MMD-B-Fair: Learning Fair Representations with Statistical Testing

Namrata Deka
dnamrata@cs.ubc.ca
University of British Columbia

Danica J. Sutherland
dsuth@cs.ubc.ca
University of British Columbia & Amii

Abstract

We introduce a method, MMD-B-Fair, to learn fair representations of data via kernel two-sample testing. We find neural features of our data where a maximum mean discrepancy (MMD) test cannot distinguish between different values of sensitive attributes, while preserving information about the target. Minimizing the power of an MMD test is more difficult than maximizing it (as done in previous work), because the test threshold's complex behavior cannot be simply ignored. Our method exploits the simple asymptotics of block testing schemes to efficiently find fair representations without requiring the complex adversarial optimization or generative modelling schemes widely used by existing work on fair representation learning. We evaluate our approach on various datasets, showing its ability to "hide" information about sensitive attributes, and its effectiveness in downstream transfer tasks.

1 INTRODUCTION

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; B. 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; Madras et al. 2018; Zhang 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 (Bounliphone et al. 2016) to give an effective criterion for driving down the test power of sensitive tests – a problem not handled well by previous work which focused 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 2, establishing all the tools we will need for our method for learning fair kernels and representations (Section 3), and emphasizing aspects important to our approach.

2 PRELIMINARIES

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; we will describe our approach's relationship to that scheme shortly.

2.1 Maximum Mean Discrepancy (MMD)

The MMD (Gretton et al. 2012) is a measure of the distance between distributions. For distributions $\mathbb{P}$ and $\mathbb{Q}$ over a domain $\mathcal{X}$ (the set of conceivable data points), the MMD is defined in terms of a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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]

$$\widehat{\mathrm{MMD}}^2_{\mathrm{U}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) = \frac{1}{m(m-1)} \sum_{i \neq j} H_{ij} \qquad (1)$$

$$H_{ij} = k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(Y_i, X_j),$$

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 (1) is to choose some kernel k a priori, then reject the null hypothesis $\mathfrak{H}_0$ that $\mathbb{P} = \mathbb{Q}$ if the scaled estimator $m\,\widehat{\mathrm{MMD}}^2_{\mathrm{U}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)$ is larger than a threshold c_α. c_α should have $\Pr_{\mathfrak{H}_0}\left(m\,\widehat{\mathrm{MMD}}^2_{\mathrm{U}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) > c_\alpha\right) \leq \alpha$, i.e. we have α probability of incorrectly rejecting $\mathfrak{H}_0$ when it is true. The estimate is scaled by m because, as m grows, this choice makes $m\,\widehat{\mathrm{MMD}}^2_{\mathrm{U}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)$ converge in distribution to a fixed distribution: an infinite mixture of χ^2 variables, with weights depending on $\mathbb{P} = \mathbb{Q}$ and k, but independent of m. We can then estimate the threshold c_α, the $(1 - \alpha)$th quantile of that distribution, 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, compute $m\,\widehat{\mathrm{MMD}}^2_{\mathrm{U}}$, 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 divides 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 $\widehat{\mathrm{MMD}}^2_{\mathrm{U}}$ on each block separately; each of those terms will be an independent unbiased estimator of the squared MMD, so we then average them, obtaining the estimator $\widehat{\mathrm{MMD}}^2_{\mathrm{B}}$.

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

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

with V_B the variance of $\widehat{\mathrm{MMD}}^2_{\mathrm{U}}$ on samples of size B (depending on $\mathbb{P}$, $\mathbb{Q}$, and k). A block test, then, can take as its test statistic $\sqrt{b}\,\widehat{\mathrm{MMD}}^2_{\mathrm{B}}$ 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 standard deviation of the b batches, 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); our primary tests will be based on U-statistics.

2.2 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 significant power. To try 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

[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.

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 lengthscale σ_ϕ is also contained in ω. These kernels have seen success across a variety of areas (e.g. A. G. Wilson et al. 2016; C.-L. Li et al. 2017; Jean et al. 2018; Y. 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

$$\Pr_{\mathfrak{H}_1}\left(m\,\widehat{\mathrm{MMD}}_{\mathrm{U}}^2 > c_\alpha\right) \to \Phi\left(\frac{\mathrm{MMD}^2 - c_\alpha/m}{\sqrt{V_m}}\right), \quad (3)$$

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) and 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) 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), 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) 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), 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 minibatch of size $n = 32$.

To roughly maximize (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

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

using H_{ij} from (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'll 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 Appendix A 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); 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.

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 (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 LEARNING FAIR REPRESENTATIONS

Let $\mathbb{P}^a$ and $\mathbb{Q}^a$ be conditional distributions on a dataset that only differ by the value of the binary feature a on which they condition: e.g. $\mathbb{P}^a$ is the distribution of data with $a = 0$, and $\mathbb{Q}^a$ the distribution with $a = 1$. Take corresponding sample sets $S_{\mathbb{P}^a}$, $S_{\mathbb{Q}^a}$. In this section we will outline our approach for learning either a fair kernel or a fair vector representation.

