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.

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 $\mathbf{Y}^*_{\mathbf{a}}$ for the outcome in the intervened SCM, also called *potential outcome*. Note that the interventional distribution $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y})$ 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?" Counterfactuals capture properties of a "subpopulation" $\mathbf{u} \sim \mathbb{P}_{\mathbf{U}|\mathbf{W}=\mathbf{w}}$ compatible with the observations.[4] Even for granular $\mathbf{W}$, 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 $\hat{\mathbf{Y}}$.

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

[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 $\mathbf{Y}^*_{\mathbf{a}}$ 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 $\mathbb{P}_{\mathbf{Y}_{\mathbf{a}}}(\mathbf{y}) = \mathbb{P}_{\mathbf{Y}_{\mathbf{a}'}}(\mathbf{y})$ almost surely, for all $\mathbf{a}, \mathbf{a}'$ in the domain of $\mathbf{A}$.

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 $\mu_{\mathbf{Y}|\mathbf{S}=\mathbf{s}} := \mathbb{E}\left[k_{\mathbf{Y}}(\,\cdot\,,\mathbf{y}) \mid \mathbf{S}=\mathbf{s}\right]$. 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 granular 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 shed light on the connection between almost sure CI (a.s.-CI), distributional CI ($\mathcal{D}$-CI) as in Definition 2.2, and CI of predictors ($\mathcal{F}$-CI) showing that $f_{\hat{\mathbf{Y}}}$ being $\mathcal{F}$-CI is equivalent to $\hat{\mathbf{Y}}$ being $\mathcal{D}$-CI conditioned on $\mathbf{X}$, rendering it also equivalent to counterfactual fairness (Kusner et al., 2017).

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 independencies implies any bounds on the difference in counterfactuals. On the contrary, the weaker notion of $\mathcal{D}$-CI in Definition 2.2 can in fact be equivalent to conditional independencies in the observational distribution (Fawkes & Evans, 2023, Lem. A.3). Albeit only proved for special cases where the counterfactual distribution is identifiable, this opens the possibility for sufficient graphical criteria for distributional CI in any graph.

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.

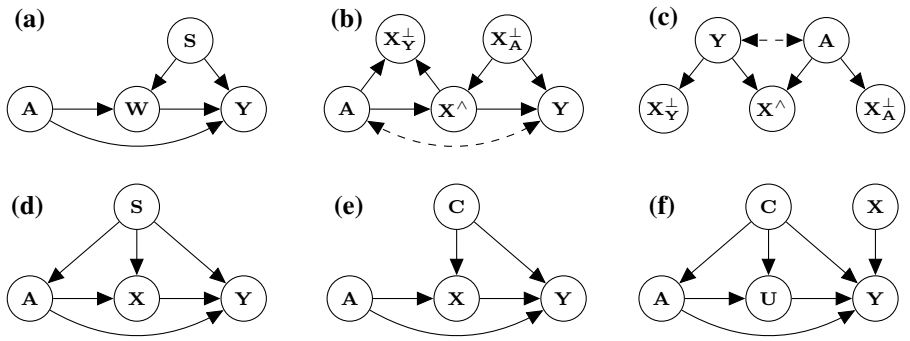

Figure 1: **(a)** Exemplary graph in which a predictor $\hat{\mathbf{Y}}$ with $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}$ is CI in $\mathbf{A}$ w.r.t. $\mathbf{W}$. **(b)-(c)** Causal and anti-causal structure from Veitch et al. (2021) where $\mathbf{X}_\mathbf{A}^\perp$ 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 $\mathbf{A} = \{$'gender', 'age'$\}$. **(f)** Causal structure for our semi-synthetic image experiments (Section 4.2), where $\mathbf{A} = \{\text{Pos.X}\}$, $\mathbf{U} = \{\text{Scale}\}$, $\mathbf{C} = \{\text{Shape}, \text{Pos.Y}\}$, $\mathbf{X} = \{\text{Color}, \text{Orientation}\}$, and $\mathbf{Y} = \{\text{Outcome}\}$.

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 $\mathcal{G}$.

