Kernel Methods for Invariant Representation Learning: Enforcing Fairness and Conditional Independence

by

Namrata Deka

B.Tech., Computer Science & Engineering, Indraprastha Institute of Information Technology, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

April 2023

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Kernel Methods for Invariant Representation Learning: Enforcing Fairness and Conditional Independence

submitted by **Namrata Deka** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Danica J. Sutherland, Assistant Professor, Computer Science, UBC
Supervisor

Mi Jung Park, Assistant Professor, Computer Science, UBC
Supervisory Committee Member

Abstract

This work introduces two novel kernel-based measures to enforce certain invariance properties in the learned representation space of a deep neural network. The first method, MMD-B-Fair, learns fair representations of data via kernel two-sample testing. It finds neural features of data where a maximum mean discrepancy (MMD) test cannot distinguish between different representations of different sensitive groups, while preserving information about the target variable to be predicted. To minimize the power of an MMD test this method exploits the simple asymptotics of a block testing scheme to address challenges presented by the complex dependency of the test threshold on the estimated MMD. Compared to existing methods on fair representation learning, MMD-B-Fair does not require generative modeling or discriminative architectural tuning, and is able to achieve competitive results on fairness benchmarks and downstream transfer. The second method, CIRCE, introduces a measure of conditional independence for multivariate continuous-valued variables that can be efficiently used as a regularizer to learn deep neural features that are conditionally independent of a known distractor Z given a target label Y. CIRCE requires just a single ridge regression from Y to kernelized features of Z, which can be done in advance. It is then only necessary to enforce independence of the learned neural features from the residuals of this regression. By contrast, earlier measures of conditional dependence require multiple regressions for each step of feature learning, resulting in severe bias and variance, and greater computational cost. CIRCE has superior performance to previous methods on challenging benchmarks, including learning conditionally invariant image features. Python implementations of both methods are made publicly available at github.com/namratadeka/mmd-b-fair and github.com/namratadeka/circe.

Lay Summary

A fundamental challenge in machine learning is the lack of robustness due to the tendency of popular learning algorithms to exploit statistical signatures that capture spurious correlations or "shortcuts" in the training data. These signatures are often absent in testing domains where the algorithms are deployed leading to failures that can be critical in sensitive applications such as autonomous driving, medical diagnosis, criminal profiling, credit assignments, etc. In this work, we present two novel methods, MMD-B-Fair and CIRCE, for learning predictive models that do not depend on spuriously correlated features thereby resulting in better performance with respect to fairness and co-variate shifts outside the training domain. We validate the accuracy and effectiveness of both methods over multiple benchmark datasets against existing algorithms and show competitive and superior performance.

Preface

This thesis presents original research conducted by Namrata Deka under the direct supervision of Dr. Danica J. Sutherland. Unless otherwise mentioned, all implementations were done by Namrata Deka. The first method, MMD-B-Fair, presented in Section 3.1 and evaluated in Section 4.1 was developed by Namrata Deka under the sole guidance of Dr. Danica J. Sutherland. The proof of the non-existence of an unbiased estimator of the normalized MMD test power objective presented in Section A.1 was completed by Dr. Danica J. Sutherland. The second method, namely CIRCE, as proposed in Section 3.2 and evaluated in Section 4.2 was developed in collaboration between Roman Pogodin, Namrata Deka and Yazhe Li under the joint supervision of Dr. Arthur Gretton, Dr. Victor Veitch and Dr. Danica J. Sutherland. The proofs for definitions of CIRCE and its kernel estimator presented in Section A.4 and Section A.5 were primarily completed by Roman Pogodin and Dr. Arthur Gretton. The experiments on synthetic datasets detailed in Section 4.2.1 were conducted by Yazhe Li. The thesis is based on the following published works:

[1] Deka, N. and Sutherland, D.J.. MMD-B-Fair: Learning Fair Representations with Statistical Testing. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 206:9564-9576, 2023.

[2] Pogodin, R*, Deka, N*, Li, Yazhe*, Sutherland, D.J., Veitch, V and Gretton, A.. Efficient Conditionally Invariant Representation Learning. *Proceedings of The 11th International Conference on Learning Representations (ICLR)*, 2023. (*equal contribution)

The first paper proposes MMD-B-Fair, a statistical two-sample testing paradigm to learn fair neural features for discrete sensitive and target variables. The second paper presents CIRCE, a kernel-based measure of conditional independence for continuous-valued target and distractor variables that can be efficiently estimated with small mini-batches to enforce conditional invariance in deep neural networks. This work was accepted as a notable top 5% at the 11th International Conference on Learning Representations, 2023.

Table of Contents

List of Tables

List of Figures

Acknowledgments

None of the work presented here would have been possible without the incredible support of my advisor Dr. Danica J. Sutherland. I am fortunate to have started my graduate school journey with such a kind, patient and understanding mentor. The guidance and encouragement I received has had an invaluable contribution to this thesis and my journey towards being a researcher.

I would also like to thank Dr. Arthur Gretton from the Gatsby Computational Neuroscience Unit at UCL, for guiding me through the project on conditional independence regularization. To work with and learn from such a revered statistician was an extremely wonderful experience.

I would like to thank my parents, Amiya and Mouchumi Deka, for the innumerable sacrifices they have made to give me the best of opportunities. Everything I might achieve is a testament to their hard work and unconditional love. Also, I would like to thank my buddy, Jimmy, for accompanying me through graduate school and providing me with an endless supply of serotonin and dopamine by forcing me to play with him and go on walks everyday without fail.

Lastly, I would like to thank the National Sciences and Engineering Resource Council of Canada, the Canada CIFAR AI Chairs program, WestGrid, SHARCNET, Calcul Québec, and the Digital Resource Alliance of Canada for funding my research and providing the required computational infrastructure for all my experiments.

Chapter 1

Introduction

1.1 Fair Representation Learning

Machine learning systems are increasingly being used for making critical and sensitive real-life decisions in domains like finance, criminal reform, hiring, health, etc. [Flores et al., 2016, Skeem and Lowenkamp, 2016, Bogen and Rieke, 2018, Chouldechova et al., 2018, Lebovits, 2018, Ledford, 2019, Wilson et al., 2019] The importance of designing non-discriminatory learning algorithms that can mitigate various biases regarding private and protected features like gender or race is crucial to building trustworthy AI systems. Often data collected from the real world are plagued with issues like under-representation of minority groups, correlated sensitive and target features, or drastic distributional shifts between training and testing phases [Gianfrancesco et al., 2018, Jo and Gebru, 2020]. All of these can lead to biased models that can make undesirable mistakes in the real world, and therefore we need to address this issue and develop systems that are robust to biases in data distributions.

Fair representation learning is one approach towards this goal, which tries to find data representations that satisfy certain fairness objectives [Zemel et al., 2013, Edwards and Storkey, 2016, Louizos et al., 2016, Zhang et al., 2018, Madras et al., 2018, Lahoti et al., 2020]. Most deep learning-based fair representation learning methods take one of two broad approaches: try to disentangle latent factors with a generative variational model then ultimately discard the sensitive factor from the representation, or mitigate bias via adversarial techniques where discriminator(s) attempt to predict the sensitive group from a learnt encoded representation. In this work, we explore a different route, using deep kernels and statistical two-sample testing.

Statistical two-sample tests are used to determine whether two sets of data samples come from the same underlying distribution. Our method is centered around the idea that if a machine learning system is fair with respect to certain protected attributes, then that system's representation of one sensitive group should not be statistically distinguishable from the other. Our method

learns fair representations by optimizing a neural network to minimize the test power – the ability of a two-sample test to correctly distinguish two sets of samples – for samples differing by the sensitive class label, while still finding a useful representation by maximizing the test power and/or classification accuracy for distinguishing "target" labels.

This framework avoids learning a generative model of the data or explicit adversarial training, by instead relying on tests based on the maximum mean discrepancy (MMD) [Gretton et al., 2012] to compare different samples of representations. We use the MMD in a novel way, combining existing work on power optimization [Sutherland et al., 2017, Liu et al., 2020] with block testing [Zaremba et al., 2013] to give an effective criterion for driving down the test power of sensitive tests – a problem not handled well by previous work which focuses only on maximizing power. Our method is supported by theoretical results as well as good empirical performance.

We first give a self-contained introduction to MMD-based testing in Section 3.1.1 and Section 3.1.2, establishing all the tools we will need for our method for learning fair kernels in Section 3.1.3 and fair representations in Section 3.1.4.

1.2 Conditional Independence Regularization

For our second method, CIRCE, we consider a setting where we wish to learn neural features of input X that are conditionally independent of correlated features Z given a label Y. In particular, our aim is to learn a representation function φ for the features such that $\varphi(X) \perp\!\!\!\perp Z \mid Y$. There are at least three motivating settings where this task arises.

1. Fairness. In this context, Z is some sensitive attribute (e.g., race or gender) and the condition $\varphi(X) \perp\!\!\!\perp Z \mid Y$ is the equalized odds condition [Mehrabi et al., 2021a].

2. Domain invariant learning. In this case, Z is a label for the environment in which the data was collected (e.g., if we collect data from multiple hospitals, Z_i labels the hospital that the ith datapoint is from). The condition $\varphi(X) \perp\!\!\!\perp Z \mid Y$ is sometimes used as a target for invariant learning [e.g., Long et al., 2018, Tachet des Combes et al., 2020, Goel et al., 2021, Jiang and Veitch, 2022]. Wang and Veitch [2022] argue that this condition is well-motivated in cases where Y causes X.

3. Causal representation learning. Neural networks may learn undesirable "shortcuts" for their tasks – e.g., classifying images based on the texture of the background. To mitigate this issue, various schemes have been proposed to force the network to use causally relevant factors in its decision [e.g., Veitch et al., 2021b, Makar et al., 2022, Puli et al., 2022]. The structural causal assumptions used in such approaches imply conditional independence relationships between the features we would like the network to use, and observed metadata

that we may wish to be invariant to. These approaches then try to learn causally structured representations by enforcing this conditional independence in a learned representation.

CIRCE specifically targets the case when X is some high-dimensional structured data – e.g., text, images, or video – and we would like to model the relationship between X and (the relatively low-dimensional) Y, Z using a neural network representation $\varphi(X)$. There are a number of existing techniques for learning conditionally invariant representations using neural networks (e.g., in all the motivating applications mentioned above). Usually, however, they rely on the labels Y being categorical with a small number of categories. We develop a method for conditionally invariant representation learning that is effective even when the labels Y and attributes Z are continuous or moderately high-dimensional.

To understand the challenge, it is helpful to contrast with the task of learning a representation φ satisfying the marginal independence $\varphi(X) \perp\!\!\!\perp Z$. To accomplish this, we might define a neural network to predict Y in the usual manner, interpret the penultimate layer as the representation φ, and then add a regularization term that penalizes some measure of dependence between $\varphi(X)$ and Z. As φ changes at each step, we would typically compute an estimate based on the samples in each mini-batch [e.g., Beutel et al., 2019, Veitch et al., 2021b]. The challenge for extending this procedure to conditional invariance is simply that it's considerably harder to measure. More precisely, as conditioning on Y "splits" the available data (if Y is categorical, naively we would measure a marginal independence for each level of Y), we require large samples to assess conditional independence. When regularizing neural network training, however, we only have the samples available in each mini-batch: often not enough for a reliable estimate of existing dependence measures.

Our technique reduces the problem of learning a conditionally independent representation to the problem of learning a marginally independent representation, following a characterization of conditional independence due to Daudin [1980]. We first construct a particular statistic $\zeta(Y,Z)$ such that enforcing the marginal independence $\varphi(X) \perp\!\!\!\perp \zeta(Y,Z)$ is (approximately) equivalent to enforcing $\varphi(X) \perp\!\!\!\perp Z \mid Y$. The construction is straightforward: given a fixed feature map $\psi(Y,Z)$ on $\mathscr{Y} \times \mathscr{Z}$ (which may be a kernel or random Fourier feature map), we define $\zeta(Y,Z)$ as the conditionally centered features, $\zeta(Y,Z) = \psi(Y,Z) - \mathbb{E}[\psi(Y,Z) \mid Y]$. We obtain a measure of conditional independence, the *Conditional Independence Regression CovariancE* (CIRCE), as the Hilbert-Schmidt Norm of the kernel covariance between $\varphi(X)$ and $\zeta(Y,Z)$.

A key point is that the conditional feature mean $\mathbb{E}[\psi(Y,Z) \mid Y]$ can be estimated offline, in advance of any neural network training, using standard methods [Song et al., 2009, Grunewalder et al., 2012, Park and Muandet, 2020, Li et al., 2022]. This makes CIRCE a suitable regularizer for any setting where the conditional independence relation $\varphi(X) \perp\!\!\!\perp Z \mid Y$ should be enforced when learning $\varphi(X)$. In particular, the learned relationship between Z and Y doesn't depend on

the mini-batch size, sidestepping the tension between small mini-batches and the need for large samples to estimate conditional dependence.

We introduce the relevant characterization of conditional independence from [Daudin, 1980] in Section 3.2.1 followed by our CIRCE criterion in Section 3.2.2.

Chapter 2

Related Work

2.1 Fairness in Machine Learning

Fair representation learning has of late (deservedly) found a lot of traction within the deep learning community [Mehrabi et al., 2021b]. The growing popularity and success of adversarial learning has resulted in a substantial number of adversarial techniques to mitigate bias and enforce group fairness by training discriminators to distinguish one sensitive group (or sub-group) from another [Edwards and Storkey, 2016, Xie et al., 2017, Zhang et al., 2018, Madras et al., 2018, Zhao et al., 2020]. However, representations learnt via adversarial approaches do not completely "hide" sensitive information as the learnt representations are dependent on the specific function classes (or architectural complexity) used for the discriminators. Variational methods, on the other hand, focus on learning disentangled latent spaces where sensitive factors can be separated from non-sensitive features [Louizos et al., 2016, Creager et al., 2019, Norouzi, 2020]. Other methods (including our proposed approach) try to enforce fairness by adding additional constraints in the learning objective to regularize the learned weights of the neural networks involved [Kamishima et al., 2012, Hajian et al., 2016, Zafar et al., 2017, Speicher et al., 2018].

There have also been, in particular, several MMD-based approaches to fair/invariant representation learning. Louizos et al. [2016] used the MMD as a regularizer to train fair variational autoencoders to impose statistical parity between embeddings across different sensitive groups. Recently, Oneto et al. [2020] used the MMD with a similar intuition to ours to learn representations that transfer better to unseen tasks in a multitask setting. Veitch et al. [2021a] use the MMD as regularizers to a classifier, choosing between the marginal and conditional form based on the causal direction of the task, to enforce counterfactual invariance. Most recently Lee et al. [2022] proposed using the MMD to perform fair principal component analysis by penalizing the measure between dimensionality-reduced distributions over different protected groups. Our approach, although similar in spirit, uses the power of MMD two-sample tests rather than the raw MMD

estimate, which avoids several pitfalls and is particularly important when simultaneous maximization and minimization are required – something not previously explored in the kernel-methods community.

In Section 4.1, we compare to several different baselines. LAFTR [Madras et al., 2018] employs an adversarial network to predict the sensitive class using the representations being simultaneously learnt by a target predictor. CFAIR [Zhao et al., 2020] conditionally aligns the representations for accuracy-fairness trade-off by using two adversaries (one for the positive class label, one for the negative). FCRL [Gupta et al., 2021] controls the mutual information between the representations and the sensitive labels with contrastive information estimators. sIPM [Kim et al., 2022] employs the sigmoid Integral Probability Metric (IPM) as the deviance measure over the learnt representations. This is perhaps the most closely related method to our approach, using an IPM measure to regularize the prediction function.

2.2 Conditional Independence Measures

We review prior work on kernel-based measures of conditional independence to determine or enforce $X \perp\!\!\!\perp Z|Y$, including those measures we compare against in our experiments in Section 4.2.

We begin with procedures based on kernel conditional feature covariances. The conditional kernel cross-covariance was first introduced as a measure of conditional dependence by Sun et al. [2007]. Following this work, a kernel-based conditional independence test (KCI) was proposed by Zhang et al. [2011]. The latter test relies on satisfying 1 leading to a statistic [1] that requires regression of $\varphi(X)$ on Y in every minibatch (as well as of Z on Y, as in our setting). More recently, Quinzan et al. [2022] introduced a variant of the Hilbert-Schmidt Conditional Independence Criterion [HSCIC; Park and Muandet, 2020] as a regularizer to learn a generalized notion of counterfactually-invariant representations [Veitch et al., 2021b]. Estimating $\mathrm{HSCIC}(X,Z|Y)$ from finite samples requires estimating the conditional mean-embeddings $\mu_{X,Z|Y}$, $\mu_{X|Y}$ and $\mu_{Z|Y}$ via regressions [Grunewalder et al., 2012]. HSCIC requires three times as many regressions as CIRCE, of which two must be done online in minibatches to account for the conditional cross-covariance terms involving X. We will compare against HSCIC in experiements, being representative of this class of methods, and having been employed successfully in a setting similar to ours.

Alternative measures of conditional independence make use of additional normalization over the measures described above. The Hilbert-Schmidt norm of the *normalized* cross-covariance was introduced as a test statistic for conditional independence by Fukumizu et al. [2008], and was used for structure identification in directed graphical models. Huang et al. [2022] proposed using the ratio of the *maximum mean discrepancy* (MMD) between $P_{X|ZY}$ and $P_{X|Y}$, and the MMD between

[1] The conditional-independence test statistic used by KCI is $\frac{1}{B}\mathrm{Tr}\left(\tilde{K}_{\ddot{X}|Y}\tilde{K}_{Z|Y}\right)$, where $\ddot{X}=(X,Y)$ and $\tilde{K}$ is a centered kernel matrix. Unlike CIRCE, $\tilde{K}_{\ddot{X}|Y}$ requires regressing $\ddot{X}$ on Y using kernel ridge regression.

the Dirac measure at X and $P_{X|Y}$, as a measure of the conditional dependence between X and Z given Y. The additional normalization terms in these statistics can result in favourable asymptotic properties when used in statistical testing. This comes at the cost of increased computational complexity, and reduced numerical stability when used as regularizers on minibatches.

Another approach, due to Shah and Peters [2020], is the Generalized Covariance Measure (GCM). This is a normalized version of the covariance between residuals from kernel-ridge regressions of X on Y and Z on Y (in the multivariate case, a maximum over covariances between univariate regressions is taken). As with the approaches discussed above, the GCM also involves multiple regressions – one of which (regressing X on Y) cannot be done offline. Since the regressions are univariate, and since GCM simply regresses Z and X on Y (instead of $\psi(Z,Y)$ and $\phi(X)$ on Y), we anticipate that GCM might provide better regularization than HSCIC on minibatches. This comes at a cost, however, since by using regression residuals rather than conditionally centered features, there will be instances of conditional dependence that will not be detectable. We will investigate this further in our experiments.

