Learning Counterfactually Invariant Predictors

Anonymous authors
Paper under double-blind review

Abstract

Notions of counterfactual invariance (CI) have proven essential for predictors that are fair, robust, and generalizable in the real world. We propose graphical criteria that yield a sufficient condition for a predictor to be counterfactually invariant in terms of a conditional independence in the observational distribution. In order to learn such predictors, we propose a model-agnostic framework, called Counterfactually Invariant Prediction (CIP), building on the Hilbert-Schmidt Conditional Independence Criterion (HSCIC), a kernel-based conditional dependence measure. Our experimental results demonstrate the effectiveness of CIP in enforcing counterfactual invariance across various simulated and real-world datasets including scalar and multi-variate settings.

1 Introduction

Invariance, or equivariance to certain data transformations, has proven essential in numerous applications of machine learning (ML), since it can lead to better generalization capabilities (Arjovsky et al., 2019; Chen et al., 2020; Bloem-Reddy & Teh, 2020). For instance, in image recognition, predictions ought to remain unchanged under scaling, translation, or rotation of the input image. Data augmentation, an early heuristic to promote such invariances, has become indispensable for successfully training deep neural networks (DNNs) (Shorten & Khoshgoftaar, 2019; Xie et al., 2020). Well-known examples of "invariance by design" include convolutional neural networks (CNNs) for translation invariance (Krizhevsky et al., 2012), group equivariant NNs for general group transformations (Cohen & Welling, 2016), recurrent neural networks (RNNs) and transformers for sequential data (Vaswani et al., 2017), DeepSet (Zaheer et al., 2017) for sets, and graph neural networks (GNNs) for different types of geometric structures (Battaglia et al., 2018).

Many applications in modern ML, however, call for arguably stronger notions of invariance based on causality. This case has been made for image classification, algorithmic fairness (Hardt et al., 2016; Mitchell et al., 2021), robustness (Bühlmann, 2020), and out-of-distribution generalization (Lu et al., 2021). The goal is invariance with respect to hypothetical manipulations of the data generating process (DGP). Various works develop methods that assume observational distributions (across environments or between training and test) to be governed by shared causal mechanisms, but differ due to various types of distribution shifts encoded by the causal model (Peters et al., 2016; Heinze-Deml et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Bühlmann, 2020; Subbaswamy et al., 2022; Yi et al., 2022; Makar et al., 2022). Typical goals include to train predictors invariant to such shifts, to learn about causal mechanisms and to improve robustness against spurious correlations or out of distribution generalization. The term "counterfactual invariance" has also been used in other out of distribution learning contexts unrelated to our task, e.g., to denote invariance to certain symmetry transformations (Mouli & Ribeiro, 2022).

While we share the broader motivation, these works are orthogonal to ours, because even though counterfactual distributions are also generated from the same causal model, they are fundamentally different from such shifts (Peters et al., 2017). Intuitively, counterfactuals are about events that did not, but could have happened had circumstances been different in a controlled way. A formal discussion of what we mean by counterfactuals is required to properly position our work in the existing literature and describe our contributions. Overall, our work provides a graphical criterion for counterfactual causal inference under specific injectivity conditions, alongside a new model-agnostic learning framework named Counterfactually Invariant Prediction (CIP), which employs a kernel-based conditional dependence measure effective across mixed variable types. We empirically assess CIP in (semi-)synthetic settings and show that it effectively enforces counterfactual invariance.

2 Problem setting and related work

2.1 Preliminaries and terminology

Definition 2.1 (Structural causal model (SCM))**.** *A structural causal model is a tuple* $\mathcal{S} = (\mathbf{U}, \mathbf{V}, F, \mathbb{P}_{\mathbf{U}})$ *such that* $\mathbf{U}$ *is a set of background variables that are exogenous to the model;* $\mathbf{V}$ *is a set of observable (endogenous) variables;* $F = \{f_V\}_{V \in \mathbf{V}}$ *is a set of functions from (the domains of)* $\mathsf{pa}(V) \cup U_V$ *to (the domain of)* V*, where* $U_V \subset \mathbf{U}$ *and* $\mathsf{pa}(V) \subseteq \mathbf{V} \setminus \{V\}$ *such that* $V = f_V(\mathsf{pa}(V), U_V)$*;* $\mathbb{P}_{\mathbf{U}}$ *is a probability distribution over the domain of* $\mathbf{U}$*. Further, the subsets* $\mathsf{pa}(V) \subseteq \mathbf{V} \setminus \{V\}$ *are chosen such that the graph* $\mathcal{G}$ *over* $\mathbf{V}$ *where the edge* $V' \to V$ *is in* $\mathcal{G}$ *if and only if* $V' \in \mathsf{pa}(V)$ *is a directed acyclic graph (DAG).*

Observational distribution. An SCM implies a unique observational distribution over $\mathbf{V}$, which can be thought of as being generated by transforming the distribution over $\mathbb{P}_{\mathbf{U}}$ via the deterministic functions in F iteratively to obtain a distribution over $\mathbf{V}$.[1]

Interventions. Given a variable $A \in \mathbf{V}$, an intervention $A \leftarrow a$ amounts to replacing f_A in F with the constant function setting A to a. This yields a new SCM, which induces the *interventional distribution* under intervention $A \leftarrow a$.[2] Similarly, we can intervene on multiple variables $\mathbf{V} \supseteq \mathbf{A} \leftarrow \mathbf{a}$. For an outcome (or prediction target) variable $\mathbf{Y} \subset \mathbf{V}$, we then write $\mathbf{Y_a}$ for the outcome in the intervened SCM, also called *potential outcome*. Note that the interventional distribution $\mathbb{P}_{\mathbf{Y_a}}(\mathbf{y})$ differs in general from the conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{A}}(\mathbf{y} \mid \mathbf{a})$.[3] This is typically the case when Y and A have a shared ancestor, i.e., they are confounded. In interventional distributions, potential outcomes are random variables via the exogenous variables $\mathbf{u}$, i.e., $Y_{\mathbf{a}}(u)$ where $\mathbf{u} \sim \mathbb{P}_{\mathbf{U}}$. Hence, interventions capture "population level" properties, i.e., the action is performed for all units $\mathbf{u}$.

Counterfactuals. Counterfactuals capture what happens under interventions for a "subset" of possible units $\mathbf{u}$ that are compatible with observations $\mathbf{W} = \mathbf{w}$ for a subset of observed variables $\mathbf{W} \subseteq \mathbf{V}$. This can be described in a three step procedure. (i) *Abduction:* We restrict our attention to units compatible with the observations, i.e., consider the new SCM $\mathcal{S}^{\mathbf{w}} = (\mathbf{U}, \mathbf{V}, F, \mathbb{P}_{\mathbf{U}|\mathbf{W}=\mathbf{w}})$. (ii) *Intervention:* Within $\mathcal{S}^{\mathbf{w}}$, perform an intervention $\mathbf{A} \leftarrow \mathbf{a}$ on some variables $\mathbf{A}$ (which need not be disjoint from $\mathbf{W}$). (iii) *Prediction:* Finally, we are typically interested in the outcome $\mathbf{Y}$ in an interventional distribution of $\mathcal{S}^{\mathbf{w}}$, which we denote by $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}=\mathbf{w}}(\mathbf{y})$ and call a *counterfactual distribution*: "Given that we have observed $\mathbf{W} = \mathbf{w}$, what would $\mathbf{Y}$ have been had we set $\mathbf{A} \leftarrow \mathbf{a}$, instead of the value $\mathbf{A}$ has actually taken?". In this notation the conditioning set $\mathbf{W}$ is conditioned on in the pre-interventional world. Counterfactuals capture properties of a "subpopulation" $\mathbf{u} \sim \mathbb{P}_{\mathbf{U}|\mathbf{W}=\mathbf{w}}$ compatible with the observations.[4] Even for fine-grained $\mathbf{W}$, there may be multiple units $\mathbf{u}$ in the support of this distribution. In contrast, "unit level counterfactuals" often considered in philosophy contrast $\mathbf{Y}^*_{\mathbf{a}}(\mathbf{u})$ with $\mathbf{Y}^*_{\mathbf{a}'}(\mathbf{u})$ for a single unit $\mathbf{u}$. Such unit level counterfactuals are too fine-grained in our setting. Hence, our used definition of counterfactual invariance is:

Definition 2.2 (Counterfactual invariance)**.** *Let* $\mathbf{A}$*,* $\mathbf{W}$ *be (not necessarily disjoint) sets of nodes in a given SCM. Then,* $\mathbf{Y}$ *is* counterfactually invariant in $\mathbf{A}$ w.r.t. $\mathbf{W}$ *if* $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}=\mathbf{w}}(\mathbf{y}) = \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{W}=\mathbf{w}}(\mathbf{y})$ *almost surely, for all* $\mathbf{a}, \mathbf{a}'$ *in the domain of* $\mathbf{A}$ *and all* $\mathbf{w}$ *in the domain of* $\mathbf{W}$.[5]

Predictors in SCMs. Ultimately, we aim at learning a predictor $\hat{\mathbf{Y}}$ for the outcome $\mathbf{Y}$. Originally, the predictor $\hat{\mathbf{Y}}$ is not part of the DGP, because we get to learn $f_{\hat{\mathbf{Y}}}$ from data. Using supervised learning, the predictor $f_{\hat{\mathbf{Y}}}$ depends both on the chosen inputs $\mathbf{X} \subset \mathbf{V}$ as well as the target $\mathbf{Y}$. However, once $f_{\hat{\mathbf{Y}}}$ is fixed, it is a deterministic function with arguments $\mathbf{X} \subset \mathbf{V}$, so $(\mathbf{U}, \mathbf{V} \cup \{\hat{\mathbf{Y}}\}, F \cup \{f_{\hat{\mathbf{Y}}}\}, \mathbb{P}_{\mathbf{U}})$ is a valid SCM and we can consider $\hat{\mathbf{Y}}$ an observed variable with incoming arrows from only $\mathbf{X}$. Hence, the definition of counterfactual invariance can be applied to the predictor $\hat{\mathbf{Y}}$.

[1] Note that all randomness stems from $\mathbb{P}_{\mathbf{U}}$. The observational distribution is well-defined and unique, essentially because every DAG allows for a topological order.

[2] The observational distribution in an intervened SCM is called interventional distribution of the base SCM.

[3] We use $\mathbb{P}$ for distributions (common in the kernel literature) and $\mathbf{Y_a}$ instead of the do notation.

[4] Note that conditioning in an interventional distribution is different from a counterfactual and our notation is quite subtle here $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y} \mid \mathbf{W} = \mathbf{w}) \neq \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}=\mathbf{w}}(\mathbf{y})$.

[5] With an abuse of notation, if $\mathbf{W} = \emptyset$ then the requirement of conditional counterfactual invariance becomes $\mathbb{P}_{\mathbf{Y_a}}(\mathbf{y}) = \mathbb{P}_{\mathbf{Y_{a'}}}(\mathbf{y})$ almost surely, for all $\mathbf{a}, \mathbf{a}'$ in the domain of $\mathbf{A}$. The "almost surely" part in our definition merely refers to the type of equality of distributions and is not related to the "almost sure" in a.s.-CI defined by Fawkes & Evans (2023).

Kernel mean embeddings (KME). Our method relies on kernel mean embeddings (KMEs). We describe the main concepts pertaining KMEs and refer the reader to Smola et al. (2007); Schölkopf et al. (2002); Berlinet & Thomas-Agnan (2011); Muandet et al. (2017) for details. Fix a measurable space $\mathscr{Y}$ with respect to a σ-algebra $\mathcal{F}_{\mathscr{Y}}$, and consider a probability measure $\mathbb{P}$ on the space $(\mathscr{Y}, \mathcal{F}_{\mathscr{Y}})$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) with a bounded kernel $k_{\mathbf{Y}} \colon \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$, i.e., $k_{\mathbf{Y}}$ is such that $\sup_{\mathbf{y} \in \mathscr{Y}} k(\mathbf{y}, \mathbf{y}) < \infty$. The kernel mean embedding $\mu_{\mathbb{P}}$ of $\mathbb{P}$ is defined as the expected value of the function $k(\,\cdot\,, \mathbf{y})$ with respect to $\mathbf{y}$, i.e., $\mu_{\mathbb{P}} \coloneqq \mathbb{E}\left[k(\,\cdot\,, \mathbf{y})\right]$. The definition of KMEs can be extended to conditional distributions (Fukumizu et al., 2013; Grünewälder et al., 2012; Song et al., 2009; 2013). Consider two random variables $\mathbf{Y}$, $\mathbf{S}$, and denote with $(\Omega_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}})$ and $(\Omega_{\mathbf{S}}, \mathcal{F}_{\mathbf{S}})$ the respective measurable spaces. These random variables induce a probability measure $\mathbb{P}_{\mathbf{Y},\mathbf{S}}$ in the product space $\Omega_{\mathbf{Y}} \times \Omega_{\mathbf{S}}$. Let $\mathcal{H}_{\mathbf{Y}}$ be a RKHS with a bounded kernel $k_{\mathbf{Y}}(\cdot, \cdot)$ on $\Omega_{\mathbf{Y}}$. We define the KME of a conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{S}}(\cdot \mid \mathbf{s})$ via $\mu_{\mathbf{Y}|\mathbf{S}=\mathbf{s}} \coloneqq \mathbb{E}\left[k_{\mathbf{Y}}(\,\cdot\,, \mathbf{y}) \mid \mathbf{S} = \mathbf{s}\right]$. Here, the expected value is taken over $\mathbf{y}$. KMEs of conditional measures can be estimated from samples (Grünewälder et al., 2012).Pogodin et al. (2022) recently proposed an efficient kernel-based regularizer for learning features of input data that allow for estimating a target while being conditionally independent of a distractor given the target. Since CIP ultimately enforces conditional independence (see Theorem 3.2), we believe it could further benefit from leveraging the efficiency and convergence properties of their technique, which we leave for future work.

2.2 Related work and contributions

While we focus on counterfactuals in the SCM framework (Pearl, 2000; Peters et al., 2016), there are different incompatible frameworks to describe counterfactuals (von Kügelgen et al., 2022; Dorr, 2016; Woodward, 2021), which may give rise to orthogonal notions of counterfactual invariance.

Research on algorithmic fairness has explored a plethora of causal "invariance" notions with the goal of achieving fair predictors (Loftus et al., 2018; Carey & Wu, 2022; Plecko & Bareinboim, 2022). Kilbertus et al. (2017) conceptually introduce a notion based on group-level interventions, which has been refined to take into account more fine-grained context by Salimi et al. (2019); Galhotra et al. (2022), who then obtain fair predictors by viewing it as a database repair problem or a causal feature selection problem, respectively. A counterfactual-level definition was proposed by Kusner et al. (2017) and followed up by path-specific counterfactual notions (Nabi & Shpitser, 2018; Chiappa, 2019), where the protected attribute may take different values along different paths to the outcome. Recently, Dutta et al. (2021) developed an information theoretic framework to decompose the overall causal influence allowing for exempted variables and properly dealing with synergies across different paths.

Our focus is on counterfactuals because they are fundamentally more expressive than mere interventions (Pearl, 2000; Bareinboim et al., 2022), but do not require a fine-grained path- or variable-level judgment of "allowed" and "disallowed" paths or variables, which may be challenging to devise in practice. Since CI already requires strong assumptions, we leave path-specific counterfactuals—even more challenging in terms of identifiability (Avin et al., 2005)—for future work. While our Definition 2.2 requires equality in distribution, Veitch et al. (2021) suggest a definition of a counterfactually invariant predictor $f_{\hat{\mathbf{Y}}}$ which requires almost sure equality of $\hat{\mathbf{Y}}^*_{\mathbf{a}}$ and $\hat{\mathbf{Y}}^*_{\mathbf{a}'}$, where we view $\hat{\mathbf{Y}}$ as an observed variable in the SCM as described above. Fawkes & Evans (2023) recently carefully formulated various precise technical definitions of what counterfactual invariance may mean in different contexts, such as "almost sure CI" (a.s.-CI), "distributional CI" ($\mathcal{D}$-CI), and "CI of predictors" ($\mathcal{F}$-CI). They provide various connections between them such as the fact that $f_{\hat{\mathbf{Y}}}$ being $\mathcal{F}$-CI is equivalent to $\hat{\mathbf{Y}}$ being $\mathcal{D}$-CI conditioned on $\mathbf{X}$, yielding an equivalence to the definition of counterfactual fairness (Kusner et al., 2017). The notion of counterfactual invariance in definition 2.2 is most closely related to $\mathcal{D}$-CI from Fawkes & Evans (2023), but we do not enforce conditioning on the intervening variable.

