
Causally Motivated Shortcut Removal Using Auxiliary Labels

Maggie Makar
University of Michigan, Ann Arbor

Ben Packer
Google Research

Dan Moldovan
Google Research

Davis Blalock
MosaicML

Yoni Halpern
Google Research

Alexander D’Amour
Google Research

Abstract

Shortcut learning, in which models make use of easy-to-represent but unstable associations, is a major failure mode for robust machine learning. We study a flexible, causally-motivated approach to training robust predictors by discouraging the use of specific shortcuts, focusing on a common setting where a robust predictor could achieve optimal *iid* generalization in principle, but is overshadowed by a shortcut predictor in practice. Our approach uses auxiliary labels, typically available at training time, to enforce conditional independences implied by the causal graph. We show both theoretically and empirically that causally-motivated regularization schemes (a) lead to more robust estimators that generalize well under distribution shift, and (b) have better finite sample efficiency compared to usual regularization schemes, even when no shortcut is present. Our analysis highlights important theoretical properties of training techniques commonly used in the causal inference, fairness, and disentanglement literatures.

1 INTRODUCTION

Despite their immense success, predictors constructed from deep neural networks (DNNs) have been shown to lack robustness under distribution shift [Beery et al., 2018, Ilyas et al., 2019, Azulay and Weiss, 2018, Geirhos et al., 2018], especially naturally occurring

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

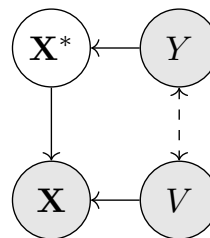


Figure 1: Causal DAG of the setting in this paper. The main label Y and auxiliary label V generate observed input \mathbf{X} , but Y only affects \mathbf{X} through \mathbf{X}^* .

distribution shifts [Taori et al., 2020]. One mechanism for this brittleness is *shortcut learning* [Geirhos et al., 2020]. Shortcut learning occurs when a predictor relies on input features that are easy to represent (i.e., shortcuts) and are predictive of the label in the training data, but whose association with the label changes under relevant distribution shifts. For example, a DNN trained for image classification could exploit correlations between the foreground object and background of images in the training distribution, and use a representation of the background as a shortcut to predict the foreground object [Beery et al., 2018, Sagawa et al., 2019]. This holds even if the foreground object alone is sufficient to achieve optimal predictive performance on the training distribution [Nagarajan et al., 2020, Sagawa et al., 2020a].

Here, we consider the problem of learning a performant predictor whose risk is invariant to interventions that change the association between irrelevant factors and the main label. Ideally, such a predictor would rely exclusively on input features that are invariant to irrelevant factors. However, identifying such invariant input features in the standard supervised learning setup is difficult, for the same reason that shortcut learning is successful: in learning setups where there are many distinct ways to construct predictors that perform well on held-out data (i.e., when the learning problem is *underspecified* [D’Amour et al., 2020]), the influence of correlated factors is difficult to disentangle

without additional supervision [Locatello et al., 2019].

For this reason, we focus on a setting where we are also given an auxiliary label that gives information about the irrelevant factor at training time. Such labels often appear in the form of metadata for the training data—for example, labels of the background of an image—but are often not available at test time. We propose an approach that exploits this auxiliary label to construct a predictor whose risk is approximately invariant across a well-defined family of test-distributions. Our method makes use of two tools from causal inference in combination: (1) weighting the training data to mimic an idealized population, and (2) enforcing an independence implied by the causal Directed Acyclic Graph (DAG) in that idealized population. While each of these approaches has been applied separately, we show here through both theoretical arguments and empirical analysis that these methods are particularly effective when applied together.

Our methodological contributions can be summarized as follows: (1) We suggest an approach to discourage shortcut learning using auxiliary labels, and specify a set of distribution shifts across which a robust model is risk-invariant. (2) We give a theoretical justification to our approach, highlighting that in some scenarios it yields models that have a lower generalization error than typical regularization schemes. We also show that our approach is robust to a set of distribution shifts. (3) We empirically validate our theoretical findings using a semi-simulated benchmark and a medical task, showing our approach has favorable in- and out-of-distribution generalization properties.

2 PRELIMINARIES

Setup. We consider a supervised learning setup where the task is to construct a predictor $f(\mathbf{X})$ parameterized by weights \mathbf{w} that predicts a label Y (e.g., foreground object) from an input \mathbf{X} (e.g., image). In addition, we have an auxiliary label V (e.g., background label) available only at training time that labels a factor of variation along which we hope the model will exhibit some invariance (e.g., background type). Throughout, we will use capital letters to denote variables, and small letters to denote their value. Our training data consist of tuples $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$ drawn from a source training distribution P_s . We restrict our focus to the case where Y and V are binary and f is a classifier. Specifically, we will consider functions f of the form $f = h(\phi(\mathbf{x}))$, where ϕ is a representation mapping and h is the final classifier.

Assumptions. We assume that P_s has a generative structure shown in Figure 1, in which the inputs \mathbf{X} are generated by the labels (Y, V) . We assume that the

labels Y and V are correlated, but not causally related; that is, an intervention on V does not imply a change in the distribution of Y , and vice versa. Such correlation often arises through the influence of an unobserved third variable such as the environment from which the data is collected. We represent this in Figure 1 with the dashed bidirectional arrow.

We assume that there is a sufficient statistic \mathbf{X}^* such that Y only affects \mathbf{X} through \mathbf{X}^* , and \mathbf{X}^* can be fully recovered from \mathbf{X} via the function $\mathbf{X}^* := e(\mathbf{X})$. However, we assume that the sufficient reduction $e(\mathbf{X})$ is unknown, so we denote \mathbf{X}^* as unobserved in Figure 1.

We also make an overlap assumption on the source distribution, P_s . Specifically we assume that $P_s(V)P_s(Y) \ll P_s(V, Y)$.

2.1 Risk Invariance and Shortcut Learning

We define the generalization risk of a function f on a distribution P as $R_P = \mathbb{E}_{X, Y \sim P}[\ell(f(X), Y)]$, where ℓ is the logistic loss.

We focus on obtaining an optimal *risk invariant* predictor, whose risk is invariant across a family of target distributions \mathcal{P} that can be obtained from P_s by interventions on the causal model in Figure 1. Specifically, we consider interventions on the confounding relationship between Y and V that keep the marginal distribution of Y constant. Each distribution in this family can be obtained by replacing the source conditional distribution $P_s(V | Y)$ with a target conditional distribution $P_t(V | Y)$:

$$\mathcal{P} = \{P_s(\mathbf{X} | \mathbf{X}^*, V)P_s(\mathbf{X}^* | Y)P_s(Y)P_t(V | Y)\}, \quad (1)$$

This family allows the marginal dependence between Y and V to change arbitrarily.

Given the family \mathcal{P} , we define the set of risk invariant predictors to be all predictors that have the same risk for all $P_t \in \mathcal{P}$, $\mathcal{F}_{\text{rinv}} = \{f : R_{P_t}(f) = R_{P'_t}(f) \quad \forall P_t, P'_t \in \mathcal{P}\}$ and an optimal risk-invariant predictor f_{rinv} to have the property $f_{\text{rinv}} \in \arg \min_{f \in \mathcal{F}_{\text{rinv}}} R_{P_t}(f) \quad \forall P_t \in \mathcal{P}$.

Any predictor that does not rely on the “shortcut” factor labeled by V will necessarily be risk invariant; thus, it is a natural property to test for and enforce for shortcut removal. Risk invariance is also independently appealing because guarantees about the performance of the predictor f_{rinv} derived under one distribution can be adapted to other distributions in \mathcal{P} .

2.2 The Unconfounded Distribution P°

Within the family of distributions \mathcal{P} , we pay special attention to the *unconfounded distribution* $P^\circ \in \mathcal{P}$

where $P^\circ(V \mid Y) := P_s(V)$. Under P° , $Y \perp\!\!\!\perp V$ and the dashed bidirectional arrow in Figure 1 can be dropped. Both our methodological approach and theoretical analysis revolve around mapping the problem of learning a risk invariant predictor under P_s to the problem of learning an optimal predictor under P° .

P° has two useful properties that are revealed by the DAG in Figure 1: (1) under the unconfounded distribution P° , the optimal predictor (with some abuse of notation) would take the form $f(\mathbf{X}^*)$, and (2) for any predictor of the form $f(\mathbf{X}^*)$, the joint distribution $P(f(\mathbf{X}^*), Y)$ (and thus the risk) is invariant across the family \mathcal{P} . Together, these imply that the optimal risk-invariant predictor $f_{\text{rinv}}(\mathbf{X}^*)$ is the optimal predictor under P° . We state this formally in Proposition 1.

Proposition 1. *Under P° , the Bayes optimal predictor is (i) only a function of \mathbf{X}^* , and (ii) an optimal risk-invariant predictor f_{rinv} with respect to \mathcal{P} .*

All proofs are shown in the appendix.

Proposition 1 motivates our approach that aims to efficiently estimate the optimal predictor under P° , even when the training data \mathcal{D} are drawn from $P_s \neq P^\circ$.

3 APPROACH

Here, we describe our approach to learning an optimal risk-invariant predictor $f_{\text{rinv}}(\mathbf{X})$ from training data $\mathcal{D} \sim P_s$. Our strategy follows the schematic $P_s \rightarrow P^\circ \rightarrow P_t$: to learn a risk invariant predictor under P_s , we map the learning problem to P° , then learn a predictor under P° that will generalize well in finite samples to any P_t in \mathcal{P} . The $P_s \rightarrow P^\circ$ step is achieved by importance weighting. The $P^\circ \rightarrow P_t$ step is achieved by a tailored regularization scheme, which we present first.

Regularization under P° . Our main innovation is a regularizer that leverages knowledge of the causal DAG to efficiently learn a classifier under P° . Specifically, we use two facts that hold under P° : (1) $V \perp\!\!\!\perp \mathbf{X}^*$, and (2) the optimal predictor is only a function of \mathbf{X}^* (Proposition 1). As we show in Section 4, this regularizer can help to reduce the sample complexity of learning without inducing bias, and directly penalizes the gap between the risk under P° and P_t .

Based on these facts, we specify a regularizer for $f = h(\phi(\mathbf{X}))$ that encourages $\phi(\mathbf{X}) \perp\!\!\!\perp V$. We do this by penalizing a distributional discrepancy between conditional distributions of the representation $P^\circ(\phi(\mathbf{X}) \mid V = 0)$ and $P^\circ(\phi(\mathbf{X}) \mid V = 1)$ that would be identical under independence. Although any number of estimable distributional discrepancy metrics could be used, here we choose to use the Maximum

Mean Discrepancy (MMD), defined as follows:

Definition 1. *Let $Z \sim P_Z$, and $Z' \sim P_{Z'}$, be two arbitrary variables. And let Ω be a class of functions $\omega : \mathcal{Z} \rightarrow \mathbb{R}$,*

$$\text{MMD}(\Omega, P_Z, P_{Z'}) = \sup_{\omega \in \Omega} (\mathbb{E}_{P_Z} \omega(Z) - \mathbb{E}_{P_{Z'}} \omega(Z')).$$

When Ω is set to be a general reproducing kernel Hilbert space (RKHS), the MMD defines a metric on probability distributions, and is equal to zero if and only if $P_Z = P_{Z'}$. Throughout, we will assume that our predictor f and our loss ℓ are contained in Ω , and in practice choose Ω to be the RKHS induced by the radial basis function (RBF) kernel. We will use the shorthand $\text{MMD}(P_Z, P_{Z'})$ to denote $\text{MMD}(\Omega, P_Z, P_{Z'})$. See Gretton et al. [2012] for a review of MMD and its empirical estimators.

Weighting to Recover P° . When the training data is drawn from some $P_s \neq P^\circ$, we weight the data to obtain empirical risk and MMD expressions that are unbiased estimates of the expressions we would obtain if $\mathcal{D} \sim P^\circ$, and proceed as before. We define weights

$$u(y, v) = \frac{P_s(Y = y)P_s(V = v)}{P_s(Y = y, V = v)}, \quad (2)$$

such that for each example, $u_i := u(y_i, v_i)$. We use \tilde{u}_i to denote u_i after normalization such that $\sum_i \tilde{u}_i = 1$. For any distribution P_s , these are importance weights that map expectations under P_s to expectations under P° . In the appendix, we show that the reweighted risk is an unbiased estimator of the risk under P° , i.e., that $\mathbb{E}_{P_s} [\hat{R}_{P_s}^u(f)] = R_{P^\circ}(f)$, where $\hat{R}_{P_s}^u(f) = \sum_i \tilde{u}_i \ell(f(\mathbf{x}_i), y_i)$, and $R_{P^\circ}(f) = \mathbb{E}_{\mathbf{X}, Y \sim P^\circ} [\ell(f(\mathbf{X}), Y)]$.

Method. Our final objective combines these weighting and regularization components. Both components of our approach are necessary together to achieve invariance and efficiency: MMD regularization without reweighting would lead to biased estimation/inaccurate models (proposition A6 in the appendix). Reweighting alone would remove the dependence on the shortcut in the asymptotic regime (proposition 1), but as is typical for reweighted estimators, it leads to high variance, which is improved by using the MMD regularizer.