Chapter 3

Methodology

3.1 Learning Fair Representations with Statistical Testing

3.1.1 Maximum Mean Discrepancy (MMD)

Based on *i.i.d.* samples $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$ from distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively, the two-sample testing problem asks whether $S_{\mathbb{P}}, S_{\mathbb{Q}}$ come from the same distribution: does $\mathbb{P} = \mathbb{Q}$? We use the null hypothesis testing framework, i.e. ask whether we can confidently say that the observed $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$ would be unlikely to be so different if $\mathbb{P} = \mathbb{Q}$.

Traditional methods for two-sample tests, including t-tests and Kolmogorov-Smirnov tests, do not scale to complex high-dimensional distributions. Another modern approach is based on classification accuracy and we will describe our approach's relationship to that scheme shortly.

The MMD [Gretton et al., 2012] is a measure of distance between distributions. For distributions $\mathbb{P}$ and $\mathbb{Q}$ over a domain $\mathscr{X}$ (the set of conceivable data points), the MMD is defined in terms of a kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ giving the "similarity" of individual data points. This kernel should be positive semi-definite, the simplest case being the linear kernel $k(x,y) = x^\top y$, and the paradigmatic example being a Gaussian kernel $k(x,y) = \exp(-\frac{1}{2\sigma^2}\|x-y\|^2)$.

If $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$, then

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}; k) = \sqrt{\mathbb{E}[k(X,X') + k(Y,Y') - 2k(X,Y)]}.$$

With a *characteristic* kernel k, such as the Gaussian, we have that $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}; k) = 0$ *if and only if* $\mathbb{P} = \mathbb{Q}$. Thus, we can run a two-sample test by estimating the MMD, and rejecting the null hypothesis that $\mathbb{P} = \mathbb{Q}$ if the estimated MMD is too large to have occurred by chance.

U-STATISTIC ESTIMATOR: Our default estimator will be the U-statistic estimator, which is unbiased for MMD^2, and has almost minimal variance among unbiased estimators:[1]

$$\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) = \frac{1}{m(m-1)} \sum_{i \neq j} H_{ij} \tag{3.1}$$
$$H_{ij} = k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(Y_i, X_j),$$

where $S_{\mathbb{P}} = \{X_1, \ldots, X_m\}, S_{\mathbb{Q}} = \{Y_1, \ldots, Y_m\}$ are *i.i.d.* samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively.

The most common scheme for testing based on (3.1) is to choose some kernel k a-priori, and then reject the null hypothesis $\mathfrak{H}_0$ that $\mathbb{P} = \mathbb{Q}$ if the scaled estimator $m\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)$ is larger than a threshold c_α. The rejection threshold, c_α, should satisfy $\Pr_{\mathfrak{H}_0}\left(m\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) > c_\alpha\right) \leq \alpha$, i.e. there is α probability of incorrectly rejecting $\mathfrak{H}_0$ when it is true. The estimate is scaled by m because, as m grows, $m\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)$ converges in distribution to an infinite mixture of χ^2 variables, with weights depending on $\mathbb{P} = \mathbb{Q}$ and k, but independent of m. The rejection threshold c_α is the $(1-\alpha)$th quantile of the distribution over $m\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)$ under $\mathfrak{H}_0$ which can be approximated with a scheme known as permutation testing, generally the preferred method in this case: randomly divide $S_{\mathbb{P}} \cup S_{\mathbb{Q}}$ into two groups, compute $m\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ between them and repeat, taking the empirical quantile of those samples [Sutherland et al., 2017].

BLOCK ESTIMATOR: An alternative approach, called B-testing by Zaremba et al. [2013], randomly splits the available samples into b blocks each containing B samples. This is more computationally efficient in its estimator and also allows avoiding permutation testing, as we will see shortly. We compute $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ on each block separately and since each of those terms will be an independent unbiased estimator of the squared MMD, we can average them to obtaining the block-based estimator $\widehat{\mathrm{MMD}}_{\mathrm{B}}^2$.

Under $\mathfrak{H}_0$, the estimate in each block converges in distribution to the kernel-dependent infinite mixture of χ^2 variables as $B \to \infty$. However, whether under $\mathfrak{H}_0$ or $\mathfrak{H}_1$, the average of b of these independent estimates will converge to a normal distribution by the central limit theorem:

$$\sqrt{b}(\widehat{\mathrm{MMD}}_{\mathrm{B}}^2 - \mathrm{MMD}^2) \xrightarrow{d} \mathcal{N}(0, V_B), \tag{3.2}$$

with V_B being the variance of $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ on samples of size B (depending on $\mathbb{P}$, $\mathbb{Q}$, and k). A block test, then, can take as its test statistic $\sqrt{b}\widehat{\mathrm{MMD}}_{\mathrm{B}}^2$ and use a threshold of $\sqrt{V_B}\Phi^{-1}(1-\alpha)$, with Φ the CDF of a standard normal.

To use this method, it remains to estimate $\sqrt{V_B}$. Zaremba et al. [2013] simply took the sample

[1] The MVUE would simply also include the $k(X_i, Y_i)$ terms; the difference in practice is usually trivial, but this form is slightly simpler and allows exact expressions for the variance.

standard deviation of the b batches, which is justified since the sample variance converges almost surely to V_B. We will employ a different scheme in our use of the block estimator (to come). Although block tests are more computationally efficient than U-statistic tests, it turns out they are also proportionally less powerful [Ramdas et al., 2015] and therefore, our primary tests will be based on U-statistics.

3.1.2 Learning Deep Kernels

MMD tests work well when the choice of kernel k is appropriate; for complicated distributions, however, simple default choices may take unreasonable numbers of samples to obtain a significant power. For a powerful test in complex situations with realistic numbers of samples, we follow Liu et al. [2020] in seeking the best kernel from a parameterized family of *deep kernels*. Specifically, we take k_ω as a Gaussian kernel κ on the output of a featurizer network ϕ_ω, $k_\omega = \kappa_\omega(\phi_\omega(x), \phi_\omega(y))$. Here, ϕ_ω is a deep neural network that extracts features from input points x and y, whose parameters are contained within ω, and κ_ω is a Gaussian kernel on those features whose length-scale is also contained in ω. These kernels have seen success across a variety of areas [e.g. Wilson et al., 2016, Li et al., 2017, Jean et al., 2018, Li et al., 2021].

To be able to reliably distinguish two distributions, we wish to find the deep kernel with the most powerful test: the one with the highest probability of correctly rejecting the null hypothesis when the alternative is true. For a U-statistic test, this probability is asymptotically

$$\Pr_{\mathfrak{H}_1}\left(m\widehat{\mathrm{MMD}}_{\mathrm{U}}^2 > c_\alpha\right) \to \Phi\left(\frac{\mathrm{MMD}^2 - c_\alpha/m}{\sqrt{V_m}}\right), \tag{3.3}$$

where Φ is the CDF of a standard normal distribution, and V_m is the variance of the $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ estimator for samples of size m from $\mathbb{P}$ and $\mathbb{Q}$ with the kernel k [Sutherland et al., 2017, Equation 2]. The terms on the right-hand side are fixed, unknown quantities depending on $\mathbb{P}$, $\mathbb{Q}$, and k; MMD^2 and c_α do not depend on m. This formula comes from an asymptotic normality result for the estimator when $\mathrm{MMD}(\mathbb{P},\mathbb{Q};k) > 0$ [Serfling, 1980, Section 5.5].

Sutherland et al. [2017], Liu et al. [2020] conducted tests by dividing each of $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$ into "training" and "test" sets, finding a kernel approximately maximizing (3.3) on the training sets, and then using that kernel to run a standard two-sample test on the independent test sets. To roughly maximize (3.3), they maximized an estimator of $\mathrm{MMD}^2/\sqrt{V_m}$, the leading term when m grows and the test is reasonably likely to reject ($m\mathrm{MMD}^2 > c_\alpha$).

Although this was not done in prior work, it will be important for our purposes to emphasize that (3.3) is the asymptotic expression for the power of a test using m samples, and so a given k, $\mathbb{P}$, and $\mathbb{Q}$ correspond to a whole curve of asymptotic powers depending on m. Inside (3.3), both MMD^2 and c_α are independent of m, while, as we will see, V_m's dependence on m is exactly

known thanks to the well-understood theory of U-statistics. Thus, we can estimate the power of an m-sample test using a *different* number of samples n. For instance, we could get a rough estimate of the power of a large-sample test ($m = 2,000$) using a small mini-batch of size $n = 32$.

To roughly maximize (3.3), Liu et al. [2020] maximized the estimator $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2 / \sqrt{\widehat{V}_{m,\lambda}}$, where $\widehat{V}_{m,\lambda}$ estimates V_m by

$$\frac{4}{mn^3} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} H_{ij} \right)^2 - \frac{4}{mn^4} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} \right)^2 + \frac{\lambda}{m}, \tag{3.4}$$

using H_{ij} from (3.1). For Liu et al.'s purposes, m is a simple scalar multiplier on the objective and so need not be specified, but it will be important for us to keep track of it, as we will see. They further proved uniform convergence of the estimator $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2 / \sqrt{\widehat{V}_{m,\lambda}}$ to $\mathrm{MMD}^2 / \sqrt{V_m}$. Sutherland et al. [2017] used a more complex unbiased estimator for V_m [see Sutherland and Deka, 2019]; an unbiased estimator for V_m will not be unbiased for $\mathrm{MMD}^2 / \sqrt{V_m}$, however, and in fact we prove in Section A.1 that *no* unbiased estimator of that quantity exists. The biased estimator also worked better in our experiments.

Sutherland et al. [2017] further mentioned, but did not try, using the threshold from permutation testing to estimate the full quantity (3.3); this is expected to be important for small m or for tests with poor power (ignoring the c_α term means the overall asymptotic power cannot be less than 0.5). This estimator, as an empirical quantile, is almost surely differentiable and straightforward to implement in deep learning libraries. We explore this further in Section 3.1.3.

As argued by Liu et al. [2020, Section 4], learning a deep kernel for an MMD test is strictly more general than classifier two-sample tests [Kim et al., 2020, Lopez-Paz and Oquab, 2017], which train a classifier between $\mathbb{P}$ and $\mathbb{Q}$ on the training split, then check whether it has nontrivial accuracy on the test split. The added generality tends to yield better tests in practice.

3.1.3 Learning a Fair Kernel

Our goal is to find a representation invariant with respect to a binary sensitive attribute s, meaning that it cannot distinguish $\mathbb{P}^s$ and $\mathbb{Q}^s$: the distribution of data points with $s = 0$ and those with $s = 1$. To achieve this, we would like to find a kernel which, when used in a two-sample test to distinguish $\mathbb{P}^s$ and $\mathbb{Q}^s$, achieves negligible power.

If this were our only goal, however, there is a trivial solution: use, say, $k(x, y) = 1$. Instead, we would like a kernel that is also useful to distinguish *target* pairs of distributions, say ones useful for a downstream task: one that has high test power between $\mathbb{P}^t$ and $\mathbb{Q}^t$. (In practice, we also include a classification loss in our objective, but we clarify this straightforward addition later.)

One simple extension to the objective function of Liu et al. [2020] towards this goal would be to minimize an estimate of $\left((\mathrm{MMD}^t)^2 / \sqrt{V_m^t} - (\mathrm{MMD}^s)^2 / \sqrt{V_m^s} \right)$, where $(\mathrm{MMD}^a)^2$ and V_m^a

are computed for the learned kernel between $\mathbb{P}^a$ and $\mathbb{Q}^a$. However, this tends to be unable to appropriately "balance" the two objectives. If the power for the target test is near 1, but the sensitive-attribute test still has high power, this objective would still be just as satisfied by driving up $(\mathrm{MMD}^t)^2/\sqrt{V_m^t}$ – increasing the asymptotic power of the target test, but only just barely – as it would be by reducing $(\mathrm{MMD}^s)^2/\sqrt{V_m^s}$).

To put the two attributes on the same scale, then, we should consider the full asymptotic power (3.3), and subtract estimators of the two, resulting in the objective:

$$\Phi\left(\frac{(\mathrm{MMD}^t)^2 - c_\alpha^t/m}{\sqrt{V_m^t}}\right) - \Phi\left(\frac{(\mathrm{MMD}^s)^2 - c_\alpha^s/m}{\sqrt{V_m^s}}\right). \tag{3.5}$$

The thresholds c_α^s and c_α^t, can be estimated using permutation tests as suggested by Sutherland et al. [2017]. This makes the optimization substantially more computationally expensive; though it can be computed based on the same kernel matrix as $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ and $\widehat{V}_m$, it requires perhaps a hundred times as many matrix-vector multiplications as does $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$. We also found that the strong dependence between $\hat{c}_\alpha$ and $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ computed on the same samples meant that optimization was rarely able to drive the asymptotic power for the sensitive attribute test below about 0.5. Data splitting helped, but halves the effective batch size, and computational and sample complexity both suffer.

To avoid this problem, we instead optimize the power of a block test with b blocks of size B. From the central limit result (3.2), we have that the power of a block test is, letting $t_\alpha = \Phi^{-1}(1-\alpha)$ where Φ is the standard normal CDF,

$$\begin{aligned}
\rho_{b,B} &= \Pr_{\mathfrak{H}_1}\left(\sqrt{b}\,\widehat{\mathrm{MMD}}_{\mathrm{B}}^2 > \sqrt{V_B}t_\alpha\right) \\
&= \Pr_{\mathfrak{H}_1}\left(\frac{\sqrt{b}\,(\widehat{\mathrm{MMD}}_{\mathrm{B}}^2 - \mathrm{MMD}^2)}{\sqrt{V_B}} > t_\alpha - \frac{\sqrt{b}\,\mathrm{MMD}^2}{\sqrt{V_B}}\right) \\
&\to \Phi\left(\sqrt{b}\,\frac{\mathrm{MMD}^2}{\sqrt{V_B}} - t_\alpha\right).
\end{aligned} \tag{3.6}$$

The block test's constant asymptotic threshold gives us a simple form that is cheaper to compute than using the permutation test threshold in (3.3), is valid even for small values of the population power, and only uses the samples in the form of the ratio $\mathrm{MMD}^2/\sqrt{V_B}$ - which we already know can estimated effectively [Liu et al., 2020]. We can thus estimate the asymptotic power with

$$\hat{\rho}_{b,B} = \Phi\left(\sqrt{b}\,\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2}{\sqrt{\widehat{V}_{B,\lambda}}} - t_\alpha\right). \tag{3.7}$$

$\hat{\rho}_{b,B}$ will converge uniformly to $\rho_{b,B}$ over classes of deep kernels satisfying some technical as-

sumptions as a corollary of Liu et al. [2020]; proof in Section A.2.

Using (3.7), our objective to learn a fair kernel with sensitive attribute s and target attribute t is

$$\arg\min_{\omega} \left[\hat{\rho}^s_{b,B} - \hat{\rho}^t_{b,B}\right]. \tag{3.8}$$

Although we are optimizing a kernel based on the power $\rho_{b,B}$ of a block test, we do not use blocking in our estimator; we just find a more amenable objective based on the asymptotic power of a hypothetical block test – closely related to power of the U-statistic test.

3.1.4 MMD-B-Fair: Fair Representations

So far we have shown how to learn an optimal kernel that can simultaneously achieve high power for distinguishing target attributes, and low power for sensitive attributes. If we wish to learn a feature *representation* rather than a single kernel, however, it is not enough that a *particular* kernel cannot distinguish the sensitive attribute; we would ideally like that *no* usage of that representation with any kernel can distinguish between $\mathbb{P}^s$ and $\mathbb{Q}^s$, while maintaining that at least one kernel can distinguish between $\mathbb{P}^t$ and $\mathbb{Q}^t$. That is, if we separate into a representation ϕ and a kernel κ on that representation, we would like to solve

$$\min_{\phi} \left[\max_{\kappa} \hat{\rho}^s_{b,B} - \max_{\kappa} \hat{\rho}^t_{b,B}\right] \tag{3.9}$$

The objective (3.9) could be optimized with an alternating minimax optimization scheme for the parameters of κ, looking something like an MMD-GAN [Li et al., 2017, Bińkowski et al., 2018]. We find it sufficient in our experiments to use a much simpler scheme: a grid of Gaussian kernels of varying length-scales. This finds a fairer kernel than using a single Gaussian, preventing the representation ϕ from learning to just "hide" information at a very different scale than the single κ examines, while being much simpler to implement and optimize than in alternating gradient schemes for GAN like models.

3.1.5 Correlated Features

So far in our discussion, the two-sample tests are based on the distributions $\mathbb{P}^s = \mathbb{P}_{X|S=0}$ and $\mathbb{Q}^s = \mathbb{Q}_{X|S=1}$. This setting learns a representation that optimizes the demographic parity (DP), defined as

$$\mathrm{DP} = 1 - |P(\hat{T} = 1 \mid S = 0) - P(\hat{T} = 1 \mid S = 1)|.$$

In our approach, this setting has the advantage of not requiring both target and sensitive labels simultaneously for any data point in the training set, i.e., it still works if we have separate collections of data points labeled for the target and for the sensitive attribute. Moreover, it works even

if we do not have a high-confidence labeling of the sensitive attribute, but instead have rough estimates collected e.g. via randomized response methods [Warner, 1965]. The DP setting, however, struggles when the target and sensitive attributes are strongly correlated so that the sample pairs $(S_{\mathbb{P}^t}, S_{\mathbb{Q}^t})$ and $(S_{\mathbb{P}^s}, S_{\mathbb{Q}^s})$ come from very similar pairs of distributions. This makes the objective of minimizing the test power over one pair while maximizing the test power over the other very difficult.

To address this, we instead condition the sensitive pair over the target classes, and sample points from $\mathbb{P}^{s|t} = \mathbb{P}_{X|S=0,T=t}$ and $\mathbb{Q}^{s|t} = \mathbb{Q}_{X|S=1,T=t}$ for all values of T. This is now equivalent to maximizing for the equalized odds (EO) notion of fairness with respect to all distinct target classes t, defined as

$$EO = 1 - \left| P(\hat{T} = t \mid T = t, S = 0) - P(\hat{T} = t \mid T = t, S = 1) \right|.$$

This modifies the sensitive power objectives in (3.8) and (3.9) to, summing over the possible values of t,

$$\arg\min_{\omega} \left[\left(\sum_t \hat{\rho}^{s|t}_{b,B} \right) - \hat{\rho}^t_{b,B} \right], \tag{3.10}$$

$$\min_{\phi} \left[\max_{\kappa} \left(\sum_t \hat{\rho}^{s|t}_{b,B} \right) - \max_{\kappa} \hat{\rho}^t_{b,B} \right]. \tag{3.11}$$

It is well-known that perfect demographic parity, $\mathtt{DP} = 1$, is not generally compatible with perfectly equalized odds, $\mathtt{EO} = 1$ [Barocas et al., 2018]. Even so, Theorem 3.1 of Zhao et al. [2020] shows that classifiers satisfying $\mathtt{EO} = 1$ have demographic parity gaps $\Delta_{\mathtt{DP}}$ upper-bounded by the gap of a perfect classifier, and hence training with an equalized odds criterion does not strongly compromise demographic parity.