Inspired by problems in natural language processing (NLP), Veitch et al. (2021) aim at "stress-testing" models for spurious correlations. It differs from our work in that they (i) focus only on two specific graphs, and (ii) provide a *necessary* but not sufficient criterion for CI in terms of a conditional independence. Their method enforces the conditional independence via maximum mean discrepancy (MMD) (in *discrete settings only*). However, enforcing a consequence of CI, does not necessarily improve CI. Indeed, Fawkes & Evans (2023, Prop. 4.4) show that while a.s.-CI implies certain conditional independencies, no set of conditional

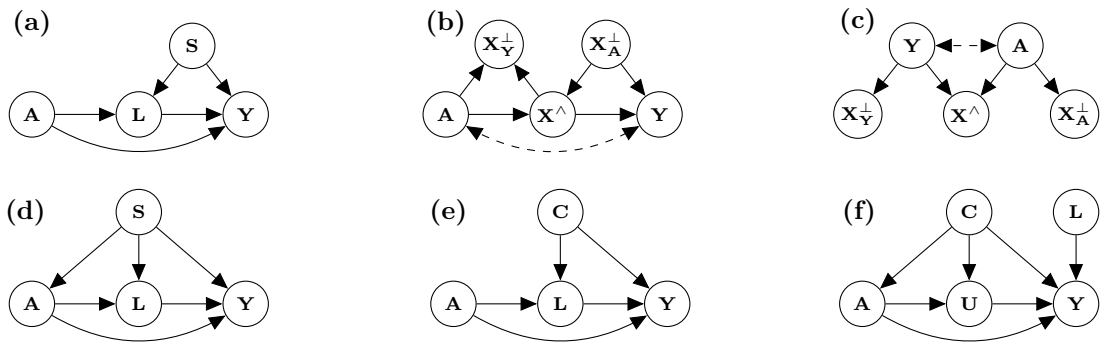

Figure 1: **(a)** Exemplary graph in which a predictor $\hat{\mathbf{Y}}$ with $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{L} \mid \mathbf{S}$ is CI in $\mathbf{A}$ w.r.t. $\{\mathbf{L}, \mathbf{S}\}$. **(b)-(c)** Causal and anti-causal structure from Veitch et al. (2021) where $\mathbf{X}_\mathbf{A}^\perp$ is not causally influenced by $\mathbf{A}$, $\mathbf{X}_\mathbf{Y}^\perp$ does not causally influence $\mathbf{Y}$, and $\mathbf{X}^\wedge$ is both influenced by $\mathbf{A}$ and influences $\mathbf{Y}$. **(d)** Assumed causal structure for the synthetic experiments, see Section 4.1 and Appendix F for details. **(e)** Assumed causal graph for the UCI Adult dataset (Section 4.3), where $\mathbf{A} = \{\text{Gender, Age}\}$. **(f)** Causal structure for our semi-synthetic image experiments (Section 4.2), where $\mathbf{A} = \{\text{Pos.X}\}$, $\mathbf{U} = \{\text{Scale}\}$, $\mathbf{C} = \{\text{Shape, Pos.Y}\}$, $\mathbf{L} = \{\text{Color, Orientation}\}$, and $\mathbf{Y} = \{\text{Outcome}\}$.

independencies implies any bounds on the difference in counterfactuals. On the contrary, distributional notions of CI such as $\mathcal{D}$-CI or our definition 2.2 are weaker than a.s.-CI (see (Fawkes & Evans, 2023, Lem. 2.4). Therefore, these weaker notions can indeed be written equivalently as conditional independencies in the observational distribution under additional assumptions about the data generating mechanism (Fawkes & Evans, 2023, Lem. A.3). For example, Fawkes & Evans (2023, Lem. A.3) shows that $\mathcal{D}$-CI can be implied by conditional independence in special settings where the counterfactual distribution is identified from the observational one. Similarly, our reduction of CI (as in definition 2.2) to conditional independence in Theorem 3.2 requires a strong injectivity assumption that essentially amounts to being able to remove exogenous uncertainty.

Contributions. We provide such a sufficient graphical criterion for $\mathcal{D}$-CI under an injectivity condition of a structural equation. Depending on the assumed causal graph, this can also come at the cost of requiring certain variables to be observed. As our main contribution, we propose a model-agnostic learning framework, called Counterfactually Invariant Prediction (CIP), using a kernel-based conditional dependence measure that also works for mixed categorical and continuous, multivariate variables. We evaluate CIP extensively in (semi-)synthetic settings and demonstrate its efficacy in enforcing counterfactual invariance even when the strict assumptions may be violated.

3 Counterfactually invariant prediction (CIP)

3.1 Sufficient criterion for counterfactual invariance

We will now establish a sufficient graphical criterion to express CI as conditional independence in the observational distribution, rendering it estimable from data. First, we need some terminology.

Graph terminology. Consider a path π (a sequence of distinct adjacent nodes) in a DAG $\mathcal{G}$. A set of nodes $\mathbf{S}$ is said to *block* π, if π contains a triple of consecutive nodes A, B, C such that one of the following hold: (i) $A \to B \to C$ or $A \leftarrow B \leftarrow C$ or $A \leftarrow B \to C$ and $B \in \mathbf{S}$; (ii) $A \to B \leftarrow C$ and neither B nor any descendent of B is in $\mathbf{S}$. Further, we call π a *causal path* between sets of nodes $\mathbf{A}, \mathbf{B}$, when it is a directed path from a node in $\mathbf{A}$ to a node in $\mathbf{B}$. A causal path π is a *proper causal path* if it only intersects $\mathbf{A}$ at the first node in π. Finally, we denote with $\mathcal{G}_\mathbf{A}$ the graph obtained by removing from $\mathcal{G}$ all incoming arrows into nodes in $\mathbf{A}$. We now define the notion of valid adjustment sets (Shpitser et al., 2010, Def. 5), which our graphical criterion for CI relies on.

Definition 3.1 (valid adjustment set)**.** *Let $\mathcal{G}$ be a causal graph and let $\mathbf{X}$, $\mathbf{Y}$ be disjoint (sets of) nodes in $\mathcal{G}$. A set of nodes $\mathbf{S}$ is a valid adjustment set for $(\mathbf{X}, \mathbf{Y})$, if (i) no element in $\mathbf{S}$ is a descendant in $\mathcal{G}_\mathbf{X}$ of any*

node $W \notin \mathbf{X}$ which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$, and (ii) $\mathbf{S}$ blocks all non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{G}$.

We now state the sufficient graphical criterion that renders CI equivalent to a conditional independence. The proof builds on known results from the literature such as the full characterization of valid adjustment sets in fully observed structural causal models, but the final statement requires a substantial original theoretical contribution, which we detail in Appendix A.

Theorem 3.2. *Let $\mathcal{G}$ be a causal graph, $\mathbf{A}$, $\mathbf{W}$ be two (not necessarily disjoint) sets of nodes in $\mathcal{G}$, such that $(\mathbf{A} \cup \mathbf{W}) \cap \mathbf{Y} = \emptyset$, let $\mathbf{S}$ be a valid adjustment set for $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$. Further, for any random variable $X \in \mathbf{W} \setminus \mathbf{A}$ denote with $g_X(\mathsf{pa}(X), U_X)$ its structural equation, and suppose that $\mathsf{pa}(X) \subseteq \mathbf{A} \cup \mathbf{W}$. Suppose that g_X is injective in the variable $\mathbf{u}$.[6] Then, in all SCMs compatible with $\mathcal{G}$, if a predictor $\hat{\mathbf{Y}}$ satisfies $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$, then $\hat{\mathbf{Y}}$ is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W}$.*

Assumptions. First, we do *not* assume the SCM (or any structural equations) to be known. We do assume the causal graph to be known. This is a standard assumption widely made in the causality literature, even though it is a strong one (Cartwright, 2007). The additional assumption of Theorem 3.2, namely injectivity of g_X is satisfied—but more general than—a wide variety of commonly considered models in causality such as the widely used Additive Noise Models (Peters et al., 2017). More broadly, Fawkes & Evans (2023, Lem. A.3) show that to achieve CI from the observational distribution and the causal graph, additional assumptions are always required.

3.2 Example use-cases of counterfactually invariant prediction

Fig. 1(a) shows an exemplary graph in which the outcome $\mathbf{Y}$ is affected by (disjoint) sets $\mathbf{A}$ (in which we want to be CI), $\mathbf{L}$ and $\mathbf{S}$ (inputs to $f_{\hat{\mathbf{Y}}}$). We consider $\mathbf{W} = \mathbf{L} \cup \mathbf{A} \cup \mathbf{S}$. Here we aim to achieve $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{L} \mid \mathbf{S}$ to obtain CI in $\mathbf{A}$ w.r.t. $\mathbf{W}$. In our synthetic experiments, we also allow $\mathbf{S}$ to affect $\mathbf{A}$, see Fig. 1(d). Let us further illustrate concrete potential applications of CI, which we later also study in our experiments.

Counterfactual fairness. Counterfactual fairness (Kusner et al., 2017) informally challenges a consequential decision: "*Would I have gotten the same outcome had my gender, race, or age been different with all else being equal*?". Here $\mathbf{Y} \subset \mathbf{V}$ denotes the outcome and $\mathbf{A} \subset \mathbf{V} \setminus \mathbf{Y}$ the *protected attributes* such as gender, race, or age—protected under anti-discrimination laws (Barocas & Selbst, 2016)—by $\mathbf{A} \subseteq \mathbf{V} \setminus \mathbf{Y}$. Collecting all remaining observed covariates into $\mathbf{W} := \mathbf{V} \setminus \mathbf{Y}$ counterfactual fairness reduces to counterfactual invariance. In experiments, we build a semi-synthetic DGP assuming the graph in Fig. 1(e) for the UCI adult dataset (Kohavi & Becker, 1996).

Robustness. CI serves as a strong notion of robustness in settings such as image classification: "*Would the truck have been classified correctly had it been winter in this situation instead of summer*?" For concrete demonstration, we use the dSprites dataset (Matthey et al., 2017) consisting of simple black and white images of different shapes (squares, ellipses, ...), sizes, orientations, and locations. We devise a DGP for this dataset with the graph depicted in Fig. 1(f).

Text classification. Veitch et al. (2021) motivate the importance of counterfactual invariance in text classification tasks. Specifically, they consider the causal and anti-causal structures depicted in Veitch et al. (2021, Fig. 1), which we replicate in Fig. 1(b,c). Both diagrams consist of protected attributes $\mathbf{A}$, observed covariates $\mathbf{X}$, and outcomes $\mathbf{Y}$. To apply our sufficient criterion to their settings, we must assume that $\mathbf{A}$ and $\mathbf{Y}$ are unconfounded. We show in Appendix G.1 that CIP still performs on par with Veitch et al. (2021) even when this assumption is violated. Theorem 3.2 provides a sufficient condition for CI (Definition 2.2) in terms of the conditional independence $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$. We next develop an operator $\mathrm{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S})$ that is (a) efficiently estimable from data, (b) differentiable, (c) a monotonic measure of conditional dependence, and (d) is zero if and only if $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$. Hence, it is a practical objective to enforce CI.

[6] The injectivity of g_X is defined as follows. Consider two pairs $\{\mathbf{p}, u\}$ and $\{\mathbf{p}, u'\}$ with $\mathbf{p}$ in the support of $\mathsf{pa}(X)$ and u, u' in the support of U_X. Suppose that $\mathbb{P}_{\mathsf{pa}(X),U_X}(\mathbf{p}, u) \neq 0$ and $\mathbb{P}_{\mathsf{pa}(X),U_X}(\mathbf{p}, u') \neq 0$. Then, it holds $g(\mathbf{p}, u) = g(\mathbf{p}, u')$ if and only if $u = u'$.

3.3 HSCIC for conditional independence

Consider two sets of random variables $\mathbf{Y}$ and $\mathbf{A} \cup \mathbf{W}$, and denote with $(\Omega_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}})$ and $(\Omega_{\mathbf{A}\cup\mathbf{W}}, \mathcal{F}_{\mathbf{A}\cup\mathbf{W}})$ the respective measurable spaces. Suppose that we are given two RKHSs $\mathcal{H}_{\mathbf{Y}}$, $\mathcal{H}_{\mathbf{A}\cup\mathbf{W}}$ over the support of $\mathbf{Y}$ and $\mathbf{A} \cup \mathbf{W}$ respectively. The tensor product space $\mathcal{H}_{\mathbf{Y}} \otimes \mathcal{H}_{\mathbf{A}\cup\mathbf{W}}$ is defined as the space of functions of the form $(f \otimes g)(\mathbf{y}, [\mathbf{a}, \mathbf{w}]) \coloneqq f(\mathbf{y})g([\mathbf{a}, \mathbf{w}])$, for all $f \in \mathcal{H}_{\mathbf{Y}}$ and $g \in \mathcal{H}_{\mathbf{A}\cup\mathbf{W}}$. The tensor product space yields a natural RKHS structure, with kernel k defined by $k(\mathbf{y} \otimes [\mathbf{a}, \mathbf{w}], \mathbf{y}' \otimes [\mathbf{a}', \mathbf{w}']) \coloneqq k_{\mathbf{Y}}(\mathbf{y}, \mathbf{y}')k_{\mathbf{A}\cup\mathbf{W}}([\mathbf{a}, \mathbf{w}], [\mathbf{a}', \mathbf{w}'])$. We refer the reader to Szabó & Sriperumbudur (2017) for more details on tensor product spaces.

Definition 3.3 (HSCIC)**.** *For (sets of) random variables* $\mathbf{Y}$*,* $\mathbf{A} \cup \mathbf{W}$*,* $\mathbf{S}$*, the HSCIC* between $\mathbf{Y}$ and $\mathbf{A} \cup \mathbf{W}$ given $\mathbf{S}$ *is defined as the real-valued random variable* $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A}\cup\mathbf{W} \mid \mathbf{S}) = H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \circ \mathbf{S}$ *where* $H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}$ *is a real-valued deterministic function, defined as* $H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}(\mathbf{s}) \coloneqq \|\mu_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}=\mathbf{s}} - \mu_{\mathbf{Y}|\mathbf{S}=\mathbf{s}} \otimes \mu_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}=\mathbf{s}}\|$ *with* $\|\cdot\|$ *the norm induced by the inner product of the tensor product space* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}\cup\mathbf{W}}$*.*

Our Definition 3.3 is motivated by, but differs slightly from Park & Muandet (2020, Def. 5.3), which relies on the Bochner conditional expected value. While it is functionally equivalent (with the same implementation, see Eq. (2)), ours has the benefit of bypassing some technical assumptions required by Park & Muandet (2020) (see Appendices C and D for details). The HSCIC has the following important property, proved in Appendix B.

Theorem 3.4 (Theorem 5.4 by Park & Muandet (2020))**.** *If the kernel* k *of* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}\cup\mathbf{W}}$ *is characteristic*[7]*,* $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = 0$ *almost surely if and only if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$*.*

Because "most interesting" kernels such as the Gaussian and Laplacian kernels are characteristic, and the tensor product of translation-invariant characteristic kernels is characteristic again (Szabó & Sriperumbudur, 2017), this natural assumption is non-restrictive in practice. Combining Theorems 3.2 and 3.4, we can now use HSCIC to reliably achieve counterfactual invariance.

Corollary 3.5. *Under the assumptions of Theorem 3.2, if* $\mathrm{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = 0$ *almost surely, then* $\hat{\mathbf{Y}}$ *is counterfactually invariant in* $\mathbf{A}$ *with respect to* $\mathbf{W}$*.*

In addition, since HSCIC is defined in terms of the MMD (Definition 3.3 and Park & Muandet (2020, Def. 5.3)), it inherits the weak convergence property, i.e., if $\mathrm{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S})$ converges to zero, then the counterfactual distributions (for different intervention values $\mathbf{a}$) weakly converge to the same distribution. We refer to Simon-Gabriel & Schölkopf (2018); Simon-Gabriel et al. (2020) for a precise characterization. Hence, as HSCIC decreases, the predictor approaches counterfactual invariance and we need not drive HSCIC all the way to zero to obtain meaningful results.

3.4 Learning counterfactually invariant predictors (CIP)

Corollary 3.5 justifies our proposed objective, namely to minimize the following loss

$$\mathcal{L}_{\mathrm{CIP}}(\hat{\mathbf{Y}}) = \mathcal{L}(\hat{\mathbf{Y}}) + \gamma \cdot \mathrm{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) \,, \qquad \text{[CIP loss]} \tag{1}$$

where $\mathcal{L}(\hat{\mathbf{Y}})$ is a task-dependent loss function (e.g., cross-entropy for classification, or mean squared error for regression) and $\gamma \geq 0$ is a parameter that regulates the trade-off between predictive performance and counterfactual invariance.

The meaning of γ and how to choose it. The second term in Eq. (1) amounts to the additional objective of CI, which is typically at odds with predictive performance within the observational distribution $\mathcal{L}$. In practice, driving HSCIC to zero, i.e., viewing our task as a constrained optimization problem, typically deteriorates predictive performance too much to be useful for prediction—especially in small data settings.[8] As the choice of γ amounts to choosing an operating point between predictive performance and CI, it cannot be selected in a data-driven fashion. As different settings call for different tradeoffs, we advocate for employing the following procedure: (i) Train an unconstrained predictor for a base predictive performance (e.g., 92% accuracy or 0.21 MSE). (ii) Fix a tolerance level α, indicating the maximally tolerable loss in predictive performance (e.g., at most 5% drop in accuracy or at most 10% increase in MSE). (iii) Perform a

[7] The tensor product kernel k is characteristic if $\mathbb{P}_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}} \mapsto \mathbb{E}_{\mathbf{y},[\mathbf{a},\mathbf{w}]}\left[k(\,\cdot\,, \mathbf{y} \otimes [\mathbf{a}, \mathbf{w}])\right]$ is injective.