Let ϕ_v denote $\{\phi(\mathbf{x}_i)\}_{i:v_i=v}$, and let u_i be as in equation 2, and $P_{\phi_v}^u$ be the weighted distribution of ϕ_v . For $\mathcal{D} \sim P_s$, and some $\alpha > 0$, the main objective to minimize is:

$$h^*, \phi^* = \underset{h, \phi}{\text{argmin}} \sum_i u_i \ell(h(\phi(\mathbf{x}_i)), y_i) \quad (3) \\ + \alpha \cdot \widehat{\text{MMD}}^2(P_{\phi_0}^u, P_{\phi_1}^u).$$

Here, $\widehat{\text{MMD}}^2$, is a weighted version of the V-statistic estimator presented in Gretton et al. [2012]. Specifically, we compute:

$$\begin{aligned} \widehat{\text{MMD}}^2 &= \sum_{i,j:v_i,v_j=0} \bar{u}_i \bar{u}_j k_\gamma(\phi_i, \phi_j) \\ &+ \sum_{i,j:v_i,v_j=1} \bar{u}_i \bar{u}_j k_\gamma(\phi_i, \phi_j) - 2 \sum_{\substack{i,j:v_i=0, \\ v_j=1}} \bar{u}_i \bar{u}_j k_\gamma(\phi_i, \phi_j), \end{aligned}$$

where $k_\gamma(x, x')$ is the radial basis function, with bandwidth γ , and the weights \bar{u}_i are a normalized version of u_i such that $\sum_{i,j:v_i,v_j=0} \bar{u}_i \bar{u}_j = \sum_{i,j:v_i,v_j=1} \bar{u}_i \bar{u}_j = \sum_{i,j:v_i=0,v_j=1} \bar{u}_i \bar{u}_j = 1$.

Cross-validation. The objective function in (3) depends on two hyperparameters: the cost of the MMD penalty α , and the penalty’s kernel bandwidth γ . Unlike many regularizers, the MMD penalty depends on the distribution of the data, and is vulnerable to overfitting, such that the estimated MMD on the training data underestimates the population MMD. For this reason, we follow a two-step cross-validation procedure. In the first step, we calculate the weighted MMD on each of the K validation folds. We test if the achieved MMDs are statistically significantly different from zero using a T-test and exclude all models with a p-value < 0.05 . In the second step, we pick the best performing model out of this subset of candidate functions.

4 THEORY

We analyze our approach in some important special cases to show how it encourages efficient learning of a risk-invariant predictor on \mathcal{P} . We focus on the generalization gap of a predictor learned under a source distribution P_s when evaluated on any target distribution $P_t \in \mathcal{P}$. We define this gap as the difference between the target risk R_{P_t} and the weighted empirical risk $\hat{R}_{P_s}^u$ for a learned function f . In keeping with our schematic $P_s \rightarrow P^\circ \rightarrow P_t$ (described in Section 3), we decompose the gap into a term about generalizing from P_s to P° , and a term about generalizing from P° to P_t :

$$\begin{aligned} R_{P_t}(f) - \hat{R}_{P_s}^u(f) &= \underbrace{R_{P_t}(f) - R_{P^\circ}(f)}_{\text{Structural risk gap } (P^\circ \rightarrow P_t)} \\ &+ \underbrace{R_{P^\circ}(f) - \hat{R}_{P_s}^u(f)}_{\text{Learning gap } (P_s \rightarrow P^\circ)} \end{aligned} \quad (4)$$

We show that constraints on the MMD between the learned representation distributions $P^\circ(\phi(\mathbf{X}) \mid V = v)$ for $v \in \{0, 1\}$, which we denote $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ)$, can translate to constraints on both of these gaps. We treat

the structural risk gap first, but spend the bulk of our effort on the finite-sample gap in the linear case.

4.1 Bounding the structural risk gap

We begin by bounding the structural risk gap, $R_{P_t}(f) - R_{P^\circ}(f)$, which characterizes how the risk for a given function f learned on data from P° is related to the risk on any target distribution $P_t \in \mathcal{P}$. Here, we show that a bound on $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ)$ translates directly to a bound on the structural risk gap in the representable case, when Y can be written as a function of the representation $\phi(\mathbf{X})$.

Proposition 2. *Suppose that $f = h(\phi(\mathbf{X}))$ is a predictor that satisfies $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ) \leq \tau$. Suppose that y is ϕ -representable, i.e., that there exists $g(\phi(\mathbf{x})) = y$, and that $g(\phi)\ell(\phi) \in \Omega$. Then $R_{P_t}(f) < R_{P^\circ}(f) + 2\tau$.*

Proposition 2 states that by encouraging small values of τ , the MMD penalty regularizes the solution toward a predictor that has similar risks on P_t and P° , reducing the gap in equation (4).

4.2 Bounding the learning gap

We now analyze how the MMD regularizer constrains the learning gap, $R_{P^\circ}(f) - \hat{R}_{P_s}^u(f)$. This gap characterizes how the weighted empirical risk, which is minimized during training on data drawn from P_s , translates to risk on the ideal distribution P° . Here, we consider the special case of linear models, where ϕ is a linear mapping, i.e., $\phi(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, and h is the sigmoid, i.e., $h(x) = \sigma(x) = 1/(1 + \exp(-x))$. Our analysis establishes some key insights about how the MMD regularizer constrains the learning problem, and when we expect it to provide significant efficiency gains. Extensions of our theoretical analysis to more complex neural networks are possible (e.g., through approaches studied in Golowich et al. [2018]).

Controlling complexity. There are two issues to consider in bounding this gap. The first is the fundamental generalization gap that we would face if we were learning f directly on “ideal” samples from P° , and the second is the additional variability incurred by importance weighting to translate the learning problem to P° to P_s , from which we actually observe samples. Regularization affects both the “ideal” and weighted learning problems in the same way, by restricting the complexity of the function class \mathcal{F} . Here, we focus on the Rademacher complexity of the function class obtained by constraining $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ)$.

Definition 2. *Let $\epsilon = \{\epsilon_i\}_{i=1}^n$ denote a vector of independent random variables drawn from the Rademacher distribution, i.e., uniform on $\{-1, 1\}$. For a function family \mathcal{F} , and $\mathcal{D} \sim P$, the Rademacher com-*

plexity for a sample of size n is defined as: $\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\mathbf{x}_i) \right]$.

In the ideal case, where $P_s = P^\circ$, the Rademacher complexity translates directly to a bound on the learning gap. Specifically, the learning gap increases as $\mathfrak{R}(\mathcal{F})$ increases (see [Mohri et al., 2018]). For the case where $P_s \neq P^\circ$, and weighting is necessary, we can obtain a bound that is similarly monotonic in $\mathfrak{R}(\mathcal{F})$ using the technique of Cortes et al. [2010]. Due to space limitations, we focus on analyzing the Rademacher complexity of the MMD regularized space in the main text and present the full generalization error bounds for both cases (when $P_s = P^\circ$ and when $P_s \neq P^\circ$) in the appendix.

Comparison with L2 Regularization To analyze how constraints on $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ)$ affect Rademacher complexity, we study how the addition of an MMD constraint shrinks a standard L_2 -regularized function class. Define the two function classes:

$$\mathcal{F}_{L_2} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A\}, \quad (5)$$

$$\mathcal{F}_{L_2, \text{MMD}} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A, \text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ) \leq \tau\}.$$

Before analyzing the precise reduction in complexity in moving from \mathcal{F}_{L_2} to $\mathcal{F}_{L_2, \text{MMD}}$, we note that this reduction is a “free lunch”: because we know that the MMD constraint is compatible with the true independence structure of P° , this reduction does not introduce bias. We formalize this in Proposition 3.

Proposition 3. *Under P° , and for \mathcal{F}_{L_2} such that $f_{\text{rinv}} \in \mathcal{F}_{L_2}$, there exists $\mathcal{F}_{L_2, \text{MMD}} \subseteq \mathcal{F}_{L_2}$ such that $f_{\text{rinv}} \in \mathcal{F}_{L_2, \text{MMD}}$. And the smallest $\mathcal{F}_{L_2, \text{MMD}}$ such that $f_{\text{rinv}} \in \mathcal{F}_{L_2, \text{MMD}}$ has $\tau = 0$.*

To study the reduction in complexity, we construct comparable Rademacher complexity upper bounds for \mathcal{F}_{L_2} and $\mathcal{F}_{L_2, \text{MMD}}$. We focus on a specific implication of the MMD constraint in the case where ϕ is linear. Here, the constraint restricts the projection of the weights \mathbf{w} onto the vector that distinguishes the conditional means: $\Delta := \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, where $\boldsymbol{\mu}_v := \mathbb{E}_{\mathbf{x} \sim P^\circ}[\mathbf{X} | V = v]$. Δ is the average change in \mathbf{X} caused by intervening to change the V under P° . Define the projection matrix $\Pi := \Delta(\Delta^\top \Delta)^{-1} \Delta^\top = \frac{\Delta \Delta^\top}{\|\Delta\|_2^2}$, which projects any vector onto Δ , and $\mathbf{w}_\perp := \Pi \mathbf{w}$ as the projection of \mathbf{w} onto the mean distinguishing vector Δ , which can be thought of as the “irrelevant” dimension of \mathbf{X} . We can directly relate $\|\mathbf{w}_\perp\|$ to the MMD constraint, in the following proposition.

Proposition 4. *Let $f(\mathbf{x}) = \sigma(\phi(\mathbf{x})) = \sigma(\mathbf{w}^\top \mathbf{x})$ be a function contained in $\mathcal{F}_{L_2, \text{MMD}}$. Then, $\|\mathbf{w}_\perp\| \leq \frac{\tau}{\|\Delta\|}$.*

Proposition 4 says that the MMD constraint limits the effect of the irrelevant components of \mathbf{w} proportionally to τ . In the image classification example, this means that the parts of \mathbf{w} that discriminate between images with different background types is constrained by τ .

Using this fact, we derive comparable bounds on the Rademacher complexity of the two function classes by splitting f in terms of how the observed features \mathbf{x} align with the mean difference vector Δ . Analogously to \mathbf{x}_\perp above, let $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$ be the orthogonal component to the mean discrepancy vector Δ .

Proposition 5. *Let $\mathbf{x}_\perp := \Pi \mathbf{x}$, $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$. For training data $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, $\sup_{\mathbf{x}_\perp} \|\mathbf{x}_\perp\|_2 \leq B_\perp$, $\sup_{\mathbf{x}_\parallel} \|\mathbf{x}_\parallel\|_2 \leq B_\parallel$, then*

$$\mathfrak{R}(\mathcal{F}_{L_2}) \leq \frac{A \sqrt{B_\parallel^2 + B_\perp^2}}{\sqrt{n}},$$

$$\mathfrak{R}(\mathcal{F}_{L_2, \text{MMD}}) \leq \frac{A \cdot B_\parallel + \tau \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}}.$$

The proof applies a standard Rademacher complexity bound for the L_2 class, and applies a more conservative strategy for $\mathcal{F}_{L_2, \text{MMD}}$ by separately bounding the worst-case terms involving \mathbf{x}_\perp and \mathbf{x}_\parallel . Nonetheless, comparing these bounds is instructive. The upper bound on $\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2})$ is smaller than that of $\mathfrak{R}(\mathcal{F}_{L_2})$ whenever τ satisfies:

$$0 \leq \tau < A \left[\sqrt{B_\parallel^2 + B_\perp^2} - B_\parallel \right] \frac{\|\Delta\|}{B_\perp}. \quad (7)$$

The ratio $\|\Delta\|/B_\perp$ characterizes how much of the variation in \mathbf{x}_\perp comes from the mean shift in \mathbf{x} induced by v . When the variation from this shift is large relative to variation in the \mathbf{x}_\parallel directions, we expect even weak MMD regularization to yield better generalization than L_2 regularization alone. This occurs in cases where V controls features that are highly salient in the input \mathbf{x} . For example, in image classification, if V denotes the background type, we expect $\|\Delta\|/B_\perp$ to be large if the background features are very different between $V = 0$ and $V = 1$, but relatively consistent within values of V . Further, if the background accounts for the majority of pixels in each image, we expect $B_\perp \gg B_\parallel$, resulting in an even stronger regularizing effect from the MMD penalty.

Why is weighting important? When $P_s \neq P^\circ$, enforcing the invariance penalty without reweighting leads to a model that is inconsistent with the causal DAG, and is hence biased (i.e., inaccurate). This means that typical invariance penalties (e.g., Krueger et al. [2021], Donini et al. [2018]) will lead to an invariance-accuracy tradeoff. We show this formally in the appendix, highlighting why enforcing the MMD penalty without reweighting when $P_s \neq P^\circ$ can introduce biased estimation. These findings have implications that extend beyond shortcut learning, e.g.,

fairness and causality where the MMD penalty is frequently used.

5 EXPERIMENTS

We now empirically demonstrate that our approach mitigates shortcut learning by showing that predictors learned with our approach are robust to distribution shifts that invalidate the shortcut. We study two tasks: predicting bird types from images, and detecting pneumonia using chest X-rays¹.

5.1 Water birds

We test our method on a semi-synthetic task adapted from Sagawa et al. [2019]. The goal is to predict the type of bird (Y) appearing in an image (\mathbf{X}) considering the background (V) to be a possible shortcut.

Data. We construct a dataset that combines images of water birds (Gulls) and land birds (Warblers) extracted from the Caltech-UCSD Birds-200-2011 (CUB) dataset [Wah et al., 2011] with water and land background extracted from the Places dataset [Zhou et al., 2017]. Additional details about data generation, hyperparameter selection, and code are included in the supplementary materials.

We examine how a model trained under a specific distribution P_s generalizes to various test distributions $P_t \in \mathcal{P}$. We consider two different cases for P_s . In the first, we generate the training data from the ideal distribution P° , with $P^\circ(Y|V=1) = P^\circ(Y|V=0) = 0.5$. This setting tests the implications of proposition 5, which suggests that our approach improves sample efficiency even under uncorrelated sampling. In the second setting, the training data is sampled from a P_s where the auxiliary label and the main label are correlated, i.e., $V \not\perp Y$. Specifically, we generate the data such that $P_s(Y=1|V=1) = P_s(Y=0|V=0) = 0.9$. For both settings, we introduce noise by randomly flipping 1% of each of the labels. We generate a number of held-out test sets, each one corresponding to a different probability of observing a waterbird with a water background, and similarly with land birds.