We will assume in this paper that the relevant attributes a have two possible values, but extensions to a small number of discrete values are straightforward.

3.1 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 es-

timate 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), and subtract estimators of the two, resulting in the objective:

$$\Phi\left(\frac{(\mathrm{MMD}^t)^2 - c_\alpha^t/m}{\sqrt{V_m^t}}\right) - \Phi\left(\frac{(\mathrm{MMD}^s)^2 - c_\alpha^s/m}{\sqrt{V_m^s}}\right). \quad (5)$$

To do so, we initially used the permutation test estimator of the threshold c_α, 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 (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,

$$\begin{aligned}\rho_{b,B} &= \Pr_{\mathfrak{H}_1}\left(\sqrt{b}\,\widehat{\mathrm{MMD}}_{\mathrm{B}}^2 > \sqrt{V_B}t_\alpha\right) \\ &= \Pr_{\mathfrak{H}_1}\left(\frac{\sqrt{b}\,(\widehat{\mathrm{MMD}}_{\mathrm{B}}^2 - \mathrm{MMD}^2)}{\sqrt{V_B}} > t_\alpha - \frac{\sqrt{b}\,\mathrm{MMD}^2}{\sqrt{V_B}}\right) \\ &\to \Phi\left(\sqrt{b}\,\frac{\mathrm{MMD}^2}{\sqrt{V_B}} - t_\alpha\right). \quad (6)\end{aligned}$$

The block test's simple asymptotic threshold gives us a simple form, which is cheaper to compute than using the permutation test threshold in (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 we can estimate effectively (Liu et al. 2020). We can thus estimate the asymptotic power with

$$\hat{\rho}_{b,B} = \Phi\left(\sqrt{b}\,\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2}{\sqrt{\widehat{V}_{B,\lambda}}} - t_\alpha\right). \quad (7)$$

$\hat{\rho}_{b,B}$ will converge uniformly to $\rho_{b,B}$ over classes of deep kernels satisfying some technical assumptions as a corollary of Liu et al. (2020); proof in Appendix B.

Using (7), our objective to learn a fair kernel with sensitive attribute s and target attribute t is

$$\operatorname*{argmin}_{\omega}\left[\hat{\rho}_{b,B}^s - \hat{\rho}_{b,B}^t\right]. \quad (8)$$

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.

3.2 Learning 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 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 ideally like to solve

$$\min_{\phi}\left[\max_{\kappa}\hat{\rho}_{b,B}^s - \max_{\kappa}\hat{\rho}_{b,B}^t\right] \quad (9)$$

Instead, it is simpler to solve the following simpler problem, as justified by Proposition 1 (proof in Appendix C):

$$\min_{\phi}\max_{\kappa}\left[\hat{\rho}_{b,B}^s - \hat{\rho}_{b,B}^t\right]. \quad (10)$$

Proposition 1. *Suppose there is a "nearly perfect" representation $\hat{\phi}$ such that (a) for all κ, $\hat{\rho}_{b,B}^s \leq \varepsilon$, and (b) for some $\hat{\kappa}$, $\hat{\rho}_{b,B}^t \geq 1-\varepsilon$. Then any solution to* (9) *is at most 4ε-suboptimal on* (10)*, and vice versa.*

The objective (10) could be optimized with an alternating minimax optimization scheme for the parameters of κ, looking something like an MMD-GAN (C.-L. Li et al. 2017; Bińkowski et al. 2018). We find it sufficient in our experiments, though, to use a much simpler scheme: a grid of Gaussian kernels of varying lengthscales. 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.

3.3 Marginal vs Conditional Power

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

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

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: 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 don't 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 solve this, we can instead condition the sensitive pair over the target distributions, and sample points from $\mathbb{P}^{s|t} = P(X \mid S = 0, T = t)$ and $\mathbb{Q}^{s|t} = P(X \mid S = 1, T = t)$ for all values of T. This is now equivalent to optimizing for equalized odds (`EqOdds`) with respect to all distinct target classes t, defined as

$$\Delta_{EQ} = \big|P(\hat{T} = t \mid T = t, S = 0) - P(\hat{T} = t \mid T = t, S = 1)\big|.$$

This modifies the objectives (8) and (10) to, summing over the possible values of t,

$$\operatorname*{argmin}_{\omega} \left[\left(\sum_t \hat{\rho}^{s|t}_{b,B}\right) - \hat{\rho}^{t}_{b,B}\right], \tag{11}$$

$$\min_{\phi} \max_{\kappa} \left[\left(\sum_t \hat{\rho}^{s|t}_{b,B}\right) - \hat{\rho}^{t}_{b,B}\right]. \tag{12}$$

It is well-known that perfect demographic parity, $\Delta_{DP} = 0$, is not generally compatible with perfectly equalized odds, $\Delta_{EQ} = 0$ (Barocas et al. 2018). Even so, Theorem 3.1 of Zhao et al. (2020) shows that classifiers satisfying $\Delta_{EQ} = 0$ have demographic parity gaps Δ_{DP} upper-bounded by the gap of a perfect classifier, and hence training with an equalized odds criterion does not strongly compromise demographic parity.

