ROBUST CAUSAL MODELING BY KERNEL-BASED CONDITIONAL INDEPENDENCE

András Strausz*
ETH Zürich
strausza@ethz.ch

Zhiheng Lyu*
University of Hong Kong
zhlyu@cs.hku.hk

Bernhard Schölkopf
MPI for Intelligent Systems
bs@tue.mpg.de

Zhijing Jin
MPI & ETH Zürich
jinzhi@ethz.ch

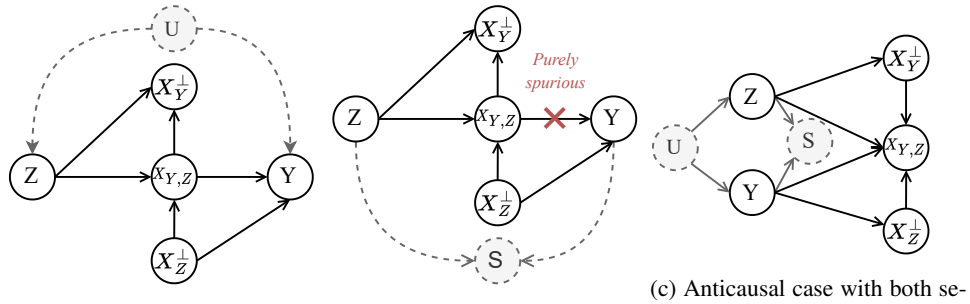
1 INTRODUCTION

The robustness of a predictor f is typically characterized by its ability to remain invariant to certain perturbations in the data. Numerous studies have defined frameworks for perturbations, proposed robustness metrics, and introduced techniques for training predictors that uphold these definitions. However, due to the lack of precise frameworks, and measures of independence for continuous, high-dimensional variables, such approaches often fall short of practical usability. For example, in the NLP domain, a common task is to remain invariant to the sentiment of some sentence when predicting it's usefulness. However, approaches for learning robust predictors for this task often made two crucial assumptions. Firstly, the choice of following either the *Equality of Odds* (Selbst et al., 2019) or *Demographic Parity* (Barocas & Selbst, 2016) framework was arbitrary. Secondly, in order to achieve these, works considered sentiment as a binary variable, which enables the use of simpler distance measures between probabilities. In this work we aim to find a more principled way to robustness with the help of causal learning and discuss how this relates to former works. Moreover, we review former works und current approaches for ensuring independence of high-dimensional variables. Lastly, we carry out experiments to measure the practical effectiveness of the proposed methods.[1]

Robustness can be formulated as an invariance criteria,for example as $f(X(z)) = f(X(z'))$, where $X(z)$ denotes that the data depends on some Z where $z, z' \sim Z$ are realizations of Z. However, this conceptual definition leaves room for how we define $P(Z)$, the relationship between Z and X, as well as how we quantify robustness. To this end, several frameworks have been researched, that connects to robustness and generally define some parts of these design choices:

1. Adversarial perturbations (Madry et al., 2018; Goodfellow et al., 2014b): This line of research quantifies robustness by identifying the smallest perturbation to the data that changes the prediction, typically in classification tasks. The data is commonly altered by adding Gaussian noise to either a specific dimension or all dimensions, though methods may vary across studies. In its simplest form, we can summarize adversarial perturbations as methods where $Z \sim \mathcal{N}(0, 1)$ and $X(z') = X + z'$. The optimization problem can be formulated as a min-max problem over $\mathcal{F}$ and $||z||$. On the other hand, the resulting data samples are often unnatural, thus predictive performance may be lost in order to gain robustness against samples that do not arise in the real world.
2. Out-of-domain testing (Hendrycks & Gimpel, 2017): here, Z is generally taken as a categorical random variable where different realizations refer to different domains. While the relation between X and Z is left undefined, robustness is measured as difference in performance between domains that have used for training and such that are unknown for the model.
3. Stress testing (Veitch et al., 2021): In this category, one explicitly defines Z and how Z effects X, even if this is often done on a high level. For example, Z can denote the

*Equal contribution.
[1]The project is open-sourced at https://github.com/strausza/CILearning



(a) Causal direction with confounding (b) Causal direction with selection (c) Anticausal case with both selection and confounding

Figure 1: Three cases discussed in Theorem 3.2 in Veitch et al. **?**

sentiment of a sentence and realizations may be $X(z)$ =*The pizza was* ***great***, $X(z')$ =*The pizza was* ***bad***. Measuring robustness similarly varies over works, and depends on whether one can generate counterfactual data samples.

