The Pokémon Theorem: Fairness Criteria Can't Catch 'Em All

Daniel Matsui Smola
Department of Computer Science
University of Washington
Seattle, WA
daniel@matsuismola.com

Alex Smola
Boson AI
Santa Clara, CA
smola@boson.ai

Abstract

Debates around algorithmic fairness rarely end. Models considered "fair" by one criterion are soon shown to be unfair by another. We prove that this is unavoidable. Satisfying any finite checklist of fairness criteria never certifies fairness itself, since the model may always violate yet another criterion. We call this the Pokémon theorem: you can't catch them all. In fact, trying to catch all criteria recovers the classical impossibility theorems. Since these violations come from differences between the groups' data, a natural remedy would be to make the data fair by learning fair features. Unfortunately, this, too, is impossible: features that scrub out sensitive group information also scrub out all predictive signal. While exact fairness is out of reach, approximate fairness is not. We derive the Pareto frontier between fairness and predictive signal for approximately fair features, and show how to attain it.

1 Introduction

Fairness of a model is usually assessed with a checklist of fairness criteria that compare how it treats two groups (e.g., men and women). A practitioner might require that the model makes positive decisions at equal rates for both groups, makes errors at equal rates for both groups, and gives scores that are calibrated within each group. Yet it has long been known that such criteria can be incompatible. In the 1960s and 1970s, researchers proposed several definitions of fairness for educational and employment tests, and showed that they cannot all hold at the same time [Darlington, 1971, Thorndike, 1971, Cole, 1973, Petersen and Novick, 1976]. The debate remains unresolved sixty years later [Hutchinson and Mitchell, 2019].

The debate intensified in 2016 with COMPAS, an algorithm used by US courts that gives each defendant a risk score predicting how likely they are to reoffend. ProPublica found that Black defendants who did not reoffend were labeled high risk nearly twice as often as white defendants [Angwin et al., 2016]. Northpointe, the developer of COMPAS, replied that the score was calibrated: among defendants with the same score, the same fraction reoffended in both groups [Dieterich et al., 2016]. Kleinberg, Mullainathan and Raghavan [KMR; Kleinberg et al., 2017] showed that these two fairness criteria are incompatible: when the groups reoffend at different rates (unequal base rates), no score can be calibrated within each group and also have equal error rates for both groups, unless it predicts perfectly. This result, together with that of Chouldechova [2017], started the modern study of impossibility results in fairness.

Since then, many more fairness criteria have been proposed, along with many more impossibility theorems [Pleiss et al., 2017, Zhao and Gordon, 2022, Barocas et al., 2023, Stewart et al., 2024]. Each theorem shows that a particular combination of criteria cannot hold together. *Is the problem specific to these combinations, or does it affect any finite checklist of criteria?*

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We prove that satisfying any finite checklist of criteria cannot certify that an algorithm is fair, which is why this sixty-year-old debate has remained unresolved. The proof uses mean embeddings in a reproducing kernel Hilbert space (RKHS) to express the difference between the two groups' data distributions and each fairness criterion as vectors. Each criterion measures a dot product with this difference vector; a value of zero means that the criterion detects no difference between the groups and is satisfied. All criteria on a checklist can therefore be satisfied while the groups differ in directions orthogonal to the subspace the criteria span. If the two groups' data distributions were identical, the difference vector would be zero, and every criterion would be satisfied whatever the algorithm, as long as it does not use the group. Unfairness thus originates in the data, not in the algorithm. This raises a second question: *can we learn features that scrub group information, yet retain predictive signal?*

We answer both questions through the following contributions:

1. **Pokémon Theorem.** No finite checklist of fairness criteria can certify fairness. We characterize the worst-case residual unfairness after the best m criteria are satisfied. These best criteria are the leading eigendirections of the data's covariance operator, pooled over both groups (Section 3).
2. **Recovering Classical Impossibilities.** Escaping the Pokémon theorem requires a dense checklist of criteria. We show that requiring one recovers the classical impossibility theorems, including KMR. With only the best m criteria, the same conclusions hold approximately, with an error that shrinks as m grows (Section 4).
3. **Impossibility of Fair Feature Learning.** Fair representation learning seeks features that scrub group information, so that every model built on them is fair, while retaining predictive signal. We show that these two goals are incompatible when base rates differ (Section 5).
4. **Approximate Fairness.** While exact fairness is impossible, approximate fairness is not: features that scrub group information only partially can retain predictive signal. We derive the Pareto frontier between fairness and predictive signal. The more the base rates differ, the more fairness must be given up for the same signal. We also give a rule that attains this frontier (Section 6).

2 Background

2.1 Fairness Criteria as Expectations

Each individual has features $X \in \mathcal{X}$, an outcome $Y \in \{0, 1\}$, such as whether a defendant reoffends, and a group $G \in \{a, b\}$. Write P_g for the distribution of X in group g. The *base rate* $p_g := \Pr[Y = 1 \mid G = g]$ is the fraction of positive outcomes in group g, and $\Delta p := p_a - p_b$ is the *base-rate gap*. We assume $\Pr[G = g] > 0$ and $0 < p_g < 1$ for both groups.

Many fairness criteria compare the expectation of a test function $c : \mathcal{X} \to \mathbb{R}$ in the two groups. A criterion is satisfied when the two expectations are equal,

$$\mathbb{E}[c(X) \mid G = a] - \mathbb{E}[c(X) \mid G = b] \;=\; 0. \tag{1}$$

For example, demographic parity requires a classifier $\hat{Y} = f(X)$ to select both groups at equal rates; its test function is $c(x) = \mathbf{1}[f(x) = 1]$ [Dwork et al., 2012]. Other criteria compare the groups within each outcome class or involve the outcome itself; we introduce them in Section 3.3. For fixed c, the map $P \mapsto \mathbb{E}_P[c(X)]$ is linear in the distribution P, even when c is nonlinear in x. We will use kernels to exploit this linearity. Nonlinear criteria are handled in Section 3.3, and learned features $Z = \Phi(X)$ in place of X are handled in Section 5.

2.2 Kernels and Mean Embeddings

Fix a positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with RKHS $\mathcal{H}$ and feature map $\phi : \mathcal{X} \to \mathcal{H}$. The reproducing property gives $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$ for $f \in \mathcal{H}$, and $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ [Aronszajn, 1950, Schölkopf and Smola, 2001]. On a compact $\mathcal{X}$, k is *universal* if $\mathcal{H}$ is dense in $C(\mathcal{X})$, so every continuous function can be uniformly approximated by RKHS functions [Steinwart, 2001].

For a distribution P on $\mathcal{X}$ with $\mathbb{E}_P[\sqrt{k(X, X)}] < \infty$, the *mean embedding* is $\mu_P := \mathbb{E}_P[\phi(X)] \in \mathcal{H}$. The reproducing property yields $\mathbb{E}_P[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}}$ [Gretton et al., 2012]. The kernel k is

characteristic when $\mu_P = \mu_Q$ holds if and only if $P = Q$ [Sriperumbudur et al., 2010]; universality implies characteristic [Gretton et al., 2012], and Gaussian, Laplace and Matérn kernels on $\mathbb{R}^d$ are characteristic. The *maximum mean discrepancy* (MMD) between P and Q is

$$\mathrm{MMD}(P, Q) \;=\; \sup_{\|f\|_{\mathcal{H}} \le 1} \big(\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\big) \;=\; \|\mu_P - \mu_Q\|_{\mathcal{H}}, \tag{2}$$

where the *witness function* $f^\star = (\mu_P - \mu_Q)/\|\mu_P - \mu_Q\|_{\mathcal{H}}$ attains the supremum for $\mu_P \neq \mu_Q$ [Gretton et al., 2012]. For a characteristic kernel, $\mathrm{MMD}(P, Q) = 0$ if and only if $P = Q$.

2.3 Fairness Criteria as Directions

Write the mean embedding of P_g and the *conditional mean discrepancy* between the groups as

$$\mu_g \;:=\; \mathbb{E}[\phi(X) \mid G = g], \qquad \delta \;:=\; \mu_a - \mu_b. \tag{3}$$

For a criterion whose test function lies in the RKHS, $c(x) = \langle v,\, \phi(x)\rangle_{\mathcal{H}}$, the reproducing property gives

$$\mathbb{E}[c(X) \mid G = a] - \mathbb{E}[c(X) \mid G = b] \;=\; \langle v,\, \delta\rangle_{\mathcal{H}}. \tag{4}$$

Satisfying the criterion means $\langle v,\, \delta\rangle_{\mathcal{H}} = 0$: its direction v is orthogonal to δ. By construction, $\mathrm{MMD}(P_a, P_b) = \|\delta\|_{\mathcal{H}}$ is the largest difference between the groups that any criterion with $\|v\|_{\mathcal{H}} \le 1$ can detect. For a characteristic kernel, $\delta = 0$ means that the groups' distributions of X are equal, so every criterion that compares them is satisfied. Criteria that compare other distributions have their own discrepancies; those within each outcome class (Section 3.3) use the class-conditional versions

$$\mu_{y,g} \;:=\; \mathbb{E}[\phi(X) \mid Y = y, G = g], \qquad \delta_y \;:=\; \mu_{y,a} - \mu_{y,b}, \tag{5}$$

where δ_y is the conditional mean discrepancy between the groups among individuals with outcome y.

2.4 Related Work

Kleinberg et al. [2017], Chouldechova [2017] and Pleiss et al. [2017] show that calibration and equal error rates are incompatible when base rates differ. Stewart et al. [2024] show that equal error rates are also incompatible with a weaker form of calibration, and Barocas et al. [2023] show that every pair among demographic parity, equal error rates and calibration is incompatible. Each result fixes a combination of criteria. We consider every finite checklist, and recover these results as the checklist grows (Section 4). Kearns et al. [2018] show that auditing a few subgroups can miss unfairness on others; we vary the criteria instead of the subgroups.

Fair representation learning seeks features that scrub group information while retaining predictive signal [Zemel et al., 2013, Edwards and Storkey, 2016, Louizos et al., 2016, Madras et al., 2018, Deka and Sutherland, 2023]. Zhao and Gordon [2022] show that, when base rates differ, any classifier built on such features must make errors, and Lechner et al. [2021] show that no features can be fair for every downstream task. We show that, when base rates differ, scrubbing group information both overall and within each outcome class removes all predictive signal (Section 5). Menon and Williamson [2018] and Wei and Niethammer [2022] study the frontier between fairness and accuracy for a single classifier. We solve the Pareto frontier between fairness and predictive signal for features, whose slope is set by the base-rate gap (Section 6).

Kernel methods have so far been used to enforce fairness [Pérez-Suay et al., 2017, 2023, Donini et al., 2018, Wei et al., 2023]. We use the same embeddings to show what finitely many constraints cannot certify.

3 The Pokémon Theorem

3.1 Qualitative Theorem

We return to the question raised in the introduction: *is the problem specific to particular combinations of criteria, or does it affect any finite checklist?* We prove that it affects any finite checklist, even when every criterion on it is satisfied exactly. We call this the Pokémon theorem: *gotta catch 'em all*, but another criterion always escapes.

Theorem 3.1 (Pokémon theorem)**.** *Let k be a characteristic kernel with RKHS $\mathcal{H}$, and suppose the groups differ, $P_a \neq P_b$. Let the checklist consist of m criteria with test functions $c_i(x) = \langle v_i,\, \phi(x)\rangle_{\mathcal{H}}$, all of which are satisfied:*

$$\mathbb{E}[c_i(X) \mid G = a] - \mathbb{E}[c_i(X) \mid G = b] \;=\; \langle v_i,\, \delta\rangle_{\mathcal{H}} \;=\; 0 \qquad \textit{for all } i = 1, \ldots, m,$$

or equivalently $\delta \perp V$ with $V := \operatorname{span}\{v_1, \ldots, v_m\}$. Then there is a criterion $v_{m+1} \in \mathcal{H}$, orthogonal to V, that is violated: $\langle v_{m+1},\, \delta\rangle_{\mathcal{H}} > 0$. An explicit choice is the MMD witness $v_{m+1} := \delta/\|\delta\|_{\mathcal{H}}$, for which

$$\langle v_{m+1},\, \delta\rangle_{\mathcal{H}} \;=\; \|\delta\|_{\mathcal{H}} \;=\; \mathrm{MMD}(P_a, P_b). \tag{6}$$

Proof. Satisfying every criterion places δ in $V^{\perp}$. Since the kernel is characteristic and $P_a \neq P_b$, $\delta \neq 0$. Hence $\delta/\|\delta\|_{\mathcal{H}}$ is a unit vector in $V^{\perp}$ with $\langle \delta/\|\delta\|_{\mathcal{H}},\, \delta\rangle_{\mathcal{H}} = \|\delta\|_{\mathcal{H}} > 0$. □

Implications. Satisfying every criterion on a checklist does not make an algorithm fair. For every finite checklist, groups that satisfy it yet differ exist (Lemma B.1). Whenever the groups' data distributions differ, some other criterion is violated, and satisfying it as well only leads to another, since the theorem applies to every finite checklist. Unlike the classical impossibilities, this holds even when the groups have equal base rates.