[8] In particular, HSCIC does not regularize an ill-posed problem, i.e., it does not merely break ties between predictors with equal $\mathcal{L}(\hat{\mathbf{Y}})$. Hence it also need not decay to zero as the sample size increases.

log-spaced binary search on γ (e.g., on $[10^{-4}, 10^4]$) to find the largest γ such that the predictive performance of the resulting predictor achieves predictive performance within the tolerance α—see Appendix F.6 for an illustration. A similar search for the optimal value of γ can be conducted, when there is a fixed requirement for a maximum tolerance of counterfactual invariance as measured by HSCIC.

Estimating the HSCIC from samples. The key benefit of HSCIC as a conditional independence measure is that it does not require parametric assumptions on the underlying probability distributions, and it is applicable for any mixed, multi-dimensional data modalities, as long as we can define positive definite kernels on them. Given n samples $\{(\hat{\mathbf{y}}_i, \mathbf{a}_i, \mathbf{w}_i, \mathbf{s}_i)\}_{i=1}^n$, denote with $\hat{K}_{\hat{\mathbf{Y}}}$ the kernel matrix with entries $[\hat{K}_{\hat{\mathbf{Y}}}]_{i,j} \coloneqq k_{\hat{\mathbf{Y}}}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j)$, and let $\hat{K}_{\mathbf{A}\cup\mathbf{W}}$ be the kernel matrix for $\mathbf{A} \cup \mathbf{W}$. We estimate the $H_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \equiv H_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}(\cdot)$ as

$$\begin{aligned}\hat{H}^2_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} &= \hat{w}^T_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \left(\hat{K}_{\hat{\mathbf{Y}}} \odot \hat{K}_{\mathbf{A}\cup\mathbf{W}}\right) \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \\ &- 2\left(\hat{w}^T_{\hat{\mathbf{Y}}|\mathbf{S}} \hat{K}_{\mathbf{Y}} \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}\right)\left(\hat{w}^T_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \hat{K}_{\mathbf{A}\cup\mathbf{W}} \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}\right) \\ &+ \left(\hat{w}^T_{\hat{\mathbf{Y}}|\mathbf{S}} \hat{K}_{\hat{\mathbf{Y}}} \hat{w}_{\hat{\mathbf{Y}}|\mathbf{S}}\right)\left(\hat{w}^T_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \hat{K}_{\mathbf{A}\cup\mathbf{W}} \hat{w}_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}}\right) ,\end{aligned} \tag{2}$$

where $\odot$ is element-wise multiplication. The functions $\hat{w}_{\hat{\mathbf{Y}}|\mathbf{S}} \equiv \hat{w}_{\hat{\mathbf{Y}}|\mathbf{S}}(\cdot)$, $\hat{w}_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \equiv \hat{w}_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}}(\cdot)$, and $\hat{w}_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \equiv \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}(\cdot)$ are found via kernel ridge regression. Caponnetto & Vito (2007) provide the convergence rates of the estimand $\hat{H}^2_{\hat{\mathbf{Y}},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}$ under mild conditions. In practice, computing the HSCIC approximation by the formula in Eq. (2) can be computationally expensive. We provide runtimes for some of our experimental setting showing that including the HSCIC term can slow down training anywhere between 2x and several 100x in Appendix F.7. To speed it up, we can use random Fourier features to approximate $\hat{K}_{\hat{\mathbf{Y}}}$ and $\hat{K}_{\mathbf{A}\cup\mathbf{W}}$ (Rahimi & Recht, 2007; Avron et al., 2017). The details on how to substantially speed up the computations with such approximations are described in Appendix E and we describe the improvements in terms of computational complexity in Appendix F.7. We emphasize that Eq. (2) allows us to consistently estimate the HSCIC *from observational i.i.d. samples, without prior knowledge of the counterfactual distributions.*

3.5 Measuring counterfactual invariance.

Besides predictive performance, e.g., mean squared error (MSE) for regression or accuracy for classification, our key metric of interest is the level of counterfactual invariance achieved by the predictor $\hat{\mathbf{Y}}$. First, we emphasize again that counterfactual distributions are generally not identified from the observational distribution (i.e., from available data) meaning that *CI is generally untestable in practice* from observational data. We can thus only evaluate CI in (semi-)synthetic settings where we have access to the full SCM and thus all counterfactual distributions.

A measure for CI must capture how the distribution of $\hat{\mathbf{Y}}^*_{\mathbf{a}'}$ changes for different values of $\mathbf{a}'$ across all conditioning values $\mathbf{w}$ (which may include an observed value $\mathbf{A} = \mathbf{a}$). We propose the **V**ariance of **C**ounter**F**actuals (VCF) as a metric of CI

$$\mathrm{VCF}(\hat{\mathbf{Y}}) \coloneqq \mathbb{E}_{\mathbf{w}\sim\mathbb{P}_{\mathbf{W}}}\left[\mathrm{var}_{\mathbf{a}'\sim\mathbb{P}_{\mathbf{A}}}\left[\mathbb{E}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}|\mathbf{W}=\mathbf{w}}[\hat{\mathbf{y}}]\right]\right] . \tag{3}$$

That is, we quantify how the average outcome varies with the interventional value $\mathbf{a}'$ at conditioning value $\mathbf{w}$ and average this variance over $\mathbf{w}$. For deterministic predictors (point estimators), which we use in all our experiments, the prediction is a fixed value for each input $\mathbb{E}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}|\mathbf{W}=\mathbf{w}}[\hat{\mathbf{y}}] = \hat{\mathbf{y}}$) and we can drop the inner expectation of Eq. (3). In this case, the variance term in Eq. (3) is zero if and only if $\mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}}|\mathbf{W}=\mathbf{w}}(\mathbf{y}) = \mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}|\mathbf{W}=\mathbf{w}}(\mathbf{y})$ almost surely. Since the variance is non-negative, the outer expectation is zero if and only if the variance term is zero almost surely, yielding the following result.

Corollary 3.6. *For point-estimators, $\hat{\mathbf{Y}}$ is counterfactually invariant in $\mathbf{A}$ w.r.t. $\mathbf{W}$ if and only if $VCF(\hat{\mathbf{Y}}) = 0$ almost surely.*

Estimating VCF in practice requires access to the DGP to generate counterfactuals. Given d i.i.d. examples $(\mathbf{w}_i)_{i=1}^d$ from a fixed observational dataset we sample k intervention values from the marginal $\mathbb{P}_{\mathbf{A}}$ and

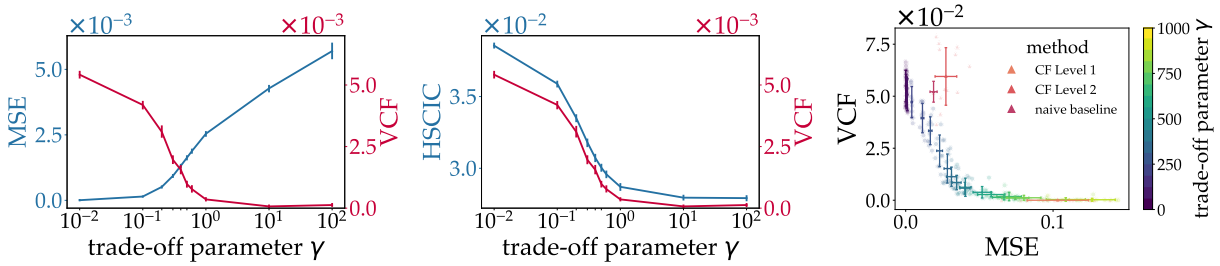

Figure 2: Results on synthetic data (see Appendix F.1 and Appendix F.2). **Left:** trade-off between MSE and counterfactual invariance (VCF). **Middle:** strong correspondence between HSCIC and VCF. **Right:** performance of CIP against baselines CF1 and CF2 and the naive baseline. As γ increases, CIP traces out a frontier characterizing the trade-off between MSE and CI. CF2 and the naive baseline are Pareto-dominated by CIP, i.e., we can pick γ to outperform CF2 in both MSE and VCF simultaneously. CF1 has zero VCF by design, but worse predictive performance than CIP at near zero VCF. Error bars are standard errors over 10 seeds.

compute corresponding predictions. The inner expectation is simply the deterministic predictor output, and we compute the empirical expectation over the d observed $\mathbf{w}$ values and empirical variances over the k sampled interventional values (for each $\mathbf{w}$). Since the required counterfactuals are by their very nature unavailable in practice, our analysis of VCF is limited to (semi-)synthetic settings. Notably, the proposed procedure for choosing γ does not require VCF. Our experiments corroborate the weak convergence property of HSCIC—small HSCIC implies small VCF. Hence, HSCIC may serve as a strong proxy for VCF and thus CI in practice.

4 Experiments

Baselines. As many existing methods focus on cruder purely observational or interventional invariances (see Section 2.2), our choice of baselines for true counterfactual invariance is highly limited. First, we compare CIP to Veitch et al. (2021) in their two limited settings (Fig. 6(b-c)) in Appendix G.1, showing that our method performs on par with theirs. Next, we compare to two methods proposed by Kusner et al. (2017) in settings where they apply. CF1 (their 'Level 1') consists of only using non-descendants of $\mathbf{A}$ as inputs to $f_{\hat{\mathbf{Y}}}$. CF2 (their 'Level 2') assumes an additive noise model and uses the residuals of descendants of $\mathbf{A}$ after regression on $\mathbf{A}$ together with non-descendants of $\mathbf{A}$ as inputs to $f_{\hat{\mathbf{Y}}}$. We refer to these two baselines as CF1 and CF2 respectively. We also compare CIP to the 'naive baseline' which consists in training a predictor ignoring $\mathbf{A}$. In settings where $\mathbf{A}$ is binary, we also compare to Chiappa (2019), devised for path-wise counterfactual fairness. Finally, we develop heuristics based on data augmentation as further possible baselines in Appendix G.2.

4.1 Synthetic experiments

First, we generate various synthetic datasets following the causal graph in Fig. 1(d). They contain (i) the prediction targets $\mathbf{Y}$, (ii) variable(s) we want to be CI in $\mathbf{A}$, (iii) covariates $\mathbf{L}$ mediating effects from $\mathbf{A}$ on $\mathbf{Y}$, and (iv) confounding variables $\mathbf{S}$. The goal is to learn a predictor $\hat{\mathbf{Y}}$ that is CI in $\mathbf{A}$ w.r.t. $\mathbf{W} := \mathbf{A} \cup \mathbf{L} \cup \mathbf{S}$. The datasets cover different dimensions for the observed variables and their correlations and are described in detail in Appendix F.

Model choices and parameters. For all synthetic experiments, we train fully connected neural networks (MLPs) with MSE loss $\mathcal{L}_{\text{MSE}}(\hat{\mathbf{Y}})$ as the predictive loss $\mathcal{L}$ in Eq. (1) for continuous outcomes $\mathbf{Y}$. We generate 10k samples from the observational distribution in each setting and use an 80 to 20 train-test split. All metrics reported are on the test set. We perform hyper-parameter tuning for MLP hyperparameters based on a random strategy (see Appendix F for details). The $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S})$ term is computed as in Eq. (2) using a Gaussian kernel with amplitude 1.0 and length scale 0.1. The regularization parameter λ for the ridge regression coefficients is set to $\lambda = 0.01$. We set $d = 1000$ and $k = 500$ in the estimation of VCF.

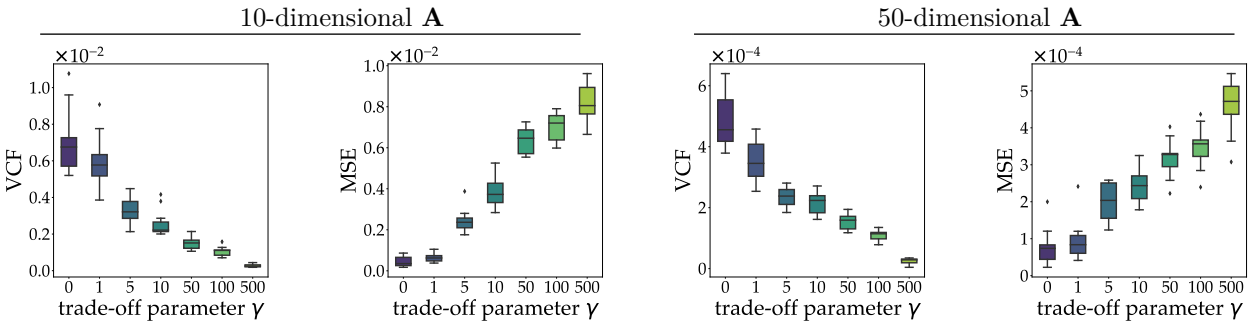

Figure 3: MSE and VCF for synthetic data (Appendix F.3) with 10- and 50-dimensional $\mathbf{A}$ for different γ and 15 random seeds per box. CIP reliably achieves CI as γ increases.

Model performance. We first study the effect of the HSCIC on accuracy and counterfactual invariance on the simulated dataset in Appendix F.1. Fig. 2 (left) depicts the expected trade-off between MSE and VCF for varying γ, whereas Fig. 2 (middle) highlights that HSCIC (estimable from observational data) is a strong proxy of counterfactual invariance measured by VCF (see discussion after Eq. (3)). Figure 2 (right) compares CIP to baselines for a simulated non-additive noise model in Appendix F.2. For a suitable choice of γ, CIP outperforms the baseline CF2 and the naive baseline in both MSE and VCF simultaneously. While CF1 achieves perfect CI by construction (VCF $= 0$), its MSE is higher than CIP at almost perfect CI (VCF near zero). To conclude, our method flexibly and reliably trades predictive performance for counterfactual invariance via a single parameter γ and Pareto-dominates existing methods. In Appendix F.2 we present extensive results on further simulated settings and compare CIP to other heuristic methods in Appendix G.2.

Effect of dimensionality of A. A key advantage of CIP is that it can deal with multi-dimensional $\mathbf{A}$. We consider simulated datasets described in Appendix F.3, where we gradually increase the dimension of $\mathbf{A}$. The results in Fig. 3 for different trade-off parameters γ and different dimensions of $\mathbf{A}$ demonstrate that CIP effectively enforces CI also for multi-dimensional $\mathbf{A}$.[9] Further results are shown in Appendix F.3.

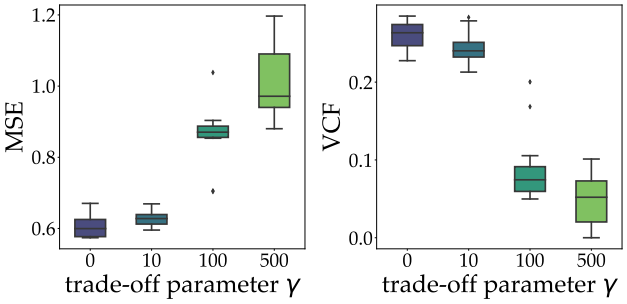

Figure 4: On the dSprites image dataset, CIP trades off MSE for VCF and achieves almost full CI as γ increases. Boxes are for 8 random seeds.

4.2 Image experiments

We consider an image classification task on the dSprites dataset (Matthey et al., 2017), with a causal model as depicted in Fig. 1(f). The full structural equations are provided in Appendix F.4. This experiment is particularly challenging due to the mixed categorical and continuous variables in $\mathbf{C}$ (`shape`, `y-pos`) and $\mathbf{L}$ (`color`, `orientation`), with continuous $\mathbf{A}$ (`x-pos`). We seek a predictor $\hat{\mathbf{Y}}$ that is CI in the x-position w.r.t. all other observed variables. Following Theorem 3.2, we achieve $\hat{\mathbf{Y}} \perp\!\!\!\perp \{\texttt{x-pos}, \texttt{scale}, \texttt{color}, \texttt{orientation}\} \mid \{\texttt{shape}, \texttt{y-pos}\}$ via the HSCIC operator. To accommodate the mixed input types, we first extract features from the images via a CNN and from other inputs via an MLP. We then use an MLP on all concatenated features for $\hat{\mathbf{Y}}$. Fig. 4 shows that CIP gradually enforces CI as γ increases and illustrates the inevitable increase of MSE.

[9] In all boxplots, boxes represent the interquartile range, the horizontal line is the median, and whiskers show minimum and maximum values, excluding outliers (dots).

