Learning Counterfactually Invariant Predictors

Francesco Quinzan[*1] **Cecilia Casolo**[*2] **Krikamol Muandet**[3] **Yucen Luo**[†4] **Niki Kilbertus**[†52]

Abstract

Notions of counterfactual invariance have proven essential for predictors that are fair, robust, and generalizable in the real world. We propose simple graphical criteria that yield a sufficient condition for a predictor to be counterfactually invariant in terms of (conditional independence in) the observational distribution. Any predictor that satisfies our criterion is provably counterfactually invariant. In order to learn such predictors, we propose a model-agnostic framework, called Counterfactual Invariance Prediction (CIP), building on a kernel-based conditional dependence measure called Hilbert-Schmidt Conditional Independence Criterion (HSCIC). 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 and Related Work

Invariance, or equivariance to certain transformations of data, 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).

Many real-world applications in modern ML, however, call for an arguably stronger notion of invariance based on causality, called *counterfactual invariance*. These applications require predictors to exhibit invariance with respect to hypothetical manipulations of the data generating process (DGP) (Peters et al., 2016; Heinze-Deml et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Bühlmann, 2020). Counterfactual invariance has the advantage that it incorporates structural knowledge of the DGP. However, enforcing counterfactual invariance is challenging in practice, because it is typically untestable in real-world observational settings unless strong prior knowledge of the DGP is available.

Inspired by problems in natural language processing (NLP), (Veitch et al., 2021) analyze two specific causal graphs (dubbed causal and anticausal, which we replicate in Figure 1(b,c)) with the goal of "stress-testing" models for spurious correlations. They develop *necessary, but not sufficient*, criteria to achieve counterfactual invariance in these two settings based only on the observational distribution. These criteria are enforced in practice for discrete conditioning variables via distribution matching using the maximum mean discrepancy (MMD). Our work differs in that we provide graphical criteria for any given causal graph and develop a *sufficient* (potentially not necessary) criterion for counterfactual invariance, again based on the observational distribution only. Hence, unlike (Veitch et al., 2021), our approach guarantees counterfactual invariance. Finally, we propose a model-agnostic learning framework, called Counterfactual Invariance Prediction (CIP), building on a kernel-based conditional dependence measure called Hilbert-Schmidt Conditional Independence Criterion (HSCIC) (Park & Muandet, 2020). CIP thus allows for mixed categorical and continuous multivariate variables.

2. Preliminaries

We denote with $\mathbf{Y} \subset \mathbf{V}$ the outcome (or prediction target), and with $\hat{\mathbf{Y}}$ a predictor for that target. Each Structural Causal Model (SCM) implies a unique *observational distribution* over $\mathbf{V}$, but it also entails interventional distributions (Pearl, 2000). 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 $A \leftarrow a$. Similarly, we can intervene on multiple variables $\mathbf{V} \supseteq \mathbf{A} \leftarrow \mathbf{a}$. 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 $\mathbb{P}_{\mathbf{Y}|\mathbf{A}}(\mathbf{y} \mid \mathbf{a})$.[1]This is typically the case when Y and A have a shared parent, i.e., are confounded. We can also condition on a set of variables $\mathbf{W} \subseteq \mathbf{V}$ in the (observational distribution of the) original

[*,†] Equal contribution [1]Department of Computer Science, University of Oxford [2]Helmholtz AI, Munich [3]CISPA–Helmholtz Center for Information Security [4]Max Planck Institute for Intelligent Systems [5]Technical University of Munich. Correspondence to: Francesco Quinzan <francesco.quinzan@cs.ox.ac.uk>.

Presented at the 2^{nd} Workshop on Formal Verification of Machine Learning, co-located with the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA., 2023.

[1]We use $\mathbb{P}$ for distributions (common in the kernel literature) and the notation $\mathbf{Y}^*_{\mathbf{a}}$ instead of $\mathbf{Y} \mid do(\mathbf{a})$ for conciseness.

SCM before performing an intervention, which we denote by $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w})$. This is 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?". Note that the sets $\mathbf{A}$ and $\mathbf{W}$ need not be disjoint.

3. Counterfactual Invariance Prediction (CIP)

3.1. Sufficient Criterion for Counterfactual Invariance

We start with the definition of counterfactual invariance.

Definition 3.1 (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 $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w}) = \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w})$ almost surely, for all $\mathbf{a}, \mathbf{a}'$ in the domain of $\mathbf{A}$ and all $\mathbf{w}$ in the domain of $\mathbf{W}$.[2]

A counterfactually invariant predictor $\hat{\mathbf{Y}}$ can then be viewed as robust to changes of $\mathbf{A}$ in the sense that the (conditional) post-interventional distribution of $\hat{\mathbf{Y}}$ does not change for different values of the intervention. We now discuss some properties of our Definition 3.1 in comparison to other notions of counterfactual invariance. First, we can condition on observations $\mathbf{W}$, which allows us to model true counterfactuals including the abduction step where we condition on observed evidence. For example, enforcing counterfactual fairness requires modeling true counterfactual (Kusner et al., 2017). This sets our definition apart, for example, from Def. 1.1 of (Veitch et al., 2021), who require $\hat{\mathbf{Y}}^*_{\mathbf{a}} = \hat{\mathbf{Y}}^*_{\mathbf{a}'}$ almost surely for all $\mathbf{a}, \mathbf{a}'$ in the domain of $\mathbf{A}$. While this condition appears stronger by enforcing equality of random variables instead of equality of distributions, in practice (Veitch et al., 2021) also enforces equality of distributions (via MMD). Moreover, since $\hat{\mathbf{Y}}^*_{\mathbf{a}}, \hat{\mathbf{Y}}^*_{\mathbf{a}'}$ are (deterministic) functions of the *same* exogenous (unobserved) random variables, distributional equality is a natural choice for counterfactual invariance. Def. 1 of (Mouli & Ribeiro, 2022) instead define counterfactually invariant *representations* of some data as being invariant under a family of pre-specified symmetry transformations of the data (based on equivalence relations).

Next, we establish a simple graphical criterion to express counterfactual invariance as conditional independence in the observational distribution of an SCM, rendering it estimable from observational data. Crucially, our main result provides sufficient conditions for counterfactual invariance.

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 $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$, and define $\mathbf{Z} := (\mathbf{S} \cup \mathbf{W}) \setminus \mathbf{A}$. Then, in all SCMs compatible with $\mathcal{G}$, if the variable $\mathbf{Y}$ satisfies $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$, then $\mathbf{Y}$ is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W}$.*

Note that the conditioning set $\mathbf{Z}$ in Theorem 3.2 depends on $\mathbf{A}$, $\mathbf{W}$, and the given valid adjustment set $\mathbf{S}$. In particular, the conditioning set $\mathbf{Z}$ need not itself be a valid adjustment set. The proof is deferred to Section 7. Crucially, our proof does *not* rely on the identification of the counterfactual distributions (e.g., via the do-calculus (Pearl, 2000)).

Theorem 3.2 provides a sufficient condition for $\mathbf{Y}$ to be counterfactually invariant (Definition 3.1). If a predictor $\hat{\mathbf{Y}}$ satisfies $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$, this is a counterfactually invariant predictor. We illustrate Theorem 3.2 with an example in Figure 1(a).

In the following, we will develop an operator denoted by $\mathrm{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z})$ that is efficiently estimable from observational data, differentiable, serves as a measure of conditional dependence, and is zero if and only if $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$. We can then use this operator as a model-agnostic objective to train counterfactually invariant predictors. Some background is required.

3.2. HSCIC for Conditional Independence

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

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

Our Definition 3.3 is heavily motivated by, but differs slightly from Def. 5.3 of (Park & Muandet, 2020), which relies on the Bochner conditional expected value. While it is functionally equivalent (with the same implementation, see Section 11), ours has the benefit of bypassing some technical assumptions required by (Park & Muandet, 2020) (see Section 9-Section 10 for details). The HSCIC has the following important property.

Theorem 3.4 (Theorem 5.4 by (Park & Muandet, 2020))**.** *If the kernel k of $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}}$ is characteristic*[3]*,* $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A} \mid \mathbf{Z}) = 0$ *almost surely if and only if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$.

A proof is in Section 8. We remark that "most interesting" kernels such as the Gaussian and Laplacian kernels are char-

[2] 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 $\mathbf{A}$.