3.2 Quantitative Theorem

Satisfying more criteria should nonetheless buy us something. *How large can the remaining unfairness be if we satisfy the best m criteria?*

When every criterion is satisfied, $\delta \perp V$, and the largest violation of any criterion with $\|v\|_{\mathcal{H}} \leq 1$ is $\|\delta\|_{\mathcal{H}} = \mathrm{MMD}(P_a, P_b)$, attained by the witness. We call it the *residual unfairness*.

Although $\mathcal{H}$ is infinite-dimensional, the variation of the data may concentrate in relatively few directions. The covariance operator of the data, pooled over both groups, ranks these directions by variance:

$$\Sigma \;:=\; \mathbb{E}_Q\big[(\phi(X) - \mu_Q) \otimes (\phi(X) - \mu_Q)\big], \qquad Q \;:=\; \tfrac{1}{2}(P_a + P_b). \tag{7}$$

For bounded k, the eigendecomposition of Σ gives eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and orthonormal eigenvectors $e_1, e_2, \ldots$ on its positive spectrum. We first use a standard source condition [Caponnetto and De Vito, 2007]:

(A1) Source condition. $\delta = \Sigma^r u$ with $\|u\|_{\mathcal{H}} \leq R$, for fixed $r, R > 0$.

This limits how much of the difference between the groups can lie in directions of small variance; with $r = 1/2$ it holds for every pair of groups (Section B.3). These vectors form the ellipsoid $B_r(R) := \{\Sigma^r u : \|u\|_{\mathcal{H}} \leq R\}$, with semi-axes $R\lambda_j^r$. The supremum below ranges over every vector in this ellipsoid (Section B.3).

Theorem 3.2 (Optimal residual)**.** *Under **(A1)**,*

$$\inf_{\dim V \leq m} \sup_{\delta \in B_r(R),\; \delta \perp V} \|\delta\|_{\mathcal{H}} \;=\; R\lambda_{m+1}^r, \tag{8}$$

attained by the m optimal criteria $e_1, \ldots, e_m$.

Proof sketch. Satisfying the first m eigendirections removes the m longest axes of the ellipsoid, leaving a residual of at most the next longest semi-axis, $R\lambda_{m+1}^r$. By pigeonhole, any m-dimensional subspace leaves a direction in the span of the first $m+1$ eigenvectors unchecked. The ellipsoid extends at least $R\lambda_{m+1}^r$ along this direction, giving the matching lower bound (Section B.2). □

Thus, satisfying the best m criteria restricts the worst-case residual unfairness to $R\lambda_{m+1}^r$. To determine how quickly it decays as we add criteria, we make a second standard assumption:

(A2) Polynomial eigendecay. $\lambda_j \asymp j^{-\alpha}$ for some $\alpha > 1$.

Corollary 3.3 (Rate with more criteria)**.** *Under **(A1)** and **(A2)**, the optimal residual in* (8) *satisfies* $R\lambda_{m+1}^r = \Theta(R\, m^{-\alpha r})$.

Implications. The optimal residual is the Kolmogorov m-width of the source class [Pinkus, 1985]. As we add criteria, the residual unfairness decays at rate $m^{-\alpha r}$; faster spectral decay (larger α) or a smoother difference between the groups (larger r) makes it decay faster. Nevertheless, the worst case remains nonzero for every finite m, even when all criteria are satisfied.

3.3 Other Criteria

Inexact criteria. Practical criteria hold only up to a tolerance, $|\langle v_i,\, \delta\rangle_{\mathcal{H}}| \le \varepsilon$. These readings constrain only the part of the difference that the checklist sees, $P_V\delta$, where P_V is the orthogonal projection onto V. For orthonormal criteria, Pythagoras gives

$$\|P_{V^\perp}\delta\|_{\mathcal{H}}^2 \;=\; \mathrm{MMD}^2(P_a, P_b) - \|P_V\delta\|_{\mathcal{H}}^2 \;\ge\; \mathrm{MMD}^2(P_a, P_b) - m\varepsilon^2. \tag{9}$$

Thus, if $\mathrm{MMD}(P_a, P_b) > \sqrt{m}\,\varepsilon$, a criterion orthogonal to the checklist is still violated (Section B.4).

Different distributions. Criteria can also be linear in distributions other than that of X:

- equal error rates compare the distribution of X conditioned on the outcome $Y = y$ in the two groups;
- calibration of a score $S = s(X)$ within each group, $\mathbb{E}[(Y - S)\,h(S) \mid G = g] = 0$ for every bounded function h, involves the joint distribution of (X, Y).

The Pokémon theorem applies to each, with δ replaced by the conditional mean discrepancy of the corresponding distributions (Section B.5).

Nonlinear criteria. Criteria such as predictive parity, which compares a ratio of two expectations, are nonlinear in the distribution. A von Mises expansion, a Taylor expansion in the distribution, writes such a criterion T as a series in differences of expectations $\Delta_j := \mathbb{E}[c_j \mid G = a] - \mathbb{E}[c_j \mid G = b]$, namely $T(P_a) - T(P_b) = \sum_j a_j\Delta_j + \frac{1}{2}\sum_{j,k} b_{jk}\Delta_j\Delta_k + \cdots$ [Fernholz, 1983]. Each Δ_j is a linear criterion. If finitely many appear, they form a finite checklist, and satisfying it makes every term zero even when the groups differ.

Infinite criteria. A checklist may also contain infinitely many criteria. If their span is not dense in $\mathcal{H}$, an orthogonal complement remains, and any difference between the groups that lies there satisfies every criterion. (Think of Hilbert's hotel: checking every other eigendirection still leaves infinitely many unchecked.) Calibration is such a family: it has infinitely many criteria, one for every bounded function h, yet each depends only on the score and the outcome, so groups whose features differ beyond the score satisfy all of them (Section B.6). A dense infinite checklist does certify; Section 4 shows where requiring one leads.

4 Recovering Classical Impossibilities

Section 3 showed that a finite checklist cannot certify fairness, and that an infinite checklist that is not dense can leave differences between the groups unchecked. A dense infinite checklist leaves none. *What happens if we use a dense infinite checklist to certify fairness?* We show that it certifies one of the three traditional notions of fairness, each of which can hold on its own. When base rates differ, however, no useful score (other than a perfect one) satisfies any two of them at once. This recovers the classical impossibility theorems, including KMR from the introduction.

4.1 Three Notions of Fairness

The classical results concern a score $S \in [0, 1]$ computed from the features X, such as the risk score of COMPAS. Each fairness criterion uses one of three distributions of the score: its overall distribution, its distribution conditioned on the outcome, or its joint distribution with the outcome. Write μ_g for the mean embedding of the score's distribution in group g, $\mu_{y,g}$ for that of its distribution in group g conditioned on $Y = y$, and $\delta := \mu_a - \mu_b$ and $\delta_y := \mu_{y,a} - \mu_{y,b}$, as in Section 2.3 with S in place of X. A dense infinite checklist for one of the three distributions certifies the corresponding traditional notion of fairness [Barocas et al., 2023]:

(C1) *Independence*: S has the same distribution in both groups, $\delta = 0$. The dense checklist of equal selection rates $\Pr[S > t \mid G = g]$ at every threshold t certifies it.

(C2) *Separation*: S has the same distribution in both groups conditioned on each outcome, $\delta_0 = \delta_1 = 0$. The dense checklist of equal error rates $\Pr[S > t \mid Y = y, G = g]$ at every threshold t certifies it.

(C3) *Sufficiency*: a score means the same in both groups. Among individuals with score s, the fraction of positive outcomes is the same in both groups, $f(s)$. The checklist $\mathbb{E}[(Y - f(S))\, h(S) \mid G = g] = 0$, for every bounded function h of the score, certifies it. These criteria use the joint distribution of (S, Y) and are checked within each group. Calibration is the case $f(s) = s$.

These checklists are dense among functions of the score, though not among functions of the features (Section 3.3), so they certify properties of the score only. When S can take infinitely many values, the Pokémon theorem shows that finitely many thresholds certify neither independence nor separation (Lemma B.1).

Each notion can hold on its own. A constant score satisfies independence (C1), and a perfect score, $S = Y$, satisfies separation (C2) and sufficiency (C3). Both are degenerate: a constant score carries no predictive signal, and a perfect score requires an outcome that can be predicted without error. We call a score useful if it carries predictive signal. *Can a useful score (other than a perfect one) satisfy any two of the notions at once?*

4.2 Why Any Two Notions Conflict

When base rates differ, no useful score (other than a perfect one) satisfies any two of the notions at once:

Theorem 4.1 (Any two notions)**.** *Suppose the base rates differ, $p_a \neq p_b$.*

(i) A score that satisfies independence (C1) and separation (C2) carries no predictive signal (the score is independent of Y).

(ii) Only a perfect score satisfies separation (C2) and sufficiency (C3): $f(S) = Y$ almost surely, and $S = Y$ if S is calibrated.

(iii) No score satisfies independence (C1) and sufficiency (C3).

Proof sketch. The law of total expectation connects the three distributions of the score: $\mu_g = p_g\, \mu_{1,g} + (1 - p_g)\, \mu_{0,g}$. Subtracting the groups gives

$$\delta \;=\; p_a\, \delta_1 + (1 - p_a)\, \delta_0 + (p_a - p_b)\, \omega, \qquad \omega \;:=\; \mu_{1,b} - \mu_{0,b}, \tag{10}$$

where ω, the predictive signal, is the difference between the score's distributions in the two outcome classes of group b.

(i): Independence (C1) asks for $\delta = 0$, and separation (C2) for $\delta_0 = \delta_1 = 0$. Together they leave $(p_a - p_b)\, \omega = 0$ in (10), so $\omega = 0$; thus the score carries no predictive signal.

(ii) and (iii): Since $f(s)$ is the fraction of positive outcomes at score s, sufficiency (C3) fixes each group's average of $f(S)$ at its base rate, $\mathbb{E}[f(S) \mid G = g] = p_g$. (iii) then follows, as independence (C1) makes these averages equal, which requires $p_a = p_b$. Under separation (C2), the same law of total expectation gives $p_a - p_b = (p_a - p_b)\, (\mathbb{E}[f(S) \mid Y = 1] - \mathbb{E}[f(S) \mid Y = 0])$. Since $f(S) \in [0, 1]$, this forces $f(S) = Y$, which proves (ii) (Section C). □

The proofs of (ii) and (iii) use only averages of $f(S)$. (ii) becomes KMR for a calibrated score, $f(s) = s$. KMR requires only equal average scores within each outcome class instead of separation:

Theorem 4.2 (KMR, Kleinberg et al., 2017)**.** *Suppose $p_a \neq p_b$. If S is calibrated within each group and has equal average scores in both groups within each outcome class, then $S = Y$ almost surely.*

Statements (i) and (iii) are incompatibilities identified by Barocas et al. [2023], and (ii) strengthens their result for separation and sufficiency (Section C). Stewart et al. [2024] weaken calibration in KMR to equal gaps between average score and base rate in the two groups; our proof of (ii) covers this case. The result of Chouldechova [2017], that a binary classifier generally cannot have both equal error rates and equal positive predictive values, follows in the same way (Section C).

Finitely many criteria. As in Section 3.2, the best m criteria of a notion are the leading eigendirections of the covariance operator of the corresponding distribution of the score, pooled over both groups. With only these, the conclusions of Theorem 4.1 hold approximately, with an error that shrinks as m grows. For example, a calibrated score that satisfies the best m criteria of separation has

$$\mathbb{E}\big[|S - Y| \;\big|\; G = g\big] \;\leq\; \frac{1}{\sqrt{m+1}\,|p_a - p_b|}.$$

This follows from Theorem 3.2 with the characteristic kernel $\min(s, s')$ on scores (Corollary C.2). In this way, the Pokémon theorem gives an approximate KMR: the more of these criteria a calibrated score satisfies, the closer it must be to a perfect score.

Implications. The Pokémon theorem is an *insufficiency*: a finite checklist of satisfied criteria certifies nothing. The classical results are *incompatibilities*: certain criteria cannot hold together. The insufficiency leads to the incompatibilities, and in this sense, the Pokémon theorem is the source of the classical impossibilities. Escaping the Pokémon theorem requires dense checklists, and when base rates differ, certifying any two notions with them forces a score without predictive signal or a perfect score, or allows no score at all.

