METRIZING FAIRNESS

YVES RYCHENER, BAHAR TAŞKESEN, AND DANIEL KUHN

Abstract. We study supervised learning problems for predicting properties of individuals who belong to one of two demographic groups, and we seek predictors that are fair according to statistical parity. This means that the distributions of the predictions within the two groups should be close with respect to the Kolmogorov distance, and fairness is achieved by penalizing the dissimilarity of these two distributions in the objective function of the learning problem. In this paper, we showcase conceptual and computational benefits of measuring unfairness with integral probability metrics (IPMs) other than the Kolmogorov distance. Conceptually, we show that the generator of any IPM can be interpreted as a family of utility functions and that unfairness with respect to this IPM arises if individuals in the two demographic groups have diverging expected utilities. We also prove that the unfairness-regularized prediction loss admits unbiased gradient estimators if unfairness is measured by the squared $\mathcal{L}^2$-distance or by a squared maximum mean discrepancy. In this case the fair learning problem is susceptible to efficient stochastic gradient descent (SGD) algorithms. Numerical experiments on real data show that these SGD algorithms outperform state-of-the-art methods for fair learning in that they achieve superior accuracy-unfairness trade-offs—sometimes orders of magnitude faster. Finally, we identify conditions under which statistical parity can improve prediction accuracy.

1. Introduction

The last decade has witnessed a surge of algorithms for decision-making systems that have a consequential impact on our daily lives. Machine learning methods are increasingly used, for example, to decide whom to grant or deny loans, college admission, bail or parole. Even though it would be natural to expect that algorithms are free of prejudice, it turns out that cutting-edge AI techniques can learn or even amplify human biases and may thus be far from fair [13, 20, 36]. The necessity to correct algorithmic biases has propelled the growing field of fair machine learning, see [5, 9, 14, 17, 42] for comprehensive surveys of the state-of-the-art in this area. As of today, there exist several mathematical definitions of algorithmic fairness. All of these definitions fall into one of three main categories: (i) notions of *group fairness* ask that different demographic groups have equal chances of securing beneficial outcomes [4, 27, 33, 51, 74, 79, 80], (ii) notions of *individual fairness* demand that individuals with similar covariates should be treated similarly [25, 62, 78], and (iii) *causality-based fairness* notions require that protected attributes such as gender or race have no causal effect on outcomes [29, 35, 40, 46, 84, 85]. This paper focuses on notions of group fairness, which are widely studied in a variety of applications. Prominent group fairness criteria include statistical parity (also known as demographic parity) [11, 27, 80], equal opportunity [33] and equalized odds [33] as well as their probabilistic counterparts [51]. All of these criteria quantify fairness in terms of the distributions of the algorithms' outputs conditional on the attributes of different demographic groups.

Contributions. We study supervised learning problems whose solutions critically impact individuals of two demographic groups, and we seek predictors that are fair in the sense of statistical parity. By definition, this means that the distributions of the predictions within the two groups should be close in Kolmogorov distance, which constitutes an IPM. Fairness can thus be enforced by penalizing the dissimilarity of the two distributions in the objective function of the learning problem. In this paper we first prove that if membership in a certain demographic group provides no information about the distribution of the learning target conditional on the features, then the optimal predictor satisfies statistical parity even if it is not explicitly enforced. We then

The authors are with the Risk Analytics and Optimization Chair, EPFL Lausanne (yves.rychener,bahar.taskesen, daniel.kuhn@epfl.ch).

argue that if training is based on a biased target, fairness constraints can have a regularizing effect and improve accuracy in view of an ideal target, for which there is no data. This reasoning provides a theoretical justification for enforcing statistical parity and may help to identify learning tasks in which fairness penalties produce desirable outcomes. Next, we flesh out conceptual and computational benefits of measuring unfairness with IPMs other than the Kolmogorov distance. On the one hand, we show that the generator of any IPM can be interpreted as a family of utility functions and that unfairness with respect to this IPM arises if individuals in the two demographic groups have diverging expected utilities. This establishes a utilitarian perspective on unfairness measures. On the other hand, we prove that unfairness-regularized learning problems are susceptible to efficient SGD algorithms if unfairness is measured by the squared $\mathcal{L}^2$-distance or by a squared maximum mean discrepancy. Unbiased gradient estimators, which are necessary for SGD to converge, are difficult to obtain because the unfairness penalty is non-linear in the distribution of the training samples. Focusing on IPMs related to U-statistics and constructing random batches of training samples, we can eliminate the systematic bias of the naïve empirical unfairness penalty. However, debiasing the unfairness penalty in this manner introduces a bias in the empirical prediction loss. We thus also need to derive a bias correction term for the empirical prediction loss. In summary, these techniques allow us to obtain unbiased gradient estimators for the overall learning objective at a low computational cost. Numerical experiments on real data show that our SGD-based approach to solve fair learning problems outperforms state-of-the-art methods in that it achieves a superior accuracy-unfairness trade-off—sometimes at significantly reduced runtimes.

Related work. Fair supervised learning models can be categorized into three main groups: (i) *preprocessing methods* correct biased data before the training process [27, 37, 76, 83, 23], (ii) *in-processing methods* incorporate fairness requirements into the training process [22, 28, 38, 79, 80, 57], and (iii) *post-processing methods* adjust a trained predictor to obey some desired fairness requirement [16, 33, 51, 49]. The methods proposed in this paper add an unfairness regularizer in the form of an IPM to the training objective, and thus they belong to the *in*-processing methods. More specific regularization schemes for incentivizing fairness are described in [8, 34, 47, 67]. The possibility to quantify unfairness via generic IPMs was informally mentioned in [50] but not systematically investigated. Fairness can also be enforced rigidly via hard constraints on the predictors [22, 75, 80, 81]. As the resulting constrained optimization problems are often non-convex and unsuitable for gradient descent algorithms, however, hard fairness constraints are typically relaxed in practice [22, 47, 67, 75, 80]. The extent to which the solutions of these relaxed problems comply with the original fairness constraints is discussed in [41]. Maximum mean deviations were already used in [47] to enforce statistical parity at an intermediate layer of a neural network predictor. As no unbiased gradient estimators are derived, however, the corresponding learning problems cannot be addressed with SGD. We emphasize that unbiased gradient estimators are instrumental for many application areas of machine learning that rely on stochastic optimization such as latent variable models [68], unrolled computations [70], federated learning [77], distributionally robust optimization [31] and generative adversarial networks [6]. However, unbiasedness has not yet been investigated in the fairness domain.

This paper is structured as follows. Section 2 introduces statistical parity, whereas Section 3 formalizes the connections between unfairness measures and IPMs. Section 4 then addresses the numerical solution of learning problems with unfairness penalties, and Section 5 reports on numerical results. Finally, Section 6 concludes. Proofs and expository examples are relegated to the appendix.

Notation. All random vectors are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and the expectation with respect to $\mathbb{P}$ is denoted by $\mathbb{E}[\cdot]$. Random vectors are denoted by capital letters (*e.g.,* X), and their realizations are denoted by the corresponding lower-case letters (*e.g.,* x). The cumulative distribution function (CDF) of a random vector $X \in \mathbb{R}$ is denoted by F_X and satisfies $F_X(x) = \mathbb{P}[X \leq x]$ for all $x \in \mathbb{R}$. Similarly, the probability distribution of X is denoted by $\mathbb{P}_X$ and satisfies $\mathbb{P}_X[B] = \mathbb{P}[X \in B]$ for all Borel sets $B \subseteq \mathbb{R}^d$. We write $X \perp Y$ to indicate that the random vectors X and Y are independent under $\mathbb{P}$. For any Borel sets $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_\mathcal{Y}}$ we denote by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ the space of all Borel-measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. Given a norm $\|\cdot\|$

on $\mathbb{R}^d$, the Lipschitz modulus of $f : \mathbb{R}^d \to \mathbb{R}$ is defined as $\mathrm{Lip}(f) = \sup_{x,x' \in \mathbb{R}^d} \{|f(x) - f(x')|/\|x - x'\| : x \neq x'\}$. The indicator function $\mathbb{1}_{\mathcal{S}}$ of a logical statement $\mathcal{S}$ evaluates to 1 if $\mathcal{S}$ is true and to 0 otherwise.

2. Fairness in Supervised Learning

We study regression and classification problems of the form

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] \tag{1}$$

that aim to predict a property $Y \in \mathcal{Y} \subseteq \mathbb{R}$ (the output) of a human being characterized by a feature vector $X \in \mathcal{X} \subseteq \mathbb{R}^d$ (the input). Here, $\mathcal{H}$ represents a family of Borel-measurable hypotheses $h : \mathcal{X} \to \mathbb{R}$, and $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ represents a lower semi-continuous loss function that quantifies the discrepancy between the predicted output $h(X)$ and the actual output Y. Throughout the paper we assume that individuals have a protected attribute $A \in \mathcal{A} = \{0, 1\}$ that encodes their race, religion, age or sex etc. Note that the protected attribute A may impact the feature vector X or even be one of the features. If discrimination with respect to A is undesired or legally forbidden, however, one should seek classifiers and regressors $h(X)$ that make no use of A. Unfortunately, deleting A from the list of features (which is sometimes referred to as 'fairness by unawareness' [5]) is not helpful because the entirety of data collected about a person typically provides enough information to infer A with high reliability. Instead of ignoring A, we require the classifiers and regressors to satisfy formal statistical notions of fairness, which are defined in terms of conditional probability distributions. One of the most widely used notions of fairness in machine learning is statistical parity.

Definition 2.1 (Statistical parity [1])**.** A hypothesis $h \in \mathcal{H}$ satisfies statistical parity (SP) at level ε if

$$|\mathbb{P}(h(X) \leq \tau | A = 0) - \mathbb{P}(h(X) \leq \tau | A = 1)| \leq \varepsilon \ \forall \tau \in \mathbb{R}. \tag{2}$$

Statistical parity is also called demographic parity [25] or disparate impact [27]. Note that a hypothesis h satisfies SP at level $\varepsilon = 0$ if and only if the conditional CDFs $F_{h(X)|A=0}$ and $F_{h(X)|A=1}$ match, which means that $h(X) \perp A$. In the special case of binary classification, we have $h(X) \perp A$ if and only if the true positive rates match across the subgroups corresponding to $A = 0$ and $A = 1$. SP at level $\varepsilon > 0$ implies similarity of the conditional CDFs $F_{h(X)|A=0}$ and $F_{h(X)|A=1}$. It is enforced, for example, via the US Equal Employment Opportunities Commission's 80% rule [69], which requires that the selection rate for any race, sex, or ethnic group be less than 80% of the rate for the group with the highest rate. The following theorem shows that the optimal solution of problem (1) automatically satisfies SP at level 0 if A carries no information about the distribution of Y conditional on X.

Theorem 2.2 (Optimality implies SP)**.** Suppose that (1) is a solvable binary classification problem with $\mathcal{Y} = \{0, 1\}$ and $\mathcal{H} = \mathcal{L}(\mathcal{X}, \mathbb{R})$. If $\mathbb{P}[Y = 1|X] \perp A$, then (1) has an optimizer $h^\star$ with $h^\star(X) \perp A$.

Theorem 2.2 readily extends to multiclass classification and regression problems at the expense of a more involved proof. It implies that if we had unlimited training data, which would enable us to infer the exact distribution of (X, Y) and thus allow us to optimize over all measurable hypotheses without overfitting, then minimizing the prediction error would implicitly ensure fairness. In practice, however, the target Y (*e.g.*, college GPA) may be a noisy or biased proxy for the ideal target $Y^\star$ (*e.g.*, intellectual potential), and Y is only used for classification because there is no data for $Y^\star$. In this case we may have $\mathbb{P}[Y = 1|X] \not\perp A$ (*e.g.*, GPA may depend on ethnicity) even though $\mathbb{P}[Y^\star = 1|X] \perp A$ (*e.g.*, intellectual potential is *in*dependent of ethnicity). In addition, there may not be enough training data to infer the distribution of (X, Y) exactly. For these reasons, the optimizer of problem (1) may violate SP even though the ideal hypothesis that minimizes the ideal prediction loss satisfies SP. Restricting the feasible set of problem (1) to hypotheses that satisfy SP thus amounts to enforcing a known property of the inaccessible ideal hypothesis. This regularization can lead to optimizers that have *smaller* prediction loss with respect to the ideal target; see Example A.1. Empirical evidence for this observation is also provided in [73]. We stress that Theorem 2.2 ceases to hold if the assumption $\mathbb{P}[Y = 1|X] \perp A$ is replaced with the simpler assumption $Y \perp A$; see Example A.2.

3. Unfairness Measures and Integral Probability Metrics

Definition 2.1 implies that if a hypothesis $h \in \mathcal{H}$ satisfies SP at level ε, then the absolute difference between the CDFs of $h(X)$ conditional on $A = 0$ and $A = 1$ is uniformly bounded by ε. Equivalently, the Kolmogorov distance [63] between $\mathbb{P}_{h(X)|A=0}$ and $\mathbb{P}_{h(X)|A=1}$ is at most ε [16]. Hence, the Kolmogorov distance between $\mathbb{P}_{h(X)|A=0}$ and $\mathbb{P}_{h(X)|A=1}$ quantifies the degree of *un*fairness of h. The Kolmogorov distance is an example of an integral probability metric (IPM) [45, 64].

Definition 3.1 (Integral probability metric)**.** Let $w \in \mathcal{L}(\mathbb{R}^n, [1, \infty))$ be a weight function, define $\mathcal{L}_w(\mathbb{R}^n, \mathbb{R})$ as the set of all functions $\psi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ with $\sup_{z\in\mathbb{R}^n} |\psi(z)|/w(z) < \infty$, and define $\mathcal{Q}_w(\mathbb{R}^n)$ as the set of all probability measures $\mathbb{Q}$ on $\mathbb{R}^n$ with $\int_{\mathbb{R}^n} w(z)\mathbb{Q}(\mathrm{d}z) < \infty$. The integral probability metric on $\mathcal{Q}_w(\mathbb{R}^n)$ with generator $\Psi \subseteq \mathcal{L}_w(\mathbb{R}^n, \mathbb{R})$ is then given by

$$\mathcal{D}_\Psi(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\psi\in\Psi} \left| \int_{\mathbb{R}^n} \psi(z)\, \mathbb{Q}_1(\mathrm{d}z) - \int_{\mathbb{R}^n} \psi(z)\, \mathbb{Q}_2(\mathrm{d}z) \right| \quad \forall \mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{Q}_w(\mathbb{R}^n).$$

Note that $\mathcal{D}_\Psi$ is a pseudo-metric for any $\Psi \subseteq \mathcal{L}_w(\mathbb{R}^n, \mathbb{R})$. Indeed, $\mathcal{D}_\Psi(\mathbb{Q}_1, \mathbb{Q}_2)$ is non-negative, symmetric in $\mathbb{Q}_1$ and $\mathbb{Q}_2$, vanishes if $\mathbb{Q}_1 = \mathbb{Q}_2$ and satisfies the triangle inequality. Moreover, $\mathcal{D}_\Psi$ is a metric if Ψ separates points in $\mathcal{Q}_w(\mathbb{R}^n)$, in which case $\mathcal{D}_\Psi(\mathbb{Q}_1, \mathbb{Q}_2)$ vanishes only if $\mathbb{Q}_1 = \mathbb{Q}_2$. The Kolmogorov distance is indeed an IPM as it is generated by the family of step-functions of the form $\psi(y) = \mathbb{1}_{y\leq\tau}$ parametrized by $\tau \in \mathbb{R}$. We emphasize, however, that other IPMs have previously been used to quantify unfairness [25]. A prominent example is the 1-Wasserstein distance.