3.4 Adding classifier loss

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.

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

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.

4 RELATED WORK

Fair representation learning has of late (deservedly) found a lot of traction within the deep learning community (Mehrabi et al. 2021). 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; Madras et al. 2018; Zhang 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. (2021) use the MMD as regularizers to a classifier, choosing between the marginal and conditiona form based on the causal direction of the task, to enforce counterfactual invariance. Most recently Lee et al. (2021) 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 5 (next), we compare to several different baselines. LAFTR (Madras et al. 2018) employs an adversarial

[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 doesn't require any samples with both s and t values.

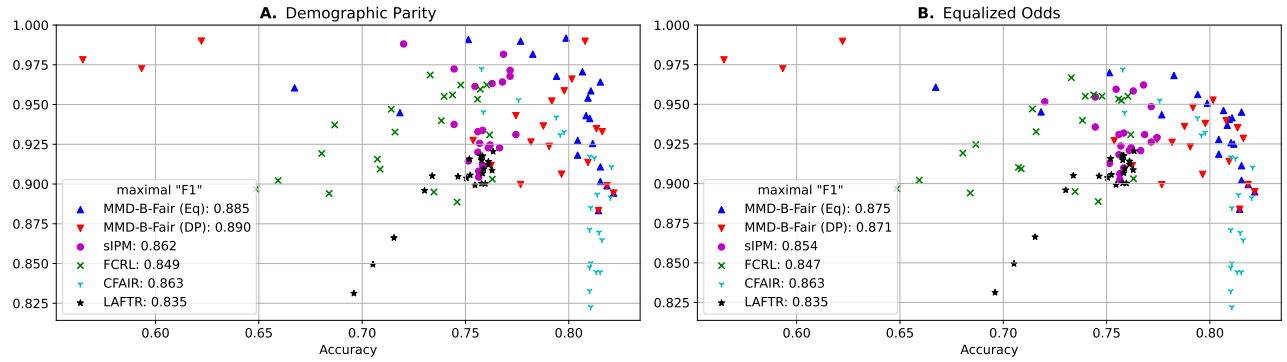


Figure 1: Fairness-accuracy measures on UCI Adult. Higher maximal "F1" score is better; 1 would indicate a point in the trade-off curve at the top-right corner.

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.

5 EXPERIMENTS

We evaluate both versions, (10) and (12), 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 predicting whether an individual's income is above $50,000 USD while being fair to their gender. The other dataset we evaluate on is the Heritage Health[3] 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.

In the main body, we present results only for fair representation learning. We explore fair kernel learning, and more model variants, in the supplemental material.

Experimental Setup We train all the models across 6 different choices of their respective fairness hyperparameters. For both versions of out method we set λ_s to $\{0, 0.1, 1, 4, 10, 100, 1000\}$ 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 hyperparameters (β and λ) proposed in their paper. We train all models with a minibatch size of 64 and report performance over three independent seeds.

[3] https://foreverdata.org/1015/

Fairness Firstly, we examine the fairness-accuracy trade-off fronts obtained by sweeping over the fairness hyperparameters in Figures 1 and 2 for Adult and Heritage Health, respectively. The x-axis is the target accuracy, computed as the average of the group-wise accuracies over the protected and unprotected sets; the y-axis reports the Demographic Parity (DP), $1 - \Delta_{DP}$, and Equalized Odds, $1 - \Delta_{EQ}$, averaged over both positive and negative target classes. Note that higher values are better. We show a scatterplot of each method's outcomes across different hyperparameter values.

The legend also indicates a score, inspired by the common F1 score for classification, of how far these tradeoff curves extend towards the top-right: the maximal value of the harmonic mean of accuracy and the fairness metric in question.

For the Adult dataset (Figure 1), our methods outperform the baselines, concurrently achieving high accuracy scores and fairness measures, and hence lying closer to the desirable top-right corner. For Heritage Health (Figure 2), the comparison is less clear, with different methods seeming to cover different parts of a similar trade-off curve even with dramatically different hyperparameter values.

Examining Learnt Representations One 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 labels is successfully hidden in the representation learning phase, then subsequent sensitive label classifiers will struggle to classify test points.

We train two-layer MLP classifiers on all the representations, and show in Figure 3a the sensitive classification per-