5 The Impossibility of Fair Feature Learning

We now turn to the second question of the introduction: *can we learn features that scrub group information, yet retain predictive signal?* Since unfairness originates in the data rather than in the algorithm, a natural remedy is to change the data. Fair representation learning learns an encoder $\Phi : \mathcal{X} \to \mathcal{Z}$ whose features $Z = \Phi(X)$ scrub the sensitive group information, so that any model built on them inherits fairness for free. Many different methods pursue this goal, using prototypes, adversaries, variational autoencoders, and MMD penalties [Zemel et al., 2013, Edwards and Storkey, 2016, Louizos et al., 2016, Madras et al., 2018, Deka and Sutherland, 2023, Jovanović et al., 2023]. *Why are there so many methods?* We show that when base rates differ, it is impossible to learn truly fair features that retain predictive signal.

What fair features need. For any model built on Z to inherit fairness, the distribution of Z must be insensitive to the group: Z must satisfy independence (C1) (often called parity), the same distribution in both groups, and separation (C2), the same distribution in both groups within each outcome class. To be useful, Z must also retain predictive signal: different distributions in the two outcome classes.

Theorem 5.1 (Impossibility of fair feature learning)**.** *Suppose the base rates differ, $p_a \neq p_b$. If the features $Z = \Phi(X)$ satisfy independence (C1) and separation (C2), they carry no predictive signal.*

Proof sketch. This is Theorem 4.1(i) with Z in place of S (Section D.1). □

Implications. This closes the door on fair feature learning as usually posed: when base rates differ, features that make every model built on them fair carry no predictive signal. It also explains the variety of methods: each approximates, in its own way, a goal that no method can reach. Useful features must give up some independence or some separation. *How much predictive signal can features keep if they give up only a little fairness?* Section 6 answers this question.

6 Approximate Fairness

Features that give up some fairness can retain predictive signal, but only in proportion to the unfairness they permit. We prove this bound and give a simple rule that attains it.

We write μ_g, $\mu_{y,g}$, δ and δ_y as in Section 4, with the features Z in place of the score S and a bounded characteristic kernel on $\mathcal{Z}$. We measure how far the features are from independence (C1) by $\varepsilon := \|\delta\|_{\mathcal{H}}$, and how far they are from separation (C2) by $\rho := \max_y \|\delta_y\|_{\mathcal{H}}$. Their predictive signal in group g is $\omega_g := \mu_{1,g} - \mu_{0,g}$, the difference between their distributions in the two outcome classes, so that ω_b is the ω of (10).

Theorem 6.1 (Fairness and predictive signal)**.** *Suppose $p_a \neq p_b$. In each group g,*

$$\|\omega_g\|_{\mathcal{H}} \;\leq\; \frac{\varepsilon + \rho}{|p_a - p_b|}. \tag{11}$$

Proof sketch. By (10) with Z in place of S, $(p_a - p_b)\,\omega_b = \delta - p_a\delta_1 - (1 - p_a)\,\delta_0$. Taking norms gives $|p_a - p_b|\,\|\omega_b\|_{\mathcal{H}} \le \varepsilon + \rho$, and exchanging the groups gives the bound for ω_a (Section D.2). □

Implications. To keep predictive signal of size $\|\omega_g\|_{\mathcal{H}}$, features must allow unfairness $\varepsilon + \rho$ of at least $|p_a - p_b| \cdot \|\omega_g\|_{\mathcal{H}}$. The more the base rates differ, the more fairness must be given up for the same signal. Theorem 5.1 is the case $\varepsilon = \rho = 0$, where the bound forces the predictive signal to zero.

This bound limits every encoder. *Can we attain this bound?* Let $\|\omega\|_{\text{sep}}$ be the supremum of $\|\omega_g\|_{\mathcal{H}}$ over features that satisfy separation (C2).

Proposition 6.2 (The bound is attained)**.** *Suppose $p_a \neq p_b$. For every $\beta < |p_a - p_b|\,\|\omega\|_{\text{sep}}$, there exist (randomized) features with $\rho = 0$ and $\varepsilon = \beta$ whose predictive signal is $\|\omega_a\|_{\mathcal{H}} = \|\omega_b\|_{\mathcal{H}} = \beta/|p_a - p_b|$.*

Proof sketch. Features that satisfy separation ($\rho = 0$) lie exactly on the bound: (10) leaves $\delta = (p_a - p_b)\,\omega_b$, so $\varepsilon = |p_a - p_b|\,\|\omega_g\|_{\mathcal{H}}$. To lower their independence violation ε, replace the features by a constant value z_0 for a random fraction $1 - \lambda$ of individuals. This shrinks every difference between their distributions by the factor λ, so ε and the signal $\|\omega_g\|_{\mathcal{H}}$ shrink together, ρ stays zero, and the features stay on the bound. Starting from separated features with signal close to $\|\omega\|_{\text{sep}}$ reaches every point below it (Section D.2). □

Implications. Up to $\|\omega\|_{\text{sep}}$, the Pareto frontier between unfairness $\varepsilon + \rho$ and predictive signal $\|\omega_g\|_{\mathcal{H}}$ is the straight line of Theorem 6.1, and the construction above attains every point on it. Beyond $\|\omega\|_{\text{sep}}$, more signal requires giving up separation, and the frontier depends on the data.

Computing the frontier. For a binary decision, these quantities are familiar rates, up to a common constant: ε is the gap in selection rates between the groups, ρ is the larger of the gaps in true- and false-positive rates, and $\|\omega_g\|_{\mathcal{H}}$ is $|\,\mathrm{TPR}_g - \mathrm{FPR}_g\,|$.

For decisions based on a score, the frontier is then the solution of a linear program:

(1) Estimate $\Pr[Y = 1 \mid X]$, with scores computed on rows the model was not fit to.

(2) In each group g, predict a positive outcome, $\hat{Y} = 1$, when the score exceeds a threshold t, where t is drawn at random, with probabilities $w_{g,t}$, from the values the score takes on group g's held-out data. The selection rates, TPR_g and FPR_g are then averages of the rates at each threshold, so they are linear in the weights $w_{g,t}$.

(3) For each budget β, maximize $\mathrm{TPR}_g - \mathrm{FPR}_g$ subject to $\varepsilon + \rho \le \beta$.

With $\rho = 0$ imposed, the program gives the most informative separated decision, at $\|\omega\|_{\text{sep}}$, and replacing it by a constant decision for a random fraction of individuals traces the straight line below it. With $\rho > 0$ allowed, larger budgets trace the rest of the frontier, which bends below the line beyond $\|\omega\|_{\text{sep}}$. Section 7 computes this frontier on three datasets and compares common fairness methods with it. Hardt et al. [2016] use the same rules, a separate randomized threshold in each group, to minimize error under equal error rates.

7 Experiments

We measure the residual left by the best m criteria (Theorem 3.2), how close fair feature methods come to the bound of Theorem 6.1, and the frontier of Section 6 on three datasets: Adult Income (men and women), COMPAS (Black and white defendants) and ACS PUMS income data from California (white and non-white respondents). Section E gives the methods and further results.

Residual after the best m criteria. *How much unfairness remains after satisfying the best m criteria?* Figure 1A takes the best m criteria (the leading eigendirections of the pooled covariance Σ, Theorem 3.2) and measures the true residual unfairness: the share of $\mathrm{MMD}^2(P_a, P_b)$ that these criteria do not see. The red line is the worst-case residual presented in Theorem 3.2, computed from the estimated eigenvalues of Σ with $r = 1/2$ and $R = 2$. This source condition holds for any two groups (Proposition B.2), and we see the worst case (red line) bound the true residual (blue line) across all three datasets. We can also see that the worst-case lines are roughly straight on the log–log

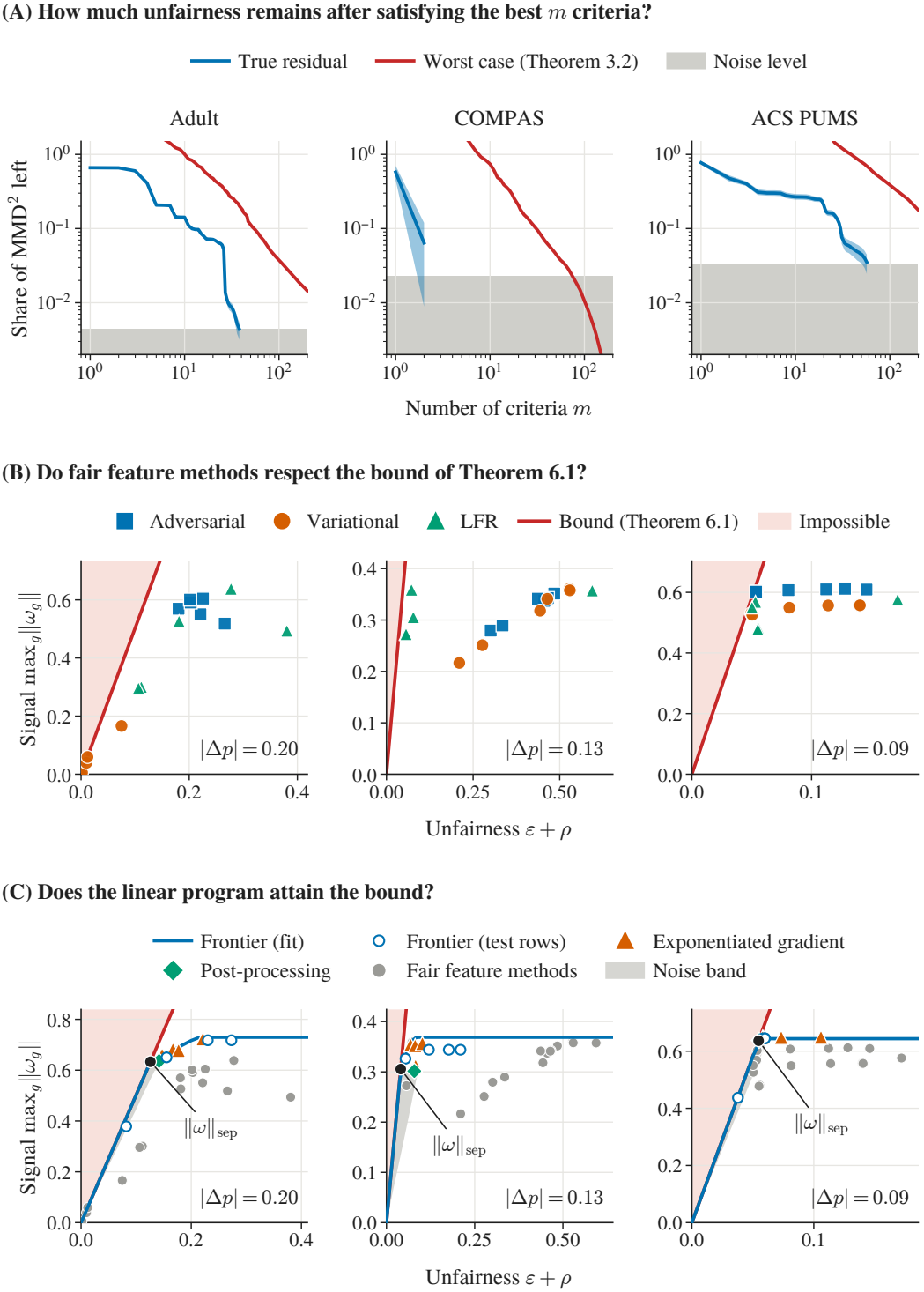


Figure 1: **(A)** True residual unfairness after the best m criteria (blue) and the worst-case residual of Theorem 3.2 (red). **(B)** Predictive signal vs. unfairness of fair feature methods, one point per setting of their fairness weight. **(C)** The frontier from the linear program, fit on training and validation rows (line) and measured on test rows (circles). In (B) and (C), the red line is the bound of Theorem 6.1, and no rule can lie in the shaded region above it. Test-row values are averaged over ten random splits of each dataset, and blue bands are 95% intervals.

plot, which is the case under polynomial eigendecay **(A2)**. Residuals inside the gray band are too small to measure. On COMPAS, the residual falls into this band after only two criteria.