Definition 3.2 (Wasserstein distance)**.** If $w(z) = 1 + \|z\|$ for some norm $\|\cdot\|$ on $\mathbb{R}^n$, then the 1-Wasserstein distance (or Kantorovich distance) between $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{Q}_w(\mathbb{R}^n)$ is

$$\mathcal{W}(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)} \int_{\mathbb{R}^n\times\mathbb{R}^n} \|z - z'\|\, \pi(\mathrm{d}z, \mathrm{d}z'),$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ denotes the set of all joint distributions or couplings of the random vectors $Z \in \mathbb{R}^n$ and $Z' \in \mathbb{R}^n$ with marginal distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively.

By the classical Kantorovich-Rubinstein theorem, the 1-Wasserstein distance is indeed an IPM.

Lemma 3.3 (Kantorovich-Rubinstein theorem [71, Remark 6.5])**.** The 1-Wasserstein distance $\mathcal{W}$ coincides with the IPM $\mathcal{D}_\Psi$ generated by the set $\Psi = \{\psi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}) : \mathit{Lip}(\psi) \leq 1\}$ of all Lipschitz continuous test functions with Lipschitz modulus of at most 1.

For univariate distributions, the Wasserstein Distance reduces to the $\mathcal{L}^1$-distance [58].

Definition 3.4 ($\mathcal{L}^p$-distance)**.** If $n = 1$ and $w(z) = 1 + |z|$, then the $\mathcal{L}^p$-distance of $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{Q}_w(\mathbb{R})$ is given by the $\mathcal{L}^p$-norm distance of their CDFs $F_{\mathbb{Q}_1}$ and $F_{\mathbb{Q}_2}$, respectively, that is,

$$d_p(\mathbb{Q}_1, \mathbb{Q}_2) = \|F_{\mathbb{Q}_1} - F_{\mathbb{Q}_2}\|_{\mathcal{L}^p} = \left(\int_{\mathbb{R}} |F_{\mathbb{Q}_1}(z) - F_{\mathbb{Q}_2}(z)|^p \, dz \right)^{1/p}$$

for $p \in [1, \infty)$ and $d_\infty(\mathbb{Q}_1, \mathbb{Q}_2) = \|F_{\mathbb{Q}_1} - F_{\mathbb{Q}_2}\|_{\mathcal{L}^\infty} = \sup_{z\in\mathbb{R}} |F_{\mathbb{Q}_1}(z) - F_{\mathbb{Q}_2}(z)|$ for $p = \infty$.

It can be shown that any $\mathcal{L}^p$-distance on a space of univariate distributions is in fact an IPM.

Lemma 3.5 (Duals of $\mathcal{L}^p$-distances [21, Lemma 1])**.** For any pair of conjugate exponents $p, q \in [1, \infty)$ with $1/p + 1/q = 1$, the $\mathcal{L}^p$-distance d_p on $\mathcal{Q}_w(\mathbb{R}^1)$ coincides with the IPM $\mathcal{D}_\Psi$ generated by the set $\Psi = \{\psi \in W_0^{1,q}(\mathbb{R}) : \|\psi'\|_{\mathcal{L}^q} \leq 1\}$, where $W_0^{1,q}(\mathbb{R})$ denotes the Sobolev space of all absolutely continuous functions $\psi : \mathbb{R} \to \mathbb{R}$ whose derivative ψ' has a finite $\mathcal{L}^q$-norm.

For $p = \infty$, the $\mathcal{L}^p$-distance collapses to the Kolmogorov distance. The squared $\mathcal{L}^2$-distance is sometimes called the Cramér distance [19], which can violate the triangle inequality and is thus only a semi-metric. The

two-fold multiple of the squared $\mathcal{L}^2$-distance is also known as the energy distance [3, 66]. Hence, the square root of the (univariate) energy distance is an IPM. The square root of the (multivariate) energy distance is also an instance of a maximum mean discrepancy (MMD).

Definition 3.6 (Maximum mean discrepancy)**.** If $K \in \mathcal{L}(\mathbb{R}^n \times \mathbb{R}^n, \mathbb{R})$ is a positive definite symmetric kernel and $w \in \mathcal{L}(\mathbb{R}^n, [1, \infty))$ satisfies $\sup_{z \in \mathbb{R}^n} K(z, z')/w(z) < \infty$ for all $z' \in \mathbb{R}^n$, then the maximum mean discrepancy between $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{Q}_w(\mathbb{R}^n)$ relative to K is given by

$$d_{MMD}(\mathbb{Q}_1, \mathbb{Q}_2) = \bigg(\int_{\mathbb{R}^n \times \mathbb{R}^n} K(z, z')\, \mathbb{Q}_1(\mathrm{d}z) \mathbb{Q}_1(\mathrm{d}z') + \int_{\mathbb{R}^n \times \mathbb{R}^n} K(z, z')\, \mathbb{Q}_2(\mathrm{d}z) \mathbb{Q}_2(\mathrm{d}z') - 2 \int_{\mathbb{R}^n \times \mathbb{R}^n} K(z, z')\, \mathbb{Q}_1(\mathrm{d}z) \mathbb{Q}_2(\mathrm{d}z') \bigg)^{\frac{1}{2}}.$$

Lemma 3.7 ([64])**.** The MMD distance d_{MMD} induced by K matches the IPM $\mathcal{D}_\Psi$ induced by the unit ball $\Psi = \{\psi \in \mathbb{H}_K : \|\psi\|_{\mathbb{H}_K} \leq 1\}$ in the reproducing kernel Hilbert space $\mathbb{H}_K$ corresponding to K.

If K is any kernel satisfying $\|z - z'\| = K(z, z) + K(z', z') - 2K(z, z')$, *e.g.*, if K is the distance-induced kernel (see Appendix C.1), then $2 \cdot d^2_{\mathrm{MMD}}$ reduces to the energy distance [59, Theorem 22]. Another popular IPM is the total variation distance.

Definition 3.8 (Total variation distance)**.** The total variation distance of $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{Q}_1(\mathbb{R}^n)$ is $\mathcal{TV}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{B \in \mathcal{B}(\mathbb{R}^n)} |\mathbb{Q}_1(B) - \mathbb{Q}_2(B)|$, where $\mathcal{B}(\mathbb{R}^n)$ is the Borel σ-algebra on $\mathbb{R}^n$.

By construction, the total variation distance is an IPM generated by the indicator functions of the Borel sets $B \in \mathcal{B}(\mathbb{R}^n)$. One can thus show that its maximal generator [45, Definition 3.1] is given by $\Psi = \{\psi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}) : \|\psi\|_{\mathcal{L}^\infty} \leq 1\}$. A summary of the discussed IPMs is given in Table 1.

TABLE 1. Generators of commonly used IPMs

IPM $\mathcal{D}_\Psi$	Generator Ψ
Kolmogorov Distance	$\{\psi \in \mathcal{L}(\mathbb{R}, \mathbb{R}) : \exists \tau \in \mathbb{R} \text{ s.t. } \psi(y) = \mathbb{1}_{y \leq \tau}\ \forall y \in \mathbb{R}\}$
Wasserstein Distance	$\{\psi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}) : \mathrm{Lip}(\psi) \leq 1\}$
$\mathcal{L}^p$-Distance	$\{\psi \in W_0^{1,q}(\mathbb{R}) : \|\psi'\|_{\mathcal{L}^q} \leq 1\}$, where $1/p + 1/q = 1$
$\sqrt{\text{Cramér Distance}}$	$\{\psi \in W_0^{1,2}(\mathbb{R}) : \|\psi'\|_{\mathcal{L}^2} \leq 1\}$
$\sqrt{\text{Energy Distance}}$	$\{\psi \in W_0^{1,2}(\mathbb{R}) : \|\psi'\|_{\mathcal{L}^2} \leq \sqrt{2}\}$
Maximum Mean Discrepancy	$\{\psi \in \mathbb{H}_K : \|\psi\|_{\mathbb{H}_K} \leq 1\}$
Total Variation Distance	$\{\psi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}) : \|\psi\|_{\mathcal{L}^\infty} \leq 1\}$

Different IPMs induce different unfairness measures. Observe first that each IPM $\mathcal{D}_\Psi$ of Table 1 satisfies the identity of indiscernibles. This means that $\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h(X)|A=1}) = 0$ if and only if $\mathbb{P}_{h(X)|A=0}$ coincides with $\mathbb{P}_{h(X)|A=1}$ or, put differently, if and only if h satisfies SP at level 0. Hence, all IPMs agree on what it means for a hypothesis to be perfectly fair. However, they attribute different levels of unfairness to hypotheses with $\mathbb{P}_{h(X)|A=0} \neq \mathbb{P}_{h(X)|A=1}$. Fix now an IPM $\mathcal{D}_\Psi$ defined on $\mathcal{Q}_w$, and assume that $\mathbb{P}_{h(X)} \in \mathcal{Q}_w$ for all $h \in \mathcal{H}$. In this case, the test functions $\psi \in \Psi$ can be viewed as utility functions, which are routinely used to model preferences under uncertainty [72]. Specifically, if the prediction $h(X)$ impacts the well-being of a person with feature vector X, *e.g.*, if $h(X)$ determines the person's salary, then the expected utility $\mathbb{E}[\psi(h(X))|A = a]$ quantifies the person's satisfaction with the prediction $h(X)$ if that person belongs to class $a \in \mathcal{A}$. We can now use a hypothetical experiment to introduce a notion of *utilitarian fairness*. Imagine that you are asked to assess a hypothesis h before birth, that is, before knowing any of your own personal traits such as your feature vector X, class A or utility function ψ. In this situation, it is natural to call h ε-fair if the expected utilities conditional on $A = 0$ and $A = 1$ differ at most by ε for any $\psi \in \Psi$, *i.e.*,

$$|\mathbb{E}[\psi(h(X))|A = 0] - \mathbb{E}[\psi(h(X))|A = 1]| \leq \varepsilon \quad \forall \psi \in \Psi. \tag{3}$$

This utilitarian perspective gives a physical interpretation to unfairness measures induced by IPMs.

Lemma 3.9 (Utilitarian fairness)**.** For any given hypothesis $h \in \mathcal{H}$ and any given IPM $\mathcal{D}_\Psi$ on $\mathcal{Q}_w$ such that $\mathbb{P}_{h(X)} \in \mathcal{Q}_w$, we have that $\mathcal{D}_\Psi(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}) \leq \varepsilon$ if and only if (3) holds.

In summary, any IPM $\mathcal{D}_\Psi$ can be used to measure the deviation from perfect statistical parity, and its generator Ψ can be viewed as a family of utility functions. The degree of unfairness that $\mathcal{D}_\Psi$ assigns to any given hypothesis thus always has a utilitarian interpretation. We conclude that, conceptually, there is no objective reason for preferring the Kolmogorov metric (which is used in the standard definition of SP in [1], for example) over other IPMs. In the remainder of the paper we will argue, however, that other IPMs have distinct computational advantages over the Kolmogorov metric.

4. Numerical Solution of Fair Learning Problems

From now on we assume that L is convex and that $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$, where Θ is an open subset of a Euclidean space, while the parametric hypothesis $h_\theta(x)$ is Lipschitz continuous in x for every fixed θ and Lipschitz continuous in θ for every fixed x. For example, h_θ could be a linear hypothesis with gradient θ or a multi-layer neural network whose weight matrices are encoded by θ. To find an optimal trade-off between prediction loss and unfairness, we can solve the fair learning problem

$$\min_{\theta\in\Theta} \; \mathbb{E}[L(h_\theta(X), Y)] + \mathcal{U}(h_\theta), \tag{4}$$

which differs from (1) only in that it's objective function involves an unfairness penalty in the form of $\mathcal{U}(h_\theta) = \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1}))$, where $\mathcal{D}_\Psi$ is an IPM and $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ is a smooth and non-decreasing regularization function. Below we will discuss how to solve problem (4) to local optimality when the distribution $\mathbb{P}_{(X,Y,A)}$ of X, Y and A is only indirectly observable through training samples.

4.1. Empirical Risk Minimization.

Assume that $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, is a stochastic process of independent and identically distributed (i.i.d.) training samples, all of which follow the probability distribution $\mathbb{P}_{(X,Y,A)}$. Define now $\tau_t^a \in \mathbb{N}$ for each $t \in \mathbb{N}$ and $a \in \mathcal{A}$ via the recursion $\tau_t^a = \inf\{i > \tau_{t-1}^a : \hat{A}_i = a\}$ initialized with $\tau_0^a = 0$. Thus, τ_t^a represents the index of the t-th sample in class a. Note that τ_t^a is a stopping time in the sense that the event $\{\tau_t^a = i\}$ belongs to the σ-algebra generated by $\hat{X}_1, \ldots, \hat{X}_i$ for any $i \in \mathbb{N}$. To exclude trivialities, we assume from now on that $\mathbb{P}[A = a] > 0$ for all $a \in \mathcal{A}$, which implies that τ_t^a is $\mathbb{P}$-almost surely finite. Next, define $\hat{X}_t^a = \hat{X}_{\tau_t^a}$ as the feature vector of the t-th sample in class a. By [15, Lemma 5.3.4], the random vectors $\hat{X}_t^a$, $t \in \mathbb{N}$, are i.i.d. for all $a \in \mathcal{A}$. A simple generalization of the same argument shows that these random vectors are independent across all $t \in \mathbb{N}$ and $a \in \mathcal{A}$. In addition, $\hat{X}_t^a$ follows the probability distribution $\mathbb{P}_{X|A=a}$ of X conditional on $A = a$ because

$$\begin{aligned}
\mathbb{P}\left[\hat{X}_t^a \leq x\right] = \mathbb{P}\left[\hat{X}_1^a \leq x\right] &= \sum_{i\in\mathbb{N}} \mathbb{P}\left[\hat{X}_i \leq x,\, \tau_1^a = i\right] \\
&= \sum_{i\in\mathbb{N}} \mathbb{P}\left[\hat{X}_i \leq x,\, \hat{A}_1 \neq a, \ldots, \hat{A}_{i-1} \neq a, \hat{A}_i = a\right] \\
&= \sum_{i\in\mathbb{N}} \mathbb{P}\left[X \leq x,\, A = a\right] \mathbb{P}\left[A \neq a\right]^{i-1} = \frac{\mathbb{P}\left[X \leq x,\, A = a\right]}{1 - \mathbb{P}[A \neq a]} = \mathbb{P}\left[X \leq x | A = a\right]
\end{aligned}$$

for all $x \in \mathcal{X}$, where the second equality exploits the law of total probability, the fourth equality holds because the original training samples are i.i.d., and the sixth equality follows from the observation that $1 - \mathbb{P}[A \neq a] = \mathbb{P}[A = a]$. Next, for any fixed $N \in \mathbb{N}$ define $T_N^a = \max_{t\geq 0}\{t : \tau_t^a \leq N\}$, which counts how many out of the first N training samples belong to class a. One can show that the random counters T_N^0 and T_N^1 are independent of $\hat{X}_t^a$ for all $t \in \mathbb{N}$ and $a \in \mathcal{A}$. Using these notational conventions, we then introduce the conditional empirical distributions

$$\hat{\mathbb{P}}_{N,\theta}^a = \frac{1}{T_N^a} \sum_{t=1}^{T_N^a} \delta_{h_\theta(\hat{X}_t^a)} \quad \forall a \in \mathcal{A},$$

where δ_z denotes the Dirac measure at $z \in \mathbb{R}$, and the empirical risk minimization problem

$$\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L(h_\theta(\hat{X}_i),\hat{Y}_i) + \rho\left(\mathcal{D}_\Psi(\hat{\mathbb{P}}^0_{N,\theta},\hat{\mathbb{P}}^1_{N,\theta})\right). \tag{ERM}$$

Problem (ERM) is generically non-convex even if the hypotheses depend linearly on θ, and even if L and ρ are convex. Indeed, if $\mathcal{D}_\Psi$ is the MMD distance and ρ is quadratic, for example, then one can show that the fairness penalty in (ERM) represents a sum of convex and concave functions of θ. The lack of convexity is of little concern if the hypotheses represent multi-layer neural networks, which already display a non-trivial dependence on θ. We therefore propose to solve problem (ERM) to local optimality via gradient descent-based algorithms. Problem (ERM) is indeed amenable to such algorithms for commonly used hypothesis classes (*e.g.* linear hypotheses or neural networks with piecewise linear activation functions) and standard loss functions provided that $\rho(\mathcal{D}_\Psi(\mathbb{Q}_1,\mathbb{Q}_2))$ is piecewise differentiable with respect to the support points of any discrete distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$.