3.1.6 A Fair Predictor

Representations with strong power on a target task are likely able to strongly distinguish at least some portion of samples as belonging to a certain value of t. If our final goal is to train a classifier, though, it will help to try to ensure our representation can classify all points well, by adding a standard classification loss for t to our objectives, e.g.

$$\min_{\phi, g} \left[\max_{\kappa} \lambda_s \left(\sum_t \hat{\rho}^{s|t}_{b,B} \right) - \max_{\kappa} \lambda_t \hat{\rho}^t_{b,B} + \lambda_{\text{cls}} L^t(g \circ \phi) \right],$$

where g is a classifier on ϕ, $L(g \circ \phi, t)$ is the cross-entropy loss of the classifier $g(\phi(x))$ with labels t,[2] and λ_s, λ_t, λ_{cls} control the relative regularization strengths. We perform an ablation study

[2]For the equalized-odds objective, we evaluate the classification loss on all samples. For the demographic parity version, we only evaluate it on the points from $S_{\mathbb{P}^t}$ and $S_{\mathbb{Q}^t}$, to ensure the method does not require any samples with

showing the significance of the classifier loss in Section 4.1

3.2 Conditionally Invariant Representation Learning

The approach described in Section 3.1 cannot handle the scenario when the sensitive feature is continuous and/or high-dimensional. This is because the MMD-based measure is computed by splitting the data on the value of the categorical sensitive variable. In this section, we instead explore the setting described in Section 1.2 where both the distractor as well as the target are continuous-valued variables and the desired representations must be conditionally independent of the distractor Z given the target Y.

3.2.1 Conditional Independence

We begin with a natural definition of conditional independence for real random variables:

Definition 3.2.1 (Daudin, 1980)**.** X and Z are Y-conditionally independent, $X \perp\!\!\!\perp Z \mid Y$, if for all test functions $g \in L^2_{XY}$ and $h \in L^2_{ZY}$, i.e. for all square-integrable functions of (X,Y) and (Z,Y) respectively, we have almost surely in Y that

$$\mathbb{E}_{XZ}\left[g(X,Y)\,h(Z,Y)\,|\,Y\right] = \mathbb{E}_X\left[g(X,Y)\,|\,Y\right]\mathbb{E}_Z\left[h(Z,Y)\,|\,Y\right]. \tag{3.12}$$

The following classic result provides an equivalent formulation:

Proposition 1 (Daudin, 1980)**.** *X and Z are Y-conditionally independent if and only if it holds for all test functions $g \in E_1 = \left\{g \in L^2_{XY} \mid \mathbb{E}_X\left[g(X,Y)\,|\,Y\right] = 0\right\}$ and $h \in E_2 = \left\{h \in L^2_{ZY} \mid \mathbb{E}_Z\left[h(Z,Y)\,|\,Y\right] = 0\right\}$ that*

$$\mathbb{E}[g(X,Y)\,h(Z,Y)] = 0. \tag{3.13}$$

Daudin [1980] notes that this condition can be further simplified (see Theorem A.3.3 for a proof):

Proposition 2 (Equation 3.9 of Daudin 1980)**.** *X and Z are Y-conditionally independent if and only if it holds for all $g \in L^2_X$ and $h \in E_2 = \left\{h \in L^2_{ZY} \mid \mathbb{E}_Z\left[h(Z,Y)\,|\,Y\right] = 0\right\}$ that*

$$\mathbb{E}[g(X)\,h(Z,Y)] = 0. \tag{3.14}$$

An equivalent way of writing this last condition (see Theorem A.4.1 for a formal proof) is:

$$\text{for all } g \in L^2_X \text{ and } h \in L^2_{ZY}, \quad \mathbb{E}\left[g(X)\left(h(Z,Y) - \mathbb{E}_{Z'}\left[h(Z',Y)\,|\,Y\right]\right)\right] = 0. \tag{3.15}$$

both s and t values.

The reduction to g not depending on Y is crucial for our method: when we are learning the representation $\varphi(X)$, then evaluating the conditional expectations $\mathbb{E}_X[g(\varphi(X),Y)\,|\,Y]$ from 1 on every minibatch in gradient descent requires impractically many samples, but $\mathbb{E}_Z[h(Z,Y)\,|\,Y]$ does not depend on X and so can be pre-computed before training the network.

3.2.2 CIRCE: Conditional Independence Regression CovariancE

The characterization (3.15) of conditional independence is still impractical, as it requires checking all pairs of square-integrable functions g and h. We will now transform this condition into an easy-to-estimate measure that characterizes conditional independence, using kernel methods.

A *kernel* $k(x,x')$ is a symmetric positive-definite function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$. A kernel can be represented as an inner product $k(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathscr{H}}$ for a *feature vector* $\phi(x) \in \mathscr{H}$, where $\mathscr{H}$ is a reproducing kernel Hilbert space (RKHS). These are spaces $\mathscr{H}$ of functions $f : \mathscr{X} \to \mathbb{R}$, with the key *reproducing property* $\langle \phi(x), f \rangle_{\mathscr{H}} = f(x)$ for any $f \in \mathscr{H}$. For M points we denote K_X. a row vector of $\phi(x_i)$, such that K_{Xx} is an $M \times 1$ matrix with $k(x_i,x)$ entries and K_{XX} is an $M \times M$ matrix with $k(x_i,x_j)$ entries. For two separable Hilbert spaces $\mathscr{G}, \mathscr{F}$, a Hilbert-Schmidt operator $A : \mathscr{G} \to \mathscr{F}$ is a linear operator with a finite Hilbert-Schmidt norm

$$\|A\|^2_{\mathrm{HS}(\mathscr{G},\mathscr{F})} = \sum\nolimits_{j\in J} \|Ag_j\|^2_{\mathscr{F}}\,, \tag{3.16}$$

where $\{g_j\}_{j\in J}$ is an orthonormal basis of $\mathscr{G}$ (for finite-dimensional Euclidean spaces, obtained from a linear kernel, A is just a matrix and $\|A\|_{\mathrm{HS}}$ its Frobenius norm). The Hilbert space $\mathrm{HS}(\mathscr{G},\mathscr{F})$ includes in particular the rank-one operators $\psi \otimes \phi$ for $\psi \in \mathscr{F}$, $\phi \in \mathscr{G}$, representing outer products,

$$[\psi \otimes \phi]g = \psi \langle \phi, g \rangle_{\mathscr{G}}\,, \qquad \langle A, \psi \otimes \phi \rangle_{\mathrm{HS}(\mathscr{G},\mathscr{F})} = \langle \psi, A\,\phi \rangle_{\mathscr{F}}\,. \tag{3.17}$$

See Gretton [2022, Lecture 5] for further details.

We next introduce a kernelized operator which (for RKHS functions g and h) reproduces the condition in (3.15), which we call the Conditional Independence Regression CovariancE (CIRCE).

Definition 3.2.2 (CIRCE operator)**.** Let $\mathscr{G}$ be an RKHS with feature map $\phi : \mathscr{X} \to \mathscr{G}$, and $\mathscr{F}$ an RKHS with feature map $\psi : (\mathscr{Z} \times \mathscr{Y}) \to \mathscr{F}$, with both kernels bounded: $\sup_x \|\phi(x)\| < \infty$, $\sup_{z,y} \|\psi(z,y)\| < \infty$. Let X, Y, and Z be random variables taking values in $\mathscr{X}$, $\mathscr{Y}$, and $\mathscr{Z}$ respectively. The *CIRCE operator* is

$$C^c_{XZ|Y} = \mathbb{E}\left[\phi(X) \otimes \left(\psi(Z,Y) - \mathbb{E}_{Z'}\left[\psi(Z',Y)\,|\,Y\right]\right)\right] \in \mathrm{HS}(\mathscr{G},\mathscr{F}). \tag{3.18}$$

For any two functions $g \in \mathscr{G}$ and $h \in \mathscr{F}$, Equation 3.18 gives rise to the same expression as in (3.15),

$$\left\langle C^c_{XZ|Y}, g \otimes h \right\rangle_{\text{HS}} = \mathbb{E}\left[g(X) \left(h(Z,Y) - \mathbb{E}_{Z'}\left[h(Z',Y) \,|\, Y \right] \right) \right]. \tag{3.19}$$

The assumption that the kernels are bounded in Theorem 3.2.2 guarantees Bochner integrability [Steinwart and Christmann, 2008, Def. A.5.20], which allows us to exchange expectations with inner products as above: the argument is identical to that of Gretton [2022, Lecture 5] for the case of the unconditional feature covariance. For unbounded kernels, Bochner integrability can still hold under appropriate conditions on the distributions over which we take expectations, e.g. a linear kernel works if the mean exists, and energy distance kernels may have well-defined feature (conditional) covariances when relevant moments exist [Sejdinovic et al., 2013].

Our goal now is to define a kernel statistic which is zero iff the CIRCE operator $C^c_{XZ|Y}$ is zero. One option would be to seek the functions, subject to a bound such as $\|g\|_{\mathscr{G}} \leq 1$ and $\|f\|_{\mathscr{F}} \leq 1$, that maximize (3.19); this would correspond to computing the largest singular value of $C^c_{XZ|Y}$. For unconditional covariances, the equivalent statistic corresponds to the Constrained Covariance, whose computation requires solving an eigenvalue problem [e.g. Gretton et al., 2005b, Lemma 3]. We instead follow the same procedure as for unconditional kernel dependence measures, and replace the spectral norm with the Hilbert-Schmidt norm [Gretton et al., 2005a]: both are zero when $C^c_{XZ|Y}$ is zero, but as we will see in Section 3.2.3 below, the Hilbert-Schmidt norm has a simple closed-form empirical expression, requiring no optimization.

Next, we show that for rich enough RKHSes $\mathscr{G}, \mathscr{F}$ (including, for instance, those with a Gaussian kernel), the Hilbert-Schmidt norm of $C^c_{XZ|Y}$ characterizes conditional independence.

Theorem 3.2.3. *For $\mathscr{G}$ and $\mathscr{F}$ with L^2-universal kernels [see, e.g., Sriperumbudur et al., 2011],*

$$\|C^c_{XZ|Y}\|_{\text{HS}} = 0 \quad \textit{if and only if} \quad X \perp\!\!\!\perp Z \mid Y. \tag{3.20}$$

The "if" direction is immediate from the definition of $C^c_{XZ|Y}$. The "only if" direction uses the fact that the RKHS is dense in L^2 , and therefore if (3.19) is zero for all RKHS elements, it must be zero for all L^2 functions. See **??** for the proof. Therefore, minimizing an empirical estimate of $\|C^c_{XZ|Y}\|_{\text{HS}}$ will approximately enforce the conditional independence we need.

Definition 3.2.4. For convenience, we define $\text{CIRCE}(X,Z,Y) = \|C^c_{XZ|Y}\|^2_{\text{HS}}$.

In the next section, we construct a differentiable estimator of this quantity from samples.

3.2.3 Empirical Estimate of the CIRCE Regularizer

To estimate CIRCE, we first need to estimate the conditional expectation $\mu_{ZY|Y}(y) = \mathbb{E}_Z[\psi(Z,y)\,|\,Y=y]$. We define[3] $\psi(Z,Y) = \psi(Z)\otimes\psi(Y)$, which for radial basis kernels (e.g. Gaussian, Laplace) is L_2-universal for (Z,Y).[4] Therefore, $\mu_{ZY|Y}(y) = \mathbb{E}_Z[\psi(Z)\,|\,Y=y]\otimes\psi(y) = \mu_{Z|Y}(y)\otimes\psi(y)$. The CIRCE operator can be written as

$$C^c_{XZ|Y} = \mathbb{E}\left[\phi(X)\otimes\psi(Y)\otimes\left(\psi(Z)-\mu_{Z|Y}(Y)\right)\right] \tag{3.21}$$

Algorithm 1 Estimation of CIRCE

Holdout data $\{(z_i,y_i)\}_{i=1}^M$, mini-batch $\{(x_i,z_i,y_i)\}_{i=1}^B$

Holdout data

Leave-one-out for λ (ridge parameter) and σ_y (parameters of Y kernel):

$\lambda,\ \sigma_y = \arg\min \sum_{i=1}^M \frac{\left\|\psi(z_i)-K_{y_iY}(K_{YY}+\lambda I)^{-1}K_Z\,.\right\|^2_{\mathscr{H}_Z}}{\left(1-\left(K_{YY}(K_{YY}+\lambda I)^{-1}\right)_{ii}\right)^2}$

$W_1 = (K_{YY}+\lambda I)^{-1},\ W_2 = W_1K_{ZZ}W_1$

Mini-batch

Compute kernel matrices $K_{xx}, K_{yy}, K_{yY}, K_{yZ}$ (x,y,z: mini-batch, Y,Z: holdout)

$\hat{K}^c = K_{yy}\odot\left(K_{zz}-K_{yY}W_1K_{Zz}-(K_{yY}W_1K_{Zz})^\top+K_{yY}W_2K_{Yy}\right)$

$\text{CIRCE} = \frac{1}{B(B-1)}\text{Tr}\left(K_{xx}\hat{K}^c\right)$

We need two datasets to compute the estimator: a holdout set of size M used to estimate conditional expectations, and the main set of size B (e.g., a mini-batch). The holdout dataset is used to estimate conditional expectation $\mu_{ZY|Y}$ with kernel ridge regression. This requires choosing the ridge parameter λ and the kernel parameters for Y. We obtain both of these using leave-one-out cross-validation; we derive a closed form expression for the error by generalizing the result of Bachmann et al. [2022] to RKHS-valued "labels" for regression.

The following theorem defines an empirical estimator of the Hilbert-Schmidt norm of the empirical CIRCE operator, and establishes the consistency of this statistic as the number of training samples B, M increases. The proof and a formal description of the conditions may be found in Section A.5.2

Theorem 3.2.5. *The following estimator of CIRCE for B points and M holdout points (for the*

[3] We abuse notation in using ψ to denote feature maps of (Y,Z), Y, and Z; in other words, we use the argument of the feature map to specify the feature space, to simplify notation.

[4] Fukumizu et al. [2008, Section 2.2] show this kernel is *characteristic*, and Sriperumbudur et al. [2011, Figure 1 (3)] that being characteristic implies L_2 universality in this case.

conditional expectation):

$$\widehat{CIRCE} = \frac{1}{B(B-1)} \mathrm{Tr}\left(K_{XX}\left(K_{YY} \odot \hat{K}^c_{ZZ}\right)\right). \tag{3.22}$$

converges as $O_p(1/\sqrt{B} + 1/M^(\beta - 1)/(2(\beta + p)))$*, when the regression in Equation A.16 is well-specified.* K_{XX} *and* K_{YY} *are kernel matrices of* X *and* Y*; elements of* K^c_{ZZ} *are defined as* $K^c_{zz'} = \left\langle \psi(z) - \mu_{Z|Y}(y), \psi(z') - \mu_{Z|Y}(y') \right\rangle$*;* $\beta \in (1,2]$ *characterizes how well-specified the solution is and* $p \in (0,1]$ *describes the eigenvalue decay rate of the covariance operator over* Y*.*

The notation $O_p(A)$ roughly states that with any constant probability, the estimator is $O(A)$. The algorithm is summarized in Algorithm 1. We can further improve the computational complexity for large training sets with random Fourier features [Rahimi and Recht, 2007]; see Section A.6.

We can use of our empirical CIRCE as a regularizer for conditionally independent regularization learning, where the goal is to learn representations that are conditionally independent of a known distractor Z. We switch from X to an *encoder* $\varphi_\theta(X)$. If the task is to predict Y using some loss $L(\varphi_\theta(X), Y)$, the CIRCE regularized loss with the regularization weight $\gamma > 0$ is as follows:

$$\min_\theta L(\varphi_\theta(X), Y) + \gamma \mathrm{CIRCE}(\varphi_\theta(X), Z, Y). \tag{3.23}$$

Chapter 4

Evaluation

4.1 Evaluating MMD-B-Fair

We evaluate both versions, (3.9) and (3.11), of our proposed regularizer; we call these MMD-B-Fair (DP) and MMD-B-Fair (Eq). We also evaluate baselines sIPM [Kim et al., 2022], FCRL [Gupta et al., 2021], CFAIR [Zhao et al., 2020] and LAFTR [Madras et al., 2018]. One testbed is the widely used UCI Adult dataset [Dua and Graff, 2017] – a structured dataset to predict whether an individual has income above $50,000 USD while being fair to their gender. We also evaluate performance on COMPAS[1] which contains criminal records of over 5000 people living in Florida. The task is to predict recidivism (binary) within the next two years while being sensitive to the race of an individual (also binary). The final dataset we evaluate on is the Heritage Health[2] dataset, which contains records of insurance claims and physician information of over 60,000 patients. The primary task is to predict Charlson index - an estimate of the risk of a patient's death over the next ten years - without being biased by the age at which they first claimed an insurance cover.

We present results of fairness-accuracy trade-offs and various downstream tasks along with an ablation study to investigate the importance of all of the terms in our loss function.

EXPERIMENTAL SETUP: We train all the algorithms across different choices of their respective fairness hyper-parameters. For both versions of our method we set λ_s to $\{0, 0.1, 1, 10, 100, 1000, 10000\}$ with a fixed λ_t and λ_{cls} of 1. For sIPM, CFAIR and LAFTR we set the regularization strength to the same set of values as λ_s, and for FCRL we use a subset of the hyper-parameters (β and λ) proposed in their paper. We train all models with a mini-batch size of 64 and report the average performance over ten independent seeds. Wherever possible, the encoder architecture is shared across different methods. We perform a χ^2-test of independence between the sensitive

[1] github.com/propublica/compas-analysis
[2] foreverdata.org/1015/

Dataset		Train	Val	Test
Adult	χ^2	1177.9	238.5	**0.0**
	p-value	3.96e-258	8.33e-54	**1.0**
COMPAS	χ^2	26.032	5.263	20.944
	p-value	3.35e-07	0.021	4.72e-06
Heritage Health	χ^2	6565.2	1606.9	8260.8
	p-value	0	0	0

Table 4.1: χ^2-test of independence between target and sensitive variables in the data.