Fair features. *Do fair feature methods respect the bound of Theorem 6.1?* Theorem 6.1 says that features can gain predictive signal only by giving up fairness: moving up in Figure 1B (more predictive signal) requires moving right (more unfairness). The red line draws this bound, $\max_g \|\omega_g\|_{\mathcal{H}} = (\varepsilon + \rho)/|p_a - p_b|$, and no features can lie above it. Its slope is steepest on ACS PUMS, whose base rates differ least. We train adversarial [Edwards and Storkey, 2016], variational [Louizos et al., 2016] and LFR [Zemel et al., 2013] fair feature encoders, each with a weight that trades fairness against accuracy. This weight scales the adversary's loss, an MMD penalty between the groups, or LFR's parity term. For each value of the weight, Figure 1B plots the predictive signal that the features retain against the unfairness they allow, measured on a classifier trained on the features. As Theorem 6.1 guarantees, every point lies below the red line; most lie well below it.

Computing the frontier. *Does the linear program attain the bound?* Figure 1C computes the frontier with the linear program from Section 6, on scores from gradient-boosted trees [Chen and Guestrin, 2016] or a neural network, whichever predicts better on validation data. We fit the program on training and validation rows (line), and measure the same rules on test rows (circles). The frontier follows the bound (red line) up to the black point $\|\omega\|_{\text{sep}}$, the most signal a rule keeps with exact separation ($\rho = 0$), as Proposition 6.2 guarantees. Beyond it, more signal requires giving up separation. The gray band shows where a rule exactly on the bound would land on the test rows by chance. Exact separation costs the group that is harder to predict nothing: on all 30 pairs of dataset and split, $\|\omega\|_{\text{sep}}$ equals the most signal any threshold rule on the score achieves in that group. Beyond the black point, the extra signal comes only from the group that is easier to predict. Exponentiated gradient [Agarwal et al., 2018], optimizing the same objective, lands near the frontier at tight tolerances, and post-processing [Hardt et al., 2016] solves the same program. The gray points repeat the fair feature methods of panel B.

8 Discussion

Scope and open problems. We consider two groups and a binary outcome; individual and causal notions of fairness lie outside this framework. A finite-dimensional class of criteria, such as all linear combinations of d features, can be certified by d criteria; the Pokémon theorem concerns equality of the groups' distributions (Section B.1). Learned features are often reused for tasks whose outcomes are unknown when the features are learned. Separation depends on each task's outcome, and no features that allow accurate predictions are fair for every task [Lechner et al., 2021]; which features are best for a family of tasks remains open.

Fairness as budget allocation. Exact fairness is impossible, but that does not let algorithm designers off the hook. The difference δ between the groups acts as an unfairness budget. A designer decides how to spend it: on predictive signal, or on fairness criteria. The two compete. By the law of total expectation (10),

$$\delta \;=\; p_a\,\delta_1 + (1 - p_a)\,\delta_0 + (p_a - p_b)\,\omega,$$

so the budget splits between differences within each outcome class and the predictive signal ω. Keeping signal of size $\|\omega\|_{\mathcal{H}}$ therefore leaves unfairness $\varepsilon+\rho$ of at least $|p_a-p_b|\cdot\|\omega\|_{\mathcal{H}}$ (Theorem 6.1). Which criteria to spend it on depends on the social setting: equal error rates may matter most in criminal justice, and equal selection rates in hiring. Following a checklist blindly can use up the budget on criteria that matter little in the application. A fair model needs more than a satisfied checklist: it needs criteria chosen deliberately for the people it affects.

Acknowledgments and Disclosure of Funding

We thank Yu-Xiang Wang for suggesting that a connection between the Pokémon theorem and KMR should exist, and Ryan Tibshirani for helpful comments and feedback on the manuscript. This work received no external funding. The authors declare no competing interests.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.1090/S0002-9947-1950-0051437-7.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. URL https://fairmlbook.org/.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019. doi: 10.1147/JRD.2019.2942287.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047.

Nancy S. Cole. Bias in selection. *Journal of Educational Measurement*, 10(4):237–255, 1973. doi: 10.1111/j.1745-3984.1973.tb00802.x.

Richard B. Darlington. Another look at "cultural fairness". *Journal of Educational Measurement*, 8 (2):71–82, 1971. doi: 10.1111/j.1745-3984.1971.tb00908.x.

Namrata Deka and Danica J. Sutherland. MMD-B-Fair: Learning fair representations with statistical testing. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pages 9564–9576, 2023. URL https://proceedings.mlr.press/v206/deka23a.html.

William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe Inc., July 2016.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 6478–6490, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

Michele Donini, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 2796–2806, 2018. URL https://papers.neurips.cc/paper/7544-empirical-risk-minimization-under-fairness-constraints.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*, pages 214–226. ACM, 2012. doi: 10.1145/2090236.2090255.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR 2016)*, 2016. URL https://openreview.net/forum?id=ByuP8yZRb.

Luisa Turrin Fernholz. *von Mises Calculus for Statistical Functionals*, volume 19 of *Lecture Notes in Statistics*. Springer, 1983. doi: 10.1007/978-1-4612-5604-5.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Zaïd Harchaoui, Francis Bach, and Éric Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems 20*, pages 609–616, 2008.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 3315–3323, 2016. URL https://papers.neurips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, pages 49–58, 2019. doi: 10.1145/3287560.3287600.

Nikola Jovanović, Mislav Balunović, Dimitar Iliev Dimitrov, and Martin Vechev. FARE: Provably fair representation learning with practical certificates. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *PMLR*, pages 15401–15420, 2023. URL https://proceedings.mlr.press/v202/jovanovic23a.html.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. doi: 10.1007/s10115-011-0463-8.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572, 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. Impossibility results for fair representations. *arXiv preprint arXiv:2107.03483*, 2021.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR 2016)*, 2016. URL https://iclr.cc/archive/www/2016.html.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *PMLR*, pages 3384–3393, 2018. URL https://proceedings.mlr.press/v80/madras18a.html.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018)*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 2018. URL https://proceedings.mlr.press/v81/menon18a.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019. URL https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Machine Learning and Knowledge Discovery in Databases (ECML-PKDD 2017), Part I*, volume 10534 of *Lecture Notes in Computer Science*, pages 339–355. Springer, 2017. doi: 10.1007/978-3-319-71249-9_21.

Adrián Pérez-Suay, Paula Gordaliza, Jean-Michel Loubes, Dino Sejdinovic, and Gustau Camps-Valls. Fair kernel regression through cross-covariance operators. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=MyQ1e1VQQ3.

Nancy S. Petersen and Melvin R. Novick. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 13(1):3–29, 1976. doi: 10.1111/j.1745-3984.1976.tb00178.x.

Allan Pinkus. *n-Widths in Approximation Theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete, 3. Folge*. Springer-Verlag, Berlin, 1985. doi: 10.1007/978-3-642-69894-1.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 2000. doi: 10.7551/mitpress/1113.003.0008.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5680–5689, 2017. URL https://papers.neurips.cc/paper/7151-on-fairness-and-calibration.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2001. ISBN 9780262194754. doi: 10.7551/mitpress/4175.001.0001. URL https://mitpress.mit.edu/9780262194754/learning-with-kernels/.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL https://jmlr.org/papers/v11/sriperumbudur10a.html.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001. URL https://jmlr.org/papers/v2/steinwart01a.html.

Rush T. Stewart, Benjamin Eva, Shanna Slank, and Reuben Stern. An impossibility theorem for base rate tracking and equalized odds. *Analysis*, 84(4):778–787, 2024. doi: 10.1093/analys/anae039.

Robert L. Thorndike. Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70, 1971. doi: 10.1111/j.1745-3984.1971.tb00907.x.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of AI systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023. URL https://www.jmlr.org/papers/v24/23-0389.html.

Shaokui Wei, Jiayin Liu, Bing Li, and Hongyuan Zha. Mean parity fair regression in RKHS. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *PMLR*, pages 4602–4628, 2023. URL https://proceedings.mlr.press/v206/wei23a.html.

Susan Wei and Marc Niethammer. The fairness–accuracy Pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):287–302, 2022. doi: 10.1002/sam.11560.

Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, volume 28 of *PMLR*, pages 325–333, 2013. URL https://proceedings.mlr.press/v28/zemel13.html.

Han Zhao and Geoffrey J. Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022. URL https://www.jmlr.org/papers/v23/21-1427.html.

A Kernels in Brief

This appendix explains the kernel tools of Section 2.2 for readers who are new to them; Schölkopf and Smola [2001] and Gretton et al. [2012] give full treatments.

Feature map and reproducing property. A kernel $k(x, x')$ measures how similar two individuals are. The Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2))$, for example, is close to 1 when x and x' are close and decays to 0 as they move apart. A positive-definite kernel is an inner product between features, $k(x, x') = \langle \phi(x),\, \phi(x') \rangle_{\mathcal{H}}$, where the feature map ϕ takes values in a Hilbert space $\mathcal{H}$, the RKHS. The elements of $\mathcal{H}$ are functions on $\mathcal{X}$, and evaluating one is again an inner product, $f(x) = \langle f,\, \phi(x) \rangle_{\mathcal{H}}$. This reproducing property makes every test function $f \in \mathcal{H}$ a vector in the same space as the features. For the Gaussian kernel, $\mathcal{H}$ is infinite-dimensional, and its functions approximate every continuous function on a compact set.

Mean embeddings and characteristic kernels. Averaging the features over a distribution P gives its mean embedding $\mu_P = \mathbb{E}_P[\phi(X)]$, a single vector that represents P, and the expectation of any $f \in \mathcal{H}$ is $\mathbb{E}_P[f(X)] = \langle f,\, \mu_P \rangle_{\mathcal{H}}$. How much of P the mean embedding records depends on the kernel. For the linear kernel $k(x, x') = x^\top x'$, $\phi(x) = x$, and μ_P is the ordinary mean, which many distributions share. A kernel is characteristic when the mean embedding records the whole distribution: $\mu_P = \mu_Q$ only if $P = Q$. The Gaussian, Laplace and Matérn kernels on $\mathbb{R}^d$ are characteristic. With a characteristic kernel, the difference $\delta = \mu_a - \mu_b$ is zero exactly when the groups' distributions are equal.

MMD and the witness. The MMD is the length of this difference, $\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}$. By the Cauchy–Schwarz inequality, it is the largest difference in expectation that a test function with $\|f\|_{\mathcal{H}} \le 1$ can detect, and the witness, the unit vector along $\mu_P - \mu_Q$, detects it. The MMD can be estimated from kernel evaluations alone. The unbiased estimate of $\mathrm{MMD}^2(P, Q)$ averages k over pairs of distinct points within each sample, adds the two averages, and subtracts twice the average over pairs drawn across the two samples [Gretton et al., 2012].

Covariance operator. For a distribution Q, the covariance operator $\Sigma = \mathbb{E}_Q[(\phi(X) - \mu_Q) \otimes (\phi(X) - \mu_Q)]$, with $(f \otimes f')\, v := \langle f',\, v \rangle_{\mathcal{H}}\, f$, generalizes the covariance matrix: for a test function $v \in \mathcal{H}$, $\langle v,\, \Sigma v \rangle_{\mathcal{H}}$ is the variance of $v(X)$ under Q. For a bounded kernel, Σ has eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ with a finite sum, and orthonormal eigenvectors $e_1, e_2, \ldots$. Among unit test functions orthogonal to $e_1, \ldots, e_{j-1}$, the eigenvector e_j has the largest variance, λ_j. Since the eigenvalues have a finite sum, they decay, and a few directions carry most of the variation. On n samples, they are the eigenvalues of the centered kernel matrix divided by n, as in kernel principal component analysis [Schölkopf and Smola, 2001].

B Proofs for Section 3

This appendix proves the results of Section 3. Throughout, k is a bounded measurable kernel with separable RKHS $\mathcal{H}$, so that mean embeddings and covariance operators exist, and $\kappa^2 := \sup_x k(x, x)$.

B.1 Groups That Satisfy a Checklist yet Differ

Theorem 3.1 assumes that the groups differ. The lemma below shows that, for every finite checklist, there are two groups that satisfy it yet differ, by an amount we can choose. It reweights one group by a function that every criterion misses: m criteria and the total probability impose only $m + 1$ conditions, so a reweighting with $m + 2$ free values exists.

Lemma B.1 (Realizability)**.** *Let k be a bounded characteristic kernel on a separable metric space $\mathcal{X}$, and let P be a distribution on $\mathcal{X}$ whose support contains infinitely many points. For every finite checklist $v_1, \ldots, v_m \in \mathcal{H}$, there is $\gamma_{\max} > 0$ such that, for every $\gamma \in (0, \gamma_{\max}]$, there are group distributions $P_a = P$ and $P_b \ne P_a$ that satisfy every criterion, $\langle v_i,\, \delta \rangle_{\mathcal{H}} = 0$, and have $\mathrm{MMD}(P_a, P_b) = \gamma$. The distribution P_b has a bounded, nonnegative density with respect to P.*

Proof. The criteria are bounded functions, since $|v_i(x)| = |\langle v_i,\, \phi(x) \rangle_{\mathcal{H}}| \le \kappa \|v_i\|_{\mathcal{H}}$.