In this work we focus on robustness in the natural sense, where perturbed examples are not adversially generated but following some natural law. We discuss works for all parts: (Section 2) how X and Z are related and how Z may be correlated with Y, (Section 3) what are ways to learn a predictor f that is invariant w.r.t. Z given the framework, and finally (Section 4) we show empirically how the different design choices, especially the choice of $p(Z)$ effect all former parts and thus the evaluation of a robust predictor.

2 A CAUSAL FRAMEWORK FOR ROBUST LEARNING

Whilst the target of robustness must be defined for every domain specifically, Veitch et al. (2021) note that on a high-level, the data generating process inherently implies certain independence relationships between the prediction $Y = f(X)$ and the confounder Z, given that f is invariant w.r.t Z. In specific, the authors conceptually define $X_Y^\perp$, $X_Z^\perp$ and $X_{Y \wedge Z}$ that denote the "parts of the data" that is independent of Y, of Z or influenced by both of them respectively. Moreover, they discuss the cases where X causes Y (causal) and where Y causes X (anti-causal). They also differentiate between the type of non-causal associations between the confound variable Z and Y. The resulting CSMs are depicted in Figure 1. The authors use this conceptual separation of X and define counterfactual invariance as follows:

Proposition 1 (Counterfactually invariant predictor) *A predictor f is counterfactually invariant to Z if $f(X(z)) = f(X(z'))$ for $z, z' \in Z$ holds almost everywhere. Moreover, f is counterfactually invariant if and only if f is $X_Z^\perp$-measurable.*

By reading d-separation, the following independence relations can be deduced, assuming that f is in fact counterfactually invariant:

Proposition 2 (Implications of a counterfactually invariant predictor) *Assume that $f(X)$ is a counterfactually invariant predictor:*

1. *Under the anticausal graph, $f(X) \perp\!\!\!\perp Z|Y$.*
2. *Under the causal graph, if Y and Z are not subject to selection (but possibly confounded), $f(X) \perp\!\!\!\perp Z$*
3. *Under the causal graph, if the association between Y and Z satisfies that $Y \perp\!\!\!\perp X|X_z^\perp, Z$ (called purely spurious) and Y and Z are not subject to confounding (but possibly to selection), $f(X) \perp\!\!\!\perp Z|Y$*

A direct way to incorporate this signature of counterfactual invariance into learning is to regularize the learning objective accordingly. Veitch et al. (2021) also follows this method and show that en-

forcing the signature leads to increased robustness using both synthetic and real datasets. However, they only discuss classification problems with binary Z. This simplifies the independence criterion as this can now be computed as the distance between the respective conditional probabilities. To this end, they use MMD (Gretton et al., 2012).

2.1 WEAKENING COUNTERFACTUAL INVARIANCE

It is worth to mention, that counterfactual invariance must not necessarily be defined as equality of random variables (i.e. $f(X(z)) = f(X(z'))$ for $z, z' \in Z$) but one may settle on equality in probability (i.e. $P_Z(f(X(z))) = P_{Z'}(f(X(z')))$). Whether this formulation allows for sufficient conditions is an open question, an approach to this can be found in Quinzan et al. (2023) (under work at the time of the publication of this report). Another approach is to weaken the invariance criteria by restricting the invariance to certain transformations of the data (Mouli & Ribeiro, 2022).

3 INDEPENDENCE MEASURES FOR LEARNING

In the following we introduce three metrics to measure conditional independence for continuous variables. In particular, we are interested in a metric that satisfies $M(X, Y, Z) = 0$ if and only if $X \perp\!\!\!\perp Z|Y$. In the context of learning, we aim to achieve $\hat{Y} \perp\!\!\!\perp Z|Y$ where $\hat{Y}$ is the output of some parameterized model f_θ and Z is a protected condition.