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

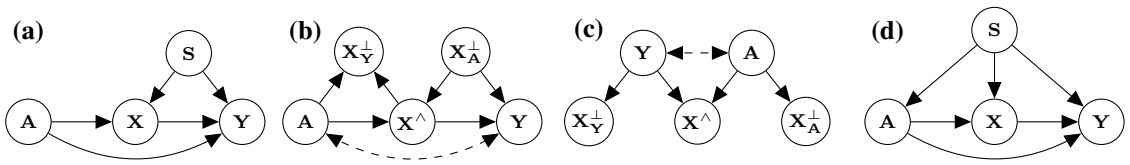

Figure 1. **(a) An example for Theorem 3.2**. Any predictor $\hat{\mathbf{Y}}$ such that $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$ with $\mathbf{Z} = \mathbf{X} \cup \mathbf{S}$ is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{X}$. **(b)-(c) Causal and anti-causal structure** as in (Veitch et al., 2021). The variable $\mathbf{X}$ is decomposed in three parts. $\mathbf{X}_{\mathbf{A}}^{\perp}$ is the part of $\mathbf{X}$ that is not causally influenced by $\mathbf{A}$, $\mathbf{X}_{\mathbf{Y}}^{\perp}$ is the part that does not causally influence $\mathbf{Y}$, and $\mathbf{X}^{\wedge}$ is the remaining part that is both influenced by $\mathbf{A}$ and that influences $\mathbf{Y}$. **(d) Causal structure** for the synthetic experiments (see Section 4.1).

acteristic. Furthermore, if kernels are translation-invariant and characteristic, then their tensor product is also a characteristic kernel (Szabó & Sriperumbudur, 2017). Hence, 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. *Consider an SCM with causal diagram $\mathcal{G}$ and fix two (not necessarily disjoint) sets of nodes $\mathbf{A}$, $\mathbf{W}$. Let $\mathbf{Z}$ be a set of nodes as in Theorem 3.2. Then, if it holds* $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A} \mid \mathbf{Z}) = 0$ *almost surely, then $\mathbf{Y}$ is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W}$.*

In Section 11, the estimation of HSCIC from samples is introduced. We emphasize that this procedure allows us to consistently estimate the HSCIC *from observational i.i.d. samples, without prior knowledge of the counterfactual distributions*.

3.3. Learning Counterfactually Invariant Predictors

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

$$\mathcal{L}_{\mathrm{CIP}}(\hat{\mathbf{Y}}) = \mathcal{L}(\hat{\mathbf{Y}}) + \gamma \cdot \mathrm{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z}) \,, \qquad (1)$$

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. The second term in Equation (1) does not act as a regularizer, i.e., we do not aim at overcoming an ill-posedness of, e.g., multiple models with equal training loss $\mathcal{L}$. Instead, it is an additional objective, typically in conflict with $\mathcal{L}$. As a result, γ does not need to decay to zero as the sample size increases and it is impossible to select an "optimal value" based on data alone. In practice, driving HSCIC to zero to ensure perfect counterfactual invariance typically deteriorates the predictive performance on observational data useless compared to an unconstrained model. Typical use-cases either have strict requirements on predictive performance (e.g., at most 5% loss in accuracy compared to an unconstrained model) and/or may tolerate small violations of invariance (which cannot be avoided in a finite data setting in any case). Hence, a practically useful approach consists of: (i) learn a collection of CIP models on the Pareto frontier of invariance and predictive performance using different values of γ; (ii) choose one of these models depending on pre-specified criteria in the application context such as the most invariant model within a required predictive performance, or the most predictive model within a certain tolerance regarding invariance.

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 $\hat{\mathbf{Y}}$. Such a measure must capture how the distribution of $\hat{\mathbf{Y}}_{\mathbf{a}}^{*}$ changes for different values of $\mathbf{a}$ across all conditioning values $\mathbf{w}$. We quantify this in a single scalar, which we call the **V**ariance of **C**ounter**F**actuals (VCF)

$$\mathrm{VCF}(\hat{\mathbf{Y}}) = \mathbb{E}_{\mathbf{w} \sim \mathbb{P}_{\mathbf{W}}} \left[\mathrm{var}_{\mathbf{a}' \sim \mathbb{P}_{\mathbf{A}}} [\mathbb{E}_{\hat{\mathbf{Y}}_{\mathbf{a}'}^{*} \mid \mathbf{W}=\mathbf{w}} [\hat{\mathbf{y}} \mid \mathbf{w}]] \right] . \quad (2)$$

That is, we look at 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, i.e., point estimators, which we use in all our experiments, $\hat{\mathbf{Y}}$ is constant and we can drop the inner expectation of Equation (2). In this case, the variance term in Equation (2) is zero if and only if $\mathbb{P}_{\hat{\mathbf{Y}}_{\mathbf{a}}^{*} \mid \mathbf{W}}(\mathbf{y} \mid \mathbf{w}) = \mathbb{P}_{\hat{\mathbf{Y}}_{\mathbf{a}'}^{*} \mid \mathbf{W}}(\mathbf{y} \mid \mathbf{w})$ almost surely. Since the variance is non-negative, the outer expectation is zero if and only if the variance term is zero almost surely. Hence, $\mathrm{VCF}(\hat{\mathbf{Y}}) = 0$ almost surely is equivalent to counterfactual invariance.

Crucially, VCF requires access to ground-truth counterfactual distributions, which by their very nature are unavailable in practice (neither for training nor at test time). Hence, we can only assess VCF, as a direct measure of counterfactual invariance, in synthetic scenarios. Our experiments demonstrate that HSCIC (estimable from the observed data) empirically serves as a proxy for VCF.

4. Experiments

In this section, we aim to demonstrate the effectiveness of the proposed method in enforcing counterfactual invariance across different simulated datasets. We compare CIP with established baselines to showcase its competitive results in preserving counterfactual invariance. In Section 13.3 and Section 13.4, the method is also respectively applied to image and real-world datasets.

Baselines Since counterfactually invariant training has not received much attention yet, our choice of baselines for ex-

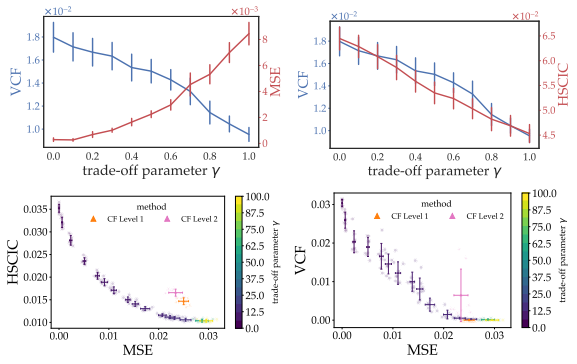

Figure 2. **(Top-left) trade-off between the accuracy and counterfactual invariance.** We observe that the VCF decreases, as the MSE increases. Vertical bars denote standard errors over 10 different random seeds. **(Top-right) Correspondence between the HSCIC and the VCF**, for increasing γ. Again, vertical bars denote standard errors over 10 different random seeds. **(Bottom) Performance of CIP against baselines CF1 and CF2 on a synthetic dataset (see Section 13.2).** Notably, the HSCIC-MSE frontier traced out for different values of the trade-off parameter (which is available in purely observational settings) can guide the desired choice of γ as it closely mimics the VCF-MSE frontier. CF2 is Pareto-dominated by this frontier, i.e., we can pick γ to outperform CF2 in both MSE and VCF simultaneously.

perimental comparison is limited. We benchmarked CIP against (Veitch et al., 2021) in the specific causal and anticausal settings of Figure 3(b-c) in Section 13.5, showing that our method performs on par with theirs. Since counterfactual fairness is a special case of counterfactual invariance, we also compare against two methods proposed by (Kusner et al., 2017) (in applicable settings). We compare to the *Level 1* (only use non-descendants of $\mathbf{A}$ as inputs to $\hat{\mathbf{Y}}$) and the *Level 2* (assume an additive noise model and in addition to non-descendants, only use the residuals of descendants of $\mathbf{A}$ after regression on $\mathbf{A}$ as inputs to $\hat{\mathbf{Y}}$) approaches of (Kusner et al., 2017). We refer to these two baselines as CF1 and CF2 respectively.

4.1. Synthetic Experiments

We begin our empirical assessment of HSCIC, by generating various synthetic datasets following the causal graph in Figure 1(d). The datasets are composed of four sets of observed continuous variables: (i) the prediction target $\mathbf{Y}$, (ii) the variable(s) we want to be counterfactually invariant in $\mathbf{A}$, (iii) covariates that mediate 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 counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W} := \mathbf{A} \cup \mathbf{X} \cup \mathbf{S}$. Following the notation of Theorem 3.2, we have $\mathbf{Z} = \mathbf{X} \cup \mathbf{S}$. We consider various synthetic datasets for this case, which mainly differ in the dimension of the observed variables and their correlations. All datasets are described in detail in Section 13.

Model performance. We first perform a set of experiments to study the effect of the HSCIC, and to highlight the trade-off between accuracy and counterfactual invariance. For this set of experiments, we generate a dataset as described in Section 13.1. Figure 2 (top-left) shows the values attained by the VCF and MSE for increasing γ, demonstrating the expected trade-off in raw predictive performance and enforcing counterfactual invariance. Finally, Figure 2 (top-right) highlights the usefulness of HSCIC as a measure of counterfactual invariance, being in strong agreement with VCF (see discussion after Equation (2)).

Comparison with baselines. We compare CIP against baselines in different simulated settings. In Figure 2(bottom), the results for a non-additive noise model data-generating mechanism are shown. For a suitable choice of γ, CIP outperforms the baseline CF2 in both MSE and VCF simultaneously. While CF1 satisfies counterfactual invariance perfectly by construction (VCF $= 0$), its MSE is generally higher in comparison to other possible choices of the parameter γ that still achieve high levels of counterfactual invariance. Our method provides to flexibly trade predictive performance for counterfactual invariance via a single tuning knob λ and Pareto-dominates existing methods. In Section 13.2, the results in another simulated setting are presented.