A reweighting that every criterion misses. Choose $m+2$ distinct points in the support of P, and disjoint open balls $A_1, \dots, A_{m+2}$ around them; each ball has $P(A_j) > 0$. A step function $h = \sum_j c_j \mathbf{1}_{A_j}$ must satisfy $\int h\, dP = 0$ and $\int v_i\, h\, dP = 0$ for $i = 1, \dots, m$. These are $m+1$ homogeneous linear equations in the $m+2$ unknowns c_j, so they have a nonzero solution.

The reweighted groups satisfy every criterion. For $0 < t \le 1/\|h\|_\infty$, let $P_a := P$ and $dP_b := (1 + th)\, dP$. Since $1 + th \ge 0$ and $\int h\, dP = 0$, P_b is a probability distribution, and $P_b \ne P_a$ because $h \ne 0$ on a set of positive probability. The difference between the groups is $\delta = -t\, g_h$ with $g_h := \int \phi\, h\, dP$, so $\langle v_i,\, \delta\rangle_{\mathcal{H}} = -t \int v_i\, h\, dP = 0$ for every i.

The difference is nonzero and can have any size up to $\gamma_{\max}$. Write $h = h^+ - h^-$ for its positive and negative parts. Since $\int h\, dP = 0$, both parts have the same mass $c_0 > 0$, so $Q^\pm := c_0^{-1} h^\pm\, dP$ are probability distributions. They live on disjoint sets, so they differ, and since k is characteristic, $g_h = c_0\,(\mu_{Q^+} - \mu_{Q^-}) \ne 0$. Hence $\mathrm{MMD}(P_a, P_b) = t\|g_h\|_{\mathcal{H}}$ takes every value in $(0, \gamma_{\max}]$ as t ranges over $(0, 1/\|h\|_\infty]$, with $\gamma_{\max} := \|g_h\|_{\mathcal{H}}/\|h\|_\infty$. □

Scope of the lemma. The kernel enters only to show that the groups differ in MMD; the criteria need only be bounded. The lemma therefore also holds for test functions outside $\mathcal{H}$, such as a classifier's decisions or the threshold indicators $\mathbf{1}[s > t]$ of a score, and within an outcome class, by reweighting the distribution of X given $Y = y$ in one group. Applied to a score that takes infinitely many values, it shows that finitely many thresholds certify neither independence (C1) nor separation (C2) (Section 4.1). The assumption of infinite support cannot be dropped. If P takes only d values, $d - 1$ criteria that, together with the constant function, span all functions on these values certify $P_a = P_b$ among distributions on them. Likewise, a kernel with a D-dimensional feature map, such as the linear kernel on $\mathbb{R}^D$, is not characteristic when $\mathcal{X}$ has more than $D+1$ points. Then D criteria that span its directions certify only that D expectations agree: the lemma, applied to these D coordinates, gives groups that agree on all of them yet differ.

What a checklist certifies. Certifying a class of criteria differs from certifying that the groups have equal distributions. A finite-dimensional class is certified by criteria that span it; the d averages of the input features, for example, certify every criterion that is linear in the features. Equal distributions cannot be certified by finitely many criteria when the data take infinitely many values, and this is what the Pokémon theorem concerns (Section 8). Section E.7 measures both on the three datasets. The witness of Theorem 3.1 shows that the groups' distributions differ; whether the difference matters for a particular decision depends on the application.

B.2 Proof of Theorem 3.2 and Corollary 3.3

The proof computes the residual left by the leading eigendirections, then shows that no other m criteria leave less. For a bounded kernel and separable $\mathcal{H}$, Σ is positive and trace-class. Its eigenvectors e_j with positive eigenvalues span the closure of its range, which contains the ellipsoid $B_r(R)$. If Σ has rank at most m, then $\lambda_{m+1} = 0$, and criteria that span its range force $\delta = 0$, so both sides of (8) vanish. Otherwise $\lambda_{m+1} > 0$.

Upper bound. Let $V_m := \mathrm{span}\{e_1, \dots, e_m\}$, and let $\delta = \Sigma^r u \in B_r(R)$ satisfy every criterion in V_m. With $u_j := \langle u,\, e_j\rangle_{\mathcal{H}}$, the criteria give $\langle e_j,\, \delta\rangle_{\mathcal{H}} = \lambda_j^r u_j = 0$ for $j \le m$, so

$$\|\delta\|_{\mathcal{H}}^2 \;=\; \sum_{j>m} \lambda_j^{2r} u_j^2 \;\le\; \lambda_{m+1}^{2r} \|u\|_{\mathcal{H}}^2 \;\le\; R^2 \lambda_{m+1}^{2r},$$

with equality at $\delta = R\lambda_{m+1}^r e_{m+1}$.

Lower bound. Let V be any subspace with $\dim V \le m$. The span of $e_1, \dots, e_{m+1}$ has dimension $m+1$, so it contains a unit vector $w = \sum_{j \le m+1} w_j e_j$ orthogonal to V. The vector $\delta^\star := R\lambda_{m+1}^r w$ satisfies every criterion in V and has norm $R\lambda_{m+1}^r$. It lies in $B_r(R)$, since $\delta^\star = \Sigma^r u$ with $u = R\sum_{j \le m+1} w_j (\lambda_{m+1}/\lambda_j)^r e_j$, and

$$\|u\|_{\mathcal{H}}^2 \;=\; R^2 \sum_{j \le m+1} w_j^2 \Big(\frac{\lambda_{m+1}}{\lambda_j}\Big)^{2r} \;\le\; R^2.$$

The two bounds prove (8). Dropping the requirement $\delta \perp V$ and measuring the part of δ orthogonal to V gives the same value, which is the Kolmogorov m-width of the ellipsoid [Pinkus, 1985]. Under **(A2)**, $\lambda_{m+1}^r \asymp (m+1)^{-\alpha r} \asymp m^{-\alpha r}$, which proves Corollary 3.3.

B.3 When the Source Condition Holds

The source condition **(A1)** limits how much of the difference between the groups can lie in directions of small variance. With $r = 1/2$, it holds for every pair of groups, with a radius set by how different they are. Let $h := dP_a/dQ - dP_b/dQ$ be the difference between the groups' densities with respect to their average $Q = \frac{1}{2}(P_a + P_b)$. Each density lies between 0 and 2, because $Q \geq \frac{1}{2}P_g$, so $|h| \leq 2$.

Proposition B.2 (The source condition holds with $r = 1/2$)**.** *The difference between the groups is* $\delta = \Sigma^{1/2}u$ *with* $\|u\|_{\mathcal{H}} \leq \|h\|_{L^2(Q)}$*, and*

$$\|h\|_{L^2(Q)}^2 \;=\; 2\int \frac{(dP_a - dP_b)^2}{d(P_a + P_b)} \;\leq\; 4\,\mathrm{TV}(P_a, P_b) \;\leq\; 4,$$

where TV *is the total variation distance. Hence* ***(A1)*** *holds with* $r = 1/2$ *and* $R = 2$ *for every pair of groups.*

Proof. Define $\Gamma g := \mathbb{E}_Q[g(X)\,(\phi(X) - \mu_Q)]$ for $g \in L^2(Q)$. Its adjoint is $\Gamma^* v = \langle v,\, \phi(\cdot) - \mu_Q\rangle_{\mathcal{H}}$, so $\Gamma\Gamma^* = \Sigma$. Since $\mathbb{E}_Q[h] = 0$, $\delta = \mathbb{E}_Q[h(X)\,\phi(X)] = \Gamma h$. The polar decomposition gives $\Gamma = \Sigma^{1/2}W$ with $\|W\| \leq 1$, so $\delta = \Sigma^{1/2}u$ with $u = Wh$ and $\|u\|_{\mathcal{H}} \leq \|h\|_{L^2(Q)}$. The identity for $\|h\|_{L^2(Q)}^2$ follows from $dQ = \frac{1}{2}d(P_a + P_b)$, and $|h| \leq 2$ gives $\|h\|_{L^2(Q)}^2 \leq 2\,\mathbb{E}_Q|h| = 2\int |dP_a - dP_b| = 4\,\mathrm{TV}(P_a, P_b)$. □

For a characteristic kernel, the radius is exact, $\|\Sigma^{-1/2}\delta\|_{\mathcal{H}} = \|h\|_{L^2(Q)}$ [Harchaoui et al., 2008], so no smaller radius works for $r = 1/2$.

When the worst case is attained. Theorem 3.2 takes the worst case over every vector in the ellipsoid $B_r(R)$, and not every such vector is a difference between two distributions. For the best m criteria, a pair of distributions attains the worst case when the eigenfunctions are bounded. Let $\psi_j := \langle \phi(\cdot) - \mu_Q,\, e_j\rangle_{\mathcal{H}}/\lambda_j^{1/2}$. These functions have mean zero and are orthonormal in $L^2(Q)$. Suppose $|\psi_{m+1}| \leq M$. Then $P_a := (1 + \psi_{m+1}/M)\,Q$ and $P_b := (1 - \psi_{m+1}/M)\,Q$ are distributions whose average is Q, so their pooled covariance is Σ. Their difference is $\delta = (2/M)\,\Gamma\psi_{m+1} = (2/M)\,\lambda_{m+1}^{1/2} e_{m+1}$. It satisfies the criteria $e_1, \ldots, e_m$, and its residual unfairness $(2/M)\,\lambda_{m+1}^{1/2}$ equals the worst case of Theorem 3.2 over $B_{1/2}(R)$ with $R = 2/M$. Since ψ_{m+1} has unit norm in $L^2(Q)$, $M \geq 1$, so this radius is at most the radius 2 of Proposition B.2. Without a bound on the eigenfunctions, no pair of distributions need attain the worst case over the ellipsoid.

B.4 Inexact Criteria

When criteria hold only up to a tolerance, $|\langle v_i,\, \delta\rangle_{\mathcal{H}}| \leq \varepsilon$, they bound only the part $P_V\delta$ of the difference that the checklist sees. Pythagoras gives $\|P_{V^\perp}\delta\|_{\mathcal{H}}^2 = \mathrm{MMD}^2(P_a, P_b) - \|P_V\delta\|_{\mathcal{H}}^2$, which is (9). For orthonormal criteria, $\|P_V\delta\|_{\mathcal{H}}^2 = \sum_{i=1}^m \langle v_i,\, \delta\rangle_{\mathcal{H}}^2 \leq m\varepsilon^2$. For criteria that are not orthonormal, let $K_{ij} := \langle v_i,\, v_j\rangle_{\mathcal{H}}$ be their Gram matrix and $z_i := \langle v_i,\, \delta\rangle_{\mathcal{H}}$ their readings. Then $\|P_V\delta\|_{\mathcal{H}}^2 = z^\top K^\dagger z$, where $K^\dagger$ is the Moore–Penrose pseudoinverse, so $\|P_V\delta\|_{\mathcal{H}} \leq \varepsilon\sqrt{m/\lambda_{\min}^+(K)}$, where $\lambda_{\min}^+(K)$ is the smallest positive eigenvalue of K. Whenever the residual is nonzero, the unit vector $P_{V^\perp}\delta/\|P_{V^\perp}\delta\|_{\mathcal{H}}$ is a criterion orthogonal to the checklist, and it detects a difference of $\|P_{V^\perp}\delta\|_{\mathcal{H}}$ between the groups.

B.5 Criteria in Different Distributions

Section 3.3 applies the Pokémon theorem to criteria that compare distributions other than that of X. Each criterion compares the expectation of a test function between the groups under one of three distributions. By the reproducing property, each comparison is an inner product with the difference between the groups' mean embeddings of that distribution.

The distributions of X and of X given Y. Demographic parity compares $\mathbb{E}[c(X) \mid G = g]$ between the groups, with $c(x) = \mathbf{1}[f(x) = 1]$. Equal error rates compare $\mathbb{E}[c(X) \mid Y = y, G = g]$ with the same c: the true-positive rate for $y = 1$ and the false-positive rate for $y = 0$. For $c = \langle v,\, \phi(\cdot)\rangle_{\mathcal{H}}$, the comparisons are $\langle v,\, \delta\rangle_{\mathcal{H}}$ and $\langle v,\, \delta_y\rangle_{\mathcal{H}}$, with δ_y from (5).