We use ResNet-50 [He et al., 2016], pretrained on ImageNet, and fine tuned for our task. All models in this paper are implemented in TensorFlow [Abadi et al., 2015]. We present the results from 20 simulations. In each simulation, we generate different train/test splits, and different bird-background combinations.

Baselines. We compare the following methods in our experiments: (1) **wMMD-T** is our main proposal, which minimizes (3), and applies the two-step cross

validation process described in Section 3. (2) **cMMD-T** is similar to our main approach but does not incorporate the weights u , and penalizes an MMD penalty that is conditional on the main label Y . Enforcing the conditional independence is consistent with the causal DAG so we expect this baseline to give unbiased estimators as is shown in Veitch et al. [2021]. (3) **L2-S** is the standard DNN trained to minimize the empirical risk, with an L2 penalty on the weights. (4) **wL2-S**: similar to L2-S but also incorporates weighting using \mathbf{u} as defined in (2). (5) **Rand-Aug-S**: a baseline that aims for robustness by augmenting the data at training time using random flips and rotations. (6) **Rex**: a recently suggested method that imposes an invariance penalty without re-weighting [Krueger et al., 2021]. We show the results from Rex in the appendix since it significantly under performs compared to other models.

For L2-S, wL2-S, Rand-Aug-S, and Rex cross-validation is done in the standard way by picking the model that has the best accuracy on the validation set.

In the uncorrelated setting, we only present the unweighted variants of all the models, since the weights are roughly constant across data points in that setting.

Ablations. We study ablated variants of our approach: (1) **wMMD-S** is similar to wMMD-T, but uses standard cross-validation rather than our two-step cross-validation. (2) **MMD-T** minimizes a variant of equation 3 that does not utilize \mathbf{u} in the logistic loss or the MMD penalty. At validation time hyperparameters are picked using our two-step cross-validation approach using \mathbf{u} -weighted metrics. (3) **MMD-S** is similar to MMD-T, but does standard cross-validation. (4) **MMD-uT** is similar to MMD-T, but uses unweighted validation metrics in our two-step cross-validation.

Results: Sampling from the ideal distribution.

Figure 2 (left) shows the results from the first setting, where the training data is sampled from the ideal, uncorrelated distribution P° . The x -axis shows $P(Y=1|V=1) = P(Y=0|V=0)$ at test time, while the y -axis shows the corresponding mean AUROC, averaged over 20 simulations. The vertical dashed line shows the conditional probability at training time. Both variants of our proposed approach, with standard and two-step cross-validation outperform other baselines within the training distribution and when there is distribution shift. This conforms with our theory that suggests that even when the data are sampled from the ideal distribution, using a causally-motivated regularization scheme leads to more efficient models, which translates into better performance in finite samples. Despite being consistent with the causal DAG, cMMD-

¹Our code is available at github.com/mymakar/causally_motivated_shortcut_removal

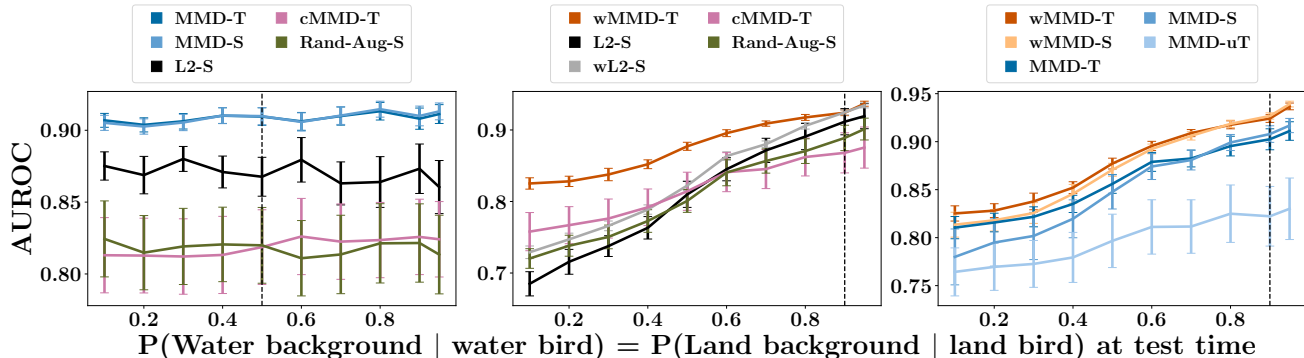


Figure 2: **In all plots** x -axis shows $P(Y|V)$ at test time under different shifted distributions, y -axis shows AUROC on test data, and vertical dashed line shows $P(Y|V)$ at training time. **(Left)** Training data sampled from the ideal (uncorrelated) distribution. MMD-regularized models outperform baselines within, and outside the training distribution. **(Middle)** Training data sampled from a correlated distribution. MMD-regularized model outperform baselines showing better robustness against distribution shifts at test time. **(Right)** Ablation study shows how different components of our suggested approach (wMMD-T) contribute to improved performance.

T requires conditioning on Y to estimate the MMD penalty. We conjecture that the need to slice the data on Y leads to unstable estimates of the conditional MMD when using standard batch sizes. Additional experimental results that support this conjecture are included in the appendix. In addition, all models are robust to distribution shift when trained on data from P° . This conforms with proposition 1, which states that under P° , optimal predictors are risk-invariant.

Results: Sampling from a correlated distribution. Figure 2 (middle) shows the results from the setting where the training data is sampled a correlated distribution with $P_s(Y = 1|V = 1) = P_s(Y = 0|V = 0) = 0.9$. The x , and y axes are similar to figure 2 (left). Our approach (wMMD-T) outperforms other models especially at high divergence from the training distribution. Out of all the non-MMD regularized baselines, the weighted L2-regularized model performs best. This suggests that minimizing the empirical risk on the \mathbf{u} -reweighted distribution contributes to model robustness. Conclusions about cMMD-T are similar to the uncorrelated distribution setting. Studying performance metrics other than the AUROC (e.g., Brier score or logistic loss) yields the same conclusions (see the appendix section F).

The ablation study in Figure 2 (right) shows that the largest increase in performance is attributable to using the weighted MMD penalty at training time, since the two weighted MMD variants outperform the two unweighted variants. Within those two groups, the two-step cross validation with weighted metrics outperforms the others, especially in terms of robustness. This shows that when training models using the MMD penalty, it is important to take into consideration that, unlike L2-norm regularization, the penalty depends on the training data, and is prone to overfitting. Recall that MMD-uT strictly enforces the MMD penalty

without addressing the fact that the training distribution has been sampled from a correlated distribution. This yields a fairly robust predictor that has poor performance. This conforms with our findings stated in the appendix proposition A6, which imply that there will be a bias-robustness trade-off if the correlated sampling is not corrected.

5.2 Chest X-rays

For a less synthetic test, we adapt an experiment from Jabbour et al. [2020] where the task of predicting pneumonia ($Y = 1$) vs. no clinical findings ($Y = 0$) from chest X-rays (\mathbf{X}) considering sex to be a possible shortcut (V).

Data. We conduct this analysis on a publicly available dataset, CheXpert [Irvin et al., 2019]. At training time, we under-sample women who did not have pneumonia to create setting where sex-related attributes are possible shortcuts to predict pneumonia. Specifically, we sample the data such that $P_s(Y = 1) = 0.3$, and the majority of women have pneumonia whereas the majority of men do not have pneumonia i.e., $P(Y = 1|V = 1) = P(Y = 0|V = 0) = 0.9$. For this task, we use DensNet-121 [Huang et al., 2017], pretrained on ImageNet, and fine tuned for our specific task. We use DenseNet because it was shown to outperform other commonly used architectures on the CheXpert dataset [Irvin et al., 2019, Ke et al., 2021, Jabbour et al., 2020]. We present the results from 5 simulations. In each simulation, we generate different train/test splits. Additional details about the training are presented in the appendix.

Results. We evaluate the performance of our approach and baselines similar to those outlined in section 5.1 in the **no shift** setting where the test data is sampled from the same biased distribution as the training data, and a **shift** setting where the test data

Model	AUROC (STE)	
	Shift	No shift
wMMD-T	0.75 (0.006)	0.85 (0.007)
MMD-uT	0.7 (0.013)	0.75 (0.033)
cMMD-T	0.71 (0.024)	0.73 (0.05)
L2-S	0.62 (0.066)	0.71 (0.063)
wL2-S	0.69 (0.019)	0.86 (0.01)
Rex	0.57 (0.036)	0.6 (0.038)

Table 1: CheXpert results show that our approach (wMMD-T) outperforms others when there is a distribution shift and performs comparably to the L2-regularized DNN that utilizes our proposed weighting scheme (wL2-S) in distribution.

comes from a distribution where $P(Y = 1|V = 1) = P(Y = 0|V = 0) = 0.5$. The results in table 1 show that our proposed approach (wMMD-T) outperforms all others when there is a distribution shift and performs comparably in-distribution to the L2-regularized DNN that utilizes our proposed weighting scheme (wL2-S). The performance of MMD-uT, cMMD-T and Rex does not change as much as other models across the two distributions, signaling robustness. However, MMD-uT and Rex under-perform in terms of accuracy, which is expected since they enforce an invariance penalty that is inconsistent with the causal DAG and are subject to a invariance-accuracy trade-off. Meanwhile, cMMD-T is consistent with the causal DAG, but under-performs in practice. This is consistent with our conjecture that the cMMD-T penalty is harder to estimate using small batches.

6 CONNECTIONS TO EXISTING WORK

Shortcut learning. Discouraging shortcut learning using data augmentation (e.g., rotation, translation, cropping) has been suggested by multiple authors [Hendrycks et al., 2020, Yin et al., 2019, Lyle et al., 2020, Lopes et al., 2019, Cubuk et al., 2018]. This approach can work if a generator for shortcut transformations is available at training time. However, if this set of generated transformations does not include interventions that disrupt the shortcut, the desired robustness might not be achieved, as evidenced by the empirical performance of the random augmentation baseline presented in the experiments section. Our approach does not require generating shortcut transformations. Instead, our approach leads to invariant models by leveraging the auxiliary labels to inform the relevant transformations the main label should be independent to. Other work views shortcuts as a consequence of model overparameterization. For example,

Sagawa et al. [2020b] observe that overparameterization exacerbates the reliance on spurious correlations, and suggest an approach similar to wL2-S, which is outperformed by our model.

Fairness and invariance. Our work sheds light on properties of invariant representations, which are used in the fairness literature [Madras et al., 2018, Donini et al., 2018]. One key distinction between our work and fairness literature is that our invariance penalty is designed to be consistent with the causal DAG. This leads to estimators that are asymptotically optimal. By contrast, in the the fairness literature, invariance constraints are often motivated by external criteria, and may be incompatible with the causal DAG. In these cases, a tradeoff can arise between the fairness criterion and accuracy (see proposition A6 in the appendix). Thus our work contributes to ongoing investigations about whether there is a trade-off between fairness and accuracy [Zhang et al., 2019, Calders et al., 2009, Johansson et al., 2019, Dutta et al., 2020, Zhao and Gordon, 2019]. Note that the MMD-uT and the cMMD-T baselines presented in the experiments section correspond to penalties that enforce demographic parity and equalized odds respectively [Hardt et al., 2016].

Our theoretical analysis is related to analysis of generalization in “Fair ERM” appearing in Donini et al. [2018], but has a different focus. Specifically, our analysis (Proposition 4) highlights that an invariance constraint can *improve* efficiency by reducing the complexity of a learning problem through regularization of “orthogonal” dimensions.

Similar invariance penalties have been suggested in causality literature [Shalit et al., 2017, Johansson et al., 2016], and domain shift literature [Tzeng et al., 2014, Long et al., 2015]. While proposition 2 bears some similarity to statements presented in the domain adaptation literature [Long et al., 2015, Ben-David et al., 2007, 2010], our work is distinct in that we do not aim to generalize to a specific target domain. Instead, we aim to build models that generalize across a *family* of target domains. One consequence of this is that, unlike unsupervised domain adaptation, we do not require access to examples from a target domain.

Causally-motivated invariance. In other work, Arjovsky et al. [2019] propose an invariant risk minimization (IRM) approach inspired by ideas from causality. Unlike our approach, IRM does not explicitly penalize dependence on the redundant dimensions, but instead relies on the idea that the invariant risk minimizer should achieve the lowest error across datasets sampled from different target distributions P_t . As others (e.g., Guo et al. [2021], Rosenfeld et al. [2020]) noted,

when the family of functions is as flexible as DNNs, it is possible to find a predictor that achieves the objective of IRM but is not robust. Similar to us, Krueger et al. [2021] suggest a model for risk extrapolation (Rex) in the anti-causal settings. Their method does not correct for biased sampling and hence has the same limitations as the unweighted MMD models presented in the experiments. Rex results presented in the appendix confirm that it consistently performs worse than our approach.

Similar to this work, Subbaswamy and Saria [2018] develop methods to create estimators that are stable across distribution shifts by appealing to causal DAGs and without requiring access to samples from the target distribution. However, they do so by identifying stable features or “conditioning sets” that only contain variables with stable relationships with the target label. Such an approach, which assumes that the input features \mathbf{X} are interpretable, would not be appropriate in a more general setting where the input features are image pixels.