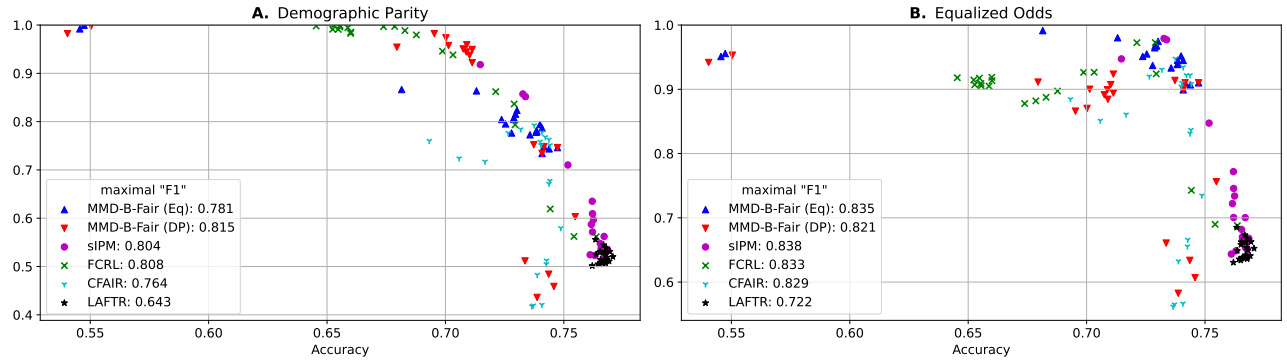
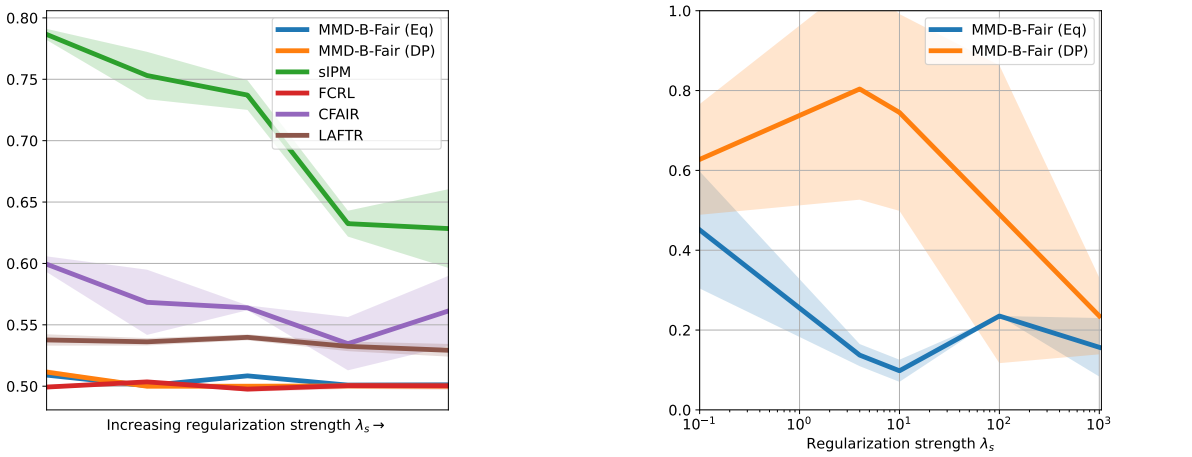


Figure 2: Fairness-accuracy measures on Heritage Health. Higher maximal "F1" score is better.



(a) Classification accuracy; 50% is ideal.

(b) Empirical power of an optimized MMD test; 0.05 is ideal.

Figure 3: Distinguishing sensitive attribute distributions on Adult, for varying regularization strengths.

formance across the regularization strengths used to train the fair models. Both the DP and Eq version of our method, as well as FCRL, are able to maintain the desired random accuracy score of 50% over sensitive labels across all regularization strengths. Other methods do not.

Checking whether this accuracy is 50% is essentially a classifier-based two-sample test (Lopez-Paz and Oquab 2017) between $\mathbb{P}^a$ and $\mathbb{Q}^a$, based on the learnt representation. We can 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-based test (Liu et al. 2020). For models with classification accuracies significantly above 50%, this power will be near-perfect, but it might be that even if few individual points can be correctly classified, a two-sample test might be able to distinguish the distributions as a whole. We run this check for our methods on Adult in Figure 3b, using a Gaussian kernel with learnt lengthscale on a one-layer MLP architecture trained to roughly maximize the asymptotic power as in Liu et al. (2020) (maximizing $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2 / \sqrt{\widehat{V}_{m,\lambda}}$). We then evaluate the empirical power of the test, how many times it rejects the null hypothesis, while repeating the test with 64 samples at a time.

Strikingly, two-sample tests here are far more sensitive measures of attribute leakage than classification accuracy, achieving nontrivial power across all regularization strengths. The equalized odds version of our method in particular manages to make the representations significantly less distinguishable, however.

Figure 4 shows t-SNE visualizations of latent space embeddings, further demonstrating that our method's representations separate the target attribute well and make the sensitive attribute difficult to distinguish.

Appendix D conducts this analysis for Heritage Health.