The joint distribution of (X, Y). Calibration involves the outcome itself, so it needs a kernel on pairs (x, y). We use $k((x, y), (x', y')) := k(x, x')\,\mathbf{1}[y = y']$. Its feature map is $\psi(x, y) := \phi(x) \otimes e_y$, with e_0, e_1 the standard basis of $\mathbb{R}^2$, and its RKHS consists of the functions $f(x, y) = f_y(x)$ with $f_0, f_1 \in \mathcal{H}$. Write $\nu_g := \mathbb{E}[\psi(X, Y) \mid G = g]$ for the mean embedding of group g's joint distribution. Its two parts are $\Pr[Y = y \mid G = g]\,\mu_{y,g}$ for $y = 0, 1$, the embeddings of the distribution of X jointly with $Y = y$. If k is injective on finite signed measures, as the Gaussian, Laplace and Matérn kernels are, each part determines that distribution, so the kernel on pairs is characteristic.

Calibration of a score $S = s(X)$ within group g, $\mathbb{E}[Y \mid S, G = g] = S$, holds if and only if $\mathbb{E}[c_h(X, Y) \mid G = g] = 0$ for every bounded function h, where $c_h(x, y) := (y - s(x))\,h(s(x))$, since $\mathbb{E}[c_h(X, Y) \mid G = g] = \mathbb{E}[(\mathbb{E}[Y \mid S, G = g] - S)\,h(S) \mid G = g]$. It suffices to check h in any class that is dense in L^2 of the distribution of S in group g. When $x \mapsto (y - s(x))\,h(s(x))$ lies in $\mathcal{H}$ for $y = 0, 1$, the test function c_h lies in the RKHS on pairs, and $\mathbb{E}[c_h(X, Y) \mid G = g] = \langle c_h,\, \nu_g\rangle$. Calibration within both groups then gives $\langle c_h,\, \nu_a - \nu_b\rangle = 0$ for every such h. Calibration is therefore a family of criteria in the joint distribution, and the Pokémon theorem applies to it with $\nu_a - \nu_b$ in place of δ. Its member $h \equiv 1$ says that each group's average score equals its base rate. Section B.6 shows that this family, although infinite, certifies nothing beyond the score.

B.6 Infinitely Many Criteria

Satisfying every criterion in a family $C \subset \mathcal{H}$ makes δ orthogonal to the closed span of C. If this span is all of $\mathcal{H}$, then $\delta = 0$, and for a characteristic kernel $P_a = P_b$: a dense family certifies. If instead some eigenvector e_j of the covariance operator of a distribution Q is orthogonal to every criterion in C, the family certifies nothing, however many criteria it contains. With ψ_j as in Section B.3, $|\psi_j| \le 2\kappa/\lambda_j^{1/2}$, so for $0 < t \le \lambda_j^{1/2}/(2\kappa)$ the distributions $P_a := (1 + t\,\psi_j)\,Q$ and $P_b := (1 - t\,\psi_j)\,Q$ exist. They differ, with $\delta = 2t\lambda_j^{1/2} e_j$, and satisfy every criterion in C.

Density is sufficient for certification but not necessary. For a kernel that is injective on finite signed measures, the family $\{c \in \mathcal{H} : c(x_0) = 0\}$ is not dense, since $\phi(x_0)$ is orthogonal to it, yet it certifies. Satisfying it forces $\delta = t\,\phi(x_0)$ for some $t \in \mathbb{R}$, the mean embedding of t times the point mass at x_0. Injectivity then makes $P_a - P_b$ equal to t times this point mass, and since both groups have total mass one, $t = 0$.

Calibration. Calibration of a score $S = s(X)$ within each group is the family $\mathbb{E}[(Y - S)\,h(S) \mid G = g] = 0$ over bounded functions h (Section B.5). It contains infinitely many criteria, but each depends on the data only through the pair (S, Y). The proposition below shows that two groups can agree on every such criterion yet have different features.

Proposition B.3 (Criteria that see only the score and the outcome)**.** *Let s be measurable, and let π be a joint distribution of (X, Y) under which X carries information beyond (S, Y): some bounded test function c has $\Pr_\pi[c(X) \neq \mathbb{E}_\pi[c(X) \mid S, Y]] > 0$. Then there is a joint distribution π' that gives every bounded function of (S, Y) the same expectation as π, but a different distribution of X.*

Proof. Let $\bar{c}(S, Y) := \mathbb{E}_\pi[c(X) \mid S, Y]$ and $h(x, y) := c(x) - \bar{c}(s(x), y)$. Then h is bounded, nonzero on a set of positive probability, and $\mathbb{E}_\pi[h \mid S, Y] = 0$. For $0 < t \le 1/\|h\|_\infty$, let $d\pi' := (1 + th)\,d\pi$, a probability distribution. For every bounded function ξ of (S, Y), $\mathbb{E}_{\pi'}[\xi(S, Y)] = \mathbb{E}_\pi[\xi(S, Y)] + t\,\mathbb{E}_\pi[\xi(S, Y)\,\mathbb{E}_\pi[h \mid S, Y]] = \mathbb{E}_\pi[\xi(S, Y)]$.

Under π', the distribution of X has density $1 + t\,(c(x) - Tc(x))$ with respect to that under π, where $Tc(x) := \mathbb{E}_\pi[\bar{c}(S, Y) \mid X = x]$. If $Tc = c$, then $\mathbb{E}_\pi[c(X)^2] = \mathbb{E}_\pi[c(X)\,\bar{c}(S, Y)] = \mathbb{E}_\pi[\bar{c}(S, Y)^2]$, so $\mathbb{E}_\pi[(c(X) - \bar{c}(S, Y))^2] = 0$, contrary to the assumption. Hence the two distributions of X differ. □

Taking π and π' as the joint distributions of the two groups, both groups have the same base rate, the same distribution of scores and the same error rates at every threshold. Every calibration criterion

takes the same value in both groups, so the score is calibrated in one group exactly when it is calibrated in the other. Yet their features differ, and with a characteristic kernel the Pokémon witness detects the difference.

C Proofs for Section 4

This appendix proves the results of Section 4. Throughout, $S \in [0,1]$ is a score, the kernel on scores is bounded and characteristic, and μ_g, $\mu_{y,g}$, δ and δ_y are the mean embeddings of the score's distributions, as in Section 4.1. For each group g, $\bar{g}$ denotes the other group. Identity (10) also holds with a and b exchanged,

$$\delta \;=\; p_b\,\delta_1 + (1-p_b)\,\delta_0 + (p_a - p_b)\,(\mu_{1,a} - \mu_{0,a}), \tag{12}$$

as substituting $\mu_{y,b} = \mu_{y,a} - \delta_y$ into $\delta = \mu_a - \mu_b$ shows.

C.1 Dense Checklists Certify the Three Notions

Section 4.1 states that a dense checklist for each distribution of the score certifies the corresponding notion. Write P_a and P_b for the groups' distributions of S, and $Q := \frac{1}{2}(P_a + P_b)$.

Independence (C1). The difference $P_a - P_b$ has a density with respect to Q that is bounded by 2 in absolute value and integrates to zero. The constant function and the indicators $\mathbf{1}[s > t]$ span the step functions, which are dense in $L^2(Q)$. Equal selection rates at every threshold make the density orthogonal in $L^2(Q)$ to each of these functions, hence to all of $L^2(Q)$, so it is zero. The two distributions are equal, and $\delta = 0$.

Separation (C2). The same argument within each outcome class shows that equal error rates at every threshold give $\delta_0 = \delta_1 = 0$.

Sufficiency (C3). For every bounded h, $\mathbb{E}[(Y - f(S))\,h(S) \mid G = g] = \mathbb{E}[(\mathbb{E}[Y \mid S, G = g] - f(S))\,h(S) \mid G = g]$. This vanishes for every h if and only if $\mathbb{E}[Y \mid S, G = g] = f(S)$ almost surely.

C.2 Proofs of Theorems 4.1 and 4.2

Averages of the score. Parts (ii) and (iii) of Theorem 4.1 and Theorem 4.2 use only averages of the score. Write $\bar{S}_{y,g} := \mathbb{E}[S \mid Y = y, G = g]$ for the average score in outcome class y of group g. Its distances from a perfect score, $\mathrm{FNR}_g := 1 - \bar{S}_{1,g}$ and $\mathrm{FPR}_g := \bar{S}_{0,g}$, are the score's error rates; for a binary decision they are the usual false-negative and false-positive rates. Let $d_y := \bar{S}_{y,a} - \bar{S}_{y,b}$ be the difference between the groups' average scores in class y, and $\beta_g := p_g - \mathbb{E}[S \mid G = g]$ the gap between group g's base rate and its average score. Calibration within group g gives $\beta_g = 0$.

The law of total expectation gives $p_g - \beta_g = \mathbb{E}[S \mid G = g] = p_g\,\bar{S}_{1,g} + (1-p_g)\,\bar{S}_{0,g}$ for each group. For $g = a$, substitute $\bar{S}_{y,b} = \bar{S}_{y,a} - d_y$ into the equation for b and subtract it from the equation for a. This gives $(p_a - p_b)(\bar{S}_{1,a} - \bar{S}_{0,a}) + p_b\,d_1 + (1-p_b)\,d_0 = (p_a - p_b) - (\beta_a - \beta_b)$, and since $\bar{S}_{1,a} - \bar{S}_{0,a} = 1 - \mathrm{FNR}_a - \mathrm{FPR}_a$,

$$(p_a - p_b)\,(\mathrm{FNR}_g + \mathrm{FPR}_g) \;=\; (\beta_a - \beta_b) + p_{\bar{g}}\,d_1 + (1 - p_{\bar{g}})\,d_0. \tag{13}$$

The case $g = b$ is symmetric.

Proof of Theorem 4.1. (i) Independence and separation give $\delta = \delta_0 = \delta_1 = 0$, so (10) leaves $(p_a - p_b)\,\omega = 0$, and $\mu_{1,b} = \mu_{0,b}$. Since the kernel is characteristic, S has the same distribution in the two outcome classes of group b. By separation, the same holds in group a, so the four distributions of S given Y and G are equal. The distribution of S given $Y = y$ mixes those of the two groups, so it is the same for $y = 0$ and $y = 1$, and S is independent of Y.

(ii) Let $T := f(S) \in [0,1]$. Sufficiency gives $\mathbb{E}[Y \mid T, G = g] = \mathbb{E}[\mathbb{E}[Y \mid S, G = g] \mid T, G = g] = T$, so T is calibrated within each group, and $\beta_a = \beta_b = 0$ for T. Separation of S implies separation of T, a function of S, so $d_0 = d_1 = 0$ for T. Then (13) for T gives $\mathrm{FNR}_g + \mathrm{FPR}_g = 0$ in each group. Both terms are nonnegative, so $\mathbb{E}[1 - T \mid Y = 1, G = g] = 0$ and $\mathbb{E}[T \mid Y = 0, G = g] = 0$, and since $T \in [0,1]$, $T = Y$ almost surely. A calibrated score has $f(s) = s$, so $S = Y$.

(iii) Sufficiency gives $\mathbb{E}[f(S) \mid G = g] = \mathbb{E}[Y \mid G = g] = p_g$, and independence makes these averages equal, so $p_a = p_b$. □

Proof of Theorem 4.2, and the version of Stewart et al. [2024]. KMR assume calibration within each group, so $\beta_a = \beta_b = 0$, and equal average scores within each outcome class, so $d_0 = d_1 = 0$. Then (13) gives $\mathrm{FNR}_g + \mathrm{FPR}_g = 0$, and $S = Y$ almost surely as in (ii). The proof uses only $\beta_a = \beta_b$, so it also proves the theorem of Stewart et al. [2024], who replace calibration by equal gaps between average score and base rate in the two groups. □

Relation to the original statements. Kleinberg et al. [2017] state their theorem for binned, possibly randomized scores, and conclude that the instance admits perfect prediction; their proof shows that the score itself is perfect. In Barocas et al. [2023], (i) and (iii) of Theorem 4.1 are Propositions 3 and 2. Their Proposition 4 states that separation and sufficiency cannot both hold when every event of (G, S, Y) has positive probability, an assumption that excludes perfect scores; (ii) shows that the perfect score is the only exception.

Equal error rates and equal predictive values. For a binary decision $\hat{Y}$ that is positive with positive probability, separation means common true- and false-positive rates TPR and FPR in both groups. Equal positive predictive value means $\mathbb{E}[(Y - \mathrm{PPV})\,\hat{Y} \mid G = g] = 0$ in both groups, a criterion linear in the distribution. Since $\mathbb{E}[Y\hat{Y} \mid G = g] = p_g\,\mathrm{TPR}$ and $\mathbb{E}[\hat{Y} \mid G = g] = p_g\,\mathrm{TPR} + (1 - p_g)\,\mathrm{FPR}$, it reads $\frac{p_g}{1-p_g}\,\mathrm{TPR}\,(1 - \mathrm{PPV}) = \mathrm{FPR}\ \mathrm{PPV}$. The right side is the same in both groups, so when $p_a \neq p_b$, $\mathrm{TPR}\,(1 - \mathrm{PPV}) = 0$: either $\mathrm{TPR} = 0$, or $\mathrm{PPV} = 1$ and $\mathrm{FPR} = 0$. This is the result of Chouldechova [2017]: unless a decision has no true positives or no false positives, equal error rates and equal predictive values are incompatible when base rates differ.

