authorized users are spouses, scoring models have traditionally treated all authorized-user accounts like the applicant's own accounts [4]. This enables a strategic intervention known as 'piggybacking': so-called 'credit repair' companies add credit applicants as authorized users (in name only, without receiving a card) on a stranger's old, well-maintained account for a fee.[9] The account's favorable history then enters the applicant's credit record, while their repayment behavior remains unchanged. In our terms, the applicant intervenes at a cost on a covariate that is predictive of Y but leaves Y unchanged, in order to increase $f(X)$, as in Thm. 3 (ii). Brevoort et al. [4] estimate that adding one such account moves 28% of 'subprime' and 35% of 'near-prime' borrowers into a higher credit-quality segment. FICO announced in 2007 that its next model would exclude authorized-user accounts; after warnings that exclusion might violate Regulation B, FICO instead reduced the weight of these accounts [4]. This is an example of a prediction-intervention game in which the leader initially chose to exclude a covariate from its model to protect itself against strategic interventions by followers, even though the covariate is predictive in some subgroups: excluding it leaves overall predictive accuracy statistically unchanged, but significantly reduces accuracy for applicants with two or fewer accounts of their own [4].

A second example is web page ranking, where a search engine scores the relevance of a page from covariates including its content and metadata. The `keywords` meta tag lists terms the page owner places in the HTML. It is invisible to visitors and can be changed independently of the page's content. After widespread 'keyword stuffing' (filling the tag with irrelevant terms to attract traffic), Google excluded it from its web ranking algorithm [13].

B Extended related work

We expand on the related work discussion from § 1.

Strategic classification and performative prediction. Strategic classification [5, 24] models followers who respond to a deployed classifier f by modifying their features with some cost c to obtain

[9] According to Brevoort et al. [4, Footnote 2], applicants pay between $1,000 and $2,000 for this service, and the holder of the rented account can earn about $200 per month.

a more favorable prediction, typically through the pointwise best response $x^f \in \operatorname{argmax}_{x'} f(x') - c(x, x')$. The leader's goal is to learn an f that is robust under the induced distribution shift. The leader's and follower's objectives are typically asymmetric in this Stackelberg formulation; Ehrenberg et al. [16] make this explicit and study a game in which the leader minimizes a worst-case loss over a set of different utility-maximizing opponents, without a causal structure over X and Y. A related question is how the choice of predictor creates different incentives for the followers and distinguishes gaming the system from genuine improvement: Kleinberg and Raghavan [31] model followers as investing effort into different actions that affect the observed features, and ask which decision rules incentivize followers to invest in actions a leader views as valuable rather than in gaming; Chen et al. [8] distinguish improvable, manipulable, and immutable features, with gaming defined as changes to manipulable features that by assumption do not affect Y. They do not specify a full SCM over (X, Y, E) that determines how the follower's actions propagate through the whole system. Performative prediction [23, 46] considers the broader phenomenon that deployed models shift the data distribution; Mendler-Dünner et al. [40] address identifiability of the causal effect of predictions. These works do not generally specify the follower's moves through an SCM. Perdomo et al. [46] formalize the dependence of the data on the deployed model by an abstract function $D(\cdot)$ from models to distributions and evaluate f by its performative risk, the risk of f under $D(f)$. In our game, $D(f)$ is the distribution $\mathbb{P}_e$ induced by a best response $e \in \mathcal{E}^*(f)$, and the leader's objective in (1) is the performative risk, with a supremum over best responses. $D(\cdot)$ is generally unknown to the decision maker, and the results of Perdomo et al. [46] concern retraining procedures such as repeated risk minimization, which retrains on the distribution induced by the previously deployed model and, under regularity conditions, converges to a fixed point of retraining. Such a fixed point need not minimize the performative risk, which accounts for the change of the distribution with the predictor. These retraining procedures deploy a sequence of predictors and observe the induced distributions, which the leader in our game cannot do, since f is chosen before the follower responds. We instead specify how $D(\cdot)$ arises, through the follower's interventions on the children of E in an SCM, and use the invariances implied by this structure. Thus, the methods proposed in this paper can compare predictors before deployment (possibly using a single dataset), whereas retraining methods in performative prediction require a sequence of datasets. Other game-theoretic interactions between predictors and strategic agents have been studied in market settings [27].

Causal Stackelberg formulations. Several works model Stackelberg games with SCMs. Shavit et al. [59] formulate strategic linear regression in a linear SCM, and learn linear predictors targeting one of three specific objectives: minimizing the leader's risk under strategic follower responses, incentivizing followers to genuinely improve their outcomes, or estimating the parameters of the outcome model. Their algorithms deploy multiple predictors and observe the resulting follower responses to learn the final predictor. In contrast, our predictor is defined without observing follower interventions. We further allow a general SCM, cover regression and classification under any loss strictly consistent for the mean functional, and place no specific functional form on the follower's response: their setup fixes it through an effort conversion matrix and a quadratic cost, whereas ours constrains the follower only through the graph $\mathcal{G}$ and the action sets A_j. Miller et al. [41] develop a causal framework that distinguishes gaming from improvement and prove that designing classifiers which incentivize improvement is at least as hard as orienting the edges of the underlying causal graph. Their contribution is conceptual and concerns the difficulty of incentive design, while we focus on the leader's optimal choice of f given an SCM. Yan et al. [68] assume an unknown causal graph and follower manipulations constrained by a fixed budget. Their algorithm iteratively deploys predictors and observes strategic responses to recover the graph, and their leader objective trades off prediction accuracy with outcome improvement. Our predictor does not require iterated deployment, and in the unknown-graph case (§ 5) we use multiple training environments instead of observed strategic responses. Horowitz and Rosenfeld [26] partition covariates into causal and non-causal predictive components. Their follower is an individual who modifies covariates to change the prediction in its favor, and only changes to causal variables propagate to Y. We, instead, allow for a more flexible model of the follower: we may specify which variables the follower may intervene on, and which other variables their interventions may depend on. The latter can induce additional edges in the graph, so our setup accommodates settings in which the deployment graph extends the training graph. Unlike the work by Horowitz and Rosenfeld [26], we do not estimate the causal function of the response but work with invariant sets; furthermore, we admit regression in addition to classification, allow for more general losses and for more general follower utilities; finally, Horowitz and Rosenfeld [26] do not provide optimality results in their setting. Kulynych [35] extends an SCM

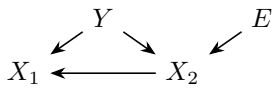
over (X, Y) by including the deployed predictor as an additional node M, with edges $M \to X$ but no edge to Y. Instead of modeling a follower optimizing its own utility, the deployment-induced shift in X is encoded directly in the structural assignments of the SCM. In this setting, they study conditions under which the Markov-blanket predictor remains optimal under the shift it induces. In our formulation, the post-deployment distribution instead arises from a follower choosing interventions to maximize an objective distinct from the leader's risk, which the leader anticipates when choosing f. Mishra et al. [42] study interventions in causal games [22], graphical models of multi-agent decision-making, and give a set of primitive interventions from which arbitrary interventions can be composed; one application represents a leader's commitment to a policy, fixed before the follower acts, as an intervention on the leader's decision rule. As in our setting, a leader commits to a policy to which a follower responds, and interventions may add edges to the graph. Our game could be represented as a causal game with the intervened variables as the follower's decision variables and the prediction as the leader's. They study mechanism design, whereas we study a specific prediction problem and analyze the predictive risk and worst-case optimality of predictors based on invariant subsets.

Causal distribution generalization and invariant subsets. Invariance-based causal prediction [9, 38, 47] uses subsets S for which $Y \mid X_S$ is stable across environments. Pfister et al. [48] introduce the stable blanket. For regression, invariant subsets have been motivated by worst-case performance: Rojas-Carulla et al. [51] show that for each invariant subset S, f_S minimizes the worst-case squared loss over the family $\mathcal{Q}_S$ of distributions sharing the conditional $Y \mid X_S$. With Prop. 15 in § C.6, we generalize this result to binary classification and regression with any loss strictly consistent for the mean functional. This result is not directly applicable to the optimization problem (5), as the family $\mathcal{Q}_S$ depends on S and may strictly contain $\{\mathbb{P}_e\}_{e\in\mathcal{E}}$, in which case f_S guards against shifts that cannot occur under $\mathcal{E}$. The result does not inform us about which invariant subset to use, either. We instead analyze worst-case optimality directly over $\{\mathbb{P}_e\}_{e\in\mathcal{E}}$ and give conditions under which the stable blanket is optimal among all measurable predictors. Our worst-case analysis is related to a line of work taking a distributional-robustness view in which the uncertainty set comes from SCM interventions [6, 39, 60, 64]; this connects to the broader DRO literature [17, 34, 56]. Through the bound (4), prediction-intervention games provide a concrete example of such a worst-case problem with the distribution family $\{\mathbb{P}_e\}_{e\in\mathcal{E}}$ determined by the SCM. These works do not directly apply to our setting, however: their uncertainty sets (interventions of a fixed form on a fixed DAG or balls around a reference distribution in the case of DRO) generally do not coincide with $\{\mathbb{P}_e\}_{e\in\mathcal{E}}$. Anchor regression [53] attains worst-case optimality over a class of bounded shift interventions. Other function-view criteria include IRM [1] and instrument-based independent residuals [54]. We do not use a function view but a subset view, as discussed in § C.4.

Comparison to Subbaswamy et al. [63] and Lee et al. [37]. Subbaswamy et al. [63] assume knowledge of a causal graph, possibly with hidden confounders, in which the mechanisms of a known set M of mutable variables may change arbitrarily; this defines an uncertainty set Γ of distributions. Their proposed full conditional surgery predictor $f_s(x) \coloneqq \mathbb{E}[Y \mid X_{M^C} = x_{M^C}, \mathrm{do}(X_M = x_M)]$ intervenes on M and conditions on the remaining covariates; it is not always identifiable from observational data. Their Alg. 2 therefore searches over predictors $\mathbb{E}[Y \mid X_Z = x_Z, \mathrm{do}(X_M = x_M)]$ for conditioning sets $Z \subseteq M^C$, retains those that are identifiable from the observational distribution, and selects by the lowest held-out source loss. Our setting maps to this framework with $M = \mathrm{CH}(E)$. Since $\mathrm{SB}(Y)$ is a stable conditioning set in their terminology, $f_{\mathrm{SB}(Y)}$ is among the predictors their method considers [63, Lem. 1 and Cor. 1]. However, as the following example in our fully observed setting shows, their selection rule, evaluated with population source risk, can return a predictor with higher worst-case risk than $f_{\mathrm{SB}(Y)}$ in a setting that satisfies Asm. A1, where $f_{\mathrm{SB}(Y)}$ is worst-case optimal by Thm. 5.

Example 8. *Consider the DAG of Ex. 7, the squared loss, and the SCM*

$$
\begin{aligned}
Y &\sim \mathrm{Unif}\{-1, 0, 1\},\\
X_2 &\coloneqq h(Y, E, \varepsilon_2) \in \{0, 1\},\\
X_1 &\coloneqq (1 - X_2)\varepsilon_1 + X_2 \mathbf{1}\{Y \geq 0\},
\end{aligned}
$$



where $\varepsilon_1 \sim \mathrm{Ber}(1/2)$ *is independent of* (Y, ε_2). $\mathcal{E}$ *consists of a training environment* e_{tr} *with* $\mathbb{P}(X_2 = 1 \mid Y = \pm 1) = 3/4$ *and* $\mathbb{P}(X_2 = 1 \mid Y = 0) = 1/4$*, and two deployment environments*

e_1 *and* e_2 *with* $X_2 := \mathbf{1}\{Y = 0\}$ *and* $X_2 := 0$*, respectively. Here,* $M = \{2\}$ *and* $\mathrm{SB}(Y) = \varnothing$*, so* $f_{\mathrm{SB}(Y)} \equiv \mathbb{E}[Y] = 0$*.*

The two identifiable candidates are $f_{\mathrm{SB}(Y)}$*, with the empty conditioning set, and* f_s*, with conditioning set* $\{1\}$*: under* $\mathrm{do}(X_2 = 0)$*,* $X_1 = \varepsilon_1$ *is independent of* Y*, and under* $\mathrm{do}(X_2 = 1)$*,* $X_1 = \mathbf{1}\{Y \geq 0\}$*, so*

$$f_s(x) = \mathbb{E}[Y \mid X_1 = x_1, \mathrm{do}(X_2 = x_2)] = x_2\big(-1 + \tfrac{3}{2}x_1\big) \quad \textit{for } x \in \{0,1\}^2.$$

The MSEs in $(e_{\mathrm{tr}}, e_1, e_2)$ *are* $(1/4, 3/4, 2/3)$ *for* f_s *and* $(2/3, 2/3, 2/3)$ *for* $f_{\mathrm{SB}(Y)}$*. The selection rule therefore returns* f_s *because of its lower source MSE, although its worst-case MSE over* $\mathcal{E}$ *is* $3/4$*, compared to* $2/3$ *for* $f_{\mathrm{SB}(Y)}$*. The assumptions of Subbaswamy et al. [63, Thms. 3 and 4] do not hold here (*f_s *depends on* x_2*, so* $\mathbb{P}(Y \mid X_1, \mathrm{do}(X_2))$ *would be a function of their* M*).*

A further difference concerns the uncertainty set: the guarantees of Subbaswamy et al. [63, Thms. 3 and 4] are worst-case guarantees for the squared error over Γ, whereas our uncertainty set $\{\mathbb{P}_e\}_{e\in\mathcal{E}}$ may be a strict subset of Γ, and interventions that add permitted parent edges can yield distributions outside Γ. Moreover, Γ contains every change of the mechanisms of M and thus strong interventions (as in Asm. A1 or A2) by construction, whereas $\{\mathbb{P}_e\}_{e\in\mathcal{E}}$ need not; our results in § 3 assume neither Asm. A1 nor A2. Besides using a different uncertainty set, Subbaswamy et al. [63] restrict their guarantees to the squared error and to the worst case, whereas all our results apply to any loss strictly consistent for the mean, and Lem. 2 and Thm. 3 cover non-adversarial follower objectives. In our setting, under condition (6), $\mathrm{SB}(Y)$ is the Markov blanket of Y after removing the edges into M, so $f_s = f_{\mathrm{SB}(Y)}$ and the condition of their Thm. 3 holds. Lee et al. [37] build on the framework of Subbaswamy et al. [63] with a polynomial-time greedy algorithm, but do not prove worst-case optimality for the returned predictor. The methods of Subbaswamy et al. [63] and Lee et al. [37] rely on conditional independencies in interventional distributions, which Shpitser and Pearl [61] study under the name of dormant independencies.

C Additional theoretical results

C.1 Individual followers with shift interventions and target-specific observations

The argument of § 2.1 assumes for simplicity that the individual follower observes only variables X_O with $O \subseteq \mathrm{ND}(\mathrm{CH}(E))$ and that $\mathrm{PA}(j) \setminus \{E\} \subseteq O \cup \mathrm{CH}(E)$ for every intervention target X_j, $j \in \mathrm{CH}(E)$. Here, we extend the argument to individual followers like the ones in the setting of § 6.1, which differs from § 2.1 in two ways. First, the interventions are (bounded) shifts on X_1 and X_4, so the follower observes the values X_1 and X_4 would have taken before the intervention. Second, the follower sets X_4 after observing Y, which is a descendant of the earlier intervention target X_1 and hence not in $\mathrm{ND}(\mathrm{CH}(E))$. We therefore consider target-specific observation sets O_j instead of the single O of § 2.1: the follower can use the realized y to set X_4, but not to set X_1 (otherwise, this would create a cycle and a timing issue). § F.5 compares individual and population followers empirically.

Observations. We consider an individual follower who sets the intervention targets X_j, $j \in \mathrm{CH}(E)$, along a fixed topological order of $\mathcal{G}_0$. To allow for shift interventions (which we consider in § 6.1 and § F.5), we now introduce auxiliary variables: for each X_j, $j \in \mathrm{CH}(E)$, we introduce X'_j, that is, the value of X_j without the shift intervention on X_j, once the upstream targets have been set; e.g., if $X_j := h(X_{\mathrm{PA}(j)}, \varepsilon_j) + \delta_E$, then $X'_j := h(X_{\mathrm{PA}(j)}, \varepsilon_j)$ and $X_j = X'_j + \delta_E$. Before intervening on X_j, the follower observes the values x_{O_j} and x'_j for a set O_j of variables. Fig. 5 shows $\mathcal{G}_0$ from our running example in Fig. 1 together with the auxiliary variables. The leader does not observe the X'_j's. We impose two conditions on O_j: (i) its variables precede X_j in the topological order and include $\mathrm{PA}(j) \setminus \{E\}$, and (ii) it contains the earlier intervention targets X_i and their observation sets O_i. By (i), the variables in O_j are not affected by the intervention on X_j; (ii) means when the follower chooses an intervention, they have

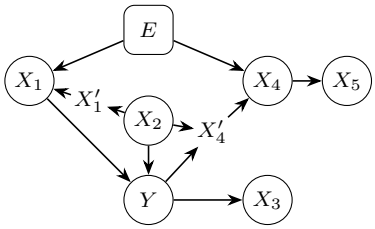


Figure 5: $\mathcal{G}_0$ of Fig. 1 with the auxiliary variables X'_1 and X'_4.

access to their earlier observations and choices. For example, in Fig. 5, the sets $O_1 = \{2\}$ and $O_4 = \{1, 2, 3, Y\}$ satisfy (i) and (ii). In this case, an individual follower may manipulate X_1 using the observations x_2 and x'_1, and X_4 using x_1, x_2, x_3, y, and x'_4. Fig. 5 shows only $\mathcal{G}_0$ together with the auxiliary variables; using observations of X_1 and X_3 for the intervention on X_4 makes them parents of X_4 in the deployment DAG $\mathcal{G}$ (dashed in Fig. 1).