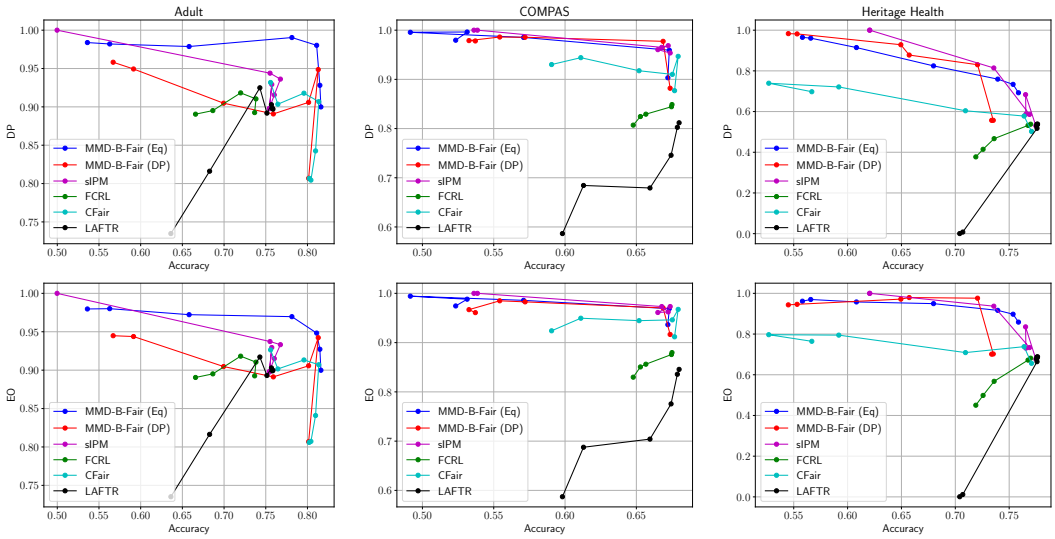

Figure 4.1: Fairness-accuracy trade-off curves on the test set of (right) Adult, (middle) COMPAS and (left) Heritage Health. Higher values for all metrics are better.

and target attributes to better understand the performance over each dataset. The test statistics and respective p-values within each split is shown in Table 4.1. In the Adult dataset there is a co-variate shift between the train and test domains where the target and sensitive variables goes from being strongly dependent in the train set to being completely independent in the test set.

4.1.1 Fairness-Accuracy Tradeoff

Firstly, we examine the fairness-accuracy tradeoff fronts obtained by sweeping over the fairness hyper-parameters in Figure 4.1. The x-axis is the target accuracy; the y-axis reports the Demographic Parity (`DP`) and Equalized Odds (`EO`), averaged over both positive and negative target classes. Note that higher values are better.

For the Adult dataset (Figure 4.1, left), MMD-B-Fair (Eq) outperforms the baselines, concur-

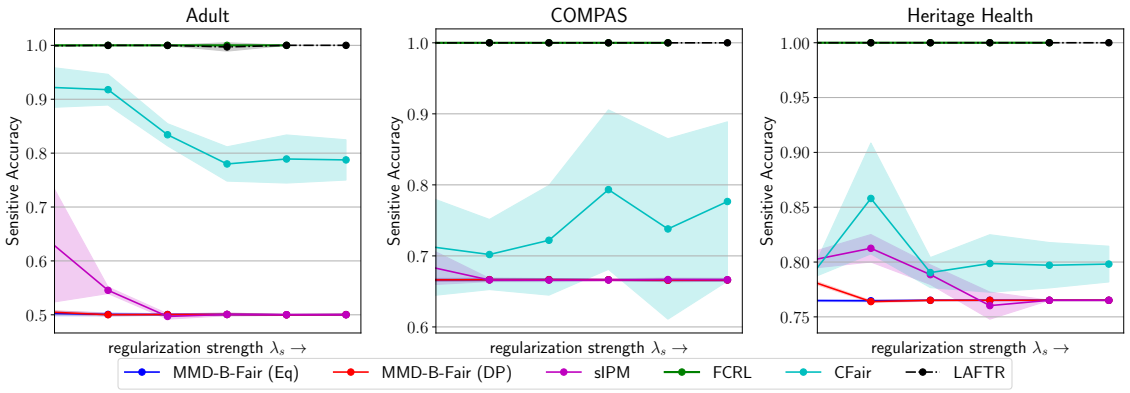

Figure 4.2: Downstream sensitive label classification over fair representations. Majority class probabilities: Adult: 0.5, COMPAS: 0.66, Heritage Health: 0.76.

rently achieving high accuracy scores and fairness measures. Recall the co-variate shift across the train and test split in this dataset further highlighting the robustness of our method compared to others. In the absence of co-variate shift across splits, both of our methods and sIPM perform equally well on the COMPAS (Figure 4.1, middle) and Heritage Health (Figure 4.1, right) datasets.

4.1.2 Examining Learnt Representations

A popular method for evaluating fair models is to examine if the learnt representations contain enough information to predict the sensitive labels: if all information regarding the sensitive attributes is successfully hidden in the representation learning phase, then subsequent classifiers will struggle to discriminate between the sensitive classes and learn to assign the majority class label to each sample to maximize the classification accuracy. This accuracy will be equal to the probability of the majority class label. On the test set, these probabilities are 0.5 for Adult, 0.66 for COMPAS and 0.76 for Heritage Health. We train MLP classifiers over the learnt representations, and show in Figure 4.2 the sensitive classification performance as a function of the fairness regularization strengths used to train the underlying fair models. Both versions of our method, as well as sIPM, are able to maintain the desired accuracy score equal to the fraction of the majority sensitive label in the test set for each dataset. sIPM converges to the ideal accuracy at sightly higher regularization strengths compared to MMD-B-Fair, while classifiers over representations from FCRL and LAFTR easily achieve perfect sensitive accuracy scores of 100% even with strong regularization indicating their failure to be invariant to sensitive information.

Checking the sensitive accuracy is essentially a classifier-based two-sample test [Lopez-Paz and Oquab, 2017] between $\mathbb{P}^s$ and $\mathbb{Q}^s$ based on the learnt representations. We also try using a more sensitive measure of whether these representations are the same: the power of an MMD two-sample test with a learned kernel, which is more general and often more powerful than a classifier-

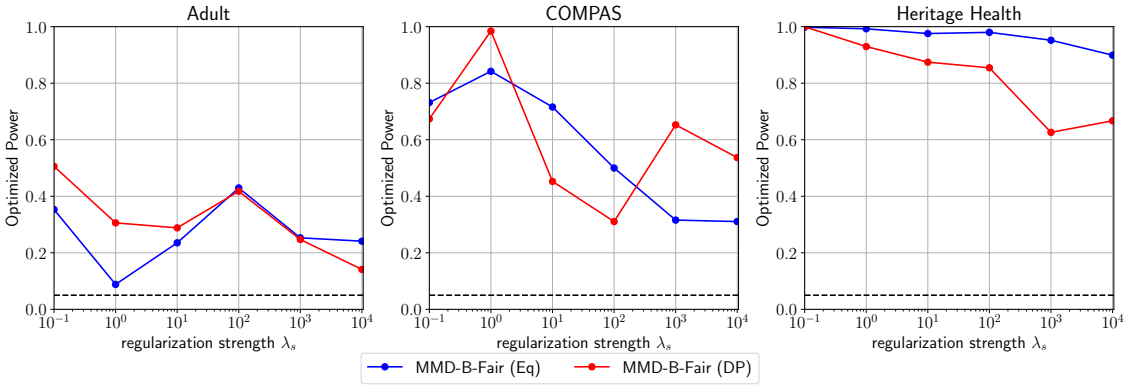

Figure 4.3: Empirical test power with an optimized kernel to maximize sensitive power over learnt representations.

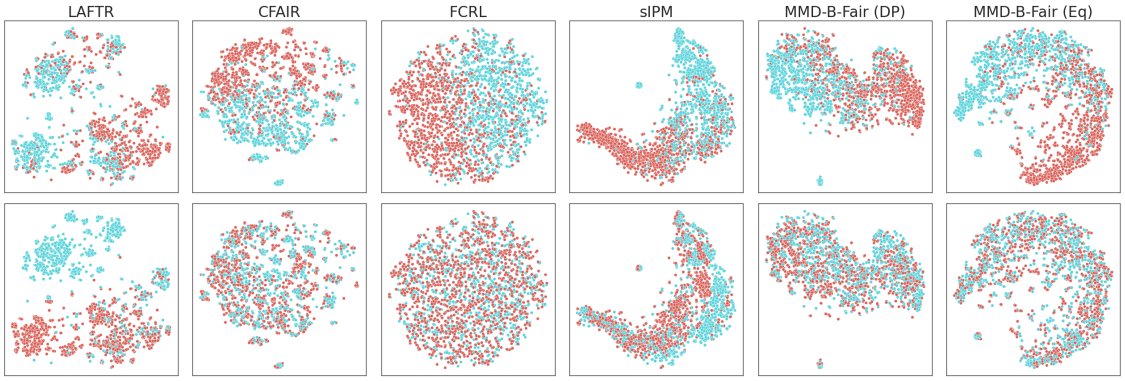

Figure 4.4: t-SNE visualizations of Adult representations, colored by target attribute (top) and sensitive attribute (bottom).

based test [Liu et al., 2020]. For models with classification accuracies significantly above 50%, this power will be near-perfect as it might be that even if few individual points can be correctly classified, a two-sample test will be able to distinguish the distributions as a whole. We run this check for our methods in Figure 4.3, using a Gaussian kernel with a learnt length-scale over a one-layer MLP architecture trained to roughly maximize the asymptotic power $\hat{\rho}^s_{b,B}$ operating on top of the fair representations as input. We then evaluate the empirical power of this test i.e., how many times it rejects the null hypothesis, while repeating the test with 64 samples at a time. As expected, two-sample tests are far more sensitive measures of attribute leakage than classification accuracy.

Figure 4.4 shows t-SNE visualizations of learnt latent space embeddings, further demonstrating that our method's representations separate the target attribute well and make the sensitive attribute difficult to distinguish.

4.1.3 Downstream Fair Transfer

A major goal of fair *representation* learning, rather than simply finding a fair classifier, is to be able to use the same representations for more than one potential downstream task. We would like our representations to have good (and fair) performance for classifiers when trained on tasks unknown at the representation learning time, even for downstream classifiers that are trained without any concern about fairness at all: the representations should enforce it.

To model this situation, we take representations learned to predict Charlson Index on Heritage Health and use them to predict each of five Primary Condition Groups, which were left out in the original representation learning phase. We train these classifiers without regard to fairness by simply minimizing the cross-entropy loss.

Table 4.2: Using Heritage Health representations to predict various downstream tasks. **Red** marks the best result per row, blue second-best, and green third-best

Transfer Label		LAFTR	CFAIR	FCRL	sIPM	MMD-B-Fair (DP)	MMD-B-Fair (Eq)
MSC2a3	acc	57.2	62.5	58.0	**72.8**	71.3	70.3
	DP	52.3	65.1	**99.2**	69.3	72.2	84.5
	Eq	57.4	70.1	**98.0**	69.9	71.8	86.6
METAB3	acc	**72.9**	72.2	53.9	72.4	70.7	69.4
	DP	52.3	65.1	**97.7**	54.5	65.6	82.1
	Eq	61.3	77.1	**97.6**	63.4	74.6	92.1
ARTHSPHIN	acc	66.4	65.9	59.3	**70.6**	67.5	67.8
	DP	52.3	65.1	**98.0**	74.6	83.0	87.7
	Eq	54.9	70.1	**98.1**	76.7	84.9	90.0
NEUMENT	acc	64.4	61.9	60.1	**68.0**	67.1	67.3
	DP	52.3	65.1	**99.1**	72.9	86.8	94.5
	Eq	54.9	69.7	**97.5**	73.2	86.7	95.4
MISCHRT	acc	71.0	67.3	69.3	**73.5**	73.0	72.5
	DP	52.3	65.1	**98.6**	85.0	87.2	96.4
	Eq	59.4	79.0	**98.2**	88.5	88.6	97.5

Table 4.2 shows the resulting accuracy scores with respect to the transfer labels and fairness scores with respect to the original sensitive labels of downstream classifiers trained on each representation. With these representations, MMD-B-Fair (Eq) provides stronger fairness results than any competitor except FCRL (which is quite inaccurate), while being more accurate than any competitor except sIPM (which is quite unfair).

4.1.4 Ablation Study

Since our objective consists of three terms - target classification loss, sensitive power and target power - we perform an ablation study in Figure 4.5 to ascertain the contribution of each term to

learning fair representations that can achieve high target accuracy. When the classification loss is turned off by setting λ_{cls} to 0, we see from the tradeoff curve that a downstream classifier trained on top of the learnt representations fail to achieve a good accuracy score. Turning off the target power instead (by setting $\lambda_t = 0$) does not have this effect, however the fairness metrics are slightly impacted at the high accuracy regime. Supposing this is not a significant drop in fairness measures, we also train a model that directly minimizes the normalized sensitive MMD instead of the power (which, recall from our discussion in Section 3.1.3 was used to balance both sensitive and target terms when used together). However, in this case we observe that the MMD measure by itself overwhelms the classifier leading to representations that are perfectly fair but come at the cost of random target classification performance.

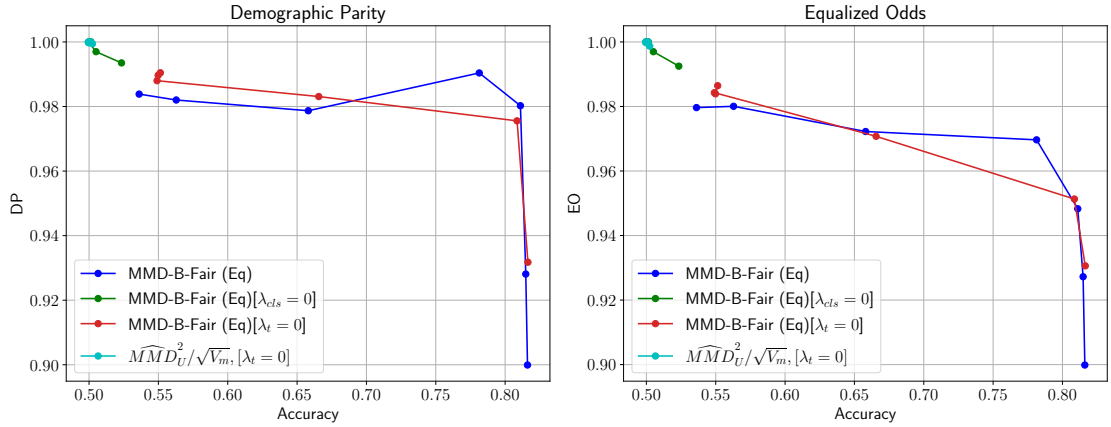

Figure 4.5: Performance ablation w.r.t. different loss terms on Adult.

4.2 Evaluating CIRCE

To evaluate CIRCE, we conduct experiments addressing two settings: (1) synthetic data of moderate dimension, to study effectiveness of CIRCE at enforcing conditional independence under established settings (as envisaged for instance in econometrics or epidemiology); and (2) high dimensional image data, with the goal of learning image representations that are robust to domain shifts. We compare performance over all experiments with HSCIC [Quinzan et al., 2022] and GCM [Shah and Peters, 2020].

4.2.1 Synthetic Data

We first evaluate performance on the synthetic datasets proposed by Quinzan et al. [2022]: these use the structural causal model (SCM) shown in Figure 4.6, and comprise 2 univariate and 2 mul-

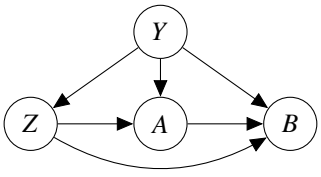

Figure 4.6: Causal structure for synthetic datasets.

tivariate cases (see Section A.7 for details). Given samples of A, Y
and Z, the goal is to learn a predictor $\hat{B} = \varphi(A,Y,Z)$ that is counterfactually invariate to Z. Achieving this requires enforcing conditional independence $\varphi(A,Y,Z) \perp\!\!\!\perp Z|Y$. For all experiments on synthetic data, we used a fully connected network with 9 hidden layers. The inputs of the network were A, Y and Z. The task is to predict B and the network is learned with the MSE loss. For each test case, we generated 10k examples, where 8k were used for training and 2k for evaluation. Data were normalized with zero mean and unit standard deviation. The rest of experimental details is provided in Section A.7.

We report in-domain MSE loss, and measure the level of counterfactual invariance of the predictor using the VCF [Quinzan et al., 2022, eq. 4; lower is better]. Given $X = (A,Y,Z)$,

$$\text{VCF} := \mathbb{E}_{x\sim\mathbf{X}}\left[\mathbb{V}_{z'\sim\mathbf{Z}}\left[\mathbb{E}_{\hat{B}^*_{z'}|X}\left[\hat{B}|X=x\right]\right]\right]. \tag{4.1}$$

$P_{\hat{B}^*_{z'}|X}$ is the counterfactual distribution of $\hat{B}$ given $X = x$ and an intervention of setting z to z'.

Univariate Cases Table 4.3 summarizes the in-domain MSE loss and VCF comparing CIRCE to baselines. Without regularization, MSE loss is low in-domain but the representation is not invariant to changes of Z. With regularization, all three methods successfully achieve counterfactual invariance in these simple settings, and exhibit similar in-domain performance.

Case	No Reg		GCM		HSCIC		CIRCE	
	MSE	VCF	MSE	VCF	MSE	VCF	MSE	VCF
1	2.03e-4	0.180	0.198	2.59e-06	0.197	2.08e-11	0.197	8.77e-08
2	0.027	0.258	1.169	9.07e-07	1.168	3.08e-11	1.168	7.37e-11

Table 4.3: MSE loss and VCF for univariate synthetic datasets. Comparison of representation without conditional independence regularization against regularization with GCM, HSCIC and CIRCE.

Multivariate Cases We present results on 2 multivariate cases: case 1 has high dimensional Z and case 2 has high dimensional Y. For each multivariate case, we vary the number of dimensions $d = \{2,5,10,20\}$. To visualize the trade-offs between in-domain performance and invariant representation, we plot the Pareto front of MSE loss and VCF. With high dimensional Z (Figure 4.7A), CIRCE and HSCIC have a similar trade-off profile, however it is notable that GCM needs to sacrifice more in-domain performance to achieve the same level of invariance. This may be because the GCM statistic is a maximum over normalized covariances of univariate residuals, which can be less effective in a multivariate setting. For high dimensional Y (Figure 4.7B), the regression

from Y to $\psi(Z)$ is much harder. We observe that HSCIC becomes less efficient with increasing d until at $d = 20$ it fails completely, while GCM still sacrifices more in-domain performance than CIRCE.

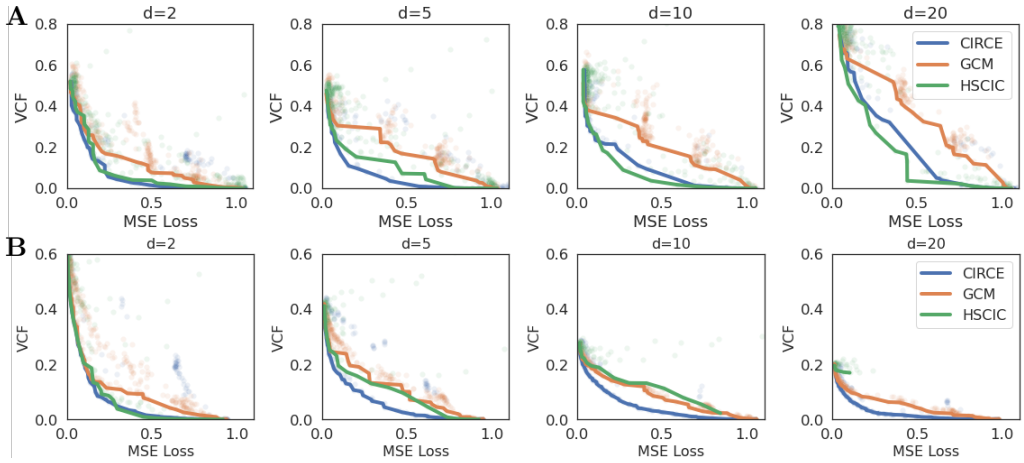

Figure 4.7: Pareto fronts of MSE and VCF for multivariate synthetic dataset. **A**: case 1; **B**: case 2.