4.2. **Stochastic Approximation.** The state-of-the-art method for solving empirical risk minimization problems over large datasets and complex hypothesis spaces is the stochastic gradient descent (SGD) algorithm [55] and its variants such as mini-batch SGD, Adam [39, 53, 86] or Adadelta [82]. All of these fast algorithms require *unbiased* estimators for the gradients of the objective function of problem (4). Unfortunately, the empirical risk itself (that is, the objective function of (ERM)) constitutes a *biased* estimator for the objective function of (4). This is a direct consequence of the following lemma.

Lemma 4.1 (The empirical unfairness penalty is biased)**.** For all $N \in \mathbb{N}$ and $\theta \in \Theta$ we have

$$\mathbb{E}\left[\rho\left(\mathcal{D}_\Psi(\hat{\mathbb{P}}^0_{N,\theta},\hat{\mathbb{P}}^1_{N,\theta})\right)\middle| T^0_N, T^1_N \geq 1\right] \geq \rho\left(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0},\mathbb{P}_{h_\theta(X)|A=1})\right).$$

The inequality is strict if ρ increases strictly and $\mathbb{P}_{h_\theta(X)|A=0} = \mathbb{P}_{h_\theta(X)|A=1}$ is no Dirac distribution.

Conditioning on $T^a_N \geq 1$ ensures that the empirical distribution $\hat{\mathbb{P}}^a_{N,\theta}$ is well-defined. Lemma 4.1 shows that the empirical risk provides a *biased* estimator for the true risk, which suggests that the gradients of the empirical risk provide *biased* estimators for the gradients of the true risk. Nevertheless, the fair statistical learning problem (4) is amenable to efficient SGD-type algorithms when $\mathcal{D}_\Psi = d_{\mathrm{MMD}}$ is an MMD metric and $\rho(z) = \lambda z^2$ is a quadratic penalty function with $\lambda \geq 0$. We will now show that, in this special case, one can construct unbiased gradient estimators by using *random* batches of training samples. To this end, we set $\tau_1 = 1$ and define τ_b for $b \geq 2$ recursively as the smallest integer satisfying $\tau_b - \tau_{b-1} \geq \bar{N}$ such that the set $\mathcal{I}_b = \{\tau_b, \dots, \tau_{b+1} - 1\}$ contains the indices of at least two training samples of each class $a \in \mathcal{A}$. By construction, $|\mathcal{I}_b|$ is not smaller than a given target batch size $\bar{N}$. Defining $\mathcal{I}^a_b = \{t \in \mathcal{I}_b : \hat{A}_t = a\}$, we further have $|\mathcal{I}^a_b| \geq 2$ for each $a \in \mathcal{A}$. Using a similar reasoning as in Section 4.1, one can show that the index τ_b of the first sample in the b-th batch is $\mathbb{P}$-almost surely finite and constitutes a stopping time for every $b \in \mathbb{N}$. Conditional on $|\mathcal{I}^a_b| = N^a$, one can also show that $\{\hat{X}_i : i \in \mathcal{I}^a_b\}$ is a family of N^a i.i.d. training features governed by $\mathbb{P}_{X|A=a}$ for every $a \in \mathcal{A}$. By construction, both the cardinality $|\mathcal{I}_b|$ of the b-th batch and the cardinality $|\mathcal{I}^a_b|$ of its subfamily corresponding to any class $a \in \mathcal{A}$ is random. Then,

$$\hat{U}_b(\theta) = \sum_{a\in\mathcal{A}} \sum_{i,j\in\mathcal{I}^a_b, i\neq j} \frac{K(h_\theta(\hat{X}_i),h_\theta(\hat{X}_j))}{|\mathcal{I}^a_b|(|\mathcal{I}^a_b|-1)} - 2\sum_{i\in\mathcal{I}^0_b, j\in\mathcal{I}^1_b} \frac{K(h_\theta(\hat{X}_i),h_\theta(\hat{X}_j))}{|\mathcal{I}^0_b||\mathcal{I}^1_b|}$$

is an unbiased estimator for the squared MMD distance between $\mathbb{P}_{h_\theta(X)|A=0}$ and $\mathbb{P}_{h_\theta(X)|A=1}$; see Proposition B.1. Note that if the number of training samples in each class $a \in \mathcal{A}$ was *deterministic* and not smaller than 2, then $\hat{U}_b(\theta)$ would reduce to a classical U-statistic and thus constitute the minimum variance unbiased estimator for the squared MMD distance [60, § 5]. Given a stream of i.i.d. training samples $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, however, there is always a positive probability that a batch of deterministic cardinality contains less than two samples of one class, in which case $\hat{U}_b(\theta)$ is not defined. Hence, working with batches of random cardinality

seems unavoidable to correct the bias in the empirical unfairness penalty. Unfortunately, this randomness introduces a bias in the empirical prediction error $|\mathcal{I}_b|^{-1}\sum_{i\in\mathcal{I}_b} L(h_\theta(\hat{X}_i),\hat{Y}_i)$; see Proposition B.3. While one could construct an unbiased estimator for the prediction error using only the first $\bar{N}$ training samples in the b-th batch, this would amount to sacrificing the last $|\mathcal{I}_b|-\bar{N}$ training samples and thus be data-inefficient. We circumvent this problem by introducing bias correction terms defined via the auxiliary function

$$\Delta(N,n) = \mathbb{1}_{\{N=\bar{N}\}} + \frac{N}{2(N-1)}\mathbb{1}_{\{N>\bar{N},\, n=2\}} + \frac{N}{N-1}\mathbb{1}_{\{N>\bar{N},\, n=N-2\}},$$

where $N\in\{\bar{N},\bar{N}+1,\ldots\}$ and $n\in\{2,\ldots,N-2\}$. Using this definition, we can prove that

$$\hat{R}_b(\theta) = \frac{1}{|\mathcal{I}_b|}\sum_{a\in\mathcal{A}}\sum_{i\in\mathcal{I}_b^a}\Delta(|\mathcal{I}_b|,|\mathcal{I}_b^a|)L(h_\theta(\hat{X}_i),\hat{Y}_i)$$

constitutes an unbiased estimator for $\mathbb{E}[L(h_\theta(X),Y)]$; see Proposition B.5. In summary, Propositions B.1 and B.5 imply that $\hat{R}_b(\theta)+\lambda\hat{U}_b(\theta)$ is an unbiased estimator for the objective function of the fair learning problem (4). Unbiased gradient estimators are available under the following assumption.

Assumption 4.2 (Uniform integrability)**.** The loss function L as well as the kernel function K are piecewise differentiable, and every $\theta\in\Theta$ has a neighborhood $\Theta_0\subseteq\Theta$ such that

$$\mathbb{E}\left[\sup_{\theta\in\Theta_0'}\left\|\nabla_\theta K(h_\theta(\hat{X}_1),h_\theta(\hat{X}_2))\right\|_2\right]<\infty \quad\text{and}\quad \mathbb{E}\left[\sup_{\theta\in\Theta_0'}\left\|\nabla_\theta L(h_\theta(\hat{X}_1),\hat{Y}_1)\right\|_2\right]<\infty,$$

where Θ_0' is the subset of Θ_0 on which the gradients exist.

Theorem 4.3 (Unbiased gradient estimators)**.** If $\mathcal{D}_\Psi = d_{\mathrm{MMD}}$ is an MMD metric, $\rho(z)=\lambda z^2$ is a quadratic penalty function with $\lambda\geq 0$ and Assumption 4.2 holds, then $\nabla_\theta\hat{R}_b(\theta)+\lambda\nabla_\theta\hat{U}_b(\theta)$ constitutes an unbiased estimator for the gradient of the objective function of problem (4).

In practice, one may use automated differentiation to evaluate the gradient estimators of Theorem 4.3 (*e.g.*, using Autograd in PyTorch or GradientTape in TensorFlow). As these estimators are unbiased, stochastic gradient descent-type algorithms are guaranteed to converge, in expectation, to a stationary point of problem (4) provided that the learning rates are chosen appropriately [30, 43].

5. Numerical Experiments

We now assess the accuracy-unfairness trade-offs and runtimes of different fair learning models on five standard datasets (Drug [26], Communities&Crime (CC) [54, 24], Compas [36], Adult [24] and Student Grades [18]). The algorithm for solving (4) is implemented in PyTorch [48], and all experiments are run on an Intel i7-10700 CPU (2.9GHz) computer with 32GB RAM. The corresponding codes are available at `https://github.com/RAO-EPFL/Metrizing-Fairness`.

Classification. In all classification experiments, we solve an instance of (4), where L is the cross entropy loss [44, § 16.5.4], $\mathcal{D}_\Phi = d_2$ and $\rho(z)=2\lambda z^2$. As $\sqrt{2}\cdot d_2$ coincides with the MMD metric corresponding to the distance-induced kernel, Theorem 4.3 implies that the objective of (4) admits unbiased gradient estimators. The hypothesis space $\mathcal{H}$ is set to the family of all neural networks with one hidden layer accommodating 16 nodes, ReLU (Rectified Linear Unit) activation functions at the hidden layer and sigmoid activation functions at the output layer. The resulting problem instance is solved to local optimality using the Adam optimizer [39] (*i.e,*, an accelerated SGD method). Below we refer to this approach as *metrized fair learning* (MFL). All hyperparameters are detailed in Appendix D. We compare our approach against three baselines: (i) a convexified empirical logistic regression model with relaxed fairness constraints proposed by Zafar et al. [80], (ii) an instance of (ERM) with cross entropy loss proposed by Cho et al. [12], which is solved with the Adam optimizer using biased gradient estimators, and (iii) an instance of (ERM) with cross entropy loss and a Sinkhorn divergence-based unfairness penalty proposed by Oneto et al. [47], which is solved via gradient

descent. Detailed information on these baselines is provided in Appendix D. In all models we sweep the unfairness penalty parameter λ from 10^{-5} to 10 in 25 equal steps on a log scale.

TABLE 2. Test AUC values (mean ± standard deviation) and training times averaged over 10 runs

Dataset	Performance	Zafar et al. [80]	Cho et al. [12]	Oneto et al. [47]	MFL (NN)
Drug	AUC	0.67±0.04	0.79±0.03	**0.83±0.01**	0.82±0.01
	Time	0.35 secs	10.54 secs	416.18 secs	6.66 secs
CC	AUC	0.79±0.01	0.82±0.01	0.79±0.01	**0.83±0.01**
	Time	81.05 secs	13.52 secs	217.27 secs	8.39 secs
Compas	AUC	0.63±0.02	**0.67±0.01**	**0.67±0.01**	0.66±0.01
	Time	0.44 secs	17.33 secs	2299.37 secs	9.99 secs
Adult	AUC	0.80±0.00	0.77±0.01	-	**0.84±0.00**
	Time	29.45 secs	141.35 secs	-	149.61 secs

Figure A.2 in Appendix D visualizes the accuracy-unfairness trade-off of the trained hypotheses on test data, where unfairness is measured by the Kolmogorov metric as in the basic definition of SP. Table 2 reports the areas under the respective accuracy-unfairness curves (AUC; see Appendix D for a precise definition) as well as the training times averaged over 10 independent replications of the same experiment with randomly permuted data. In order to check if the differences between average AUC values are significant, the reported standard deviations must be divided by $\sqrt{10}$ to obtain the standard deviations of the averages. The method by Oneto et al. [47] failed to train on the Adult dataset due to memory constraints. Note that MFL performs favorably vis-à-vis its competitors in that it often attains the highest AUC values, while training significantly faster than the other neural network-based methods (*i.e.*, those by Cho et al. [12] and Oneto et al. [47]). The method by Zafar et al. [80] relies on convex optimization and thus provides fast computation when the input dimension d is low, but it cannot learn non-linear input-output relations. As Oneto et al. [47] do not have access to unbiased gradients, they resort to classical gradient descent algorithms to solve the empirical risk minimization models. For large training datasets, the time necessary for training is thus no longer competitive.

Regression. When applied to regression problems, the MFL approach replaces the cross entropy loss with the the mean squared error. In neural network regression, $\mathcal{H}$ is set to the family of all neural networks with one hidden layer accommodating 20 nodes, ReLU activation functions at the hidden layer and linear activation functions at the output layer. In linear regression, on the other hand, $\mathcal{H}$ comprises all linear hypotheses. All other modeling choices remain as in the classification experiments; for details see Appendix D. We compare our MFL approach against two methods by Berk et al. [8], which train a linear regressor by solving a convex optimization problem implemented in CVXPY [2]. We sweep the unfairness penalty parameter λ in our MFL model from 10^{-5} to 10^{3} for the Student Math and Student Portugese datasets, and from 10^{-5} to 10^{2} for the CC dataset, all in 50 equal steps on a log scale. Similarly, in the methods by Berk et al. [8], we sweep the unfairness penalty parameter λ from 10^{-2} to 10^{5} in 50 equal steps on a log scale.

Figure A.4 in Appendix D visualizes the trade-off between the goodness of fit (measured by the coefficient of determination R^2) and the unfairness (measured by the Kolmogorov metric) of the trained hypotheses on test data. Table 3 reports the corresponding AUC values and training times averaged over 10 independent simulation runs. We observe that MFL is at least one order of magnitude faster than the baseline methods and consistently attains superior AUC values.