5. Discussion and Future Work

We developed a method to learn counterfactually invariant predictors $\hat{\mathbf{Y}}$, i.e., predictors that remain invariant in changes of certain covariates (conditioned on observed evidence). First, we presented a novel sufficient graphical criterion to characterize counterfactual invariance and reduce it to conditional independence in the observational distribution. Our method (CIP) does not require identifiability of the counterfactual distribution. We then built on kernel mean embeddings and the Hilbert-Schmidt Conditional Independence Criterion to devise an efficiently estimable, model-agnostic objective to practically train counterfactually invariant predictors. This choice allowed us to deal with mixed continuous/categorical, multi-dimensional variables. We demonstrated the efficacy of CIP in regression and classification tasks involving simulation studies, images, and in a fairness application on tabular data, where it outperforms existing baselines.

The main limitation of our work, shared by all studies in this domain, is the assumption that the causal graph is known. Another limitation is that our methodology is applicable only when our graphical criterion is satisfied, requiring a certain set of variables to be observed (albeit unobserved confounders are not generally excluded).

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 blocking set. Lastly, our graphical criterion and KME-based objective can also

be useful for causal representation learning, where one aims to isolate causally relevant, autonomous factors underlying the data-generating process of a given dataset.

6. Acknowledgments

This project received partial support from ELSA: European Lighthouse on Secure and Safe AI project (grant agreement No. 101070617 under UK guarantee).

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, volume 70, pp. 253–262, 2017. 13, 14

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

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020. 1

Caponnetto, A. and Vito, E. D. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. 13

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 1, 15

Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019. 18

Chiappa, S. and Pacchiano, A. Fairness with continuous optimal transport. *arXiv preprint arXiv:2101.02084*, 2021. 18, 19

Çinlar, E. and ðCınlar, E. *Probability and stochastics*, volume 261. Springer, 2011. 11

Dinculeanu, N. *Vector integration and stochastic integration in Banach spaces*, volume 48. John Wiley & Sons, 2000. 11

Drineas, P., Mahoney, M. W., and Cristianini, N. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(12), 2005. 15

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, volume 20, 2007. 12, 13, 21

Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, volume 3734, pp. 63–77, 2005. 21

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018. 1

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. URL `http://arxiv.org/abs/1312.6114`. 18

Kohavi, R. and Becker, B. Uci adult data set. *UCI Meachine Learning Repository*, 1996. 18

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017. 2, 4

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. 16

Mouli, S. C. and Ribeiro, B. Asymmetry learning for counterfactually-invariant classification in OOD tasks. In *Proc. of ICLR*, 2022. 2

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, 2018. 18

Park, J. and Muandet, K. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, pp. 21247–21259, 2020. 1, 2, 10, 11, 12

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. 1, 2, 8

Peters, J., Bühlmann, P., and Meinshausen, N. 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. 1

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2007. 13

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. 1

Shpitser, I. and Pearl, J. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008. 8

Shpitser, I. and Pearl, J. 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. 7, 8

Shpitser, I., VanderWeele, T. J., and Robins, J. M. 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. 7, 9

Si, S., Hsieh, C.-J., and Dhillon, I. Memory efficient kernel approximation. In *International Conference on Machine Learning*, pp. 701–709. PMLR, 2014. 15

Szabó, Z. and Sriperumbudur, B. K. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:233:1–233:29, 2017. 2, 3

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, pp. 16196–16208, 2021. 1, 2, 3, 4, 20, 21

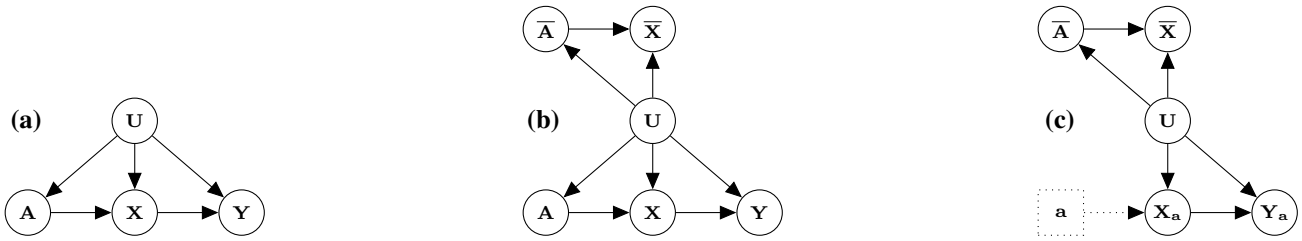

Figure 3. **(a) A causal graph** $\mathcal{G}$, which embeds information for the random variables of the model in the pre-interventional world. **(b) The corresponding graph** $\mathcal{G}'$ for the set $\mathbf{W} = \{\mathbf{A}, \mathbf{X}\}$. The variables $\overline{\mathbf{A}}$ and $\overline{\mathbf{X}}$ are *copies* of $\mathbf{A}$ and $\mathbf{X}$ respectively. **(c) The post-interventional graph** $\mathcal{G}'_{\mathbf{a}}$**.** By construction, any intervention of the form $\mathbf{A} \leftarrow \mathbf{a}$ does not affect the group $\overline{\mathbf{W}} = \{\overline{\mathbf{A}}, \overline{\mathbf{X}}\}$.

7. Proof of Theorem 3.2

7.1. 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 $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$, and define $\mathbf{Z} := (\mathbf{S} \cup \mathbf{W}) \setminus \mathbf{A}$. Then, in all SCMs compatible with $\mathcal{G}$, if the variable $\mathbf{Y}$ satisfies $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$, then $\mathbf{Y}$ 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 $\mathbf{A} \leftarrow \mathbf{a}$. Hence, we study the identification of conditional counterfactual distributions using a diagram that embeds the causal relationships between the pre- and the post-interventional world. After defining this diagram, we prove that some conditional measures in this new model provide an estimate for $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w})$. We then combine this result with the properties of $\mathbf{Z}$ to prove the desired result.

7.2. 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 7.1. 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 7.2. 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 $\mathbf{Y}$. (ii) $\mathbf{S}$ blocks all non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{G}$.

Definition 7.2 is a useful graphical criterion for the identifiability of counterfactual distributions. In fact, following Corollary 1 by (Shpitser et al., 2010), if $\mathbf{S}$ satisfies the adjustment criterion relative to $(\mathbf{A}, \mathbf{Y})$, then it holds

$$\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y}) = \int \mathbb{P}_{\mathbf{Y}|\mathbf{A},\mathbf{S}}(\mathbf{y} \mid \mathbf{a}, \mathbf{s}) d\mathbb{P}_{\mathbf{S}}. \tag{3}$$

Furthermore, this identifiability criterion is *complete*. That is, consider any graph $\mathcal{G}$ and a set of nodes $\mathbf{S}$ that do not fulfill the valid adjustment criterion with respect to $(\mathbf{A}, \mathbf{Y})$. Then, there exists a model inducing $\mathcal{G}$ such that Equation (3) does not hold (see Theorem 3 by (Shpitser et al., 2010)).

7.3. 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 7.3 and the concept of d-separation as follows.

Definition 7.3 (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 7.3 as a graphical criterion for conditional independence (Pearl, 2000).

Lemma 7.4 (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 $\mathcal{G}$.

7.4. A graphical characterization of conditional counterfactual distributions

We study the relationships between the pre-interventional model corresponding to a causal diagram $\mathcal{G}$ and the post-interventional model, inducing a diagram $\mathcal{G}_{\mathbf{a}}$ after an intervention $\mathbf{A} \leftarrow \mathbf{a}$. A natural way to study this relationship is to use the counterfactual graph (Shpitser & Pearl, 2008). However, the construction of the counterfactual graph is rather intricate. For our purposes it is sufficient to consider a simpler construction, generalizing the work by (Shpitser & Pearl, 2009).

Consider an SGM with causal graph $\mathcal{G}$, and fix a set of observed random variables of interest $\mathbf{W}$. Denote with $\mathsf{de}(\mathbf{A})$ all descendants of $\mathbf{A}$ in $\mathcal{G}$. Furthermore, for each node N of $\mathcal{G}$, denote with $\mathsf{an}(N)$ the set of all its ancestral variables. We define the corresponding graph $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ in the following steps:

1. Define $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ to be the same graph as $\mathcal{G}$.
2. For each node $N \in \mathbf{A} \cup \mathbf{W}$, add a new duplicate node $\overline{N}$ to $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$.
3. For each node $N \in \mathbf{A} \cup \mathbf{W}$ and for each ancestral variable $P \in \mathsf{an}(N) \setminus (\mathbf{A} \cup \mathbf{W})$ such that $P \in \mathsf{de}((\mathbf{A} \cup \mathbf{W}))$, add a new duplicate node $\overline{P}$ to $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$.
4. For each duplicate node $\overline{N}$ and for each parent $P \in \mathsf{pa}(N)$, if a duplicate node $\overline{P}$ was added in steps 2-3, then add an edge $\overline{P} \to \overline{N}$; otherwise add an edge $P \to \overline{N}$.
5. For each duplicate node $\overline{N}$, add an edge $U_N \to \overline{N}$.