Backward induction. We derive the follower's optimal choices recursively, starting from the last target. Consider first the last intervention target X_j in the topological order. Given their observation (x_{O_j}, x'_j), the follower chooses the value c_j for X_j that maximizes $\mathbb{E}_{c_j}[k(f, X, Y) \,|\, X_{O_j} = x_{O_j}, X'_j = x'_j]$, where, as in (2), the subscript indicates that X_j is set to c_j; the expectation is over the remaining noise variables. The optimal value depends on the observation, so there is a function g_j with $c_j = g_j(x_{O_j}, x'_j)$ (we assume throughout that the maximizers exist and can be chosen measurably). Consider next the second-to-last intervention target X_l. Given their observation (x_{O_l}, x'_l), the follower chooses the value c_l for X_l that maximizes $\mathbb{E}_{c_l}[k(f, X, Y) \,|\, X_{O_l} = x_{O_l}, X'_l = x'_l]$, where now the structural assignment of X_j is replaced by $g_j(X_{O_j}, X'_j)$; since X_{O_j} depends on c_l, this takes into account that the choice for X_j adapts to c_l. Again, the optimal value defines a function $g_l(x_{O_l}, x'_l)$. Continuing backwards along the topological order defines g_j for all $j \in \mathrm{CH}(E)$.[10] The functions $(g_j)_{j \in \mathrm{CH}(E)}$ therefore describe the optimal behavior of the individual follower. They can be computed before any value is realized, by maximizing, for every possible observation, the expectation over the noise variables not yet realized at that stage; the topological order only determines when their inputs become available. As in § 2.1, they also maximize the population objective $\mathbb{E}_{(g_j)_{j \in \mathrm{CH}(E)}}[k(f, X, Y)]$, where the subscript indicates that the structural assignment of every X_j is replaced by $g_j(X_{O_j}, X'_j)$, because at every stage they maximize the conditional expectation for every value of the observation.

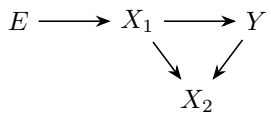
Population follower. A population of such individual followers hence acts like a population follower whose structural assignment of X_j is $g_j(X_{O_j}, X'_j)$; since X'_j is a function of $X_{\mathrm{PA}(j)}$ and ε_j, this is a function of $X_{\mathrm{PA}(j)}$, of X_{A_j} with $A_j \coloneqq O_j \setminus \mathrm{PA}(j)$, and of ε_j. As in § 2, A_j must not contain Y; that is, the follower may observe Y before setting X_j only if $Y \in \mathrm{PA}(j)$. The new edges $X_k \to X_j$, $k \in A_j$, follow the topological order, so the augmented graph is acyclic, and ε_j, which is independent of all other noise variables, serves as the exogenous noise of the new assignment. In conclusion, as in § 2.1, a population of such individual followers has, from the perspective of the leader and the game's value, the same effect as a single population follower in an augmented graph.

C.2 A follower objective under which the parents beat the stable blanket

Thm. 3 covers two classes of follower objectives. For other objectives, the ordering can reverse.

Example 9 (Parents beat the stable blanket)**.** *Let ℓ be the Brier score and consider the following SCM with $\mathcal{E} = \{0, 1\}$ and $\varepsilon_Y \sim \mathrm{Unif}([0, 1])$:*

$$\begin{aligned} X_1 &\coloneqq E \in \{0, 1\}, \\ Y &\coloneqq \mathbf{1}\{\varepsilon_Y \leq 1/2 - X_1/6\}, \\ X_2 &\coloneqq X_1 Y. \end{aligned}$$



Then $\mathrm{PA}(Y) = \{1\}$ and $\mathrm{SB}(Y) = \{1, 2\}$, with $f_{\mathrm{PA}(Y)}(x) = 1/2 - x_1/6$ and $f_{\mathrm{SB}(Y)}(x) = (1 - x_1)/2 + x_1 x_2$. The risks in $e = 0$ and $e = 1$ are $(1/4, 2/9)$ for $f_{\mathrm{PA}(Y)}$ and $(1/4, 0)$ for $f_{\mathrm{SB}(Y)}$, respectively. Let the follower maximize $V_e^{\mathrm{foll}}(f) = -\mathbb{E}_e[f(X)^2]$. Under $f_{\mathrm{PA}(Y)}$, the follower chooses $e = 1$ ($-1/9 > -1/4$); under $f_{\mathrm{SB}(Y)}$, $e = 0$ ($-1/4 > -1/3$). The leader's game objective (1) is therefore $2/9$ for $f_{\mathrm{PA}(Y)}$ and $1/4$ for $f_{\mathrm{SB}(Y)}$. This objective lies outside Cases (i) and (ii) of Thm. 3.

[10] When the follower chooses some c_j, they have already set the intervention targets X_i preceding X_j in the topological order, using functions g_i that the backward induction determines only afterwards (if c_j depended on these functions, the backward induction would be circular). Which functions these are does not matter for c_j: the observed x_i equals $g_i(x_{O_i}, x'_i)$ with x_{O_i} observed by (ii), so knowing g_i gives information only about x'_i, that is, about ε_i, which enters only X_i, whose value is already known.

Ex. 9 also shows that the optimal leader strategy can depend on the follower utility V^{foll}. Since $\text{DE}(Y)_{\text{forb}} = \varnothing$, Lem. 2 yields $R_e(f) \geq R_e(f_{\text{SB}(Y)})$ for all measurable f and all $e \in \mathcal{E}$. Therefore, for adversarial followers with $V_e^{\text{foll}}(f) = R_e(f)$, we know that $f_{\text{SB}(Y)}$ is an optimal leader strategy (see also the proof of Thm. 3 (i)). For $V_e^{\text{foll}}(f) = -\mathbb{E}_e[f(X)^2]$, however, $f_{\text{SB}(Y)}$ is not optimal, since $f_{\text{PA}(Y)}$ is strictly better. Hence, whether $f_{\text{SB}(Y)}$ is optimal depends on V^{foll}.

C.3 Subset-level formulation

Remark 10. *The following subset-level result follows from Lem. 2: for all $S \subseteq \text{DE}(Y)^C_{\text{forb}}$,*

$$\sup_{e_1, e_2 \in \mathcal{E}} \mathbb{E}_{e_1}\Big[\ell\big(Y, \mathbb{E}_{e_2}[Y \mid X_S]\big)\Big] \geq \sup_{e \in \mathcal{E}} R_e(f_{\text{SB}(Y)}). \tag{8}$$

By Thm. 5, Eq. (8) *holds for all $S \subseteq \{1, \dots, d\}$ under Asm. A1 or A2. Consequently, under faithfulness, Pfister et al. [48, Thm. 3.5] follows as a corollary, since every subset S d-separating Y and E satisfies $S \subseteq \text{DE}(Y)^C_{\text{forb}}$ by part (a) of their proof.*

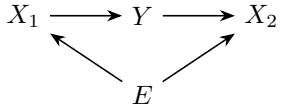
C.4 Worst-case optimal functions

In this work, we study predictors $f_S(x) = \mathbb{E}[Y \mid X_S = x_S]$ defined by an invariant subset S (the 'subset view'). Thm. 5 shows that, under Asm. A1 or A2, $f_{\text{SB}(Y)}$ is worst-case optimal over all measurable functions $f \colon \mathcal{X} \to \mathcal{A}$ (the 'function view'). We outline three reasons why the subset view is the right choice in our setting.

(1) To solve Problem (5), Invariant Risk Minimization (IRM) [1] takes a function view and jointly learns a representation $\Phi \colon \mathcal{X} \to \mathcal{H}$ and a predictor $w \colon \mathcal{H} \to \mathcal{A}$. The representation Φ is required to be invariant in the sense that the optimal predictor w on top of Φ is the same across environments. While failure cases of IRM are known [29, 52], we use the 0–1 loss in the following example to make the point that IRM's invariance criterion can be weaker than the conditional invariance $Y \perp\!\!\!\perp E \mid X_S$ of Def. 1.

Example 11 (Weak IRM invariance under 0–1 loss)**.** *Consider the following SCM*

$$\begin{aligned} X_1 &\coloneqq \mathbf{1}\{\varepsilon_1 \leq E/10\}, \\ Y &\coloneqq \mathbf{1}\{\varepsilon_Y \leq 1/4 + X_1/2\}, \\ X_2 &\coloneqq Y \oplus \mathbf{1}\{\varepsilon_2 \leq E/10\}, \end{aligned}$$



where $\varepsilon_1, \varepsilon_Y, \varepsilon_2 \sim \text{Unif}([0,1])$, mutually independent, and $\oplus$ denotes addition modulo 2. Set $\mathcal{E}_{\text{tr}} = \{1,2,3\}$ and $\mathcal{E} = \mathcal{E}_{\text{tr}} \cup \{8\}$.

Subset method: only $S = \{1\}$ is invariant, with $\mathbb{P}_e(Y = 1 \mid X_1 = 1) = 3/4$ and $\mathbb{P}_e(Y = 1 \mid X_1 = 0) = 1/4$ for all $e \in \mathcal{E}$. The classifier $f_{\{1\}}$ achieves $3/4$ accuracy in every environment.

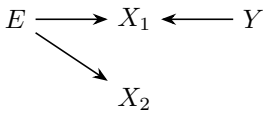
IRM: consider the representation $\Phi(x) \coloneqq x_2$. Although $\mathbb{P}_e(Y = 1 \mid X_2 = 1)$ varies across $\mathcal{E}_{\text{tr}}$, it stays above $1/2$ in every training environment, so the classifier $w = \text{id}$ is Bayes-optimal under the 0–1 loss in all $e \in \mathcal{E}_{\text{tr}}$. The resulting accuracy ranges from $9/10$ to $7/10$ across $\mathcal{E}_{\text{tr}}$, exceeding the $3/4$ of $f_{\{1\}}$ on average. In the test environment $e = 8$, the accuracy of $w \circ \Phi$ drops to $1/5$.

In Ex. 11, IRM accepts Φ as invariant because the Bayes-optimal classifier under the 0–1 loss does not change across $\mathcal{E}_{\text{tr}}$, although $Y \not\perp\!\!\!\perp E \mid \Phi(X)$. This illustrates that function-based invariance criteria can be weaker than the invariance $Y \perp\!\!\!\perp E \mid X_S$ of Def. 1. Here, $w \circ \Phi$ is outperformed by the subset-based $f_{\{1\}}$ in worst-case accuracy on $\mathcal{E}$. However, one could tackle this by modifying the original IRM as follows: we could replace the 0–1 loss with a loss strictly consistent for the mean functional in IRM: under such a loss, IRM would reject Φ since $\mathbb{E}_e[Y \mid \Phi(X)]$ varies across $\mathcal{E}_{\text{tr}}$. In this case, the representation $X \mapsto X_1$ would still satisfy the IRM criterion.

(2) The subset view and the function view differ in the class of test interventions they are robust to. For example, if E has two children, the function view typically requires that the joint distribution of these children shifts with E on $\mathcal{E}$ as it does on $\mathcal{E}_{\text{tr}}$; a subset method, by contrast, often comes with guarantees even when the two children of E are intervened on separately; we make this precise with the following example, where the function view exploits a pattern on $\mathcal{E}_{\text{tr}}$ that does not persist on $\mathcal{E}$.

Example 12 (Worst-case suboptimality of an invariant f)**.** *Let $\mathcal{Y} = \mathbb{R}$, $\ell(y,q) = (y-q)^2$, and consider the following SCM over (X_1, X_2, Y, E) whose structural assignments on $\mathcal{E}_{\mathrm{tr}}$ are*

$$\begin{aligned} Y &\coloneqq \varepsilon_Y, \\ X_1 &\coloneqq \delta_E + Y + \varepsilon_1, \\ X_2 &\coloneqq \delta_E + \varepsilon_2, \end{aligned}$$



where $\delta_E \in \mathbb{R}$ and $\varepsilon_Y, \varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Here, $\mathrm{SB}(Y) = \varnothing$, so $f_{\mathrm{SB}(Y)} = \mathbb{E}[Y] = 0$. The representation $\Phi(X) \coloneqq X_1 - X_2 = Y + \varepsilon_1 - \varepsilon_2$ is independent of E on $\mathcal{E}_{\mathrm{tr}}$, and the predictor $f(x) \coloneqq \mathbb{E}[Y \mid \Phi = x_1 - x_2] = (x_1 - x_2)/3$ is optimal for the squared loss on top of Φ. For all $e \in \mathcal{E}_{\mathrm{tr}}$, $R_e(f) = \frac{2}{3} < 1 = R_e(f_{\mathrm{SB}(Y)})$.

On $\mathcal{E}_{\mathrm{tr}}$, f satisfies several invariance properties: (i) $Y \perp\!\!\!\perp E \mid f(X)$, and thus $\mathbb{E}_e[Y \mid \Phi(X)]$ does not depend on $e \in \mathcal{E}_{\mathrm{tr}}$, so Φ satisfies the IRM invariance criterion of Arjovsky et al. [1, Def. 3]; (ii) $E \perp\!\!\!\perp f(X)$; (iii) invariant residuals $Y - f(X) \perp\!\!\!\perp E$, the invariance notion used by Saengkyongam et al. [54]; and (iv) invariance of the risks.

We now extend the SCM to $\mathcal{E}$: for all $e \in \mathcal{E}_{\mathrm{tr}}$, let $\mathcal{E}$ contain an environment e' in which the structural assignment of X_1 is replaced by $X_1 \coloneqq \tilde{\varepsilon}_1 \sim \mathcal{N}(0,1)$, independent of $(Y, \varepsilon_1, \varepsilon_2)$, while the assignment of X_2 is unchanged. Then Asm. A2 holds. In environment e', $f(X) = (\tilde{\varepsilon}_1 - \delta_{e'} - \varepsilon_2)/3$ is independent of Y. Furthermore, $R_{e'}(f) = 1 + (2 + \delta_{e'}^2)/9 > 1 = R_{e'}(f_{\mathrm{SB}(Y)})$, and hence, $\sup_{e \in \mathcal{E}} R_e(f) > \sup_{e \in \mathcal{E}} R_e(f_{\mathrm{SB}(Y)})$, that is, Inequality (7) is strict.

In Ex. 12, f exploits a pattern on $\mathcal{E}_{\mathrm{tr}}$ that does not extend to $\mathcal{E}$. Our setup in § 2 makes few structural assumptions on the interventions, and in this regime the subset view is robust under the intervention class considered here. Thm. 5 formalizes this. If the interventions are restricted to a known class, the function view can instead be preferable: anchor regression [53], for instance, is worst-case optimal over a specific class of bounded shift interventions. But in this paper, we focus on the more general case.

(3) The subset view is arguably more interpretable and transparent: it combines feature selection with standard supervised learning, rather than an opaque function of all covariates. In regulated settings such as credit scoring or clinical decision support, this simplifies auditing and certifying that prohibited features are not used, and makes individual predictions easier to explain. Only the variables in the chosen invariant subset $S \subseteq \{1, \dots, d\}$ need to be measured at deployment. Finally, the subset view can make the follower's potential actions more transparent, as the choice of which subset of covariates is used to construct f influences the followers' behavior: as in § 1, using only causal features (such as safety equipment) means a follower can change their prediction only by genuinely changing Y, whereas including manipulable non-causal features (such as car color) lets them change the prediction without changing Y.

We return to Ex. 12 to show that the bound (4) and the game (1) can order predictors differently.

Remark 13 (Bound (4) versus the game)**.** *Consider the SCM of Ex. 12 with $\mathcal{E} = \{e_0, e_1\}$, where $\delta_{e_0} = \delta_{e_1} = -1$ and e_1 additionally replaces the assignment of X_1 by $X_1 \coloneqq \tilde{\varepsilon}_1$ as above, so that Asm. A2 holds, and let the follower maximize $V_e^{\mathrm{foll}}(f) = -\mathbb{E}_e[f(X)]$. For $f(x) = (x_1 - x_2)/3$, we have $\mathbb{E}_{e_0}[f(X)] = 0$ and $\mathbb{E}_{e_1}[f(X)] = 1/3$, so $\mathcal{E}^*(f) = \{e_0\}$, whereas $\mathbb{E}_e[f_{\mathrm{SB}(Y)}(X)] = 0$ for both e, so $\mathcal{E}^*(f_{\mathrm{SB}(Y)}) = \mathcal{E}$. The game objective and the bound are therefore*

$$\sup_{e \in \mathcal{E}^*(f)} R_e(f) = R_{e_0}(f) = \tfrac{2}{3}, \qquad \sup_{e \in \mathcal{E}} R_e(f) = R_{e_1}(f) = \tfrac{4}{3},$$

$$\sup_{e \in \mathcal{E}^*(f_{\mathrm{SB}(Y)})} R_e(f_{\mathrm{SB}(Y)}) = 1, \qquad \sup_{e \in \mathcal{E}} R_e(f_{\mathrm{SB}(Y)}) = 1.$$

The bound is thus smaller for $f_{\mathrm{SB}(Y)}$, in line with Thm. 5, while the game objective is smaller for f.

C.5 Generalization to Ex. 7

The following proposition generalizes Ex. 7.