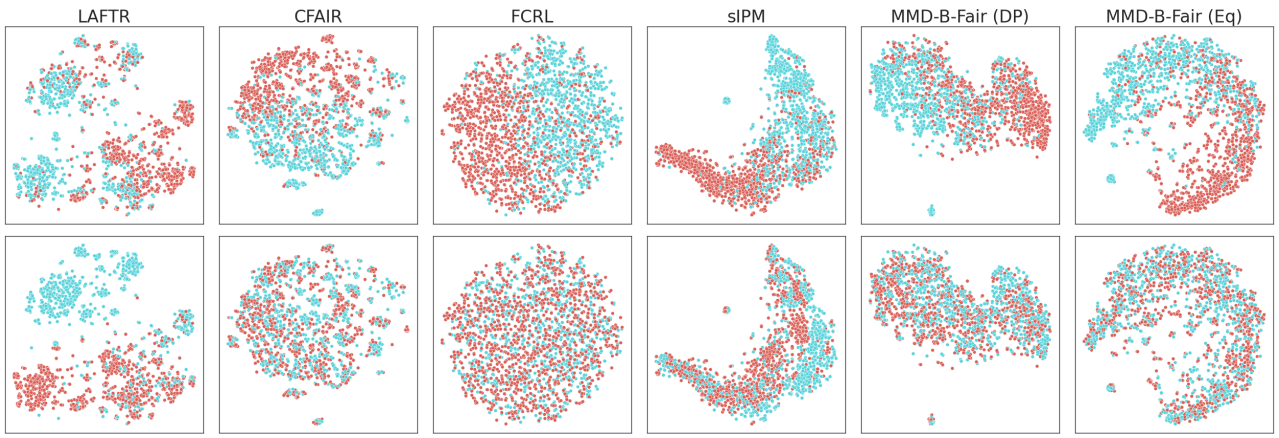


Figure 4: t-SNE visualizations of Adult representations, colored by target attribute (top) and sensitive attribute (bottom).

Fair Transfer Learning One major aim of fair *representation* learning, rather than simply finding a fair classifier, is to be able to use the same representation for more than one potential downstream task. We would like our representation to give good (and fair) performance for classifiers when trained on tasks unknown at representation learning time, even classifiers that are trained without any concern about fairness at all: the representation 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 not used in the original representation learning phase. We train these classifiers without regard to fairness by simply minimizing the cross-entropy loss.

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

Table 1 shows the resulting accuracy and fairness scores 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).

6 DISCUSSION

We proposed a method for learning fair kernels as well as representations built off of two-sample testing – a different paradigm than previous approaches to learning fair representations. Our approach 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.

Our method performs well compared to previous approaches based on adversarial learning and generative modelling. We provide two different versions of our approach – the marginal (demographic parity) version which can be trained using weak set-level labels from disjoint datasets, albeit at a disadvantage when dealing with correlated features, and a conditional (equalized odds) version, which can handle correlation between features well. We also show that, compared to previous approaches, our representations transfer well to new tasks with respect to both accuracy and fairness.

Areas for future work include building in support for continuous-valued sensitive attributes via the Hilbert-Schmidt Independence Criterion (Gretton et al. 2008) and extending to related applications like domain adaptation, invariant feature learning, causal representation learning, and so on.

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2018). "Fairness and Machine Learning Limitations and Opportunities."

Bickel, P. J. and E. L. Lehmann (1969). "Unbiased Estimation in Convex Families." *The Annals of Mathematical Statistics* 40.5, pp. 1523–1535.

Bińkowski, Mikołaj, Danica J. Sutherland, Michael Arbel, and Arthur Gretton (2018). "Demystifying MMD GANs." *ICLR*. arXiv: 1801.01401.

Bogen, Miranda and Aaron Rieke (2018). "Help wanted: an examination of hiring algorithms, equity, and bias."

Bounliphone, Wacha, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton (2016). "A Test of Relative Similarity For Model Selection in Generative Models." *ICLR*. arXiv: 1511.04581.

Chouldechova, Alexandra, Diana Benavides Prado, Oleksandr Fialko, and Rhema Vaithianathan (2018). "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." *FAT*.

Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel (2019). "Flexibly Fair Representation Learning by Disentanglement." arXiv: 1906.02589.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Edwards, Harrison and Amos J. Storkey (2016). "Censoring Representations with an Adversary." *ICLR*. arXiv: 1511.05897.

Flores, Anthony W., Kristin A. Bechtel, and Christopher T. Lowenkamp (2016). "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, p. 38.

Gianfrancesco, Milena A, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk (2018). "Potential biases in machine learning algorithms using electronic health record data." *JAMA internal medicine* 178.11, pp. 1544–1547.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A kernel two-sample test." *The Journal of Machine Learning Research* 13.1, pp. 723–773.

Gretton, Arthur, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola (2008). "A Kernel Statistical Test of Independence." *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc.

Gupta, Umang, Aaron Ferber, Bistra N. Dilkina, and Greg Ver Steeg (2021). "Controllable Guarantees for Fair Outcomes via Contrastive Information Estimation." *AAAI*.

Hajian, Sara, Francesco Bonchi, and Carlos Castillo (2016). "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Jean, Neal, Sang Michael Xie, and Stefano Ermon (2018). "Semi-supervised Deep Kernel Learning: Regression with Unlabeled Data by Minimizing Predictive Variance." *NeurIPS*. arXiv: 1805.10407.

Jo, Eun Seo and Timnit Gebru (2020). "Lessons from archives: Strategies for collecting sociocultural data in machine learning." *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 306–316.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma (2012). "Considerations on Fairness-Aware Data Mining." *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 378–385.

Kim, Dongha, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim (2022). "Learning fair representation with a parametric integral probability metric." *ICML*.