4.2.2 Image Data

We next evaluate our method on two high-dimensional image datasets: d-Sprites (Matthey et al. [2017]) which contains images of 2D shapes generated from six independent latent factors; and the Extended Yale-B Face dataset (Georghiades et al. [2001]) of faces of 28 individuals under varying camera poses and illumination. We use both datasets with the causal graph in Figure 4.8 where the image X is directly caused by the target variable Y and a distractor Z. There also exists a strong non-causal association between Y and Z in the training set (denoted by the dashed edge).

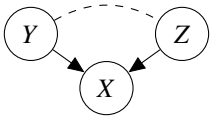

Figure 4.8: Causal structure for dSprites and Yale-B. Dashed line denotes a non-causal association between nodes.

The basic setting is as follows: for the in-domain (train) samples, the observed Y and Z are correlated through the true Y as

$$Y \sim P_Y,\ \xi_z \sim \mathcal{N}(0,\sigma_z), \qquad Z = \beta(Y) + \xi_z, \tag{4.2}$$

$$Y' = Y + \xi_y, \xi_y \sim \mathcal{N}(0,\sigma_y), \quad Z' = f_z(Y,Z,\xi_z), \quad X = f_x(Y',Z'). \tag{4.3}$$

Y and Z are observed; f_z is the structural equation for Z' (in the simplest case $Z' = Z$); f_x is the generative process of X. Y' and Z' represent noise added during generation and are unobserved.

A regular predictor would take advantage of the association β between Z and Y during training,

since this is a less noisy source of information on Y. For unseen out-of-distribution (OOD) regime, where Y and Z are uncorrelated, such solution would be incorrect.

Therefore, our task is to learn a predictor $\hat{Y} = \varphi(X)$ that is conditionally independent of Z: $\varphi(X) \perp\!\!\!\perp Z | Y$, so that during the OOD/testing phase when the association between Y and Z ceases to exist, the model performance is not harmed as it would be if $\varphi(X)$ relied on the "shortcut" Z to predict Y.

For all image experiments we use the AdamW (Loshchilov and Hutter [2019]) optimizer and anneal the learning rate with a cosine scheduler (details in Section A.8). We select the hyper-parameters of the optimizer and scheduler via a grid search to minimize the in-domain validation set loss.

d-Sprites Dataset

Of the six independent generative factors in d-Sprites, we choose the y-coordinate of the object as our target Y and the x-coordinate of the object in the image as our distractor variable Z. Our neural network consists of three convolutional layers interleaved with max pooling and leaky ReLU activations, followed by three fully-connected layers with 128, 64, 1 unit(s) respectively.

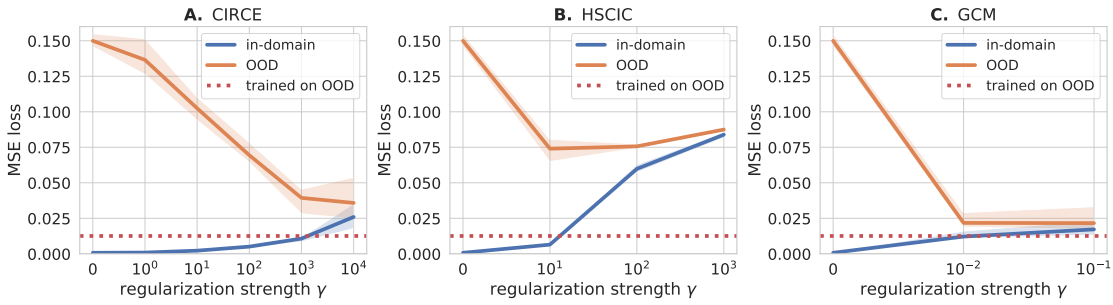

Figure 4.9: dSprites (linear). Blue: in-domain test loss; orange: out-of-domain loss (OOD); red: loss for OOD-trained encoder. Solid lines: median over 10 seeds; shaded areas: min/max values.

Linear dependence We sample images from the dataset as per the linear relation $Z' = Z = Y + \xi_z$. We then translate all sampled images (both in-domain and OOD) vertically by ξ_y, resulting in an observed object coordinate of $(Z, Y + \xi_y)$. In this case, linear residual methods, such as GCM, are able to sufficiently handle the dependence as the residual $Z - \mathbb{E}[Z|Y] = \xi_z$ is correlated with Z – which is the observed x-coordinate. As a result, penalizing the cross-covariance between $\varphi(X) - \mathbb{E}[\varphi(X)|Y]$ and $Z - \mathbb{E}[Z|Y]$ will also penalize the network's dependence on the observed x-coordinate to predict Y.

In Figure 4.9 we plot the in-domain and OOD losses over a range of regularization strengths and demonstrate that indeed GCM is able to perform quite well with a linear function relating Z to Y. CIRCE is comparable to GCM with strong regularization and outperforms HSCIC. To get the optimal OOD baseline we train our network on an OOD training set where Y and Z are uncorrelated.

Non-linear dependence To demonstrate the limitation of GCM, which simply regresses Z on Y instead of $\psi(Z,Y)$ on Y, we next address a more complex nonlinear dependence $\beta(Y) = 0$ and $Z' = Y + \alpha Z^2$. The observed coordinate of the object in the image is $(Y + \alpha\xi_z^2, Y + \xi_y)$. For a

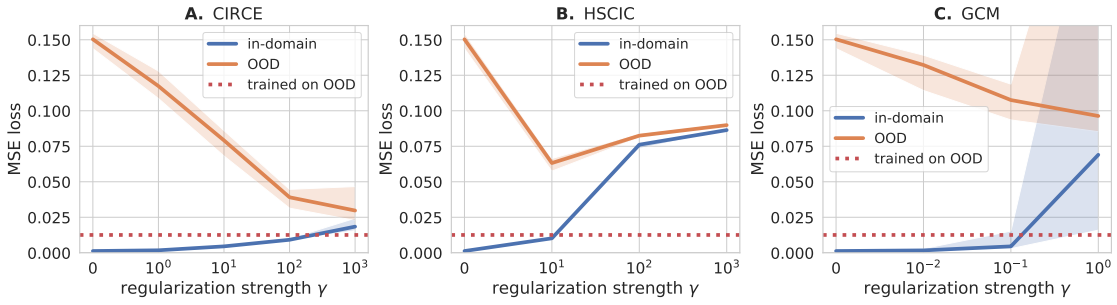

Figure 4.10: dSprites (non-linear). Blue: in-domain test loss; orange: out-of-domain loss (OOD); red: loss for OOD-trained encoder. Solid lines: median over 10 seeds; shaded areas: min/max values.

small α, the unregularized network will again exploit the shortcut, i.e. the observed x-coordinate, in order to predict Y. The linear residual, if we don't use features of Z, is $Z - \mathbb{E}[Z|Y] = \xi_z$, which is uncorrelated with $Y + \alpha\xi_z^2$, because $\mathbb{E}[\xi_z^3] = 0$ due to the symmetric and zero-mean distribution of ξ_z. As a result, penalizing cross-covariance with the linear residual (as done by GCM) will not penalize solutions that use the observed x-coordinate to predict Y. Whereas CIRCE which uses a feature map $\psi(Z)$ can capture higher order features. Results are shown in Figure 4.10: we see again that CIRCE performs best, followed by HSCIC, with GCM doing poorly. Curiously, GCM performance does still improve slightly on OOD data as regularization increases - we conjecture that the encoder $\varphi(X)$ may extract non-linear features of the coordinates. However, GCM is numerical unstable for large regularization weights, which might arise from combining a ratio normalization and a max operation in the statistic.

Extended Yale-B Dataset

Finally, we evaluate CIRCE as a regressor for supervised tasks on the natural image dataset of Extended Yale-B Faces. The task here is to estimate the camera pose Y from image X while being

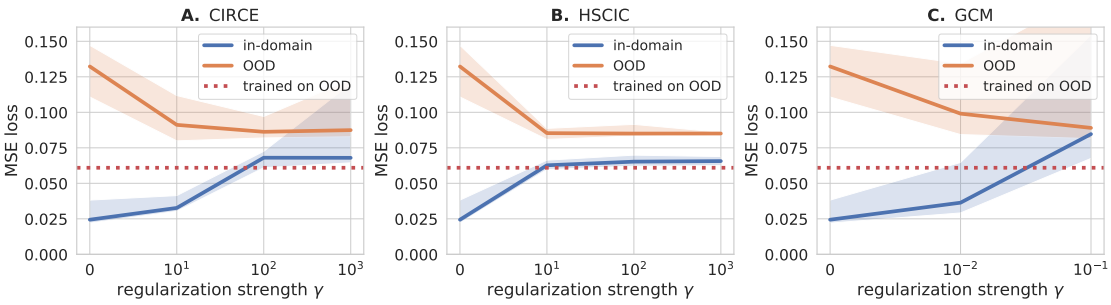

Figure 4.11: Yale-B. Blue: in-domain test loss; orange: out-of-domain loss (OOD); red: loss for OOD-trained encoder. Solid lines: median over 10 seeds; shaded areas: min/max values.

conditionally independent of the illumination Z which is represented as the azimuth angle of the light source with respect to the subject. Since, these are natural images, we use the ResNet-18 [He et al., 2016] model pre-trained on ImageNet [Deng et al., 2009] to extract image features, followed by three fully-connected layers containing 128, 64 and 1 unit(s) respectively. Here we sample the training data according to the non-linear relation $Z' = Z = 0.5(Y + \varepsilon Y^2)$, where ε is either $+1$ or -1 with equal probability. In this case $\mathbb{E}[Z|Y] = 0.5Y + 0.5Y^2\,\mathbb{E}[\varepsilon|Y] = 0.5Y$, and thus the linear residuals depend on Y. (In experiments, Y and ε are re-scaled to be in the same range. We avoid it here for simplicity.) Note that GCM can in principle find the correct solution using a linear decoder. Results are shown in Figure 4.11. CIRCE shows a small advantage over HSCIC in OOD performance for the best regularizer choice. GCM suffers from numerical instability in this example, which leads to poor performance.

Chapter 5

Conclusions

This thesis explored the use of kernel-based measures to learn deep representations that are invariant to known sensitive and/or nuisance variables in the data under two settings:

1. when both the sensitive, Z, and target variables, Y, are categorical, and therefore the data can be easily split according to the values of these features, we introduce MMD-B-Fair which optimizes *deep kernels* using the asymptotic block-power of a MMD two-sample test;

2. when the nuisance, Z, and target, Y, variables are continuous and/or high-dimensional and we require conditionally independent features, we introduce CIRCE which reduces the problem to enforcing marginal independence between the neural features and the residuals of a regression task that can be carried out offline.

Using MMD-B-Fair to learn fair representations under the notion of both demographic parity (DP) and equalized odds (EO) is a different paradigm than previous approaches to learning fair representations as it combines two-sample techniques in a novel way, using the U-statistic estimator to estimate the power of a block test, which may also be useful for other testing approaches where one may need to minimize a test power. Our method performs well compared to previous approaches based on adversarial learning and generative modelling when the dependency between the target and sensitive attributes is not the same in the train and test sets, i.e., when the i.i.d. assumption is violated. Downstream tasks like fair transfer learning also achieve a better balance between fairness and accuracy when using our learnt representations.

CIRCE provides an efficient way to learn conditionally invariant representations using mini-batches by offloading the regression of the nuisance variable on the target variable to a pre-training step. We achieve this by characterizing Daudin [1980]'s one-sided definition of conditional independence with an easy-to-estimate measure using the Hilbert Schmidt norm of the covariance operator between the neural features of the input and the residuals of a kernel ridge regression

from Y to Z. In contrast, alternative conditional independence measures when used as regularizers require additional regression steps on each mini-batch, resulting in higher variance criterions which can be less effective in complex learning tasks.

The following would be a couple of interesting questions to tackle for future work:

1. is CIRCE a statistically significant measure on a given dataset, so as to employ it as a statistic for a test of conditional dependence?

2. if Z is high-dimensional and complex, is it possible to use adversarial optimization to learn features of Z that Y must regress to based on Z's alignment with the high-dimensional input variable?

In conclusion, this work has demonstrated the use of kernel-methods under various settings to provide mathematically principled approaches to obtaining invariant representations with deep neural networks which can have numerous applications in sub-fields like fairness, domain adaptation, causal structure learning etc.

Bibliography

G. Bachmann, T. Hofmann, and A. Lucchi. Generalization through the lens of leave-one-out error. In *ICLR*, 2022. → pages 18, 47

S. Barocas, M. Hardt, and A. Narayanan. Fairness and machine learning limitations and opportunities. 2018. → page 14

A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *AIES*, 2019. → page 3

P. J. Bickel and E. L. Lehmann. Unbiased estimation in convex families. *The Annals of Mathematical Statistics*, 40(5):1523 – 1535, 1969. → page 39

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. In *ICLR*, 2018. → pages 13, 39

M. Bogen and A. Rieke. Help wanted: an examination of hiring algorithms, equity, and bias. Upturn, 2018. → page 1

A. Chouldechova, D. B. Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAT*, 2018. → page 1

E. Creager, D. Madras, J.-H. Jacobsen, M. A. Weis, K. Swersky, T. Pitassi, and R. S. Zemel. Flexibly fair representation learning by disentanglement. In *ICML*, 2019. → page 5

J. Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980. → pages 3, 4, 15, 31, 42, 43, 44

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. → page 30

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. → page 20

H. Edwards and A. J. Storkey. Censoring representations with an adversary. In *ICLR*, 2016. → pages 1, 5

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *JMLR*, 21:205–1, 2020. → page 50

A. W. Flores, K. A. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". *Federal Probation*, 80:38, 2016. → page 1

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NeurIPS*, 2008. → pages 6, 18, 47

A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE T-PAMI*, 23(6):643–660, 2001. → page 27

M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11): 1544–1547, 2018. → page 1

K. Goel, A. Gu, Y. Li, and C. Ré. Model patching: Closing the subgroup performance gap with data augmentation. In *ICLR*, 2021. → page 2

A. Gretton. Introduction to RKHS, and some simple kernel algorithms. Lecture Notes, Gatsby Computational Neuroscience Unit, 2022. URL http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html. → pages 16, 17

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pages 63–77, 2005a. → page 17

A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *JMLR*, 6:2075–2129, 2005b. → page 17

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, pages 723–773, 2012. → pages 2, 8

S. Grunewalder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *ICML*, 2012. → pages 3, 6, 46, 47

U. Gupta, A. Ferber, B. N. Dilkina, and G. V. Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *AAAI*, 2021. → pages 6, 20

S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. *KDD*, 2016. → page 5

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016. → page 30

Z. Huang, N. Deb, and B. Sen. Kernel partial correlation coefficient — a measure of conditional dependence. *JMLR*, 23(216):1–58, 2022. → page 6

N. Jean, S. M. Xie, and S. Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *NeurIPS*, 2018. → page 10

Y. Jiang and V. Veitch. Invariant and transportable representations for anti-causal domain shifts, 2022. → page 2

E. S. Jo and T. Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *FAccT*, pages 306–316, 2020. → page 1

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Considerations on fairness-aware data mining. *ICDM Workshops*, pages 378–385, 2012. → page 5

D. Kim, K. Kim, I. Kong, I. Ohn, and Y. Kim. Learning fair representation with a parametric integral probability metric. In *ICML*, 2022. → pages 6, 20

I. Kim, A. Ramdas, A. Singh, and L. Wasserman. Classification accuracy as a proxy for two sample testing. *Annals of Statistics*, 2020. → page 11

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. → page 53

I. Klebanov, I. Schuster, and T. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020. → page 46

P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, pages 728–740, 2020. → page 1

H. Lebovits. Automating inequality: How high-tech tools profile, police, and punish the poor. *Public Integrity*, 21:448 – 452, 2018. → page 1

H. Ledford. Millions of black people affected by racial bias in health-care algorithms. *Nature*, 574:608–609, 2019. → page 1

J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo. Fast and Efficient MMD-Based Fair PCA via Optimization over Stiefel Manifold. In *AAAI*, 2022. → page 5

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *NeurIPS*, 2017. → pages 10, 13

Y. Li, R. Pogodin, D. J. Sutherland, and A. Gretton. Self-supervised learning with kernel dependence maximization. In *NeurIPS*, 2021. → page 10

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Optimal rates for regularized conditional mean embedding learning. *arXiv preprint arXiv:2208.01711*, 2022. → pages 3, 46, 47, 50, 51

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *ICML*, pages 6316–6326, 2020. → pages 2, 10, 11, 12, 13, 23, 41, 42

M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, volume 31, 2018. → page 2

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. In *ICLR*, 2017. → pages 11, 22

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. → page 28

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. In *ICLR*, 2016. → pages 1, 5

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *ICML*, pages 3384–3393, 2018. → pages 1, 5, 6, 20

M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally motivated shortcut removal using auxiliary labels. In *AISTATS*, 2022. → page 2

L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset, 2017. URL https://github.com/deepmind/dsprites-dataset/. → page 27

C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989. → page 49

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 2021a. → page 2

N. Mehrabi, F. Morstatter, N. A. Saxena, K. Lerman, and A. G. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54:1 – 35, 2021b. → page 5

M. Mollenhauer and P. Koltai. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020. → page 46

S. Norouzi. Variational fair information bottleneck. 2020. → page 5

L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In *NeurIPS*, 2020. → page 5

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *NeurIPS*, 2020. → pages 3, 6, 46

A. M. Puli, L. H. Zhang, E. K. Oermann, and R. Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *ICLR*, 2022. → page 2

F. Quinzan, C. Casolo, K. Muandet, N. Kilbertus, and Y. Luo. Learning counterfactually invariant predictors. *arXiv preprint arXiv:2207.09768*, 2022. → pages 6, 25, 26

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NeurIPS*, 2007. → pages 19, 52, 53

A. Ramdas, S. J. Reddi, B. Poczos, A. Singh, and L. Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing, 2015. → page 10

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2702, 2013. → page 17

R. Serfling. *Approximation Theorems of Mathematical Statistics*. 1980. → page 10

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. → pages 7, 25

J. L. Skeem and C. T. Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54:680–712, 2016. → page 1

L. Song, J. Huang, A. J. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions. In *ICML*, 2009. → page 3

T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. K. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *KDD*, 2018. → page 5

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *JMLR*, 12:2389–2410, 2011. → pages 17, 18, 45, 47

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008. → page 17

X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu. A kernel-based causal learning algorithm. In *ICML*, pages 855–862, 2007. → page 6

D. J. Sutherland and N. Deka. Unbiased estimators for the variance of MMD estimators, 2019. → pages 11, 40

D. J. Sutherland, H.-Y. F. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017. → pages 2, 9, 10, 11, 12

R. Tachet des Combes, H. Zhao, Y.-X. Wang, and G. J. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *NeurIPS*, 2020. → page 2

V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. In *NeurIPS*, 2021a. → page 5

V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *NeurIPS*, 2021b. → pages 2, 3, 6

Z. Wang and V. Veitch. A unified causal view of domain invariant representation learning. In *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2022. → page 2

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. → page 14

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *AISTATS*, 2016. → page 10

B. Wilson, J. Hoffman, and J. H. Morgenstern. Predictive inequity in object detection, 2019. → page 1

Q. Xie, Z. Dai, Y. Du, E. H. Hovy, and G. Neubig. Controllable invariance through adversarial feature learning. In *NIPS*, 2017. → page 5

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *WWW*, 2017. → page 5

W. Zaremba, A. Gretton, and M. Blaschko. B-tests: Low variance kernel two-sample tests. In *NeurIPS*, 2013. → pages 2, 9

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013. → page 1

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *AIES*, 2018. → pages 1, 5

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2011. → page 6

H. Zhao, A. Coston, T. Adel, and G. J. Gordon. Conditional learning of fair representations. In *ICLR*, 2020. → pages 5, 6, 14, 20

Appendix A

Supporting Materials

A.1 Non-existence of an unbiased estimator of alternate MMD test power

Proposition 3. *For any fixed kernel k, let* $J(\mathbb{P},\mathbb{Q}) = \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})/\sqrt{V_m(\mathbb{P},\mathbb{Q})}$ *for some* $m > 2$. *Let* $\mathscr{P}$ *be some class of distributions such that* $\{(1-\alpha)\mathbb{P}_0 + \alpha\mathbb{P}_1 : \alpha \in [0,1]\} \subseteq \mathscr{P}$, *where* $\mathbb{P}_0 \neq \mathbb{P}_1$ *are two distributions with* $\mathrm{MMD}(\mathbb{P}_0,\mathbb{P}_1) > 0$. *Then no estimator of J can be unbiased on* $\mathscr{P}$.