3.1 MMD

For completeness, we include the way to enforce conditional indeoendence for binary Z and Y using MMD Gretton et al. (2012), which was the method included in Veitch et al. (2021). This will be used as a baseline to measure the effectiveness of lifting up the binary condition.

Objective 1 (MMD)

$$\arg\min_{h \in \mathcal{H}} \mathcal{L}(h, X, Y, Z) + \lambda C(X, Y, Z) \tag{1}$$

with

$$C = \textit{MMD}(P(f(X)|Z=0, Y=0), P(f(X)|Z=1, Y=0)) + \tag{2}$$
$$\textit{MMD}(P(f(X)|Z=0, Y=1), P(f(X)|X=1, Y=1)) \tag{3}$$

3.2 HGR

A natural way to express dependence of random variables is given by the maximum correlation coefficient (Rényi, 1959):

$$\text{HGR}(U, V) = \sup_{f,g} \rho(f(U), g(V)) \tag{4}$$

for $U \in \mathcal{U}, V \in \mathcal{V}$ random variables, f, g square integrable functions and the Pearson's correlation coefficient ρ. To extend this to conditional dependence one may consider $U \sim P_{U|Z}$ and $V \sim P_{V|Z}$. To compute any dependence measure of this style the feature functions f and g has to capture all higher moments where associations may exist between U and V.

3.3 DIRECT APPROXIMATION OF HGR

In Mary et al. (2019) an information theoretic approach is given to approximate HGR using Witsenhausen's characterization (Witsenhausen, 1975).

Theorem 1 (Witsenhausen) *Suppose U and V are discrete random variables and consider*

$$Q(u, v) = \frac{\pi(u, v)}{\sqrt{\pi(u)}\sqrt{\pi(v)}}$$

where π, π_U and π_V are the joint and marginal distributions of U and V. Then,

$$\textit{HGR}(U, V) = \sigma_2(Q) \tag{5}$$

with $\sigma_2(Q)$ denoting the second largest eigenvalue of Q.

Moreover, this extends to the continuous case assuming compactness on Q. To use the above for learning, Mary et al. (2019) show the upper bound $\text{HGR}^2 \leq \chi^2(\pi_{U,V}, \pi_U \otimes \pi_V)$ which similarly holds both in the discrete and continuous case. This is then utilized for learning as:

Objective 2 (HGR)

$$\arg\min_{h \in \mathcal{H}} \mathcal{L}(h, X, Y, Z) + \lambda ||\chi^2(\pi_{\hat{Y}|Y,Z|Y}, \pi_{Y\hat{|}Y} \otimes \pi_{Z|Y})||_1 \tag{6}$$

We will use HGR to measure conditional independence between continuous 1D random variables. While KDE can also handle higher dimensions, its performance tends to diminish, as discussed in Chapter 4 of Hart et al. (2000).

3.4 Kernel-based Conditional Covariance Measure

A line of work developed multiple characterization for dependence or conditional dependence based on measuring (conditional) cross-correlation in Reproducing Kernel Hilbert Spaces. Fukumizu et al. (2007) proposes to use the Hilbert-Schmidt norm of the normalized conditional cross-covariance operator and show a kernel-free integral expression in the limit of infinite data. They formulate the following requirement for kernels to be expressive enough:

Definition 1 (Characteristic Kernel) *Let $(\mathcal{X}; \mathcal{B})$ be a measurable space, X a random variable on $\mathcal{X}$ with distribution P , and $(\mathcal{H}; k)$ an RKHS on $\mathcal{X}$ with k integrable. The mean element of $\mathcal{X}$ on $\mathcal{H}$ is defined by the unique element $m_X \in \mathcal{H}$ such that $\forall f \in \mathcal{H} : \langle m_X; f \rangle_{\mathcal{H}} = \mathbb{E}[f(X)]$. Letting $\mathcal{P}$ be the family of all probabilities on $(\mathcal{X}; \mathcal{B})$, we define the map M_k by $M_k : \mathcal{P} \to \mathcal{H}$, $P \to m_P$. The kernel k is said to be characteristic if the map M_k is injective, or equivalently, if the condition $\forall f \in \mathcal{H} : \mathbb{E}_Q[f(X)] = \mathbb{E}_P[f(X)]$ implies $P = Q$.*