Lahoti, Preethi, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi (2020). *Fairness without Demographics through Adversarially Reweighted Learning*. arXiv: 2006.13114.

Lebovits, Hannah (2018). "Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor." *Public Integrity* 21, pp. 448–452.

Ledford, Heidi (2019). "Millions of black people affected by racial bias in health-care algorithms." *Nature* 574, pp. 608–609.

Lee, Junghyun, Gwangsun Kim, Matt Olfat, Mark A. Hasegawa-Johnson, and Chang Dong Yoo (2021). *Fast and Efficient MMD-based Fair PCA via Optimization over Stiefel Manifold*. arXiv: 2109.11196.

Li, Chun-Liang, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos (2017). "MMD GAN: Towards Deeper Understanding of Moment Matching Network." *NeurIPS*. arXiv: 1705.08584.

Li, Yazhe, Roman Pogodin, Danica J. Sutherland, and Arthur Gretton (2021). "Self-Supervised Learning with Kernel Dependence Maximization." *NeurIPS*. arXiv: 2106.08320.

Liu, Feng, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland (2020). "Learning deep kernels for non-parametric two-sample tests." *International Conference on Machine Learning*. PMLR, pp. 6316–6326.

Lopez-Paz, David and Maxime Oquab (2017). "Revisiting Classifier Two-Sample Tests." *ICLR*. arXiv: 1610.06545.

Louizos, Christos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel (2016). "The Variational Fair Autoencoder." *ICLR*. arXiv: 1511.00830.

Madras, David, Elliot Creager, Toniann Pitassi, and Richard S. Zemel (2018). "Learning Adversarially Fair and Transferable Representations." *ICML*.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan (2021). "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54, pp. 1–35.

Norouzi, Sajad (2020). "Variational Fair Information Bottleneck."

Oneto, L., Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil (2020). "Exploiting MMD and Sinkhorn Divergences for Fair and Transferable Representation Learning." *NeurIPS*.

Ramdas, Aaditya, Sashank J. Reddi, Barnabas Poczos, Aarti Singh, and Larry Wasserman (2015). *Adaptivity and Computation-Statistics Tradeoffs for Kernel and Distance based High Dimensional Two Sample Testing*. arXiv: 1508.00655.

Serfling, Robert (1980). *Approximation Theorems of Mathematical Statistics*.

Skeem, Jennifer L and Christopher T. Lowenkamp (2016). "RISK, RACE, AND RECIDIVISM: PREDICTIVE BIAS AND DISPARATE IMPACT*: RISK, RACE, AND RECIDIVISM." *Criminology* 54, pp. 680–712.

Speicher, Till, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Kumar Singla, Adrian Weller, and Muhammad Bilal Zafar (2018). "A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Sutherland, Danica J. and Namrata Deka (2019). *Unbiased estimators for the variance of MMD estimators*. arXiv: 1906.02104.

Sutherland, Danica J., Hsiao-Yu Fish Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton (2017). "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy." *ICLR*.

Veitch, Victor, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein (2021). *Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests*. arXiv: 2106.00545.

Warner, Stanley L. (1965). "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60.309, pp. 63–69.

Wilson, Andrew Gordon, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing (2016). "Deep kernel learning." *AISTATS*. arXiv: 1511.02222.

Wilson, Benjamin, Judy Hoffman, and Jamie H. Morgenstern (2019). "Predictive Inequity in Object Detection." *ArXiv* abs/1902.11097.

Xie, Qizhe, Zihang Dai, Yulun Du, Eduard H. Hovy, and Graham Neubig (2017). "Controllable Invariance through Adversarial Feature Learning." *NIPS*.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi (2017). "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment." *Proceedings of the 26th International Conference on World Wide Web*.

Zaremba, Wojciech, Arthur Gretton, and Matthew B. Blaschko (2013). "B-tests: Low Variance Kernel Two-Sample Tests." *Advances in Neural Information Processing Systems*. arXiv: 1307.1954.

Zemel, Richard S., Ledell Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork (2013). "Learning Fair Representations." *ICML*.

Zhang, B., Blake Lemoine, and Margaret Mitchell (2018). "Mitigating Unwanted Biases with Adversarial Learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Zhao, Han, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon (2020). "Conditional Learning of Fair Representations." *ICLR*. arXiv: 1910.07162.

A Non-existence of an unbiased estimator

Proposition 2. *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 $\mathcal{P}$ be some class of distributions such that $\{(1-\alpha)\mathbb{P}_0 + \alpha\mathbb{P}_1 : \alpha \in [0,1]\} \subseteq \mathcal{P}$, 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 $\mathcal{P}$.*

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,

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

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

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

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 $\mathcal{P}$.

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

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

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

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

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

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

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 don't 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. □

B Uniform convergence of our objective

We show here that optimizing the approximated block-test power from (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 that with probability at least $1-\delta$ over the choice of n samples used in the estimators

$$\sup_{k \in \mathcal{K}} \left| \frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2}{\sqrt{n \widehat{V}_{n, n \cdot n^{-1/3}}}} - \frac{\mathrm{MMD}^2}{\sqrt{\lim_{m \to \infty} m V_m}} \right| \leq \alpha(\mathcal{K}, \mathbb{P}, \mathbb{Q}, n, \delta) \tag{13}$$

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 $\mathcal{K}$ is the class of considered kernels; note that mV_m converges to a constant.