An illustration of this graph is presented in Figure 3. We denote with $\overline{\boldsymbol{\mathcal{H}}}$ the set of duplicate nodes that were added to $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$. We can naturally define structural equations for the new variables $\overline{N}$ as

$$\overline{N} = f_N(\mathsf{pa}(\overline{N}), U_N),$$

with f_N the structural equation for N in the original model, and $\mathsf{pa}(\overline{N})$ the parents of $\overline{N}$ in the newly define graph $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$. Note that each random variable $\overline{N}$ is a *copy* of the corresponding N, in the sense that $\overline{N} = N$ almost surely. Importantly, the following lemma holds.

Lemma 7.5. *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, $\mathbf{S}$ satisfies the adjustment criterion relative to $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$ in $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$.*

Proof. We prove the claim, by showing that all non-causal paths in $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ from $\mathbf{A} \cup \mathbf{W}$ to $\mathbf{Y}$ are blocked by $\mathbf{S}$. Indeed, if $\mathbf{S}$ satisfies the adjustment criterion relative to $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$ in $\mathcal{G}$, then condition (i) of the adjustment criterion Definition 7.2 relative to $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$ in $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ is satisfied. Let π be any such non-causal path in $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ from $\mathbf{A} \cup \mathbf{W}$ to $\mathbf{Y}$. If π does not cross any duplicate node, then it is blocked by $\mathbf{S}$. Otherwise, without loss of generality, we can decompose π in three paths, which we refer to as π_1, π_2, and π_3. The path π_1 starts from a node in $\mathbf{A} \cup \mathbf{W}$ of $\mathcal{G}$, and it terminates in $\overline{\boldsymbol{\mathcal{H}}}$. The path π_2 only contains nodes in a node in $\overline{\boldsymbol{\mathcal{H}}}$, and the path π_3 starts from a node of $\overline{\boldsymbol{\mathcal{H}}}$, and it terminates in $\mathbf{Y}$. The paths π_1 and π_3 necessarily contain paths of the form $\overline{N} \leftarrow P$ or $\overline{N} \leftarrow U_N \to N$, with $\overline{N} \in \overline{\boldsymbol{\mathcal{H}}}$, P and N nodes of $\mathcal{G}$, and U_P a latent variable. By construction, no node $\overline{N} \in \overline{\boldsymbol{\mathcal{H}}}$ belongs to the adjustment set $\mathbf{S}$. Hence, the path π contains a fork of three nodes, with the central node, or any descendants of the central node, are included in $\mathbf{S}$. Hence, the path π is blocked.

□

We further prove the following lemma.

Lemma 7.6 (Following Theorem 4 by (Shpitser et al., 2010))**.** *Define the sets* $\mathbf{X} = \mathbf{W} \setminus \mathbf{A}$ *and* $\overline{\mathbf{X}} = \overline{\mathbf{W}} \setminus \overline{\mathbf{A}}$*. 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* $\mathbf{Y}^*_{\mathbf{a},\mathbf{x}} \perp\!\!\!\perp \overline{\mathbf{A}}, \overline{\mathbf{X}} \mid \mathbf{S}$ *for any intervention intervention* $\mathbf{A}, \mathbf{X} \leftarrow \mathbf{a}, \mathbf{x}$*.*

Proof. By Lemma 7.5, $\mathbf{S}$ satisfies the adjustment criterion relative to $(\mathbf{A} \cup \mathbf{W}, \mathbf{Y})$ in $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$. Equivalently, $\mathbf{S}$ satisfies the adjustment criterion relative to $(\mathbf{A} \cup \mathbf{X}, \mathbf{Y})$ in $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$. Hence, by the sufficiency of the adjustment criterion (Theorem 4 by (Shpitser et al., 2010)), it hold $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}, \mathbf{X} \mid \mathbf{S}$ in the graph $(\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}})_{\mathbf{a},\mathbf{x}}$, which is obtained from $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ by performing an intervention $\mathbf{A}, \mathbf{X} \leftarrow \mathbf{a}, \mathbf{x}$. By definition, the group of random variables $\overline{\mathbf{A}}$ and $\overline{\mathbf{X}}$ in $(\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}})_{\mathbf{a},\mathbf{x}}$ are copies of the pre-interventional variables $\mathbf{A}$, $\mathbf{X}$ in $(\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}})_{\mathbf{a},\mathbf{x}}$. It follows that $\mathbf{Y} \perp\!\!\!\perp \overline{\mathbf{A}}, \overline{\mathbf{X}} \mid \mathbf{S}$ in the graph $(\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}})_{\mathbf{a},\mathbf{x}}$ or, equivalently, that $\mathbf{Y}^*_{\mathbf{a},\mathbf{x}} \perp\!\!\!\perp \overline{\mathbf{A}}, \overline{\mathbf{X}} \mid \mathbf{S}$, as claimed. □

7.5. Proof of Theorem 3.2

We can identify conditional counterfactual distributions in $\mathcal{G}$, by identifying distributions on $\mathcal{G}'$. We can combine this observation with the notion of a valid adjustment set to derive a closed formula for the identification of the distributions of interest.

Proof of Theorem 3.2. Following the notation of Lemma 7.6, define the sets $\mathbf{X} = \mathbf{W} \setminus \mathbf{A}$, $\overline{\mathbf{X}} = \overline{\mathbf{W}} \setminus \overline{\mathbf{A}}$, and let $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$ be the augmented graph obtained by adding duplicate nodes. Note that, using this notation, the assumption that $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$ can be written as $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{X}, \mathbf{S}$. Denote with $\overline{\mathbb{P}}$ the induced measure on $\mathcal{G}'_{\mathbf{A}\cup\mathbf{W}}$. Suppose that it holds

$$\overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) = \int \mathbb{P}_{\mathbf{Y}|\mathbf{A},\mathbf{X},\mathbf{S}}(\mathbf{y} \mid \mathbf{a}', \mathbf{x}, \mathbf{s}) d\overline{\mathbb{P}}_{\mathbf{S}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{s} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) \tag{4}$$

for any intervention $\mathbf{A} \leftarrow \mathbf{a}$, and for any possible value $\mathbf{w}$ attained by $\mathbf{W}$. Assuming that Equation (4) holds, we have that

$$\begin{aligned}
\overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) &= \int \mathbb{P}_{\mathbf{Y}|\mathbf{A},\mathbf{X},\mathbf{S}}(\mathbf{y} \mid \mathbf{a}', \mathbf{x}, \mathbf{s}) d\overline{\mathbb{P}}_{\mathbf{S}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{z} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) && \text{(assuming Equation (4))} \\
&= \int \mathbb{P}_{\mathbf{Y}|\mathbf{A},\mathbf{X},\mathbf{S}}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}, \mathbf{s}) d\overline{\mathbb{P}}_{\mathbf{S}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{s} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) && (\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{X}, \mathbf{S}) \\
&= \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a},\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}). && \text{(assuming Equation (4))}
\end{aligned} \tag{5}$$

To conclude, define the set $\overline{\mathbf{T}} = \overline{\mathbf{A}} \setminus \overline{\mathbf{W}}$. It follows that

$$\begin{aligned}
\overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{W}}}(\mathbf{y} \mid \overline{\mathbf{w}}) &= \int \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) d\overline{\mathbb{P}}_{\overline{\mathbf{T}}|\overline{\mathbf{W}}}(\overline{\mathbf{t}} \mid \overline{\mathbf{w}}) && \text{(by conditioning)} \\
&= \int \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a},\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) d\overline{\mathbb{P}}_{\overline{\mathbf{T}}|\overline{\mathbf{W}}}(\overline{\mathbf{t}} \mid \overline{\mathbf{w}}) && \text{(by Equation (5))} \\
&= \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a},\mathbf{x}}|\overline{\mathbf{W}}}(\mathbf{y} \mid \overline{\mathbf{w}}). && \text{(by unconditioning)}
\end{aligned}$$

Since $\mathbf{X} \subseteq \mathbf{W}$, from the inequalities above it holds

$$\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}'}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w}) = \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w}) = \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{W}}}(\mathbf{y} \mid \overline{\mathbf{w}}) = \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a},\mathbf{x}}|\overline{\mathbf{W}}}(\mathbf{y} \mid \overline{\mathbf{w}}) = \mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w}),$$

as claimed. The proof of Theorem 3.2 thus boils down to proving Equation (4). To this end, we use the valid adjustment property of $\mathbf{S}$. Note that by Lemma 7.6 it holds $\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}} \perp\!\!\!\perp \overline{\mathbf{A}}, \overline{\mathbf{X}} \mid \mathbf{S}$. Hence,

$$\begin{aligned}
&\overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) \\
&= \int \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\overline{\mathbf{A}},\overline{\mathbf{X}},\mathbf{S}}(\mathbf{y} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}, \mathbf{s}) d\overline{\mathbb{P}}_{\mathbf{S}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{s} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) && \text{(by conditioning)} \\
&= \int \overline{\mathbb{P}}_{\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}}|\mathbf{S}}(\mathbf{y} \mid \mathbf{s}) d\overline{\mathbb{P}}_{\mathbf{S}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{s} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}) && (\mathbf{Y}^*_{\mathbf{a}',\mathbf{x}} \perp\!\!\!\perp \overline{\mathbf{A}}, \overline{\mathbf{X}} \mid \mathbf{S}) \\
&= \int \mathbb{P}_{\mathbf{Y}|\mathbf{A},\mathbf{X},\mathbf{S}}(\mathbf{y} \mid \mathbf{a}', \mathbf{x}, \mathbf{s}) d\overline{\mathbb{P}}_{\mathbf{S}|\overline{\mathbf{A}},\overline{\mathbf{X}}}(\mathbf{s} \mid \overline{\mathbf{a}}, \overline{\mathbf{x}}), && \text{(by Lemma 7.5)}
\end{aligned}$$

and Equation (4) follows. □

8. 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 $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}}$ is characteristic*[4]*,* $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A} \mid \mathbf{Z}) = 0$ *almost surely if and only if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$.