Proof. We follow Bińkowski et al. [2018] in using the broad approach of Bickel and Lehmann [1969]. Let $\mathbb{P}_\alpha = (1-\alpha)\mathbb{P}_0 + \alpha\mathbb{P}_1$ denote a mixture between $\mathbb{P}_0$ and $\mathbb{P}_1$.

Suppose there is some unbiased estimator $\hat{J}(X,Y)$, meaning that for some finite n_1 and n_2,

$$\mathbb{E}_{\substack{X\sim\mathbb{P}^{n_1} \\ Y\sim\mathbb{Q}^{n_2}}} \hat{J}(X,Y) = J(\mathbb{P},\mathbb{Q}).$$

Then, for any fixed $\mathbb{Q} \in \mathscr{P}$, the function

$$\begin{aligned} R(\alpha) &= J(\mathbb{P}_\alpha,\mathbb{Q}) \\ &= \int\cdots\int \hat{J}(X,Y)\,\mathrm{d}\mathbb{P}_\alpha(X_1)\cdots\mathrm{d}\mathbb{P}_\alpha(X_{n_1})\,\mathrm{d}\mathbb{Q}^{n_2}(Y) \\ &= \int\cdots\int \hat{J}(X,Y)\,[(1-\alpha)\mathrm{d}\mathbb{P}_0(X_1) + \alpha\mathrm{d}\mathbb{P}_1(X_1)]\cdots\mathrm{d}\mathbb{Q}^{n_2}(Y) \\ &= (1-\alpha)^{n_1}\mathbb{E}_{\substack{X\sim\mathbb{P}_0^{n_1} \\ Y\sim\mathbb{Q}^{n_2}}}[\hat{J}(X,Y)] + \cdots + \alpha^{n_1}\mathbb{E}_{\substack{X\sim\mathbb{P}_1^{n_1} \\ Y\sim\mathbb{Q}^{n_2}}}[\hat{J}(X,Y)] \end{aligned}$$

must be a polynomial in α.

But, if we pick $\mathbb{Q} = \mathbb{P}_1$, we will show that

$$R(\alpha) = \frac{\mathrm{MMD}^2(\mathbb{P}_\alpha,\mathbb{P}_1)}{\sqrt{V_m(\mathbb{P}_\alpha,\mathbb{P}_1)}}$$

is not a polynomial, and thus no unbiased estimator can exist on $\mathscr{P}$.

To do this, we will need some notation, and some unfortunately tedious calculations. Let

$$\begin{aligned}
\mathbb{P}_\alpha &= (1-\alpha)\mathbb{P}_0 + \alpha\mathbb{P}_1 \\
\mu_\alpha &= \mathbb{E}_{X\sim\mathbb{P}_\alpha} k(X,\cdot) = (1-\alpha)\mu_0 + \alpha\mu_1 \\
C_\alpha &= \mathbb{E}_{X\sim\mathbb{P}_\alpha} k(X,\cdot)\otimes k(X,\cdot) = (1-\alpha)C_0 + \alpha C_1,
\end{aligned}$$

where μ_α is the kernel mean embedding of $\mathbb{P}_\alpha$, and C_α its (uncentered) covariance operator. Here $k(x,\cdot)$ is the embedding of the point x into the RKHS corresponding to the kernel k, satisfying $\langle k(x,\cdot), k(y,\cdot)\rangle = k(x,y)$, and $a\otimes b$ is the outer product of two vectors in a Hilbert space, a linear operator such that $[a\otimes b]c = a\langle b,c\rangle$.

The numerator of $R(\alpha)$ is

$$\mathrm{MMD}(\mathbb{P}_\alpha,\mathbb{P}_1)^2 = \|(1-\alpha)\mu_0 + \alpha\mu_1 - \mu_1\|^2 = (1-\alpha)^2\,\mathrm{MMD}(\mathbb{P}_0,\mathbb{P}_1).$$

The denominator is much more complex, but equation (2) of Sutherland and Deka [2019] shows that

$$\begin{aligned}
V_m(\mathbb{P}_\alpha,\mathbb{P}_1) = \frac{2}{m(m-1)}\Big[& \\
& 2(m-2)\langle\mu_\alpha, C_\alpha\mu_\alpha\rangle - (2m-3)\|\mu_\alpha\|^2 \\
& 2(m-2)\langle\mu_1, C_1\mu_1\rangle - (2m-3)\|\mu_1\|^2 \\
& + 2(m-2)\langle\mu_1, C_\alpha\mu_1\rangle + 2(m-2)\langle\mu_\alpha, C_1\mu_\alpha\rangle - 2(2m-3)\langle\mu_\alpha,\mu_1\rangle^2 \\
& - 4(m-1)\langle\mu_\alpha, (C_\alpha + C_1)\mu_1\rangle + 4(m-1)\left(\|\mu_\alpha\|^2 + \|\mu_1\|^2\right)\langle\mu_\alpha,\mu_1\rangle \\
& + \mathbb{E}_{(X,X')\sim\mathbb{P}_\alpha^2} k(X,X')^2 + \mathbb{E}_{(Y,Y')\sim\mathbb{P}_1^2} k(Y,Y')^2 + 2\mathbb{E}_{X\sim\mathbb{P}_\alpha, Y\sim\mathbb{P}_1} k(X,Y)^2\Big].
\end{aligned}$$

We need not give a full expansion of V_m in terms of α; we will merely show that it is of degree three. Since the ratio of a degree-two polynomial with the square root of a degree-three polynomial cannot possibly be itself polynomial, that will suffice to show that $R(\alpha)$ is not polynomial, and hence no unbiased estimator exists.

To see this, notice that μ_α and C_α are each linear in α, so that any term containing fewer than three such terms, e.g. $\|\mu_\alpha\|^2$ or $\langle\mu_\alpha, C_1\mu_\alpha\rangle$, cannot possibly be of degree three and so is not relevant to our goal. The expectations of squared kernels are also not relevant: the highest-order in terms of α is

$$\mathbb{E}_{X,X'\sim\mathbb{P}_\alpha} k(X,X')^2 = (1-\alpha)^2\mathbb{E}_{X,X'\sim\mathbb{P}_0} k(X,X')^2 + 2\alpha(1-\alpha)\mathbb{E}_{\substack{X\sim\mathbb{P}_0\\X'\sim\mathbb{P}_1}} k(X,X')^2 + \alpha^2\mathbb{E}_{X,X'\sim\mathbb{P}_1} k(X,X')^2$$

which is $\mathcal{O}(\alpha^2)$, abusing notation slightly to mean "terms of degree 2 or lower in α." This leaves us

$$V_m(\mathbb{P}_\alpha,\mathbb{P}_1) = \frac{2}{m(m-1)}\Big[2(m-2)\langle\mu_\alpha, C_\alpha\mu_\alpha\rangle + 4(m-1)\|\mu_\alpha\|^2\langle\mu_\alpha,\mu_1\rangle\Big] + \mathcal{O}(\alpha^2).$$

We can find the α^3 terms by

$$\begin{aligned}
\langle\mu_\alpha, C_\alpha\mu_\alpha\rangle &= (1-\alpha)\langle\mu_\alpha, C_\alpha\mu_0\rangle + \alpha\langle\mu_\alpha, C_\alpha\mu_1\rangle \\
&= \alpha\langle\mu_\alpha, C_\alpha(\mu_1-\mu_0)\rangle + \mathcal{O}(\alpha^2) \\
&= \alpha^2\langle\mu_\alpha, (C_1-C_0)(\mu_1-\mu_0)\rangle + \mathcal{O}(\alpha^2) \\
&= \alpha^3\langle\mu_1-\mu_0, (C_1-C_0)(\mu_1-\mu_0)\rangle + \mathcal{O}(\alpha^2)
\end{aligned}$$

and

$$\begin{aligned}
\|\mu_\alpha\|^2\langle\mu_\alpha,\mu_1\rangle &= \alpha\langle\mu_\alpha,\mu_\alpha\rangle\langle\mu_1-\mu_0,\mu_1\rangle + \mathcal{O}(\alpha^2) \\
&= \alpha^2\langle\mu_\alpha,\mu_1-\mu_0\rangle\langle\mu_1-\mu_0,\mu_1\rangle + \mathcal{O}(\alpha^2) \\
&= \alpha^3\langle\mu_1-\mu_0,\mu_1-\mu_0\rangle\langle\mu_1-\mu_0,\mu_1\rangle + \mathcal{O}(\alpha^2).
\end{aligned}$$

Because we assumed $\mathrm{MMD}(\mathbb{P}_0,\mathbb{P}_1) > 0$, we have $\mu_1 \neq \mu_0$. Thus these two terms cancel only if

$$\langle\mu_1-\mu_0, [(m-2)(C_1-C_0) + 2(m-1)(\mu_1-\mu_0)\otimes\mu_1](\mu_1-\mu_0)\rangle = 0.$$

Now, suppose we had defined $R(\alpha)$ with $\mathbb{Q} = \mathbb{P}_\beta$ rather than $\mathbb{P}_1$ for some other $\beta \in [0,1]$. The only relevant thing that changes is that the lone μ_1 above becomes μ_β; the numerator stays quadratic in α. Thus, if the terms cancel for μ_1, we can simply choose a different μ_β for which they do not cancel, which will always be possible. Thus the denominator is the square root of a degree-three polynomial, $R(\alpha)$ is not a polynomial, and no unbiased estimator can exist. □

A.2 Uniform convergence of our MMD power objective

We show here that optimizing the approximated block-test power from (3.7) with a finite number of samples from each conditional distribution works, i.e. as m increases, our power estimate converges uniformly over the parameter space towards an optimal solution.

Liu et al. [2020] proved the following with probability at least $1-\delta$ over the choice of n samples used in the estimators:

$$\sup_{k\in\mathscr{K}}\left|\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2}{\sqrt{n\widehat{V}_{n,n\cdot n^{-1/3}}}}-\frac{\mathrm{MMD}^2}{\sqrt{\lim_{m\to\infty}mV_m}}\right|\leq\alpha(\mathscr{K},\mathbb{P},\mathbb{Q},n,\delta) \tag{A.1}$$

for some function α (given asymptotically in their Theorem 6 and Proposition 9, or with full constants in their Theorem 11 and Proposition 23; see also their Remarks 24 and 25). Here $\mathscr{K}$ is the class of considered kernels; note that mV_m converges to a constant.

Notice from (3.4) that, for any m and ℓ, $\widehat{V}_{\ell,\lambda}=\frac{m}{\ell}\widehat{V}_{m,\lambda}$. Thus we can rewrite (3.7) as

$$\hat{\rho}_{b,B}=\Phi\left(\sqrt{b}\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2}{\sqrt{\widehat{V}_{B,\lambda}}}-t_\alpha\right)=\Phi\left(\sqrt{bB}\,\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2}{\sqrt{n\widehat{V}_{n,\lambda}}}-t_\alpha\right)=\Phi\left(\sqrt{bB}\,\hat{J}_\lambda-t_\alpha\right),$$

where we defined $\hat{J}_\lambda=\widehat{\mathrm{MMD}}_{\mathrm{U}}^2/\sqrt{n\widehat{V}_{n,\lambda}}$.

Defining $J=\mathrm{MMD}^2/\sqrt{\lim_{m\to\infty}mV_m}$, we can now rewrite (A.1) more compactly as showing that, with probability at least $1-\delta$, $\sup_{k\in\mathscr{K}}|\hat{J}_{n^{2/3}}-J|\leq\alpha(\mathscr{K},\mathbb{P},\mathbb{Q},n,\delta)$.

Also, notice from (3.6) that $\rho_{b,B}\to\Phi(\sqrt{bB}J-t_\alpha)=:R_{b,B}$, the asymptotic power of a test with b blocks of size B.

Finally, the function $x\mapsto\Phi(\sqrt{bB}x-t_\alpha)$ is Lipschitz continuous:

$$\left|\frac{\partial}{\partial x}\Phi(\sqrt{bB}x-t_\alpha)\right|=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\sqrt{bB}x-t_\alpha\right)^2\right)\leq\frac{1}{\sqrt{2\pi}}.$$

Thus, applying this function to each of the terms in (A.1) yields that, when we use $\lambda=n^{2/3}$,

$$\sup_{k\in\mathscr{K}}|\hat{\rho}_{b,B}-R_{b,B}|\leq\frac{1}{\sqrt{2\pi}}\alpha(\mathscr{K},\mathbb{P},\mathbb{Q},n,\delta).$$

This shows uniform convergence of each $\hat{\rho}_{b,B}$ to the relevant asymptotic power. By a union bound, this immediately implies uniform convergence of the objective (3.8), or (3.9) for a finite class of "top-level" kernels κ (as we use here), to the corresponding term based on asymptotic powers. (Convergence of (3.9) over an infinite class of κ would also follow with a similar argument to that of Liu et al..)

A.3 Conditional independence definitions

We first repeat the proof of the main theorem in Daudin 1980, as the missing proofs we need for the alternative definitions of independence rely on the main one.

Theorem A.3.1 (Theorem 1 of Daudin 1980)**.** *Define* $E_1 = \{g : g \in L^2_{XY}, \mathbb{E}[g|Y] = 0\}$, $E_2 = \{h : h \in L^2_{YZ}, \mathbb{E}[h|Y] = 0\}$. *Then, the following two conditions are equivalent:*

$$\begin{aligned} \mathbb{E}[g_1 h_1] &= 0 && \forall g_1 \in E_1, \forall h_1 \in E_2, \\ \mathbb{E}[gh|Y] &= \mathbb{E}[g|Y]\mathbb{E}[h|Y] && \forall g \in L^2_{XY}, \forall h \in L^2_{YZ}. \end{aligned}$$

Proof. Necessary condition: $\mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y] \implies \mathbb{E}[g_1 h_1] = 0$

Because $E_1 \subseteq L^2_{XY}$ and $E_2 \subseteq L^2_{YZ}$, for $g_1 \in E_1$ and $h_1 \in E_2$ we have

$$\begin{aligned} &\mathbb{E}[g_1 h_1|Y] = \mathbb{E}[g_1|Y]\mathbb{E}[h_1|Y] = 0 \\ \implies &\mathbb{E}[g_1 h_1] = \mathbb{E}_Y[\mathbb{E}[g_1 h_1|Y]] = 0. \end{aligned}$$

Sufficient condition: $\mathbb{E}[g_1 h_1] = 0 \implies \mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y]$

Let $g' = g - \mathbb{E}[g|Y]$ where $g \in L^2_{XY}$ and $h' = h - \mathbb{E}[h|Y]$ where $h \in L^2_{XY}$. Then, $g' \in E_1$ and $h' \in E_2$

$$\begin{aligned} \mathbb{E}[g'h'] &= \mathbb{E}[(g - \mathbb{E}[g|Y])(h - \mathbb{E}[h|Y])] \\ &= \mathbb{E}[gh - h\mathbb{E}[g|Y] - g\mathbb{E}[h|Y] + \mathbb{E}[g|Y]\mathbb{E}[h|Y]] \\ &= \mathbb{E}_Y[\mathbb{E}[(gh - h\mathbb{E}[g|Y] - g\mathbb{E}[h|Y] + \mathbb{E}[g|Y]\mathbb{E}[h|Y])|Y]] \\ &= \mathbb{E}_Y[\mathbb{E}[gh|Y] - \mathbb{E}[g|Y]\mathbb{E}[h|Y]] = 0. \end{aligned} \tag{A.2}$$

Let B be a Borel set of the image space of Y, $g* = gI_B$ where I_B is an indicator function of B. We have $\int g^{*2} dP = \int g^2 I_B dP = \int_B g^2 dP \leq \int g^2 dP < \infty$, therefore $g^* \in L^2_{XY}$. Using Equation A.2,

$$\begin{aligned} &\mathbb{E}_Y[\mathbb{E}[g^* h|Y] - \mathbb{E}[g^*|Y]\mathbb{E}[h|Y]] \\ =&\mathbb{E}_Y[\mathbb{E}[ghI_B|Y] - \mathbb{E}[gI_B|Y]\mathbb{E}[h|Y]] \\ =&\int_B \mathbb{E}[gh|Y]\,dP - \int_B \mathbb{E}[g|Y]\mathbb{E}[h|Y]\,dP = 0 \end{aligned}$$

So $\mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y]$ almost surely. □

Corollary A.3.2 (Equation 3.8 of Daudin 1980)**.** *The following two conditions are equivalent:*

$$\begin{aligned} \mathbb{E}[gh_1] &= 0 && \forall g \in L^2_{XY}, \forall h_1 \in E_2, \\ \mathbb{E}[gh|Y] &= \mathbb{E}[g|Y]\mathbb{E}[h|Y] && \forall g \in L^2_{XY}, \forall h \in L^2_{YZ}. \end{aligned}$$

Proof. Necessary condition is identical to the previous proof.