Notice from (4) that, for any m and ℓ, $\widehat{V}_{\ell,\lambda} = \frac{m}{\ell} \widehat{V}_{m,\lambda}$. Thus we can rewrite (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 (13) more compactly as showing that, with probability at least $1-\delta$, $\sup_{k\in\mathcal{K}} |\hat{J}_{n^{2/3}} - J| \le \alpha(\mathcal{K}, \mathbb{P}, \mathbb{Q}, n, \delta)$.

Also, notice from (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:

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

Thus applying this function to each of the terms in (13) yields that, when we use $\lambda = n^{2/3}$,

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

Uniform convergence of each $\hat{\rho}_{b,B}$ to the relevant asymptotic power immediately implies uniform convergence of the objective (8) or (10) to the term based on asymptotic powers, by a union bound.

C Proof of Proposition 1

Proof. For clarity, we will use $\hat{\rho}^a_{\phi,\kappa}$ to denote $\hat{\rho}^a_{b,B}$ for the kernel $\phi \circ \kappa$, with a either s or t.

By assumption, $\hat{\phi}$ achieves a value of at most $2\varepsilon - 1$ on (9). Thus any solution $\tilde{\phi}$ to (9) must be at least that good. Then, since $\max_\kappa \hat{\rho}^t_{\tilde{\phi},\kappa} \le 1$, necessarily $\max_\kappa \hat{\rho}^s_{\tilde{\phi},\kappa} \le 2\varepsilon$. Similarly, since $\max_\kappa \hat{\rho}^t_{\tilde{\phi},\kappa} \ge 0$, $\max_\kappa \hat{\rho}^t_{\tilde{\phi},\kappa} \ge 1 - 2\varepsilon$. Thus $\max_\kappa \hat{\rho}^s_{\tilde{\phi},\kappa} - \hat{\rho}^t_{\tilde{\phi},\kappa} \le 4\varepsilon - 1$, 4ε worse than the optimal value of -1.

For the other direction, note that $(\hat{\phi}, \hat{\kappa})$ also achieves value at most $2\varepsilon - 1$ on (10). Thus, any solution $(\tilde{\phi}, \tilde{\kappa})$ must also achieve value at most $2\varepsilon - 1$. Like before, this implies that $\hat{\rho}^s_{\tilde{\phi},\tilde{\kappa}} \le 2\varepsilon$ and $\hat{\rho}^s_{\tilde{\phi},\tilde{\kappa}} \ge 1 - 2\varepsilon$. Moreover, if there were any κ with $\hat{\rho}^s_{\tilde{\phi},\kappa} > 2\varepsilon$, it would necessarily have $\hat{\rho}^s_{\tilde{\phi},\kappa} - \hat{\rho}^t_{\tilde{\phi},\kappa} > 2\varepsilon - 1$, contradicting that $\tilde{\kappa}$ maximizes $\hat{\rho}^s_{\tilde{\phi},\kappa} - \hat{\rho}^t_{\tilde{\phi},\kappa}$. Thus $\tilde{\phi}$ achieves a value of at most $4\varepsilon - 1$ on (9), again 4ε worse than the optimal value of -1. □

D Further experimental results

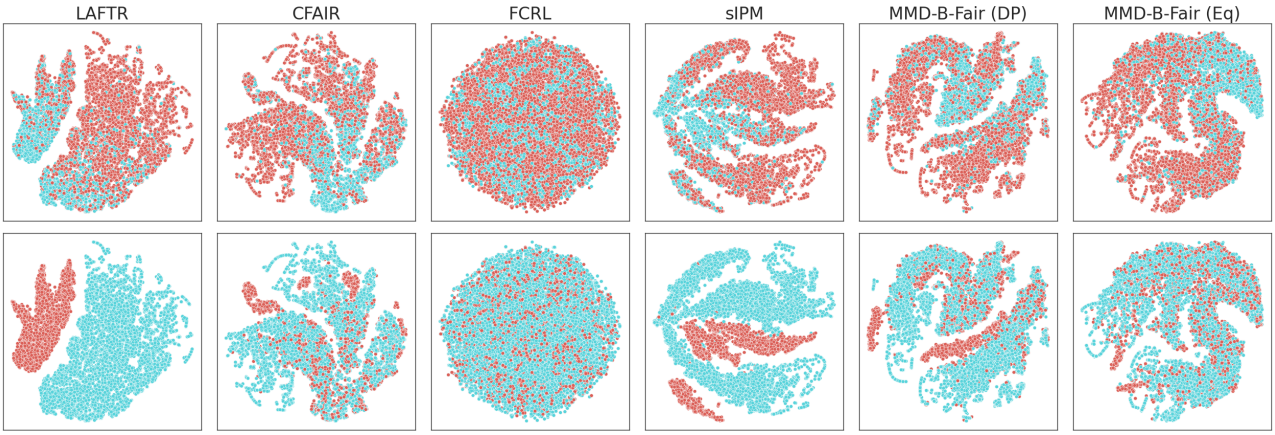
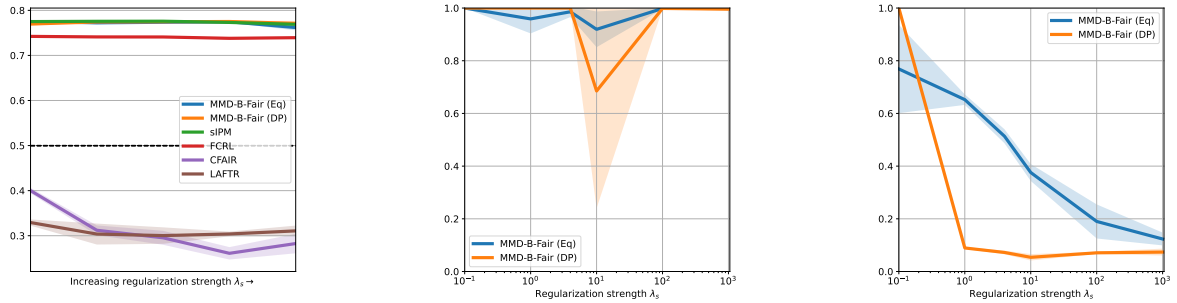