The main building block of the measure is the cross-covariance operator $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ defined as $\langle g, \Sigma_{YX} F \rangle_{\mathcal{H}_{\mathcal{Y}}} = \text{Cov}[f(X), g(Y)]$ for random variables X, Y on $\mathcal{X} \times \mathcal{Y}$ with RKHSs $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}$ with integrable kernels $k_{\mathcal{X}}, k_{\mathcal{Y}}$. Moreover, it is shown that $\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$, where V_{YX} is unique and bounded with unit norm and called the *normalized cross-covariance operator*. Assuming a third random variable Z on $\mathcal{Z}$ with RKHS $(\mathcal{H}_{\mathcal{Z}}; k_{\mathcal{Z}})$ the previous definition is expanded to the normalized conditional cross-covariance operator defined as $V_{YX|Z} = V_{YX} - V_{YZ} V_{ZX}$.

Theorem 2 *Let $\ddot{X} = (X, Z)$ and $k_{\ddot{\mathcal{X}}} = k_{\mathcal{X}} k_{\mathcal{Z}}$ and assume that the product $k_{\ddot{\mathcal{X}}} k_{\mathcal{Y}}$ is characteristic on $(\mathcal{X} \times \mathcal{Z}) \times \mathcal{Y}$ and $\mathcal{H}_{\mathcal{Z}} + \mathbb{R}$ is dense in $L^2(P_{\mathcal{Z}})$. Then*

$$V_{Y\ddot{X}|Z} = 0 \iff X \perp\!\!\!\perp Y | Z$$

Fukumizu et al. (2007) shows that the above measure is Hilbert-Schmidt and proposes to use the measure $I^{COND}(X, Y|Z) = ||V_{Y\ddot{X}|Z}||_{HS}$. It's empirical counterpart can be expressed in terms of the corresponding centered Gram matrices, and thus used for learning as follows:

Objective 3 (Kernelized Conditional Cross Covariance (KCCC))

$$\arg\min_{h \in \mathcal{H}} \mathcal{L}(h, X, Y, Z) + \lambda \hat{I}_n^{COND}(X, Y|Z) \tag{7}$$

which is computed using the centered Gram matrices $R_{\cdot}$ as

$$\hat{I}_n^{COND}(X, Y|Z) = ||\hat{V}_{Y\ddot{X}|Z}||_{HS} = \text{Tr}[R_{\ddot{Y}} R_{\ddot{X}} - 2R_{\ddot{Y}} R_{\ddot{X}} R_Z + R_{\ddot{Y}} R_Z R_{\ddot{X}} R_Z] \tag{8}$$

3.5 CIRCE

Pogodin et al. (2022) approach conditional independence from a slightly different direction, taking advantage of the following definition of conditional independence, which is a reformulation of the definition due to Daudin (1980):

Theorem 3 *X and Y are Z-conditionally independent if and only if it holds that for all $g \in L^2_X$ and $h \in L^2_{ZY}$ that $\mathbb{E}[g(X)(h(Z,Y) - \mathbb{E}_{Z'}[h(Z',Y)|Y])] = 0$*

This formulation has the advantage that the inner expectation can be precomputed as it is independent of X. Analogously to 3.4 they show that if the spaces $\mathcal{G}$ and $\mathcal{F}$ are RKHSs with bounded feature maps $\phi : \mathcal{X} \to \mathcal{G}$ and $\psi : (\mathcal{Z} \times \mathcal{Y}) \to \mathcal{F}$ then the following operator is Hilbert-Schmidt:

$$C_{XY|Z} = \mathbb{E}[\phi(X) \otimes (\psi(Z,Y) - \mathbb{E}_{Z'}[\psi(Z',Y)|Y])] \in HS(\mathcal{G},\mathcal{F}) \tag{9}$$

Moreover, they establish that zero norm of the operator similarly implies conditional independence:

Theorem 4 *For $\mathcal{G}$ and $\mathcal{F}$ with L^2-universal (Sriperumbudur et al., 2011) kernels we have*

$$||C_{XY|Z}||_{HS} = 0 \iff X \perp\!\!\!\perp Z|Y \tag{10}$$

For radial basis kernels Sriperumbudur et al. (2011) show that characteristicity implies universality, connecting the two theories. The above definition gives rise to a third regularizer for measuring conditional invariance:

Objective 4

$$\arg\min_{h \in \mathcal{H}} \mathcal{L}(h,X,Y,Z) + \lambda \hat{C}_{XY|Z} \tag{11}$$

where the empirical counterpart $\hat{C}_{XY|Z}$ is computed as:

$$\frac{1}{B(B-1)} Tr(K_{XX}(K_{YY} \odot K_{ZZ}) \tag{12}$$

As described earlier, only K_{XX} must be computed for every minibatch, $\mathbb{E}_{Z'}[h(Z',Y)|Y])]$ is estimated on heldout data with kernel ridge regression. The authors propose to use leave-on-out cross validation to obtain the kernel and ridge parameters. For the full algorithm we refer the reader to the original paper Pogodin et al. (2022).

4 EMPIRICAL COMPARISON

In the experimental section we both aim to (1) compare the different independence measure for continuous confounder variable as well as (2) evaluate the advantage of lifting up the binary or 1D restriction on Z.

4.1 EXPERIMENTAL DETAILS

We experiment with the "Communities and Crimes" dataset (Redmond & Deane, 2002) and the Amazon reviews dataset (McAuley et al., 2015). For the crime dataset, we use a MLP that we train with a learning rate of 0.001 using the Adam optimizer. For the reviews dataset, we first create embeddings of the reviews with the `BERT Base` model (Devlin et al., 2018), and then apply a single linear layer, trained with the same parameters as before.

4.2 CRIME DATASET

In the "Communities and Crime" dataset (Redmond & Deane, 2002), the primary task is to predict the crime rate for specific regions in the U.S. using a combination of socio-economic, law enforcement and crime statistics. In this problem, Z is considered as the ratio of different ethnicities in the

Regularizer	Weight	$\mathbf{R^2}$	**p** ($*10^{-2}$)	KCCC	Circe
None	0	0.57196	0	79.00524	0.00911
KCCC	0.001	0.56923	0	67.31462	0.00928
	0.003	0.48096	0	68.38218	0.00886
	0.005	0.40843	0	64.22034	0.00903
	0.007	0.40993	0	66.37890	0.00884
	0.01	0.33253	0.0007	55.59844	0.00858
Circe	10	0.53384	0	79.67969	0.00896
	30	0.52796	0	79.03073	0.00879
	50	0.45208	0.04411	80.53570	0.00842
	70	0.35230	0.56986	81.47501	0.00823
	90	0.26100	0.69175	80.86659	0.00829
HGR	0.1	0.55586	0.00001	80.03571	0.00883
	0.5	0.55858	0.00017	81.29045	0.00863
	1	0.51590	0.00098	82.62284	0.00868
	2	0.45225	0	78.60696	0.00868
	5	0.35418	0.00088	77.79527	0.00848
MMD	0.01	0.56490	0	78.18658	0.00896
	0.05	0.53423	0	82.05845	0.00892
	0.1	0.53213	0	81.64593	0.00876
	0.5	0.43871	0	85.40918	0.00931
	1	0.41252	0	81.35972	0.00871

Table 1: Results for the Communities and Crime dataset.

neighborhood, originally a 4-dimensional feature vector. One generally assumes a causal relationship between X and Y and selection. Moreover, the purely spurious assumption by the framework due to Veitch et al. (2021) is also natural, therefore we enforce conditional independence.

We want to see how using the 4D Z helps over using measures that were developed only for 1D variable, and thus we must reduce Z's dimensionality. We compare KCCC and Circe where we use the full Z to HGR where only the *racePtBlack* attribute was used. Further, for MMD we binarize this using the sign function after feature normalization.

As the relationship between Z and X is even conceptually unknown, we cannot measure directly the robustness of f w.r.t Z. Instead, we cross-validate the measures, and compute KCCC and Circe on the test set using the full Z, as well as run a conditional independence test as introduced by Zhang et al. (2011).

Results: We see that neither of the metrics show improvements in the conditional independence for MMD, and only a slight increase for HGR. Using KCCC as the regularizer, it's effect can be measured on the test set but it is not reflected in a significant increase in the p-value. On the other hand, Circe leads to larger p-values, as well as a deacrease if measured by itself, on the test set. However, this trend cannot be seen in KCCC. This disparity is interesting, and may stem from the strong dependence on the kernel parameter chosen. We discuss new measures in Section **??** that are currently under development for conditional independence and, instead of relying on kernels, uses neural networks to learn the necessary representation.