TABLE 3. Test AUC values (mean ± standard deviation) for the regression problems over 10 runs

Dataset	Performance	Berk et al. [8] (individual)	Berk et al. [8] (group)	MFL (Linear)	MFL (NN)
Student Math	AUC	0.53±0.13	0.43±0.18	0.62±0.10	**0.63±0.11**
	Time	1.48 secs	0.21 secs	4.07 secs	4.69 secs
Student Portugese	AUC	0.68±0.04	0.64±0.09	**0.73±0.04**	0.70±0.05
	Time	4.67 secs	1.10 secs	6.12 secs	7.06 secs
CC	AUC	0.36±0.03	0.50±0.03	0.52±0.02	**0.54±0.02**
	Time	710.23 secs	168.49 secs	11.39 secs	16.54 secs

6. CONCLUDING REMARKS

In this paper, we show that the generator of an IPM can be identified with a family of utility functions, which implies that IMPs have intuitive appeal as fairness regularizers that promote statistical parity. We also demonstrate that particular IPMs such as the $\mathcal{L}^2$-distance, the square root of the energy distance or any MMD metric are computationally attractive because they lead to fair learning models that are susceptible to efficient batch SGD-type algorithms. These algorithms often outperform state-of-the-art methods for fair learning on standard datasets. We believe that our intuitive and scalable approach to fair learning is easy to use, thus hopefully boosting the uptake of fair learning models. At the same time, we acknowledge certain risks and limitations of the results and methods presented in this paper. For example, more research is required to design fair learning processes that are immunized against adversarial attacks and missing data. This research might benefit from the observation that the severity of adversarial attacks can also be quantified by IPMs [61]. In addition, we observed that the performance of a fair predictor can be highly sensitive to the choice of the unfairness penalty parameter λ. For a nuanced discussion on how to tune λ via a two-step validation process we refer to [22]. Finally and most importantly, it should be kept in mind that not every application may require fairness constraints or that the imposed fairness constraints may have undesirable effects. Example A.3 in Appendix A describes a scenario in which the imposition of statistical parity does not necessarily improve the well-being of the members of the protected group. Therefore, when deploying fair learning models in the real world, one should be mindful about the choice of the most suitable fairness regularizer. For a critical assessment of arguments for and against statistical parity we refer to [52]. We emphasize that the methods developed in this paper readily extend to other notions of fairness such as equal opportunity [33] or equalized odds [33].

Appendix

This appendix is organized as follows. Section A describes stylized examples that offer additional insights and complement our theoretical results, Section B contains all proofs omitted from the main text, Section C provides background information on distance-induced kernels and highlights a few straightforward generalizations of our main results, whereas Section D details the datasets as well as the training processes used in the numerical experiments and reports on additional simulation results.

Appendix A. Examples

We first show that enforcing SP can increase predictive accuracy when training on a biased target.

Example A.1 (SP increases accuracy)**.** Consider a classification problem of the form (1) with feature space $\mathcal{X} = \{0,1\}^2$, label space $\mathcal{Y} = \{0,1\}$, 0-1 loss function $L(\hat{y}, y) = \mathbb{1}_{\hat{y}-y}$ and hypothesis space $\mathcal{H} = \mathcal{L}(\mathcal{X}, \mathcal{Y})$, where $X_2 = A$ is a protected attribute. We assume that X_1 and X_2 are independent and uniformly distributed on $\{0,1\}$ and that the goal is to predict the true target $Y^\star$, which satisfies

$$\mathbb{P}[Y^\star = 1|X] = \begin{cases} 0.3 & \text{if } X_1 = 0, \\ 0.7 & \text{if } X_1 = 1. \end{cases}$$

This implies that $\mathbb{E}[Y^\star|X] \perp A$. By Theorem 2.2, the optimal predictor of $Y^\star$ thus satisfies SP. Next, assume that Y is a biased target that approximates $Y^\star$, and let the conditional distribution of Y given X be determined by the success probabilities in Table A.4a. As $\mathbb{E}[Y|X] = \mathbb{P}[Y = 1|X] \not\perp A$, the optimal predictor of Y may violate SP. Assume now that we have infinitely many training samples for the biased target Y but none for the true target $Y^\star$. Hence, the optimal predictor of Y can be computed, whereas the one of $Y^\star$ is inaccessible. A simple calculation shows that the unique optimal solution $h^\star$ of problem (1) with target Y is completely characterized by the information in Table A.4b.

Table A.4. Success probabilities and optimal decisions for Example A.1

(a) $\mathbb{P}[Y = 1|X]$

	$X_1 = 0$	$X_1 = 1$
$X_2 = 1$	0.3	0.7
$X_2 = 0$	0.4	0.4

(b) $h^\star(X)$

	$X_1 = 0$	$X_1 = 1$
$X_2 = 1$	0	1
$X_2 = 0$	0	0

Note that $\mathbb{P}[h^\star(X) = 1|A = 0] = 0$ but $\mathbb{P}[h^\star(X) = 1|A = 1] = 0.5$, which confirms that $h^\star$ violates SP. Next, we solve a 'fair' variant of (1) with the extra constraint (2) for $\varepsilon = 0$, which restricts attention to hypotheses that are perfectly fair with respect to the SP criterion. A tedious but routine calculation reveals that the optimal SP-fair hypothesis is $h^\star_{\mathrm{SP}}(x) = x_1$. Clearly, enforcing fairness must deteriorate the accuracy of the optimal predictor in view of the biased target Y. A direct calculation shows indeed that $\mathbb{E}[L(h^\star_{\mathrm{SP}}(X), Y)] = 0.4 > 0.35 = \mathbb{E}[L(h^\star(X), Y)]$. However, enforcing fairness improves the accuracy of the optimal predictor in view of the true target $Y^\star$. Another direct calculation shows indeed that $\mathbb{E}[L(h^\star_{\mathrm{SP}}(X), Y^\star)] = 0.3 < 0.4 = \mathbb{E}[L(h^\star(X), Y^\star)]$. In summary, if (non-statistical) domain knowledge suggests that the protected attribute A carries no information about the distribution of $Y^\star$ conditional on X, then we know that the true optimal prediction must satisfy SP. If we are forced to train in view of a biased target Y because we have no training samples for $Y^\star$, we can obtain a better prediction of $Y^\star$ by enforcing SP. This is plausible because enforcing SP amounts to enforcing a known property of the true optimal (but inaccessible) predictor. We emphasize that fairness constraints and unfairness penalties do not always improve accuracy in view of $Y^\star$. However, empirical evidence suggests that this is often the case [73].

Next, we show that $Y \perp A$ does not guarantee that the optimizer of problem (1) satisfies SP.

TABLE A.5. Success probabilities and optimal decisions for Example A.2

(A) $\mathbb{P}[Y=1|X]$

	$X_1=0$	$X_1=1$
$X_2=1$	0.4	0.9
$X_2=0$	0.55	0.3

(B) $h^\star(X)$

	$X_1=0$	$X_1=1$
$X_2=1$	0	1
$X_2=0$	1	0

Example A.2 ($Y \perp A$ does not imply SP)**.** Consider a classification problem of the form (1) with feature space $\mathcal{X}=\{0,1\}^2$, label space $\mathcal{Y}=\{0,1\}$, 0-1 loss function $L(\hat{y},y)=\mathbb{1}_{\hat{y}-y}$ and hypothesis space $\mathcal{H}=\mathcal{L}(\mathcal{X},\mathcal{Y})$, where $X_2=A$ is a protected attribute. Assume that X_1 is independent of A and that its marginal distribution is given by $\mathbb{P}[X_1=0]=0.8$ and $\mathbb{P}[X_1=1]=0.2$. The distribution of A is irrelevant for this example. Finally, the conditional distribution of Y given X is completely determined by the success probabilities in Table A.5a. Based on this information, it is easy to verify that

$$\mathbb{P}[Y=1|A=a]=\sum_{x_1\in\{0,1\}}\mathbb{P}[Y=1|X_1=x_1,\,A=a]\,\mathbb{P}[X_1=x_1]=0.5\quad\forall a\in\mathcal{A},$$

and thus we have $Y \perp A$. Similarly, we find

$$\mathbb{P}[\,\mathbb{P}[Y=1|X]=0.4\,|\,A=a]=\begin{cases}0.2 & \text{if } a=1,\\ 0 & \text{if } a=0,\end{cases}$$

which is sufficient to imply that $\mathbb{P}[Y=1|X] \not\perp A$. Thus, Theorem 2.2 does not apply. Thanks to its simplicity, problem (1) can be solved analytically, and its unique optimal solution $h^\star$ is fully characterized by the information in Table A.5b. Next, one readily verifies that

$$\mathbb{P}[h^\star(X)=1|A=a]=\mathbb{P}[X_1=a]=\begin{cases}0.2 & \text{if } a=1,\\ 0.8 & \text{if } a=0,\end{cases}$$

which reveals that $h^\star(X) \not\perp A$. Hence, $h^\star$ fails to satisfy SP even though Y is independent of A.

While $Y \perp A$ does *not* induce SP, we know from Theorem 2.2 that the condition $\mathbb{P}[Y=1|X] \perp A$ is sufficient to induce SP. From Example A.2 it is thus clear that $Y \perp A$ does *not* imply $\mathbb{P}[Y=1|X] \perp A$. Conversely, $\mathbb{P}[Y=1|X] \perp A$ does also *not* imply $Y \perp A$ in general. To see this, just assume that $Y=A$ and that X is independent of Y. In this case $\mathbb{P}[Y=1|X] \perp A$ is satisfied, but $Y \perp A$ is not.

Finally, we show that enforcing SP can reduce both accuracy and fairness by any reasonable standard.

Example A.3 (Enforcing SP can reduce accuracy and fairness)**.** Consider a least squares regression problem of the form (1) with $L(\hat{y},y)=(\hat{y}-y)^2$ that aims to predict the skill level $Y\in[0,1]$ of a job candidate based on a feature vector $X=(X_1,X_2)$, where $X_1\in[0,1]$ and $X_2=A\in\{0,1\}$ represent the candidate's normalized college GPA and age group, respectively. For example, suppose that $X_2=1$ if the candidate's age is at most 40 years and that $X_2=0$ otherwise. The hypothesis space $\mathcal{H}$ comprises all Borel-measurable functions h from $\mathcal{X}=[0,1]\times\{0,1\}$ to $\mathcal{Y}=[0,1]$. A correct and fair prediction of the skill level is critical because it will determine the candidate's salary. In the following, we set $p_a=\mathbb{P}[A=a]$ for all $a\in\mathcal{A}$, and we assume that the skill level satisfies $Y=AX_1+(1-A)S$, where S denotes the candidate's work experience, which is not observed. Thus, the skill level matches the GPA for junior candidates and the work experience for senior candidates. We also assume that both X_1 and S are uniformly distributed on $[0,1]$ and that X_1, S and A are mutually independent. Hence, the distribution of Y conditional on $A=a$ coincides with the uniform distribution on $[0,1]$ irrespective of $a\in\mathcal{A}$, that is, Y is independent of A. Moreover, one can show that the optimal value of (1) amounts to $p_0/12$, which is uniquely attained by the hypothesis

$$h^\star(x)=\mathbb{E}[Y|X=x]=\begin{cases}\frac{1}{2} & \text{if } x_2=a=0,\\ x_1=y & \text{if } x_2=a=1.\end{cases}$$

This confirms that, even though the protected attribute A is independent of Y, the optimal hypothesis may display a non-trivial dependence on A. As the conditional distributions of $h^\star(X)$ differ across the two age groups, the SP criterion is violated. Next, we solve a 'fair' variant of (1) with the extra constraint (2) for $\varepsilon = 0$, which restricts attention to hypotheses that are perfectly fair with respect to the SP criterion. A tedious but routine calculation reveals that the optimal SP-fair hypothesis is

$$h^\star_{\mathrm{SP}}(x) = \mathbb{E}[Y|X_1 = x_1] = \tfrac{1}{2} + p_1(x_1 - \tfrac{1}{2}),$$

whose objective function value in (1) exceeds that of $h^\star$ because

$$\mathbb{E}[L(h^\star_{\mathrm{SP}}(X), Y)] = \tfrac{p_0}{12} \cdot \tfrac{1}{p_1}\left[1 + 10p_1 - 9p_1^2 + 4p_1^3\right] + \tfrac{p_0^2 p_1}{12} > \tfrac{p_0}{12} = \mathbb{E}[L(h^\star(X), Y)].$$

Note that under $h^\star_{\mathrm{SP}}$ the predicted skill level grows with the GPA independent of the protected attribute. Thus, the distribution of $h^\star_{\mathrm{SP}}(X)$ conditional on $A = a$ coincides with the uniform distribution on the interval $[\frac{1}{2}(1-p_1), \frac{1}{2}(1+p_1)]$ irrespective of $a \in \mathcal{A}$, indicating that the representatives of the two age groups have the same prospects of being hired into a particular salary band. This confirms that $h^\star_{\mathrm{SP}}(X)$ is indeed fair according to SP. Nevertheless, the prediction $h^\star_{\mathrm{SP}}(X)$ does not seem more fair than $h^\star(X)$ by any reasonable standard. While $h^\star_{\mathrm{SP}}(X)$ is perfectly correlated with the skill level of junior candidates, it is independent of the skill level of senior applicants. In other words, $h^\star_{\mathrm{SP}}$ enforces SP by making purely random predictions that have no bearing on the actual qualifications of senior applicants, which suggests that their earning a high salary is tantamount to winning a lottery. Under the hypothesis $h^\star$ that minimizes the prediction loss without constraints, on the other hand, all senior candidates are treated equally, which seems more 'fair' than affording them a random salary.

Appendix B. Proofs

This section contains all proofs omitted from the main text and provides several auxiliary results.

Proof of Theorem 2.2. By the law of iterated conditional expectations, problem (1) is equivalent to

$$\min_{h \in \mathcal{H}} \mathbb{E}[\mathbb{E}[L(h(X), Y)|X]].$$

Recall now that the loss function L is lower semi-continuous, which implies via Fatou's Lemma that the conditional expectation $\mathbb{E}[L(\hat{y}, Y)|X = x]$ is lower semi-continuous in $\hat{y}$. Recall also that the hypothesis space $\mathcal{H} = \mathcal{L}(\mathcal{X}, \mathbb{R})$ contains all real-valued Borel-measurable functions, and problem (1) is assumed to be solvable. Therefore, [56, Theorem 14.60] implies that problem (1) is equivalent to

$$\mathbb{E}\left[\min_{\hat{y} \in \mathbb{R}} \mathbb{E}[L(\hat{y}, Y)|X]\right]$$

and that it admits an optimal solution $h^\star \in \mathcal{H}$ that satisfies

$$h^\star(X) = \arg\min_{\hat{y} \in \mathbb{R}} \mathbb{E}[L(\hat{y}, Y)|X] \quad \mathbb{P}\text{-a.s.} \tag{A.5}$$

Next, observe that the objective function

$$\mathbb{E}[L(\hat{y}, Y)|X] \;=\; \mathbb{P}[Y = 1|X]\, L(\hat{y}, 1) + (1 - \mathbb{P}[Y = 1|X])\, L(\hat{y}, 0)$$

of the parametric optimization problem (A.5) depends on X only indirectly through the conditional probability $\mathbb{P}[Y = 1|X]$, which is assumed to be independent of the protected attribute A. Thus, the optimal prediction $h^\star(X)$ depends on X only indirectly through $\mathbb{P}[Y = 1|X]$, too. Using $\mathrm{I}(Z; Z')$ to denote the mutual information between two random variables Z and Z', we then find

$$0 \leq \mathrm{I}(h^\star(X); A) \leq \mathrm{I}(\mathbb{P}[Y = 1|X]; A) = 0,$$

where the second inequality exploits the data processing inequality, and the equality holds because $\mathbb{P}[Y = 1|X]$ and A are independent. Thus, we may conclude that $\mathrm{I}(h^\star(X); A) = 0$, which implies that $h^\star(X)$ and A are also independent. This observation completes the proof. □