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

We first prove that if $\mathrm{HSCIC}(\mathbf{Y}, \mathbf{A} \mid \mathbf{Z}) = 0$ almost surely, then it holds $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$. To this end, consider an event $\Omega' \subseteq \Omega_{\mathbf{X}}$ that occurs almost surely, and such that it holds $(H_{\mathbf{Y},\mathbf{A}|\mathbf{X}} \circ \mathbf{X})(\omega) = 0$ for all $\omega \in \Omega'$. Fix a sample $\omega \in \Omega'$, and consider the corresponding value $\mathbf{z}_\omega = \mathbf{Z}(\omega)$, in the support of $\mathbf{Z}$. It holds

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

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

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

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

The claim follows. □

9. 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).

9.1. 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 9.1. 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

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

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

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 9.2 (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 9.2 can be used to promote conditional independence. To this end, we use the notion of a regular version.

Definition 9.3 (Regular Version, following Definition 2.4 by (Çinlar & ðCınlar, 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 9.2 can be used to promote conditional independence.

Theorem 9.4 (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 9.4. In this work, HSCIC is not used for conditional independence testing but as a conditional independence measure.

9.2. 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 9.5 (Following Proposition 2.5 by (Çinlar & ðCınlar, 2011))**.** *Following the notation introduced in Definition 9.3, 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 $\mathbb{E}[k(\cdot, \mathbf{y}) \mid \mathbf{Z}]$.

As a consequence of Theorem 9.5, we prove the following result.

Lemma 9.6. *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 9.5 there exists an event $\Omega \subseteq \Omega_{\mathbf{Z}}$ that occurs almost surely such that

$$\mu_{\mathbf{Y}|\mathbf{Z}}(\omega) = \mathbb{E}[k(\mathbf{y}, \cdot) \mid \mathbf{Z}](\omega) = \int k(\mathbf{y}, \cdot) dQ_\omega(\mathbf{y}), \tag{6}$$

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

$$\begin{aligned}
\mu_{\mathbf{Y}|\mathbf{Z}}(\omega) &= \int k(\mathbf{x}, \cdot) dQ_\omega(\mathbf{x}) && \text{(it follows from Equation (6))} \\
&= \int k(\mathbf{x}, \cdot) d\mathbb{P}_{\mathbf{X}|\mathbf{A}}(\mathbf{x} \mid \{\mathbf{A} = \mathbf{A}(\omega)\}) && (Q_\omega(\mathbf{y}) = \mathbb{P}_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \{\mathbf{Z} = \mathbf{Z}(\omega)\})) \\
&= \mu_{\mathbf{X}|\{\mathbf{A}=\mathbf{A}(\omega)\}}, && \text{(by definition as in Section 2)}
\end{aligned}$$

as claimed. □

As a consequence of Lemma 9.6, we can prove that the definition of the HSCIC by (Park & Muandet, 2020) is equivalent to ours. The following corollary holds.

Corollary 9.7. *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 9.7.

10. Conditional independence and 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 10.1. *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 10.2 (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 10.1.

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

$$\Lambda_{\mathbf{Y},\mathbf{A}|\mathbf{Z}} \coloneqq \Lambda_{\mathbf{Y},\mathbf{A}} - \Lambda_{\mathbf{Y},\mathbf{Z}} \circ \Lambda_{\mathbf{Z},\mathbf{A}}. \tag{7}$$

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 10.3 (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*

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

Here, $\Lambda_{\mathbf{Y},\ddot{\mathbf{A}}|\mathbf{Z}}$ *is an operator defined as in Equation* (7)*.*

By Theorem 10.3, 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 10.3 requires to embed the variable $\mathbf{Z}$ in an RKHS. In contrast, CIP only requires the embedding of the variables $\mathbf{Y}$ and $\mathbf{A}$.

11. Estimating the HSCIC from samples.

Given n samples $\{(\hat{\mathbf{y}}_i, \mathbf{a}_i, \mathbf{z}_i)\}_{i=1}^n$, denote with $\hat{K}_{\hat{\mathbf{Y}}}$ the kernel matrix with entries $[\hat{K}_{\hat{\mathbf{Y}}}]_{i,j} \coloneqq k_{\mathbf{Y}}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j)$, and let $\hat{K}_{\mathbf{A}}$ be the kernel matrix for $\mathbf{A}$. We estimate the $H_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{X}} \equiv H_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{X}}(\cdot)$ as

$$\begin{aligned}\hat{H}^2_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}} = \hat{w}^T_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}} \left(\hat{K}_{\hat{\mathbf{Y}}} \odot \hat{K}_{\mathbf{A}}\right) \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}} \qquad (8)\\ - 2\left(\hat{w}^T_{\hat{\mathbf{Y}}|\mathbf{Z}} \hat{K}_{\mathbf{Y}} \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{X}}\right)\left(\hat{w}^T_{\mathbf{A}|\mathbf{X}} \hat{K}_{\mathbf{A}} \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}}\right)\\ + \left(\hat{w}^T_{\hat{\mathbf{Y}}|\mathbf{Z}} \hat{K}_{\hat{\mathbf{Y}}} \hat{w}_{\hat{\mathbf{Y}}|\mathbf{Z}}\right)\left(\hat{w}^T_{\mathbf{A}|\mathbf{Z}} \hat{K}_{\mathbf{A}} \hat{w}_{\mathbf{A}|\mathbf{Z}}\right) , \qquad (9)\end{aligned}$$

where $\odot$ is element-wise multiplication. The functions $\hat{w}_{\hat{\mathbf{Y}}|\mathbf{Z}} \equiv \hat{w}_{\hat{\mathbf{Y}}|\mathbf{Z}}(\cdot)$, $\hat{w}_{\mathbf{A}|\mathbf{Z}} \equiv \hat{w}_{\mathbf{A}|\mathbf{Z}}(\cdot)$, and $\hat{w}_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}} \equiv \hat{w}_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}}(\cdot)$ are found via kernel ridge regression. (Caponnetto & Vito, 2007) provide the convergence rates of the estimand $\hat{H}^2_{\hat{\mathbf{Y}},\mathbf{A}|\mathbf{Z}}$ under mild conditions. In practice, computing the HSCIC approximation by the formula in Equation (8) can be computationally expensive. To speed it up, we can use random Fourier features to approximate the matrices $\hat{K}_{\hat{\mathbf{Y}}}$ and $\hat{K}_{\mathbf{A}}$ (Rahimi & Recht, 2007; Avron et al., 2017).

12. 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 12.1 (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* $\mathbb{P}_k(\boldsymbol{\eta})$ *such that*

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

Since both the kernel k and the probability distribution $\mathbb{P}_k$ are real-valued functions, the integrand in Theorem 12.1 ca be replaced by the function $\cos \boldsymbol{\eta}^T(\mathbf{z} - \mathbf{z}')$, and we obtain the following formula

$$k(\mathbf{z}, \mathbf{z}') = \int \cos \boldsymbol{\eta}^T(\mathbf{z} - \mathbf{z}') d\mathbb{P}_k = \mathbb{E}\left[\cos \boldsymbol{\eta}^T(\mathbf{z} - \mathbf{z}')\right], \qquad (10)$$

where the expected value is taken with respect to the distribution$\mathbb{P}_k(\boldsymbol{\eta})$. This equation allows to approximate the kernel $k(\mathbf{z}, \mathbf{z}')$, via the empirical mean of points $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_l$ sampled independently according to $\mathbb{P}_k$. In fact, it is possible to prove exponentially fast convergence of an empirical estimate for $\mathbb{E}\left[\cos \boldsymbol{\eta}^T(\mathbf{z} - \mathbf{z}')\right]$, as shown in the following theorem.

Theorem 12.2 (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* $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_l$ *sampled independent according to the distribution* $\mathbb{P}_k$*. Define the function*

$$\hat{k}(\mathbf{z}, \mathbf{z}') \coloneqq \frac{1}{l} \sum_{j=1}^{l} \cos \boldsymbol{\eta}_j^T(\mathbf{z} - \mathbf{z}'),$$

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 12.2, 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

$$\zeta_{k,l}(\mathbf{z}) \coloneqq \frac{1}{\sqrt{l}}\left[\cos \boldsymbol{\eta}_1^T\mathbf{z}, \ldots, \cos \boldsymbol{\eta}_l^T\mathbf{z}\right] \tag{11}$$

with $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_l$ sampled independent according to the distribution $\mathbb{P}_k$.

We can approximate the Kernel matrix using the functions defined as in Equation (11). 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 12.3. 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 $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_l$ is sufficiently large.