4.3 Amazon Reviews dataset

In the Amazon reviews dataset we predict the usefulness of a review based on the review text. Conditional independence in enforced w.r.t. the sentiment of the review. The task clearly resembles a causal relationship between X and Y. It can be argued that there is a selection effect, as people may tend to flag positive reviews more often as helpful then negative ones.

We measure sentiment through three proxy methods. First, we introduce a binary confounder where Z is set to positive for reviews with more than 3 stars and negative for those with fewer than 3 stars;

Regularizer	Confounder	Weight	Accuracy	VCF
None	None	0	0.775	0.29960
KCCC	continuous	0.005	0.765	0.28751
		0.007	0.785	0.27986
		0.01	0.81	0.27598
		0.03	0.735	0.27403
		0.05	0.61	0.31496
	discrete	0.005	0.81	0.27683
		0.007	0.79	0.28762
		0.01	0.8	0.27845
		0.03	0.795	0.26795
		0.05	0.670	0.30312
Circe	continuous	1	0.78	0.28036
		3	0.755	0.28294
		5	0.72	0.28759
		7	0.695	0.28836
		10	0.695	0.29141
	discrete	1	0.8	0.27930
		3	0.765	0.28661
		5	0.75	0.29456
		7	0.745	0.29493
		10	0.73	0.29761
HGR	continuous	0.01	0.78	0.29802
		0.03	0.78	0.29876
		0.05	0.77	0.30146
		0.07	0.78	0.30572
		0.1	0.77	0.30249
	discrete	0.01	0.775	0.29929
		0.03	0.78	0.29703
		0.05	0.775	0.29796
		0.07	0.77	0.29720
		0.1	0.77	0.29624
MMD	binary	0.01	0.77500	0.29694
		0.03	0.795	0.28760
		0.05	0.79	0.28073
		0.07	0.78	0.28953
		0.1	0.770	0.29851

Table 2: Results for the Amazon reviews dataset.

reviews with exactly 3 stars are excluded. Second, we derive a discrete value for Z by normalizing the star rating against the maximum possible rating. Lastly, we employ a sentiment analyzer to extract the sentiment of a sentence based on its adjectives. Detailed specifications of these methods can be found in Appendix A.

Whilst the ground-truth relationship between X and Z is unknown, one may assume that the sentiment of a sentence is determined by the adjectives it contains. Thus, to create perturbed data, we swap the adjectives by it's antonyms. In specific, we gather all possible antonyms for every adjective in the sentence, select a random subset of the adjectives that we will alter, and create the new sentence by changing the chosen adjectives by a random antonym of theirs. For every sentence we create at most 20 such samples. To measure the robustness of the predictor, we measure the variance of the predictor over the perturbed dataset:

$$\text{VCF} = \mathbb{E}_{x\sim X}[\mathbb{V}_{z\sim Z}[f(x(z))]] \tag{13}$$

Results: We observe only small differences between the VCF scores independent of the measure or the proxy for sentiment used. The lowest VCF score is achieved by KCCC for the discrete confounder, but similar results are achieved for the continuous case too. We note, that this may not only depend on kernel parameters or the way the proxy is computed, but also on how we create counterfactual samples.

4.4 Discussion and Outlook

In this exploratory work we discussed how causal models can be used to argue for necessary conditions for a robust predictor. A strong assumption of the framework due to Veitch et al. (2021) is the purely spuriousness in the causal case with selection. Eastwood et al. (2023) discuss how to include spurious features that may still contain information about Y. Further research could establish connections between the two works and experiment with how the methods proposed by the latter may be included in learning a robust predictor.

We also examined several measures for conditional independence based on kernel methods. A significant limitation of these methods is the challenge in selecting appropriate kernel parameters, for which there is not yet a theoretically justified approach. Concurrent work due to Heiner et al. proposes to use neural networks and learn sufficient representations (for current reference, see Kremer et al. (2022)). These representations then serve as a basis for measuring conditional independence. The optimization task evolves into a minimax problem concerning both the predictor parameters and the representation parameters. Although this method bypasses the challange of kernel parameter selection, optimizing such objectives can pose challenges, as seen in GAN training Goodfellow et al. (2014a). A future direction would be to compare this novel metric with previously established ones.