Proof of Lemma 3.9. The claim is an immediate consequence of Definition 3.1 and the relation

$$\mathbb{E}[\psi(h(X))|A=a] = \int_{\mathbb{R}} \psi(\hat{y})\, \mathbb{P}_{h(X)|A=a}(\mathrm{d}\hat{y}) \quad \forall a \in \mathcal{A},$$

which follows from the variable transformation $\hat{y} \leftarrow f(\omega)$, where $f : \Omega \to \mathbb{R}$ is defined through $f(\omega) = h(X(\omega))$ for all $\omega \in \Omega$, and the observation that $\mathbb{P}_{h(X)|A=a}$ is the pushforward measure of the conditional probability measure $\mathbb{P}[\cdot|A=a]$ under the transformation f. □

Proof of Lemma 4.1. To avoid clutter, we use $\mathbb{P}_\theta^a$ as a shorthand for $\mathbb{P}_{h_\theta(X)|A=a}$, $a \in \mathcal{A}$. As the penalty function ρ is convex and non-decreasing, a repeated application of Jensen's inequality yields

$$\begin{aligned}
\mathbb{E}\left[\rho\left(\mathcal{D}_\Psi(\hat{\mathbb{P}}^0_{N,\theta}, \hat{\mathbb{P}}^1_{N,\theta})\right) \middle| T_N^0, T_N^1 \geq 1\right] &\geq \rho\left(\mathbb{E}\left[\mathcal{D}_\Psi(\hat{\mathbb{P}}^0_{N,\theta}, \hat{\mathbb{P}}^1_{N,\theta}) \middle| T_N^0, T_N^1 \geq 1\right]\right) \\
&= \rho\left(\mathbb{E}\left[\sup_{\psi\in\Psi} \left| \int_{\mathbb{R}} \psi(z)\, \hat{\mathbb{P}}^0_{N,\theta}(\mathrm{d}z) - \int_{\mathbb{R}} \psi(z)\, \hat{\mathbb{P}}^1_{N,\theta}(\mathrm{d}z) \right| \middle| T_N^0, T_N^1 \geq 1\right]\right) \\
&\geq \rho\left(\sup_{\psi\in\Psi} \left| \mathbb{E}\left[\int_{\mathbb{R}} \psi(z)\, \hat{\mathbb{P}}^0_{N,\theta}(\mathrm{d}z) - \int_{\mathbb{R}} \psi(z)\, \hat{\mathbb{P}}^1_{N,\theta}(\mathrm{d}z) \middle| T_N^0, T_N^1 \geq 1\right]\right|\right) \\
&= \rho\left(\sup_{\psi\in\Psi} \left| \mathbb{E}\left[\frac{1}{T_N^0}\sum_{t=1}^{T_N^0} \psi(h_\theta(\hat{X}_t^0)) - \frac{1}{T_N^1}\sum_{t=1}^{T_N^1} \psi(h_\theta(\hat{X}_t^0)) \middle| T_N^0, T_N^1 \geq 1\right]\right|\right) \\
&= \rho\left(\sup_{\psi\in\Psi} \left| \int_{\mathbb{R}} \psi(z)\, \mathbb{P}^0_\theta(\mathrm{d}z) - \int_{\mathbb{R}} \psi(z)\, \mathbb{P}^1_\theta(\mathrm{d}z) \right|\right) = \rho\left(\mathcal{D}_\Psi(\mathbb{P}^0_\theta, \mathbb{P}^1_\theta)\right),
\end{aligned}$$

where the first two equalities follow from the definitions of the IPM $\mathcal{D}_\Psi$ and the empirical distributions $\hat{\mathbb{P}}^0_{N,\theta}$ and $\hat{\mathbb{P}}^1_{N,\theta}$, respectively, while the third equality exploits the independence of $\hat{X}_t^a$ and T_N^a for every $t \in \mathbb{N}$ and $a \in \mathcal{A}$. Thus, the desired inequality follows. It remains to be shown that this inequality is strict if ρ is strictly increasing and if $\mathbb{P}_\theta^1 = \mathbb{P}_\theta^0$ and is *not* a Dirac distribution. In this case, as $T_N^0 + T_N^1 = N$, the law of total expectation implies that

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{D}_\Psi(\hat{\mathbb{P}}^0_{N,\theta}, \hat{\mathbb{P}}^1_{N,\theta}) \middle| T_N^0, T_N^1 \geq 1\right] \\
&\quad = \sum_{t_0=1}^{N-1} \mathbb{E}\left[\mathcal{D}_\Psi(\hat{\mathbb{P}}^0_{N,\theta}, \hat{\mathbb{P}}^1_{N,\theta}) \middle| T_N^0 = t_0\right] \mathbb{P}[T_N^0 = t_0] > 0 = \mathcal{D}_\Psi(\mathbb{P}^0_\theta, \mathbb{P}^1_\theta),
\end{aligned}$$

where the inequality holds because all terms in the sum on the second line are strictly positive. Indeed, we have $\mathbb{P}[T_N^0 = t_0] > 0$ for every $t_0 = 1, \ldots, N-1$ because of our assumption that $\mathbb{P}[A=a] > 0$ for all $a \in \mathcal{A}$. Similarly, as the support of the data-generating distributions $\mathbb{P}_\theta^1 = \mathbb{P}_\theta^0$ is not a singleton, the distance between the empirical distributions $\hat{\mathbb{P}}^0_{N,\theta}$ and $\hat{\mathbb{P}}^1_{N,\theta}$ conditional on $T_N^0 = t_0$ is strictly positive with a strictly positive probability for every $t_0 = 1, \ldots, N-1$. The second equality in the above expression holds because $\mathbb{P}_\theta^1 = \mathbb{P}_\theta^0$. The claim now follows because ρ is strictly increasing. □

To prove Theorem 4.3, we first show that $\hat{U}_b(\theta)$ constitutes an *unbiased* estimator for the unfairness regularizer in the objective function of problem (4) (Proposition B.1). This result critically relies on the randomized construction of the b-th batch of training samples. Next, we demonstrate that the empirical prediction loss $|\mathcal{I}_b|^{-1}\sum_{i\in\mathcal{I}_b} L(h_\theta(\hat{X}_i), \hat{Y}_i)$ constitutes a *biased* estimator for the true expected prediction loss in the objective function of problem (4) (Lemma B.2 and Proposition B.3). This is another (yet undesirable) consequence of the randomized construction of the b-th batch of training samples. Finally, we prove that $\hat{R}_b(\theta)$ constitutes an *unbiased* estimator for the expected prediction loss (Lemma B.4 and Proposition B.5). This result critically relies on the construction of the bias correction term $\Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)$, which is needed to counteract the effects of the randomness of the b-th batch. Theorem 4.3 is finally proved by combining Proposition B.1 and Proposition B.5.

Proposition B.1. The statistic $\hat{U}_b(\theta)$ constitutes an unbiased estimator for $d^2_{MMD}(\mathbb{P}_{X|A=0}, \mathbb{P}_{X|A=1})$.

Proof. As in the proof of Lemma 4.1, we use $\mathbb{P}_\theta^a$ as a shorthand for $\mathbb{P}_{h_\theta(X)|A=a}$ for every $a \in \mathcal{A}$ in order to avoid clutter. To prove that $\hat{U}_b(\theta)$ is an unbiased estimator for $d^2_{\mathrm{MMD}}(\mathbb{P}^0_\theta, \mathbb{P}^1_\theta)$, first note that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{|\mathcal{I}_b^a|(|\mathcal{I}_b^a|-1)} \sum_{i,j\in\mathcal{I}_b^a,\, i\neq j} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j))\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\frac{1}{|\mathcal{I}_b^a|(|\mathcal{I}_b^a|-1)} \sum_{i,j\in\mathcal{I}_b^a,\, i\neq j} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j))\,\middle|\, |\mathcal{I}_b^a|\right]\right] \\
&\quad = \int_{\mathbb{R}^n\times\mathbb{R}^n} K(h_\theta(x), h_\theta(x'))\, \mathbb{P}_\theta^a(\mathrm{d}x)\, \mathbb{P}_\theta^a(\mathrm{d}x') \quad \forall a \in \mathcal{A},
\end{aligned}$$

where the second equality holds because, conditional on $|\mathcal{I}_b^a|$, the samples $\hat{X}_i$, $i \in \mathcal{I}_b^a$, are independent and governed by the distribution $\mathbb{P}_\theta^a$. Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{2}{|\mathcal{I}_b^0||\mathcal{I}_b^1|} \sum_{i\in\mathcal{I}_b^0,\, j\in\mathcal{I}_b^1} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j))\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\frac{2}{|\mathcal{I}_b^0||\mathcal{I}_b^1|} \sum_{i\in\mathcal{I}_b^0,\, j\in\mathcal{I}_b^1} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j))\,\middle|\, |\mathcal{I}_b^0|,\, |\mathcal{I}_b^1|\right]\right] \\
&\quad = 2\int_{\mathbb{R}^n\times\mathbb{R}^n} K(h_\theta(x), h_\theta(x'))\, \mathbb{P}_\theta^0(\mathrm{d}x)\, \mathbb{P}_\theta^1(\mathrm{d}x').
\end{aligned}$$

The claim then follows directly from the definitions of $\hat{U}_b(\theta)$ and $d^2_{\mathrm{MMD}}(\mathbb{P}^0_\theta, \mathbb{P}^1_\theta)$. □

In the following we use p_a as a shorthand for the marginal probability $\mathbb{P}[A = a]$ for each $a \in \mathcal{A}$ in order to avoid clutter. The next lemma is needed to prove that the empirical prediction loss is biased.

Lemma B.2. Fix any $a \in \mathcal{A} = \{0, 1\}$, and define $a' = 1 - a$. Then, the statistic $|\mathcal{I}_b^a|/|\mathcal{I}_b|$ constitutes a biased estimator for the class probability p_a, that is, we have $\mathbb{E}\left[|\mathcal{I}_b^a|/|\mathcal{I}_b|\right] = p_a(1 + \beta_a)$, where

$$\beta_a = p_{a'}^{\bar{N}-1} - p_{a'} p_a^{\bar{N}-2} + \frac{2p_a}{p_{a'}^2}\left(\log(p_a) + \sum_{N=1}^{\bar{N}} \frac{p_{a'}^N}{N}\right) - \frac{2p_{a'}^2}{p_a^3}\left(\log(p_{a'}) + \sum_{N=1}^{\bar{N}} \frac{p_a^N}{N}\right).$$

Proof. We first characterize the joint distribution of the two random variables $|\mathcal{I}_b|$ and $|\mathcal{I}_b^a|$. As the batch $\mathcal{I}_b$ must contain at least $\bar{N}$ samples in total, it is clear that its cardinality $|\mathcal{I}_b|$ may only adopt an integer value $N \geq \bar{N}$. In addition, as $\mathcal{I}_b$ contains at least two samples of each class, it is also clear that $|\mathcal{I}_b^a|$ may only adopt an integer value $n = 2, \ldots, N-2$. The probabilities $\mathbb{P}[|\mathcal{I}_b| = N, |\mathcal{I}_b^a| = n]$ of the possible scenarios (N, n) can be calculated as follows. If $N = \bar{N}$, then we have

$$\mathbb{P}[|\mathcal{I}_b| = \bar{N},\, \mathcal{I}_b^a = n] = \begin{cases} \binom{\bar{N}}{n} p_a^n p_{a'}^{\bar{N}-n} & \text{if } n = 2, \ldots, \bar{N}-2, \\ 0 & \text{otherwise.} \end{cases} \tag{A.6a}$$

Indeed, the probability that the batch $\mathcal{I}_b$ contains n samples of class a and $\bar{N} - n$ samples of class a' in a particular order is $p_a^n p_{a'}^{\bar{N}-n}$, and there are $\binom{\bar{N}}{n}$ possibilities of ordering these samples. Assume next that the batch $\mathcal{I}_b$ contains $N > \bar{N}$ samples. In this case we have

$$\mathbb{P}[|\mathcal{I}_b| = N,\, \mathcal{I}_b^a = n] = \begin{cases} (N-1)\, p_a^2\, p_{a'}^{N-2} & \text{if } n = 2, \\ (N-1)\, p_a^{N-2}\, p_{a'}^2 & \text{if } n = N-2, \\ 0 & \text{otherwise.} \end{cases} \tag{A.6b}$$

Note that the batch cardinality $N = |\mathcal{I}_b|$ can only exceed $\bar{N}$ if the first $\bar{N}$ samples do not include at least two representatives of each class. If there is none or only one sample of class a among the first $\bar{N}$ samples, for example, we continue to add samples until the batch includes exactly two samples of class a. Thus, the last (*i.e.*, the N-th) sample of the batch must belong to class a, and there must be exactly one other sample

belonging to class a. This other sample can reside in any of the first $N-1$ positions in the batch. Thus, the probability of the event $|\mathcal{I}_b^a|=2$ is $(N-1)p_a^2 p_{a'}^{N-2}$. If there is none or only one sample of class a' among the first $\bar{N}$ samples, then a similar reasoning applies.

We are now ready to calculate the expected value of the estimator $|\mathcal{I}_b^a|/|\mathcal{I}_b|$. By (A.6), we have

$$\begin{aligned}\mathbb{E}\left[\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\right] &= \sum_{N=\bar{N}}^{\infty}\sum_{n=2}^{N-2}\frac{n}{N}\mathbb{P}[|\mathcal{I}_b|=N,\,|\mathcal{I}_b^a|=n] \\ &= \sum_{n=2}^{\bar{N}-2}\frac{n}{\bar{N}}\,\mathbb{P}\left[|\mathcal{I}_b=\bar{N}|,\,|\mathcal{I}_b^a|=n\right] + \sum_{N=\bar{N}+1}^{\infty}\frac{2}{N}\,\mathbb{P}\left[|\mathcal{I}_b=N|,\,|\mathcal{I}_b^a|=2\right] \\ &\qquad + \sum_{N=\bar{N}+1}^{\infty}\frac{N-2}{N}\,\mathbb{P}\left[|\mathcal{I}_b=N|,\,|\mathcal{I}_b^a|=N-2\right].\end{aligned} \tag{A.7}$$