4.3 Fairness with continuous protected attributes

Finally, we apply CIP to the widely-used UCI Adult dataset (Kohavi & Becker, 1996) and compare it against a 'naive baseline' which simply ignores $\mathbf{A}$, CF1, CF2, and path-specific counterfactual fairness (PSCF) (Chiappa, 2019). Like CF1, PSCF always achieves VCF $= 0$ by design. We explicitly acknowledge the shortcomings of the UCI Adult dataset to reason about social justice (Ding et al., 2021). Instead, we chose it due to previous investigations into plausible causal structures based on domain knowledge (Zhang et al., 2017). The task is to predict whether an individual's income is above a threshold based on demographic information, including protected attributes. We follow Nabi & Shpitser (2018); Chiappa (2019), where a causal structure is assumed for a subset of the variables as in Fig. 1(e) (see Appendix F.5 and Fig. 8 for details). We choose gender and age as the protected attributes $\mathbf{A}$, collect marital status, level of education, occupation, working hours per week, and work class into $\mathbf{L}$, and combine the remaining observed attributes in $\mathbf{C}$. Our aim is to learn a predictor $\hat{\mathbf{Y}}$ that is CI in $\mathbf{A}$ w.r.t. $\mathbf{W} = \mathbf{C} \cup \mathbf{L}$. Achieving (causal) fairness for (mixed categorical and) continuous protected attributes is under active investigation (Mary et al., 2019; Chiappa & Pacchiano, 2021), but directly supported by CIP.

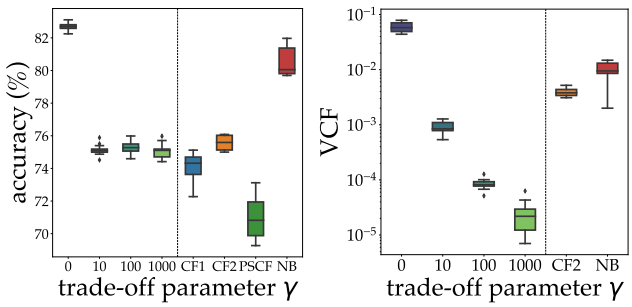

Figure 5: Accuracy and VCF on the Adult dataset. CIP achieves better VCF than CF2 and the naive baseline (NB), improved in accuracy compared to PSCF and is on par with CF1 in accuracy at VCF ≈ 0.

We use an MLP with binary cross-entropy loss for $\hat{\mathbf{Y}}$. Since this experiment is based on real data, the true counterfactual distribution cannot be known. Following Chiappa & Pacchiano (2021) we estimate a possible SCM by inferring the posterior distribution over the unobserved variables using variational autoencoders (Kingma & Welling, 2014). Figure 5 shows that CIP gradually achieves CI and even manages to keep a constant accuracy after an initial drop. It Pareto-dominates CF2 and PSCF and achieves comparable accuracy to CF1 when reaching VCF ≈ 0. While the naive baseline is more accurate for $\gamma \geq 5$ than CIP, CIP can achieve better VCF.

5 Discussion and future work

We developed CIP, a method to learn counterfactually invariant predictors $\hat{\mathbf{Y}}$. First, we presented a sufficient graphical criterion to characterize counterfactual invariance and reduced it to conditional independence in the observational distribution under an injectivity assumption of a causal mechanism. We then built on kernel mean embeddings and the Hilbert-Schmidt Conditional Independence Criterion to devise an efficiently estimable, differentiable, model-agnostic objective to train CI predictors for mixed continuous/categorical, multi-dimensional variables. We demonstrated the efficacy of CIP in extensive empirical evaluations on various regression and classification tasks.

A key limitation of our work, shared by all studies in this domain, is the assumption that the causal graph is known. Guaranteeing CI necessarily requires strong untestable additional assumptions, but we demonstrated that CIP performs well empirically even when these are violated. Moreover, the computational cost of estimating HSCIC may limit the scalability of CIP to very high-dimensional settings even when using efficient random Fourier features. While the increased robustness of counterfactually invariant predictors are certainly desirable in many contexts, this presupposes the validity of our assumptions. Thus, an important direction for future work is to assess the sensitivity of CIP to misspecifications of the causal graph or insufficient knowledge of the required adjustment set. Lastly, we envision our graphical criterion and KME-based objective to be useful also for causal representation learning to isolate causally relevant, autonomous factors underlying the data.

References

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

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, volume 70, pp. 253–262, 2017.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pp. 507–556. 2022.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *The Journal of Machine Learning Research*, 21:90–1, 2020.

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

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5:892837, 2022.

Nancy Cartwright. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press, 2007.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Silvia Chiappa and Aldo Pacchiano. Fairness with continuous optimal transport. *arXiv preprint arXiv:2101.02084*, 2021.

Erhan Çinlar. *Probability and stochastics*, volume 261. Springer, 2011.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Nicolae Dinculeanu. *Vector integration and stochastic integration in Banach spaces*, volume 48. John Wiley & Sons, 2000.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Cian Dorr. Against counterfactual miracles. *The Philosophical Review*, 125(2):241–286, 2016.

Petros Drineas and Michael W. Mahoney. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(72):2153–2175, 2005. URL http://jmlr.org/papers/v6/drineas05a.html.

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. Fairness under feature exemptions: Counterfactual and observational measures. *IEEE Transactions on Information Theory*, 67(10):6675–6710, 2021.

Jake Fawkes and Robin J Evans. Results on counterfactual invariance. *arXiv preprint arXiv:2307.08519*, 2023.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, volume 20, 2007.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. In *Proceedings of the 2022 International Conference on Management of Data*, pp. 276–285, 2022.

Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, volume 3734, pp. 63–77, 2005.

Steffen Grünewälder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *International conference on machine learning*, 2012.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

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

Cho-Jui Hsieh, Si Si, and Inderjit S Dhillon. Fast prediction for large-scale kernel machines. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/de03beffeed9da5f3639a621bcab5dd4-Paper.pdf.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems*, volume 30, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. URL http://arxiv.org/abs/1312.6114.

Ronny Kohavi and Barry Becker. Uci adult data set. *UCI Meachine Learning Repository*, 1996.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 739–766. PMLR, 2022.

Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, volume 97, pp. 4382–4391. PMLR, 2019.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

S. Chandra Mouli and Bruno Ribeiro. Asymmetry learning for counterfactually-invariant classification in OOD tasks. In *Proc. of ICLR*, 2022.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2): 1–141, 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, pp. 21247–21259, 2020.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

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.

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

Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J Sutherland, Victor Veitch, and Arthur Gretton. Efficient conditionally invariant representation learning. *arXiv preprint arXiv:2212.08645*, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2007.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Ilya Shpitser and Judea Pearl. Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence,*, pp. 514–521, 2009.

Ilya Shpitser, Tyler J. VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence,*, pp. 527–536, 2010.

Si Si, Cho-Jui Hsieh, and Inderjit S. Dhillon. Memory efficient kernel approximation. *Journal of Machine Learning Research*, 18(20):1–32, 2017. URL http://jmlr.org/papers/v18/15-025.html.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44): 1–29, 2018.

Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *under review*, 2020.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Le Song, Jonathan Huang, Alexander J. Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*, volume 382, pp. 961–968, 2009.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30 (4):98–111, 2013.

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.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:233:1–233:29, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, pp. 16196–16208, 2021.

Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. *arXiv preprint arXiv:2211.00472*, 2022.

James Woodward. *Causation with a human face: Normative theory and descriptive psychology.* Oxford University Press, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268, 2020. URL https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf.

Mingyang Yi, Ruoyu Wang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Breaking correlation shift via conditional invariant regularizer. In *International Conference on Learning Representations*, 2022.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in data release. In *International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344, 2017.

A Proof of Theorem 3.2

To provide a mostly self-contained proof, we start by stating the necessary definitions and repeat required known results from the literature. Appendices A.1 to A.3 contain these preliminary definitions and results, whereas the final statement and proof provided in Appendix A.4 is our original contribution (making use of the preliminaries stated before).

A.1 Overview of the proof techniques

We restate the main theorem for completeness.

Theorem 3.2. *Let $\mathcal{G}$ be a causal graph, $\mathbf{A}$, $\mathbf{W}$ be two (not necessarily disjoint) sets of nodes in $\mathcal{G}$, such that $(\mathbf{A} \cup \mathbf{W}) \cap \mathbf{Y} = \emptyset$, let $\mathbf{S}$ be a valid adjustment set for $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$. Further, for any random variable $X \in \mathbf{W} \setminus \mathbf{A}$ denote with $g_X(\mathsf{pa}(X), U_X)$ its structural equation, and suppose that $\mathsf{pa}(X) \subseteq \mathbf{A} \cup \mathbf{W}$. Suppose that g_X is injective in the variable $\mathbf{u}$.[10] Then, in all SCMs compatible with $\mathcal{G}$, if a predictor $\hat{\mathbf{Y}}$ satisfies $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$, then $\hat{\mathbf{Y}}$ is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W}$.*

Our proof technique generalizes the work of Shpitser & Pearl (2009). To understand the proof technique, note that conditional counterfactual distributions of the form $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w})$ involve quantities *from two different worlds*. The variables $\mathbf{W}$ belong to the pre-interventional world, and the interventional variable $\mathbf{Y}_{\mathbf{a}}$ belongs to the world after performing the intervention $\mathbf{A} \leftarrow \mathbf{a}$. The variable $\mathbf{Y}^*_{\mathbf{a}|\mathbf{W}}$ refers to the counterfactual of $\mathbf{Y}_{\mathbf{a}}$ after conditioning in the pre-interventional world on $\mathbf{W}$.

A.2 Identifiability of counterfactual distributions

In this section, we discuss a well-known criterion for the identifiability of conditional distributions, which we will then use to prove Theorem 3.2. To this end, we use the notions of a blocked path and valid adjustment set, which we restate for clarity.

Definition A.1. *Consider a path π of causal graph $\mathcal{G}$. A set of nodes $\mathbf{Z}$ blocks π, if π contains a triple of consecutive nodes connected in one of the following ways: $N_i \to Z \to N_j$, $N_i \leftarrow Z \to N_j$, with $N_i, N_j \notin \mathbf{Z}$, $Z \in \mathbf{Z}$, or $N_i \to M \leftarrow N_j$ and neither M nor any descendent of M is in $\mathbf{Z}$.*

Using this definition, we define the concept of a valid adjustment set.

Definition 3.1 (valid adjustment set)**.** *Let $\mathcal{G}$ be a causal graph and let $\mathbf{X}$, $\mathbf{Y}$ be disjoint (sets of) nodes in $\mathcal{G}$. A set of nodes $\mathbf{S}$ is a valid adjustment set for $(\mathbf{X}, \mathbf{Y})$, if (i) no element in $\mathbf{S}$ is a descendant in $\mathcal{G}_{\mathbf{X}}$ of any node $W \notin \mathbf{X}$ which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$, and (ii) $\mathbf{S}$ blocks all non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{G}$.*

Definition 3.1 is a useful graphical criterion for the identifiability of counterfactual distributions. In fact, the following theorem holds:

Theorem A.2 (Theorem 4 by Shpitser et al. (2010))**.** *Let $\mathcal{G}$ be a causal graph, $\mathbf{A}$, $\mathbf{W}$ be two (not necessarily disjoint) sets of nodes in $\mathcal{G}$, such that $(\mathbf{A} \cup \mathbf{W}) \cap \mathbf{Y} = \emptyset$. Define the set $\mathbf{X} = \mathbf{W} \setminus \mathbf{A}$. Suppose that a set of nodes $\mathbf{S}$ satisfies the adjustment criterion relative to $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$ in $\mathcal{G}$. Then, it holds $\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'} \perp\!\!\!\perp \mathbf{A}, \mathbf{X} \mid \mathbf{S}$ for any intervention $\mathbf{A}, \mathbf{X} \leftarrow \mathbf{a}', \mathbf{x}'$.*

A.3 d-separation and conditional independence

In this section, we discuss a well-known criterion for conditional independence, which we will then use to prove Theorem 3.2. We use the notion of a blocked path, as in Definition A.3 and the concept of d-separation as follows.

Definition A.3 (d-Separation)**.** *Consider a causal graph $\mathcal{G}$. Two sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ of $\mathcal{G}$ are said to be d-separated by a third set $\mathbf{S}$ if every path from any node of $\mathbf{X}$ to any node of $\mathbf{Y}$ is blocked by $\mathbf{S}$.*

We use the notation $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{S}$ to indicate that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{S}$ in $\mathcal{G}$. We use Definition A.3 as a graphical criterion for conditional independence (Pearl, 2000).

[10] The injectivity of g_X is defined as follows. Consider two pairs $\{\mathbf{p}, u\}$ and $\{\mathbf{p}, u'\}$ with $\mathbf{p}$ in the support of $\mathsf{pa}(X)$ and u, u' in the support of U_X. Suppose that $\mathbb{P}_{\mathsf{pa}(X),U_X}(\mathbf{p}, u) \neq 0$ and $\mathbb{P}_{\mathsf{pa}(X),U_X}(\mathbf{p}, u') \neq 0$. Then, it holds $g(\mathbf{p}, u) = g(\mathbf{p}, u')$ if and only if $u = u'$.

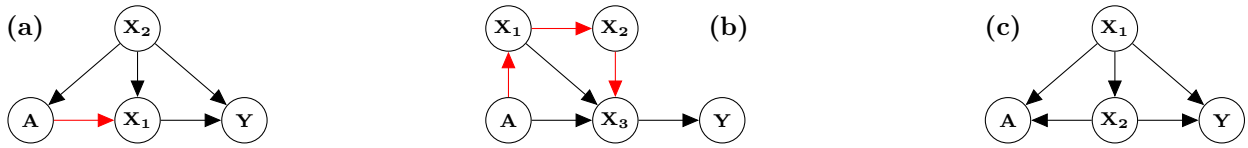

Figure 6: (a) In this causal graph the elements of the set $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}\}$ have depth $\mathsf{depth}(\mathbf{X_1}) = 1$ and $\mathsf{depth}(\mathbf{X_2}) = 0$. (b) In this causal graph the elements of the set $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}\}$ have depth $\mathsf{depth}(\mathbf{X_1}) = 1$, $\mathsf{depth}(\mathbf{X_2}) = 2$, $\mathsf{depth}(\mathbf{X_3}) = 3$. (c) In this causal graph the elements of the set $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}\}$ have depth $\mathsf{depth}(\mathbf{X_1}) = \mathsf{depth}(\mathbf{X_2}) = 0$.

Lemma A.4 (Markov Property)**.** *Consider a causal graph $\mathcal{G}$, and suppose that two sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ of $\mathcal{G}$ are d-separated by $\mathbf{S}$. Then, $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{S}$ in any model induced by the graph $\mathcal{G}$.*

The Markov Property is also referred to as d-separation property. We use the notation $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{S}$ to indicate that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{S}$ in $\mathcal{G}$.

A.4 Proof of Theorem 3.2

Before discussing the proof of Theorem 3.2, we prove an additional auxiliary result. All the following theoretical results are novel and based on the previous sections.

Lemma A.5. *Let $\mathcal{G}$ be a causal graph, $\mathbf{A}$, $\mathbf{X}$ be two disjoint sets of nodes in $\mathcal{G}$. Suppose that any r.v. $X \subseteq \mathbf{X}$ is defined by a structural equation of the form $X = g(\mathsf{pa}(X), \mathbf{U_X})$, with $\mathsf{pa}(X) \subseteq \mathbf{X} \cup \mathbf{A}$ and $\mathbf{U_X}$ exogenous independent noise. Furthermore, suppose that g is injective in $\mathbf{U_X}$. Then, for any observational values $\mathbf{a}, \mathbf{x}$ with $\mathbb{P}_{\mathbf{X},\mathbf{A}}(\mathbf{x}, \mathbf{a}) \neq 0$, there exist a value $\mathbf{u}$ in the support of $\mathbf{U_X}$ such that $\mathbb{P}_{\mathbf{U_X}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{u}) = \mathbb{1}_{\mathbf{u}}$. Here, $\mathbb{1}_{\mathbf{u}}$ is the Dirac δ-distribution.*

Proof. In this proof, we use the following additional notation. Consider a set of nodes $\mathbf{T} \subseteq \mathbf{A} \cup \mathbf{X}$, and let $\mathbf{a}$ be a point in the support of $\mathbf{A}$. We denote with $\mathbf{a_T}$ the restriction of $\mathbf{a}$ to $\mathbf{A} \cap \mathbf{T}$. Similarly, for a point $\mathbf{x}$ in the support of $\mathbf{X}$, we denote with $\mathbf{x_T}$ the restriction of $\mathbf{a}$ to $\mathbf{X} \cap \mathbf{T}$. Furthermore, given two points $\mathbf{a}$, $\mathbf{x}$ as above, we define $\mathbf{a} \cup \mathbf{x}$ as the only point in the support of $\mathbf{A} \cup \mathbf{X}$ such that $(\mathbf{a} \cup \mathbf{x}) \mid_{\mathbf{A}} = \mathbf{a}$ and $(\mathbf{a} \cup \mathbf{x}) \mid_{\mathbf{X}} = \mathbf{x}$ respectively.