Sufficient condition: $\mathbb{E}[gh_1] = 0 \implies \mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y]$

Let $h' = h - \mathbb{E}[h|Y]$ where $h \in L^2_{YZ}$, then $h' \in E_2$

$$\begin{aligned}
\mathbb{E}[gh'] &= \mathbb{E}[g(h - \mathbb{E}[h|Y])] \\
&= \mathbb{E}[gh - g\mathbb{E}[h|Y]] \\
&= \mathbb{E}_Y[\mathbb{E}[(gh - g\mathbb{E}[h|Y])|Y]] \\
&= \mathbb{E}_Y[\mathbb{E}[gh|Y] - \mathbb{E}[g\mathbb{E}[h|Y]|Y]] \\
&= \mathbb{E}_Y[\mathbb{E}[gh|Y] - \mathbb{E}[g|Y]\mathbb{E}[h|Y]] = 0.
\end{aligned}$$

Using the same argument as for app:th:daudin$_m$*ain*,$\mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y]$ almost surely. □

Corollary A.3.3 (Equation 3.9 of Daudin 1980)**.** *The following two conditions are equivalent:*

$$\begin{aligned}
&\mathbb{E}\left[g'h_1\right] = 0 && \forall g' \in L^2_X, \forall h_1 \in E_2, \\
&\mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y] && \forall g \in L^2_{XY}, \forall h \in L^2_{YZ}.
\end{aligned}$$

Proof. Necessary condition: As $E_2 \subseteq L^2_{YZ}$ and $L^2_X \subseteq L^2_{XY}$,

$$\mathbb{E}\left[g'h_1|Y\right] = \mathbb{E}\left[g'|Y\right]\mathbb{E}[h_1|Y] = 0.$$

Sufficient condition: $\mathbb{E}[g'h_1] = 0 \implies \mathbb{E}[gh|Y] = \mathbb{E}[g|Y]\mathbb{E}[h|Y]$

Take a simple function $g_a = \sum_{i=1}^n a_i I_{A_i}$ for an integrable Borel set A_i in XY. As integrable simple functions are dense in L^2_{XY}, we only need to prove the condition for all g_a.

In our case, the indicator function decomposes as $I_{A_i} = I_{A_i^X} I_{A_i^Y}$, and therefore for $g_i = a_i I_{A_i^X}$

$$g_a = \sum_i^n g_i I_{A_i^Y}.$$

Therefore,

$$\mathbb{E}[g_a h_1] = \mathbb{E}\left[\sum_{i=1}^n I_{A_i^Y}\mathbb{E}[g_i h_1|Y]\right] = \mathbb{E}\left[\sum_{i=1}^n I_{A_i^Y} \cdot 0\right] = 0.$$

As simple functions are dense in L^2_{XY}, we immediately have $\mathbb{E}[gh_1] = 0 \; \forall g \in L^2_{XY}, h_1 \in E_2$. Applying Theorem A.3.2 concludes the proof.

□

A.4 CIRCE definition

First, we need a more convenient function class:

Lemma A.4.1. *The function class* $E_2 = \left\{h \in L^2_{ZY}, \mathbb{E}\left[h \,|\, Y\right] = 0\right\}$ *coincides with the function class* $E_2' = \left\{h' = h - \mathbb{E}\left[h \,|\, Y\right], h \in L^2_{ZY}\right\}$.

Proof. $E_2 \subseteq E_2'$: any $h \in E_2$ is in L^2_{ZY} and has the form $h = h - \mathbb{E}\left[h \,|\, Y\right]$ by construction because the last term is zero.

$E_2' \subseteq E_2$: first, any $h' \in E_2'$ satisfies $\mathbb{E}\left[h' \,|\, Y\right] = 0$ by construction. Second,

$$\int (h')^2 \, d\mu(Z,Y) = \int (h - \mathbb{E}\left[h \,|\, Y\right])^2 \, d\mu(Z,Y) \tag{A.3}$$

$$= \int \left(h^2 - 2h\,\mathbb{E}\left[h \,|\, Y\right] + \left(\mathbb{E}\left[h \,|\, Y\right]\right)^2\right) d\mu(Z,Y) \tag{A.4}$$

$$= \int \left(h^2 - \left(\mathbb{E}\left[h \,|\, Y\right]\right)^2\right) d\mu(Z,Y) < +\infty, \tag{A.5}$$

as $h \in L^2_{ZY}$ and the second term is non-positive. □

Proof of th:cond$_c$ov$_m$eans$_i$ndep For the "if" direction, we simply "pull out" the Y expectation in the definition of the CIRCE operator and apply conditional independence:

$$C^c_{XZ|Y} = \mathbb{E}_Y\Big[\mathbb{E}_X[\phi(X) \mid Y] \otimes \underbrace{\left(\mathbb{E}_Z[\psi(Z,Y) \mid Y] - \mathbb{E}_{Z'}[\psi(Z',Y) \mid Y]\right)}_{0}\Big] = 0.$$

For the other direction, first, $\|C^c_{XQ}\|_{\mathrm{HS}} = 0$ implies that for any $g \in \mathscr{G}$ and $h \in \mathscr{F}$,

$$\mathbb{E}\left[g\left(h - \mathbb{E}\left[h \,|\, Y\right]\right)\right] = 0 \tag{A.6}$$

by Cauchy-Schwarz.

Now, we use that an L_2-universal kernel is dense in L^2 by definition (see Sriperumbudur et al. [2011]). Therefore, for any $g \in L^2_X$ and $h \in L^2_{ZY}$, for any $\varepsilon > 0$ we can find $g_\varepsilon \in \mathscr{G}$ and $h_\varepsilon \in \mathscr{F}$ such that

$$\|g - g_\varepsilon\|_2 \leq \varepsilon, \; \|h - h_\varepsilon\|_2 \leq \varepsilon\,. \tag{A.7}$$

For the L^2 function, we can now write the conditional independence condition as

$$\begin{aligned}\mathbb{E}\left[g\left(h-\mathbb{E}\left[h\,|\,Y\right]\right)\right] &= \mathbb{E}\left[\left(g\pm g_\varepsilon\right)\left(h\pm h_\varepsilon-\mathbb{E}\left[h\pm h_\varepsilon\,|\,Y\right]\right)\right] &\text{(A.8)}\\ &= 0+\mathbb{E}\left[\left(g-g_\varepsilon\right)\left(h-h_\varepsilon-\mathbb{E}\left[h-h_\varepsilon\,|\,Y\right]\right)\right] &\text{(A.9)}\\ &\quad+\mathbb{E}\left[g_\varepsilon\left(h-h_\varepsilon-\mathbb{E}\left[h-h_\varepsilon\,|\,Y\right]\right)\right]-\mathbb{E}\left[\left(g-g_\varepsilon\right)\left(h_\varepsilon-\mathbb{E}\left[h_\varepsilon\,|\,Y\right]\right)\right]. &\text{(A.10)}\end{aligned}$$

The first term is zero because $\|C^c_{XQ}\|_{\mathrm{HS}}=0$. For the rest, we need to apply Cauchy-Schwarz:

$$\begin{aligned}\mathbb{E}\left[\left(g-g_\varepsilon\right)\left(h-h_\varepsilon\right)\right] &\leq \|g-g_\varepsilon\|_2\,\|h-h_\varepsilon\|_2\leq\varepsilon^2 &\text{(A.11)}\\ \mathbb{E}\left[\left(g-g_\varepsilon\right)\left(\mathbb{E}\left[h-h_\varepsilon\,|\,Y\right]\right)\right] &\leq \|g-g_\varepsilon\|_2\,\|h-h_\varepsilon\|_2\leq\varepsilon^2, &\text{(A.12)}\end{aligned}$$

where in the last inequality we used that $\mathbb{E}\left[\left(\mathbb{E}\left[X\,|\,H\right]\right)^2\right]\leq\mathbb{E}\left[X^2\right]$ for conditional expectations.

Similarly, also using the reverse triangle inequality,

$$\mathbb{E}\left[g_\varepsilon\left(h-h_\varepsilon\right)\right]\leq\varepsilon\,\|g_\varepsilon\|_2\leq\varepsilon\left(\|g\|_2+\varepsilon\right). \tag{A.13}$$

Repeating this calculation for the rest of the terms, we can finally apply the triangle inequality to show that

$$\begin{aligned}\left|\mathbb{E}\left[g\left(h-\mathbb{E}\left[h\,|\,Y\right]\right)\right]\right| &\leq 2\varepsilon^2+2\varepsilon\left(\|g\|_2+\varepsilon\right)+2\varepsilon\left(\|h\|_2+\varepsilon\right) &\text{(A.14)}\\ &= 2\varepsilon\left(3\varepsilon+\|g\|_2+\|h\|_2\right). &\text{(A.15)}\end{aligned}$$

As $\|g\|_2$ and $\|h\|_2$ are fixed and finite, we can make the bound arbitrary small, and hence $\mathbb{E}\left[g\left(h-\mathbb{E}\left[h\,|\,Y\right]\right)\right]=0$. □

A.5 Proofs for CIRCE estimators

A.5.1 Estimating the conditional mean embedding

We will construct an estimate of the term $\mathbb{E}_Z\left[\psi(Z,Y)\,|\,Y\right]$ that appears inside CIRCE, as a function of Y. We summarize the established results on conditional feature mean estimation: see [Grunewalder et al., 2012, Park and Muandet, 2020, Mollenhauer and Koltai, 2020, Klebanov et al., 2020, Li et al., 2022] for further details. To learn $\mathbb{E}\left[\psi(Q)\,|\,Y\right]$ for some feature map $\psi(q)\in\mathscr{H}_Q$ and random variable Q (both to be specified shortly), we can minimize the following loss:

$$\hat{\mu}_{Q|Y,\lambda}(y)=\operatorname*{arg\,min}_{F\in\mathscr{G}_{QY}}\sum_{i=1}^{N}\|\psi(q_i)-F(y_i)\|^2_{\mathscr{H}_Q}+\lambda\|F\|^2_{\mathscr{G}_{QY}}, \tag{A.16}$$

where $\mathscr{G}_{QY}$ is the space of functions from Y to $\mathscr{H}_Q$. The above solution is said to be *well-specified* when there exists a Hilbert-Schmidt operator $A^* \in HS(\mathscr{H}_Y, \mathscr{H}_Q)$ such that $F^*(y) = A^*\psi(y)$ for all $y \in \mathscr{Y}$, where $\mathscr{H}_Y$ is the RKHS on $\mathscr{Y}$ with feature map $\psi(y)$ [Li et al., 2022].

We now consider the case relevant to our setting, where $Q := (Z,Y)$. We define[1] $\psi(Z,Y) = \psi(Z) \otimes \psi(Y)$, which for radial basis kernels (e.g. Gaussian, Laplace) is L_2-universal for (Z,Y).[2] We then write $\mathbb{E}_Z[\psi(Z,y)\,|\,Y=y] = \mathbb{E}_Z[\psi(Z)\,|\,Y=y] \otimes \psi(y)$. The conditional feature mean $\mathbb{E}[\psi(Z)\,|\,Y]$ can be found with kernel ridge regression [Grunewalder et al., 2012, Li et al., 2022]:

$$\mu_{Z|Y}(y) \equiv \mathbb{E}[\psi(Z)\,|\,Y](y) \approx K_{yY}\left(K_{YY} + \lambda I\right)^{-1} K_{Z\cdot}. \tag{A.17}$$

where $K_{Z\cdot}$ indicates a "matrix" with rows $\psi(z_i)$, $(K_{YY})_{i,j} = k(y_i,y_j)$, and $(K_{yY})_i = k(y,y_i)$. Note that we have used the argument of k to identify which feature space it pertains to – i.e., the kernel on Z need not be the same as that on Y.

We can find good choices for the Y kernel and the ridge parameter λ by minimizing the leave-one-out cross-validation error. In kernel ridge regression, this is almost computationally free, based on the following version of a classic result for scalar-valued ridge regression. The proof generalizes the proof of Theorem 3.2 of Bachmann et al. [2022] to RKHS-valued outputs.

Theorem A.5.1 (Leave-one-out for kernel mean embeddings)**.** *Denote the predictor trained on the full dataset as $F_{\mathscr{S}}$, and the one trained without the i-th point as F_{-i}. For $\lambda > 0$ and $A \equiv K_{YY}\left(K_{YY} + \lambda I\right)^{-1}$, the leave-one-out (LOO) error for eq:kme_lossis $\frac{1}{N}\sum_{i=1}^N \|\psi(z_i) - F_{-i}(y_i)\|^2_{\mathscr{H}_Z} = \frac{1}{N}\sum_{i=1}^N \frac{\|\psi(z_i) - F_{\mathscr{S}}(y_i)\|^2_{\mathscr{H}_Z}}{(1-A_{ii})^2}$* .(A.18)

Proof. Denote the full dataset $\mathscr{S} = \{(y_i,z_i)\}_{i=1}^M$; the dataset missing the i-th point is denoted $\mathscr{S}_{-i}$. Prediction on the full dataset takes the form $F(Y) = AK_{Z\cdot}$.

Consider the prediction obtained without the M-th point (w.l.o.g.) but evaluated on y_M: $F_{-M}(y_M)$. Define a new dataset $\mathscr{Z} = \mathscr{S}_{-M} \cup \{(y_M, F_{-M}(y_M))\}$ and compute the loss for it:

$$L_{\mathscr{Z}}(F_{-M}) = \sum_{i=1}^{M-1} \|\psi(z_i) - F_{-M}(y_i)\|^2_{\mathscr{H}_Z} + \|F_{-M}(y_M) - F_{-M}(y_M)\|^2_{\mathscr{H}_Z} + \lambda\|F_{-M}\|^2_{\mathscr{G}_{ZY}} \tag{A.19}$$

$$= L_{\mathscr{S}_{-M}}(F_{-M}) \le L_{\mathscr{S}_{-M}}(F) \le L_{\mathscr{S}_{-M}}(F) + \|F_{-M}(y_M) - F(y_M)\|^2_{\mathscr{H}_Z} \le L_{\mathscr{Z}}(F), \tag{A.20}$$

where the first inequality is due to F_{-M} minimizing $L_{\mathscr{S}_{-M}}$. Therefore, F_{-M} also minimizes $L_{\mathscr{Z}}$. As A in the prediction expression $F_{\mathscr{S}}(Y) = AK_{Z\cdot}$ depends only on Y, and not on Z, F_{-M} has to have

[1] We abuse notation in using ψ to denote feature maps of (Y,Z), Y, and Z; in other words, we use the argument of the feature map to specify the feature space, to simplify notation.

[2] Fukumizu et al. [2008, Section 2.2] show this kernel is *characteristic*, and Sriperumbudur et al. [2011, Figure 1 (3)] that being characteristic implies L_2 universality in this case.

the same form as the full prediction:

$$F_{-M}(Y) = AK_{\tilde{Z}}\,., \quad K_{\tilde{z}_i,.} = \begin{cases} \psi(z_i), & i < M\,, \\ F_{-M}(y_M), & i = M\,. \end{cases} \tag{A.21}$$

This allows us to solve for $F_{-M}(y_M)$:

$$\begin{aligned} F_{-M}(y_M) &= K_{y_M Y}\left(K_{YY} + \lambda\, I\right)^{-1} K_{\tilde{Z}}\,. = \sum_{i=1}^{M} A_{Mi}\psi(z_i) && \text{(A.22)} \\ &= \sum_{i=1}^{M-1} A_{Mi}\psi(z_i) + A_{MM}\psi(z_i) \pm A_{MM}\psi(z_M) && \text{(A.23)} \\ &= \sum_{i=1}^{M} A_{Mi}\psi(z_i) - A_{MM}\psi(z_M) + A_{MM}\psi(z_i) && \text{(A.24)} \\ &= F_{\mathscr{S}}(y_M) - A_{MM}\psi(z_M) + A_{MM}\psi(z_i) && \text{(A.25)} \\ &= F_{\mathscr{S}}(y_M) - A_{MM}\psi(z_M) + A_{MM}F_{-M}(y_M)\,. && \text{(A.26)} \end{aligned}$$

As A_{MM} is a scalar, we can solve for $F_{-M}(y_M)$:

$$F_{-M}(y_M) = \frac{F_{\mathscr{S}}(y_M) - A_{MM}\psi(z_M)}{1 - A_{MM}} \tag{A.27}$$

Therefore,

$$\begin{aligned} \psi(z_M) - F_{-M}(y_M) &= \frac{(1 - A_{MM})\psi(z_M) - F_{\mathscr{S}}(y_M) + A_{MM}\psi(z_M)}{1 - A_{MM}} && \text{(A.28)} \\ &= \frac{\psi(z_M) - F_{\mathscr{S}}(y_M)}{1 - A_{MM}}\,. && \text{(A.29)} \end{aligned}$$

Taking the norm and summing this result over all points (not just M) gives the LOO error. □

A.5.2 CIRCE estimators

Lemma A.5.2. *For B points and $K^c_{zz'} = \langle \psi(z) - \mathbb{E}\left[Z\,|\,Y\right](y),\, \psi(z') - \mathbb{E}\left[Z\,|\,Y\right](y')\rangle$, the CIRCE estimator*

$$\widehat{\|C^c_{XZ|Y}\|^2}_{\mathrm{HS}} = \frac{1}{B(B-1)}\mathrm{Tr}\left(K_{XX}(K_{YY}\odot)\,K^c_{ZZ}\right) \tag{A.30}$$

has $O(1/B)$ bias and $O_p(1/\sqrt{B})$ deviation from the mean for any fixed probability of the deviation.