Our work is most similar and complementary to Veitch et al. [2021], where the authors also note that the causal structure of a problem has implications for distributional robustness. Unlike Veitch et al. [2021], we focus on the specific anti-causal setting described in figure 1, and provide a finite sample analysis highlighting the robustness and efficiency of our estimator. We note that our results extend to non anti-causal DAGs, so long as the correlation between Y , and V is purely spurious, which is consistent with the findings of Veitch et al. [2021].

7 DISCUSSION

We presented an approach to build models that are invariant to distribution shifts using auxiliary labels. Guided by our theoretical insight, we suggested a causally-motivated regularization scheme to train robust, and accurate models. We showed that our approach empirically outperforms others.

Limitations. Our approach requires *a priori* knowledge of a shortcut that might be exploited by a predictor. This is to be expected: the choice of which source of variation to exclude from a predictor is highly context-specific and problem-specific. If a predictor exhibit a lack of robustness and shortcut learning is suspected, practitioners could apply exploratory techniques (e.g., interpretability method such as saliency maps Simonyan et al. [2013] or Shapley values Lundberg and Lee [2017], Wang et al. [2021]) to surface factors that the model may rely on. Ultimately, however, this requires judgment about the problem domain. A second limitation is the focus on one binary

auxiliary variable. For multiple/categorical V our approach could be modified in several ways. First, for a categorical V with m categories, the MMD term in equation (4) can be replaced with $m!/2(m-2)!$ MMD terms, each corresponding to a pairwise comparison between two groups defined by V . Alternatively, we would define the invariance penalty with respect to the Hilbert Schmidt independence criterion (HSIC), which allows for higher cardinality [Gretton et al., 2007].

Acknowledgements

We thank the reviewers for their insightful comments. We thank John Guttag and the members of the Clinical and Applied Machine Learning group at MIT for their detailed feedback. Part of this work was done while MM was an intern at Google Research.

References

- Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.
- Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.
- Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.
- Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, pages 7538–7550, 2018.
- Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

- Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.
- Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020a.
- Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124, 2019.
- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.
- David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.
- Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2796–2806, 2018.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- Victor Veitch, Alexander D’Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.
- Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR, 2020.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings*

- of the *Conference on Health, Inference, and Learning*, pages 116–124, 2021.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.
- Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*, 2019.
- Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.
- Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.
- Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.
- Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *arXiv preprint arXiv:2005.04345*, 2020b.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.
- Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.
- Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.
- Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.
- Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pages 2803–2813. PMLR, 2020.
- Han Zhao and Geoffrey J Gordon. Inherent trade-offs in learning fair representations. *arXiv preprint arXiv:1906.08386*, 2019.
- Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
- Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085, 2017.
- Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.
- Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.
- Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.
- Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.
- Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.
- Adarsh Subbaswamy and Suchi Saria. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. *arXiv preprint arXiv:1808.03253*, 2018.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualizing image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

- Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.
- Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2021.
- Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007.
- Michel Ledoux. Isoperimetry and gaussian analysis. In *Lectures on probability theory and statistics*, pages 165–294. Springer, 1996.
- Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In *International Conference on Machine Learning*, pages 6661–6671. PMLR, 2020.
- Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Supplementary Material: Causally Motivated Shortcut Removal Using Auxiliary Labels

A SOCIETAL IMPACT

Our approach could be used in certain fairness applications where invariance to a sensitive label is favorable. We caution that even though our approach outperforms the baselines, it is imperfect in that it still exhibits some dependence on the features related to the sensitive (auxiliary) label as shown in figure 2 (middle). As with most learned models, human audits of the model’s output are necessary in such high-stakes applications.

B PROOFS FOR SECTION 2

To prove proposition 1, we first formally state the assumption of invertability mentioned in section 2. Specifically, the fact that \mathbf{X}^* is a function of \mathbf{X} only implies that \mathbf{X}^* is invertible, i.e. for all \mathbf{X} , \mathbf{X}^* can be exactly recovered from \mathbf{X} . We state that formally in the following assumption

Assumption A1. (Invertability) *There exists some function e such that $\mathbf{X}^* = e(\mathbf{X})$ for all \mathbf{X} .*

Proposition A1 (Restated proposition 1). *Under P° , the Bayes optimal predictor is (i) only a function of \mathbf{X}^* , and (ii) an optimal risk-invariant predictor f_{rinv} with respect to \mathcal{P} .*

Proof. Under P° , \mathbf{X}^* d -separates Y from \mathbf{X} , so $\mathbb{E}_{P^\circ}[Y | \mathbf{X}] = \mathbb{E}_{P^\circ}[Y | \mathbf{X}^*]$. Thus, the population risk minimizer is only a function of \mathbf{X}^* .

By the assumption A1, we have that $\mathbf{X}^* = e(\mathbf{X})$ and hence \mathbf{X}^* can be perfectly recovered from \mathbf{X} . This means that $\mathbb{E}_{P^\circ}[Y | \mathbf{X}^*]$ can be written as a function of \mathbf{X} , i.e., we can define $g(\mathbf{X}) = \mathbb{E}_{P^\circ}[Y | e(\mathbf{X})]$. Thus, the Bayes optimal classifier $f(\mathbf{X})$, which is a function of $\mathbb{E}_{P^\circ}[Y | \mathbf{X}] = \mathbb{E}_{P^\circ}[Y | e(\mathbf{X})]$, can be written (with some abuse of notation) as $f(\mathbf{X}^*)$ (that is, a function that only varies with the value of $\mathbf{X}^* = e(\mathbf{X})$).

Thus, the risk is also invariant. To see that note the following:

$$\begin{aligned} R^\circ(f) &= \int_{\mathbf{X}, Y} \ell(f(\mathbf{X}), Y) P^\circ(\mathbf{X} | \mathbf{X}^*, V) P^\circ(\mathbf{X}^* | Y) P^\circ(Y) P^\circ(V) dY d\mathbf{X} \\ &= \int_{\mathbf{X}^*, Y} \ell(f(\mathbf{X}^*), Y) P^\circ(\mathbf{X}^* | Y) P^\circ(Y) dY d\mathbf{X}^* \\ &= \int_{\mathbf{X}^*, Y} \ell(f(\mathbf{X}^*), Y) P(\mathbf{X}^* | Y) P(Y) dY d\mathbf{X}^* \\ &= R_P(f), \end{aligned}$$

for any $P \in \mathcal{P}$.

Because this classifier is optimal under P° , no other risk invariant classifier can obtain a lower risk across \mathcal{P} ; thus this classifier is an optimal risk invariant classifier. \square

C PROOFS FOR SECTION 3

We show that the reweighted risk is an unbiased estimator of the risk under P° , i.e., that

$$R_{P_s}^{\mathbf{u}}(f) = \mathbb{E}_{P_s} \left[\hat{R}_{P_s}^{\mathbf{u}}(f) \right] = R_{P^\circ}(f).$$

For any P_s , the \mathbf{u} -weighted risk is equal to the risk under the corresponding unconfounded distribution P° . That is, $R_{P_s}^{\mathbf{u}} := \mathbb{E}_{P_s}[u(Y, V)\ell(f(\mathbf{X}, Y))] = R_{P^\circ}$.

To see this, note that the conditional distribution $P_s(\mathbf{X} | Y, V)$ is invariant across the family \mathcal{P} defined in (1). Thus, the risk conditional on Y and V , $R_{P_s|y,v} := \mathbb{E}_{P_s}[\ell(f(\mathbf{X}), Y) | Y = y, V = v]$, does not change with P_s .

$$\begin{aligned} R_{P_s}^u(f) &:= \mathbb{E}_{P_s}[u(Y, V)\ell(f(\mathbf{X}), Y)] = \mathbb{E}_{P_s}[\mathbb{E}_{P_s}[u(Y, V)\ell(f(\mathbf{X}), Y) | Y = y, V = v]] \\ &= \sum_{y,v} P_s(Y = y, V = v)u(y, v)R_{P|y,v} = \sum_{y,v} P_s(Y = y)P_s(V = v)R_{P_s^\circ|y,v} \\ &= \mathbb{E}_{P^\circ}[\mathbb{E}_{P^\circ}[R_{P^\circ|y,v}]] = R_{P^\circ}. \end{aligned}$$

D PROOFS FOR SECTION 4.1

Lemma A1. *For training data $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, and a corresponding learned $f = h(\phi(\mathbf{x}))$ with expected risk $R_{P^\circ}(f)$, suppose that y is ϕ -representable, i.e., that there exists $g(\phi(\mathbf{x})) = y$, and that $g(\phi)\ell(\phi) \in \Omega$. Then for all y :*

$$P(Y = y)[R_{0y}^\circ - R_{1y}^\circ] \leq \tau,$$

where $R_{vy}^\circ := \mathbb{E}_{\mathbf{x} \sim P^\circ}[\ell(f(\mathbf{x}), y) | V = v, Y = y]$

Proof. Without loss of generality, suppose that for $y = 1$,

$$P(Y = y)[R_{0y}^\circ - R_{1y}^\circ] = \tau_1 > \tau.$$

Then due to the fact that $\text{MMD} \leq \tau$, and by assumption that $\ell \in \Omega$

$$\begin{aligned} P(Y = 0)[R_{00}^\circ - R_{10}^\circ] &\leq \tau - \tau_1 \\ P(Y = 0)[R_{00}^\circ - R_{10}^\circ] &< 0 \\ R_{00}^\circ - R_{10}^\circ &< 0. \end{aligned}$$

Using the shorthand $R_{\Delta 0}^\circ := R_{00}^\circ - R_{10}^\circ$, the above inequality implies that $-R_{\Delta 0}^\circ > 0$.

Let $\dot{\ell}(\phi) = (2g(\phi) - 1) \cdot \ell(\phi)$, and $\dot{R}^\circ := \mathbb{E}[\dot{\ell}(\phi)]$. By assumption, we have that $\dot{\ell}$ is also $\in \Omega$. However,

$$\begin{aligned} \text{MMD}(\dot{\ell}, P_{\phi_0}, P_{\phi_1}) &= P(Y = 0)[\dot{R}_{00}^\circ - \dot{R}_{10}^\circ] + P(Y = 1)[\dot{R}_{01}^\circ - \dot{R}_{11}^\circ] \\ &= P(Y = 0)[-(R_{00}^\circ - R_{10}^\circ)] + P(Y = 1)[R_{01}^\circ - R_{11}^\circ] \\ &= P(Y = 0)[-R_{\Delta 0}^\circ] + \tau_1 > \tau. \end{aligned}$$

This contradicts the MMD condition; that for all functions in Ω , $\text{MMD} \leq \tau$

□

Proposition A2 (Restated proposition 2). *Suppose that $f = h(\phi(\mathbf{X}))$ is a predictor that satisfies $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ) \leq \tau$. Suppose that y is ϕ -representable, i.e., that there exists $g(\phi(\mathbf{x})) = y$, and that $g(\phi)\ell(\phi) \in \Omega$. Then*

$$R_{P_t}(f) < R_{P^\circ}(f) + 2\tau.$$

Proof. We will use $P_{v|y}(v) := P(V = v | Y = y)$, $P_y(y) = P(Y = y)$, $P_v(v) = P(V = v)$, and R_{vy}° as defined in lemma A1.

Note that

$$\begin{aligned} R_{P^\circ} &= \sum_y P_y^\circ(y) [P_{v|y}^\circ(0)R_{0,y}^\circ + P_{v|y}^\circ(1)R_{1,y}^\circ] \\ &= \sum_y P_y(y) [P_v^\circ(0)R_{0,y}^\circ + P_v^\circ(1)R_{1,y}^\circ]. \end{aligned}$$

And:

$$R_P = \sum_y P_y(y) [P_{v|y}(0)R_{0,y}^\circ + P_{v|y}(1)R_{1,y}^\circ].$$

Taking the difference between the two:

$$\begin{aligned} R_P - R_{P^\circ} &= \sum_y P_y(y) [(P_{v|y}(0) - P_v^\circ(0))R_{0,y}^\circ + (P_{v|y}(1) - P_v^\circ(1))R_{1,y}^\circ] \\ &= \sum_y P_y(y) [(P_{v|y}(0) - P_v^\circ(0))R_{0,y}^\circ + ((1 - P_{v|y}(0)) - (1 - P_v^\circ(0)))R_{1,y}^\circ] \\ &= \sum_y P_y(y) [(P_{v|y}(0) - P_v^\circ(0))R_{0,y}^\circ - (P_{v|y}(0) - P_v^\circ(0))R_{1,y}^\circ] \\ &= \sum_y P_y(y) [\beta_y R_{0,y}^\circ - \beta_y R_{1,y}^\circ] \\ &= \sum_y \beta_y P_y(y) [R_{0,y}^\circ - R_{1,y}^\circ] \\ &\leq \sum_y \beta_y \tau \\ &= \beta \cdot \tau \end{aligned}$$

where in the fourth equality, we use the shorthand $\beta_y := P_{v|y}(0) - P_v^\circ(0)$, and $-1 < \beta_y < 1$. The first inequality follows from lemma A1, and the last equality follows from setting $\beta = \sum_y \beta_y$. Since $\beta \leq 2$, this completes our proof. \square

E PROOFS FOR SECTION 4.2

Proposition A3. (Restated proposition 3) For $\mathcal{D} \sim P^\circ$, and for any for any \mathcal{F}_{L_2} such that $f_{\text{rinv}} \in \mathcal{F}_{L_2}$, there exists a $\mathcal{F}_{\text{MMD},L_2} \subseteq \mathcal{F}_{L_2}$ such that $f_{\text{rinv}} \in \mathcal{F}_{\text{MMD},L_2}$. And the smallest $\mathcal{F}_{\text{MMD},L_2}$ such that $f_{\text{rinv}} \in \mathcal{F}_{\text{MMD},L_2}$ has $\text{MMD} = 0$.