Next, we may use (A.6a) to reformulate the first sum in (A.7) as

$$\begin{aligned}\sum_{n=2}^{\bar{N}-2}\frac{n}{\bar{N}}\,\mathbb{P}\left[|\mathcal{I}_b=\bar{N}|,\,|\mathcal{I}_b^a|=n\right] &= \sum_{n=2}^{\bar{N}-2}\frac{n}{\bar{N}}\cdot\frac{\bar{N}!}{n!(\bar{N}-n)!}\,p_a^n p_{a'}^{\bar{N}-n} \\ &= p_a\sum_{n=2}^{\bar{N}-2}\binom{\bar{N}-1}{n-1}p_a^{n-1}p_{a'}^{\bar{N}-n} = p_a\sum_{n=1}^{\bar{N}-3}\binom{\bar{N}-1}{n}p_a^{n}p_{a'}^{\bar{N}-1-n} \\ &= p_a\left(1-p_a^{\bar{N}-1}-p_{a'}^{\bar{N}-1}-(\bar{N}-1)p_a^{\bar{N}-2}p_{a'}\right),\end{aligned}$$

where the last equality follows from the binomial expansion of $(p_a+p_{a'})^{\bar{N}-1}=1$. Smilarly, we may then use (A.6b) to reformulate the second sum in (A.7) as

$$\begin{aligned}\sum_{N=\bar{N}+1}^{\infty}\frac{2}{N}\,\mathbb{P}\left[|\mathcal{I}_b=N|,\,|\mathcal{I}_b^a|=2\right] &= \sum_{N=\bar{N}+1}^{\infty}\frac{2(N-1)}{N}\,p_a^2 p_{a'}^{N-2} \\ &= \frac{2p_a^2}{p_{a'}^2}\left(\sum_{N=\bar{N}+1}^{\infty}p_{a'}^N - \sum_{N=\bar{N}+1}^{\infty}\frac{p_{a'}^N}{N}\right) \\ &= \frac{2p_a^2}{p_{a'}^2}\left(\frac{p_{a'}^{\bar{N}+1}}{1-p_{a'}} - \left(\sum_{N=1}^{\infty}\frac{p_{a'}^N}{N} - \sum_{N=1}^{\bar{N}}\frac{p_{a'}^N}{N}\right)\right) \\ &= \frac{2p_a^2}{p_{a'}^2}\left(\frac{p_{a'}^{\bar{N}+1}}{1-p_{a'}} + \log(1-p_{a'}) + \sum_{N=1}^{\bar{N}}\frac{p_{a'}^N}{N}\right) \\ &= \frac{2p_a^2}{p_{a'}^2}\left(\frac{p_{a'}^{\bar{N}+1}}{p_a} + \log(p_a) + \sum_{N=1}^{\bar{N}}\frac{p_{a'}^N}{N}\right),\end{aligned} \tag{A.8}$$

where the third equality exploits the standard formula for infinite geometric series, which applies because $p_{a'}<1$, while the fourth equality follows from power series representation of $\log(1-p_{a'})$ (*i.e.*, the Newton-Mercator series). The last equality holds because $p_a = 1-p_{a'}$. Finally, we may use (A.6a) once again to reformulate the third sum in (A.7) as

$$\begin{aligned}\sum_{N=\bar{N}+1}^{\infty}\frac{N-2}{N}\,\mathbb{P}\left[|\mathcal{I}_b=N|,\,|\mathcal{I}_b^a|=N-2\right] &= \sum_{N=\bar{N}+1}^{\infty}\frac{N-2}{N}(N-1)\,p_a^{N-2}p_{a'}^2 \\ &= \sum_{N=\bar{N}+1}^{\infty}(N-1)\,p_a^{N-2}p_{a'}^2 - \sum_{N=\bar{N}+1}^{\infty}\frac{2}{N}(N-1)\,p_a^{N-2}p_{a'}^2.\end{aligned} \tag{A.9}$$

The first sum in (A.9) is equivalent to

$$\sum_{N=\bar{N}+1}^{\infty}(N-1)p_a^{N-2}p_{a'}^2 = p_{a'}^2\,\frac{\mathrm{d}}{\mathrm{d}p_a}\left(\sum_{N=\bar{N}+1}^{\infty}p_a^{N-1}\right) = p_{a'}^2\frac{\mathrm{d}}{\mathrm{d}p_a}\left(\sum_{N=\bar{N}}^{\infty}p_a^{N}\right)$$
$$= p_{a'}^2\,\frac{\mathrm{d}}{\mathrm{d}p_a}\left(\frac{p_a^{\bar{N}}}{1-p_a}\right) = \bar{N}p_{a'}p_a^{\bar{N}-1} + p_a^{\bar{N}},$$

where we use again the standard formula for the geometric series. By swapping the roles of a and a' and then repeating the derivations in (A.8), one can further show that the second term in (A.9) equals

$$\sum_{N=\bar{N}+1}^{\infty}\frac{2}{N}(N-1)p_a^{N-2}p_{a'}^2 = \frac{2p_{a'}^2}{p_a^2}\left(\frac{p_a^{\bar{N}+1}}{p_{a'}} + \log(p_{a'}) + \sum_{N=1}^{\bar{N}}\frac{p_a^N}{N}\right).$$

Substituting the formulas for the different sums into (A.7) finally yields

$$\mathbb{E}\left[\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\right] = p_a\left(1 - p_a^{\bar{N}-1} - p_{a'}^{\bar{N}-1} - (\bar{N}-1)p_a^{\bar{N}-2}p_{a'}\right) + \frac{2p_a^2}{p_{a'}^2}\left(\frac{p_{a'}^{\bar{N}+1}}{p_a} + \log(p_a) + \sum_{N=1}^{\bar{N}}\frac{p_{a'}^N}{N}\right)$$
$$+ \bar{N}p_{a'}p_a^{\bar{N}-1} + p_a^{\bar{N}} - \frac{2p_{a'}^2}{p_a^2}\left(\frac{p_a^{\bar{N}+1}}{p_{a'}} + \log(p_{a'}) + \sum_{N=1}^{\bar{N}}\frac{p_a^N}{N}\right) = p_a(1+\beta_a),$$

where β_a is defined as in the statement of the lemma. □

Armed with Lemma B.2, we are now prepared to show that the empirical prediction loss with respect to all the samples in the (random) batch $\mathcal{I}_b$ constitutes a biased estimator for the true expected loss.

Proposition B.3. The empirical prediction loss $|\mathcal{I}_b|^{-1}\sum_{i\in\mathcal{I}_b}L(h_\theta(\hat{X}_i),\hat{Y}_i)$ constitutes a biased estimator for the true expected loss $\mathbb{E}[L(h_\theta(X),Y)]$, that is, defining β_a as in Lemma B.2, we have

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}_b|}\sum_{i\in\mathcal{I}_b}L(h_\theta(\hat{X}_i),\hat{Y}_i)\right] = \mathbb{E}[L(h_\theta(X),Y)] + \sum_{a\in\mathcal{A}}\beta_a\cdot p_a\cdot\mathbb{E}[L(h_\theta(X),Y)|A=a].$$

Proof. The expected empirical loss can be reformulated as

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}_b|}\sum_{i\in\mathcal{I}_b}L(h_\theta(\hat{X}_i),\hat{Y}_i)\right] = \mathbb{E}\left[\frac{1}{|\mathcal{I}_b|}\mathbb{E}\left[\sum_{i\in\mathcal{I}_b}L(h_\theta(\hat{X}_i),\hat{Y}_i)\Big||\mathcal{I}_b^0|,|\mathcal{I}_b^1|\right]\right]$$
$$= \mathbb{E}\left[\sum_{a\in\mathcal{A}}\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\mathbb{E}\left[\sum_{i\in\mathcal{I}_b^a}\frac{1}{|\mathcal{I}_b^a|}L(h_\theta(\hat{X}_i),\hat{Y}_i)\Big||\mathcal{I}_b^a|\right]\right]$$
$$= \sum_{a\in\mathcal{A}}\mathbb{E}\left[\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\right]\mathbb{E}[L(h_\theta(X),Y)|A=a],$$

where the first equality follows from law of iterated conditional expectations, whereas the second equality holds because the sets $\mathcal{I}_b^a$, $a\in\mathcal{A}$, form a partition of $\mathcal{I}_b$. The third equality exploits our earlier insight that, conditional on fixing the number $|\mathcal{I}_b^a|$ of samples from class a in the current batch, $\{(\hat{X}_i,\hat{Y}_i)\}_{i\in\mathcal{I}_b^a}$ represent independent samples from $\mathbb{P}_{(X,Y)|A=a}$, implying that the sample average $|\mathcal{I}_b^a|^{-1}\sum_{i\in\mathcal{I}_b^a}L(h_\theta(\hat{X}_i),\hat{Y}_i)$ constitutes an unbiased estimator for $\mathbb{E}[L(h_\theta(X),Y)|A=a]$. Recall now from Lemma B.2 that $\mathbb{E}[|\mathcal{I}_b^a|/|\mathcal{I}_b|] = p_a(1+\beta_a)$. Thus, we have

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}_b|}\sum_{i\in\mathcal{I}_b}L(h_\theta(\hat{X}_i),\hat{Y}_i)\right] = \sum_{a\in\mathcal{A}}p_a\,(1+\beta_a)\,\mathbb{E}[L(h_\theta(X),Y)|A=a],$$

and this observation completes the proof. □

Next, we show that the empirical class probability $|\mathcal{I}_b^a|/|\mathcal{I}_b|$ admits an explicit correction term that eliminates the bias identified in Lemma B.2. To this end, we define $\Delta(N,n)$ as in the main paper.

Lemma B.4. For any fixed $a \in \mathcal{A}$, the adjusted empirical class probability $\Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}$ constitutes an unbiased estimator for the class probability p_a.

Proof. Recall from the proof of Lemma B.2 that $|\mathcal{I}_b|$ can adopt any integer value $N \geq \bar{N}$. In addition, recall that if $N = \bar{N}$, then $|\mathcal{I}_b^a|$ can adopt any integer value $n \in \{2, \ldots, \bar{N}-2\}$ and that if $N > \bar{N}$, then $|\mathcal{I}_b^a|$ can adopt only one of the two integer values $n \in \{2, \bar{N}-2\}$. By the definition of $\Delta(N,n)$, the expected value of the adjusted empirical class probability thus satisfies

$$\mathbb{E}\left[\Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\right] = \sum_{n=2}^{\bar{N}-2} \frac{n}{\bar{N}}\,\mathbb{P}\left[|\mathcal{I}_b = \bar{N}|, |\mathcal{I}_b^a| = n\right] + \sum_{N=\bar{N}+1}^{\infty} \frac{1}{N-1}\,\mathbb{P}\left[|\mathcal{I}_b = N|, |\mathcal{I}_b^a| = 2\right] + \sum_{N=\bar{N}+1}^{\infty} \frac{N-2}{N-1}\,\mathbb{P}\left[|\mathcal{I}_b = N|, |\mathcal{I}_b^a| = N-2\right]. \tag{A.10}$$

For ease of notation, we henceforth use a' as a shorthand for $1-a$. From the proof of Lemma B.2 we already know that the first sum in (A.10) evaluates to $p_a(1 - p_a^{\bar{N}-1} - p_{a'}^{\bar{N}-1} - (\bar{N}-1)p_a^{\bar{N}-2}p_{a'})$. By (A.6b), we may then reformulate the second sum in (A.10) as

$$\sum_{N=\bar{N}+1}^{\infty} \frac{1}{N-1}\,\mathbb{P}\left[|\mathcal{I}_b = N|, |\mathcal{I}_b^a| = 2\right] = p_a^2 \sum_{N=\bar{N}+1}^{\infty} p_{a'}^N = p_a^2 \frac{p_{a'}^{\bar{N}-1}}{1-p_{a'}} = p_a p_{a'}^{\bar{N}-1},$$

where the second equality exploits the formula for infinite geometric series, which applies because $p_{a'} < 1$, and the third equality follows from the observation that $p_{a'} = 1 - p_a$. Using (A.6b) once again, the third sum in (A.10) can be re-expressed as

$$\begin{aligned}\sum_{N=\bar{N}+1}^{\infty} \frac{N-2}{N-1}\,\mathbb{P}\left[|\mathcal{I}_b = N|, |\mathcal{I}_b^a| = N-2\right] &= p_{a'}^2 \sum_{N=\bar{N}+1}^{\infty} (N-2)p_a^{N-2} \\ &= p_{a'}^2\, p_a \frac{\mathrm{d}}{\mathrm{d}p_a}\left(\sum_{N=\bar{N}+1}^{\infty} p_a^{N-2}\right) = p_{a'}^2\, p_a \frac{\mathrm{d}}{\mathrm{d}p_a}\left(\frac{p_a^{\bar{N}-1}}{1-p_a}\right) = (\bar{N}-1)p_{a'}\, p_a^{\bar{N}-1} + p_a^{\bar{N}},\end{aligned}$$

where the third equality follows again from the formula for geometric series. Substituting the formulas for the different sums into (A.10) finally shows that $\mathbb{E}[\Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)|\mathcal{I}_b^a|/|\mathcal{I}_b|] = p_a$. □

Using Lemma B.4, we can now show that the adjusted empirical prediction loss with respect to all the samples in the (random) batch $\mathcal{I}_b$ constitutes an unbiased estimator for the true expected loss.

Proposition B.5. The statistic $\hat{R}_b(\theta)$ constitutes an unbiased estimator for $\mathbb{E}[L(h_\theta(X), Y)]$.

Proof. The expectation of the adjusted empirical prediction loss satisfies

$$\begin{aligned}\mathbb{E}\left[\sum_{a\in\mathcal{A}} \Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)\frac{1}{|\mathcal{I}_b|}\sum_{i\in\mathcal{I}_b^a} L(h_\theta(\hat{X}_i), \hat{Y}_i)\right] &= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\,\mathbb{E}\left[\frac{1}{|\mathcal{I}_b^a|}\sum_{i\in\mathcal{I}_b^a} L(h_\theta(\hat{X}_i), \hat{Y}_i)\middle| |\mathcal{I}_b^0|, |\mathcal{I}_b^1|\right]\right] \\ &= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)\frac{|\mathcal{I}_b^a|}{|\mathcal{I}_b|}\,\mathbb{E}[L(h_\theta(X), Y)|A = a]\right] \\ &= \sum_{a\in\mathcal{A}} \mathbb{P}[A = a]\,\mathbb{E}[L(h_\theta(X), Y)|A = a] = \mathbb{E}[L(h_\theta(X), Y)],\end{aligned}$$

where the first equality follows from law of iterated conditional expectations and the observation that the sets $\mathcal{I}_b^a$, $a \in \mathcal{A}$, form a partition of $\mathcal{I}_b$. The second equality exploits our earlier insight that, conditional on fixing the number $|\mathcal{I}_b^a|$ of samples from class a in the current batch, $\{(\hat{X}_i, \hat{Y}_i)\}_{i \in \mathcal{I}_b^a}$ represent independent samples from $\mathbb{P}_{(X,Y)|A=a}$, implying that the sample average $|\mathcal{I}_b^a|^{-1} \sum_{i \in \mathcal{I}_b^a} L(h_\theta(\hat{X}_i), \hat{Y}_i)$ constitutes an unbiased estimator for $\mathbb{E}[L(h_\theta(X), Y)|A = a]$. Finally, the forth equality holds thanks to Lemma B.4. Thus, the claim follows. □

We are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. Propositions B.1 and B.5 imply that $\hat{R}_b(\theta) + \lambda \hat{U}_b(\theta)$ represents an unbiased estimator for the objective function of the fair learning problem (4). To prove the theorem statement, it thus remains to be shown that gradients and expectations can be interchanged. However, this follows immediately from [32, Lemma 1], which applies thanks to Assumption 4.2. □

Appendix C. Discussions

This section contains useful background information on kernels and discusses possible extensions of our main results. Specifically, in Section C.1 we briefly elucidate the connection between the energy distance and the family of MMD metrics, and in Section C.2 we sketch extensions of the proposed IPM-based fairness regularizers to other fairness criteria beyond statistical parity.

C.1. Distance-Induced Kernel.

Fix any norm $\|\cdot\|$ on $\mathbb{R}^n$, and set $w(z) = 1 + \|z\|$. The energy distance with respect to $\|\cdot\|$ between two probability distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{Q}_w(\mathbb{R}^n)$ is defined as [3, 66]

$$\mathcal{E}(\mathbb{Q}_1, \mathbb{Q}_2) = 2 \int_{\mathbb{R}\times\mathbb{R}} \|z - z'\| \, \mathbb{Q}_1(\mathrm{d}z) \, \mathbb{Q}_2(\mathrm{d}z') - \int_{\mathbb{R}\times\mathbb{R}} \|z - z'\| \, \mathbb{Q}_1(\mathrm{d}z) \, \mathbb{Q}_1(\mathrm{d}z') - \int_{\mathbb{R}\times\mathbb{R}} \|z - z'\| \, \mathbb{Q}_2(\mathrm{d}z) \, \mathbb{Q}_2(\mathrm{d}z').$$

Recall that the univariate energy distance (for $n = 1$) is intimately connected to the $\mathcal{L}^2$-distance through the relation $\mathcal{E}(\mathbb{Q}_1, \mathbb{Q}_2) = 2d_2^2(\mathbb{Q}_1, \mathbb{Q}_2)$ [65, Theorem 1]. In addition, the multivariate energy distance (for $n \geq 1$) can also be expressed in terms of an MMD for a suitable choice of the underlying kernel [59, Theorem 2]. To keep this paper self-contained, we derive this expression below.