Furthermore, we denote with $\mathbf{X} = g_{\mathbf{X}}(\mathsf{pa}(\mathbf{X}), \mathbf{U_X})$ the structural equation for the joint random variable $\mathbf{X}$. In this equation, $\mathbf{U_X}$ corresponds to the joint exogenous noise. Note that the function $g_X(\mathsf{pa}(X), \mathbf{U_X})$ is injective in $\mathbf{U_X}$, by the injectivity assumption on the structural equation $g_X(\mathsf{pa}(X), \mathbf{U_X})$ for each $X \subseteq \mathbf{X}$. Define the function $h(\mathbf{p})$ over the support of $\mathsf{pa}(\mathbf{X})$ as

$$h(\mathbf{a}, \mathbf{x}) := \begin{cases} \mathbf{u} \text{ such that } g_{\mathbf{X}}((\mathbf{a} \cup \mathbf{x}) \mid_{\mathsf{pa}(\mathbf{X})}, \mathbf{u}) = \mathbf{x} & \text{if } \mathbb{P}_{\mathsf{pa}(\mathbf{X}),\mathbf{U_X}}((\mathbf{a} \cup \mathbf{x}) \mid_{\mathsf{pa}(\mathbf{X})}, \mathbf{u}) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that by the injectivity of $g_{\mathbf{X}}$, the function h as above is well-defined. Since $\mathsf{pa}(\mathbf{X}) \subseteq \mathbf{X} \cup \mathbf{A}$, then it holds $\mathbb{P}_{\mathbf{U_X}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{u}) = \mathbb{1}_{h(\mathbf{a},\mathbf{x})}$. The claim follows, by defining $\mathbf{u} := h(\mathbf{a}, \mathbf{x})$. □

We use lemma A.5 to prove the following result.

Lemma A.6. *Under the assumptions of Lemma A.5, fix a set $\mathbf{T} \subseteq \mathbf{A} \cup \mathbf{X}$. Then, for any intervention $\mathbf{A} \leftarrow \mathbf{a}'$ and observational values $\mathbf{a}, \mathbf{x}$ with $\mathbb{P}_{\mathbf{X},\mathbf{A}}(\mathbf{x}, \mathbf{a}) \neq 0$, there exist a point $\mathbf{t}$ in the support of $\mathbf{T}$ such that $\mathbb{P}_{\mathbf{T}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \mathbb{1}_{\mathbf{t}}$.*

Proof. In this proof, we use the notation introduced in Lemma A.5. Consider a set of nodes $\mathbf{T} \subseteq \mathbf{A} \cup \mathbf{X}$, and let $\mathbf{a}$ be a point in the support of $\mathbf{A}$. We denote with $\mathbf{a_T}$ the restriction of $\mathbf{a}$ to $\mathbf{A} \cap \mathbf{T}$. We prove the claim with a induction argument. To this end, we define $\mathsf{depth}(\mathbf{T})$ as the length of the longest directed path π in $\mathcal{G}$ from any node in $\mathbf{A} \cup \mathbf{U_T}$ to any node in $\mathbf{T}$ (see Figure 6).[11]

[11] Note that since $\mathsf{pa}(X) \in \mathbf{X} \cup \mathbf{A}$ for all $X \in \mathbf{T}$, then any such path π only consists of nodes that are in $\mathbf{X} \cup \mathbf{A} \cup \mathbf{U_T}$.

(Base case). We first assume that with $\mathsf{depth}(\mathbf{T}) = 0$. We further assume w.l.o.g. that $\mathbf{T} \cap \mathbf{A} = \emptyset$, since the variables $\mathbf{T} \cap \mathbf{A}$ are fixed by the intervention $\mathbf{A} \leftarrow \mathbf{a}'$. Denote with $\mathbf{T} = g_{\mathbf{T}}(\mathsf{pa}(\mathbf{T}), \mathbf{U}_{\mathbf{T}})$ the structural equation for the joint random variable $\mathbf{T}$. Since $\mathsf{depth}(\mathbf{T}) = 0$, it holds $\mathsf{pa}(\mathbf{T}) \subseteq \mathbf{A}$.[12] Then, the random variable $\mathbf{T}^*_{\mathbf{a}'}$ is defined by the formula $\mathbf{T}^*_{\mathbf{a}'} = g_{\mathbf{T}}(\mathbf{a}' |_{\mathbf{T}}, \mathbf{U}_{\mathbf{T}})$. By Lemma A.5, there exist a value $\mathbf{u}'$ in the support of $\mathbf{U}_{\mathbf{T}}$ such that $\mathbb{P}_{\mathbf{U}_{\mathbf{T}}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{u}) = \mathbb{1}_{\mathbf{u}'}$. It follows that $\mathbb{P}_{\mathbf{T}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \mathbb{1}_{\mathbf{t}'}$ with $\mathbf{t} := g_{\mathbf{T}}(\mathbf{a}' |_{\mathbf{T}}, \mathbf{u}')$.

(Inductive step). We assume by induction that the claim holds for all $\mathbf{T}$ such that $\mathsf{depth}(\mathbf{T}) \leq k$, and we prove the claim for a set $\mathbf{T}$ of depth $\mathsf{depth}(\mathbf{T}) = k + 1$. Since $\mathsf{depth}(\mathsf{pa}(\mathbf{T})) < \mathsf{depth}(\mathbf{T}) = k + 1$, by the inductive hypothesis, it holds

$$\mathbb{P}_{\mathsf{pa}(\mathbf{T})^*_{a'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{x}) = \mathbb{1}_{\mathbf{t}'}, \tag{4}$$

for a point $\mathbf{t}'$ in the support of $\mathsf{pa}(\mathbf{T})$. Again, consider the structural equation $\mathbf{T} = g_{\mathbf{T}}(\mathsf{pa}(\mathbf{T}), \mathbf{U}_{\mathbf{T}})$ for the joint random variable $\mathbf{T}$. In this case, the random variable $\mathbf{T}^*_{a'}$ is defined by the formula $\mathbf{T}^*_{\mathbf{a}'} = g_{\mathbf{T}}(\mathsf{pa}(\mathbf{T})^*_{a'}, \mathbf{U}_{\mathbf{T}})$. By Lemma A.5, there exist a value $\mathbf{u}'$ in the support of $\mathbf{U}_{\mathbf{T}}$ such that $\mathbb{P}_{\mathbf{U}_{\mathbf{T}}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{u}) = \mathbb{1}_{\mathbf{u}'}$. Combining this observation with Eq. (4), we get $\mathbb{P}_{\mathbf{T}^*_{a'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \mathbb{1}_{g_{\mathbf{T}}(\mathbf{t}',\mathbf{u}')}$. The claim follows by defining $\mathbf{t} := g_{\mathbf{T}}(\mathbf{t}', \mathbf{u}')$. □

We use the results above to prove a third auxiliary lemma.

Lemma A.7. *Under the assumptions of Lemma A.5 and Lemma A.6, for any intervention $\mathbf{A} \leftarrow \mathbf{a}'$ and observational values $\mathbf{a}, \mathbf{x}$, there exist an intervention $\mathbf{X} \leftarrow \mathbf{x}'$ such that $\mathbb{P}_{\mathbf{Y}_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) = \mathbb{P}_{\mathbf{Y}_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y})$.*

Proof. First note that by lemma A.6 there exists a point $\mathbf{x}'$ such that $\mathbb{P}_{\mathbf{X}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) = \mathbb{1}_{\mathbf{x}'}$. We choose this point to define the intervention $\mathbf{X} \leftarrow \mathbf{x}'$ as in the statement of this lemma. Denote with $\mathcal{G}_{\mathbf{a}'}$ the causal graph after performing the intervention $\mathbf{A} \leftarrow \mathbf{a}'$. Furthermore, denote with $\mathcal{G}_{\mathbf{a}',\mathbf{x}'}$ the causal graph after performing the interventions $\mathbf{A} \leftarrow \mathbf{a}'$ and $\mathbf{X} \leftarrow \mathbf{x}'$. We first prove that for any node $\mathbf{V}$ of $\mathcal{G}_{\mathbf{a}'}$ and $\mathcal{G}_{\mathbf{a}',\mathbf{x}'}$ it holds

$$\mathbb{P}_{\mathbf{V}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{v}) = \mathbb{P}_{\mathbf{V}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{X}^*_{\mathbf{a}'}=\mathbf{x}'}(\mathbf{v}). \tag{5}$$

In fact, it holds

$$\begin{aligned}
\mathbb{P}_{\mathbf{V}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{v}) &= \int \mathbb{P}_{\mathbf{V}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{X}^*_{\mathbf{a}'}=\mathbf{x}'}(\mathbf{v}) d\mathbb{P}_{\mathbf{X}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{x}') && \text{(by marginalization)} \\
&= \int \mathbb{P}_{\mathbf{V}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{X}^*_{\mathbf{a}'}=\mathbf{x}'}(\mathbf{v}) d\mathbb{1}_{\mathbf{x}'} && \text{(by lemma A.6 with } \mathbf{T} = \mathbf{X}) \\
&= \mathbb{P}_{\mathbf{V}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{X}^*_{\mathbf{a}'}=\mathbf{x}'}(\mathbf{v}). && \text{(by marginalization)}
\end{aligned}$$

Hence, Eq. (5) holds. It follows that

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) &= \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{X}^*_{\mathbf{a}'}=\mathbf{x}'}(\mathbf{y}) && \text{(Eq. (5) with } \mathbf{V} = \mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}) \\
&= \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{X}^*_{\mathbf{a}'}=\mathbf{x}'}(\mathbf{y}) && \text{(Axiom of Consistency (Pearl, 2000))} \\
&= \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}). && \text{(Eq. (5) with } \mathbf{V} = \mathbf{Y}^*_{\mathbf{a}'})
\end{aligned}$$

Hence, the claim holds. □

We now prove our main result.

Proof of Theorem 3.2. Following the notation of Theorem A.2, define the set $\mathbf{X} = \mathbf{W} \setminus \mathbf{A}$. Note that, using this notation, the assumption $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}, \mathbf{W} \mid \mathbf{S}$ can be written as $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}, \mathbf{X} \mid \mathbf{S}$. Suppose that it holds

$$\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) = \int \mathbb{P}_{\mathbf{Y}|\mathbf{A}=\mathbf{a}',\mathbf{X}=\mathbf{x}',\mathbf{S}=\mathbf{s}}(\mathbf{y}) d\mathbb{P}_{\mathbf{S}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{s}) \tag{6}$$

[12] Here, $\mathsf{pa}(\mathbf{T})$ can also be the empty set. In this case, $\mathbf{T} = \mathbf{U}_{\mathbf{T}}$ is an exogenous random variable and the proof follows.

for any intervention $\mathbf{A}, \mathbf{X} \leftarrow \mathbf{a}', \mathbf{x}'$, and for any possible value $\mathbf{w}$ attained by $\mathbf{W}$. Assuming that Eq. (6) holds, we have that

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) &= \int \mathbb{P}_{\mathbf{Y}|\mathbf{A}=\mathbf{a}',\mathbf{X}=\mathbf{x}',\mathbf{S}=\mathbf{s}}(\mathbf{y}) d\mathbb{P}_{\mathbf{S}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{s}) && \text{(assuming Eq. (6))} \\
&= \int \mathbb{P}_{\mathbf{Y}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}'',\mathbf{S}=\mathbf{s}}(\mathbf{y}) d\mathbb{P}_{\mathbf{S}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{s}) && (\mathbf{Y} \perp\!\!\!\perp \mathbf{A}, \mathbf{X} \mid \mathbf{S}) \\
&= \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a},\mathbf{x}''}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}), && \text{(assuming Eq. (6))} \qquad (7)
\end{aligned}$$

for any value $\mathbf{x}''$ in the support of $\mathbf{X}$. To conclude, define the set $\mathbf{T} = \mathbf{A} \setminus \mathbf{W}$. Note that $\mathbf{W} \cup \mathbf{T} = \mathbf{A} \cup \mathbf{X}$. It follows that

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{W}=\mathbf{w}}(\mathbf{y}) &= \int \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) d\mathbb{P}_{\mathbf{T}|\mathbf{W}=\mathbf{w}}(\mathbf{t}) && \text{(by conditioning)} \\
&= \int \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) d\mathbb{P}_{\mathbf{T}|\mathbf{W}=\mathbf{w}}(\mathbf{t}) && \text{(by Lemma A.7)} \\
&= \int \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a},\mathbf{x}''}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) d\mathbb{P}_{\mathbf{T}|\mathbf{W}=\mathbf{w}}(\mathbf{t}) && \text{(by Eq. (7))} \\
&= \int \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) d\mathbb{P}_{\mathbf{T}|\mathbf{W}=\mathbf{w}}(\mathbf{t}) && \text{(by Lemma A.7)} \\
&= \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}=\mathbf{w}}(\mathbf{y}). && \text{(by unconditioning)} \qquad (8)
\end{aligned}$$

Note that the claim follows follows from Eq. (8). The proof of Theorem 3.2 thus boils down to proving Eq. (6).

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{y}) \\
&= \int \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x},\mathbf{S}=\mathbf{s}}(\mathbf{y}) d\mathbb{P}_{\mathbf{S}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{s}) && \text{(by conditioning)} \\
&= \int \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}'}|\mathbf{A}=\mathbf{a}',\mathbf{X}=\mathbf{x}',\mathbf{S}=\mathbf{s}}(\mathbf{y}) d\mathbb{P}_{\mathbf{S}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{s}) && \text{(by Theorem A.2)} \\
&= \int \mathbb{P}_{\mathbf{Y}^*|\mathbf{A}=\mathbf{a}',\mathbf{X}=\mathbf{x}',\mathbf{S}=\mathbf{s}}(\mathbf{y}) d\mathbb{P}_{\mathbf{S}|\mathbf{A}=\mathbf{a},\mathbf{X}=\mathbf{x}}(\mathbf{s}) && \text{(Axiom of Consistency (Pearl, 2000))}
\end{aligned} \qquad (9)$$

and Eq. (6) follows. □

B Proof of Theorem 3.4

We prove that the HSCIC can be used to promote conditional independence, using a similar technique as Park & Muandet (2020). The following theorem holds.

Theorem 3.4 (Theorem 5.4 by Park & Muandet (2020))**.** *If the kernel k of $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}\cup\mathbf{W}}$ is characteristic*[13]*, $HSCIC(\mathbf{Y}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = 0$ almost surely if and only if $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$.*

Proof. By definition, we can write $\text{HSCIC}(\mathbf{Y}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}} \circ \mathbf{S}$, where $H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}$ is a real-valued deterministic function. Hence, the HSCIC is a real-valued random variable, defined over the same domain $\Omega_{\mathbf{S}}$ of the random variable $\mathbf{X}$.

We first prove that if $\text{HSCIC}(\mathbf{Y}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = 0$ almost surely, then it holds $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$. To this end, consider an event $\Omega' \subseteq \Omega_{\mathbf{X}}$ that occurs almost surely, and such that it holds $(H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{X}} \circ \mathbf{X})(\omega) = 0$ for all

[13]The tensor product kernel k is characteristic if $\mathbb{P}_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}} \mapsto \mathbb{E}_{\mathbf{y},[\mathbf{a},\mathbf{w}]}\left[k(\,\cdot\,, \mathbf{y} \otimes [\mathbf{a}, \mathbf{w}])\right]$ is injective.

$\omega \in \Omega'$. Fix a sample $\omega \in \Omega'$, and consider the corresponding value $\mathbf{s}_\omega = \mathbf{S}(\omega)$, in the support of $\mathbf{S}$. It holds

$$\begin{aligned}
&\int k(\mathbf{y} \otimes [\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{Y}, \mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega} = \mu_{\mathbf{Y}, \mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega && \text{(by definition)} \\
&= \mu_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \otimes \mu_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega && \text{(since } \omega \in \Omega') \\
&= \int k_{\mathbf{Y}}(\mathbf{y}, \ \cdot \ ) d\mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \otimes \int k_{\mathbf{A} \cup \mathbf{W}}([\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega} && \text{(by definition)} \\
&= \int k_{\mathbf{Y}}(\mathbf{y}, \ \cdot \ ) \otimes k_{\mathbf{A} \cup \mathbf{W}}([\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \mathbb{P}_{\mathbf{A} \cup \mathbf{W} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega}, && \text{(by Fubini's Theorem)}
\end{aligned}$$

with $k_{\mathbf{Y}}$ and $k_{\mathbf{A} \cup \mathbf{W}}$ the kernels of $\mathcal{H}_{\mathbf{Y}}$ and $\mathcal{H}_{\mathbf{A} \cup \mathbf{W}}$ respectively. Since the kernel k of the tensor product space $\mathcal{H}_{\mathbf{Y}} \otimes \mathcal{H}_{\mathbf{A} \cup \mathbf{W}}$ is characteristic, then the kernels $k_{\mathbf{Y}}$ and $k_{\mathbf{A} \cup \mathbf{W}}$ are also characteristic. Hence, it holds $\mathbb{P}_{\mathbf{Y}, \mathbf{A} | \mathbf{S} = \mathbf{s}_\omega} = \mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \mathbb{P}_{\mathbf{A} | \mathbf{S} = \mathbf{s}_\omega}$ for all $\omega \in \Omega'$. Since the event Ω' occurs almost surely, then $\mathbb{P}_{\mathbf{Y}, \mathbf{A} | \mathbf{S} = \mathbf{s}_\omega} = \mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \mathbb{P}_{\mathbf{A} | \mathbf{S} = \mathbf{s}_\omega}$ almost surely, that is $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$.