Proposition 14. *Assume $\mathcal{Y} = \mathbb{R}$ and $\ell(y,q) = (y-q)^2$. Let $\mathcal{G}$ be a DAG over $(X_1, \dots, X_d, Y, E)$ with E a source node and no edge $E \to Y$, and let $\bar{\mathcal{G}}$ be the induced subgraph of $\mathcal{G}$ on*

$(X_1, \dots, X_d, Y)$. *If* (6) *fails, then there exists an SCM with graph* $\mathcal{G}$ *and induced distribution* $\mathbb{P}$*, and a set of environments* $\mathcal{E}$ *such that (i) there exists* $e_0 \in \mathcal{E}$ *under which* $\mathbb{P}_{e_0}$ *is faithful w.r.t.* $\bar{\mathcal{G}}$*, (ii) the second part of Asm. A2 holds (that is, for all* $e \in \mathcal{E}$ *there exists an environment* $e' \in \mathcal{E}$ *for which every* X_j*,* $j \in \mathrm{CH}_{\mathrm{int}}(Y)$*, is replaced by exogenous noise independent of all other exogenous noises, with all other structural assignments unchanged), and (iii) there exists a measurable* $f \colon \mathcal{X} \to \mathbb{R}$ *with* $R_e(f) < R_e(f_{\mathrm{SB}(Y)})$ *for all* $e \in \mathcal{E}$*.*

A proof can be found in § E.7.

C.6 Generalizing the result by Rojas-Carulla et al. [51]

The following proposition generalizes Thm. 4 by Rojas-Carulla et al. [51] to losses strictly consistent for the mean (and, in particular, to classification).

Proposition 15 (Adversarial optimality)**.** *Let* S *be an invariant subset of covariates, and let* $N \coloneqq \{1, \dots, d\} \setminus S$*. Let* μ *be a dominating measure on* $\mathcal{X}$ *such that* $\mu = \mu_S \otimes \mu_N$ *for some measures* μ_S *on* $\mathcal{X}_S$ *and* μ_N *on* $\mathcal{X}_N$ *(such as Lebesgue or counting measures). Define* $\mathcal{Q}_S$ *to be the family of all joint distributions* $\mathbb{Q}$ *of* (X, Y) *on* $\mathcal{X} \times \mathcal{Y}$ *such that*

(i) $\mathbb{Q}(Y \in \cdot \,|\, X_S) = \mathbb{P}(Y \in \cdot \,|\, X_S)$ $\quad$ $\mathbb{Q}$*-almost surely,*

(ii) the marginal distribution of X *under* $\mathbb{Q}$ *is absolutely continuous with respect to* μ*.*

Assume $\sup_{\mathbb{Q} \in \mathcal{Q}_S} \mathbb{E}_{\mathbb{Q}}[\ell(Y, f_S(X))] < \infty$*. Then the predictor* f_S *satisfies*

$$f_S \in \operatorname*{argmin}_{\substack{f \colon \mathcal{X} \to \mathcal{A} \\ \text{measurable}}} \sup_{\mathbb{Q} \in \mathcal{Q}_S} \mathbb{E}_{(X,Y) \sim \mathbb{Q}}\big[\ell(Y, f(X))\big].$$

This result is insufficient to answer the question in § 4.1 whether $f_{\mathrm{SB}(Y)}$ is worst-case optimal among all measurable predictors, as, in our setting, the considered distributions do not change with S. A proof can be found in § E.8.

C.7 The stable blanket under hidden variables

Our main results assume a fully observed DAG (§ 2). For the light tunnel experiment (§ 6.2), we hypothesize a hidden common cause of the response and the sensor measurements (Fig. 3, right). Here, we use the stable frontier construction of Xiang [67, Secs. 3 and 6] to extend Lem. 2 and Thm. 3 to graphs with hidden variables and to justify the subset used as the stable blanket in our experiment. Unlike in the fully observed case, the graph alone does not necessarily determine the most predictive subset among those whose invariance follows from the Markov property.

Graphs with hidden variables. Consider an acyclic SCM over (X, H, Y, E) with hidden variables H, where, as in § 2, E is a source node whose children are the intervention targets. We assume that there is no edge from E to Y or into H. For simplicity, we assume that the follower's interventions do not add edges to the graph (in the notation of § 2, $A_j = \varnothing$ for all $j \in \mathrm{CH}(E)$). We work with the latent projection of this graph onto (X, Y, E) [65, Sec. 3.1.3]. This is an acyclic directed mixed graph [ADMG; 49] $\mathcal{M}$ that contains the edge $i \to j$ if the DAG contains a directed path from i to j whose intermediate nodes, if any, are all hidden, and the edge $i \leftrightarrow j$ if the DAG contains a path $i \leftarrow \cdots \leftarrow H_k \rightarrow \cdots \rightarrow j$ whose intermediate nodes are all hidden. In $\mathcal{M}$, a node k is a collider on a path if both adjacent edges have an arrowhead at k (as in $\rightarrow k \leftarrow$ or $\leftrightarrow k \leftrightarrow$), and a path is m-blocked by S if it contains a non-collider in S or a collider k such that neither k nor any of its descendants is in S [49, Sec. 2.1]. We write $E \perp_m Y \,|\, S$ if every path between E and Y in $\mathcal{M}$ is m-blocked by S (m-separation). For disjoint sets of observed nodes, m-separation in $\mathcal{M}$ is equivalent to d-separation in the DAG [50, Prop. 1], so the Markov property of the SCM yields

$$E \perp_m Y \,|\, S \quad \Longrightarrow \quad E \perp\!\!\!\perp Y \,|\, X_S,$$

that is, a set S that m-separates E and Y is invariant (Def. 1). We write $\mathrm{DIS}(j)$ for the district of j in $\mathcal{M}$, that is, the set containing j and all nodes connected to j by paths of bidirected edges, and $\mathrm{DIS}(A) = \bigcup_{j \in A} \mathrm{DIS}(j)$ for a node set A.

Stable frontiers. Following Xiang [67], we now construct an analogue of the stable blanket in $\mathcal{M}$. Intuitively, we want a set of covariates that m-separates E and Y and, like the Markov blanket, is informative about Y without containing irrelevant covariates. Without hidden variables, the stable blanket (3) satisfies these requirements: it equals the Markov blanket of Y in the graph obtained by removing $\mathrm{DE}(Y)_{\mathrm{forb}}$, that is, the intervened children of Y and their descendants, and the removal ensures that the resulting set d-separates E and Y (Lem. 19). With hidden variables, we proceed in the same way: we determine which paths from E to Y the Markov blanket of Y in $\mathcal{M}$ does not block, and we remove covariates so that the Markov blanket of the remaining graph blocks them. By Xiang [67, Prop. 3.3], in an ADMG the Markov blanket of Y within a single environment (that is, for fixed E) is

$$\Big(\mathrm{DIS}(Y) \cup \mathrm{PA}(\mathrm{DIS}(Y)) \cup \mathrm{DIS}(\mathrm{CH}(Y)) \cup \mathrm{PA}(\mathrm{DIS}(\mathrm{CH}(Y)))\Big) \setminus \{Y, E\}. \tag{9}$$

As in § 3.2, the Markov blanket of Y in $\mathcal{M}$ does not necessarily m-separate E and Y: the paths from E to Y that it does not block are exactly those of the two forms

$$E \to X_j \leftrightarrow \cdots \leftrightarrow Y \qquad \text{and} \qquad E \to X_j \leftrightarrow \cdots \leftrightarrow X_k \leftarrow Y, \tag{10}$$

including the shortest cases $E \to X_j \leftrightarrow Y$ and $E \to X_j \leftarrow Y$; nodes of such a path may also be joined by directed edges, but the path uses the bidirected edges. Every covariate on such a path is a collider on it and lies in $\mathrm{DIS}(Y)$ or $\mathrm{DIS}(\mathrm{CH}(Y))$, and hence in (9). Conversely, if a path from E to Y contains a non-collider, then the non-collider closest to Y lies in (9) and blocks the path; otherwise, all intermediate nodes are colliders, so consecutive ones are joined by bidirected edges and the path has one of the two forms.

A covariate on a path of the form (10) is called an admissible collider if it has no directed edge to either of its neighbors on the path and no directed path to Y.[11] The first condition makes the covariate a collider on every path through the same nodes, and the second ensures that removing the covariate and its descendants keeps all ancestors of Y in the graph. Following Xiang [67, Setting 6.1], we assume that every path of the form (10) contains an admissible collider (which is satisfied in § 6.2). This assumption is necessary for the existence of a subset that m-separates E and Y [67, Props. 3.8 and 6.3], and the proof of Prop. 16 below shows that it is also sufficient. It fails, for example, if $E \to X_1 \to Y$ and $X_1 \leftrightarrow Y$: the only covariate X_1 on the path $E \to X_1 \leftrightarrow Y$ has a directed edge to Y and thus violates both conditions, and indeed no subset m-separates E and Y. Xiang [67] constructs the analogue of the stable blanket in three steps.

1. Select an admissible collider set, that is, a set O of admissible colliders that contains at least one admissible collider of every path of the form (10) [67, Def. 6.4].
2. Remove O and its descendants: let $N(O) \subseteq \{1, \dots, d\}$ denote the retained covariates, that is, those that are neither in O nor descendants of O, and let $\mathcal{M}_O$ denote the subgraph of $\mathcal{M}$ induced by $N(O) \cup \{Y, E\}$.
3. Take the Markov blanket of Y in $\mathcal{M}_O$: this is the stable frontier $S(O)$ for O, given by (9) with all sets computed in $\mathcal{M}_O$ [67, Def. 3.12, Thm. 3.14, and Sec. 6].

For a fixed O, $f_{S(O)}$ has the smallest risk in every environment among all predictors based on $X_{N(O)}$ (Prop. 16 (ii) below), but different choices of O retain different covariates and can yield different frontiers. Ex. 17 below goes through the three steps for two such choices. If all choices yield the same frontier, Xiang [67] calls it the stable blanket. Without hidden variables, the only paths of the form (10) are $E \to X_j \leftarrow Y$ with $j \in \mathrm{CH}_{\mathrm{int}}(Y)$, so $O = \mathrm{CH}_{\mathrm{int}}(Y)$ is the only admissible collider set and $S(O) = \mathrm{SB}(Y)$.

Prediction guarantees. The following result extends Lem. 2 and Thm. 3 to graphs with hidden variables, for a fixed admissible collider set. $\mathrm{PA}(Y)$ denotes the parents of Y in $\mathcal{M}$.

Proposition 16. *Let O be an admissible collider set in $\mathcal{M}$.*

(i) $N(O)$ and $S(O)$ are invariant, and $\mathrm{PA}(Y) \subseteq S(O) \subseteq N(O)$.

[11] In the terminology of Xiang [67], paths of the first and second form in (10) correspond to intervened districts (Def. 6.1) and intervened sub-districts (Def. 3.5), and admissible colliders on them to district colliders (Def. 6.2) and sub-district colliders (Def. 3.6); we use one term for both forms because we treat them alike.

(ii) For all $\sigma(X_{N(O)})$-measurable $f\colon \mathcal{X} \to \mathcal{A}$ and all $e \in \mathcal{E}$, we have $R_e(f) \geq R_e(f_{S(O)})$.

(iii) If $\mathrm{PA}(Y)$ is invariant, which holds if $E \perp_m Y \mid \mathrm{PA}(Y)$, then Thm. 3 holds with $S(O)$ in place of $\mathrm{SB}(Y)$.

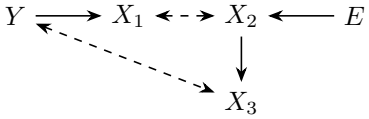
In a DAG, $E \perp_{\mathcal{G}} Y \mid \mathrm{PA}(Y)$ always holds because E is a source node and not a parent of Y. In an ADMG, this can fail, so the invariance of $\mathrm{PA}(Y)$ in (iii) is an assumption.[12]

If Asm. A1 holds with $S(O)$, $O \cup \mathrm{DE}(O)$, and $N(O)$ in place of $\mathrm{SB}(Y)$, $\mathrm{DE}(Y)_{\mathrm{forb}}$, and $\mathrm{DE}(Y)^C_{\mathrm{forb}}$, then $f_{S(O)}$ solves Problem (5), that is, $\sup_{e\in\mathcal{E}} R_e(f) \geq \sup_{e\in\mathcal{E}} R_e(f_{S(O)})$ for all measurable f. The proof of Thm. 5 applies unchanged, since it uses only the invariance of both sets and the identity $\mathbb{E}_e[Y \mid X_{N(O)}] = f_{S(O)}(X)$. We leave extensions of Asm. A2 and Lem. 4 open.

The graph does not necessarily determine the most predictive m-separating subset. Two admissible collider sets can yield frontiers that are not nested, and the graph alone need not determine the most predictive subset among those whose invariance follows from it.

Example 17. *Consider the SCM with parameter $a > 0$ and ADMG $\mathcal{M}$ shown on the right:*

$$Y \coloneqq \varepsilon_Y, \qquad X_1 \coloneqq aY + \varepsilon_1,$$
$$X_2 \coloneqq \delta_E + \varepsilon_2, \qquad X_3 \coloneqq X_2 + \varepsilon_3,$$



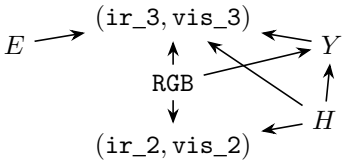
where $\delta_E \in \mathbb{R}$ shifts X_2 depending on E, and the noise variables are centered, jointly Gaussian, and independent of E, with unit variances and the only nonzero cross-covariances $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = \mathrm{Cov}(\varepsilon_Y, \varepsilon_3) = 1/2$ due to hidden common causes. The only path of the form (10) is $E \to X_2 \leftrightarrow X_1 \leftarrow Y$. Both X_2 and X_1 are admissible colliders on it. Choosing $O_A \coloneqq \{2\}$ yields the frontier $S_A \coloneqq S(O_A) = \{1\}$; choosing $O_B \coloneqq \{1\}$ yields $S_B \coloneqq S(O_B) = \{2, 3\}$. Both S_A and S_B m-separate E and Y and are thus invariant. For the squared loss, the two predictors and their risks are

$$f_{S_A}(X) = \frac{aX_1}{1+a^2}, \qquad R_e(f_{S_A}) = \frac{1}{1+a^2},$$
$$f_{S_B}(X) = \frac{X_3 - X_2}{2}, \quad R_e(f_{S_B}) = \frac{3}{4}.$$

For $a = 2$, f_{S_A} has the smaller risk, and for $a = 1/2$, f_{S_B} has the smaller risk, while the graph is the same in both cases.

Of the eight subsets of $\{1, 2, 3\}$ in Ex. 17, exactly $\varnothing$, $\{1\}$, $\{2\}$, and $\{2, 3\}$ m-separate E and Y, and the predictors of $\varnothing$ and $\{2\}$ have risk 1, so the most predictive m-separating subset is S_A or S_B, depending on a.[13]

The light tunnel. We now apply the above to the hypothesized graph of the light tunnel, reproduced on the right from Fig. 3. In the corresponding ADMG $\mathcal{M}$, the nodes Y, `ir_2`, `vis_2`, `ir_3`, `vis_3` are pairwise connected by bidirected edges, `RGB` is a parent of each of them, and Y and E are parents of `ir_3` and `vis_3`. The assumption that every path of the form (10) contains an admissible collider holds: every such path starts with $E \to$ `ir_3` or $E \to$ `vis_3`, and because `ir_3` and `vis_3` have no children, each is an admissible collider on every such path containing it. Hence $O^* \coloneqq \{$`ir_3`, `vis_3`$\}$ is an admissible collider set. Conversely, the paths $E \to$ `ir_3` $\leftrightarrow Y$ and $E \to$ `vis_3` $\leftrightarrow Y$ contain no other covariate, so every admissible collider set contains O^*. The retained covariates are $N(O^*) = \{$`RGB`, `ir_2`, `vis_2`$\}$, and in $\mathcal{M}_{O^*}$ we have



$$\mathrm{DIS}(Y) = \{Y, \texttt{ir_2}, \texttt{vis_2}\}, \qquad \mathrm{PA}(\mathrm{DIS}(Y)) = \{\texttt{RGB}\}, \qquad \mathrm{CH}(Y) = \varnothing,$$

[12] For example, in the ADMG with edges $E \to X_1 \to X_2 \leftrightarrow Y$ and $X_2 \to X_3 \to Y$, conditioning on $\mathrm{PA}(Y) = \{X_3\}$ opens the path $E \to X_1 \to X_2 \leftrightarrow Y$ at the collider X_2.

[13] The two frontiers in Ex. 17 cannot be merged into an m-separating subset, but an invariant function can use both. With $Z \coloneqq X_3 - X_2$, the distribution of (Y, X_1, Z) does not depend on e, and $g(X) \coloneqq \mathbb{E}[Y \mid X_1, Z] = (3aX_1 + 2Z)/(3a^2+4)$ has risk $3/(3a^2+4)$, which is smaller than both $R_e(f_{S_A})$ and $R_e(f_{S_B})$ for every $a > 0$; see § C.4 for a discussion of such function-based predictors.

so $S(O^*) = \{\texttt{RGB}, \texttt{ir_2}, \texttt{vis_2}\}$. Other admissible collider sets yield smaller frontiers; for example, $O^* \cup \{\texttt{ir_2}\}$ yields $\{\texttt{RGB}, \texttt{vis_2}\}$. The frontier is thus not unique, so there is no stable blanket in the sense of Xiang [67], but $S(O^*)$ contains every other frontier (in § 6.2, we thus write $\mathrm{SB}(Y)$ for this largest frontier in a slight abuse of notation). By Prop. 16, $f_{\mathrm{SB}(Y)}$ has the smallest risk in every environment among all predictors that do not use `ir_3` or `vis_3`. Moreover, every path from E to Y starts with $E \to \texttt{ir_3}$ or $E \to \texttt{vis_3}$ followed by an edge pointing into that node, so a set m-separates E and Y if and only if it contains neither `ir_3` nor `vis_3`. Hence, every m-separating subset is contained in $N(O^*)$, and $f_{\mathrm{SB}(Y)}$ also has the smallest risk in every environment among all predictors based on m-separating subsets. In particular, $\mathrm{PA}(Y) = \{\texttt{RGB}\}$ is invariant, so $f_{\mathrm{SB}(Y)}$ weakly dominates $f_{\texttt{RGB}}$ in the game (1) for both follower classes of Thm. 3. Consistent with this, $\hat{f}_{\texttt{RGB,ir_2,vis_2}}$ has a lower MSE than $\hat{f}_{\texttt{RGB}}$ while the MSE of both predictors remains essentially constant under the follower's interventions, and SC recovers exactly this subset from data (Fig. 4).