Definition C.1 (Distance-induced kernel [59])**.** The distance-induced kernel $K \in \mathcal{L}(\mathbb{R}^n \times \mathbb{R}^n, \mathbb{R})$ corresponding to the norm $\|\cdot\|$ on $\mathbb{R}^n$ and the anchor point $z_0 \in \mathbb{R}^n$ is given by

$$K(z, z') = \frac{1}{2} \left(\|z - z_0\| + \|z' - z_0\| - \|z - z'\| \right).$$

The distance-induced kernel is positive definite if and only if the norm $\|\cdot\|$ is of negative type ([7, Lemma 2.1]), such as the Euclidean norm [59, Proposition 3]. In addition, it satisfies the identity $\|z - z'\| = K(z, z) + K(z', z') - 2K(z, z')$ irrespective of the anchor point z_0. This implies that

$$\begin{aligned}
\mathcal{E}(\mathbb{Q}_1, \mathbb{Q}_2) &= 2 \int_{\mathbb{R}^n\times\mathbb{R}^n} K(z, z) + K(z', z') - 2K(z, z') \, \mathbb{Q}_1(\mathrm{d}z) \, \mathbb{Q}_2(\mathrm{d}z') \\
&\quad - \int_{\mathbb{R}^n\times\mathbb{R}^n} K(z, z) + K(z', z') - 2K(z, z') \, \mathbb{Q}_1(\mathrm{d}z) \, \mathbb{Q}_1(\mathrm{d}z') \\
&\quad - \int_{\mathbb{R}^n\times\mathbb{R}^n} K(z, z) + K(z', z') - 2K(z, z') \, \mathbb{Q}_2(\mathrm{d}z) \, \mathbb{Q}_2(\mathrm{d}z') \\
&= 2 \int_{\mathbb{R}^n\times\mathbb{R}^n} K(z, z') \, \mathbb{Q}_1(\mathrm{d}z) \, \mathbb{Q}_1(\mathrm{d}z') + 2 \int_{\mathbb{R}^n\times\mathbb{R}^n} K(z, z') \, \mathbb{Q}_2(\mathrm{d}z) \, \mathbb{Q}_2(\mathrm{d}z') \\
&\quad - 4 \int_{\mathbb{R}^n\times\mathbb{R}^n} K(z, z') \, \mathbb{Q}_1(\mathrm{d}z) \, \mathbb{Q}_2(\mathrm{d}z')
\end{aligned}$$

$$= 2 \cdot d^2_{\mathrm{MMD}}(\mathbb{Q}_1, \mathbb{Q}_2),$$

where d_{MMD} is the MMD corresponding to K. This derivation shows that the distance-induced kernel corresponding to any fixed anchor point and, in fact, any mixture of distance-induced kernels corresponding to different anchor points generates the energy distance.

C.2. **Extensions to other Fairness Criteria.** The ideas of this paper can be generalized along several dimensions. For example, statistical parity may be enforced at the level of the losses instead of the output distributions. In this case, statistical parity requires the conditional distributions $\mathbb{P}_{L(h(X),Y)|A=a}$ to be similar across all $a \in \mathcal{A}$. In addition, our techniques readily extend to popular fairness notions specific to classification problems such as equal opportunity [33], probabilistic equal opportunity [51] or log-probabilistic equal opportunity [67]. These fairness criteria impose similarity of the distributions $\mathbb{P}_{h(X)|Y=1,A=a}$ across all $a \in \mathcal{A}$. Conditioning on $Y = 1$ requires no new ideas but means that only the positive training samples are used to estimate the fairness of a hypothesis. Another fairness criterion used in classification is equalized odds [33], which asks the distributions $\mathbb{P}_{h(X)|Y=y,A=a}$ to be similar across all $a \in \mathcal{A}$ for every fixed $y \in \mathcal{Y}$. This criterion can be dealt with by introducing separate unfairness penalties for all $y \in \mathcal{Y}$. Last but not least, our methods can also be used if the protected attribute has more than two realizations or if there is more than one protected attribute. In this case it is again expedient to introduce multiple unfairness penalties. Details are omitted for brevity of exposition.

Appendix D. Experiments

In this section we provide descriptions of the datasets underlying our numerical experiments (Section D.1), details of the training processes (Section D.2) and a definition of the AUC metric that we use to quantify the accuracy-unfairness trade-off (Section D.3). Finally, we provide further details on the numerical results of Section 5 and report on additional numerical experiments (Section D.4).

D.1. Datasets.

- **Drug**:[1] This dataset contains records for 1885 respondents. Each respondent is described by 12 features including personality type, level of education, age, gender, country of residence and ethnicity. Additionally, each respondent's self-declared drug usage history is recorded. That is, for each drug the respondents declare the time of the last consumption (possible responses are: never, over a decade ago, in the last decade/year/month/week, or on the last day). The classification task is to predict the response 'never used' versus 'others' (*i.e.*, 'used') for heroin, and we treat race as the protected attribute. A detailed description of the dataset is given in [26].
- **Communites&Crime (CC)**:[2] This dataset contains socio-economic, law enforcement, and crime data for 1,994 different communities in the US described by 99 attributes. We use this dataset both for regression and classification experiments. The regression task is to predict the number of violent crimes per 100,000 residents, and the classification task is to predict whether the incidence of violent crime in a community exceeds the national average. As in [11], we create a binary protected attribute by thresholding the percentage of black residents at the median across all communities.
- **Compas**:[3] This dataset contains records on 10,000 criminal defendants in Broward County, Florida. The features correspond to the information used by the popular COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm to predict recidivism within 2 years of the original offense. Specifically, each defendant is described by 7 features including age, gender and prior offenses as well as race, which we use as the protected attribute. The classification task is to predict whether a defendant

[1] https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
[2] http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime
[3] https://www.kaggle.com/danofer/compass

will reoffend. The full dataset consists of three subsets, and we only use the one that focuses on violent recividism, which comprises 6,172 records.

- **Adult**:[4] This dataset comprises 45,222 records for different individuals described by 14 features including age, work class, education, marital status, gender, race and yearly income. The classification task is to predict whether or not a person's income exceeds 50,000$ per year. We use gender as the protected attribute. The dataset is split into fixed training and test sets that contain 32,561 and 12,661 records, respectively. We use these prescribed training and test sets for our experiments.
- **Student Grades**:[5] This dataset contains information about the academic performance and about demographic, social and school-related features of students of two Portuguese high schools. After converting all categorical features into binary variables via one-hot encoding, each student is described by 40 features including gender as the protected attribute. The regression task is to predict the final grade (on a scale from 0 to 20). From the full dataset we extract two partial datasets used to predict the final grade in Portuguese language (Student Portuguese, comprising 649 records) and in Mathematics (Student Math, comprising 395 records), respectively.

All datasets are randomly partitioned into training and test sets containing 75% and 25% of the samples, respectively. The only exception is the Adult dataset, where this partition is predefined.

D.2. **Specifics of the Training Processes. Classification.** This section provides additional information on the training processes of the proposed MFL method as well as the three baseline methods by Zafar et al. [80], Cho et al. [12] and Oneto et al. [47]. For all datasets, we use thresholding at 0.5 for converting scores to labels.

TABLE A.6. Hyperparameters of the MFL method used in the classification experiments

Dataset	Drug	Communities&Crime	Compas	Adult
Batch Size	128	128	2,048	2,048
(β_1, β_2) (Adam)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Learning rate (LR)	5e-4	5e-4	5e-4	5e-4
LR decay factor	0.99	0.99	0.99	0.99
Number of epochs	500	500	500	500

- The code of the MFL method is available from https://github.com/RAO-EPFL/Metrizing-Fairness. Table A.6 lists all hyperparameters.
- The code of the method by Zafar et al. [80] can be downloaded from https://github.com/mbilalzafar/fair-classification. This method is at a disadvantage because it focuses on linear hypotheses to ensure that the training problem constitutes a convex quadratic program.
- The code of the method by Cho et al. [12] is available from https://proceedings.neurips.cc/paper/2020/file/ac3870fcad1cfc367825cda0101eee62-Supplemental.zip. The underlying hypothesis space is the class of all neural networks with one hidden layer accommodating 16 nodes with ReLU activation functions. Note that our MFL method relies on the exact same hypothesis space. For the Adult and Compas datasets, we adopt the hyperparameters proposed in [12, Supplementary Material, § 5.2]. For the Adult dataset, we thus set the learning rate to 1e-1 with a decay rate of 0.98, the batch size to 512 and the number of epochs to 200, and for the Compas dataset we set the learning rate to 5e-4 without decay, the batch size to 2,048 and the number of epochs to 500. For all other datasets not considered in [12], we use the default hyperparameters predefined in the code, that is, we set the learning rate to 2e-4 without decay, the batch size

[4] http://archive.ics.uci.edu/ml/datasets/Adult
[5] https://archive.ics.uci.edu/ml/datasets/student+performance

to 2,048 and the number of epochs to 500. Throughout all experiments, we use the default Adam parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). All of these parameters are chosen to ensure best comparability with MFL.

- Since there is no publicly available code of the method by Oneto et al. [47], we implemented and calibrated it ab initio. As in the MFL method, we define the hypothesis space as the class of neural networks with one hidden layer accommodating 16 nodes with ReLU activation functions, and we use the cross-entropy loss for training. In addition, we set the regularization parameter of the Sinkhorn divergence to 0.1. To ensure a fair comparison with MFL, which uses no regualizers in addition to the unfairness penalty, we further set the weight of the Tikhonov regularization term to 0. Finally, we set the total number of iterations of the gradient descent algorithm to 500, initialize the learning rate as 0.1 and set the decay factor of the learning rate to 0.99. For full implementation details see https://github.com/RAO-EPFL/Metrizing-Fairness.

Regression. Here we provide additional information on the training processes of the proposed MFL method as well as the two baseline methods by Berk et al. [8].

TABLE A.7. Hyperparameters of the MFL method used in the regression experiments

Dataset	Communities&Crime	Math	Portugese
Batch Size	128	128	128
(β_1, β_2) (Adam)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Learning rate (LR)	1e-4	1e-3	1e-3
LR decay factor	None	None	None
Number of epochs	1,000	2,000	2,000

- The code of the proposed MFL method is available from https://github.com/RAO-EPFL/Metrizing-Fairness. Table A.7 lists all hyperparameters.
- Since there is no publicly available code for the two methods by Berk et al. [8], we reimplemented and calibrated them ab initio using CVXPY [2]. To ensure a fair comparison with MFL, which uses no regualizers in addition to the unfairness penalty, we set the weight of the Tikhonov regularization term to 0. Full implementation details can be found in our GitHub repository https://github.com/RAO-EPFL/Metrizing-Fairness.

D.3. **AUC Metric.** Each method for fair statistical learning considered in this paper involves an accuracy-unfairness trade-off parameter. The accuracy-SP plots shown in Section D.4.1 below are obtained by sweeping this parameter in equal steps on a logarithmic scale. The most appropriate search grid may depend on the method and the dataset at hand (concrete specifications are given in the main paper). At each grid point we compute an optimal hypothesis on the training set and evaluate its accuracy (correct classification rate for classification tasks or coefficient of determination for regression tasks) as well as its SP-unfairness measure (using the Kolmogorov distance as in Definition 2.1) on the test set. The resulting tuples are conveniently represented as points in the unfairness-accuracy plane; see the gray dots in Figure A.1 for a schematic. By the definitions of the correct classification rate, the coefficient of determination and the SP-unfairness measure based on the Kolmogorov distance, all of these points must reside inside the unit box $[0, 1]^2$. The red area visualizes all Pareto-dominated points, which attain a smaller accuracy as well as a higher unfairness than at least one of the gray dots, and its boundary can be interpreted as the corresponding Pareto frontier. The *area under the curve* (AUC) is then defined as the size of the red area, which is necessarily a number in $[0, 1]$. For an ideal classifier, the red

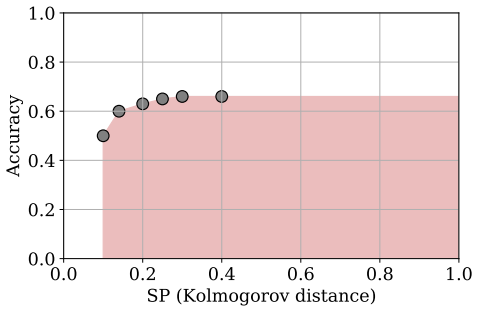
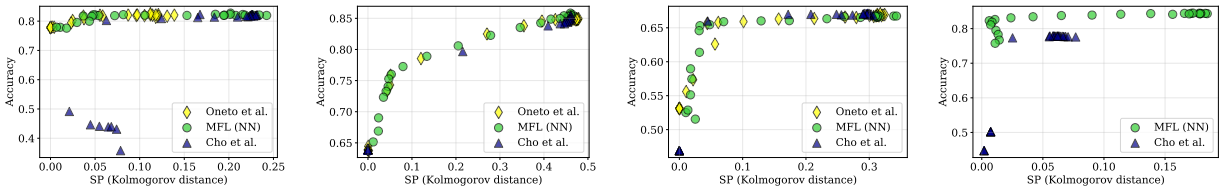


FIGURE A.1. Definition of the AUC value



(A) Drug (B) Communities&Crime (C) Compas (D) Adult

FIGURE A.2. Accuracy-unfairness trade-off for neural network-based classification models

area would span the whole graph, which corresponds to an AUC value of 1. Such a classifier would attain perfect accuracy at zero unfairness. We use the function `sklearn.metrics.auc` from the Scikit-learn toolbox in Python [10] to compute the AUC values.

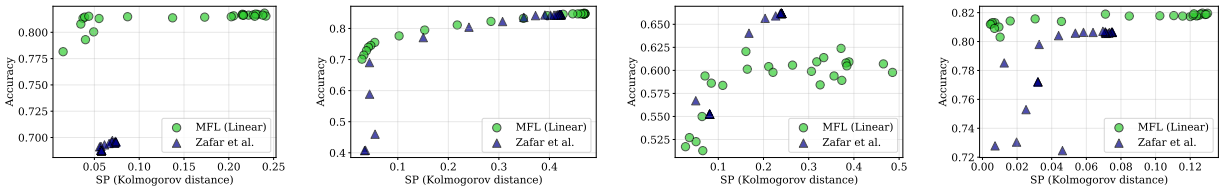
D.4. **Additional Numerical Results.** Section 4 in the main text describes how problem (4) can be solved to local optimality in an *online setting*, that is, when we have access to an infinite stream of training samples drawn independently from $\mathbb{P}_{(X,A,Y)}$. In practice, however, we are often given a finite training dataset in tabular form. In this *offline setting*, the best we can hope for is to solve the empirical version of problem (4), where $\mathbb{P}_{(X,A,Y)}$ corresponds to the discrete empirical distribution on the given training samples. The methods of Section 4 can readily be applied to find a local minimizer of this empirical learining problem because an infinite stream of independent samples from $\mathbb{P}_{(X,A,Y)}$ can be generated by randomly drawing samples from the given dataset. However, the empirical version of problem (4) can also be solved with a simpler SGD algorithm that uses batches of a fixed deterministic size. This is possible because, instead of generating a single data stream containing samples of both protected classes in a random succession, we can generate two separate data streams, each containing exclusively samples from one protected class. Note that the class probabilities $p_a = \mathbb{P}[A = a]$ for $a \in \mathcal{A}$ are easy to compute under the empirical distribution by simply counting the numbers of samples of each class a in the entire training dataset. It is then natural to construct batches of a deterministic size $\bar{N}$ that contain exactly $\lceil p_0 \bar{N} \rceil$ samples of class 0 and $\bar{N} - \lceil p_0 \bar{N} \rceil$ samples of class 1. In this case the batch cardinalities $|\mathcal{I}_b|$, $|\mathcal{I}_b^0|$ and $|\mathcal{I}_b^1|$ become deterministic, and one readily verifies that the gradient estimator derived in Section 4 remains unbiased.