Assume now that $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$. By definition there exists an event $\Omega'' \subseteq \Omega_{\mathbf{S}}$ such that $\mathbb{P}_{\mathbf{Y}, \mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega} = \mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \mathbb{P}_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega}$ for all samples $\omega \in \Omega''$, with $\mathbf{s}_\omega = \mathbf{S}(\omega)$. It holds

$$\begin{aligned}
\mu_{\mathbf{Y}, \mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega} &= \int k(\mathbf{y} \otimes [\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{Y}, \mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega && \text{(by definition)} \\
&= \int k(\mathbf{y} \otimes [\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \mathbb{P}_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega && \text{(since } \omega \in \Omega') \\
&= \int k_{\mathbf{Y}}(\mathbf{y}, \ \cdot \ ) k_{\mathbf{A} \cup \mathbf{W}}([\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \mathbb{P}_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega} && \text{(by definition of } k) \\
&= \int k_{\mathbf{Y}}(\mathbf{y}, \ \cdot \ ) d\mathbb{P}_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \otimes \int k_{\mathbf{A} \cup \mathbf{W}}([\mathbf{a}, \mathbf{w}], \ \cdot \ ) d\mathbb{P}_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega} && \text{(by Fubini's Theorem)} \\
&= \mu_{\mathbf{Y} | \mathbf{S} = \mathbf{s}_\omega} \otimes \mu_{\mathbf{A} \cup \mathbf{W} | \mathbf{S} = \mathbf{s}_\omega}. && \text{(by definition)}
\end{aligned}$$

The claim follows. □

C Conditional kernel mean embeddings and the HSCIC

The notion of conditional kernel mean embeddings has already been studied in the literature. We show that, under stronger assumptions, our definition is equivalent to the definition by Park & Muandet (2020). In this section, without loss of generality we will assume that $\mathbf{W} = \emptyset$ and we will refer to the conditioning set as $\mathbf{Z}$.

C.1 Conditional kernel mean embeddings and conditional independence

We show that, under stronger assumptions, the HSCIC can be defined using the Bochner conditional expected value. The Bochner conditional expected value is defined as follows.

Definition C.1. *Fix two random variables $\mathbf{Y}$, $\mathbf{Z}$ taking value in a Banach space $\mathcal{H}$, and denote with $(\Omega, \mathcal{F}, \mathbb{P})$ their joint probability space. Then, the Bochner conditional expectation of $\mathbf{Y}$ given $\mathbf{Z}$ is any $\mathcal{H}$-valued random variable $\mathbf{X}$ such that*

$$\int_E \mathbf{Y} d\mathbb{P} = \int_E \mathbf{X} d\mathbb{P}$$

for all $E \in \sigma(\mathbf{Z}) \subseteq \mathcal{F}$, with $\sigma(\mathbf{Z})$ the σ-algebra generated by $\mathbf{Z}$. We denote with $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ the Bochner expected value. Any random variable $\mathbf{X}$ as above is a version of $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.

The existence and almost sure uniqueness of the conditional expectation are shown in Dinculeanu (2000). Given a RKHS $\mathcal{H}$ with kernel k over the support of $\mathbf{Y}$, Park & Muandet (2020) define the corresponding conditional kernel mean embedding as

$$\mu_{\mathbf{Y} | \mathbf{Z}} \coloneqq \mathbb{E}[k(\cdot, \mathbf{y}) \mid \mathbf{Z}].$$

Note that, according to this definition, $\mu_{\mathbf{Y} | \mathbf{Z}}$ is an $\mathcal{H}$-valued random variable, not a single point of $\mathcal{H}$. Park & Muandet (2020) use this notion to define the HSCIC as follows.

Definition C.2 (The HSCIC according to Park & Muandet (2020))**.** *Consider (sets of) random variables* $\mathbf{Y}$*,* $\mathbf{A}$*,* $\mathbf{Z}$*, and consider two RKHS* $\mathcal{H}_{\mathbf{Y}}$*,* $\mathcal{H}_{\mathbf{A}}$ *over the support of* $\mathbf{Y}$ *and* $\mathbf{A}$ *respectively. The HSCIC between* $\mathbf{Y}$ *and* $\mathbf{A}$ *given* $\mathbf{Z}$ *is defined as the real-valued random variable*

$$\omega \mapsto \left\| \mu_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\omega) - \mu_{\mathbf{Y}|\mathbf{Z}}(\omega) \otimes \mu_{\mathbf{A}|\mathbf{Z}}(\omega) \right\|,$$

for all samples ω *in the domain* $\Omega_{\mathbf{Z}}$ *of* $\mathbf{Z}$*. Here,* $\|\cdot\|$ *the metric induced by the inner product of the tensor product space* $\mathcal{H}_{\mathbf{Y}} \otimes \mathcal{H}_{\mathbf{Z}}$*.*

We show that, under more restrictive assumptions, Definition C.2 can be used to promote conditional independence. To this end, we use the notion of a regular version.

Definition C.3 (Regular Version, following Definition 2.4 by Çinlar (2011))**.** *Consider two random variables* $\mathbf{Y}$*,* $\mathbf{Z}$*, and consider the induced measurable spaces* $(\Omega_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}})$ *and* $(\Omega_{\mathbf{Z}}, \mathcal{F}_{\mathbf{Z}})$*. A regular version* Q *for* $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}$ *is a mapping* $Q\colon \Omega_{\mathbf{Z}} \times \mathcal{F}_{\mathbf{Y}} \to [0, +\infty]\colon (\omega, \mathbf{y}) \mapsto Q_\omega(\mathbf{y})$ *such that: (i) the map* $\omega \mapsto Q_\omega(\mathbf{x})$ *is* $\mathcal{F}_{\mathbf{A}}$*-measurable for all* $\mathbf{y}$*; (ii) the map* $\mathbf{y} \mapsto Q_\omega(\mathbf{y})$ *is a measure on* $(\Omega_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}})$ *for all* ω*; (iii) the function* $Q_\omega(\mathbf{y})$ *is a version for* $\mathbb{E}\left[\mathbb{1}_{\{\mathbf{Y}=\mathbf{y}\}} \mid \mathbf{Z}\right]$*.*

The following theorem shows that the random variable as in Definition C.2 can be used to promote conditional independence.

Theorem C.4 (Theorem 5.4 by Park & Muandet (2020))**.** *With the notation introduced above, suppose that the kernel* k *of the tensor product space* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}}$ *is characteristic. Furthermore, suppose that* $\mathbb{P}_{\mathbf{Y},\mathbf{A}|\mathbf{X}}$ *admits a regular version. Then,* $\left\| \mu_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\omega) - \mu_{\mathbf{Y}|\mathbf{Z}}(\omega) \otimes \mu_{\mathbf{A}|\mathbf{Z}}(\omega) \right\| = 0$ *almost surely if and only if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$*.*

Note that the assumption of the existence of a regular version is essential in Theorem C.4. In this work, HSCIC is not used for conditional independence testing but as a conditional independence measure.

C.2 Equivalence with our approach

The following theorem shows that under the existence of a regular version, conditional kernel mean embeddings can be defined using the Bochner conditional expected value. To this end, we use the following theorem.

Theorem C.5 (Following Proposition 2.5 by Çinlar (2011))**.** *Following the notation introduced in Definition C.3, suppose that* $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}(\cdot \mid \mathbf{Z})$ *admits a regular version* $Q_\omega(\mathbf{y})$*. Consider a kernel* k *over the support of* $\mathbf{Y}$*. Then, the mapping*

$$\omega \mapsto \int k(\cdot, \mathbf{y}) dQ_\omega(\mathbf{y})$$

is a version of $\mathbb{E}\left[k(\cdot, \mathbf{y}) \mid \mathbf{Z}\right]$*.*

As a consequence of Theorem C.5, we prove the following result.

Lemma C.6. *Fix two random variables* $\mathbf{Y}$*,* $\mathbf{Z}$*. Suppose that* $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}$ *admits a regular version. Denote with* $\Omega_{\mathbf{Z}}$ *the domain of* $\mathbf{Z}$*. Then, there exists a subset* $\Omega \subseteq \Omega_{\mathbf{Z}}$ *that occurs almost surely, such that* $\mu_{\mathbf{Y}|\mathbf{Z}}(\omega) = \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{Z}(\omega)}$ *for all* $\omega \in \Omega$*. Here,* $\mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{Z}(\omega)}$ *is the embedding of conditional measures as in Section 2.*

Proof. Let $Q_\omega(\mathbf{y})$ be a regular version of $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}$. Without loss of generality we may assume that it holds $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \{\mathbf{Z} = \mathbf{Z}(\omega)\}) = Q_\omega(\mathbf{y})$. By Theorem C.5 there exists an event $\Omega \subseteq \Omega_{\mathbf{Z}}$ that occurs almost surely such that

$$\mu_{\mathbf{Y}|\mathbf{Z}}(\omega) = \mathbb{E}[k(\mathbf{y}, \ \cdot\ ) \mid \mathbf{Z}](\omega) = \int k(\mathbf{y}, \ \cdot\ ) dQ_\omega(\mathbf{y}), \tag{10}$$

for all $\omega \in \Omega$. Then, for all $\omega \in \Omega$ it holds

$$\begin{aligned}
\mu_{\mathbf{Y}|\mathbf{Z}}(\omega) &= \int k(\mathbf{x}, \ \cdot\ ) dQ_\omega(\mathbf{x}) && \text{(it follows from Eq. (10))} \\
&= \int k(\mathbf{x}, \ \cdot\ ) d\mathbb{P}_{\mathbf{X}|\mathbf{A}}(\mathbf{x} \mid \{\mathbf{A} = \mathbf{A}(\omega)\}) && (Q_\omega(\mathbf{y}) = \mathbb{P}_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \{\mathbf{Z} = \mathbf{Z}(\omega)\})) \\
&= \mu_{\mathbf{X}|\{\mathbf{A}=\mathbf{A}(\omega)\}}, && \text{(by definition as in Section 2)}
\end{aligned}$$

as claimed. □

As a consequence of Lemma C.6, we can prove that the definition of the HSCIC by Park & Muandet (2020) is equivalent to ours. The following corollary holds.

Corollary C.7. *Consider (sets of) random variables* $\mathbf{Y}$, $\mathbf{A}$, $\mathbf{Z}$, *and consider two RKHS* $\mathcal{H}_{\mathbf{Y}}$, $\mathcal{H}_{\mathbf{A}}$ *over the support of* $\mathbf{Y}$ *and* $\mathbf{A}$ *respectively. Suppose that* $\mathbb{P}_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot \mid \mathbf{Z})$ *admits a regular version. Then, there exists a set* $\Omega \subseteq \Omega_{\mathbf{A}}$ *that occurs almost surely, such that*

$$\left\|\mu_{\mathbf{X},\mathbf{A}|\mathbf{Z}}(\omega) - \mu_{\mathbf{X}|\mathbf{Z}}(\omega) \otimes \mu_{\mathbf{A}|\mathbf{Z}}(\omega)\right\| = (H_{\mathbf{Y},\mathbf{A}|\mathbf{Z}} \circ \mathbf{Z})(\omega).$$

Here, $H_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}$ *is a real-valued deterministic function, defined as*

$$H_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\mathbf{z}) \coloneqq \left\|\mu_{\mathbf{Y},\mathbf{A}|\mathbf{Z}=\mathbf{z}} - \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}} \otimes \mu_{\mathbf{A}|\mathbf{Z}=\mathbf{z}}\right\|,$$

and $\|\cdot\|$ *is the metric induced by the inner product of the tensor product space* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}}$.

We remark that the assumption of the existence of a regular version is essential in Corollary C.7.

D The cross-covariance operator

In this section, we show that under additional assumptions, our definition of conditional KMEs is equivalent to the definition based on the cross-covariance operator, under more restrictive assumptions. The definition of KMEs based on the cross-covariance operator requires the use of the following well-known result.

Lemma D.1. *Fix two RKHS* $\mathcal{H}_{\mathbf{X}}$ *and* $\mathcal{H}_{\mathbf{Z}}$, *and let* $\{\varphi_i\}_{i=1}^{\infty}$ *and* $\{\psi_j\}_{j=1}^{\infty}$ *be orthonormal bases of* $\mathcal{H}_{\mathbf{X}}$ *and* $\mathcal{H}_{\mathbf{Z}}$ *respectively. Denote with* $\mathsf{HS}(\mathcal{H}_{\mathbf{X}}, \mathcal{H}_{\mathbf{Z}})$ *the set of Hilbert-Schmidt operators between* $\mathcal{H}_{\mathbf{X}}$ *and* $\mathcal{H}_{\mathbf{Z}}$. *There is an isometric isomorphism between the tensor product space* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{Z}}$ *and* $\mathsf{HS}(\mathcal{H}_{\mathbf{X}}, \mathcal{H}_{\mathbf{Z}})$, *given by the map*

$$T\colon \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_{i,j}\varphi_i \otimes \psi_j \mapsto \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_{i,j}\langle\,\cdot\,, \varphi_i\rangle_{\mathcal{H}_{\mathbf{X}}}\psi_j.$$

For proof of this result see i.e., Park & Muandet (2020). This lemma allows us to define the cross-covariance operator between two random variables, using the operator T.

Definition D.2 (Cross-Covariance Oprator)**.** *Consider two random variables* $\mathbf{X}$, $\mathbf{Z}$. *Consider corresponding mean embeddings* $\mu_{\mathbf{X},\mathbf{Z}}$, $\mu_{\mathbf{X}}$ *and* $\mu_{\mathbf{Z}}$, *as defined in Section 3. The cross-covariance operator is defined as* $\Sigma_{\mathbf{X},\mathbf{Z}} \coloneqq T(\mu_{\mathbf{X},\mathbf{Z}} - \mu_{\mathbf{X}} \otimes \mu_{\mathbf{Z}})$. *Here,* T *is the isometric isomorphism as in Lemma D.1.*

It is well-known that the cross-covariance operator can be decomposed into the covariance of the marginals and the correlation. That is, there exists a unique bounded operator $\Lambda_{\mathbf{Y},\mathbf{Z}}$ such that

$$\Sigma_{\mathbf{Y},\mathbf{Z}} = \Sigma_{\mathbf{Y},\mathbf{Y}}^{1/2} \circ \Lambda_{\mathbf{Y},\mathbf{Z}} \circ \Sigma_{\mathbf{Z},\mathbf{Z}}^{1/2}$$

Using this notation, we define the *normalized conditional cross-covariance operator*. Given three random variables $\mathbf{Y}$, $\mathbf{A}$, $\mathbf{Z}$ and corresponding kernel mean embeddings, this operator is defined as

$$\Lambda_{\mathbf{Y},\mathbf{A}|\mathbf{Z}} \coloneqq \Lambda_{\mathbf{Y},\mathbf{A}} - \Lambda_{\mathbf{Y},\mathbf{Z}} \circ \Lambda_{\mathbf{Z},\mathbf{A}}. \tag{11}$$

This operator was introduced by Fukumizu et al. (2007). The normalized conditional cross-covariance can be used to promote statistical independence, as shown in the following theorem.

Theorem D.3 (Theorem 3 by Fukumizu et al. (2007))**.** *Following the notation introduced above, define the random variable* $\ddot{\mathbf{A}} \coloneqq (\mathbf{A}, \mathbf{Z})$. *Let* $\mathbb{P}_{\mathbf{Z}}$ *be the distribution of the random variable* $\mathbf{Z}$, *and denote with* $L^2(\mathbb{P}_{\mathbf{Z}})$ *the space of the square integrable functions with probability* $\mathbb{P}_{\mathbf{Z}}$. *Suppose that the tensor product kernel* $k_{\mathbf{Y}} \otimes k_{\mathbf{A}} \otimes k_{\mathbf{Z}}$ *is characteristic. Furthermore, suppose that* $\mathcal{H}_{\mathbf{Z}} + \mathbb{R}$ *is dense in* $L^2(\mathbb{P}_{\mathbf{Z}})$. *Then, it holds*

$$\Lambda_{\mathbf{Y},\ddot{\mathbf{A}}|\mathbf{Z}} = 0 \quad \textit{if and only if} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}.$$

Here, $\Lambda_{\mathbf{Y},\ddot{\mathbf{A}}|\mathbf{Z}}$ *is an operator defined as in Eq. (11).*

By Theorem D.3, the operator $\Lambda_{\mathbf{Y},\ddot{\mathbf{A}}|\mathbf{Z}}$ can also be used to promote conditional independence. However, CIP is more straightforward since it requires less assumptions. In fact, Theorem D.3 requires to embed the variable $\mathbf{Z}$ in an RKHS. In contrast, CIP only requires the embedding of the variables $\mathbf{Y}$ and $\mathbf{A}$.

E Random fourier features

Random Fourier features is an approach to scaling up kernel methods for shift-invariant kernels (Rahimi & Recht, 2007). Recall that a shift-invariant kernel is a kernel of the form $k(\mathbf{z}, \mathbf{z}') = h_k(\mathbf{z} - \mathbf{z}')$, with h_k a positive definite function.

Fourier features are defined via the following well-known theorem.