Theorem 12.4 (Theorem 7 by (Avron et al., 2017))**.** *Fix a constant* $\gamma \leq 1/2$*. Consider* n *samples* $\mathbf{z}_1, \ldots, \mathbf{z}_n$*, 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* $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_l$ *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 12.2, with $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_l \sim \mathcal{N}(0, \Sigma)$, with $\Sigma \coloneqq \sigma^{-1}I$, with I the identity matrix. The functions $\zeta_{k,l}(\mathbf{z})$ can be defined accordingly.

13. Experiment settings

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 Equation (1) for continuous outcomes $\mathbf{Y}$. We generate 10k samples from the observational distribution in each setting and use an 80 to 20 train to test split. All metrics reported are on the test set. We perform hyper-parameter tuning for MLP hyperparameters based on a random strategy (see Section 13 for details). The $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z})$ term is computed as in Equation (8) 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. Additional information on the experiments is now provided.

13.1. Dataset for model performance with the use of the HSCIC

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

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

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

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

where $\varepsilon_{\mathbf{A}} \sim \mathcal{N}(0,1)$ and $\varepsilon_{\mathbf{Y}}, \varepsilon_{\mathbf{X}} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1)$.
In the first experiment, Figure 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 300 epochs were used.
The kernel ridge regression estimation generally requires $O(n^3)$ and $O(n^2)$ complexity, with n the size of the dataset. However, these bounds can be significantly improved by using, i.e., Fourier Features (see, 12). By using Fourier features, the resulting approximate kernel ridge regression estimator can be computed in $O(ns^2)$ and $O(ns)$ memory. Here, s is a parameter determining the accuracy of the approximation. In practice, s can be set to be significantly smaller than the problem size, resulting in a dramatic speed-up. Other methods for efficient kernel computation include the popular Nystrom approximation (Drineas et al., 2005), (Chen et al., 2020), and Memory-Efficient Kernel Approximation (MEKA) (Si et al., 2014).

13.2. 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 Figure 2 (Scenario 1) is the following:

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

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

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

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

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

$$\mathbf{X} = \left(\frac{1}{2}\mathbf{Z} + \mathbf{A}\right)\cdot\varepsilon_{\mathbf{X}}$$

$$\mathbf{Y} = \sin(\mathbf{Z}) + \mathbf{A} + \mathbf{X} + \frac{1}{5}\varepsilon_{\mathbf{Y}},$$

where $\varepsilon_{\mathbf{A}}, \varepsilon_{\mathbf{X}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\varepsilon_{\mathbf{Y}} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1)$. Figure 4 shows the performance of CIP against baselines CF1 and CF2 in Scenario 2. In Table 1, the results of MSE, HSCIC and VCF are presented. In this table, both Scenario 1 and Scenario 2 were considered. The results shown in Figure 4 and Table 1 are 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

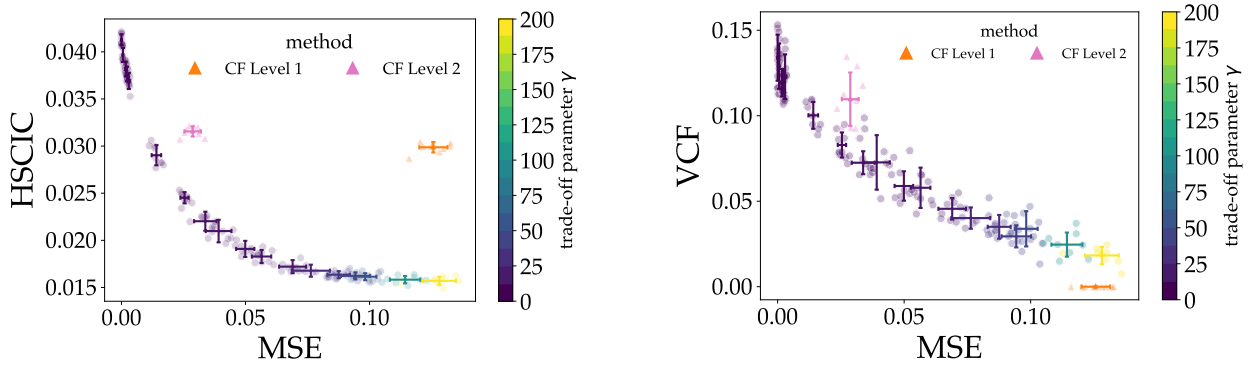

Figure 4. **Results of MSE, HSCIC operator and VCF in comparison with CF1 and CF2 for Scenario 2**. The plot shows the results for 10 different seeds, along with the mean and standard deviations. CF2 is Pareto-dominated by the VCF-MSE frontier, we can hence pick a γ value to outperform CF2 in both accuracy and counterfactual invariance simultaneously.

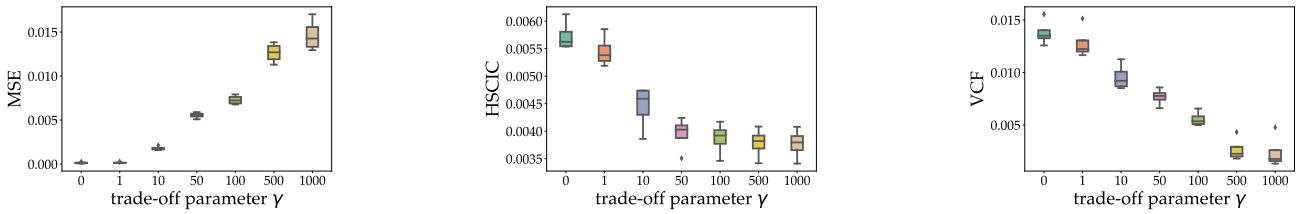

Figure 5. **Results of MSE, HSCIC operator and VCF for multi-dimensional variable experiment with dimA** $= 5$.

used for the prediction of $\hat{\mathbf{Y}}$ and the one used for the prediction of the $\mathbf{X}$ 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.

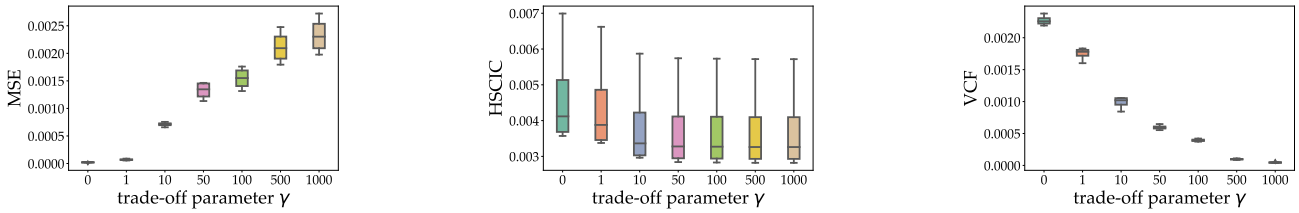

Figure 6. **Results of MSE, HSCIC operator and VCF for multi-dimensional variable experiment with dimA** $= 10$.

13.3. Image Experiment

We consider the image classification task on the dSprites dataset (Matthey et al., 2017). Since this dataset is fully synthetic and labelled, we consider a causal model as depicted in Figure 10(right). The full structural equations are provided later, where we assume a causal graph over the determining factors of the image, and essentially look up the corresponding image in the simulated dataset. This experiment is particularly challenging due to the mixed categorical and continuous variables in $\mathbf{C}$ (`shape`, `y-pos`) and $\mathbf{X}$ (`color`, `orientation`), continuous $\mathbf{A}$ (`x-pos`). Our goal is to learn a predictor $\hat{\mathbf{Y}}$ that is counterfactually invariant in the x-position with respect to all other observed variables. Following Theorem 3.2, we seek to achieve $\hat{\mathbf{Y}} \perp\!\!\!\perp$ `x-pos` $\mid \{$`shape`, `y-pos`, `scale`$\}$ via the HSCIC operator. To accommodate the mixed input types, $\hat{\mathbf{Y}}$ puts an MLP on top of features extracted from the images via convolutional layers concatenated with features extracted from the remaining inputs via an MLP. Figure 8 demonstrates that HSCIC achieves improved VCF as γ increases up to a certain point while affecting MSE, an inevitable trade-off.

Table 1. **Performance of the HSCIC against baselines** CF1 and CF2 on two synthetic datasets. Notably, for γ within $[1, 5]$ in Scenario 1 CIP outperforms CF2 in **MSE** and **VCF** simultaneously. Similarly, in Scenario 2 this holds for γ within $[1, 2]$.