C.8 Discussion of assumptions

The assumption on A_j in Lem. 4. Lem. 4 assumes $A_j \subseteq \mathrm{DE}_{\mathcal{G}_0}(Y)^C_{\mathrm{forb}} \setminus \{Y, E\}$ for all $j \in \mathrm{CH}_{\mathcal{G}_0}(E)$ and that $\mathcal{G}$ is acyclic. The sets A_j restrict only the new parents of an intervention target X_j: an added edge $X_k \to X_j$ means that the new mechanism of X_j uses the realized value of X_k, so A_j specifies which new inputs are available before X_j is set. In Fig. 1, the new mechanism of X_4 may use X_1 and X_3 (dashed edges), but not X_5, which is realized only after X_4 is set. None of X_3, X_4, and X_5 can become a new parent of X_1, since they are consequences of X_1 and the added edge would create a cycle. In this example, the restriction to $\mathrm{DE}_{\mathcal{G}_0}(Y)^C_{\mathrm{forb}}$ excludes nothing beyond acyclicity, because X_4 and X_5 are downstream of X_1, and X_5 of X_4.

The assumption is realistic for actions chosen before their consequences are observed, such as changes to an application before submitting it or a treatment chosen from baseline measurements, but less so for adaptive settings, such as changing one feature after observing the system's response to a previous change. Excluding Y as a new parent is plausible when Y is latent or realized later, but not when followers act after observing the outcome. If the restriction on A_j fails, the results of § 3 and § 4 other than Lem. 4 still apply to the acyclic $\mathcal{G}$ under their stated assumptions, and $\mathrm{SB}_{\mathcal{G}}(Y) \subseteq \mathrm{SB}_{\mathcal{G}_0}(Y)$ still holds because new edges only enter intervention targets. If the inclusion is strict, however, $\mathrm{SB}_{\mathcal{G}_0}(Y)$ contains a descendant of an intervened child X_i of Y, and conditioning on it opens the path $E \to X_i \leftarrow Y$, so a leader who reads the stable blanket off $\mathcal{G}_0$ uses a set that is in general no longer invariant in $\mathcal{G}$.

Asm. A1 and A2. Asm. A1 and A2 do not restrict the intervention model of § 2; they require that $\mathcal{E}$ contains, for each environment, a corresponding strong intervention, not that every environment is one. Only Thm. 5 and its extensions in the appendix use them; Prop. 6 uses a relaxed form.

Asm. A1 requires, for each e, an environment e' that acts like e on $X_{\mathrm{SB}(Y)}$ (same marginal distribution) and in which $X_{\mathrm{DE}(Y)_{\mathrm{forb}}}$ is uninformative about Y given $X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}$. If $\mathcal{E}$ contains, for example, only bounded soft or additive interventions, $X_{\mathrm{DE}(Y)_{\mathrm{forb}}}$ can be informative about Y given $X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}$ in every environment, so no e' satisfies condition (ii) of Asm. A1. Asm. A2 additionally restricts the graph and is plausible when (a) every intervened child of Y can be set to exogenous noise or a fixed value, (b) all other mechanisms are retained, and (c) no non-intervened child of Y is downstream of an intervened child. In a health setting, for example, Asm. A2 holds if the children of a health outcome Y are sensor readings of Y or questionnaires answered by the patients themselves, manipulating one does not affect the others, and $\mathcal{E}$ contains, for each e, the environment that sets all intervened children jointly to independent noise while leaving the other mechanisms of e unchanged.

Prop. 14 shows for regression that for every DAG violating (6), there exists an SCM with this DAG for which $f_{\mathrm{SB}(Y)}$ is not worst-case optimal. At the same time, for every DAG there is an SCM with that graph for which $f_{\mathrm{SB}(Y)}$ is worst-case optimal: let Y be a deterministic function of $X_{\mathrm{PA}(Y)} \subseteq X_{\mathrm{SB}(Y)}$ (constant if $\mathrm{PA}(Y) = \varnothing$), so that $f_{\mathrm{SB}(Y)} = Y$ almost surely in every environment. Thus, (6) cannot be dropped from a worst-case guarantee for $f_{\mathrm{SB}(Y)}$ that holds for all SCMs with a given graph, but the graph alone cannot decide worst-case optimality for a given SCM.

D Stabilized regression and classification

D.1 Invariance tests for classification

We now provide detailed descriptions of the invariance tests that can be used for the methods described in § 5. All tests target the conditional independence $H_{0,S}: Y \perp\!\!\!\perp E \mid X_S$ and return a p-value used as the invariance score $s_{\text{inv}}(S)$. Each test can be used with different base models (such as logistic regression, random forests, or gradient boosting).

D.1.1 Residual-based tests

Residual-based tests fit a model $\hat{f}$ for $\mathbb{E}[Y \mid X_S]$ on data pooled across environments, compute residuals $R = Y - \hat{f}(X_S)$, and test whether the distribution of R depends on E.

Invariant Residual Distribution Test (IRD). Tests whether the mean of R varies across environments using a one-way ANOVA F-test [adapted from 25].

TRAM-GCM. Applies the generalized covariance measure [58] to the residuals of both predicting Y from X_S and E from X_S, testing whether their expected product is zero [33].

D.1.2 Predictive-performance-based tests

Predictive-performance-based tests check whether E carries additional predictive information about Y beyond X_S, or vice versa.

Invariant Target Prediction Test (ITP). Compares the ROC AUC of a classifier predicting Y from (X_S, E) against one using X_S and permuted observations of E via DeLong's test [14] for correlated ROC curves [15, 25]. A significant AUC decrease upon permuting E rejects invariance.

Invariant Environment Prediction Test (IEP). Reverses the direction and tests whether Y helps predict E given X_S. It compares the cross-entropy loss of a model predicting E from (X_S, Y) against one using X_S and permuted observations of Y via a paired t-test [adapted from 25].

D.2 Algorithm

Stabilized regression [48] uses a two-stage procedure to determine the weights $\hat{\omega}_S$ for $S \subseteq \{1, \dots, d\}$. We now propose stabilized classification, which is directly ported from stabilized regression and proceeds in the same way.

Stage 1: invariance. First, we estimate the collection of invariant subsets by computing an invariance score $s_{\text{inv}}(S)$ for each subset $S \subseteq \{1, \dots, d\}$. Then, we define the empirically invariant subsets as $\hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}} := \left\{S \subseteq \{1, \dots, d\} \mid s_{\text{inv}}(S) \geq \alpha_{\text{inv}}\right\}$, where the cutoff $\alpha_{\text{inv}} \in (0, 1)$ is a tuning parameter. § D.1 discusses suitable scores $s_{\text{inv}}(S)$ for this step. Because iterating over all 2^d subsets becomes computationally prohibitive even for moderate d, an optional variable screening step (e.g., using ℓ_1-penalized logistic regression) can be applied to reduce the candidate covariate set before computing the invariance scores.

Stage 2: prediction performance. We estimate the subsets with best prediction performance among those in $\hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}}$. We evaluate the prediction performance of each $S \in \hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}}$ using a score $s_{\text{pred}}(S)$, such as the negative binary cross-entropy loss on the training data. We then define the final set of selected subsets as $\hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}} := \left\{S \in \hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}} \mid s_{\text{pred}}(S) \geq c(\alpha_{\text{pred}})\right\}$, where $c(\alpha_{\text{pred}})$ is a cutoff threshold. Following Pfister et al. [48], we determine $c(\alpha_{\text{pred}})$ using a bootstrap procedure. Let $\hat{S}_{\max} := \operatorname{argmax}_{S \in \hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}}} s_{\text{pred}}(S)$ denote the invariant subset with best prediction performance. We generate B bootstrap samples from the training data, compute the prediction scores for $\hat{S}_{\max}$ on each sample, and set $c(\alpha_{\text{pred}})$ to the α_{pred}-quantile of these bootstrap scores. Therefore, the hyperparameters $\alpha_{\text{inv}}, \alpha_{\text{pred}} \in (0, 1)$ control the size of the ensemble. Finally, we use uniform weights

Algorithm 1 Stabilized classification (SC)

Require: data $\mathcal{D} = (x_i, y_i, e_i)_{i=1}^n$ from $\mathcal{E}_{\text{tr}}$; hyperparameters $\alpha_{\text{inv}}, \alpha_{\text{pred}} \in (0,1)$, $B \in \mathbb{N}$

(optional) perform variable screening to reduce d
for all $S \subseteq \{1, \dots, d\}$ **do**
    compute invariance score $s_{\text{inv}}(S)$ using $\mathcal{D}$
end for
$\hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}} \leftarrow \{S \subseteq \{1, \dots, d\} \mid s_{\text{inv}}(S) \geq \alpha_{\text{inv}}\}$ ▷ invariance
for all $S \in \hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}}$ **do**
    fit base classifier $\hat{f}_S$ on $\mathcal{D}$
    compute prediction score $s_{\text{pred}}(S)$ on $\mathcal{D}$
end for
$\hat{S}_{\max} \leftarrow \operatorname{argmax}_{S \in \hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}}} s_{\text{pred}}(S)$
compute cutoff $c(\alpha_{\text{pred}})$ as the α_{pred}-quantile of B bootstrap scores for $\hat{S}_{\max}$
$\hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}} \leftarrow \{S \in \hat{\mathcal{I}}_{\mathcal{E}_{\text{tr}}} \mid s_{\text{pred}}(S) \geq c(\alpha_{\text{pred}})\}$ ▷ prediction performance
assign weights $\hat{\omega}_S \leftarrow \frac{1}{|\hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}}|} \mathbf{1}\{S \in \hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}}\}$

Output: base estimators $\hat{f}_S$ and weights $\hat{\omega}_S$ defining the ensemble $\hat{f}^{\text{SC}}$

$\hat{\omega}_S$ over the selected subsets in $\hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}}$: $\hat{\omega}_S := \frac{1}{|\hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}}|}$ for $S \in \hat{\mathcal{C}}_{\mathcal{E}_{\text{tr}}}$ and $\hat{\omega}_S := 0$ otherwise. Algorithm 1 summarizes the full procedure.

E Proofs

The following lemma serves as a generalization of the orthogonality of the conditional expectation (used for regression ($Y \in \mathbb{R}$) with the MSE) to regression and binary classification with loss functions strictly consistent for the mean functional. It is used for several results.

Lemma 18. *Let Y and W be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with Y taking values in $\mathcal{Y}$ and integrable. Let Z be a $\sigma(W)$-measurable random variable taking values in $\mathcal{A}$. If ℓ is strictly consistent for the mean functional, then*

$$\mathbb{E}\big[\ell(Y, Z)\big] \;\geq\; \mathbb{E}\big[\ell\big(Y, \mathbb{E}[Y \mid W]\big)\big],$$

where $\mathbb{E}[\cdot]$ denotes expectation under $\mathbb{P}$. If both expectations are finite, equality holds if and only if $Z = \mathbb{E}[Y \mid W]$ almost surely.

Proof of Lem. 18. For all $q \in \mathcal{A}$ and probability distributions F on $\mathcal{Y}$ with finite mean $\mathbb{E}_{Y \sim F}[Y]$, define

$$L(F, q) := \mathbb{E}_{Y \sim F}[\ell(Y, q)].$$

Let F_W denote the conditional distribution of Y given W [28, Thm. 8.5]. Its mean equals $\mathbb{E}[Y \mid W]$ almost surely. Conditional on W, the random variable Z acts as a constant since Z is $\sigma(W)$-measurable, so almost surely,

$$\mathbb{E}[\ell(Y, Z) \mid W] = L(F_W, Z).$$

By strict consistency of ℓ, for every distribution F with finite mean, $q \mapsto L(F, q)$ is uniquely minimized at $q = \mathbb{E}_{Y \sim F}[Y]$. Since this holds for every F, it holds pointwise for the random conditional distribution F_W for almost every $\omega \in \Omega$. Thus, almost surely,

$$L(F_W, Z) \geq L(F_W, \mathbb{E}[Y \mid W]) = \mathbb{E}\big[\ell(Y, \mathbb{E}[Y \mid W]) \mid W\big],$$

with equality if and only if $Z = \mathbb{E}[Y \mid W]$ almost surely. Taking expectations, $\mathbb{E}[\ell(Y, Z)] \geq \mathbb{E}[\ell(Y, \mathbb{E}[Y \mid W])]$. For the equality case, assume both expectations are finite and equal. Then the nonnegative random variable $D := \mathbb{E}[\ell(Y, Z) \mid W] - \mathbb{E}[\ell(Y, \mathbb{E}[Y \mid W]) \mid W]$ satisfies $\mathbb{E}[D] = 0$, hence $D = 0$ almost surely, and pointwise strictness gives $Z = \mathbb{E}[Y \mid W]$ almost surely. The converse is immediate. □

E.1 Proof of Lem. 2

We first state and prove a lemma. The d-separation statement for $\mathrm{SB}(Y)$ is also shown in the proof of Pfister et al. [48, Thm. 3.5]; we include the argument for completeness and extend it to $\mathrm{DE}(Y)^C_{\mathrm{forb}}$.

Lemma 19. *The subsets* $\mathrm{DE}(Y)^C_{\mathrm{forb}}$ *and* $\mathrm{SB}(Y)$ *satisfy the d-separations*

$$Y \perp_{\mathcal{G}} E \,|\, \mathrm{DE}(Y)^C_{\mathrm{forb}} \qquad \textit{and} \qquad Y \perp_{\mathcal{G}} E \,|\, \mathrm{SB}(Y).$$

Remark 20. *The statement and proof of Lem. 19 only use that E is a source node in $\mathcal{G}$ with no edge $E \to Y$, and apply unchanged to any DAG over $(X_1, \dots, X_d, Y, E)$ with these properties, provided* $\mathrm{DE}(Y)_{\mathrm{forb}}$ *and* $\mathrm{SB}(Y)$ *are computed in that DAG.*

Proof of Lem. 19. We prove both statements simultaneously. Let $S \in \{\mathrm{DE}(Y)^C_{\mathrm{forb}}, \mathrm{SB}(Y)\}$ and recall that $\mathrm{PA}(Y) \subseteq S \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$ from (3). Consider an arbitrary path from E to Y in $\mathcal{G}$. Since E is a source node and not a parent of Y, the path contains an intermediate node i adjacent to Y. We distinguish two cases.

Case 1. $E \to \cdots\, i \to Y$**.** Then $i \in \mathrm{PA}(Y) \subseteq S$, so the path is blocked.

Case 2. $E \to \cdots\, i \leftarrow Y$**.** Since E is a source node, the path contains at least one collider.

- If $i \in \mathrm{DE}(Y)_{\mathrm{forb}}$: let k be the collider closest to Y on the path (possibly $k = i$). The segment from Y to k is directed ($Y \to i \to \cdots \to k$), so $k \in \mathrm{DE}(i) \cup \{i\}$. Since $i \in \mathrm{DE}(Y)_{\mathrm{forb}}$, and because descendants of nodes in $\mathrm{DE}(Y)_{\mathrm{forb}}$ are again in $\mathrm{DE}(Y)_{\mathrm{forb}}$, we have $\mathrm{DE}(k) \cap \mathrm{DE}(Y)^C_{\mathrm{forb}} = \varnothing$. Hence neither k nor any descendant of k lies in $S \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$, and the path is blocked.
- If $i \in \mathrm{DE}(Y)^C_{\mathrm{forb}}$: then $i \in \mathrm{CH}(Y) \cap \mathrm{DE}(Y)^C_{\mathrm{forb}} \subseteq S$. If i is not a collider, the path is blocked because $i \in S$. If i is a collider, the path has the form $E \to \cdots j \to i \leftarrow Y$ for some j with $i \neq j \neq E$ (if $j = E$, then $i \in \mathrm{CH}(E) \cap \mathrm{CH}(Y) = \mathrm{CH}_{\mathrm{int}}(Y)$, contradicting $i \in \mathrm{DE}(Y)^C_{\mathrm{forb}}$). Then $j \in \mathrm{PA}(\mathrm{CH}(Y) \cap \mathrm{DE}(Y)^C_{\mathrm{forb}}) \setminus \{Y\} \subseteq \mathrm{SB}(Y)$ by (3), and $j \in \mathrm{DE}(Y)^C_{\mathrm{forb}}$ since $\mathrm{DE}(Y)_{\mathrm{forb}}$ is closed under descendants. Thus $j \in S$, and since j is a non-collider on the path, the path is blocked.