Figure 5: t-SNE visualizations of Heritage Health representations, colored by target attribute (top) and sensitive attribute (bottom).

D.1 Heritage Health Learnt Representations

Here, we examine the learnt representations obtained for the Heritage Health dataset. Figure 5 shows the t-SNE visualizations of the latent space embeddings. We examine the information contained in the learnt representations about the



(a) Sensitive classification accuracy; 0.5 is ideal.

(b) Empirical power of an **optimized** MMD test; 0.05 is ideal.

(c) Empirical power of an MMD test; 0.05 is ideal.

Figure 6: Distinguishing sensitive attribute distributions on Heritage Health, for varying regularization strengths.

sensitive attribute in Figure 6. Sensitive classification performance of a two-layer MLP trained on the representations from each method is shown in Figure 6a. Compared to the Adult dataset, on Heritage Health all methods perform poorly, i.e. they are able to predict the sensitive labels with non-random accuracy indicating that hiding sensitive information on this dataset is a challenging task. Figure 6b plots the empirical power of an optimized MMD test after fine-tuning a 1-layer MLP along with a Gaussian kernel to maximize the sensitive power using the learnt features. Ideally, this should correspond to the significance level α (0.05) if the features did not contain distinguishing information regarding S. However, representations learnt by both our methods irrespective of regularization strengths did not meet this requirement. We were able to find deep kernels that could very easily distinguish $S_{\mathbb{P}^s}$ from $S_{\mathbb{Q}^s}$, despite the fact that a MMD test directly on the representations had the desired behavior for high regularization strengths (Figure 6c).

D.2 Fair Kernels

Using Equation (8) and its conditional version from Equation (12), (we refer to them as MMD-B-Fair-Kernel (DP) and MMD-B-Fair-Kernel (Eq)) we train fair kernels and use them for downstream classification with support vector machines (SVMs) with a radial basis function(r.b.f.) kernel to classify the sensitive and target labels. The SVMs are trained (3-fold cross-validation over the test split) on the features obtained from the trained featurizer network ϕ_ω, and the bandwidth of the SVM's r.b.f. kernel is set to the learnt Gaussian bandwidth σ_ϕ. Figure 7 shows both target and sensitive label accuracies over increasing regularization strengths λ_s on the Adult dataset. As is expected, sensitive classification accuracy is near-random (50%) regardless of λ_s while target classification accuracies decrease with stronger regularization. MMD-B-Fair-Kernel (Eq) performs slightly better with respect to having near-random sensitive accuracy than MMD-B-Fair-Kernel (DP).

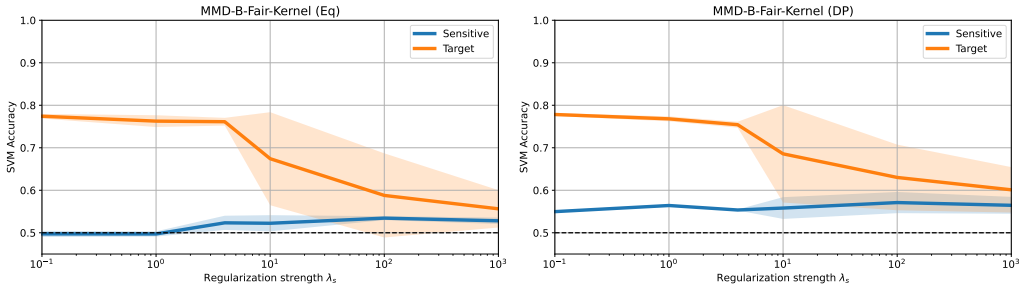


Figure 7: SVM on Adult using learnt deep kernel from Equation (8). Ideal value for sensitive is 0.5.