	Scenario 1			**Scenario 2**		
	MSE $\times 10^3$	**HSCIC** $\times 10^3$	**VCF** $\times 10^3$	**MSE** $\times 10^3$	**HSCIC** $\times 10^2$	**VCF** $\times 10^2$
$\gamma = 0$	0.01 ± 0.00	35.22 ± 0.87	30.37 ± 0.94	0.01 ± 0.01	4.12 ± 0.05	13.39 ± 1.35
$\gamma = 0.1$	0.05 ± 0.01	34.54 ± 0.85	29.32 ± 1.74	0.04 ± 0.01	4.10 ± 0.06	13.34 ± 1.41
$\gamma = 0.2$	0.24 ± 0.07	33.50 ± 1.10	27.67 ± 0.88	0.10 ± 0.02	4.07 ± 0.06	12.67 ± 0.68
$\gamma = 0.3$	0.61 ± 0.08	32.01 ± 1.11	25.93 ± 1.88	0.21 ± 0.04	4.03 ± 0.06	12.64 ± 0.72
$\gamma = 0.4$	1.28 ± 0.10	30.36 ± 1.13	24.20 ± 1.72	0.49 ± 0.05	4.01 ± 0.07	12.44 ± 1.20
$\gamma = 0.5$	2.36 ± 0.25	28.13 ± 1.10	20.32 ± 2.08	0.59 ± 0.13	3.97 ± 0.09	12.44 ± 0.73
$\gamma = 0.6$	3.70 ± 0.20	25.69 ± 0.78	20.01 ± 2.60	0.84 ± 0.15	3.90 ± 0.07	12.12 ± 0.80
$\gamma = 0.7$	5.10 ± 0.26	23.56 ± 0.62	18.96 ± 2.44	1.24 ± 0.27	3.87 ± 0.08	12.09 ± 0.74
$\gamma = 0.8$	6.39 ± 0.30	21.86 ± 0.75	18.08 ± 3.01	1.73 ± 0.35	3.81 ± 0.08	11.93 ± 0.70
$\gamma = 0.9$	7.72 ± 0.60	22.00 ± 0.83	16.57 ± 3.22	2.21 ± 0.46	3.76 ± 0.08	11.90 ± 0.70
$\gamma = 1.0$	9.11 ± 0.60	18.87 ± 0.81	14.58 ± 1.62	2.96 ± 0.42	3.69 ± 0.08	11.28 ± 1.30
$\gamma = 2.0$	17.29 ± 0.92	13.05 ± 0.43	4.03 ± 1.67	14.09 ± 1.91	2.90 ± 0.10	10.22 ± 0.73
$\gamma = 3.0$	20.73 ± 0.77	11.60 ± 0.30	1.46 ± 1.11	25.42 ± 1.62	2.42 ± 0.11	8.29 ± 0.67
$\gamma = 4.0$	22.27 ± 0.99	11.17 ± 0.32	0.76 ± 0.24	33.80 ± 4.52	2.20 ± 0.05	7.25 ± 0.86
$\gamma = 5.0$	23.17 ± 0.98	10.94 ± 0.30	0.50 ± 0.22	39.16 ± 5.15	2.09 ± 0.10	7.27 ± 1.59
$\gamma = 7.0$	24.48 ± 1.07	10.70 ± 0.32	0.46 ± 0.13	49.90 ± 3.67	1.90 ± 0.11	5.89 ± 0.86
$\gamma = 10.0$	25.40 ± 1.09	10.58 ± 0.32	0.24 ± 0.08	56.49 ± 3.88	1.82 ± 0.07	5.79 ± 1.27
$\gamma = 50.0$	28.70 ± 1.13	10.37 ± 0.32	0.13 ± 0.09	98.23 ± 4.53	1.61 ± 0.03	3.39 ± 1.03
$\gamma = 100.0$	29.54 ± 1.27	10.36 ± 0.32	0.01 ± 0.01	114.3 ± 6.67	1.58 ± 0.03	2.46 ± 0.50
CF1	25.50 ± 0.98	14.68 ± 0.05	0	125.8 ± 5.64	2.98 ± 0.05	0
CF2	23.39 ± 1.39	16.57 ± 0.10	6.45 ± 4.32	28.71 ± 2.38	3.16 ± 0.05	10.96 ± 1.56

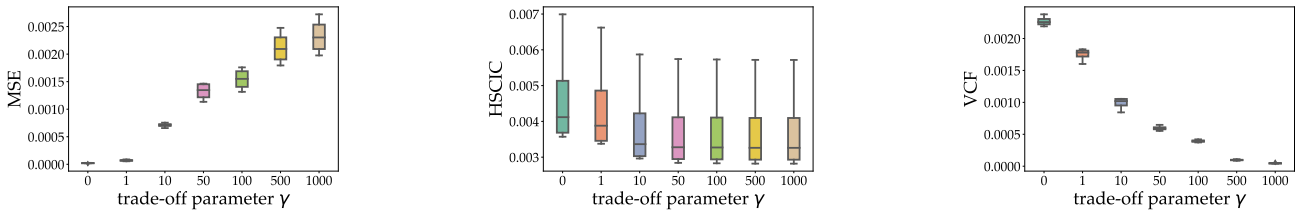

Figure 7. **Results of MSE, HSCIC operator and VCF for multi-dimensional variable experiment with dimA** $= 20$.

The simulation procedure for the results shown in Section 13.3 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 3, 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. In Figure 9 the results are presented for higher values of γ, with a specific emphasis on the interplay between accuracy and counterfactual invariance. The means and standard deviations corresponding to 8 seeds can be found in Table 2. As evidenced by the results for $\gamma = 500$,

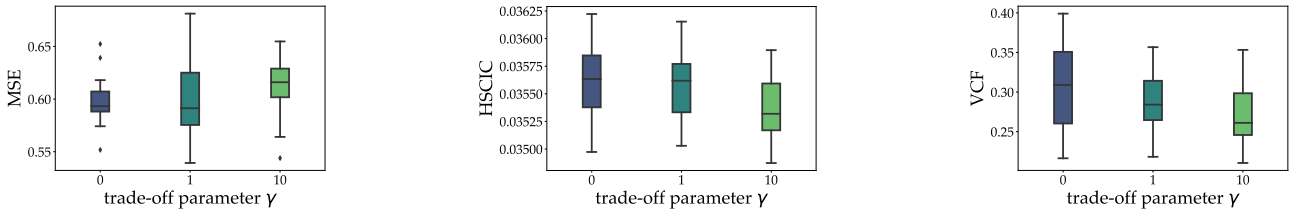

Figure 8. **Results of MSE, HSCIC operator and VCF for the dSprites image dataset experiment**. The HSCIC operator decreases steadily with higher values of γ. Similarly, a necessary increase of MSE can be observed. For both $\gamma = 1$ and $\gamma = 10$ an overall decrease of VCF is observed compared to the unconstrained setting. Boxes represent the interquartile range (IQR), the horizontal line is the median, and whiskers show the minimum and maximum values, excluding outliers. Outliers are represented as dots. The results correspond to 12 seeds.

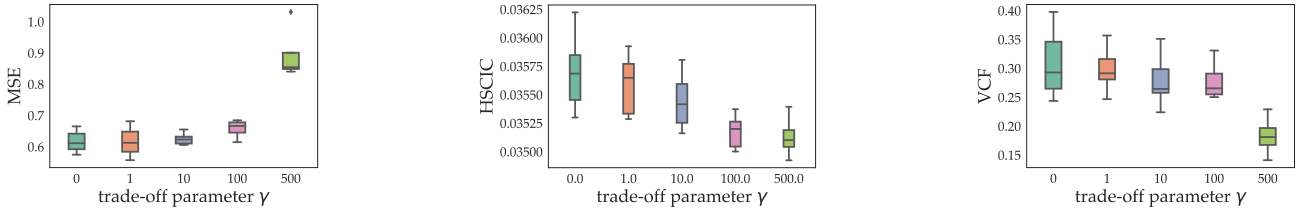

Figure 9. **Results of MSE, HSCIC operator and VCF for the dSprites image dataset experiment.** The HSCIC operator decreases with higher values of γ. Similarly, a necessary increase of MSE can be observed. A decrease of VCF is observed compared to the unconstrained setting.

there is a clear trade-off between these two factors, with a notable loss in accuracy leading to a significant improvement in counterfactual invariance, as indicated by the low VCF metric.

13.4. Fairness with continuous protected attributes

We then apply CIP to the widely-used UCI Adult dataset (Kohavi & Becker, 1996) and we compare it against CF1 and CF2. The goal of this task is to predict whether an individual's income is above a certain threshold based on demographic information, including protected attributes. We follow (Nabi & Shpitser, 2018; Chiappa, 2019), where a subset of variables are selected from the dataset and a causal structure is assumed as in Figure 1(e) (see Section 13.4 and Figure 10 for details). We choose gender (considered binary in this dataset) and age (considered continuous) as the protected attributes $\mathbf{A}$. We denote the marital status, level of education, occupation, working hours per week, and work class jointly by $\mathbf{X}$ and combine the remaining observed attributes in $\mathbf{C}$.

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. Hence, we follow (Chiappa & Pacchiano, 2021) and estimate a possible true SCM by inferring the posterior distribution over the unobserved variables using variational autoencoders (Kingma & Welling, 2014). Figure 11 (left) highlights once more that the HSCIC operator is in agreement with the VCF, again trading off accuracy. Furthermore, we observe that that the HSCIC has better accuracy than CF1 and CF2, and better VCF then CF2. Figure 11 (right) presents the counterfactual distribution (i.e., Equation (2) before taking the outer expectation) for one seed for different trade-off parameters. It shows that CIP achieves more counterfactually fair outcome distributions (more mass of the VCF distribution near zero) than an unconstrained classifier ($\gamma = 0$).