□

Proof of Lem. 2. By Lem. 19 and the global Markov property, $\mathrm{DE}(Y)^C_{\mathrm{forb}}$ and $\mathrm{SB}(Y)$ are invariant subsets. In particular, for all $e \in \mathcal{E}$, the conditional mean functions representing $\mathbb{E}_e[Y \,|\, X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}]$ and $\mathbb{E}_e[Y \,|\, X_{\mathrm{SB}(Y)}]$ do not depend on e. By definition of $\mathrm{SB}(Y)$ [48, Def. 3.4], for all $j \in \mathrm{DE}(Y)^C_{\mathrm{forb}} \setminus \mathrm{SB}(Y)$ the d-separation $j \perp_{\mathcal{G}} Y \,|\, \mathrm{SB}(Y)$ holds. Thus, $\big(\mathrm{DE}(Y)^C_{\mathrm{forb}} \setminus \mathrm{SB}(Y)\big) \perp_{\mathcal{G}} Y \,|\, \mathrm{SB}(Y)$, so by the global Markov property,

$$X_{\mathrm{DE}(Y)^C_{\mathrm{forb}} \setminus \mathrm{SB}(Y)} \perp\!\!\!\perp Y \,|\, X_{\mathrm{SB}(Y)}$$

holds under $\mathbb{P}_e$ for all $e \in \mathcal{E}$. Doob's conditional independence property [28, Thm. 8.9] then yields that almost surely,

$$\mathbb{E}[Y \,|\, X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}] = \mathbb{E}[Y \,|\, X_{\mathrm{SB}(Y)}, X_{\mathrm{DE}(Y)^C_{\mathrm{forb}} \setminus \mathrm{SB}(Y)}] = \mathbb{E}[Y \,|\, X_{\mathrm{SB}(Y)}] = f_{\mathrm{SB}(Y)}(X). \tag{11}$$

Let $e \in \mathcal{E}$ and let $f \colon \mathcal{X} \to \mathcal{A}$ be $\sigma(X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}})$-measurable. Applying Lem. 18 under $\mathbb{P}_e$ with $W = X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}$ and $Z = f(X)$ gives

$$R_e(f) \geq \mathbb{E}_e\big[\ell(Y, \mathbb{E}[Y \,|\, X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}])\big] \overset{(11)}{=} R_e(f_{\mathrm{SB}(Y)}).$$

Since e and f were arbitrary, we obtain the result. □

E.2 Proof of Thm. 3

Proof of Thm. 3. For (i), $\mathcal{E}^*(f) = \operatorname{argmax}_{e\in\mathcal{E}} R_e(f)$, so $\sup_{e\in\mathcal{E}^*(f)} R_e(f) = \sup_{e\in\mathcal{E}} R_e(f)$ for every f. By Lem. 2, $R_e(f_{\mathrm{SB}(Y)}) \le R_e(f_{\mathrm{PA}(Y)})$ for all $e \in \mathcal{E}$, and taking suprema gives the claim.

For (ii), since $\mathrm{PA}(Y)$ and $\mathrm{SB}(Y)$ are invariant, the tower property gives $\mathbb{E}_e[f_{\mathrm{PA}(Y)}(X)] = \mathbb{E}_e[f_{\mathrm{SB}(Y)}(X)] = \mathbb{E}_e[Y]$ for all $e \in \mathcal{E}$. Hence $V_e^{\mathrm{foll}}(f_{\mathrm{PA}(Y)}) = V_e^{\mathrm{foll}}(f_{\mathrm{SB}(Y)})$ for all e, which implies $\mathcal{E}^*(f_{\mathrm{PA}(Y)}) = \mathcal{E}^*(f_{\mathrm{SB}(Y)})$. Therefore

$$\sup_{e\in\mathcal{E}^*(f_{\mathrm{SB}(Y)})} R_e(f_{\mathrm{SB}(Y)}) \le \sup_{e\in\mathcal{E}^*(f_{\mathrm{PA}(Y)})} R_e(f_{\mathrm{PA}(Y)}),$$

where the inequality uses Lem. 2. □

E.3 Proof of Lem. 4

Proof of Lem. 4. Since $A_j \subseteq \mathrm{DE}_{\mathcal{G}_0}(Y)^C_{\mathrm{forb}} \setminus \{Y, E\}$ for all j, we have $\mathrm{CH}_{\mathcal{G}}(Y) = \mathrm{CH}_{\mathcal{G}_0}(Y)$, $\mathrm{CH}_{\mathcal{G}}(E) = \mathrm{CH}_{\mathcal{G}_0}(E)$, and $\mathrm{CH}_{\mathrm{int}}(Y)$ is the same in $\mathcal{G}$ and $\mathcal{G}_0$.

First, we show $\mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}} = \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$. The inclusion $\mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}} \subseteq \mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}}$ holds because adding edges can only enlarge the set of descendants. For the reverse, assume by contradiction that there exists $i \in \mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}}$ with $i \notin \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$. Then, there exists a directed path in $\mathcal{G}$ from some $j \in \mathrm{CH}_{\mathrm{int}}(Y)$ to i that uses at least one edge only existing in $\mathcal{G}$ (since this path does not exist in $\mathcal{G}_0$). Let $k \to l$ be the first such edge on the path (with $k \in A_l \subseteq \mathrm{DE}_{\mathcal{G}_0}(Y)^C_{\mathrm{forb}}$). The segment $j \to \cdots \to k$ uses only edges of $\mathcal{G}_0$, so $k \in \mathrm{DE}_{\mathcal{G}_0}(\mathrm{CH}_{\mathrm{int}}(Y)) \cup \mathrm{CH}_{\mathrm{int}}(Y) = \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$. This contradicts $k \in A_l \subseteq \mathrm{DE}_{\mathcal{G}_0}(Y)^C_{\mathrm{forb}}$.

Second, we show $\mathrm{SB}_{\mathcal{G}}(Y) = \mathrm{SB}_{\mathcal{G}_0}(Y)$. By (3), the stable blanket $\mathrm{SB}_{\mathcal{G}_0}(Y)$ is the union of $\mathrm{PA}_{\mathcal{G}_0}(Y)$, $\mathrm{CH}_{\mathcal{G}_0}(Y) \setminus \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$, and $\mathrm{PA}_{\mathcal{G}_0}(\mathrm{CH}_{\mathcal{G}_0}(Y) \setminus \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}) \setminus \{Y\}$. Regarding the first set, $\mathrm{PA}_{\mathcal{G}}(Y) = \mathrm{PA}_{\mathcal{G}_0}(Y)$ because no intervention on Y takes place. The second is unchanged in $\mathcal{G}$ since $\mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}} = \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$ and $\mathrm{CH}_{\mathcal{G}}(Y) = \mathrm{CH}_{\mathcal{G}_0}(Y)$. The third is unchanged in $\mathcal{G}$ since every new edge points into $\mathrm{CH}_{\mathcal{G}_0}(E)$, and $\mathrm{CH}_{\mathcal{G}_0}(E) \cap (\mathrm{CH}_{\mathcal{G}_0}(Y) \setminus \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}) = \varnothing$ since $\mathrm{CH}_{\mathcal{G}_0}(E) \cap \mathrm{CH}_{\mathcal{G}_0}(Y) = \mathrm{CH}_{\mathrm{int}}(Y) \subseteq \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$. Hence, no new edge enters $\mathrm{CH}_{\mathcal{G}_0}(Y) \setminus \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}}$, and $\mathrm{PA}_{\mathcal{G}}(\mathrm{CH}_{\mathcal{G}}(Y) \setminus \mathrm{DE}_{\mathcal{G}}(Y)_{\mathrm{forb}}) = \mathrm{PA}_{\mathcal{G}_0}(\mathrm{CH}_{\mathcal{G}_0}(Y) \setminus \mathrm{DE}_{\mathcal{G}_0}(Y)_{\mathrm{forb}})$.

Finally, $Y \perp_{\mathcal{G}} E \,|\, \mathrm{SB}_{\mathcal{G}_0}(Y)$ follows since $\mathrm{SB}_{\mathcal{G}}(Y) = \mathrm{SB}_{\mathcal{G}_0}(Y)$ and by Lem. 19 and Rmk. 20, $Y \perp_{\mathcal{G}} E \,|\, \mathrm{SB}_{\mathcal{G}}(Y)$. □

E.4 Proof of Thm. 5

We first state and prove a lemma.

Lemma 21 (Asm. A2 implies Asm. A1)**.** *If Asm. A2 holds, then for every $e \in \mathcal{E}$ the environment e' provided by Asm. A2 satisfies both conditions of Asm. A1.*

Proof of Lem. 21. Fix an arbitrary $e \in \mathcal{E}$ and let $e' \in \mathcal{E}$ be the corresponding environment provided by Asm. A2, in which for all $j \in \mathrm{CH}_{\mathrm{int}}(Y)$, we set $X_j := \tilde{\varepsilon}_j$ with $\tilde{\varepsilon}_j$ independent of all other exogenous noises, removing incoming edges to $\mathrm{CH}_{\mathrm{int}}(Y)$ in $\mathcal{G}$. All other structural assignments are unchanged. Let $\mathcal{G}'$ denote the resulting graph. Unless stated otherwise, all graph-theoretic notation such as $\mathrm{PA}(\cdot)$ refers to the original graph $\mathcal{G}$. Quantities defined with respect to $\mathcal{G}'$ are marked explicitly.

We first verify that Asm. A1 (i) holds. Since $\mathrm{DE}(Y)_{\mathrm{forb}}$ is closed under descendants, no node in $\mathrm{SB}(Y) \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$ is a descendant of any node in $\mathrm{CH}_{\mathrm{int}}(Y)$. Because only the structural assignments of $\mathrm{CH}_{\mathrm{int}}(Y)$ are modified in e', the marginal distribution of $X_{\mathrm{SB}(Y)}$ is the same under $\mathbb{P}_e$ and $\mathbb{P}_{e'}$.

To verify Asm. A1 (ii), we need to verify $Y \perp\!\!\!\perp X_{\mathrm{DE}(Y)_{\mathrm{forb}}} \,|\, X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}$ under $\mathbb{P}_{e'}$. Since $\mathbb{P}_{e'}$ is generated by the modified SCM, it is Markov with respect to $\mathcal{G}'$. It therefore suffices to show that $Y \perp_{\mathcal{G}'} \mathrm{DE}(Y)_{\mathrm{forb}} \,|\, \mathrm{DE}(Y)^C_{\mathrm{forb}}$. Consider an arbitrary path in $\mathcal{G}'$ from Y to some $j \in \mathrm{DE}(Y)_{\mathrm{forb}}$, and let i be the node adjacent to Y on this path.

- If $Y \leftarrow i$, then $i \in \mathrm{PA}(Y) \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$, so the path is blocked by $\mathrm{DE}(Y)^C_{\mathrm{forb}}$.

- If $Y \to i$ with $i \in \mathrm{CH}(Y) \setminus \mathrm{DE}(Y)^C_{\mathrm{forb}}$, then $i \in \mathrm{DE}(Y)_{\mathrm{forb}}$, and condition (6) gives $i \in \mathrm{CH}_{\mathrm{int}}(Y)$. But in $\mathcal{G}'$ the edge $Y \to i$ is deleted, so this path does not exist.
- If $Y \to i$ with $i \in \mathrm{CH}(Y) \cap \mathrm{DE}(Y)^C_{\mathrm{forb}}$, then if i is a non-collider, the path is blocked since $i \in \mathrm{DE}(Y)^C_{\mathrm{forb}}$. If i is a collider, the path has the form $Y \to i \leftarrow k \cdots$ for some $k \in \mathrm{PA}_{\mathcal{G}'}(i) \setminus \{Y\}$. Since $i \in \mathrm{CH}(Y) \cap \mathrm{DE}(Y)^C_{\mathrm{forb}}$, the node i is not intervened on, hence $\mathrm{PA}_{\mathcal{G}'}(i) = \mathrm{PA}(i)$. Therefore, $k \in \mathrm{PA}(\mathrm{CH}(Y) \cap \mathrm{DE}(Y)^C_{\mathrm{forb}}) \setminus \{Y\} \subseteq \mathrm{SB}(Y) \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$ by (3). Since k is a non-collider on the path, the path is blocked.

Since every path is blocked, $Y \perp_{\mathcal{G}'} \mathrm{DE}(Y)_{\mathrm{forb}} \mid \mathrm{DE}(Y)^C_{\mathrm{forb}}$, and the global Markov property gives $Y \perp\!\!\!\perp X_{\mathrm{DE}(Y)_{\mathrm{forb}}} \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}$ under $\mathbb{P}_{e'}$. □

Proof of Thm. 5. By Lem. 21, Asm. A2 implies Asm. A1, so it suffices to prove the claim under Asm. A1. Let $f\colon \mathcal{X} \to \mathcal{A}$ be an arbitrary measurable predictor. Fix $e \in \mathcal{E}$ and let $e' \in \mathcal{E}$ be the corresponding environment provided by Asm. A1. Applying Lem. 18 under $\mathbb{P}_{e'}$ with $W = X$ and $Z = f(X)$,

$$R_{e'}(f) \geq \mathbb{E}_{e'}\big[\ell(Y, \mathbb{E}_{e'}[Y \mid X])\big]. \tag{12}$$

By Asm. A1 (ii), $Y \perp\!\!\!\perp X_{\mathrm{DE}(Y)_{\mathrm{forb}}} \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}$ under $\mathbb{P}_{e'}$, so by Doob's conditional independence property [28, Thm. 8.9], $\mathbb{E}_{e'}[Y \mid X] = \mathbb{E}_{e'}[Y \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}]$ almost surely. Since $\mathrm{DE}(Y)^C_{\mathrm{forb}}$ is invariant by Lem. 19 and the global Markov property, $\mathbb{E}_{e'}[Y \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}] = \mathbb{E}[Y \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}]$, and (11) from the proof of Lem. 2 gives $\mathbb{E}[Y \mid X_{\mathrm{DE}(Y)^C_{\mathrm{forb}}}] = f_{\mathrm{SB}(Y)}(X)$ almost surely. Substituting into (12) yields

$$R_{e'}(f) \geq R_{e'}(f_{\mathrm{SB}(Y)}).$$

It remains to show $R_{e'}(f_{\mathrm{SB}(Y)}) = R_e(f_{\mathrm{SB}(Y)})$. Since the marginal distribution of $X_{\mathrm{SB}(Y)}$ and the conditional distribution of $Y \mid X_{\mathrm{SB}(Y)}$ are the same under $\mathbb{P}_e$ and $\mathbb{P}_{e'}$ by Asm. A1 (i) and the invariance of $\mathrm{SB}(Y)$, the distribution of $(Y, X_{\mathrm{SB}(Y)})$ is the same under $\mathbb{P}_{e'}$ and $\mathbb{P}_e$. Since $\ell(Y, f_{\mathrm{SB}(Y)}(X))$ is a measurable function of $(Y, X_{\mathrm{SB}(Y)})$, the risks are equal.

Consequently, $\sup_{e'' \in \mathcal{E}} R_{e''}(f) \geq R_{e'}(f) \geq R_e(f_{\mathrm{SB}(Y)})$. Since e was arbitrary, taking the supremum over e on the right-hand side yields (7). □

E.5 Proof of Prop. 6

Proof of Prop. 6. Fix an arbitrary predictor $f\colon \mathcal{X} \to \mathcal{A}$, $e \in \mathcal{E}$, and let $(e'_j)_{j \geq 1}$ be the associated sequence from the assumption. For all j, by Lem. 18 applied under $\mathbb{P}_{e'_j}$ with $W = X$ and $Z = f(X)$,

$$R_{e'_j}(f) \geq R_{e'_j}(\mu_{e'_j}).$$

Since $\mathrm{SB}(Y)$ is invariant, the conditional distribution of $Y \mid X_{\mathrm{SB}(Y)}$ is the same in every environment. Together with condition (i), the joint distribution of $(Y, X_{\mathrm{SB}(Y)})$ is the same under $\mathbb{P}_e$ and $\mathbb{P}_{e'_j}$. Because $\ell(Y, f_{\mathrm{SB}(Y)}(X))$ is a measurable function of $(Y, X_{\mathrm{SB}(Y)})$, it follows that for all j,

$$R_{e'_j}\big(f_{\mathrm{SB}(Y)}\big) = R_e\big(f_{\mathrm{SB}(Y)}\big).$$

By condition (ii), for all $\varepsilon > 0$ there exists N such that for all $j \geq N$, $R_{e'_j}\big(f_{\mathrm{SB}(Y)}\big) - R_{e'_j}(\mu_{e'_j}) < \varepsilon$. Consequently,

$$R_{e'_j}(f) \geq R_e\big(f_{\mathrm{SB}(Y)}\big) - \Big(R_{e'_j}\big(f_{\mathrm{SB}(Y)}\big) - R_{e'_j}(\mu_{e'_j})\Big) \geq R_e\big(f_{\mathrm{SB}(Y)}\big) - \varepsilon.$$

Therefore, for all $\varepsilon > 0$, we have $\sup_{e'' \in \mathcal{E}} R_{e''}(f) \geq R_e\big(f_{\mathrm{SB}(Y)}\big) - \varepsilon$. Letting $\varepsilon \to 0$ and taking the supremum over $e \in \mathcal{E}$ (e was arbitrary) yields (7). □

E.6 Details for Ex. 7

Details for Ex. 7. In this example, $X_1 \in \mathrm{CH}(Y) \cap \mathrm{DE}(\mathrm{CH}_{\mathrm{int}}(Y)) \setminus \mathrm{CH}_{\mathrm{int}}(Y)$, so condition (6) fails. Since $\mathrm{SB}(Y) = \varnothing$ and $Y \sim \mathrm{Ber}(1/2)$, the stable blanket classifier is constant: $f_{\mathrm{SB}(Y)} = \mathbb{P}(Y = 1) = 1/2$. Consequently, for all $e \in \mathcal{E}$, $R_e(f_{\mathrm{SB}(Y)}) = \mathbb{E}_e[(Y - 1/2)^2] = 1/4$.