Proof. We prove the existence of a subset, $\mathcal{F}_{\text{MMD},L_2} \subset \mathcal{F}_{L_2}$ by giving an example of such a subset. Consider

$$\mathcal{F}_{L_2, \text{MMD}} = \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A, \text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ) = 0\},$$

Clearly, $\mathcal{F}_{\text{MMD},L_2} \subset \mathcal{F}_{L_2}$. We will now show that any $f_{\text{rinv}} \in \mathcal{F}_{L_2}$ is also $\in \mathcal{F}_{\text{MMD},L_2}$.

By the definition of f_{rinv} , any $f_{\text{rinv}} \in \mathcal{F}_{L_2}$ must satisfy $f_{\text{rinv}}(\mathbf{x}) \perp v$. Then $T_1(f_{\text{rinv}}(\mathbf{x})) \perp T_2(v)$ for any transformations T_1, T_2 . Taking T_1 to be the inverse of the sigmoid function, σ^{-1} , and T_2 to be the identity transformation, we get that $\sigma^{-1}(f_{\text{rinv}}(\mathbf{x})) = \sigma^{-1}(\sigma(\mathbf{w}^\top \mathbf{x})) = \mathbf{w}^\top \mathbf{x} \perp v$. This implies that $p(\mathbf{w}^\top \mathbf{x} | v = 0) = p(\mathbf{w}^\top \mathbf{x} | v = 1)$, which in turn implies that $\text{MMD}(P_{\phi_0}^\circ, P_{\phi_1}^\circ) = 0$, where $\phi(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. \square

Proposition A4. (Restated proposition 4). Let $f(\mathbf{x}) = \sigma(\phi(\mathbf{x})) = \sigma(\mathbf{w}^\top \mathbf{x})$ be a function contained in $\mathcal{F}_{L_2, \text{MMD}}$. Then,

$$\|\mathbf{w}_\perp\| \leq \frac{\tau}{\|\Delta\|}. \quad (8)$$

Proof. Note that τ must be greater than 0. If not, let ω' be the function that achieves the max difference $\tau' < 0$. Then we can define $\omega'' = -\omega'$, which achieves $\tau'' = -\tau' > 0$, which is a contradiction. This means that for all $\omega \in \Omega$,

$$\tau \geq |\mathbb{E}[\omega(\mathbf{x}_i) | v_i = 0] - \mathbb{E}[\omega(\mathbf{x}_i) | v_i = 1]|$$

Taking $\omega(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$,

$$\tau \geq |\mathbb{E}[\mathbf{w}^\top \mathbf{x}_i | v_i = 0] - \mathbb{E}[\mathbf{w}^\top \mathbf{x}_i | v_i = 1]| = |\mathbf{w}^\top \Delta|.$$

Note that $\|\mathbf{w}_\perp\| = \frac{|\mathbf{w}^\top \Delta|}{\|\Delta\|}$, which completes our proof. \square

Proposition A5. (Restated Proposition 5). Let $\mathbf{x}_\perp := \Pi \mathbf{x}$, $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$. For training data $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, $\sup_{\mathbf{x}_\perp} \|\mathbf{x}_\perp\|_2 \leq B_\perp$, $\sup_{\mathbf{x}_\parallel} \|\mathbf{x}_\parallel\|_2 \leq B_\parallel$,

$$\mathfrak{R}(\mathcal{F}_{L_2}) \leq \frac{A\sqrt{B_\parallel^2 + B_\perp^2}}{\sqrt{n}},$$

and

$$\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2}) \leq \frac{A \cdot B_\parallel + \tau \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}}.$$

Proof. First, we derive the bound on $\mathfrak{R}(\mathcal{F}_{L_2})$

$$\begin{aligned} \mathfrak{R}(\mathcal{F}_{L_2}) &= \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_2 \leq A} \frac{1}{n} \sum_i \epsilon_i \mathbf{w}^\top \mathbf{x}_i \right] \\ &= \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_2 \leq A} \frac{1}{n} \sum_i \epsilon_i \mathbf{w}^\top (\mathbf{x}_{\perp i} + \mathbf{x}_{\parallel i}) \right] \end{aligned}$$

Following the usual derivations (e.g., see Mohri et al. [2018]), we get the desired result for $\mathfrak{R}(\mathcal{F}_{L_2})$. Next, we derive the bound on $\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2})$.

$$\begin{aligned} \mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2}) &= \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_2 \leq A} \frac{1}{n} \sum_i \epsilon_i \mathbf{w}^\top \mathbf{x}_i \right] \\ &= \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_2 \leq A} \frac{1}{n} \sum_i \epsilon_i (\Pi \mathbf{w}^\top \mathbf{x}_i + (1 - \Pi) \mathbf{w}^\top \mathbf{x}_i) \right] \\ &\leq \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\substack{\mathbf{w}_\parallel: \|\mathbf{w}_\parallel\|_2 \leq A \\ \mathbf{w}_\perp: \|\mathbf{w}_\perp\|_2 \leq A}} \frac{1}{n} \sum_i \epsilon_i \mathbf{w}_\perp^\top \mathbf{x}_{\perp i} + \epsilon_i \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \right] \\ &\leq \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}_\perp: \|\mathbf{w}_\perp\|_2 \leq A} \frac{1}{n} \sum_i \epsilon_i \mathbf{w}_\perp^\top \mathbf{x}_{\perp i} \right] + \mathbb{E}_{\mathcal{D}} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}_\parallel: \|\mathbf{w}_\parallel\|_2 \leq A} \frac{1}{n} \sum_i \epsilon_i \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \right], \end{aligned}$$

where the last inequality follows by the subadditivity of the supremum. Again, following the usual derivations (e.g., see Mohri et al. [2018]), we get the required result for $\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2})$ \square

In the following proposition we show that when sampling from a biased distribution, the smallest τ' that does not introduce bias is greater than 0.

Proposition A6. Let $\mathcal{F}'_{L_2, \text{MMD}} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A, \text{MMD}(P_{\phi_0}, P_{\phi_1}) \leq \tau'\}$ be the smallest function class that contains f_{rinv} . Then $\tau' = c \cdot A$ for some $c > 0$, and the corresponding upper bound on the Rademacher complexity of $\mathcal{F}'_{\text{MMD}, L_2}$ is

$$\mathfrak{R}(\mathcal{F}'_{\text{MMD}, L_2}) \leq \frac{A \cdot B_\parallel + c \cdot A \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}}.$$

Proof. By proposition 3, we have that the smallest MMD regularized function class that contains f_{rinv} when $\mathcal{D} \sim P^\circ$ has MMD = 0. And by proposition 4 we have in that function class $\|\mathbf{w}_\perp\| = 0$, i.e., \mathbf{w}_\perp is the 0 vector.

$$\begin{aligned}
 \tau' &\geq \left\| \mathbb{E}[\mathbf{w}^\top \mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{w}^\top \mathbf{x}_i \mid v_i = 1] \right\| \\
 &= \left\| \mathbb{E}[\mathbf{w}_\perp^\top \mathbf{x}_{\perp i} + \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 0] - \mathbb{E}[\mathbf{w}_\perp^\top \mathbf{x}_{\perp i} + \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 1] \right\| \\
 &= \left\| \mathbb{E}[\mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 0] - \mathbb{E}[\mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 1] \right\| \\
 &= \left\| \mathbf{w}_\parallel (\mathbb{E}[\mathbf{x}_{\parallel i} \mid v_i = 0] - \mathbb{E}[\mathbf{x}_{\parallel i} \mid v_i = 1]) \right\| \\
 &= \left\| \mathbf{w}_\parallel (1 - \Pi) (\mathbb{E}[\mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{x}_i \mid v_i = 1]) \right\| \\
 &= \|\mathbf{w}_\parallel\| \left\| (1 - \Pi) (\mathbb{E}[\mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{x}_i \mid v_i = 1]) \right\| \\
 &= A \|(1 - \Pi)\Delta_P\|,
 \end{aligned}$$

where the fifth equality holds because the two vectors are scalar multiples of the same vector (they are both projections onto the vector orthogonal to Δ) so Cauchy-Schwartz holds with equality. Also note that $\|(1 - \Pi)\Delta_P\| = 0$ if and only if $\Delta_P = \Delta$, i.e., $P = P^\circ$. So $\|(1 - \Pi)\Delta_P\| > 0$.

The upper bound on the Rademacher complexity follows immediately by plugging the upper bound on τ' into $\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2})$ obtained in proposition 5. \square

Notes on proposition A6

1. One important implication of proposition A6 is that naively implementing the MMD penalty when the data are sampled from some $P_s \neq P^\circ$ will lead to a bias-invariance tradeoff. For any $\tau' < c \cdot A$, invariance comes at the cost of bias. To understand why, note that when $\tau' < c \cdot A$, the optimal function does not exist in the candidate function class. Recall that when sampling from P° , we could allow τ to go to zero without introducing bias.
2. Comparing the upper bound on $\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2})$ in proposition 5 to $\mathfrak{R}(\mathcal{F}'_{\text{MMD}, L_2})$ in proposition A6 shows that the statistical complexity increases when sampling from $P_s \neq P^\circ$. Consequently, the generalization error is higher (i.e., less favorable) when sampling from a non-ideal, correlated distribution.

E.1 Full generalization error statements

We start with the simpler case, where $\mathcal{D} \sim P^\circ$.

Proposition A7. *For a dataset $\mathcal{D} \sim P^\circ$, and loss bounded above by $M > 0$, the finite-sample gap*

$$R_{P^\circ}(f) - \hat{R}_{P^\circ}(f) \leq \frac{A \cdot B_\parallel + \tau \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

Proof. The proof is a straightforward application of theorem 11.3 in Mohri et al. [2018]. The result can be immediately acquired by substituting μ in theorem 11.3 with 1 (since the logistic loss is 1-Lipschitz), and $\mathfrak{R}_m(\mathcal{H})$ with $\mathfrak{R}(\mathcal{F}_{\text{MMD}, L_2})$ from proposition 5. \square

To get a similar statement for the case where $\mathcal{D} \sim P_s \neq P^\circ$, and reweighting is needed, we need to apply the techniques for estimating the generalization error of reweighted estimators presented in Cortes et al. [2010]. To apply the Cortes results, we need to construct a discretization or a covering of $\mathcal{F}_{L_2, \text{MMD}}$, defined next.

Definition A1. *Given any function class \mathcal{F} , a metric D on the elements of \mathcal{F} , and $\varepsilon > 0$, we define a covering number $\mathcal{N}(\mathcal{F}, D, \varepsilon)$ as the minimal number m of functions $f_1, f_2, \dots, f_m \in \mathcal{F}$, such that for all $f \in \mathcal{F}$, $\min_{i=1, \dots, m} D(f_i, f) \leq \varepsilon$, with*

$$D(f, f') = \sqrt{\frac{1}{n} \sum_i (f(\mathbf{x}_i) - f'(\mathbf{x}_i))^2}.$$

Our statement also makes use of Gaussian complexities, defined next.

Definition A2. For a function family \mathcal{F} , the empirical Gaussian complexity is defined as:

$$\mathfrak{G}(\mathcal{F}) = \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\eta} \left[\sup_{f \in \mathcal{F}} \eta_i f(\mathbf{x}_i) \right]$$

We are now ready to present the metric entropy of the discretized hypothesis space in this next lemma.

Lemma A2. Let $\mathbf{x}_{\perp} := \Pi \mathbf{x}$, $\mathbf{x}_{\parallel} := (I - \Pi) \mathbf{x}$, $\sup_{\mathbf{x}_{\perp}} \|\mathbf{x}_{\perp}\|_2 \leq B_{\perp}$, $\sup_{\mathbf{x}_{\parallel}} \|\mathbf{x}_{\parallel}\|_2 \leq B_{\parallel}$, D, ε as is defined in A1. For $\varepsilon, c', c'' > 0$:

$$\begin{aligned} & \log(\mathcal{N}(\mathcal{F}_{L_2, \text{MMD}}, D, \varepsilon)) \\ & \leq c'' \left(\frac{c' \sqrt{\log(n)} \cdot \left(A \cdot B_{\parallel} + \tau \frac{B_{\perp}}{\|\Delta\|} \right)}{\varepsilon} \right)^2 \end{aligned}$$

Proof. We construct our argument relying on Sudakov's minoration, and the bound between Gaussian and Rademacher complexities. Specifically, by Ledoux [1996], for some $c' > 0$:

$$\begin{aligned} \mathfrak{G}_m(\mathcal{F}_{L_2, \text{MMD}}) & \leq c' \sqrt{\log(n)} \cdot \mathfrak{R}(\mathcal{F}_{L_2, \text{MMD}}) \\ & \leq c' \sqrt{\log(n)} \cdot \frac{A \cdot B_{\parallel} + \tau \frac{B_{\perp}}{\|\Delta\|}}{\sqrt{n}}, \end{aligned}$$

where the last inequality follows from plugging in the results from proposition 5. By Sudakov's minoration (see Ledoux [1996] theorem 3.18), for some universal constant $c'' > 0$,

$$\begin{aligned} \log(\mathcal{N}(\mathcal{F}_{L_2, \text{MMD}}, D, \varepsilon)) & \leq c'' \left(\frac{\sqrt{n} \cdot \mathfrak{G}_m(\mathcal{F}_{L_2, \text{MMD}})}{\varepsilon} \right)^2 \\ & \leq c'' \left(\frac{c' \sqrt{\log(n)} \cdot \left(A \cdot B_{\parallel} + \tau \frac{B_{\perp}}{\|\Delta\|} \right)}{\varepsilon} \right)^2 \end{aligned}$$

□

Finally, to use the Cortes et al. [2010] results, we need bounded divergence between the source and the target distribution. Recall that we are trying to bound the finite-sample gap, so the source distribution is P_s , and the target distribution is P° . We can bound this divergence because of the overlap assumption stated in section 2. As a consequence of the overlap assumption, we have that:

$$\sup u(y, v) = \sup \frac{P^{\circ}(Y | V)}{P_s(Y | V)} = 2^{\Xi_{\infty}(P^{\circ} || P_s)} = C_{P_s}, \quad (9)$$

where $\Xi_k(p || q)$ is the k^{th} -order Rényi divergence, and the second equality follows by applying the Bayes rule, and the definition of the Rényi divergence. It will be convenient to denote $2^{\Xi_k(p || q)}$ by $\Lambda_k(p || q)$. Since $2^{\Xi_{k-1}(P^{\circ} || P_s)} < 2^{\Xi_k(P^{\circ} || P_s)}$, we have $\Lambda_2(P^{\circ} || P_s) < C_{P_s}$. Following similar work (e.g., Makar et al. [2020]), we will assume that the weights \mathbf{u} are known, or can be perfectly estimated from the data. In other words, we do not consider estimation error that might arise because of poor estimation of \mathbf{u} . Work by Foster and Syrgkanis [2019] has shown that under mild assumptions, the error due to estimation of \mathbf{u} from finite samples only results in a fourth order dependence in the final classifier, and hence does not greatly affect our derived generalization bounds.