The pre-processing of the UCI Adult dataset was based upon the work of (Chiappa & Pacchiano, 2021). Referring to the causal graph in Figure 10, 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 $\mathbb{P}(\mathbf{H_m}|V)$, $\mathbb{P}(\mathbf{H_r}|V)$, $\mathbb{P}(\mathbf{H_l}|V)$ 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 function. The training loss of the variational autoencoder consists of a reconstruction term (Mean-Squared Error for continuous variables and Cross-Entropy

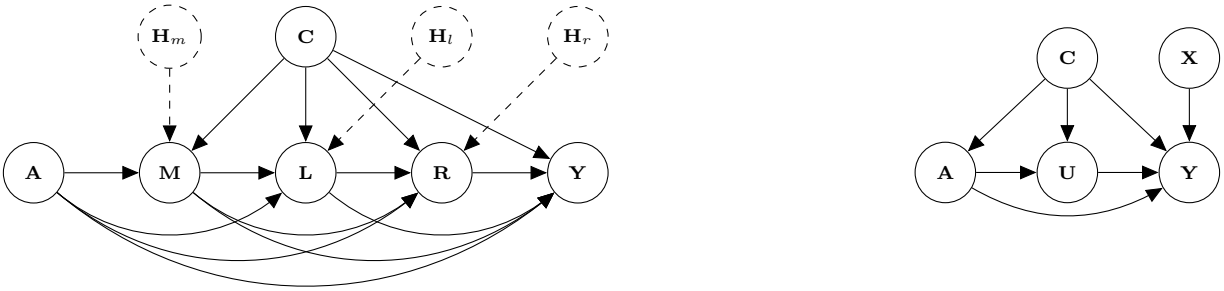

Figure 10. **(Left) Assumed causal graph for the Adult dataset**, as in (Chiappa & Pacchiano, 2021). The variables $\mathbf{H}_m$, $\mathbf{H}_l$, $\mathbf{H}_r$ are unobserved, and jointly trained with the predictor $\hat{\mathbf{Y}}$. **(Right) Causal structure for the constructed dSprites ground truth**, where $\mathbf{A} = \{\text{Pos.X}\}$, $\mathbf{U} = \{\text{Scale}\}$, $\mathbf{C} = \{\text{Shape}, \text{Pos.Y}\}$, $\mathbf{X} = \{\text{Color}, \text{Orientation}\}$, and $\mathbf{Y} = \{\text{Outcome}\}$.

Table 2. **Results of MSE, HSCIC and VCF for the dSprites image dataset experiment.** The results present the mean and standard deviation for 8 seeds.

	MSE $\times 10^1$	**HSCIC** $\times 10^3$	**VCF** $\times 10^2$
$\gamma = 0$	6.07 ± 0.26	35.79 ± 0.31	3.15 ± 0.43
$\gamma = 1$	6.15 ± 0.23	35.55 ± 0.24	2.98 ± 0.35
$\gamma = 10$	6.24 ± 0.17	35.44 ± 0.25	2.80 ± 0.44
$\gamma = 100$	6.57 ± 0.27	35.17 ± 0.13	2.77 ± 0.30
$\gamma = 500$	8.95 ± 0.64	35.13 ± 0.18	1.82 ± 0.33

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 Figure 10(Left), $\mathbf{A}$ includes the variables age and gender, $\mathbf{C}$ includes nationality and race, $\mathbf{M}$ marital status, $\mathbf{L}$ 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 Equation (1) but Binary Cross-Entropy loss ($\mathcal{L}_{\mathbf{BCE}}$) is used instead of Mean-Squared Error loss:

$$\mathcal{L}_{\text{CIP}}(\hat{\mathbf{Y}}) = \mathcal{L}_{\mathbf{BCE}}(\hat{\mathbf{Y}}) + \gamma \cdot \text{HSCIC}\left(\hat{\mathbf{Y}}, \{\text{Age}, \text{Gender}\} \middle| \mathbf{Z}\right) , \tag{12}$$

where the set $\mathbf{S} = \{\text{Race}, \text{Nationality}\}$ blocks all the non-causal paths from $\mathbf{W} \cup \mathbf{A}$ to $\mathbf{Y}$ and $\mathbf{Z} = (\mathbf{S} \cup \mathbf{W}) \setminus \mathbf{A}$. In this example we have $\mathbf{W} = \{\mathbf{C} \cup \mathbf{M} \cup \mathbf{L} \cup \mathbf{R}\}$. The results in Figure 11 (center, right) refer to one run with conditioning set $\mathbf{S} = \{\text{Race}, \text{Nationality}\}$. The results in Figure 11 (left) are the average and standard deviation of four random seeds.

	Accuracy (%)	**HSCIC** $\times 10^2$	**VCF** $\times 10^2$
$\gamma = 0$	83.8 ± 0.2	3.65 ± 0.05	6.53 ± 0.96
$\gamma = 0.5$	82.7 ± 0.4	3.53 ± 0.01	5.96 ± 0.92
$\gamma = 1$	82.9 ± 0.4	3.54 ± 0.01	4.39 ± 0.62
$\gamma = 10$	82.0 ± 0.5	3.09 ± 0.01	0.29 ± 0.01
CF1	75.02 ± 0.0	0.02 ± 0.00	0.00 ± 0.00
CF2	75.81 ± 0.0	0.03 ± 0.00	0.54 ± 0.00

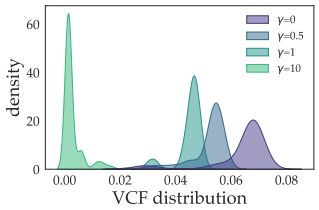

Figure 11. **(Left)** Results on accuracy, HSCIC and VCF, showing a strong decrease in VCF as γ increases at the cost of only a moderate drop in accuracy. We further observe that that the HSCIC has overall better accuracy, and better VCF then CF2. **(Right)** Distribution of VCF values (unnormalized) for different choices of γ for one seed. We observe less variance and more mass near zero for $\gamma \geq 0$. Notably, for $\gamma = 10$ we have substantial increase in counterfactual invariance, as evidenced by the values in the Table.

Table 3. **Architecture of the convolutional neural network** used for the image dataset.

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

13.5. Baseline Experiments

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 Figure 1(b)):

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

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

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

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 Figure 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 Figure 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 Figure 1(b), CIP

Table 4. **Results of the MSE, HSCIC, VCF of CIP and the baseline (Veitch et al., 2021)** applied to the causal and anti-causal structure in Figure 1(b-c). Although the graphical assumptions are not satisfied, CIP shows an overall decrease of HSCIC, 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 $\times 10^2$	**HSCIC** $\times 10^2$	**VCF**	**MSE** $\times 10^2$	**HSCIC** $\times 10^3$	**VCF**
$\gamma = 0.5$	4.48 ± 0.31	3.60 ± 0.21	0.19 ± 0.02	4.50 ± 0.40	4.54 ± 0.15	0.19 ± 0.02
$\gamma = 1.0$	5.00 ± 0.36	3.43 ± 0.12	0.17 ± 0.01	5.45 ± 0.41	4.42 ± 0.13	0.18 ± 0.02

	CIP			**(Veitch et al., 2021)**		
	MSE $\times 10^2$	**HSCIC** $\times 10^2$	**VCF**	**MSE** $\times 10^2$	**HSCIC** $\times 10^3$	**VCF**
$\gamma = 0.5$	1.16 ± 0.01	3.22 ± 0.16	1.49 ± 0.16	1.01 ± 0.01	4.55 ± 0.22	1.71 ± 0.26
$\gamma = 1.0$	1.37 ± 0.02	3.20 ± 0.16	1.28 ± 0.19	0.99 ± 0.01	4.54 ± 0.22	1.88 ± 0.28

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

	VCF $\times 10^2$	**MSE** $\times 10^2$
data augmentation	3.12 ± 0.16	0.003 ± 0.001
causal-based data augmentation	3.04 ± 0.16	0.013 ± 0.012
CIP$_{\gamma=0.5}$	1.550 ± 0.13	0.044 ± 0.022
CIP$_{\gamma=1}$	0.95 ± 0.19	0.19 ± 0.072

enforces $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{X})$ to become small, gradually enforcing $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A}|\mathbf{X}$. Differently, (Veitch et al., 2021) enforces as independence criterion $\text{HSIC}(\hat{\mathbf{Y}}, \mathbf{A})$. 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)). In the anti-causal graphical setting presented in Figure 1(c), the objective term used in CIP is again $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{X})$, 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, $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{X}, \mathbf{Z})$ 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.

13.6. Comparison Heuristic Methods Experiments

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 Section 13. 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 $\mathbf{Z}, \mathbf{X}, \mathbf{Y}$ **unchanged**. Differently, in the latter *causal-based data augmentation* method, we also take into account the causal structure given by the known DAG. Indeed, when manipulating the variable $\mathbf{A}$, its descendants (in this example $\mathbf{X}$) will also change. In this experiment, a predictor for $\mathbf{X}$ as $\hat{\mathbf{X}} = f_\theta(\mathbf{A}, \mathbf{Z})$ 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 $\mathbf{X}$ 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

Table 6. Results of MSE and **HSCIC** of the benchmarks CF1 and CF2 on the UCI Adult dataset for three random seeds.

	MSE	**HSCIC**	**VCF**
Kusner Level 1	75.02	0.02	0
Kusner Level 2	75.81	0.03	0.54

guarantee to provide counterfactually invariant predictors. The results of an empirical comparison are shown in Table 5. 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 $\gamma = 1$).

In these experiments, a dataset of $n = 3000$ is used, along with $k = 500$ and $d = 500$. The architecture used for predicting $\mathbf{X}$ 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.