Define $f\colon \{0,1\}^2 \to [0,1]$ by $f(0,0) = 0$, $f(1,0) = 1$, $f(0,1) = p$, $f(1,1) = 1-p$. Fix an arbitrary $e \in \mathcal{E}$. On the event $\{X_2 = 0\}$, we have $X_1 = Y$, so $(Y - f(X_1, 0))^2 = 0$. On the event $\{X_2 = 1\}$, $X_1 = Y \oplus \varepsilon_1$ with $\varepsilon_1 \sim \mathrm{Ber}(p)$ independent of (Y, X_2, E). With probability $1-p$, $\varepsilon_1 = 0$, so $X_1 = Y$ and $(Y - f(Y,1))^2 = p^2$. With probability p, $\varepsilon_1 = 1$, so $X_1 = 1 - Y$ and $(Y - f(1-Y,1))^2 = (1-p)^2$. Hence $\mathbb{E}[(Y - f(X_1,1))^2 \mid X_2 = 1] = (1-p)\,p^2 + p\,(1-p)^2 = p(1-p)$, which does not depend on the environment. Combining both cases, for all $e \in \mathcal{E}$,

$$R_e(f) = \mathbb{E}_e\big[(Y - f(X_1, X_2))^2\big] = \mathbb{P}_e(X_2 = 1) \cdot p(1-p) \le p(1-p).$$

Since $0 < p < 1/2$, we have $p(1-p) < 1/4 = R_e(f_{\mathrm{SB}(Y)})$. □

E.7 Proof of Prop. 14

Proof of Prop. 14. Since (6) fails, there exists $i \in \mathrm{CH}(Y) \cap \mathrm{DE}(\mathrm{CH}_{\mathrm{int}}(Y)) \setminus \mathrm{CH}_{\mathrm{int}}(Y)$. Consequently, $i \notin \mathrm{CH}(E)$, and $i \notin \mathrm{SB}(Y)$ since $i \in \mathrm{DE}(Y)_{\mathrm{forb}}$.

Choose nonzero coefficients $b_{j,k}$ for every edge $k \to j$ in $\bar{\mathcal{G}}$, and let $(\varepsilon_j)_{j \in \{Y,1,\dots,d\}}$ be jointly independent standard Gaussian random variables. Define a linear Gaussian SCM with DAG $\bar{\mathcal{G}}$:

$$Y = \sum_{k \in \mathrm{PA}_{\bar{\mathcal{G}}}(Y)} b_{Y,k} X_k + \varepsilon_Y, \qquad X_j = \sum_{k \in \mathrm{PA}_{\bar{\mathcal{G}}}(j)} b_{j,k} X_k + \varepsilon_j \quad (j = 1, \dots, d). \tag{13}$$

By the faithfulness result for linear Gaussian SCMs [62, Thm. 3.2], the coefficients can be chosen so that the induced distribution $\mathbb{P}_0$ is faithful to $\bar{\mathcal{G}}$; fix such a choice.

Extend this to an SCM with DAG $\mathcal{G}$ by adding E (independent of all noises) as a source node, taking three values $\mathcal{E} = \mathrm{supp}(E) \coloneqq \{e_0, e_0', e_1'\}$, as described in § 2. Since we condition on E to obtain $\mathbb{P}_e$, its marginal distribution is not relevant. Under e_0, we define all structural assignments to coincide with the SCM (13), so that $\mathbb{P}_{e_0} = \mathbb{P}_0$ (thus, (i) holds). Under e_0', we replace the structural assignment of every $j \in \mathrm{CH}_{\mathrm{int}}(Y)$ by $X_j \coloneqq \tilde{\varepsilon}_j$, where $(\tilde{\varepsilon}_j)_{j \in \mathrm{CH}_{\mathrm{int}}(Y)}$ are mutually independent standard Gaussian noises independent of the original noises, and all other structural assignments are unchanged. Consequently, $e' \coloneqq e_0'$ satisfies the requirement in the second part of Asm. A2 both for $e = e_0$ and for $e = e_0'$ itself (in e_0', every X_j, $j \in \mathrm{CH}_{\mathrm{int}}(Y)$, is already an exogenous noise independent of all other exogenous noises, and all other assignments trivially agree with those of e_0'); thus, (ii) holds for these two environments.

Under e_1', use the assignments of e_0', but replace ε_j by an independent standard Gaussian copy for every $j \in \mathrm{CH}(E) \setminus \mathrm{CH}_{\mathrm{int}}(Y)$. All added noises are mutually independent and independent of the original noises. This makes the assignments of the remaining children of E depend on E and gives $\mathbb{P}_{e_1'} = \mathbb{P}_{e_0'}$. For $e = e_1'$, choose $e' = e_1'$, so (ii) also holds.

We now construct a predictor with lower worst-case risk than $f_{\mathrm{SB}(Y)}$. Since $i \notin \mathrm{CH}(E)$, the SCM has the assignment $X_i = \sum_{k \in \mathrm{PA}(i)} b_{i,k} X_k + \varepsilon_i$ in every environment. Define

$$\bar{X}_i \coloneqq X_i - \sum_{k \in \mathrm{PA}(i) \setminus \{Y\}} b_{i,k} X_k,$$

so that $\bar{X}_i = b_{i,Y} Y + \varepsilon_i$ in every environment. We claim that none of Y, E, or $X_{\mathrm{SB}(Y)}$ can be descendants of i in $\mathcal{G}$. Since $i \in \mathrm{CH}(Y)$, the node Y is a parent of i and hence cannot also be a descendant in a DAG. E is a source node and therefore has no ancestors. Since $i \in \mathrm{DE}(Y)_{\mathrm{forb}}$ and all descendants of nodes in $\mathrm{DE}(Y)_{\mathrm{forb}}$ are in $\mathrm{DE}(Y)_{\mathrm{forb}}$, we have that no node in $\mathrm{SB}(Y) \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$ is a descendant of i. Unrolling the structural assignments along a topological ordering of $\mathcal{G}$, each variable is a function of the noise variables of itself and its ancestors only. Consequently, since Y, E, and every node in $X_{\mathrm{SB}(Y)}$ are not descendants of i, none of them depend on ε_i:

$$\varepsilon_i \perp\!\!\!\perp (Y, X_{\mathrm{SB}(Y)}, E). \tag{14}$$

Under $\mathbb{P}_{e_0}$, the distribution of $Y \mid X_{\mathrm{SB}(Y)} = x_{\mathrm{SB}(Y)}$ is Gaussian with mean $m(x_{\mathrm{SB}(Y)})$ and constant variance σ^2. Moreover, since $\mathrm{SB}(Y) \subseteq \mathrm{DE}(Y)^C_{\mathrm{forb}}$ and all descendants of nodes in $\mathrm{DE}(Y)_{\mathrm{forb}}$ are in $\mathrm{DE}(Y)_{\mathrm{forb}}$, no node in $\mathrm{SB}(Y)$ is a descendant of $\mathrm{CH}_{\mathrm{int}}(Y)$. As only the assignments of $\mathrm{CH}_{\mathrm{int}}(Y)$ differ between e_0 and e_0', the marginal distribution of $X_{\mathrm{SB}(Y)}$ is the same under $\mathbb{P}_{e_0}$ and

$\mathbb{P}_{e'_0}$, and hence also under $\mathbb{P}_{e'_1}$. Since $\mathrm{SB}(Y)$ is invariant, it holds in every environment that, almost surely,

$$Y \mid X_{\mathrm{SB}(Y)} = x_{\mathrm{SB}(Y)} \sim \mathcal{N}\big(m(x_{\mathrm{SB}(Y)}), \sigma^2\big).$$

Since the SCM on $\bar{\mathcal{G}}$ has independent Gaussian noises with positive variances, the covariance matrix of $(Y, X_{\mathrm{SB}(Y)})$ under e_0 is positive definite. In particular $\sigma^2 > 0$, since $\sigma^2 = 0$ would imply that Y is an affine function of $X_{\mathrm{SB}(Y)}$ almost surely, contradicting the positive definiteness.

The conditional distribution of $(Y, \bar{X}_i)$ given $X_{\mathrm{SB}(Y)}$ is determined by $Y \mid X_{\mathrm{SB}(Y)} \sim \mathcal{N}(m, \sigma^2)$ and $\bar{X}_i \mid Y \sim \mathcal{N}(b_{i,Y} Y, 1)$, both environment-independent by invariance of $\mathrm{SB}(Y)$ and (14), respectively. (Indeed, by (14), $\bar{X}_i = b_{i,Y} Y + \varepsilon_i$ is conditionally independent of $(X_{\mathrm{SB}(Y)}, E)$ given Y, so the two stated conditionals determine the conditional distribution of $(Y, \bar{X}_i)$ given $X_{\mathrm{SB}(Y)}$ in every environment.) Hence $Y \mid (X_{\mathrm{SB}(Y)}, \bar{X}_i)$ does not depend on e (because of joint Gaussianity), and the following predictor is well-defined independently of the environment:

$$f(x) \coloneqq \mathbb{E}[Y \mid X_{\mathrm{SB}(Y)} = x_{\mathrm{SB}(Y)}, \bar{X}_i = \bar{x}_i].$$

Since $(Y, \bar{X}_i)$ is jointly Gaussian conditionally on $X_{\mathrm{SB}(Y)}$,

$$f(X) = \mathbb{E}[Y \mid X_{\mathrm{SB}(Y)}, \bar{X}_i] = f_{\mathrm{SB}(Y)}(X) + \frac{b_{i,Y}\sigma^2}{b_{i,Y}^2 \sigma^2 + 1}\big(\bar{X}_i - \mathbb{E}[\bar{X}_i \mid X_{\mathrm{SB}(Y)}]\big),$$

where $b_{i,Y}\sigma^2/(b_{i,Y}^2\sigma^2 + 1)$ is nonzero. Using (14), the residual $\bar{X}_i - \mathbb{E}[\bar{X}_i \mid X_{\mathrm{SB}(Y)}]$ has variance at least $\mathrm{Var}(\varepsilon_i) = 1$. Hence, $f(X) \neq f_{\mathrm{SB}(Y)}(X)$ with positive probability under every $\mathbb{P}_e$. Applying Lem. 18 under $\mathbb{P}_e$ with $W = (X_{\mathrm{SB}(Y)}, \bar{X}_i)$ and $Z = f_{\mathrm{SB}(Y)}(X)$ yields $R_e(f_{\mathrm{SB}(Y)}) > R_e(f)$ for all $e \in \mathcal{E}$ (thus, (iii) holds). □

E.8 Proof of Prop. 15

Proof of Prop. 15. Write $x = (x_S, x_N)$ with $x_S \in \mathcal{X}_S$ and $x_N \in \mathcal{X}_N$. Let $f\colon \mathcal{X} \to \mathcal{A}$ be an arbitrary measurable predictor and $\mathbb{Q} \in \mathcal{Q}_S$. We construct a distribution $\tilde{\mathbb{Q}} \in \mathcal{Q}_S$ on $\mathcal{X} \times \mathcal{Y}$ by defining it as the product measure

$$\tilde{\mathbb{Q}} \coloneqq \mathbb{Q}_{(X_S, Y)} \otimes \mathbb{Q}_{X_N},$$

where $\mathbb{Q}_{(X_S,Y)}$ and $\mathbb{Q}_{X_N}$ denote the marginal distributions of (X_S, Y) and X_N under $\mathbb{Q}$, respectively. We verify $\tilde{\mathbb{Q}} \in \mathcal{Q}_S$:

(i) The marginal distribution of (X_S, Y) under $\tilde{\mathbb{Q}}$ equals $\mathbb{Q}_{(X_S,Y)}$, so $\tilde{\mathbb{Q}}(Y \in \cdot \mid X_S) = \mathbb{Q}(Y \in \cdot \mid X_S) = \mathbb{P}(Y \in \cdot \mid X_S)$ almost surely.

(ii) Since $\mathbb{Q}_X \ll \mu$ and $\mu = \mu_S \otimes \mu_N$, we have $\mathbb{Q}_{X_S} \ll \mu_S$ and $\mathbb{Q}_{X_N} \ll \mu_N$, hence $\tilde{\mathbb{Q}}_X = \mathbb{Q}_{X_S} \otimes \mathbb{Q}_{X_N} \ll \mu$.

By Tonelli's theorem,

$$\mathbb{E}_{\tilde{\mathbb{Q}}}[\ell(Y, f(X))] = \int_{\mathcal{X}_N} \left(\int_{\mathcal{X}_S \times \mathcal{Y}} \ell(y, f(x_S, x_N))\, d\mathbb{Q}_{(X_S,Y)}(x_S, y) \right) d\mathbb{Q}_{X_N}(x_N).$$

For each fixed $x_N \in \mathcal{X}_N$, define $g_{x_N} : \mathcal{X}_S \to \mathcal{A}$ by $g_{x_N}(z) \coloneqq f(z, x_N)$. The inner integral can be written as

$$\int_{\mathcal{X}_S \times \mathcal{Y}} \ell(y, f(x_S, x_N))\, d\mathbb{Q}_{(X_S,Y)}(x_S, y) = \mathbb{E}_{(X_S,Y)\sim\mathbb{Q}}\Big[\ell\big(Y, g_{x_N}(X_S)\big)\Big].$$

For every fixed x_N, $g_{x_N}(X_S)$ is $\sigma(X_S)$-measurable. Applying Lem. 18 under $\mathbb{Q}$ with $W = X_S$ and $Z = g_{x_N}(X_S)$, and using condition (i),

$$\mathbb{E}_{(X_S,Y)\sim\mathbb{Q}}\Big[\ell\big(Y,\, g_{x_N}(X_S)\big)\Big] \geq \mathbb{E}_{(X_S,Y)\sim\mathbb{Q}}\Big[\ell\big(Y,\, f_S(X_S)\big)\Big].$$

The right-hand side does not depend on x_N. Integrating over x_N yields

$$\mathbb{E}_{\tilde{\mathbb{Q}}}\big[\ell(Y, f(X))\big] \geq \mathbb{E}_{\mathbb{Q}}\big[\ell(Y, f_S(X_S))\big].$$

Since $\tilde{\mathbb{Q}} \in \mathcal{Q}_S$,

$$\sup_{\mathbb{Q}' \in \mathcal{Q}_S} \mathbb{E}_{\mathbb{Q}'}\big[\ell(Y, f(X))\big] \geq \mathbb{E}_{\tilde{\mathbb{Q}}}\big[\ell(Y, f(X))\big] \geq \mathbb{E}_{\mathbb{Q}}\big[\ell(Y, f_S(X_S))\big].$$

As $\mathbb{Q} \in \mathcal{Q}_S$ was arbitrary, the right-hand side can be replaced by its supremum over $\mathcal{Q}_S$. Since f was arbitrary, f_S minimizes the worst-case risk. □

E.9 Proof of Prop. 16

Proof of Prop. 16. Let $D(O) \coloneqq O \cup \mathrm{DE}(O)$, so that $N(O) = \{1, \ldots, d\} \setminus D(O)$.

Step 1: m-separation in $\mathcal{M}_O$ implies m-separation in $\mathcal{M}$. No node in $D(O)$ is Y or an ancestor of Y, since otherwise some node in O would have a directed path to Y. Since E is a source node, $E \notin D(O)$. Hence $D(O) \subseteq \{1, \ldots, d\}$, and its complement $N(O) \cup \{Y, E\}$ in $\mathcal{M}$ is closed under taking ancestors because $D(O)$ is closed under taking descendants. Now let $x, y \in N(O) \cup \{Y, E\}$, let $Z \subseteq N(O) \cup \{Y, E\}$, and let π be a path between x and y in $\mathcal{M}$ that is not m-blocked by Z; we show that π lies in $\mathcal{M}_O$ and is not m-blocked by Z in $\mathcal{M}_O$ either. Every node on π is an endpoint, a node in Z, or an ancestor of one of these: a collider on π is in Z or has a descendant in Z, and a non-collider has an edge pointing away from it on π, which starts a directed subpath of π that ends at a collider or an endpoint [see also 49, proof of Lem. 3]. Hence π and every directed path from a collider on π to Z lie in $N(O) \cup \{Y, E\}$ and thus in $\mathcal{M}_O$, so π is not m-blocked by Z in $\mathcal{M}_O$ either.

Step 2: $(E, N(O) \setminus S(O)) \perp_m Y \mid S(O)$ in $\mathcal{M}_O$, that is, every path in $\mathcal{M}_O$ between Y and a node in $\{E\} \cup (N(O) \setminus S(O))$ is m-blocked by $S(O)$. Let F denote the Markov blanket $\mathrm{DIS}(Y) \cup \mathrm{PA}(\mathrm{DIS}(Y)) \cup \mathrm{DIS}(\mathrm{CH}(Y)) \cup \mathrm{PA}(\mathrm{DIS}(\mathrm{CH}(Y)))$ computed in $\mathcal{M}_O$, so that $S(O) = F \setminus \{Y, E\} \subseteq N(O)$. We use two observations about paths in $\mathcal{M}_O$ whose intermediate nodes are all colliders.