All experiments based on real tabular datasets use the offline SGD algorithm to solve problem (4) (see Sections 5 and D.4.1). In contrast, the experiments based on synthetic data sampled from a prescribed continuous probability distribution use the online SGD algorithm (see Section D.4.2).

D.4.1. *Offline Learning.* This section provides additional information on the numerical results reported in Section 5. Figures A.2 and A.3 visualizes the trade-off between accuracy and unfairness of different fair classification methods on test data. Here, accuracy is measured by the correct classification rate, and unfairness is measured via the Kolmogorov distanced as in the standard definition of SP. All results are averaged over ten independent replications of the SGD algorithm with random seeds. All methods compared in Figure A.2 optimize over the same family of neural networks. Recall that the method by Zafar et al. [80] can only optimize over linear classifiers and is thus at a disadvantage *vis-à-vis* the neural network-based methods. In Figure A.3, we thus compare the method by Zafar et al. [80] separately against a variant of MFL that optimizes only over linear classifiers.

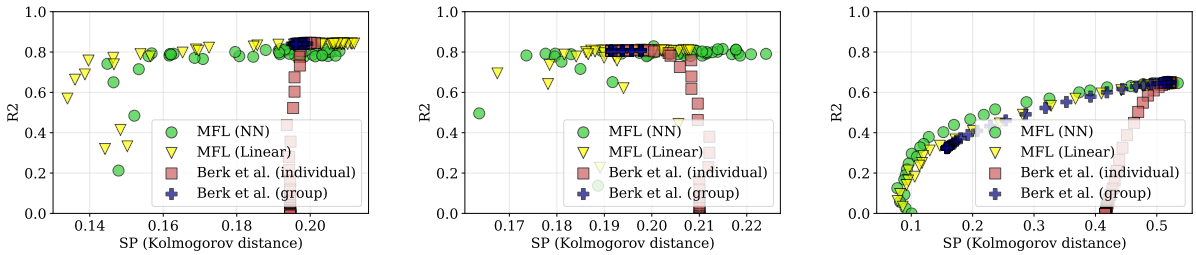
Figure A.4 visualizes the trade-off between accuracy and unfairness of different fair regression methods on test data. Here, accuracy is measured by the coefficient of determination. Otherwise, we use the same setup and conventions as in the classification experiments. Recall that both methods by Berk et al. [8] can only optimize over linear regressors. We thus compare them against two variants of the MFL approach with linear as well as neural network-based regressors, respectively.

D.4.2. *Online Learning.* In the context of online learning, training samples corresponding to the two protected classes arrive in a random order, in which case the construction of unbiased gradient estimators is tricky. We



(A) Drug (B) Communities&Crime (C) Compas (D) Adult

FIGURE A.3. Accuracy-unfairness trade-off for linear classification models



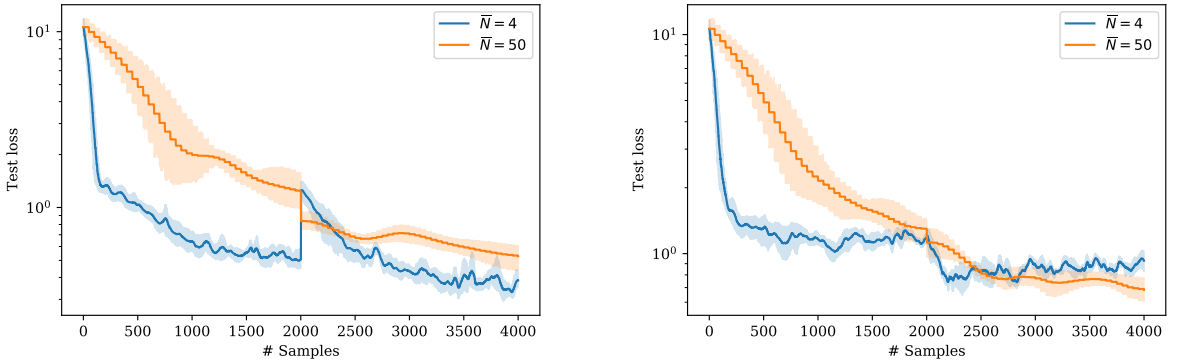
(A) Student Portugese (B) Student Math (C) Communities&Crime

FIGURE A.4. Accuracy-unfairness trade-off for linear and neural network-based regression models

will now show that even slightly biased gradient estimators can have a detrimental effect on the performance of SGD-type algorithms. Specifically, we will compare MMD-based unfairness penalties, which allow for unbiased gradient esimators, against Wasserstein-based unfairness penalties, for which we are unable to construct unbiased gradient estimators. Note that even though our gradient estimators for the Wasserstein-based unfairness penalties are biased, increasing the batch size will lead to an aggregation of a larger number of samples and thereby reduce the bias. Unfortunately, this naïve strategy to mitigate the bias would increase memory usage and result in delayed model updates that reduce the algorithm's ability to adapt to regime shifts in the data generating distribution.

Consider now an instance of the fair learning problem (4), where $L(\hat{y}, y) = (\hat{y} - y)^2$ is the square loss, and the unfairness penalty is constructed from the $\mathcal{L}^2$-distance $\mathcal{D}_\Phi = d_2$ and the quadratic regularization function $\rho(z) = 4z^2$. Hence, the unfairness penalty is proportional to the energy distance. As the $\mathcal{L}^2$-distance is a special case of an MMD metric corresponding to a distance-induced kernel, we can use the techniques developed in Section 4 to obtain unbiased batch gradients (see Theorem 4.3) and solve problem (4) with SGD. For comparison, we use SGD also to solve a variant of this problem, where the unfairness penalty is constructed from the $\mathcal{L}^1$-distance $\mathcal{D}_\Phi = d_1$ and the linear regularization function $\rho(z) = 2z$. Hence, the unfairness penalty is proportional to the Wasserstein distance. In this case the gradient of the empirical objective function corresponding to any batch of training samples provides a biased estimator for the true gradient (see Lemma 4.1). In both variants of problem (4) we set $\mathcal{H}$ to the family of all neural networks with one hidden layer accommodating 20 nodes, ReLU activation functions at the hidden layer and linear activation functions at the output layer. The resulting problems are trained in an online fashion by using the Adam optimizer [39] with learning rate 0.01. In both models we prescribe a target batch size $\bar{N}$ and construct batches that contain at least $\bar{N}$ samples overall and at least 2 samples of each class as in Section 4. This construction ensures that all gradient estimators are well-defined.

We construct synthetic training and test data as follows. We assume that X comprises $d = 10$ independent features, where X_i follows the uniform distribution on $[0, 1]$ for every $i < d$, while $X_d = A$ represents the protected attribute with $\mathbb{P}[A = a] = \frac{1}{2}$ for all $a \in \mathcal{A}$. The target Y displays a non-linear dependence on X that may change over time. Specifically, we set $Y = \max_{j\in\{1,\ldots,5\}} \langle s_j, X \rangle$, where the vectors s_j are drawn



(A) Energy distance-based unfairness penalty (B) Wasserstein distance-based unfairness penalty

FIGURE A.5. Test loss of the optimal regressors corresponding to fair learning models with different unfairness penalties and target batch sizes as a function of the number of training samples

independently from the uniform distribution on $[-2, 2]$ once before the experiment starts and then once again at a regime shift that occurs after 2,000 training samples.

Figure A.5 visualizes the test loss of the optimal regressor as a function of the number of training samples for the learning models with energy and Wasserstein distance-based unfairness penalties and for two different target batch sizes $\bar{N} \in \{4, 50\}$, respectively. The test loss is evaluated on 1,000 test samples that follow the same distribution as the last seen training sample. The solid lines in Figure A.5 represent the averages and the shaded areas visualize the standard errors corresponding to five independent replications of the same experiment. We note that larger target batch sizes result in slower convergence because an update step is only possible once sufficiently many training samples have been accumulated to form a valid batch. This observation suggests that small target batch sizes $\bar{N}$ are preferable. However, if the gradient estimators are biased, which happens when working with Wasserstein distance-based unfairness penalties, then a large $\bar{N}$ is needed to keep the bias small. Indeed, the Wasserstein distance-based method with $\bar{N} = 4$ attains a low test loss significantly faster than the one with $\bar{N} = 50$. However, after the regime shift, the test loss of the method with $\bar{N} = 50$ continues to decrease at a steady rate, whereas that of the method with $\bar{N} = 4$ saturates at a higher level. Since the energy distance admits unbiased gradient estimators, it does not suffer from such a trade-off, that is, smaller values of $\bar{N}$ are preferable both in terms of convergence speed and the test loss at equilibrium. In this case, the only reason for increasing $\bar{N}$ is to reduce gradient noise in the prediction loss and the unfairness regularizer. This follows directly from the central limit theorem.

The key insights of this section can be summarized as follows. When working with biased gradient estimators, we will either converge quickly to a low-quality solution (if $\bar{N}$ is small) or we will converge slowly to a high-quality solution (if $\bar{N}$ is large). When working with *un*biased gradient estimators, on the other hand, we can converge quickly to a high-quality solution (by choosing $\bar{N}$ small). This feature is particularly useful when the distribution of the samples changes over time.

References

[1] A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129, 2019.

[2] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[3] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[4] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org, 2019.

[6] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv:1705.10743*, 2017.

[7] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, 1984.

[8] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv:1706.02409*, 2017.

[9] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[10] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: Experiences from the scikit-learn project. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[11] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *IEEE International Conference on Data Mining*, pages 71–80, 2013.

[12] J. Cho, G. Hwang, and C. Suh. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*, pages 15088–15099, 2020.

[13] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[14] A. Chouldechova and A. Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[15] Y.-S. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer, 1997.

[16] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 7321–7331, 2020.

[17] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[18] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. In *FUture BUsiness TEChnology Conference*, pages 5–12, 2008.

[19] H. Cramér. *On the Composition of Elementary Errors: Statistical Applications*. Almqvist and Wiksell, 1928.

[20] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. https://reut.rs/2Od9fPr, 2018.

[21] J. Dedecker and F. Merlevède. The empirical distribution function for dependent variables: Asymptotic and nonasymptotic results in $\mathbb{L}_p$. *ESAIM: Probability and Statistics*, 11:102–114, 2007.

[22] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 325–333, 2018.

[23] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu. Fairness via representation neutralization. In *Advances in Neural Information Processing Systems*, pages 12091–12103, 2021.

[24] D. Dua and C. Graff. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2017.

[25] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, pages 214–226, 2012.

[26] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk. In F. Palumbo, A. Montanari, and M. Vichi, editors, *Data Science*, pages 231–242. Springer, 2017.

[27] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[28] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *SIAM International Conference on Data Mining*, pages 144–152, 2016.

[29] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[30] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[31] S. Ghosh and M. Squillante. Unbiased gradient estimation for distributionally robust learning. *arXiv:2012.12367*, 2020.

[32] P. Glasserman. Performance continuity and differentiability in Monte Carlo optimization. In *Winter Simulation Conference*, pages 518–524, 1988.

[33] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[34] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872, 2020.

[35] K. D. Johnson, D. P. Foster, and R. A. Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv:1608.00528*, 2016.

[36] S. M. Julia Angwin, Jeff Larson and L. Kirchner. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[37] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[38] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[39] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[40] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[41] M. Lohaus, M. Perrot, and U. von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369, 2020.

[42] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[43] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *Advances in Neural Information Processing Systems*, pages 1117–1128, 2020.

[44] K. P. Murphy. *Machine Learning: A probabilistic Perspective*. MIT Press, 2012.

[45] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[46] R. Nabi and I. Shpitser. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, pages 1931–1940, 2018.

[47] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting MMD and Sinkhorn divergences for fair and transferable representation learning. In *Advances in Neural Information Processing Systems*, pages 15360–15370, 2020.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[49] F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin. Post-processing for individual fairness. In *Advances in Neural Information Processing Systems*, pages 25944–25955, 2021.

[50] S. R. Pfohl, A. Foryciarz, and N. H. Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113:103621, 2021.

[51] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[52] T. Räz. Group fairness: Independence revisited. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 129–137, 2021.

[53] S. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

[54] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[55] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[56] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[57] Y. Roh, K. Lee, S. E. Whang, and C. Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021.

[58] T. Salvemini. Sul calcolo degli indici di concordanza tra due caratteri quantitativi. *Atti della VI Riunione della Società Italiana di Statistica*, 1943.

[59] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[60] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

[61] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[62] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems*, pages 8242–8251, 2019.

[63] G. R. Shorack and G. Shorack. *Probability for Statisticians*. Springer, 2000.

[64] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[65] G. Székely. $\mathcal{E}$-statistics: The energy of statistical samples. Technical Report 02-16, Bowling Green State University, Department of Mathematics and Statistics, 2002.

[66] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 11(5):1–16, 2004.

[67] B. Taskesen, V. A. Nguyen, D. Kuhn, and J. Blanchet. A distributionally robust approach to fair classification. *arXiv:2007.09530*, 2020.

[68] G. Tucker, A. Mnih, C. J. Maddison, J. Lawson, and J. Sohl-Dickstein. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pages 2624–2633, 2017.

[69] US Equal Employment Opportunity Commission. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *EEOC: Washington, DC, USA*, 1979.

[70] P. Vicol, L. Metz, and J. Sohl-Dickstein. Unbiased gradient estimation in unrolled computation graphs with persistent evolution strategies. In *International Conference on Machine Learning*, pages 10553–10563, 2021.

[71] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[72] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 2007.

[73] M. Wick, S. Panda, and J.-B. Tristan. Unlocking fairness: A trade-off revisited. In *Advances in Neural Information Processing Systems*, pages 8780–8789, 2019.

[74] R. Williamson and A. Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797, 2019.

[75] Y. Wu, L. Zhang, and X. Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, pages 3356–3362, 2019.

[76] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data*, pages 570–575, 2018.

[77] X. Yao, T. Huang, R.-X. Zhang, R. Li, and L. Sun. Federated learning with unbiased gradient aggregation and controllable meta updating. *arXiv:1910.08234*, 2019.

[78] M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*, 2020.

[79] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, pages 1171–1180, 2017.

[80] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[81] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[82] M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv:1212.5701*, 2012.

[83] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[84] J. Zhang and E. Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3671–3681, 2018.

[85] J. Zhang and E. Bareinboim. Fairness in decision-making—the causal explanation formula. In *AAAI Conference on Artificial Intelligence*, pages 2037–2045, 2018.

[86] Z. Zhang. Improved Adam optimizer for deep neural networks. In *International Symposium on Quality of Service*, pages 1–2, 2018.