C.3 Approximate Impossibilities

When the criteria hold only up to tolerances, the conclusions of Theorem 4.1 hold approximately, with an error proportional to the tolerances. The gaps β_g and d_y measure how far calibration and equal class averages are from holding, and $\|\delta\|_{\mathcal{H}}$ and $\|\delta_y\|_{\mathcal{H}}$ how far independence and separation are.

Theorem C.1 (Approximate impossibilities)**.** *Let $S \in [0, 1]$, use any bounded kernel on scores for (i), and suppose $p_a \neq p_b$. For each group g:*

(i) Independence and separation: $|p_a - p_b| \cdot \|\mu_{1,g} - \mu_{0,g}\|_{\mathcal{H}} \leq \|\delta\|_{\mathcal{H}} + p_{\bar{g}}\,\|\delta_1\|_{\mathcal{H}} + (1 - p_{\bar{g}})\,\|\delta_0\|_{\mathcal{H}}$.

(ii) Separation and calibration: $|p_a - p_b| \cdot \mathbb{E}\big[|S - Y| \;\big|\; G = g\big] \leq |\beta_a - \beta_b| + p_{\bar{g}}\,|d_1| + (1 - p_{\bar{g}})\,|d_0|$.

(iii) Independence and calibration: $|p_a - p_b| \leq |\mathbb{E}[S \mid G = a] - \mathbb{E}[S \mid G = b]| + |\beta_a - \beta_b|$.

Proof. (i) Take norms in (10) for $g = b$ and in (12) for $g = a$. (ii) Take absolute values in (13), and use $\mathbb{E}[|S - Y| \mid G = g] = p_g\,\mathrm{FNR}_g + (1 - p_g)\,\mathrm{FPR}_g \leq \mathrm{FNR}_g + \mathrm{FPR}_g$. (iii) The definition of β_g gives $p_a - p_b = (\mathbb{E}[S \mid G = a] - \mathbb{E}[S \mid G = b]) + (\beta_a - \beta_b)$. □

With all tolerances zero, these give Theorem 4.1(i), and (ii) and (iii) for a calibrated score. If each group's average score is within η of its base rate, then $|\beta_a - \beta_b| \leq 2\eta$. By Markov's inequality, (ii) also bounds the tail, $\Pr[|S - Y| > t \mid G = g] \leq (|\beta_a - \beta_b| + p_{\bar{g}}|d_1| + (1 - p_{\bar{g}})|d_0|)/(|p_a - p_b|\,t)$. Approximate forms of KMR appear in Kleinberg et al. [2017, Theorem 1.2] and Pleiss et al. [2017, Theorem 9], with multiplicative tolerances; (ii) is linear in the tolerances. Part (iii) appears in the proof of Theorem 8 of Zhao and Gordon [2022] for a binary classifier, and part (i) sharpens Theorem 6.1 for a score.

C.4 Impossibilities with the Best m Criteria

A score that satisfies only the best m criteria of a notion satisfies that notion up to a tolerance, and Theorem 3.2 bounds the tolerance. To make the bound explicit, we need a kernel on scores whose covariance operators have a bounded trace and which measures average scores. The kernel $\min(s, s')$ has both properties.

The kernel $\min(s, s')$**.** On $[0,1]$, $\min(s,s') = \int_0^1 \mathbf{1}[t < s]\,\mathbf{1}[t < s']\,dt$, so this kernel has the feature map $s \mapsto \mathbf{1}_{[0,s)}$ into $L^2[0,1]$. Its RKHS consists of the functions f with $f(0) = 0$ and $f' \in L^2[0,1]$, with $\|f\|_{\mathcal{H}}^2 = \int_0^1 f'(t)^2\,dt$, and $k(s,s) \le 1$. The mean embedding of a distribution of S is $\mu(t) = \mathbb{E}[\min(S,t)] = \int_0^t \Pr[S > \tau]\,d\tau$. It determines $\Pr[S > \tau]$ for every τ, so the kernel is characteristic. The covariance operator of any distribution of S has trace $\int_0^1 \Pr[S > t]\,(1 - \Pr[S > t])\,dt \le 1/4$. Finally, $\iota(s) := s = k(s,1)$ has $\|\iota\|_{\mathcal{H}} = 1$ and $\langle \iota, \mu \rangle_{\mathcal{H}} = \mathbb{E}[S]$, so MMDs bound differences of average scores: $|d_y| \le \|\delta_y\|_{\mathcal{H}}$ and $|\mathbb{E}[S \mid G = a] - \mathbb{E}[S \mid G = b]| \le \|\delta\|_{\mathcal{H}}$.

The best m **criteria.** The best m criteria of independence (C1) are the m leading eigendirections of the covariance operator of the score's distribution, pooled over both groups. Those of separation (C2) are, for each y, the m leading eigendirections of the covariance operator of its distribution given $Y = y$, pooled over both groups.

Corollary C.2 (Impossibilities with the best m criteria)**.** *Use the kernel* $\min(s, s')$ *on scores, let* $\eta_m := 1/\sqrt{m+1}$*, and suppose* $p_a \neq p_b$*. For each group* g*:*

(i) If S *satisfies the best* m *criteria of independence (C1) and of separation (C2), its predictive signal in group* g *satisfies* $|p_a - p_b| \cdot \|\mu_{1,g} - \mu_{0,g}\|_{\mathcal{H}} \le 2\eta_m$*.*

(ii) If S *satisfies the best* m *criteria of separation (C2), then* $|p_a - p_b| \cdot \mathbb{E}\big[|S - Y| \;\big|\; G = g\big] \le |\beta_a - \beta_b| + \eta_m$*. For a calibrated score,* $\beta_a = \beta_b = 0$*.*

(iii) If S *satisfies the best* m *criteria of independence (C1), then* $|p_a - p_b| \le |\beta_a - \beta_b| + \eta_m$*.*

Proof. By Theorem C.1 and the bounds on $|d_y|$ and on the difference of average scores above, it suffices to show $\|\delta\|_{\mathcal{H}} \le \eta_m$ and $\|\delta_y\|_{\mathcal{H}} \le \eta_m$. Satisfying the best m criteria of independence places δ orthogonal to the leading m eigenvectors of the covariance operator Σ of $Q = \frac{1}{2}(P_a + P_b)$, the pooled distribution of S. By Proposition B.2, $\delta = \Sigma^{1/2} u$ with $\|u\|_{\mathcal{H}} \le 2$, so the upper bound of Theorem 3.2 gives $\|\delta\|_{\mathcal{H}} \le 2\lambda_{m+1}(\Sigma)^{1/2}$. Since $(m+1)\,\lambda_{m+1} \le \lambda_1 + \cdots + \lambda_{m+1} \le \operatorname{tr}\Sigma \le 1/4$, this is at most $2 \cdot \frac{1}{2}(m+1)^{-1/2} = \eta_m$. The same argument, applied to the distributions of S given $Y = y$, gives $\|\delta_y\|_{\mathcal{H}} \le \eta_m$. □

As $m \to \infty$, (i) gives Theorem 4.1(i), and (ii) and (iii) give Theorem 4.1(ii) and (iii) for a calibrated score. The bounds shrink at the rate $m^{-1/2}$, or $m^{-\alpha/2}$ under **(A2)**. The best m criteria depend on the score's own distributions, so the corollary describes the best possible checklist. It is a result about rates: it shows how quickly the conclusions of Theorem 4.1 return as a score satisfies more of the best criteria. Its bounds hold for every score and do not adapt to the data, so they are not a certificate for a particular score.

D Proofs for Sections 5 and 6

This appendix proves the results of Sections 5 and 6. Throughout, the kernel on the feature space $\mathcal{Z}$ is bounded and characteristic, and $\mathcal{H}$ denotes its RKHS. As in Section 6, μ_g, $\mu_{y,g}$, δ and δ_y are the mean embeddings of the distributions of the features $Z = \Phi(X)$, not of X, and $\omega_g := \mu_{1,g} - \mu_{0,g}$. Identities (10) and (12) hold for Z as for S, since their derivation uses only the law of total expectation.

D.1 Proof of Theorem 5.1

The proof is that of Theorem 4.1(i), with the features Z in place of the score S. Independence (C1) and separation (C2) give $\delta = \delta_0 = \delta_1 = 0$, so (10) leaves $(p_a - p_b)\,\omega_b = 0$. Since $p_a \neq p_b$, $\mu_{1,b} = \mu_{0,b}$, and since the kernel is characteristic, Z has the same distribution in the two outcome classes of group b. By separation, the same holds in group a, so the four distributions of Z given Y and G are equal, and Z is independent of Y. Hence every function f of the features with $\mathbb{E}|f(Z)| < \infty$ has $\mathbb{E}[f(Z) \mid Y = 1] = \mathbb{E}[f(Z) \mid Y = 0]$, within each group and overall: no model built on Z separates the two outcome classes, even on average. □

D.2 Proofs of Theorem 6.1 and Proposition 6.2

Proof of Theorem 6.1. By (10) and (12),

$$(p_a - p_b)\,\omega_b \;=\; \delta - p_a\,\delta_1 - (1 - p_a)\,\delta_0, \qquad (p_a - p_b)\,\omega_a \;=\; \delta - p_b\,\delta_1 - (1 - p_b)\,\delta_0.$$

Taking norms, with $\|\delta\|_{\mathcal{H}} = \varepsilon$ and $\|\delta_y\|_{\mathcal{H}} \le \rho$, gives $|p_a - p_b|\,\|\omega_g\|_{\mathcal{H}} \le \varepsilon + p_{\bar{g}}\,\rho + (1 - p_{\bar{g}})\,\rho = \varepsilon + \rho$ for each group g, where $\bar{g}$ is the other group. This is (11). □

Proof of Proposition 6.2. *Separated features lie on the bound.* Let Z satisfy separation, so $\mu_{y,a} = \mu_{y,b}$ for each y, and both groups have the same signal $\omega := \omega_a = \omega_b$. Then (10) gives $\delta = (p_a - p_b)\,\omega$, so $\rho = 0$ and $\varepsilon = |p_a - p_b|\,\|\omega\|_{\mathcal{H}}$.

Mixing with a constant shrinks every difference. Fix $z_0 \in \mathcal{Z}$ and $\lambda \in [0, 1]$, and let Z' equal Z with probability λ and z_0 otherwise, using a coin independent of (X, Y, G). Each distribution of Z', overall or given Y and G, mixes the corresponding distribution of Z with the point mass at z_0. Its mean embedding is therefore $\lambda\mu + (1 - \lambda)\,\phi(z_0)$, where μ is that of Z and ϕ is the feature map of the kernel on $\mathcal{Z}$. Every difference between two of these mean embeddings is multiplied by λ: Z' satisfies separation, its independence violation is $\lambda\varepsilon$, and its signal is $\lambda\,\|\omega\|_{\mathcal{H}}$ in both groups.

Every point below $\|\omega\|_{\text{sep}}$ is reached. Let $\beta < |p_a - p_b|\,\|\omega\|_{\text{sep}}$. By the definition of the supremum, there are separated features Z with $\|\omega\|_{\mathcal{H}} > \beta/|p_a - p_b|$. Mixing them with $\lambda := \beta/(|p_a - p_b|\,\|\omega\|_{\mathcal{H}}) < 1$ gives features with $\rho = 0$, $\varepsilon = \beta$, and signal $\beta/|p_a - p_b|$ in both groups. By Theorem 6.1, no features with $\varepsilon + \rho \le \beta$ have more signal in either group. □

D.3 Error of Classifiers with Equal Error Rates

For a binary decision that satisfies separation, (10) ties the gap in selection rates to the decision's signal, and this gives a lower bound on its error. Menon and Williamson [2018] characterize the Bayes-optimal trade-off between accuracy and fairness for given constraints; the following bound holds for every distribution.

Corollary D.1 (Error under equal error rates)**.** *Suppose $p_a \neq p_b$, and let $p := \Pr[Y = 1]$. Let a decision $\hat{Y} \in \{0, 1\}$ satisfy separation, with true- and false-positive rates* TPR *and* FPR *common to both groups, and let $\Delta_{\text{sel}} := |\Pr[\hat{Y} = 1 \mid G = a] - \Pr[\hat{Y} = 1 \mid G = b]|$ be its gap in selection rates. Then*

$$\Pr[\hat{Y} \neq Y] \;\ge\; \min(p, 1 - p)\,\Big(1 - \frac{\Delta_{\text{sel}}}{|p_a - p_b|}\Big), \tag{14}$$

with equality when $p \le 1/2$ and $\text{FPR} = 0$, or $p \ge 1/2$ and $\text{TPR} = 1$.