- If such a path connects a node $w \neq Y$ to Y, then $w \in F$. If w is adjacent to Y, it is a parent of Y, a child of Y, or joined to Y by a bidirected edge, and all such nodes lie in F. Otherwise, the intermediate nodes are joined by bidirected edges and form a chain in $\mathrm{DIS}(Y)$ or in $\mathrm{DIS}(\mathrm{CH}(Y))$, depending on whether the edge at Y is bidirected or points away from Y. The edge at w has an arrowhead at the first intermediate node, so w is a member or a parent of that district.
- No such path connects E to Y. Since all edges at E point away from E and E is not adjacent to Y, such a path would have one of the forms in (10) in $\mathcal{M}_O$ and hence in $\mathcal{M}$, so it would contain a node of O by the definition of an admissible collider set, but $\mathcal{M}_O$ contains no node of O.

Now let $v \in \{E\} \cup (N(O) \setminus S(O))$ and let π be a path between v and Y in $\mathcal{M}_O$; we show that π is m-blocked by $S(O)$. Let w be the intermediate non-collider on π closest to Y, or $w \coloneqq v$ if there is none. The nodes strictly between w and Y on π are colliders and $w \neq Y$, so the two observations give $w \in F$ and $w \neq E$, and thus $w \in F \setminus \{Y, E\} = S(O)$. Since $v \notin S(O)$, we have $w \neq v$, so w is a non-collider on π in $S(O)$ and π is m-blocked by $S(O)$.

Step 3: proof of (i). Step 2 and Step 1 give $(E, N(O) \setminus S(O)) \perp_m Y \mid S(O)$ in $\mathcal{M}$, since all nodes involved lie in $N(O) \cup \{Y, E\}$. The global Markov property yields $Y \perp\!\!\!\perp (E, X_{N(O) \setminus S(O)}) \mid X_{S(O)}$ under $\mathbb{P}$, and the decomposition and weak union properties [28, Thm. 8.12] give

$$Y \perp\!\!\!\perp E \mid X_{S(O)}, \qquad Y \perp\!\!\!\perp X_{N(O) \setminus S(O)} \mid X_{S(O)}, \qquad Y \perp\!\!\!\perp E \mid X_{N(O)}, \tag{15}$$

so $S(O)$ and $N(O)$ are invariant. Moreover, $S(O) \subseteq N(O)$ by definition, and $\mathrm{PA}(Y) \subseteq N(O)$ because a parent of Y in $D(O)$ would be an ancestor of Y in $D(O)$ (which Step 1 excludes). Hence $\mathrm{PA}(Y)$ is also the set of parents of Y in $\mathcal{M}_O$, where $\mathrm{PA}(Y) \subseteq \mathrm{PA}(\mathrm{DIS}(Y))$, and since $E \notin \mathrm{PA}(Y)$, we have $\mathrm{PA}(Y) \subseteq S(O)$.

Step 4: proof of (ii). By the invariance of $N(O)$, the second independence in (15) together with Doob's conditional independence property [28, Thm. 8.9], and the definition of $f_{S(O)}$, we have for all $e \in \mathcal{E}$, $\mathbb{P}_e$-almost surely,

$$\mathbb{E}_e[Y \mid X_{N(O)}] = \mathbb{E}[Y \mid X_{N(O)}] = \mathbb{E}[Y \mid X_{S(O)}] = f_{S(O)}(X).$$

For $e \in \mathcal{E}$ and $\sigma(X_{N(O)})$-measurable $f\colon \mathcal{X} \to \mathcal{A}$, applying Lem. 18 under $\mathbb{P}_e$ with $W = X_{N(O)}$ and $Z = f(X)$ yields

$$R_e(f) \geq \mathbb{E}_e\big[\ell(Y, \mathbb{E}_e[Y \mid X_{N(O)}])\big] = R_e(f_{S(O)}),$$

where the equality uses the display above; see the proof of Lem. 2 for the same argument.

Step 5: proof of (iii). Assume that $\mathrm{PA}(Y)$ is invariant. By (i), $f_{\mathrm{PA}(Y)}$ is $\sigma(X_{N(O)})$-measurable, so (ii) gives $R_e(f_{\mathrm{PA}(Y)}) \geq R_e(f_{S(O)})$ for all $e \in \mathcal{E}$; taking suprema proves Thm. 3 for the follower objective (i). For the follower objective (ii), the tower property and the invariance of $S(O)$ and $\mathrm{PA}(Y)$ give $\mathbb{E}_e[f_{S(O)}(X)] = \mathbb{E}_e[Y] = \mathbb{E}_e[f_{\mathrm{PA}(Y)}(X)]$ for all $e \in \mathcal{E}$, and the claim follows exactly as in the proof of Thm. 3. □

F Additional details for the synthetic experiments

We now provide more details on the synthetic experiments described in § 6.1. We describe the data-generating process (§ F.1), how the leader and the follower optimize their objectives (§ F.2), an additional analysis showing that predictors that are invariant in population can be unstable under the follower's intervention at finite sample sizes, with this instability shrinking as the leader receives more training data (§ F.3), implementation details (§ F.4), and a comparison of the population follower of § 6.1 with individual followers as in § 2.1 and § C.1 (§ F.5).

F.1 Data-generating processes

All synthetic experiments described in § 6.1 use the training graph $\mathcal{G}_0$ in Fig. 1. The parent set and stable blanket are

$$\mathrm{PA}(Y) = \{X_1, X_2\}, \qquad \mathrm{SB}(Y) = \{X_1, X_2, X_3\}.$$

The all-variable predictor uses $\{X_1, X_2, X_3, X_4, X_5\}$. The follower can intervene on X_1 and X_4 by adding perturbations E_{δ_1} and E_{δ_4}, which are functions of some of the random variables (detailed below) satisfying $\max(|E_{\delta_1}|, |E_{\delta_4}|) \leq b$, where b is the intervention bound.

For the nonlinear SCM, we draw independent standard Gaussian noises and generate

$$\begin{aligned}
X_2 &= \varepsilon_2, \\
X_1 &= 0.8\tanh(X_2) + 0.6X_2 + 0.3\varepsilon_1 + E_{\delta_1}(X_2, \varepsilon_1), \\
Y &= \sin(X_1) + 0.8X_2 + 0.5X_1X_2 + 0.4\varepsilon_Y, \\
X_3 &= \tanh(1.2Y) + 0.45Y^2 + 0.5\varepsilon_3, \\
X_4 &= 2.1\tanh(Y) + 0.9Y + 0.55X_2 + 0.15YX_2 + 0.2\varepsilon_4 + E_{\delta_4}(Y, X_2, \varepsilon_4), \\
X_5 &= 1.5\tanh(X_4) + 0.75X_4 + 0.35X_4^2 + 0.2\varepsilon_5.
\end{aligned}$$

For the linear-Gaussian SCM, we use

$$\begin{aligned}
X_2 &= \varepsilon_2, \\
X_1 &= 1.1X_2 + 0.3\varepsilon_1 + E_{\delta_1}(X_2, \varepsilon_1), \\
Y &= X_1 + 0.8X_2 + 0.4\varepsilon_Y, \\
X_3 &= 1.2Y + 0.5\varepsilon_3, \\
X_4 &= 1.4Y + 0.55X_2 + 0.2\varepsilon_4 + E_{\delta_4}(Y, X_2, \varepsilon_4), \\
X_5 &= 1.3X_4 + 0.2\varepsilon_5.
\end{aligned}$$

In the additional experiment of § 6.1 with the dashed edges of Fig. 1, the X_4 perturbation is allowed to depend on $(Y, X_2, \varepsilon_4, X_1, X_3)$ instead of only (Y, X_2, ε_4).

F.2 Leader's and follower's optimization

For each subset $S \in \{\mathrm{PA}(Y), \mathrm{SB}(Y), \{1, \ldots, 5\}\}$, the leader trains a predictor f_S. In the linear-Gaussian SCM, we use the closed-form population linear regression function $\mathbb{E}[Y \mid X_S]$. In the nonlinear SCM, we fit an MLP regressor on the leader's training data and evaluate it after the follower's intervention.

The follower's intervention is parameterized by neural networks for E_{δ_1} and E_{δ_4}, with a final $\tanh$ layer enforcing the intervention bound. We consider two follower objectives:

$$\text{MSE:} \quad \max_{E_{\delta_1}, E_{\delta_4}} \mathbb{E}\big[(Y - f_S(X_S))^2\big],$$
$$\text{prediction:} \quad \min_{E_{\delta_1}, E_{\delta_4}} \mathbb{E}\big[f_S(X_S)\big].$$

After optimizing the follower objective, we always report the leader's deployment MSE

$$\mathbb{E}\big[(Y - f_S(X_S))^2\big]$$

under the induced intervention distribution.

F.3 Finite-sample regression error

The theory (for example, Thm. 3 and Thm. 5) considers population predictors such as $f_S(x_S) = \mathbb{E}[Y \mid X_S = x_S]$. For an invariant set S, this conditional mean is the same in all environments. In the finite-sample nonlinear experiment (see § F.1), however, the leader deploys an estimate

$$\hat{f}_S = f_S + \Delta_S,$$

where Δ_S is the regression error. Under squared loss, the risk decomposes into population risk and a regression-error term:

$$\begin{aligned} R_e(\hat{f}_S) &= \mathbb{E}_e\big[(Y - \hat{f}_S(X_S))^2\big] \\ &= R_e(f_S) + \mathbb{E}_e\big[\Delta_S(X_S)^2\big], \end{aligned}$$

where the cross term vanishes because of $f_S = \mathbb{E}_e[Y \mid X_S]$ and the law of iterated expectations. The population risk $R_e(f_S)$ can itself vary across environments: In both SCMs presented in § F.1, $Y = m(X_1, X_2) + 0.4\varepsilon_Y$ for a function m, with independent standard Gaussian noise unaffected by interventions, so $R_e(f_{\mathrm{PA}(Y)}) = 0.16$ for every e; but in the nonlinear SCM, X_3 is more informative about Y at some values of $\mathrm{SB}(Y) \setminus \{X_3\}$ than others, so interventions can change the average conditional variance $R_e(f_{\mathrm{SB}(Y)}) = \mathbb{E}_e[\mathrm{Var}(Y \mid X_{\mathrm{SB}(Y)})]$. The regression-error term can also change across interventions because the follower changes the marginal distribution of X_S. In particular, the follower can move probability mass toward regions where $|\Delta_S|$ is large. Larger training samples can reduce regression error and the resulting sensitivity to interventions. For Fig. 2, we computed both terms of the decomposition using the exact conditional means, available in closed form for $\mathrm{PA}(Y)$ and by numerical integration over ε_Y for $\mathrm{SB}(Y)$. The increase of the deployment MSE from $b = 0.25$ to $b = 1$ is entirely regression error for $\hat{f}_{\mathrm{PA}(Y)}$, since $R_e(f_{\mathrm{PA}(Y)})$ is constant. For $\hat{f}_{\mathrm{SB}(Y)}$ and under both follower objectives, about 80% of it is due to the change of $\mathbb{E}_e[\mathrm{Var}(Y \mid X_{\mathrm{SB}(Y)})]$ (the follower lowers X_1 for most observations, which moves Y to values at which X_3 is less informative about Y).

Regression error alone does not explain the instability of the all-variable predictor. The all-variable set contains X_4 and X_5, which lie in the descendant region of the intervened child of Y. Consequently, the relevant conditional relation may change under the follower's intervention, and therefore an increase in the leader's training sample size does not necessarily make the all-variable predictor more stable: Fig. 6 confirms this behavior. In the train-size sweep, the deployment MSE of the parent and stable-blanket predictors becomes flatter as the leader receives more data, whereas the all-variable predictor remains sensitive to the intervention bound. The same figure also shows that allowing the X_4 perturbation to depend additionally on X_1 and X_3 leaves the qualitative comparison unchanged.

F.4 Implementation details of the leader's predictor

In the linear-Gaussian SCM, the conditional expectation $\mathbb{E}[Y \mid X_S]$ is computed in closed form from the SCM's joint covariance. In the nonlinear SCM, $\hat{f}_S$ is an MLP (5 hidden layers of width 256, ReLU) trained by Adam [30] (learning rate 10^{-3}, weight decay 10^{-5}, batch size 512) on $50{,}000$ training observations, with early stopping determined on a $3{,}000$-observation validation set (patience 20, up to 500 epochs).

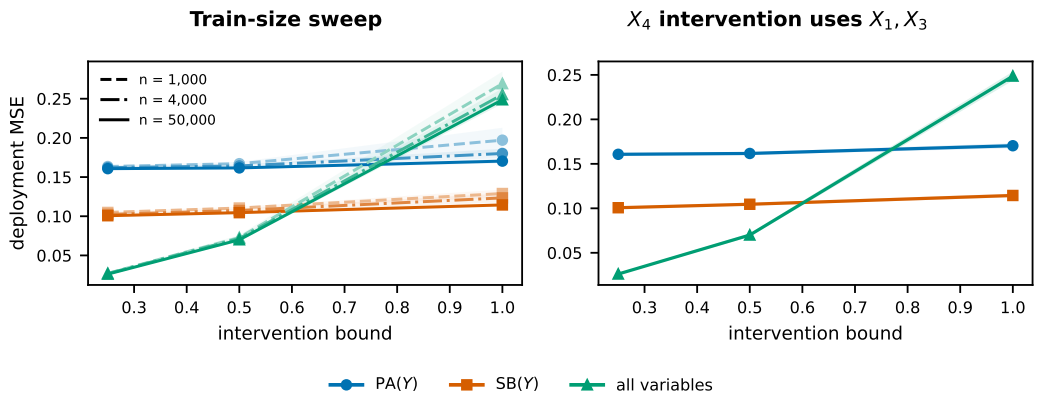


Figure 6: Additional nonlinear SCM results with a follower minimizing the expected prediction. Left: train-size sweep. Increasing the leader's sample size ($n \in \{1{,}000, 4{,}000, 50{,}000\}$) reduces the residual instability of $\mathrm{PA}(Y)$ and $\mathrm{SB}(Y)$, consistent with a decrease in finite-sample regression error. Right: allowing the X_4 intervention to additionally use X_1 and X_3 does not qualitatively change the results.

Each follower perturbation E_{δ_j} ($j \in \{1, 4\}$) is an MLP (2 hidden layers of width 64, ReLU) followed by $b \cdot \tanh(\cdot)$ to enforce $|E_{\delta_j}| \leq b$; its inputs are as specified in § F.1. With $\hat{f}_S$ frozen, E_{δ_1} and E_{δ_4} are optimized jointly by Adam (learning rate 10^{-3}) for 2,000 steps. Each step computes the follower's objective on a fresh batch of 512 observations drawn from the SCM with the current $E_{\delta_1}, E_{\delta_4}$ and backpropagates through the SCM forward pass. We use 5 random restarts and pick the parameters with the best objective value on 10,000 held-out observations.

Each curve averages 10 independent runs with shaded 95% confidence intervals; the reported deployment MSE is computed on 10,000 fresh observations under the selected $E_{\delta_1}, E_{\delta_4}$. The experiments are run on an Apple M4 Pro CPU (14 cores) using 10 parallel workers.

F.5 Comparison with individual followers

Fig. 7 compares the population follower of § 6.1 with individual followers who each choose the shift intervention given their own realization of the observed variables as in § C.1. Fig. 7 shows that the results of Fig. 2 remain almost exactly the same when the population follower is replaced by the individual followers, which empirically supports the argument of § C.1. Implementation details are given in the `README` of the code repository.[14]

G Additional details for the Causal Chambers experiments

We now provide more details on the experiment described in § 6.2. We describe the physical device and its variables (§ G.1), the data-generating process (§ G.2), the hidden-confounding hypothesis (§ G.3), implementation details for predictors deployed by the leader (§ G.4), implementation details for the follower (§ G.5), and evaluation details (§ G.6).

G.1 Causal Chambers: Physical device and variables

We collect data from the Causal Chambers by Gamella et al. [18] using the `causalchamber` API.[15] We use the light tunnel `lt_mk2_standard` containing an RGB light source (intensities `red`, `green`, `blue`), two linear polarizers at adjustable angles (`pol_1`, `pol_2`), and three pairs of infrared and visible-light sensors (`ir_j`, `vis_j` for `j` $\in \{1, 2, 3\}$) at different positions along the tunnel. Sensors at positions `j` $= 1, 2$ are unaffected by the polarizers, while sensors at position `j` $= 3$ are af-

[14] `https://github.com/LinusKuehne/prediction-intervention-games`
[15] `https://github.com/juangamella/causal-chamber-package`

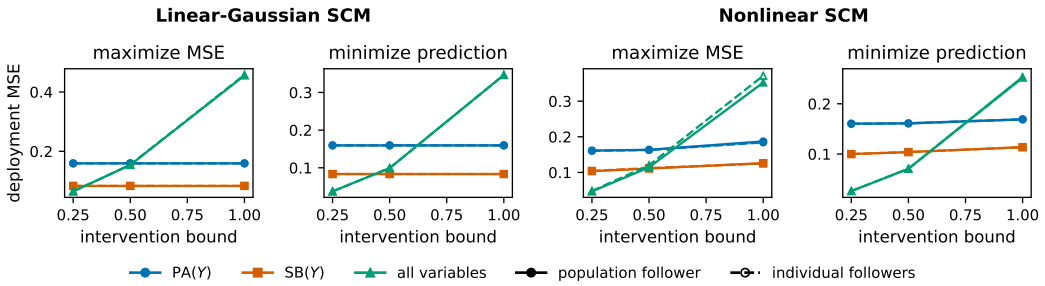


Figure 7: Population follower of § 6.1 (solid lines, filled markers) and individual followers of § C.1 (dashed lines, open markers). For each predictor, both followers are shown; the solid and dashed curves lie almost on top of each other. Left two panels: linear-Gaussian SCM with population regression predictors. Right two panels: nonlinear SCM with predictors estimated from finite data. In each pair, the follower maximizes the leader's MSE (left) or minimizes the expected prediction (right). Curves average 10 runs, with 95% confidence intervals for the mean.