Proof. The bias is straightforward:

$$\begin{aligned}
&\frac{1}{B(B-1)}\mathbb{E}\left[\mathrm{Tr}\left(K_{XX}\left(K_{YY}\odot K^c_{ZZ}\right)\right)\right]\\
=&\frac{1}{B(B-1)}\mathbb{E}\left[\sum_{i,j\neq i}K_{x_ix_j}K_{y_iy_j}K^c_{z_iz_j}\right]+\frac{1}{B(B-1)}\mathbb{E}\left[\sum_i K_{x_ix_i}K_{y_iy_i}K^c_{z_iz_i}\right]\\
=&\frac{1}{B(B-1)}\sum_{i,j\neq i}\mathbb{E}_{xx'yy'zz'}\left[K_{xx'}K_{yy'}K^c_{zz'}\right]+O\left(\frac{1}{B}\right)\\
=&\|C^c_{XQ}\|^2_{\mathrm{HS}}+O\left(\frac{1}{B}\right).
\end{aligned}$$

For the variance, first note that our estimator has bounded differences. Denote $K_{TT}=K_{YY}\odot K^c_{ZZ}$ and $t=(y,z)$, if we switch one datapoint (x_i,t_i) to (x'_i,t'_i) and denote the vectors with switch coordinates as X^i,T^i

$$\begin{aligned}
&\left|\mathrm{Tr}\left(K_{XX}K_{TT}\right)-\mathrm{Tr}\left(K_{X^iX^i}K_{T^iT^i}\right)\right|\\
&\quad=\left|K_{x_ix_i}K_{t_it_i}-K_{x'_ix'_i}K_{t'_it'_i}+2\sum_{j\neq i}\left(K_{x_jx_i}K_{t_jt_i}-K_{x_jx'_i}K_{t_jt'_i}\right)\right|\\
&\quad\leq(2+4(B-1))K_{x\,\max}K_{t\,\max}\leq(4B-2)K_{x\,\max}K_{y\,\max}K^c_{z\,\max}.
\end{aligned}$$

Therefore, for any index i

$$\begin{aligned}
&\frac{1}{B(B-1)}\left|\mathrm{Tr}\left(K_{XX}\left(K_{YY}\odot K^c_{ZZ}\right)\right)-\mathrm{Tr}\left(K_{X^iX^i}\left(K_{Y^iY^i}\odot K^c_{Z^iZ^i}\right)\right)\right|\\
\leq&\frac{4B-2}{B(B-1)}K_{x\,\max}K_{y\,\max}K^c_{z\,\max}.
\end{aligned}$$

We can now use McDiarmid's inequality [McDiarmid, 1989] with

$$c=c_i=\frac{4B-2}{B(B-1)}K_{x\,\max}K_{y\,\max}K^c_{z\,\max},$$

meaning that for any $\varepsilon>0$

$$\begin{aligned}
\mathrm{P}\left(\left|\frac{\mathrm{Tr}\left(K_{XX}K_{TT}\right)}{B(B-1)}-\mathbb{E}\frac{\mathrm{Tr}\left(K_{XX}K_{TT}\right)}{B(B-1)}\right|\geq\varepsilon\right)&\leq 2\exp\left(-\frac{2\varepsilon^2}{Bc^2}\right)\\
&=2\exp\left(-\frac{2\varepsilon^2B(B-1)^2}{(4B-2)^2K^2_{x\,\max}K^2_{y\,\max}K^{2c}_{z\,\max}}\right).
\end{aligned}$$

Therefore, for any fixed probability the deviation ε from the mean decays as $O(1/\sqrt{B})$. □

Definition A.5.3. A (β, p)-kernel for a given data distribution satisfies the following conditions (see [Fischer and Steinwart, 2020, Li et al., 2022] for precise definition using interpolation spaces):

(EVD) Eigenvalues μ_i of the covariance operator C_{YY} decay as $\mu_i \leq c \cdot i^{-1/p}$.

(EMB) For $\alpha \in (p, 1]$, the inclusion map $[\mathscr{H}_Y^\alpha \hookrightarrow L_\infty(\pi)]$ is continuous and bounded by A.

(SRC) $F \in [\mathscr{G}]^\beta$ for $\beta \in [1, 2]$ (note that $\beta < 1$ would include the misspecified setting).

Lemma A.5.4. *Consider the well-specified case of conditional expectation estimation [see Li et al., 2022]. For bounded kernels over X, Z, Y and a (β, p)-kernel over Y, $F(y) = \mathbb{E}[\psi(Z) \,|\, Y](y)$, bounded $\|F\| \leq C_F$, and M points used to estimate F, define the conditional expectation estimate as*

$$\hat{F}(y) = K_{yY} \left(K_{YY} + \lambda_M I\right)^{-1} K_Z \,., \tag{A.31}$$

where $\lambda_M = \Theta(1/M^{\beta+p})$.

Then, the estimator $\operatorname{Tr}\left(K_{XX}\hat{K}^c_{ZZ}\right)/(B(B-1))$ of the "true" CIRCE estimator (i.e., with the actual conditional expectation) deviates from the true value as $O_p(1/M^{(\beta-1)/(2(\beta+p))})$.

Proof. First, decompose the difference:

$$\begin{aligned}
&\operatorname{Tr}\left(K_{XX}K^c_{ZZ}\right) - \operatorname{Tr}\left(K_{XX}\hat{K}^c_{ZZ}\right) = \operatorname{Tr}\left(K_{XX}\left(K^c_{ZZ} - K^c_{ZZ}\right)\right) && \text{(A.32)}\\
&\quad = \operatorname{Tr}\left(K_{XX}\left[\left(K^c_{ZZ} - \hat{K}^c_{ZZ}\right) \odot K_{YY}\right]\right) = \operatorname{Tr}\left(\left[K_{XX} \odot K_{YY}\right]\left(K^c_{ZZ} - \hat{K}^c_{ZZ}\right)\right), && \text{(A.33)}
\end{aligned}$$

where in the last line we used that all matrices are symmetric.

Let's concentrate on the difference:

$$\begin{aligned}
\left(K^c_{ZZ} - \hat{K}^c_{ZZ}\right)_{ij} &= \left\langle \hat{F}(y_i) - F(y_i),\, \psi(z_j)\right\rangle + \left\langle \hat{F}(y_j) - F(y_j),\, \psi(z_i)\right\rangle && \text{(A.34)}\\
&\quad + \left\langle F(y_i), F(y_j)\right\rangle - \left\langle \hat{F}(y_i), \hat{F}(y_j) \pm F(y_j)\right\rangle && \text{(A.35)}\\
&= \left\langle \hat{F}(y_i) - F(y_i),\, \psi(z_j)\right\rangle + \left\langle \hat{F}(y_j) - F(y_j),\, \psi(z_i)\right\rangle && \text{(A.36)}\\
&\quad + \left\langle F(y_i) - \hat{F}(y_i), F(y_j)\right\rangle - \left\langle \hat{F}(y_i), \hat{F}(y_j) - F(y_j)\right\rangle && \text{(A.37)}\\
&= \left\langle F(y_i) - \hat{F}(y_i), F(y_j) - \psi(z_j)\right\rangle + \left\langle F(y_j) - \hat{F}(y_j), \hat{F}(y_i) - \psi(z_j)\right\rangle. && \text{(A.38)}
\end{aligned}$$

As we're working in the well-specified case, by definition the operator $F \in \mathscr{G}$, where $\mathscr{G}$ is a vector-valued RKHS [Li et al., 2022, Definition 1]. This implies that for the function $[K_x h](\cdot) = K(\cdot, x)h$ (where $h \in \mathscr{H}_y$),

$$\langle F(x), h\rangle = \langle F, K_x h\rangle_{\mathscr{G}}\,. \tag{A.39}$$

We can now re-write the difference as

$$\left(K^c_{ZZ} - \hat{K}^c_{ZZ}\right)_{ij} = \left\langle F - \hat{F}, K_{y_i}\left(F(y_j) - \psi(z_j)\right) + K_{y_j}\left(\hat{F}(y_i) - \psi(z_j)\right)\right\rangle_{\mathscr{G}}. \tag{A.40}$$

We can use the triangle inequality and then Cauchy-Schwarz to obtain

$$\begin{aligned}
\left|\left(K^c_{ZZ} - \hat{K}^c_{ZZ}\right)_{ij}\right| &\leq \|F - \hat{F}\|_{\mathscr{G}}\left(\left\|K_{y_i}\left(F(y_j) - \psi(z_j)\right)\right\|_{\mathscr{G}} + \left\|K_{y_j}\left(\hat{F}(y_i) - \psi(z_j)\right)\right\|_{\mathscr{G}}\right) && \text{(A.41)}\\
&= \|F - \hat{F}\|_{\mathscr{G}}\left(k(y_i, y_i)\|F(y_j) - \psi(z_j)\|_{\mathscr{H}_Z} + k(y_j, y_j)\|\hat{F}(y_i) - \psi(z_j)\|_{\mathscr{H}_Z}\right) && \text{(A.42)}\\
&\leq C_1\|F - \hat{F}\|_{\mathscr{G}}\left(C_2 + C_3\|F - \hat{F}\|_{\mathscr{G}}\right), && \text{(A.43)}
\end{aligned}$$

for some positive constants $C_{1,2,3}$ (since the kernels over both z and y are bounded, F is bounded too and hence $\|\hat{F}\| \leq \|\hat{F} - F\| + \|F\|$.

As all kernels are bounded,

$$\frac{\left|\mathrm{Tr}\left(\left[K_{XX} \odot K_{YY}\right]\left(K^c_{ZZ} - \hat{K}^c_{ZZ}\right)\right)\right|}{B(B-1)} \leq C_1C_4\|F - \hat{F}\|_{\mathscr{G}}\left(C_2 + C_3\|F - \hat{F}\|_{\mathscr{G}}\right) \tag{A.44}$$

for positive constants C_1 to C_4.

Now we can use Theorem 2 of Li et al. [2022] with $\gamma = 1$ and $\lambda = \Theta(1/M^{\beta+p})$, which shows that

$$\mathrm{P}\left(\|F - \hat{F}\|_{\mathscr{G}} \leq \tau\sqrt{K}M^{-\frac{\beta-1}{2(\beta+p)}}\right) \geq 1 - 4e^{-\tau}, \tag{A.45}$$

for some positive constant K, which gives us the $O_p(1/M^{\frac{\beta-1}{2(\beta+p)}})$ deviation. □

Now we can combine the two lemmas to prove Equation 3.2.5:

Proof of Equation 3.2.5. Combining Equation A.5.2 and Equation A.5.4 and using a union bound, we obtain the $O_p(1/\sqrt{B} + 1/M^{\frac{\beta}{2(\beta+p)}})$ rate. □

Corollary A.5.5. *For B points and M holdout points, the CIRCE estimator*

$$\widehat{CIRCE} = \frac{1}{B(B-1)}\mathrm{Tr}\left(\tilde{K}_{XX}\left(\tilde{K}_{YY} \odot \hat{\tilde{K}}^c_{ZZ}\right)\right), \quad \tilde{A} = A - \mathrm{diag}(A), \tag{A.46}$$

converges as $O_p(1/\sqrt{B} + 1/M^{\frac{\beta-1}{2(\beta+p)}})$.

Proof. This follows from the previous two proofs. □

Corollary A.5.6. *For B points and M holdout points, the CIRCE estimator*

$$\widehat{CIRCE} = \frac{1}{B(B-1)}\mathrm{Tr}\left(HK_{XX}H\left(K_{YY}\odot\hat{K}^c_{ZZ}\right)\right), \quad H = I - \frac{1}{B}1_B1_B^\top \tag{A.47}$$

has bias of $O(1/B)$ *and converges as* $O_p(1/\sqrt{B} + 1/M^{\frac{\beta-1}{2(\beta+p)}})$.

Proof. This follows from the previous two proofs and the fact that K^c is a centered matrix, meaning that in expectation $HK^cH = K^c$. □

This estimator can be less biased in practice, as $\hat{K}^c_{ZZ}$ is typically biased due to conditional expectation estimation, and $H\hat{K}^cH$ re-centers it.

A.6 Random Fourier features

Random Fourier features (RFFs) [Rahimi and Recht, 2007] allow to approximate a kernel $k(x_1,x_2) \approx \frac{1}{D}\sum_{i=1}^D r_i(x_1)^\top r_i(x_2)$, and therefore $K = RR^\top$.

The algorithm to estimate CIRCE with RFFs is provided in Algorithm 2. We sample D_0 points every L iterations, but in every batch only use D of them to reduce computational costs. It takes $O(D_0M^2 + D_0^2M)$ to compute W_1^r and W_2^r every L iterations. At each iteration, it takes $O(BD^2 + B^2D)$ to compute CIRCE. Therefore, average (per iteration) cost of RFF estimation becomes $O(\frac{D_0}{L}M^2 + \frac{D_0^2}{L}M + BD^2 + B^2D)$.

Algorithm 2 Estimation of CIRCE with random Fourier features

Holdout data $\{(z_i,y_i)\}_{i=1}^M$, mini-batch $\{(x_i,z_i,y_i)\}_{i=1}^B$
Holdout data
Leave-one-out for λ (ridge parameter) and σ_y (parameters of Y kernel):
$\lambda,\ \sigma_y = \arg\min \sum_{i=1}^M \frac{\left\|\psi(z_i) - K_{y_iY}(K_{YY}+\lambda I)^{-1}K_Z\,.\right\|^2_{\mathscr{H}_Z}}{\left(1-\left(K_{YY}(K_{YY}+\lambda I)^{-1}\right)_{ii}\right)^2}$
$W_1 = (K_{YY}+\lambda I)^{-1},\ W_2 = W_1K_{ZZ}W_1$
Every L **mini-batches**
Sample D_0 RFF $R(\cdot)$
$W_1^r = R(Y)^\top W_1 R(Z),\ W_2^r = R(Z)^\top W_2 R(Z)$
Mini-batch
Use D random RFF out of D_0
Compute $R(y), R(z)$ (mini-batch)
$\hat{K}^c = K_{yy}\odot\left(K_{zz} - R(y)W_1^rR(z)^\top - \left(R(y)W_1^rR(z)^\top\right)^\top + R(y)W_2^rR(y)^\top\right)$
$\mathrm{CIRCE} = \frac{1}{B(B-1)}\mathrm{Tr}\left(HK_{xx}H\hat{K}^c\right),\ H = I - \frac{1}{B}1_B1_B^\top$

A.7 Synthetic Data for CIRCE

We used Adam [Kingma and Ba, 2015] for optimization with batch size 256, and trained the network for 100 epochs. For experiments on univariate datasets, the learning rate was 1e-4 and weight decay was 0.3; for experiments on multivariate datasets, the learning rate was 3e-4 and weight decay was 0.1. We implemented CIRCE with random Fourier features [Rahimi and Recht, 2007] (see Section A.6) of dimension 512 for Gaussian kernels. We swept over the hyperparameters, including RBF scale, regularization weight for ridge regression, and regularization weight for the conditional independence regularization strength.

All synthetic datasets are using the same causal structure as shown in Figure 4.6. Hyperparameters sweep is listed in Table A.1 and it is the same for all test cases.

Parameter	Values	
	CIRCE and HSCIC	GCM
conditional independence γ	log space between $[1, 10^4]$;	log space between $[10^{-2}, 10^{-0.5}]$
ridge regression λ	{ 0.001, 0.01, 0.1, 1 }	
RBF scale	{ 0.001, 0.01, 0.1, 1 }	

Table A.1: Hyperparameters for CIRCE, HSCIC and GCM on synthetic datasets.

A.7.1 Univariate Cases

Structural causal model for univariate case 1:

$$
\begin{aligned}
&Y, \varepsilon_Z \sim \mathcal{N}(0,1) \\
&\varepsilon_A, \varepsilon_B \sim \mathcal{N}(0,0.1) \\
&Z = Y^2 + \varepsilon_Z \\
&A = 0.5 Z \varepsilon_A + 2Y \\
&B = 0.5 \exp(-AY) \sin(2AY) + 5Z + 0.2\varepsilon_B
\end{aligned}
$$

Structural causal model for univariate case 2:

$$
\begin{aligned}
&Y, \varepsilon_Z \sim \mathcal{N}(0,1) \\
&\varepsilon_A, \varepsilon_B \sim \mathcal{N}(0,0.1) \\
&Z = Y^2 + \varepsilon_Z \\
&A = \exp(-0.5Z^2)\sin 2Z + 2Y + 0.2\varepsilon_A \\
&B = \sin(2AY)\exp(-0.5AY) + 5Z + 0.2\varepsilon_B
\end{aligned}
$$

A.7.2 Multivariate Cases

Structural causal model for multivariate case 1:

$$
\begin{aligned}
&Y, \varepsilon_{Z_i} \sim \mathcal{N}(0,1) \\
&\varepsilon_A, \varepsilon_B \sim \mathcal{N}(0,0.1) \\
&Z_i = Y^2 + \varepsilon_{Z_i} \\
&A = \exp(-0.5Z_1) + \sum_i Z_i \sin(Y) + 0.1\varepsilon_A \\
&B = \exp(-0.5Z_2)(\sum_i Z_i) + AY + 0.1\varepsilon_B
\end{aligned}
$$

Structural causal model for multivariate case 2:

$$
\begin{aligned}
&Y_i, \varepsilon_Z \sim \mathcal{N}(0,1) \\
&\varepsilon_A, \varepsilon_B \sim \mathcal{N}(0,0.1) \\
&Z = Y^T Y + \varepsilon_Z \\
&A = \exp(-0.5Z) + \sin\sum_i Y_i Z + 0.1\varepsilon_A \\
&B = \exp(-0.5Z)Z + \sum_i Y_i + Z + AY_1 + 0.1\varepsilon_B
\end{aligned}
$$

A.8 Image Data Details

For both dSpritres and Yale-B, we choose the following training hyperparameters over the validation set and *without* regularization: weight decay (1e-4, 1e-2), learning rate (1e-4, 1e-3, 1e-2) and length of training (200 or 500 epochs). These parameters are used for all runs (including the regularized ones). For dSprites, the training set contained 589824 points, and the holdout set size was 5898 points. For Yale-B, the training set contained 11405 points, and the holdout set size was 1267 points.

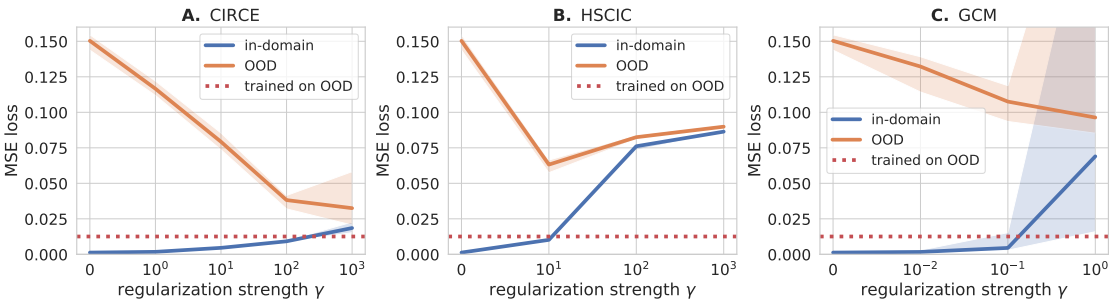

Figure A.1: dSprites with nonlinear dependence. CIRCE used holdout data in training. Blue: in-domain test loss; orange: out-of-domain loss (OOD); red: loss for OOD-trained encoder. Solid lines: median over 10 seeds; shaded areas: min/max values.

All kernels were Gaussian: $k(x,x') = \exp(-\|x-x'\|^2/(2\sigma^2))$. For Y, σ^2 from $[1.0, 0.1, 0.01, 0.001]$ and ridge regression parameter λ from $[0.01, 0.1, 1.0, 10.0, 100.0]$. The other two kernels had $\sigma^2 = 0.01$ for linear and y-cone dependencies; for the nonlinear case, the kernel over Z had $\sigma^2 = 1$ due to a different scaling of the distractor in that case.

We additionally tested a setting in which the M holdout points used for conditional expectation estimation are not removed from the training data for CIRCE. As shown in Figure A.1 for dSprites with non-linear dependence, this has little effect on the performance.