With that, we are ready to state the finite-gap bound when $P_s \neq P^{\circ}$, and reweighting is used.

Proposition A8. For $\mathcal{D} \sim P$, with $P \in \mathcal{P}$, and \mathbf{u} as defined in equation 2, C_P as defined in equation 9,

$$R_{P^{\circ}}(f) - \hat{R}_P^{\mathbf{u}}(f) \leq \frac{2C_P(\kappa(\mathcal{F}_{\text{MMD}, L_2}) + \log \frac{1}{\delta})}{2n} + \sqrt{\frac{\Lambda(P^{\circ} || P) \cdot (\kappa(\mathcal{F}) + \log \frac{1}{\delta})}{n}},$$

where

$$\kappa(\mathcal{F}_{\text{MMD}, L_2}) = c'' \left(\frac{c' \sqrt{\log(n)} \cdot (A \cdot B_{\parallel} + \tau \frac{B_{\perp}}{\|\Delta\|})}{\varepsilon} \right)^2$$

Proof. Using the bound on the metric entropy derived in lemma A2, the proof becomes a direct application of Theorem 2 in Cortes et al. [2010]. \square

F EXPERIMENTS

F.1 Waterbirds: example images

Figure 3 shows examples of the generated images.

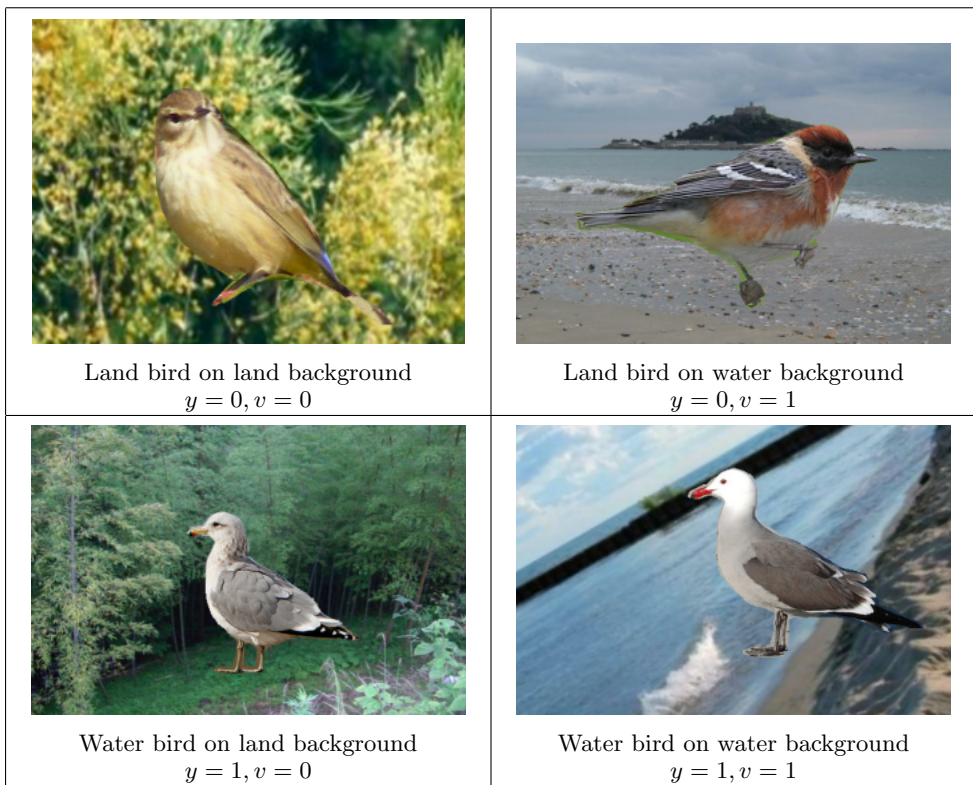


Figure 3: Examples of the generated images of water, and land birds on water, and land backgrounds

F.2 Hyperparameter setting

We train all models using Adam Kingma and Ba [2015], with learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

We train all models for 200 epochs in the waterbirds setting, and for 10 epochs in the CheXpert setting. For L_2 regularized models, we cross-validate the L_2 penalty parameter from the following values [0.0 (no regularization), 0.001, 0.0001], which is similar to values typically used for this setting [Sagawa et al., 2019, He et al., 2016]. For the MMD regularized models, we pick the optimal MMD penalty parameter, α from $[1e^3, 1e^5, 1e^7]$. We also pick the best RBF kernel bandwidth from $[1e1, 1e2, 1e3]$.

For our two-step cross-validation, we determine that a model has an MMD that is statistically insignificantly different from 0 by using a standard T-test. Specifically, we compare the mean MMD across 5 folds of validation data to 0. If the P-value of that T-test is greater than 0.05, we say that the model has an MMD that is statistically insignificantly different from 0.

F.3 Waterbirds: Noisy background

We found that the original background images frequently contain landscapes that are difficult to distinguish (e.g., water backgrounds with very small water bodies that mostly reflect the surrounding trees). To address that, we pick 200 and 300 “clean” images for each of the land and water backgrounds respectively. Using those clean images, we generate 10,000 land backgrounds, and 9,000 water backgrounds by applying random transformations (rotation, zoom, darkening/brightening) to the selected images. Figure 4 shows examples of the excluded backgrounds. To protect the privacy of the individuals in the pictures, their faces have been blurred.



Figure 4: Examples of excluded noisy backgrounds. Pictures extracted from the Places dataset Zhou et al. [2017]. Blurring is added to preserve privacy.

Here we present the results using the full (noisy) background images. Results from the main analysis largely hold, with two distinctions. First, in the ideal distribution, because the backgrounds are noisy, we see an overall higher variance in performance, so the models perform equally well with no clear “winner”. Second, we note that the cMMD-T is particularly sensitive to the level of noise in the backgrounds, while other models are overall more robust.

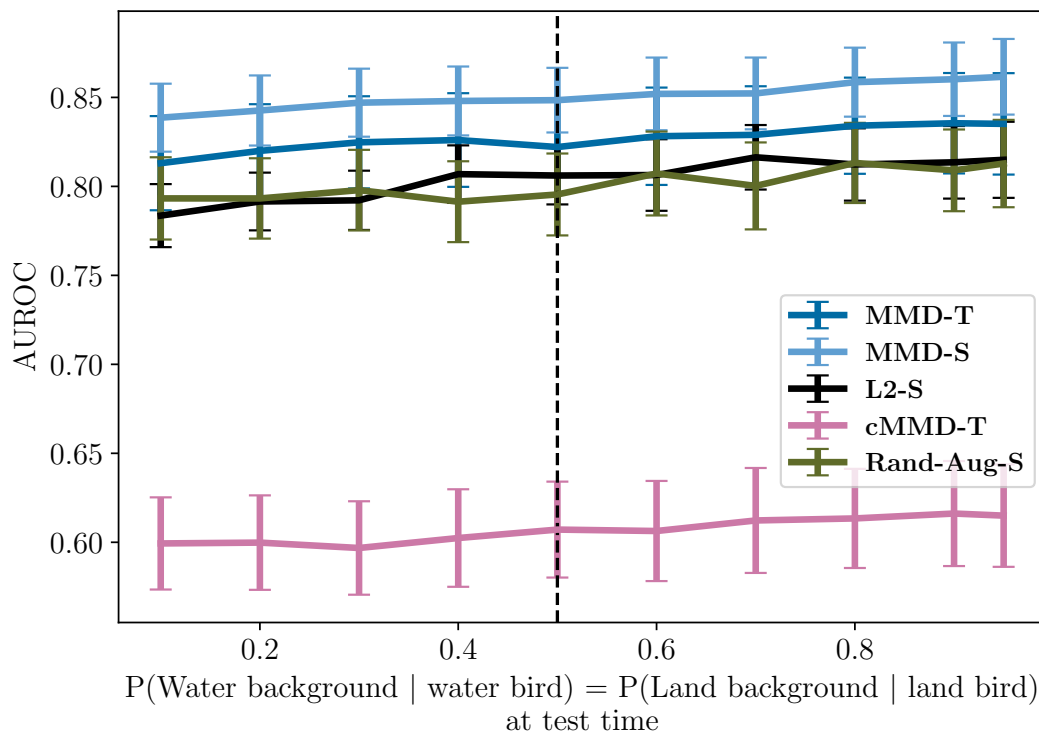


Figure 5: Training data sampled from P° , with $P^\circ(Y|V=1) = P^\circ(Y|V=0) = 0.5$ and backgrounds are sampled from a noisy set of images. x -axis shows $P(Y|V)$ at test time under different shifted distributions. y -axis shows AUROC on test data. Vertical dashed line shows training data. MMD-regularized models outperform baselines within, and outside the training distribution.

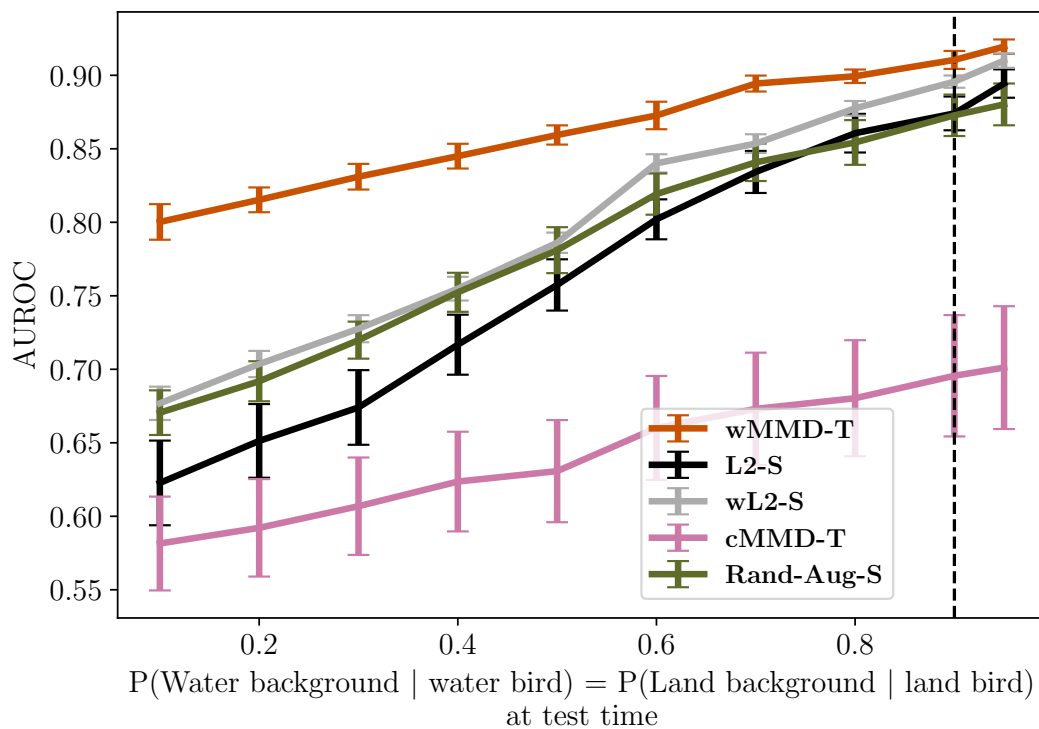


Figure 6: Training data sampled from P , with $P(Y=1|V=1) = P^\circ(Y=0|V=0) = 0.9$, and backgrounds are sampled from a noisy set of images. Vertical dashed line shows training data. x , y axes similar to figure 5. MMD-regularized models outperform baselines showing better robustness against distribution shifts at test time.