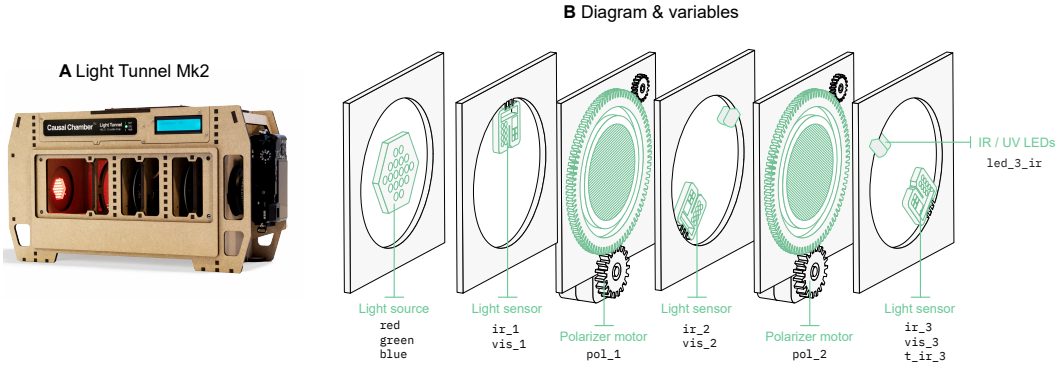


Figure 8: Overview of the Causal Chamber light tunnel (`lt_mk2_standard`). (A) Picture of the hardware. (B) Diagram of the hardware components. The displayed variables are relevant for our experiment. We use different subsets of them for classification tasks. Some are not used as covariates but as intervention targets to generate the different environments. Figure adapted from Gamella et al. [18], licensed under CC BY 4.0.

fected by the polarizers through Malus' law.[16] An infrared LED placed above the sensors at position $\texttt{j} = 3$ (`led_3_ir`) activates only during the measurement of `ir_3` and `vis_3`, and does not affect the sensors at positions $\texttt{j} = 1, 2$. Fig. 8 shows a schematic of the chamber. Tab. 1 lists all variables relevant to our experiments. We refer to Gamella et al. [18] for the complete documentation.

G.2 Data-generating process

The covariate vector is $X = (\texttt{red}, \texttt{green}, \texttt{blue}, \texttt{ir_2}, \texttt{vis_2}, \texttt{ir_3}, \texttt{vis_3})$. The light intensities are drawn independently from normal distributions truncated to $[0, 255] \subseteq \mathbb{R}$, rounded to integers:

$$\texttt{red} \sim \mathcal{N}_{[0,255]}(64, 20^2), \quad \texttt{green} \sim \mathcal{N}_{[0,255]}(32, 30^2), \quad \texttt{blue} \sim \mathcal{N}_{[0,255]}(90, 12^2).$$

We define $\texttt{RGB} := (\texttt{red}, \texttt{green}, \texttt{blue}) \in \mathbb{R}^3$. Data are generated in two stages.

1. Target generation: The `RGB` values are sampled. The chamber measures the infrared reading `ir_1`, and the target is defined as $Y := \mathbf{1}\{\texttt{ir_1} > \tau\}$, where $\tau = 12{,}500$.
2. Downstream interventions: Given the value of Y, we intervene on `ir_3` and `vis_3` by setting

$$\texttt{led_3_ir} := \beta_0^{\text{led}} + Y \cdot \beta_1^{\text{led}}, \qquad \texttt{pol_2} := \beta_0^{\text{pol}} + Y \cdot \beta_1^{\text{pol}},$$

[16] The intensity of light passing through both polarizers is proportional to $\cos^2(\texttt{pol_1} - \texttt{pol_2})$.

Table 1: Causal Chamber light tunnel variables relevant for our experiments.

Variable	Range	Description
`red`, `green`, `blue`	$\{0, \dots, 255\}$	Brightness settings for the main light source LEDs.
`ir_j` (`j` $\in \{1, 2, 3\}$)	$[0, 65535]$	Uncalibrated infrared intensity measured by sensor at pos. `j`.
`vis_j` (`j` $\in \{1, 2, 3\}$)	$[0, 65535]$	Uncalibrated visible-light intensity measured by sensor at pos. `j`.
`pol_1`, `pol_2`	$[-270, 270]$	Rotation angles for the first and second polarizer (degrees).
`led_3_ir`	$\{0, \dots, 4095\}$	Brightness of the infrared LED placed above sensor at pos. `j` $= 3$.

Table 2: Interventions in different environments. The coefficients $(\beta_0^{\text{led}}, \beta_1^{\text{led}}, \beta_0^{\text{pol}}, \beta_1^{\text{pol}})$ control the environment-dependent effect of Y on `led_3_ir` and `pol_2`.

E	β_0^{led}	β_1^{led}	β_0^{pol}	β_1^{pol}
0	0	0	0	0
1	0	20	0	50
2	0	8	0	20

rounded to integers in $\{0, \dots, 4095\}$ and $\{0, \dots, 270\}$, respectively. The coefficients $(\beta_0^{\text{led}}, \beta_1^{\text{led}}, \beta_0^{\text{pol}}, \beta_1^{\text{pol}})$ vary across environments (Tab. 2), parameterizing the exogenous variable E. The chamber then takes a second round of measurements with these updated settings (the exact values of `RGB` from the first stage are kept). Since `led_3_ir` and `pol_2` influence `ir_3` and `vis_3`, this creates an environment-dependent association between Y and the downstream sensor readings at `j` $= 3$, see Fig. 3.

Tab. 2 lists the interventions corresponding to the environments. In all three training environments, $\beta_0^{\text{led}} = \beta_0^{\text{pol}} = 0$ and $\beta_1^{\text{led}}, \beta_1^{\text{pol}} \geq 0$, so $Y = 1$ leads to higher (or equal) values of `led_3_ir` and `pol_2` than $Y = 0$. The effects of `led_3_ir` (increasing `ir_3`) and `pol_2` (decreasing `ir_3` and `vis_3`) partially cancel, resulting in a positive correlation between Y and `ir_3`, making `ir_3` a highly predictive training feature. A choice of $\beta_0^{\text{led}}, \beta_0^{\text{pol}} > 0$ and $\beta_1^{\text{led}}, \beta_1^{\text{pol}} < 0$ by a follower could reverse the relationship: $Y = 1$ would yield lower values of `led_3_ir` and `pol_2`. A classifier making predictions based on `ir_3` would perform badly in such an environment. The follower's intervention class here is more restricted than in the framework of § 2. This makes the experiment a realistic test of our framework, since in practice followers are typically limited to a particular set of actions rather than able to intervene arbitrarily.

G.3 Hidden confounding

Gamella et al. [18] provide partial causal-graph information for the light tunnel, where an edge $Z_1 \to Z_2$ means that intervening on Z_1 changes the distribution of subsequent measurements of Z_2. They note that the absence of an edge in their graphs does not preclude a causal effect or confounding by unmeasured variables. Under the DAG in Fig. 3 (middle), which we inferred from their provided information, we would expect conditional independencies of the form `ir_i` $\perp\!\!\!\perp$ `ir_j` $\mid$ `RGB` and `vis_i` $\perp\!\!\!\perp$ `vis_j` $\mid$ `RGB` to hold by the global Markov property. Tab. 3 reports p-values for the corresponding null hypotheses in the training environments described in § G.2, computed with the weighted GCM test [57]. The conditional independencies are rejected in the majority of cases. We hypothesize that this is because of hidden confounding (such as background lighting) and believe the graph in Fig. 3 (right) is a more accurate description. Indeed, this is in correspondence with the results presented in § 6.2.

G.4 Implementation details (leader)

The leader trains five predictors on data pooled from the three training environments $E \in \{0, 1, 2\}$ (Tab. 2), with 10,000 observations per environment. Four predictors are random forests (`scikit-learn` [45] defaults, 100 trees) restricted to the covariate subsets indicated by their subscript: $\hat{f}_{\texttt{RGB}}$, $\hat{f}_{\texttt{RGB,ir_2}}$, $\hat{f}_{\texttt{RGB,ir_2,vis_2}}$, and $\hat{f}_{\text{all}}$. The fifth predictor is the stabilized-classification ensemble $\hat{f}^{\text{SC}}$ (§ 5). All five predictors are held fixed after training.

Table 3: p-values of a conditional independence test (Weighted GCM test [57]) for the null hypothesis $Z_1 \perp\!\!\!\perp Z_2 \mid \texttt{RGB}$. Conditional independence is rejected in the majority of cases, and in particular when pooling the data from all training environments.

		p-value			
Z_1	Z_2	$E=0$	$E=1$	$E=2$	$E \in \{0,1,2\}$
`ir_1`	`ir_2`	< 0.001	< 0.001	< 0.001	< 0.001
`ir_2`	`ir_3`	< 0.001	0.688	< 0.001	< 0.001
`vis_1`	`vis_2`	< 0.001	< 0.001	< 0.001	< 0.001
`vis_2`	`vis_3`	< 0.001	0.081	< 0.001	< 0.001

For SC, we use $\alpha_{\text{inv}} = \alpha_{\text{pred}} = 0.05$ without tuning, enumerate all 2^d ($d = 7$) feature subsets, and apply no screening. Predictiveness is scored by out-of-sample binary cross-entropy on the pooled training environments, using out-of-bag predictions; the predictive cutoff is set by 250 bootstrap samples. Invariance is tested with the TRAM-GCM test [33], using a Python re-implementation of the non-parametric `rangerICP` variant ported from the `tramicp` package [32]. The base classifiers inside SC use 100 trees, with `scikit-learn` defaults. The invariance test uses random forests with 500 trees, a minimum leaf size of 5, and $\lfloor \sqrt{|S|} \rfloor$ candidate features per split for nonempty subsets S.

All computations run in minutes on an Apple M4 Pro CPU (14 cores) using 8 parallel workers; the dominant cost of the experiment is data collection from the chamber via the API.

G.5 Implementation details (follower)

For a given deployed predictor $\hat{f}$, the follower picks one environment from a fixed candidate set $\mathcal{E}$ of 49 environments. We denote this choice by $e^*(\hat{f})$ to emphasize its dependence on $\hat{f}$; different predictors can lead to different choices. Each environment is specified by a coefficient tuple $(\beta_0^{\text{led}}, \beta_1^{\text{led}}, \beta_0^{\text{pol}}, \beta_1^{\text{pol}})$, as in § G.2. The reference environment e_{ref} has coefficients $(0, 12, 0, 30)$; the magnitudes of its slope coefficients lie between those of the training environments $E = 1$ and $E = 2$ (Tab. 2). The 49 environments consist of the environment with coefficients $(0, 0, 0, 0)$, in which both interventions are switched off, and a 6×8 grid combining six intensity scales with eight non-trivial sign patterns for the (LED, polarizer) pair, to which e_{ref} belongs: each component takes one of three patterns: switched off, the training pattern (active for $Y = 1$), or reversed (active for $Y = 0$). Tab. 4 lists all 49 tuples.

We order the 49 environments by their distance $d := \max\{|\beta_1^{\text{led}} - 12|/25,\ |\beta_1^{\text{pol}} - 30|/60\}$ to e_{ref}; ties in d are ordered as in the released code. Across the 49 environments, the intercepts β_0^{led} and β_0^{pol} are functions of the slopes β_1^{led} and β_1^{pol}, so we compute d from the slopes only. For an intervention bound $b \in [0, 1]$, the choice set $\mathcal{E}_b \subseteq \mathcal{E}$ consists of the $\max\{1, \text{round}(49b)\}$ environments closest to e_{ref} (for $b \in \{0, 0.05, 0.1, 0.25, 0.5\}$, these are 1, 2, 5, 12, and 24 environments); in particular, we set $\mathcal{E}_{b=0} := \{e_{\text{ref}}\}$.

For each candidate environment $e \in \mathcal{E}$, we collect 1,000 observations of X from the chamber under e before the follower's selection. Across the 49 candidate environments, we reuse the same 1,000 sampled `RGB` values and their corresponding target labels Y. We randomly split the shared row indices into 500 selection and 500 evaluation observations and reuse this split across predictors and intervention bounds. The follower estimates the population mean prediction $\mathbb{E}_e[\hat{f}(X)]$ from the 500 observations reserved for environment selection and selects the environment $e^*(\hat{f}) \in \mathcal{E}_b$ minimizing this estimate.

G.6 Evaluation

After the follower's selection, we evaluate on the held-out 500 observations of X from $e^*(\hat{f})$. For each predictor $\hat{f}$ and each value of $b \in \{0, 0.05, 0.1, 0.25, 0.5\}$, we report on this evaluation set two quantities: (i) the shift in the population mean prediction relative to the reference, $\mathbb{E}_{e^*(\hat{f})}[\hat{f}(X)] - \mathbb{E}_{e_{\text{ref}}}[\hat{f}(X)]$, and (ii) the leader's MSE under the chosen environment, $R_{e^*(\hat{f})}(\hat{f})$. Both are plotted

Table 4: The 49 candidate environments of the follower, ordered by their distance d to e_{ref} (rank 1), defined in § G.5. For an intervention bound b, the choice set $\mathcal{E}_b$ consists of the environments with rank at most $\max\{1, \mathrm{round}(49b)\}$.

Rank	β_0^{led}	β_1^{led}	β_0^{pol}	β_1^{pol}	d	Rank	β_0^{led}	β_1^{led}	β_0^{pol}	β_1^{pol}	d
1	0	12	0	30	0.00	26	12	−12	0	30	0.96
2	0	6	0	15	0.25	27	12	−12	30	−30	1.00
3	0	19	0	45	0.28	28	0	12	30	−30	1.00
4	0	0	0	15	0.48	29	0	0	30	−30	1.00
5	0	0	0	30	0.48	30	19	−19	0	0	1.24
6	0	0	0	45	0.48	31	19	−19	0	45	1.24
7	0	0	0	0	0.50	32	19	−19	45	−45	1.25
8	0	6	0	0	0.50	33	0	0	45	−45	1.25
9	0	0	0	60	0.50	34	0	19	45	−45	1.25
10	0	12	0	0	0.50	35	25	−25	0	0	1.48
11	0	19	0	0	0.50	36	25	−25	0	60	1.48
12	0	25	0	60	0.52	37	25	−25	60	−60	1.50
13	0	25	0	0	0.52	38	0	25	60	−60	1.50
14	6	−6	0	0	0.72	39	0	0	60	−60	1.50
15	6	−6	0	15	0.72	40	31	−31	0	0	1.72
16	0	0	0	75	0.75	41	31	−31	0	75	1.72
17	6	−6	15	−15	0.75	42	31	−31	75	−75	1.75
18	0	6	15	−15	0.75	43	0	31	75	−75	1.75
19	0	0	15	−15	0.75	44	0	0	75	−75	1.75
20	0	31	0	75	0.76	45	35	−35	0	0	1.88
21	0	31	0	0	0.76	46	35	−35	0	84	1.88
22	0	0	0	84	0.90	47	0	0	84	−84	1.90
23	0	35	0	84	0.92	48	0	35	84	−84	1.90
24	0	35	0	0	0.92	49	35	−35	84	−84	1.90
25	12	−12	0	0	0.96						

against b in Fig. 4; for reference, the right panel also shows the leader's MSE under e_{ref} as a dotted horizontal line for each predictor.

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 describe our contributions relative to prior literature in § 1 and in § B.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We frame the setting that we consider in the abstract and in § 1. § 2 rigorously describes the exact setting and assumptions, and § 7 describes the limitations of our results.

   Guidelines:

   - The answer [N/A] 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: We rigorously define our setup in § 2, and lemma, proposition, and theorem statements include the assumptions we make. § E contains the complete proofs for all our results.

Guidelines:

- The answer [N/A] 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: § G and § F provide the details for the experiments described in § 6.

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: The code is available at `https://github.com/LinusKuehne/prediction-intervention-games`. To access the `causalchamber` API to produce data, credentials are required (`https://www.causalchamber.ai/`); we provide the produced Causal Chambers data as `.csv` files along with Python code used to produce them (thus, all experiments are reproducible).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

Answer: [Yes]

Justification: We describe our experiments in § 6 and give further details in § F and § G. We also provide the code with the full details.

Guidelines:

- The answer [N/A] 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: We report 95% confidence intervals for the main experiments in § 6.1. For the Causal Chambers application in § 6.2, we report a single physical-data evaluation because repeated API data collection is time-consuming; we present this experiment as an illustrative real-data application.

Guidelines:

- The answer [N/A] 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 experiments described in § 6 are computationally very light and can be run on a personal laptop in less than a few hours. See § F.4 and § G.4.

   Guidelines:

   - The answer [N/A] 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 reviewed the NeurIPS Code of Ethics and confirm that our work conforms to it in every respect.

   Guidelines:

   - The answer [N/A] 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: Our work is mainly of theoretical nature. We discuss real-world applications and examples in the abstract, in § 1, and in § A.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work is mostly of theoretical nature and poses no such risks.

Guidelines:

- The answer [N/A] 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: We conform to the licenses associated with the Causal Chambers data and give appropriate citations, including URLs. We cite the original papers that present practical methods subject to our experiments.

Guidelines:

- The answer [N/A] 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: [Yes]

    Justification: We document in § G.2 how we generate our Causal Chambers dataset and provide the complete Python scripts.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.