The sufficient graphical criterion that renders CI equivalent to a conditional independence then reads as follows. (The proof is deferred to 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 $\mathbf{X} := \mathbf{W} \setminus \mathbf{A}$ assume that $\mathbf{X} = g(\mathbf{X}, \mathbf{A}, \mathbf{U}_\mathbf{X})$ (which implies $\mathsf{pa}(V) \in \mathbf{X} \cup \mathbf{A}$ for all $V \in \mathbf{X} \cup \mathbf{A}$) with g injective in $\mathbf{U}_\mathbf{X}$ for all values of $\mathbf{A}$ and $\mathbf{X}$. 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 note that assuming the causal graph to be known is a standard assumption widely made in the causality literature, even though it is a strong one (Cartwright, 2007). Fawkes & Evans (2023, Prop. A.3) shows that CI cannot be decided from the observational distribution (and the causal graph) unless strong assumptions are made—they consider a case in which counterfactuals are identified. A strong assumption of Theorem 3.2, is on the arguments and injectivity of g, which ensures that any given observation $(\mathbf{x}, \mathbf{a})$ of $\mathbf{X}$ and $\mathbf{A}$ puts a point mass on a single $\mathbf{u}$ for $\mathbf{U}_\mathbf{X}$ during abduction. While this is an untestable and admittedly strong, we highlight that this is not a limitation of our works specifically, but at least comparably strong assumptions are provably required for any method that claims to guarantee CI from observational data. Therefore, we complement our theoretical results by extensive experimental evaluation demonstrating CIPS's efficacy even when some assumptions are violated. Finally, we do *not* assume the full SCM to be known.

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) and $\mathbf{X}$ (inputs to $f_{\hat{\mathbf{Y}}}$). There may be confounders $\mathbf{S}$ between $\mathbf{X}$ and $\mathbf{Y}$ and we consider $\mathbf{W} = \mathbf{X} \cup \mathbf{A} \cup \mathbf{S}$. Here we aim to achieve $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \cup \mathbf{X} \mid \mathbf{S}$ to obtain CI in $\mathbf{A}$ w.r.t. $\mathbf{X}$. 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.

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 $(f \otimes g)(\mathbf{y}, [\mathbf{a}, \mathbf{w}]) := f(\mathbf{y})g([\mathbf{a}, \mathbf{w}])$, 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 $k(\mathbf{y} \otimes [\mathbf{a}, \mathbf{w}], \mathbf{y}' \otimes [\mathbf{a}', \mathbf{w}']) := k_{\mathbf{Y}}(\mathbf{y}, \mathbf{y}')k_{\mathbf{A}\cup\mathbf{W}}([\mathbf{a}, \mathbf{w}], [\mathbf{a}', \mathbf{w}'])$. 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 $H_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}}(\mathbf{s}) := \|\mu_{\mathbf{Y},\mathbf{A}\cup\mathbf{W}|\mathbf{S}=\mathbf{s}} - \mu_{\mathbf{Y}|\mathbf{S}=\mathbf{s}} \otimes \mu_{\mathbf{A}\cup\mathbf{W}|\mathbf{S}=\mathbf{s}}\|$ 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*[6]*,* $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = 0$ *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.

[6]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.

Corollary 3.5. *Under the assumptions of Theorem 3.2, if* $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S}) = 0$ *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 $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \cup \mathbf{W} \mid \mathbf{S})$ 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

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

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.[7] 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 log-spaced binary search on γ (e.g., on $[1e-4, 1e4]$) 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.

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

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

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. 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). 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

[7] 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.

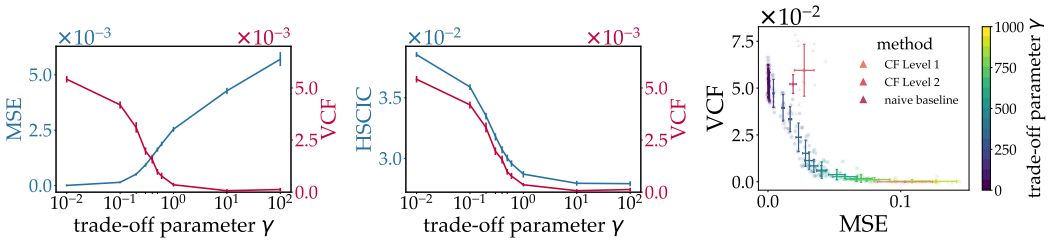

Figure 2: Results on synthetic data (see Appendix F.1 and Appendix F.2). **Left:** trade-off between MSE and counterfactual invariance (VCF). **Middle:** strong correspondence between HSCIC and VCF. **Right:** performance of CIP against baselines CF1 and CF2 and the naive baseline. As γ increases, CIP 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.