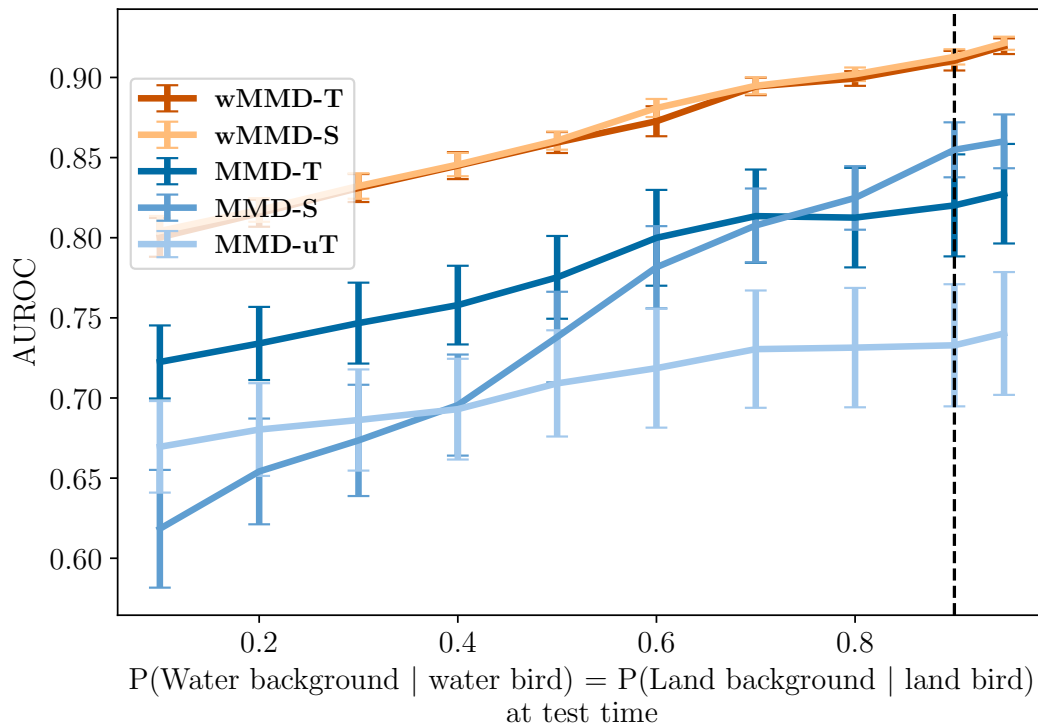


Figure 7: Training data sampled from P , with $P(Y = 1|V = 1) = P^\circ(Y = 0|V = 0) = 0.9$. x , and backgrounds are sampled from a noisy set of images. y axes similar to fig 5. An ablation study to show how different components of our suggested approach (wMMD-reg-T) contribute to improved performance.

F.4 Waterbirds: Other performance metrics

We re-examine the results from figure 2 (middle) considering performance criteria other than the AUROC. Specifically, in figure 8, we consider the logistic loss on the y -axis whereas in figure 9 we consider the Brier score on the y -axis. In both cases, lower is better. The results from both plots conform with our findings from figure 2 (middle).

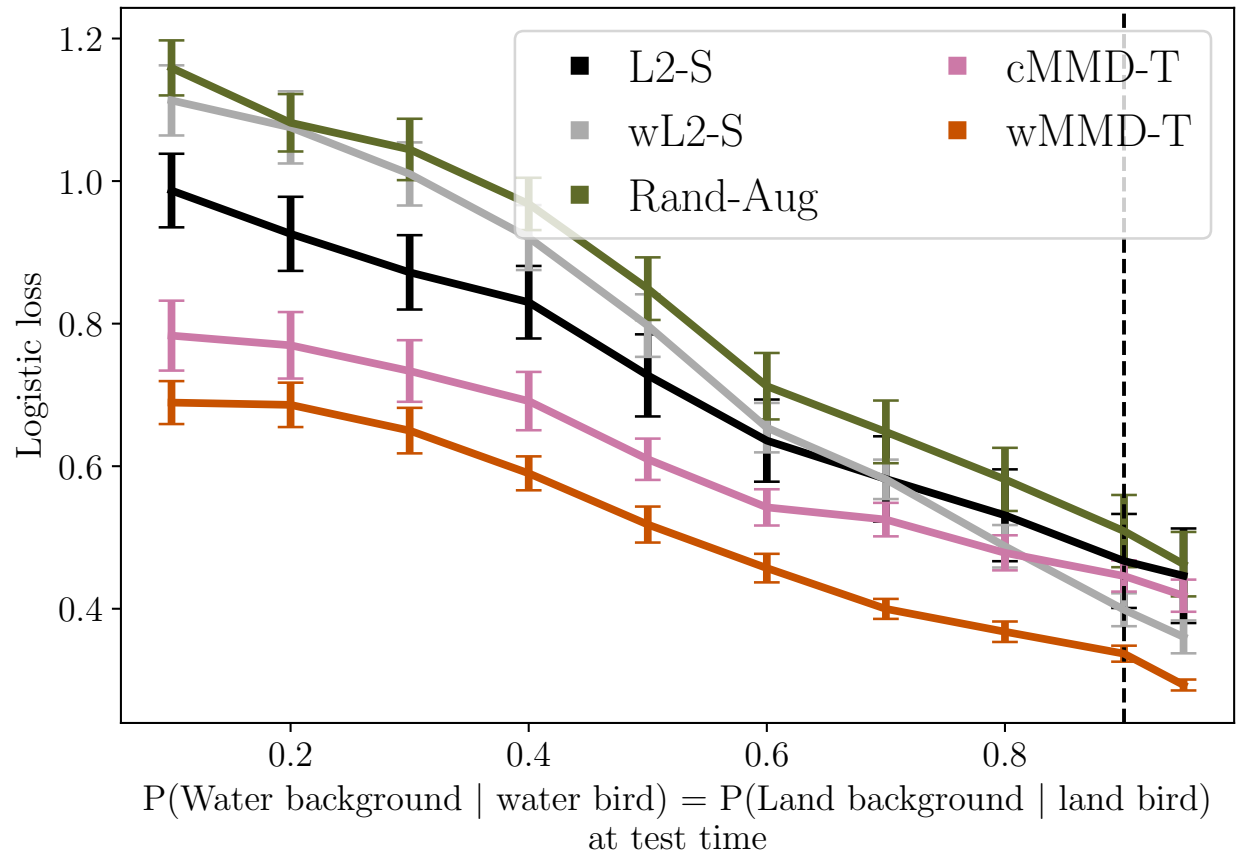


Figure 8: x -axis shows $P(Y|V)$ at test time under different shifted distributions, y -axis shows logistic loss on test data, and vertical dashed line shows $P(Y|V)$ at training time.

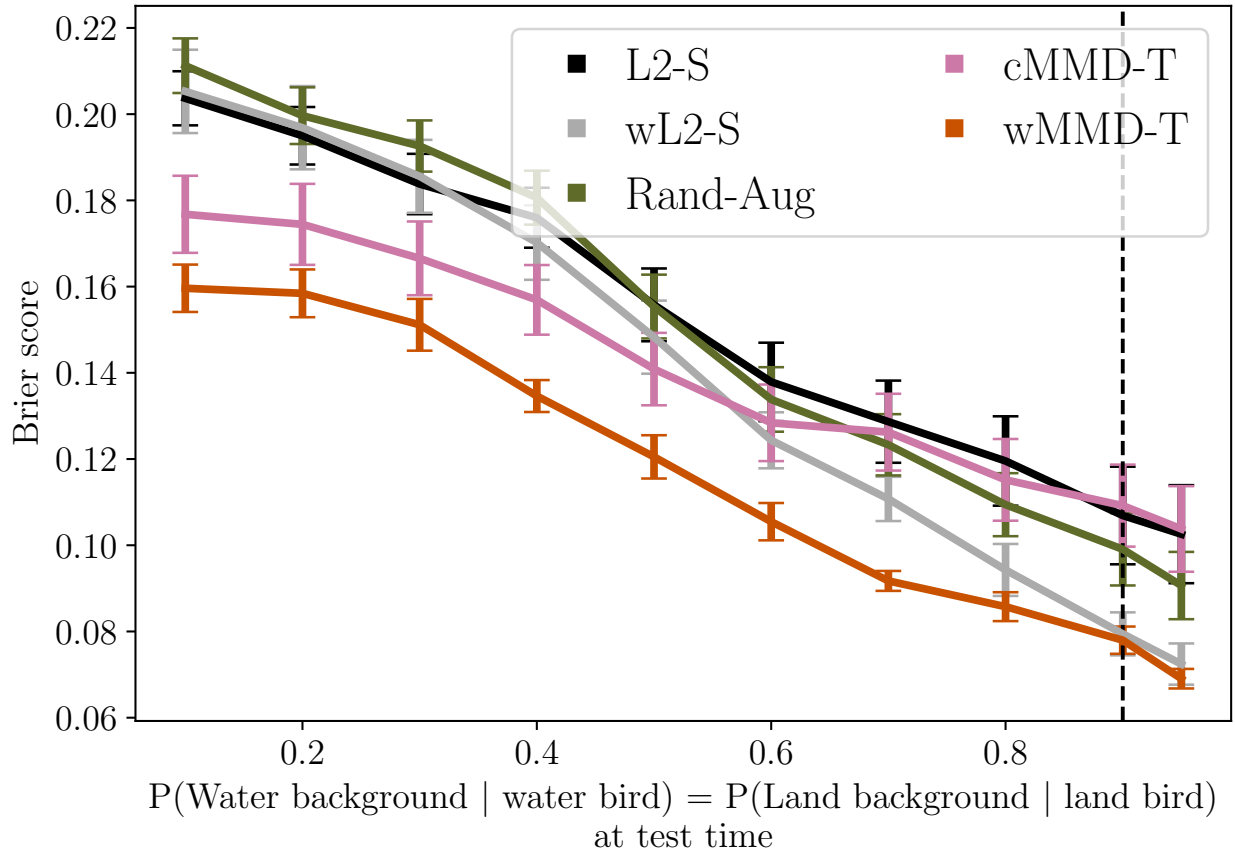


Figure 9: x -axis shows $P(Y|V)$ at test time under different shifted distributions, y -axis shows the Brier score on test data, and vertical dashed line shows $P(Y|V)$ at training time.

F.5 Waterbirds: Rex results

In this section we show the waterbirds results including Rex, which was excluded from the main paper because it significantly underperforms compared to other models. The poor performance of Rex is likely due to issues. First, it does not incorporate the weighting scheme and hence the invariance penalty is inconsistent with the causal DAG. Second, computing the objective function of Rex requires slicing the data into four different groups. Hence, it inherits the same unstable training dynamics that cMMD-T has (see section F.6).

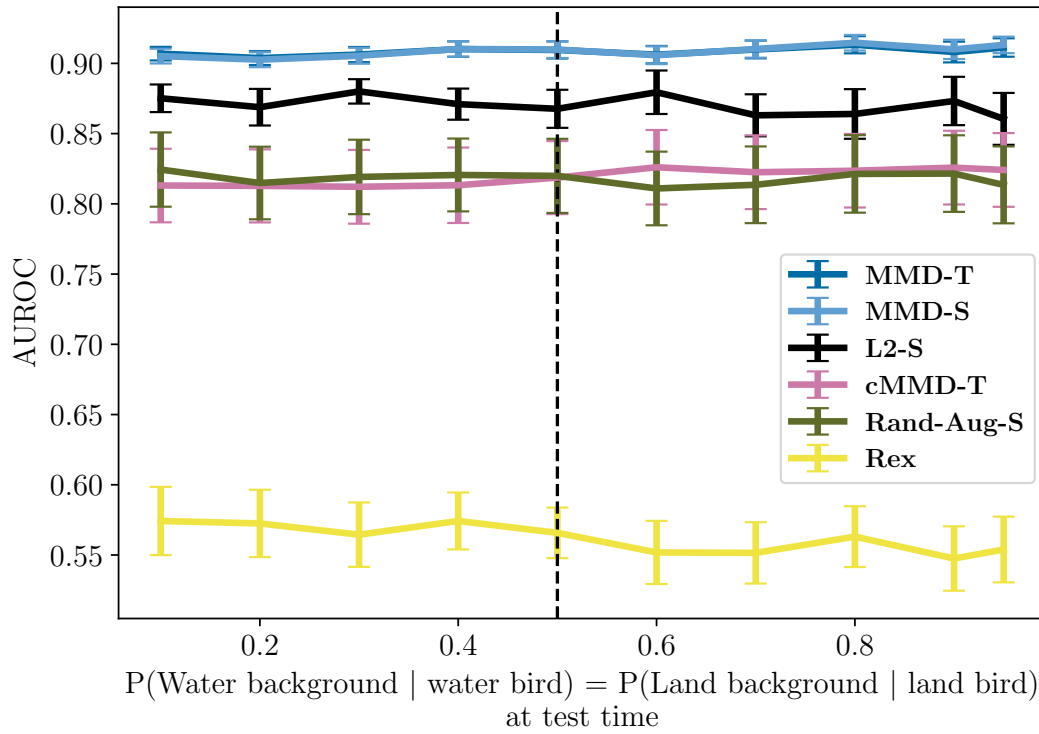


Figure 10: Training data sampled from P , with $P(Y = 1|V = 1) = P^\circ(Y = 0|V = 0) = 0.9$. Setting similar to figure 2 (left) now including results from Rex