Theorem E.1 (Bochner's Theorem)**.** *For every shift-invariant kernel of the form $k(\mathbf{z}, \mathbf{z}') = h_k(\mathbf{z} - \mathbf{z}')$ with $h_k(\mathbf{0}) = 1$, there exists a probability probability density function $\mathbb{P}_k(\eta)$ such that*

$$k(\mathbf{z}, \mathbf{z}') = \int e^{-2\pi i \eta^T (\mathbf{z} - \mathbf{z}')} d\mathbb{P}_k.$$

Since both the kernel k and the probability distribution $\mathbb{P}_k$ are real-valued functions, the integrand in Theorem E.1 ca be replaced by the function $\cos \eta^T (\mathbf{z} - \mathbf{z}')$, and we obtain the following formula

$$k(\mathbf{z}, \mathbf{z}') = \int \cos \eta^T (\mathbf{z} - \mathbf{z}') d\mathbb{P}_k = \mathbb{E}\left[\cos \eta^T (\mathbf{z} - \mathbf{z}')\right], \tag{12}$$

where the expected value is taken with respect to the distribution$\mathbb{P}_k(\eta)$. This equation allows to approximate the kernel $k(\mathbf{z}, \mathbf{z}')$, via the empirical mean of points $\eta_1, \ldots, \eta_l$ sampled independently according to $\mathbb{P}_k$. In fact, it is possible to prove exponentially fast convergence of an empirical estimate for $\mathbb{E}\left[\cos \eta^T (\mathbf{z} - \mathbf{z}')\right]$, as shown in the following theorem.

Theorem E.2 (Uniform Convergence of Fourier Features, Claim 1 by Rahimi & Recht (2007))**.** *Following the notation introduced above, fix any compact subset Ω in the domain of k, and consider points $\eta_1, \ldots, \eta_l$ sampled independent according to the distribution $\mathbb{P}_k$. Define the function*

$$\hat{k}(\mathbf{z}, \mathbf{z}') := \frac{1}{l} \sum_{j=1}^{l} \cos \eta_j^T (\mathbf{z} - \mathbf{z}'),$$

for all $(\mathbf{z}, \mathbf{z}') \in \Omega$. Then, it holds

$$\mathbb{P}\left(\sup_{\mathbf{z}, \mathbf{z}'} \left|\hat{k}(\mathbf{z}, \mathbf{z}') - k(\mathbf{z}, \mathbf{z}')\right| \geq \varepsilon\right) \leq 2^8 \sigma_k \frac{\mathsf{diam}(\Omega)}{\varepsilon} \exp\left\{-\frac{\varepsilon^2 l}{4(d+1)}\right\}.$$

Here σ_k^2 is the second moment of the Fourier transform of the kernel k, and d is the dimension of the arrays $\mathbf{z}$ and $\mathbf{z}'$.

By Theorem E.2, the estimated kernel $\hat{k}$ is a good approximation of the true kernel k on the set Ω.

Similarly, we can approximate the Kernel matrix using Random Fourier features. Following the notation introduced above, define the function

$$\zeta_{k,l}(\mathbf{z}) := \frac{1}{\sqrt{l}} \left[\cos \eta_1^T \mathbf{z}, \ldots, \cos \eta_l^T \mathbf{z}\right] \tag{13}$$

with $\eta_1, \ldots, \eta_l$ sampled independent according to the distribution $\mathbb{P}_k$.

We can approximate the Kernel matrix using the functions defined as in Eq. (13). Consider n samples $\mathbf{z}_1, \ldots, \mathbf{z}_n$, and denote with Z the $n \times l$ matrix whose i-th row is given by $\zeta_{k,l}(\mathbf{z}_i)$. Similarly, denote with Z^* the $l \times n$ matrix whose i-th column is given by $\zeta_{k,l}^*(\mathbf{z}_i)$. Then, we can approximate the kernel matrix as $\hat{K}_{\mathbf{Z}} \approx ZZ^*$.

We can also use this approximation to compute the kernel ridge regression parameters as in Section 3 using the formula $\hat{w}_{\mathbf{Y}|\mathbf{Z}}(\cdot) \approx (ZZ^* - n\lambda I)^{-1} \left[\ k_{\mathbf{Z}}(\cdot, \mathbf{z}_1), \cdots, k_{\mathbf{Z}}(\cdot, \mathbf{z}_n)\ \right]^T$. Avron et al. (2017) argue that the approximate kernel ridge regression, as defined above, is an accurate estimate of the true distribution. Their argument is based on proving that the matrix $ZZ^* - n\lambda I$ is a *good approximation* of $\hat{K}_{\mathbf{Z}} - n\lambda I$. The notion of good approximation is clarified by the following definition.

Definition E.3. *Fix two Hermitian matrices A and B of the same size. We say that a matrix A is a γ-spectral approximation of another matrix B, if it holds $(1-\gamma)B \preceq A \preceq (1+\gamma)B$. Here, the $\preceq$ symbol means that $A-(1-\gamma)B$ is positive semi-definite, and that $(1+\gamma)B - A$ is positive semi-definite.*

Avron et al. (2017) prove that $ZZ^* - n\lambda I$ is a γ- approximation of $\hat{K}_{\mathbf{Z}} - n\varepsilon I$, if the number of samples $\eta_1, \dots, \eta_l$ is sufficiently large.

Theorem E.4 (Theorem 7 by Avron et al. (2017))**.** *Fix a constant $\gamma \leq 1/2$. Consider n samples $\mathbf{z}_1, \dots, \mathbf{z}_n$, and denote with $\hat{K}_{\mathbf{Z}}$ the corresponding kernel matrix. Suppose that it holds $\|\hat{K}_{\mathbf{Z}}\|_2 \geq n\lambda$ for a constant $\lambda > 0$. Fix $\eta_1, \dots, \eta_l$ samples with*

$$l \geq \frac{8}{3\gamma^2\lambda} \ln \frac{16 \ \mathsf{tr}_\lambda(\hat{K}_{\mathbf{Z}})}{\gamma}$$

Then, the matrix $ZZ^* - n\lambda I$ is a γ- approximation of $\hat{K}_{\mathbf{Z}} - n\lambda I$ with probability at least $1-\gamma$, for all $\gamma \in (0,1)$. Here, $\mathsf{tr}_\lambda(\hat{K}_{\mathbf{Z}})$ is defined as the trace of the matrix $\hat{K}_{\mathbf{Z}}(\hat{K}_{\mathbf{Z}} + n\lambda I)^{-1}$.

We conclude this section by illustrating the use of random Fourier features to approximate a simple Gaussian kernel. Suppose that we are given a kernel of the form

$$k(\mathbf{z}, \mathbf{z}') \coloneqq \exp\left\{-\frac{1}{2}\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2\right\}.$$

Then, $k(\mathbf{z}, \mathbf{z}')$ can be estimated as in Theorem E.2, with $\eta_1, \dots, \eta_l \sim \mathcal{N}(0, \Sigma)$, with $\Sigma \coloneqq \sigma^{-1}I$, with I the identity matrix. The functions $\zeta_{k,l}(\mathbf{z})$ can be defined accordingly.

F Additional experiments and settings

This section contains detailed information on the experiments and additional results.

F.1 Dataset for model performance with the use of the HSCIC

The data-generating mechanism corresponding to the results in Fig. 2 is the following:

$$\mathbf{Z} \sim \mathcal{N}(0,1) \qquad \mathbf{A} = \mathbf{Z}^2 + \varepsilon_{\mathbf{A}}$$

$$\mathbf{L} = \exp\left\{-\frac{1}{2}\mathbf{A}^2\right\}\sin(2\mathbf{A}) + 2\mathbf{Z}\frac{1}{5}\varepsilon_{\mathbf{L}}$$

$$\mathbf{Y} = \frac{1}{2}\exp\{-\mathbf{L}\mathbf{Z}\} \cdot \sin(2\mathbf{L}\mathbf{Z}) + 5\mathbf{A} + \frac{1}{5}\varepsilon_{\mathbf{Y}},$$

where $\varepsilon_{\mathbf{A}} \sim \mathcal{N}(0,1)$ and $\varepsilon_{\mathbf{Y}}, \varepsilon_{\mathbf{L}} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1)$.
In the first experiment, Fig. 2 shows the results of feed-forward neural networks consisting of 8 hidden layers with 20 nodes each, connected with a rectified linear activation function (ReLU) and a linear final layer. Mini-batch size of 256 and the Adam optimizer with a learning rate of 10^{-3} for 1000 epochs were used.

F.2 Datasets and results for comparison with baselines

The comparison of our method CIP with the CF1 and CF2 is done on different simulated datasets. These will be referred to as Scenario 1 and Scenario 2. The data generating mechanism corresponding to the results in Fig. 2 (right) is the following:

$$\mathbf{Z} \sim \mathcal{N}(0,1) \qquad \mathbf{A} = \exp\left\{\frac{1}{2}\mathbf{Z}^2\right\} \cdot \sin(2\mathbf{Z}) + \varepsilon_{\mathbf{A}}$$

$$\mathbf{L} = (\mathbf{A} + 0.1\mathbf{Z}) \cdot \varepsilon_{\mathbf{L}}$$

$$\mathbf{Y} = \mathbf{A} + \mathbf{L} + 0.1 \cdot \sin(\mathbf{Z})$$

Table 1: **Performance of the HSCIC against baselines** CF1 and CF2 on two synthetic datasets. Notably, in both scenarios it is possible to select γ values for which CIP outperforms CF2 in **MSE** and **VCF** simultaneously.

	Scenario 1			**Scenario 2**		
	MSE $\times 10^6$	**HSCIC** $\times 10^3$	**VCF** $\times 10^3$	**MSE** $\times 10^3$	**HSCIC** $\times 10^2$	**VCF** $\times 10^2$
$\gamma = 0.001$	12 ± 9	45.38 ± 0.41	54.93 ± 7.50	0.0006	35.64 ± 0.32	5.60 ± 0.03
$\gamma = 0.01$	16 ± 12	45.35 ± 0.41	54.57 ± 7.18	0.0019	35.44 ± 0.33	5.50 ± 0.03
$\gamma = 0.1$	32 ± 20	45.11 ± 0.43	54.16 ± 7.58	0.11 ± 0.006	33.47 ± 0.36	4.46 ± 0.04
$\gamma = 0.2$	81 ± 14	44.78 ± 0.47	53.59 ± 7.90	0.42 ± 0.02	31.38 ± 0.38	3.52 ± 0.04
$\gamma = 0.3$	192 ± 33	43.92 ± 0.52	52.92 ± 7.54	0.82 ± 0.04	29.75 ± 0.34	2.50 ± 0.04
$\gamma = 0.4$	384 ± 58	43.88 ± 0.57	52.06 ± 7.25	1.21 ± 0.05	28.63 ± 0.33	1.79 ± 0.03
$\gamma = 0.5$	685 ± 133	43.26 ± 0.65	51.64 ± 7.40	1.56 ± 0.08	27.81 ± 0.26	1.1 ± 0.01
$\gamma = 0.6$	1117 ± 165	42.47 ± 0.73	50.96 ± 7.36	1.84 ± 0.11	26.87 ± 0.22	0.79 ± 0.01
$\gamma = 0.7$	1655 ± 223	42.11 ± 0.80	50.31 ± 7.44	2.11 ± 0.14	26.08 ± 0.20	0.49 ± 0.01
$\gamma = 0.8$	2225 ± 296	41.87 ± 0.84	49.76 ± 7.25	2.37 ± 0.15	25.27 ± 0.18	0.31 ± 0.01
$\gamma = 0.9$	2832 ± 372	41.52 ± 0.92	49.17 ± 7.41	2.58 ± 0.17	24.64 ± 0.16	0.21 ± 0.01
$\gamma = 1.0$	3472 ± 422	38.37 ± 0.97	48.71 ± 7.55	2.77 ± 0.19	24.21 ± 0.15	0.14 ± 0.01
CF1	10321 ± 72	41.37 ± 0.58	0 ± 0.00	4.59 ± 0.4478	25.01 ± 0.25	0 ± 0.00
CF2	2728 ± 272	41.37 ± 0.92	59.50 ± 10.35	3.97 ± 0.3479	27.03 ± 0.35	2.62 ± 0.81

where $\varepsilon_{\mathbf{A}}, \varepsilon_{\mathbf{L}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\varepsilon_{\mathbf{Y}} \overset{i.i.d.}{\sim} \mathcal{N}(0,0.1)$. This is referred to as Scenario 1. The data generating mechanism for Scenario 2 is the following:

$$\mathbf{Z} \sim \mathcal{N}(0,1) \qquad \mathbf{A} = \exp\left\{\frac{1}{2}\mathbf{Z}^2\right\} \cdot \sin(2\mathbf{Z}) + \varepsilon_{\mathbf{A}}$$

$$\mathbf{L} = \exp\left\{-\frac{1}{2}\mathbf{A}^2\right\} \cdot \varepsilon_{\mathbf{L}} + 2\mathbf{Z}$$

$$\mathbf{Y} = \frac{1}{2}\sin(\mathbf{ZL}) \cdot \exp\{-\mathbf{ZL}\} + \frac{1}{5}\varepsilon_{\mathbf{Y}},$$

where $\varepsilon_{\mathbf{A}}, \varepsilon_{\mathbf{L}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\varepsilon_{\mathbf{Y}} \overset{i.i.d.}{\sim} \mathcal{N}(0,0.1)$. Fig. 2 (right) and Table 1 present the average and standard deviation resulting from 9 random seeds runs. For CIP, the same hyperparameters as in the previous setting are used. The MLPs implemented in CF1 and CF2 used for the prediction of $\hat{\mathbf{Y}}$ and the one used for the prediction of the $\mathbf{L}$ residuals in CF2 are all designed with similar architecture and training method. The MLP models consist of 8 hidden layers with 20 nodes each, connected with a rectified linear activation function (ReLU) and a linear final layer. During training, mini-batch size of 64 and the Adam optimizer with a learning rate of 10^{-3} for 200 epochs were used.

F.3 Datasets and results for multi-dimensional variables experiments

The data-generating mechanisms for the multi-dimensional settings of Fig. 3 are now shown. Given $\dim \mathbf{A} = D_1 \geq 2$, the datasets were generated from:

$$\mathbf{Z} \sim \mathcal{N}(0,1) \qquad \mathbf{A}_i = \mathbf{Z}^2 + \varepsilon^i_{\mathbf{A}} \quad \text{for } i \in \{1, D_1\}$$

$$\mathbf{L} = \exp\left\{-\frac{1}{2}\mathbf{A}_1\right\} + \sum_{i=1}^{D_1} \mathbf{A}_i \cdot \sin(\mathbf{Z}) + 0.1 \cdot \varepsilon_{\mathbf{L}}$$

$$\mathbf{Y} = \exp\left\{-\frac{1}{2}\mathbf{A}_2\right\} \cdot \sum_{i=1}^{D_1} \mathbf{A}_i + \mathbf{LZ} + 0.1 \cdot \varepsilon_{\mathbf{Y}},$$

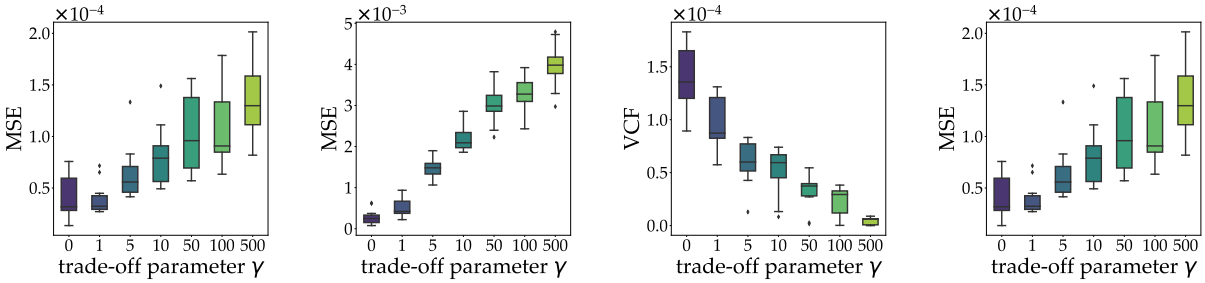

Figure 7: MSE , HSCIC, VCF for increasing dimension of $\mathbf{A}$ on synthetic data from Appendix F.3 with dimA $= 20$ (left) and dimA $= 100$ (right). All other variables are one-dimensional.

where $\varepsilon_{\mathbf{L}}, \varepsilon_{\mathbf{Y}} \overset{i.i.d}{\sim} \mathcal{N}(0, 0.1)$ and $\varepsilon_{\mathbf{A}}^{1}, ..., \varepsilon_{\mathbf{A}}^{D_1} \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$. In this experiment, the mini-batch size chosen is 512 and the same hyperparameters are used as in the previous settings. The neural network architecture is trained for 800 epochs. Fig. 7 present the results corresponding to 10 random seeds with different values of the trade-off parameter γ corresponding to different values of dimA among $\{15, 100\}$. In all of the box plots, it is evident that there exists a trade-off between the accuracy and counterfactual invariance of the predictor. As the value of γ increases, there is a consistent trend of augmenting counterfactual invariance (as evidenced by the decrease in the VCF metric). Similarly to the previous boxplots visualizations, the boxes represent the interquartile range (IQR), the horizontal line is the median, and whiskers show the minimum and maximum values, excluding the outliers (determined as a function of the inter-quartile range). Outliers are represented in the plot as dots.