We experimented with both tabular and text data, to see how lifting up the binary condition on the confounder helps to achieve a more robust predictor. Our results showed minor improvements in both experiments, but already indicated the necessity of measures that enable measuring independence w.r.t continuous, high-dimensional variables. Still the dependency on kernel parameters is a great limitations.

Lastly, it would be intriguing to see how generative models for text (such as) change if they finetuned with the proposed regularization. One analyze changes in the attention mechanism.

Acknowledgment

We thank Heiner Kremer from Max Planck Institute for the insightful discussions about the conditional independence measures.

References

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

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

J. J. Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, January 1980. ISSN 0006-3444. doi: 10.1093/biomet/67.3.581.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Cian Eastwood, Shashank Singh, Andrei Liviu Nicolicioiu, Marin Vlastelica, Julius von Kügelgen, and Bernhard Schölkopf. Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features, 2023.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS'07: Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 489–496. Curran Associates Inc., Red Hook, NY, USA, December 2007. ISBN 978-1-60560352-0. doi: 10.5555/2981562.2981624.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014a.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv (Cornell University)*, 12 2014b. URL https://arxiv.org/pdf/1412.6572.pdf.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Peter E Hart, David G Stork, and Richard O Duda. *Pattern classification*. Wiley Hoboken, 2000.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations (ICLR)*, 2017.

Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, and Bernhard Schölkopf. Functional generalized empirical likelihood estimation for conditional moment restrictions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11665–11682. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kremer22a.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv (Cornell University)*, 2 2018. URL http://arxiv.org/pdf/1706.06083.pdf.

Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-Aware Learning for Continuous Attributes and Treatments. In *International Conference on Machine Learning*, pp. 4382–4391. PMLR, May 2019. URL https://proceedings.mlr.press/v97/mary19a.html.

Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products, 2015.

S Chandra Mouli and Bruno Ribeiro. Asymmetry learning for counterfactually-invariant classification in OOD tasks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=avgclFZ221l.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J. Sutherland, Victor Veitch, and Arthur Gretton. Efficient Conditionally Invariant Representation Learning. *arXiv*, December 2022. doi: 10.48550/arXiv.2212.08645.

Francesco Quinzan, Cecilia Casolo, Krikamol Muandet, Yucen Luo, and Niki Kilbertus. Learning counterfactually invariant predictors, 2023.

Michael A. Redmond and Carla A. Deane. Communities and crime. https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime, 2002. UCI Machine Learning Repository.

A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10 (3):441–451, September 1959. ISSN 1588-2632. doi: 10.1007/BF02024507.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstractions in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *J. Mach. Learn. Res.*, 12(null):2389–2410, jul 2011. ISSN 1532-4435.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: why and how to pass stress tests. *arXiv (Cornell University)*, 5 2021. URL https://arxiv.org/pdf/2106.00545.pdf.

H. S. Witsenhausen. On Sequences of Pairs of Dependent Random Variables. *SIAM J. Appl. Math.*, 28(1):100–113, January 1975. ISSN 0036-1399. doi: 10.1137/0128010.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. UAI'11, pp. 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

A Appendix

A.1 Hyperparameter choices

We used the Gaussian kernel ($K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. For choosing kernel parameters, we swept over $\sigma = [1, 0.5, 0.1, 0.05, 0.01]$. The chosen kernel paramater was used also to measure the independence on the test set.

A.2 Amazon Reviews Dataset

We use the *Clothing Shoes and Jewelry* part of the amazon reviews dataset McAuley et al. (2015). We create 10000 examples, where drop any for which no counterfactual example could be generated; the review text, or the star-rating was massing; or received a 3-star rating.

A.2.1 Creating counterfactual sentences

We first generated a list of adjectives and their antonyms using ChatGPT V4. For every sentence, we then marked every adjectives using the NLTK's part of speech tagger (Bird et al., 2009). We then choose a random subset of the adjectives and swap them with a randomly chosen corresponding antonym. For every sentence, we create at most 20 counterfactual examples.