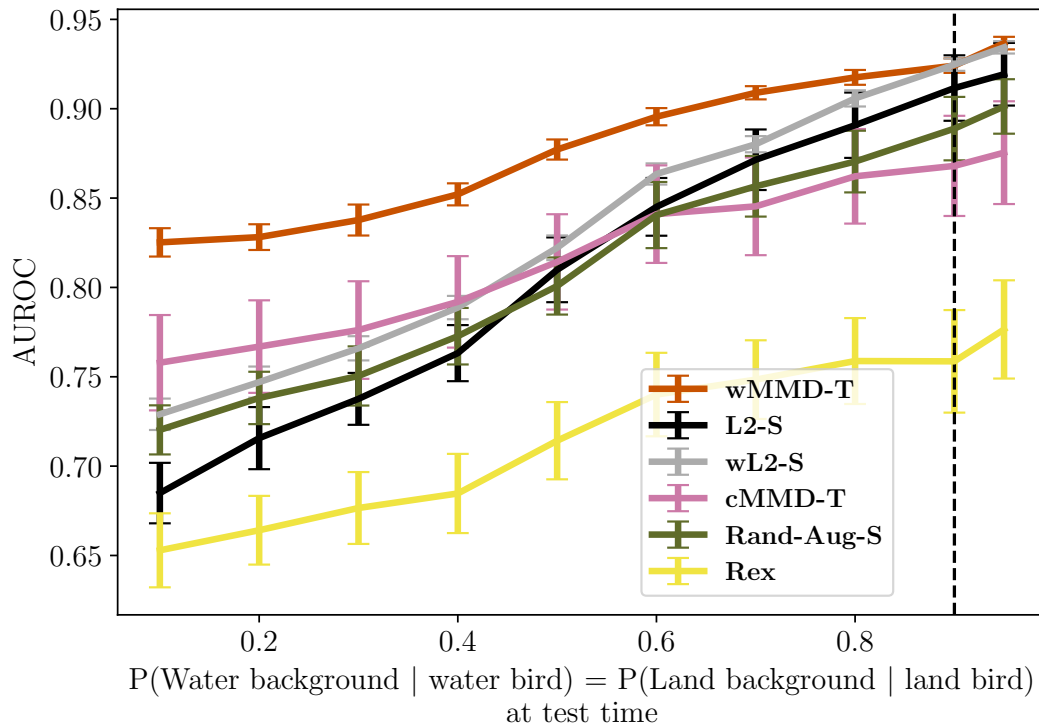


Figure 11: Training data sampled from P , with $P(Y = 1|V = 1) = P^\circ(Y = 0|V = 0) = 0.9$. x . Setting similar to figure 2 (middle, right) now including results from Rex

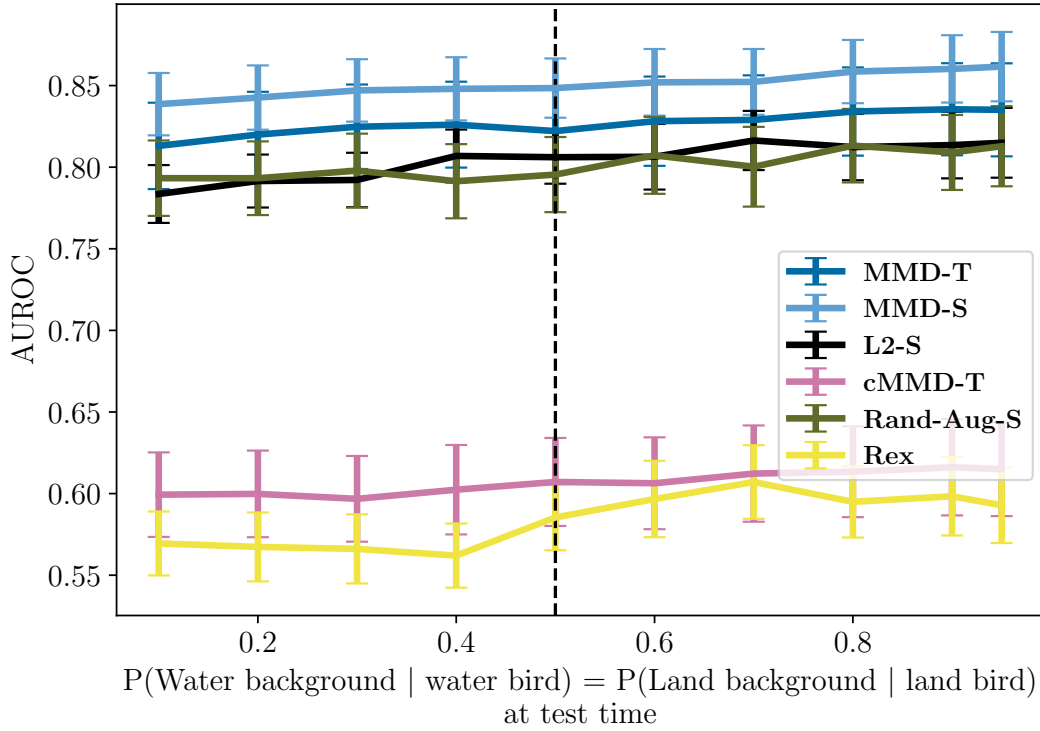


Figure 12: Training data sampled from P , with $P(Y = 1|V = 1) = P^\circ(Y = 0|V = 0) = 0.9$, and backgrounds are sampled from a noisy set of images. Results now show the performance of Rex. Setting similar to 5, now showing the performance of Rex.

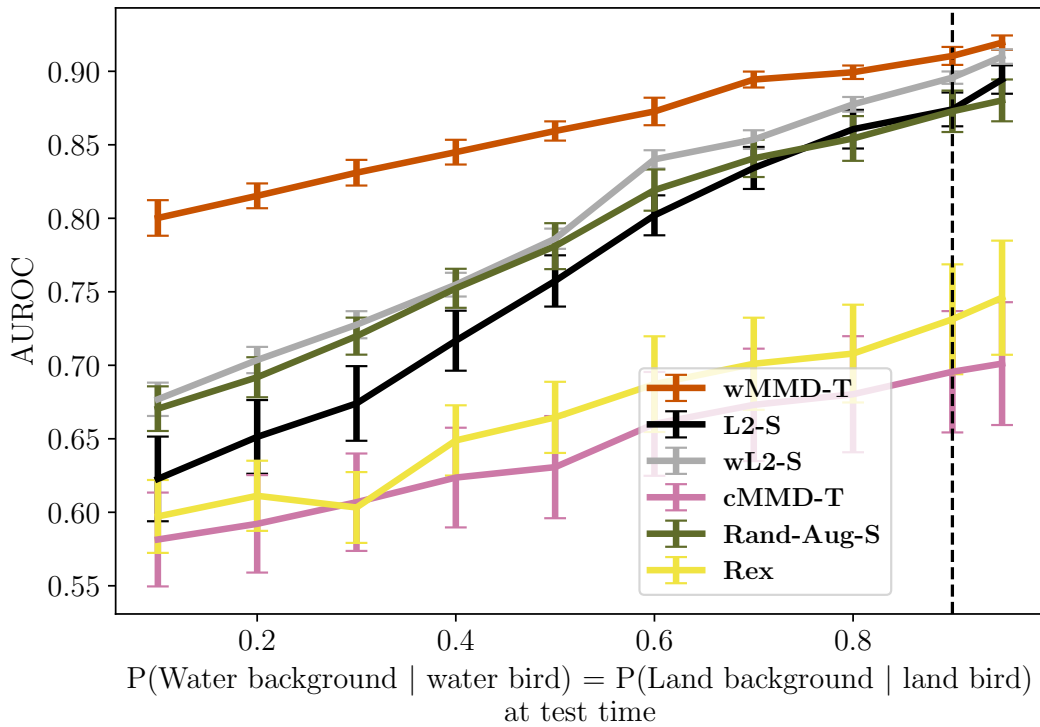


Figure 13: Training data sampled from P , with $P(Y = 1|V = 1) = P^\circ(Y = 0|V = 0) = 0.9$. x , and backgrounds are sampled from a noisy set of images. Results now show the performance of Rex. Setting similar to 6, now showing the performance of Rex.

F.6 Training dynamics of cMMD-T

Note that the cMMD-T objective is

$$h_{EO}, \phi_{EO} = \operatorname{argmin}_{h, \phi} \frac{1}{n} \sum_i \ell(h(\phi(\mathbf{x}_i)), y_i) + \alpha \cdot [\widehat{\text{MMD}}^2(P_{\phi_{0,0}}, P_{\phi_{1,0}}) + \widehat{\text{MMD}}^2(P_{\phi_{0,1}}, P_{\phi_{1,1}})], \quad (10)$$

where $P_{\phi_{v,y}} = P(\phi(\mathbf{X})|V = v, Y = y)$. This differs from the objective of our main approach which does not compute the MMD penalty conditional on Y . This means that the cMMD-T objective requires “slicing” the data into smaller subgroups and computing this MMD on those smaller subgroups leading to less stable training especially in the context of DNNs, where minibatched SGD-based training is standard. For example, if the training is being done on a batch size of 16, with $P_s(Y = 1) = 0.1$, the second MMD term in equation (10) will be computed over a sample size of roughly 2 making it a likely unreliable estimate.

Figure 14 compares the performance of our approach and cMMD-T as the batch size increases in terms of accuracy, whereas figure 15 considers the AUROC. The accuracy plot shows that, for example, at a batch size of 16, our approach vastly outperforms cMMD-T. The AUROC plot shows that the cMMD-T has higher variance at every batch size.

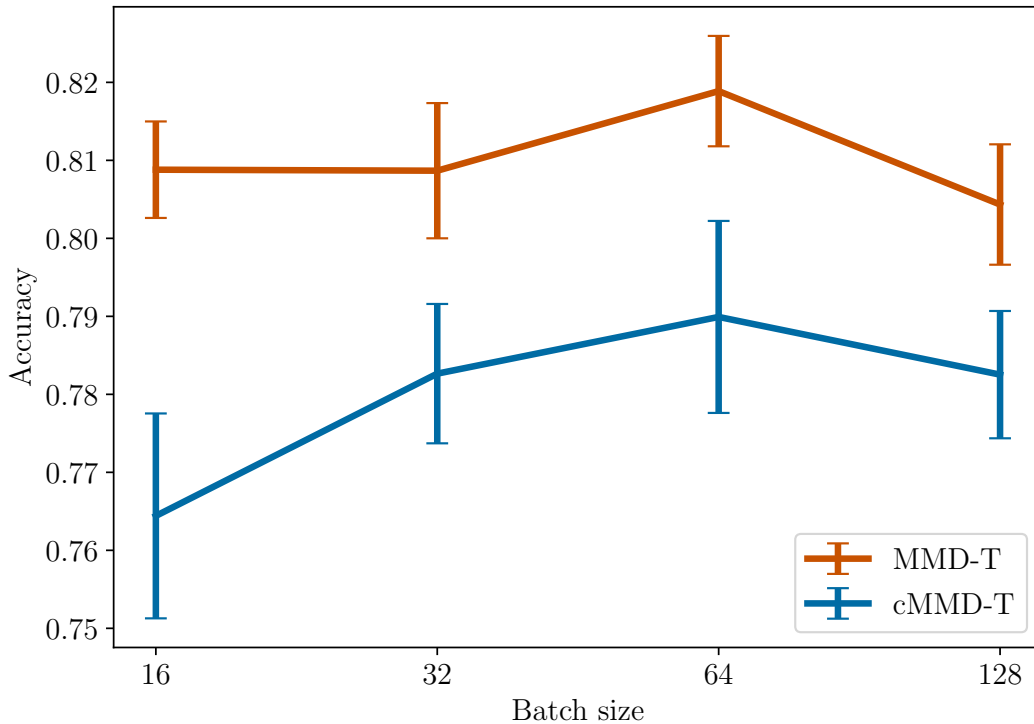


Figure 14: Waterbirds data: x -axis shows the batch size, y -axis shows the accuracy. cMMD-T has poor performance in small batch sizes due to sample splitting.

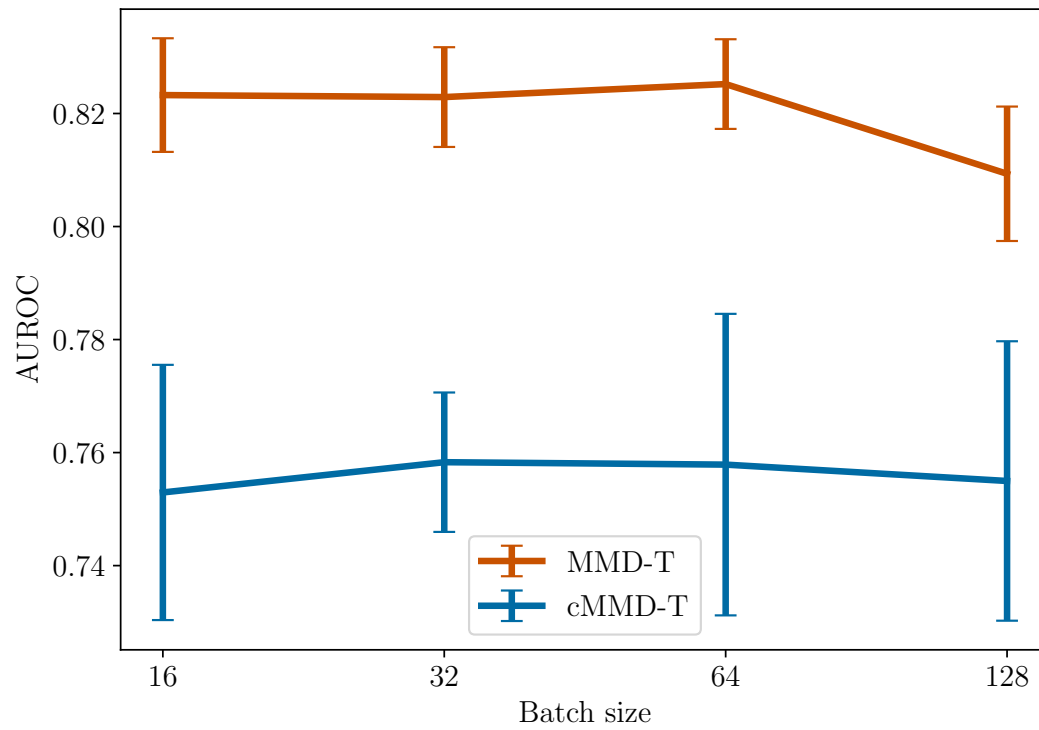


Figure 15: Waterbirds data: x -axis shows the batch size, y -axis shows the AUROC. cMMD-T has an overall higher variance due to sample splitting.