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

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

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 $\mathrm{VCF}(\hat{\mathbf{Y}}) = 0$ 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 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.

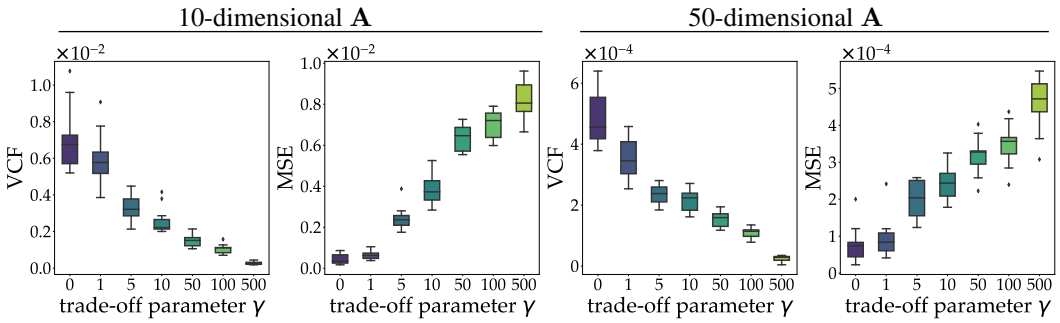

Figure 3: MSE and VCF for synthetic data (Appendix F.3) with 10- and 50-dimensional $\mathbf{A}$ for different γ and 15 random seeds per box. CIP reliably achieves CI as γ increases.

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 $\mathbf{X}$ 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. $\mathbf{W} := \mathbf{A} \cup \mathbf{X} \cup \mathbf{S}$. 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.

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 realiably 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 $\mathbf{A}$.[8] Further results are shown in Appendix F.3.

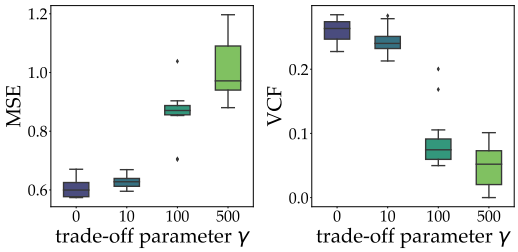

Figure 4: On the dSprites image dataset, CIP trades off MSE for VCF and achieves almost full CI as γ increases. Boxes are for 8 random seeds.

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 $\mathbf{X}$ (`color`, `orientation`), with continuous $\mathbf{A}$ (`x-pos`). We seek a

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

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.

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). We explicitly acknowledge the shortcomings of this 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 $\mathbf{X}$, 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. $\mathbf{W} = \mathbf{C} \cup \mathbf{X}$. 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.

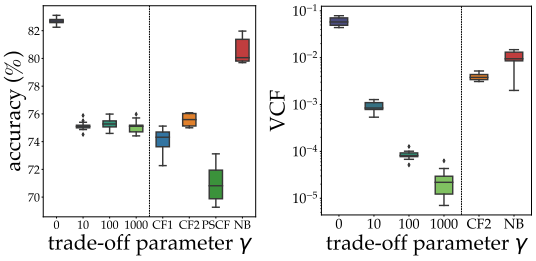

Figure 5: Accuracy and VCF on the Adult dataset. CIP achieves better VCF than CF2 and the naive baseline (NB), improved in accuracy compared to PSCF and is on par with CF1 in accuracy at VCF ≈ 0.

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. The naive baselines is more accurate than CIP for $\gamma \geq 5$ while 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.

REPRODUCIBILITY STATEMENT

The reproducibility of our work is ensured through several means. For the theoretical components, complete proofs of the stated theorems are provided in Appendix A and Appendix B. In terms of our experimental findings, Appendix F offers a detailed description of the hyperparameters and models. Moreover, to facilitate practical replication, the corresponding code is available in the supplementary material.

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 and E. ðCınlar. *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.

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.

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. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

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.

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 *The Eleventh 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.