F.4 Image dataset

The simulation procedure for the results shown in Section 4.2 is the following.

$$
\begin{aligned}
\texttt{shape} &\sim \mathbb{P}(\texttt{shape}) \\
\texttt{y-pos} &\sim \mathbb{P}(\texttt{y-pos}) \\
\texttt{color} &\sim \mathbb{P}(\texttt{color}) \\
\texttt{orientation} &\sim \mathbb{P}(\texttt{orientation}) \\
\texttt{x-pos} &= \text{round}(x), \quad \text{where } x \sim \mathcal{N}(\texttt{shape} + \texttt{y-pos}, 1) \\
\texttt{scale} &= \text{round}\Big(\Big(\frac{\texttt{x-pos}}{24} + \frac{\texttt{y-pos}}{24} \Big) \cdot \texttt{shape} + \epsilon_S \Big) \\
\mathbf{Y} &= e^{\texttt{shape}} \cdot \texttt{x-pos} + \texttt{scale}^2 \cdot \sin(\texttt{y-pos}) + \epsilon_Y,
\end{aligned}
$$

where $\epsilon_S \sim \mathcal{N}(0, 1)$ and $\epsilon_Y \sim \mathcal{N}(0, 0.01)$. The data has been generated via a matching procedure on the original dSprites dataset.

In Table 2, the hyperparameters of the layers of the convolutional neural network are presented. Each of the convolutional groups also has a ReLU activation function and a dropout layer. Two MLP architectures have been used. The former takes as input the observed tabular features. It is composed by two hidden layers of 16 and 8 nodes respectively, connected with ReLU activation functions and dropout layers. The latter takes as input the concatenated outcomes of the CNN and the other MLP. It consists of three hidden layers of 8, 8 and 16 nodes, respectively.

F.5 Fairness with continuous protected attributes

The pre-processing of the UCI Adult dataset was based upon the work of Chiappa & Pacchiano (2021). Referring to the causal graph in Fig. 8, a variational autoencoder (Kingma & Welling, 2014) was trained for each of the unobserved variables $\mathbf{H_m}$, $\mathbf{H_l}$ and $\mathbf{H_r}$. The prior distribution of these latent variables is assumed to be standard Gaussian. The posterior distributions $\mathbb{P}(\mathbf{H_m}|V)$, $\mathbb{P}(\mathbf{H_r}|V)$, $\mathbb{P}(\mathbf{H_d}|V)$ are modeled as 10-dimensional Gaussian distributions, whose means and variances are the outputs of the encoder.

The encoder architecture consists of a hidden layer of 20 hidden nodes with hyperbolic tangent activation functions, followed by a linear layer. The decoders have two linear layers with a hyperbolic tangent activation

Table 2: Architecture of the convolutional neural network used for the image dataset, as described in Appendix F.4.

layer	# filters	kernel size	stride size	padding size
convolution	16	5	2	2
max pooling	1	3	2	0
convolution	64	5	1	2
max pooling	1	1	2	0
convolution	64	5	1	2
max pooling	1	2	1	0
convolution	16	5	1	3
max pooling	1	2	2	0

function. The training loss of the variational autoencoder consists of a reconstruction term (Mean-Squared Error for continuous variables and Cross-Entropy Loss for binary ones) and the Kullback–Leibler divergence between the posterior and the prior distribution of the latent variables. For training, we used the Adam optimizer with learning rate of 10^{-2}, 100 epochs, mini-batch size 128.

The predictor $\hat{\mathbf{Y}}$ is the output of a feed-forward neural network consisting of a hidden layer with a hyperbolic tangent activation function and a linear final layer. In the training we used the Adam optimizer with learning rate 10^{-3}, mini-batch size 128, and trained for 100 epochs. The choice of the network architecture is based on the work of Chiappa & Pacchiano (2021).

The estimation of counterfactual outcomes is based on a Monte Carlo approach. Given a data point, 500 values of the unobserved variables are sampled from the estimated posterior distribution. Given an interventional value for A, a counterfactual outcome is estimated for each of the sampled unobserved values. The final counterfactual outcome is estimated as the average of these counterfactual predictions. In this experimental setting, we have $k = 100$ and $d = 1000$.

In the causal graph presented in Fig. 8, $\mathbf{A}$ includes the variables age and gender, $\mathbf{C}$ includes nationality and race, $\mathbf{M}$ marital status, $\mathbf{D}$ level of education, $\mathbf{R}$ the set of the working class, occupation, and hours per week and $\mathbf{Y}$ the income class. Compared to Chiappa & Pacchiano (2021), we include the race variable in the dataset as part of the baseline features $\mathbf{C}$. The loss function is the same as Eq. (1) but Binary Cross-Entropy loss ($\mathcal{L}_{\mathbf{BCE}}$) is used instead of Mean-Squared Error loss:

$$\mathcal{L}_{\text{CIP}}(\hat{\mathbf{Y}}) = \mathcal{L}_{\mathbf{BCE}}(\hat{\mathbf{Y}}) + \gamma \cdot \text{HSCIC}\left(\hat{\mathbf{Y}}, \{\text{age}, \text{gender}, \text{marital status}, \text{education}, \text{work}\} \middle| \mathbf{S}\right) , \tag{14}$$

where the set $\mathbf{S} = \{\text{Race}, \text{Nationality}\}$ blocks all the non-causal paths from $\mathbf{W} \cup \mathbf{A}$ to $\mathbf{Y}$. In this example we have $\mathbf{W} = \{\mathbf{C} \cup \mathbf{M} \cup \mathbf{D} \cup \mathbf{R}\}$. The results in Fig. 5 (right) refer to one run with conditioning set $\mathbf{S} = \{\text{Race}, \text{Nationality}\}$. The results correspond to 4 random seeds.

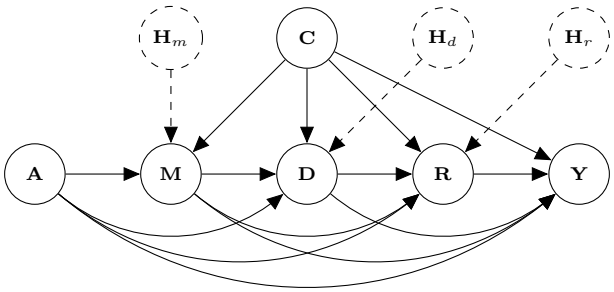

Figure 8: Assumed causal graph for the Adult dataset, as in Chiappa & Pacchiano (2021). The variables $\mathbf{H}_m$, $\mathbf{H}_d$, $\mathbf{H}_r$ are unobserved, and jointly trained with the predictor $\hat{\mathbf{Y}}$.

Table 3: Results of MSE and VCF (all times 10^2 for readability) on synthetic data of CIP with trade-off parameters depending on the chosen accuracy threshold.

	VCF $\times 10^2$	**HSCIC** $\times 10^2$
90% accuracy	3.14 ± 0.92	4.51 ± 0.72
70% accuracy	3.01 ± 0.80	4.44 ± 0.65
1% accuracy	2.91 ± 0.92	4.39 ± 0.42

F.6 Illustrating the choice of γ

In Section 3.4, we propose to choose γ to obtain a maximal level of CI within a given tolerance on predictive performance. Here, we illustrate results from running the proposed procedure that dynamically selects γ, adjusted to different predefined accuracy thresholds in a classification setting. Specifically, the algorithm chooses the largest γ value that yields an accuracy equal to or better than the threshold. As described the algorithm operates on γ values on a logarithmic scale, thereby ensuring a fine-grained search over a wide range of potential trade-off points. Table 3 shows the found trade-offs for tolerated accuracies of 90%, 70%, and 1% in the same setting as Appendix F.1.

F.7 Computational complexity and runtimes

In a dataset with $n = 1000$ data points from the setting discussed in Fig. 2, the average training time for one epoch without the regularization term is 0.003s and 1.112s with the regularization term. In these results, Adam optimizer with batch-size of 512 was used. By using smaller batch sizes, e.g. $n = 128$, the extra computational cost can further decrease. From a theoretical perspective, the estimation of the HSCIC requires kernel ridge regression (see Eq. (2) in our submission). In the high-dimensional image example, with a mini batch-size of 512, the average running time for an epoch with the regularization term is 64.03s and 34.01s without. Kernel ridge regression generally has a runtime that scales as $O(n^3)$ and memory requirements scaling like $O(n^2)$, where n is the size of the dataset. However, these bounds can be significantly improved by using, i.e., random Fourier Features (see, i.e., Rahimi & Recht (2007); Avron et al. (2017)) as detailed in Appendix E. In short, by using random Fourier features, the resulting approximate kernel ridge regression estimator can be computed in a runtime of $O(ns^2)$ with $O(ns)$ memory. Here, s is a parameter determining the accuracy of the approximation. In practice, s can be set to be much smaller than the problem size, resulting in a dramatic speed-up. Other methods for efficient kernel computation include the popular Nystrom approximation (Drineas & Mahoney, 2005; Hsieh et al., 2014), and Memory-Efficient Kernel Approximation (MEKA) (Si et al., 2017). In this work, runtimes were still reasonable for all experimental settings, which is why we did not have to resort to these faster approximations.

G Comparison with additional baselines

In this section, we compare CIP with additional baselines. These include Veitch et al. (2021) and different heuristic methods.

G.1 Baseline experiments (Veitch et al., 2021)

We provide an experimental comparison against the method by Veitch et al. (2021). To this end, we consider the following data-generating mechanism for the causal structure (see Fig. 1(b)):

$$\mathbf{Z} \sim \mathcal{N}(0,1) \qquad \mathbf{A} = \sin(0.1\mathbf{Z}) + \varepsilon_{\mathbf{A}}$$

$$\mathbf{X} = \exp\left\{-\frac{1}{2}\mathbf{A}\right\}\sin(\mathbf{A}) + \frac{1}{10}\varepsilon_{\mathbf{X}}$$

$$\mathbf{Y} = \frac{1}{10}\exp\{-\mathbf{X}\}\cdot\sin(2\mathbf{X}\mathbf{Z}) + \mathbf{A}\mathbf{A} + \frac{1}{10}\varepsilon_{\mathbf{Y}},$$

Table 4: **Results of the MSE, VCF of CIP and the baseline (Veitch et al., 2021)** applied to the causal and anti-causal structure in Fig. 1(b-c). Although the graphical assumptions are not satisfied, CIP shows an overall decrease of VCF in both of the graphical structures, performing on par with the baseline Veitch et al. (2021) in terms of accuracy and counterfactual invariance.

	CIP		**Veitch et al. (2021)**	
	MSE $\times 10^2$	**VCF**	**MSE**$\times 10^2$	**VCF**
$\gamma = 0.5$	4.58 ± 0.31	0.19 ± 0.02	4.50 ± 0.40	0.19 ± 0.02
$\gamma = 1.0$	5.60 ± 0.36	0.18 ± 0.01	5.45 ± 0.41	0.18 ± 0.02

	CIP		**Veitch et al. (2021)**	
	MSE $\times 10^2$	**VCF**	**MSE**$\times 10^2$	**VCF**
$\gamma = 0.5$	1.16 ± 0.01	1.69 ± 0.16	1.01 ± 0.01	1.71 ± 0.26
$\gamma = 1.0$	1.37 ± 0.02	1.48 ± 0.19	0.99 ± 0.01	1.88 ± 0.28

where $\varepsilon_{\mathbf{X}}, \varepsilon_{\mathbf{A}} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ and $\varepsilon_{\mathbf{Y}} \overset{i.i.d}{\sim} \mathcal{N}(0,0.1)$. The data-generating mechanism of the anti-causal structure is the following (see Fig. 1(c)):

$$\mathbf{Z} \sim \mathcal{N}(0,1) \qquad \mathbf{A} = \frac{1}{5}\sin(\mathbf{Z}) + \varepsilon_{\mathbf{A}}$$

$$\mathbf{Y} = \frac{1}{10}\sin(\mathbf{Z}) + \varepsilon_{\mathbf{Y}}$$

$$\mathbf{X} = \mathbf{A} + \mathbf{Y} + \frac{1}{10}\varepsilon_{\mathbf{X}}$$

where $\varepsilon_{\mathbf{Y}}, \varepsilon_{\mathbf{A}} \overset{i.i.d}{\sim} \mathcal{N}(0,0.1)$ and $\varepsilon_{\mathbf{X}} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. We compare our method (CIP) against the method by Veitch et al. (2021) using different values for the trade-off parameter γ. In Fig. 1(b-c) the causal and anti-causal graphical settings proposed by Veitch et al. (2021) are presented. In both of these settings there is an unobserved confounder $\mathbf{Z}$ between $\mathbf{A}$ and $\mathbf{Y}$. The graphical assumptions outlined in Theorem 3.2 of the CIP are not met in the graphical structures under examination, as the confounding path is not effectively blocked by an observed variable ($\mathbf{Z}$ is unobserved). In light of this, it is assumed in our implementation that there is no unobserved confounder. In the graphical structure Fig. 1(b), CIP enforces $\text{HSIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{X})$ to become small, gradually enforcing $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{X}$. HSIC is the Hilbert-Schmidt Independence Criterion, which is commonly used to promote independence (see, i.e., Gretton et al. (2005); Fukumizu et al. (2007)). Veitch et al. (2021) enforces as independence criterion $\text{HSIC}(\hat{\mathbf{Y}}, \mathbf{A})$, which is implied by the independence enforced in CIP. In the anti-causal graphical setting presented in Fig. 1(c), the objective term used in CIP is $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{X})$, while in the method of Veitch et al. (2021) is $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Y})$. In Table 4, the results of accuracy and VCF are presented.

In the experiments, the predictor $\hat{\mathbf{Y}}$ is a feed-forward neural network consisting of 8 hidden layers with 20 nodes each, connected with a rectified linear activation function (ReLU) and a linear final layer. Mini-batch size of 256 and the Adam optimizer with a learning rate of 10^{-4} for 500 epochs were used.

G.2 Comparison baselines heuristic methods

We provide an experimental comparison of the proposed method (CIP) with some heuristic methods, specifically data-augmentation-based methods. We consider the same data-generating procedure and causal structure as presented in Appendix F.1. The heuristic methods considered are *data augmentation* and *causal-based data augmentation*. In the former, data augmentation is performed by generating $N = 50$ samples for every data-point by sampling new values of $\mathbf{A}$ as $a_1, ..., a_N \overset{i.i.d}{\sim} \mathbb{P}_{\mathbf{A}}$ and leaving $\mathbf{Z}, \mathbf{L}, \mathbf{Y}$ unchanged. Differently, in the latter *causal-based data augmentation* method, we also take into account the causal structure

Table 5: Results of MSE and VCF (all times 10^2 for readability) on synthetic data of CIP with trade-off parameters $\gamma = 0.5$ and $\gamma = 1$ with the heuristic methods *data augmentation* and *causal-based data augmentation* and *naive prediction*.

	VCF $\times 10^3$	**MSE** $\times 10^3$
data augmentation	3.12 ± 0.16	0.03 ± 0.01
causal-based data augmentation	3.04 ± 0.16	0.13 ± 0.12
CIP ($\gamma = 0.5$)	1.05 ± 0.13	1.64 ± 0.22
CIP ($\gamma = 1.0$)	0.35 ± 0.19	2.50 ± 0.72
naive prediction (ignore A)	9.01 ± 0.02	3.01 ± 0.91

given by the known DAG. Indeed, when manipulating the variable $\mathbf{A}$, its descendants (in this example $\mathbf{L}$) will also change. In this experiment, a predictor for $\mathbf{L}$ as $\hat{\mathbf{L}} = f_\theta(\mathbf{A}, \mathbf{Z})$ is trained on 80% of the original dataset. In the data augmentation mechanism, for every data-point $\{a, x, z, y\}$, $N = 50$ samples are generated by sampling new values of $\mathbf{A}$ as $a_1, ..., a_N \overset{i.i.d}{\sim} \mathbb{P}_{\mathbf{A}}$, estimating the values of $\mathbf{L}$ as $x_1 = f_\theta(a_1, z), ..., x_N = f_\theta(a_N, z)$, while leaving the values of $\mathbf{Z}$ and $\mathbf{Y}$ unchanged. Heuristic methods such as data-augmentation methods do not theoretically guarantee to provide counterfactually invariant predictors. The results of an empirical comparison are shown in Table 5 with the average and standard deviations after 5 random seeds. It can be shown that these theoretical insights are supported by experimental results, as the VCF metric measure counterfactual invariance is lower in both of the two settings of the CIP ($\gamma = \frac{1}{2}$ and $\gamma = 1$).

A dataset of $n = 3000$ is used, along with $k = 500$ and $d = 500$. The architecture for predicting $\mathbf{L}$ and $\mathbf{Y}$ are feed-forward neural networks consisting of 8 hidden layers with 20 nodes each, connected with a rectified linear activation function (ReLU) and linear final layer. Mini-batch size of 256 and the Adam optimizer with a learning rate of 10^{-3} for 100 epochs were used.