Proof. By the law of total expectation, $\Pr[\hat{Y} = 1 \mid G = g] = \text{FPR} + p_g\,(\text{TPR} - \text{FPR})$, so $\Delta_{\text{sel}} = |p_a - p_b|\,|\,\text{TPR} - \text{FPR}\,|$. The error is

$p\,(1 - \text{TPR}) + (1 - p)\;\text{FPR} \;\ge\; \min(p, 1 - p)\,(1 - \text{TPR} + \text{FPR}) \;\ge\; \min(p, 1 - p)\,(1 - |\,\text{TPR} - \text{FPR}\,|)$,

which is (14). Both inequalities are equalities under the stated conditions. □

E Experimental Details

This appendix describes the two experiments of Section 7: E1, the residual after the best m criteria (Figure 1A), and E2, fairness and predictive signal (Figure 1B and C). Full protocols and the adversarial reviews of both experiments are released with the code.

E.1 Data, Splits and Intervals

Datasets. Adult Income (UCI) predicts income above \$50K, with men and women as the groups ($|\Delta p| = .199$). COMPAS (ProPublica) predicts two-year recidivism, with Black and white defendants as the groups, after the standard filters (5,278 rows, $|\Delta p| = .132$). ACS PUMS uses the ACSIncome task on California 2018 data from `folktables` [Ding et al., 2021], with white and non-white respondents as the groups ($|\Delta p| = .086$). The group is a function of the race column, so we remove that column from the features. The two experiments use ACS PUMS differently. For each seed, E1 draws a subsample of at most 20,000 rows, stratified by outcome and group (all rows on COMPAS), and one-hot encodes the ACS features. E2 uses numeric features on all 195,665 ACS rows.

Splits and intervals. Both experiments use the same ten seeds. In E2, each seed splits a dataset into training, validation and test rows (70/15/15), stratified by outcome and group. Intervals are 95% two-level bootstrap intervals of the mean over seeds ($B = 1000$; seeds, then rows stratified by outcome and group). Seeds re-split the same rows, so these intervals are conditional on the dataset. On COMPAS, whose seeds share most of their rows, E1 instead uses half-widths from 200 resamples of the whole dataset.

E.2 Reading Figure 1

Two held-out statistics. E2 can average its test-row values over seeds in two ways. The 10-seed mixture applies the rule of a randomly chosen seed. The mean over seeds averages what each seed's fitted rule does, which noise biases upward. A single fitted rule lies between the two. All test-row marks in Figure 1 are 10-seed mixtures, and we never compare a value of one statistic with a value of the other.

Panel A. The blue line is the residual share $r(m)$ of Section E.3 after the top m eigendirections, averaged over seeds, with its 95% interval. The red line is the worst case of Theorem 3.2 with $r = 1/2$ and $R = 2$ (Proposition B.2); on this axis it is $4\lambda_{m+1}/\operatorname{MMD}^2(P_a, P_b)$, with the eigenvalues estimated on separate rows. It is a worst case over all pairs of groups with this pooled covariance, not a prediction for this pair, so the panel does not test the theorem. The blue line stops where it falls into the gray noise level. ACS PUMS uses one-hot features without race.

Panels B and C. Both axes are in the decision units of Section 6, measured on test rows, and the red line is the bound of Theorem 6.1. The bound holds on any set of rows, so that no point lies above it follows by algebra and is not evidence for the theorem. In panel B, each point is one encoder at one setting of its fairness weight, and a logistic regression trained on its features, thresholded at $1/2$, makes the decisions. In panel C, the line is the frontier of the linear program on the rows it is fit to, and the circles apply the same rules to test rows; compare methods with the circles, not with the line. The black point is $\|\omega\|_{\text{sep}}$ on the fit rows. Below it, the circles are rescalings of one rule (Proposition 6.2) and give a single test. The gray band is where a rule exactly on the bound, refit on each seed, lands on test rows with probability 95%; it stops at $\|\omega\|_{\text{sep}}$, beyond which no rule lies on the bound. The triangles are exponentiated gradient at tolerances .005 and .02, the diamond is post-processing, and the gray points repeat panel B (Section E.4).

E.3 E1: Residual After the Best m Criteria

Protocol. We use a Gaussian kernel on standardized features with the median bandwidth ($0.5\times$ and $2\times$ in Figure 3), and pool the groups with equal weights, $Q = \frac{1}{2}(P_a + P_b)$. Each seed splits its rows three times into a part A with half of the rows, which estimates the eigendirections, and parts B_1 and B_2 with a quarter each, which give two estimates $\hat{\delta}_1$ and $\hat{\delta}_2$ of the difference between the groups. For orthonormal eigendirections $q_1, \ldots, q_m$ estimated on A, the residual share is

$$r(m) \;=\; 1 - \frac{\sum_{l \le m} \langle \hat{\delta}_1,\, q_l \rangle_{\mathcal{H}} \langle \hat{\delta}_2,\, q_l \rangle_{\mathcal{H}}}{\widehat{\operatorname{MMD}}^2_U},$$

where $\widehat{\operatorname{MMD}}^2_U$ is the unbiased estimate on $B_1 \cup B_2$. Given A, the numerator and the denominator are unbiased, so $r(m)$ can fall below zero when the true residual is below the noise. The noise level is the 90th percentile of the captured share, in absolute value, over 999 permutations of the group labels within B_1 and B_2. The witness is the difference between the groups estimated on A, used as a single criterion and evaluated on $B_1 \cup B_2$.

Results. The top m eigendirections leave 10% of MMD^2 after 12 [12, 14] criteria on Adult, 2 [2, 3] on COMPAS and 30 [30, 31] on ACS PUMS. They leave 1% after 29 [28, 32] criteria on Adult; on COMPAS and ACS PUMS, 1% is not resolved above the noise level. The worst case first falls below 1 at $m = 11$ on Adult, 8 on COMPAS and 39 on ACS PUMS. The witness, a single criterion built for the pair, leaves .0056 [.0039, .0074] of MMD^2 on Adult, .017 ($\pm$.037) on COMPAS and .042 [.024, .058] on ACS PUMS, which is .83 to .93 of its expected estimation noise at every setting. A checklist that does not see the pair thus needs 2 to 30 criteria to leave 10%, while one criterion built for the

pair leaves only noise. With numeric features on ACS PUMS, as in E2, the top m eigendirections leave 10% after 6 [6, 6] criteria and 1% after 17 [14, 20].

Input features as criteria. Figure 2 uses criteria a practitioner might check: equal averages of the input features, added one at a time. These criteria are not directions in the Gaussian RKHS, so we measure what they leave without projecting them onto it. Let $h = dP_a/dQ - dP_b/dQ$ as in Section B.3, and let h' be the result of the smallest χ^2 change to h that equalizes the first m feature averages. We report the share of $\mathrm{MMD}^2(P_a, P_b)$ left by the signed pair $(1 \pm h'/2)\, Q$. This pair is not two distributions: some rows carry negative weight, and each side holds less than full mass. With all features on the list, the share is .0059 on Adult, .023 on COMPAS and .030 on ACS PUMS with numeric features, where the pair keeps .44, .90 and .90 of its mass. Divided by the squared mass, every full list leaves about 3 to 7% of MMD^2, or 17 to 26% of the MMD; the upper end, one-hot ACS PUMS, is biased upward by its fit of 550 columns. Only ACS PUMS with numeric features is clearly above its noise level: Adult is at its level, and COMPAS and one-hot ACS PUMS are below theirs. The order of the features matters. On Adult, the list leaves 10% after 34 features in column order, but after 4 when the feature with the largest difference comes first, an order that already sees the pair.

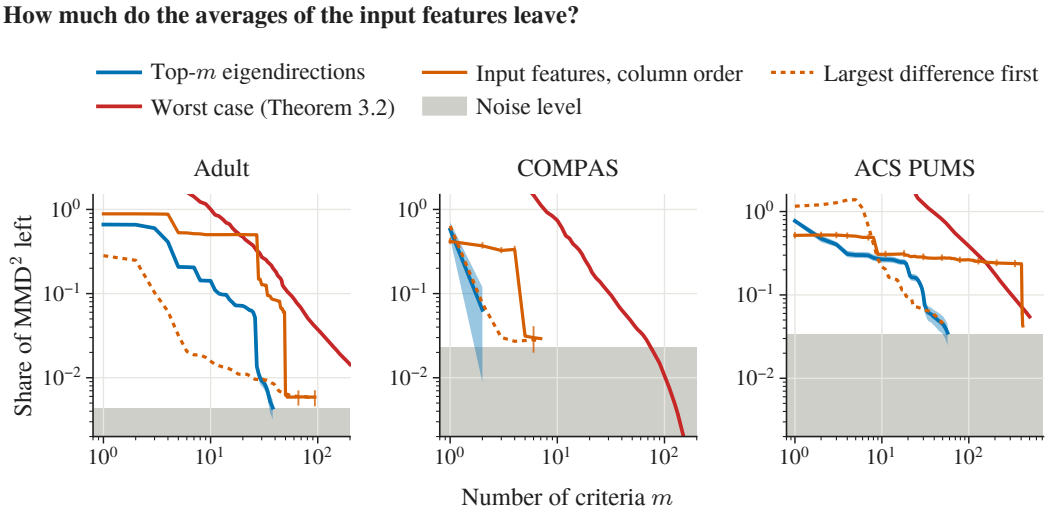


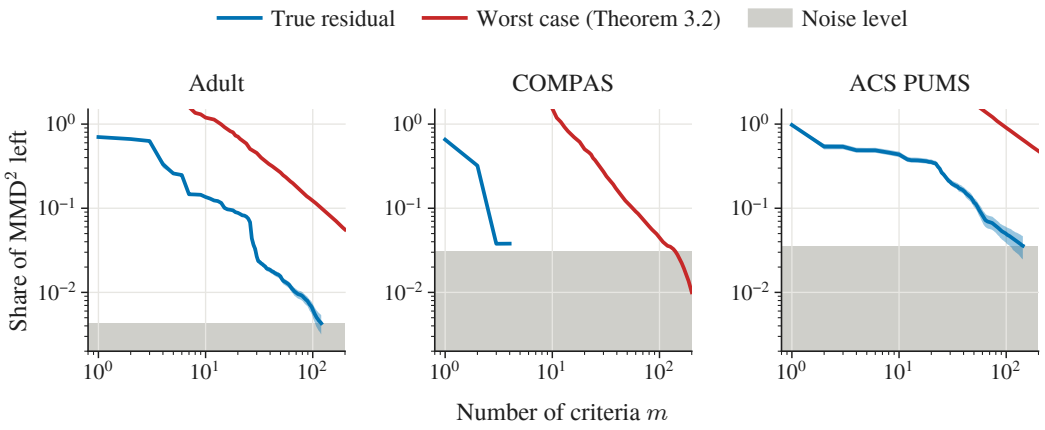
Figure 2: **Input features as criteria.** Orange shows the share of MMD^2 left by the signed pair $(1 \pm h'/2)\, Q$ after the smallest χ^2 change that equalizes the first m feature averages; that pair is not two distributions and can hold well under full mass. Solid uses the data's column order, which is arbitrary; dotted takes the largest difference first, an order that sees this pair. Blue repeats Figure 1A, the star is the witness, and the dashed line is the worst case of Theorem 3.2. Curves stop at the larger of the noise level and the witness's residual. Vertical marks are 95% intervals at grid points. Values can exceed 1: removing a subgroup can unmask a difference it was offsetting. The steep end on ACS PUMS comes from the order: the first place-of-birth columns arrive after 390 occupation columns.

E.4 E2: Fairness and Predictive Signal

Protocol. Every method fits on the same labeled rows, training and validation, and test rows are used only to evaluate. The base model is chosen by validation log-loss, unanimously over seeds: gradient-boosted trees [Chen and Guestrin, 2016] on Adult and ACS PUMS, and a neural network on COMPAS. The features never contain the group; the rules use it. The program of Section 6 is fit on 5-fold out-of-fold scores. On test rows, a rule applies to each fold model's score and averages the decisions. Below $\|\omega\|_{\mathrm{sep}}$ the rule is the blend of Proposition 6.2; above it, the program maximizes the signal of one group, then that of the other. The noise band is that of an exactly efficient rule refit on each seed, widened by $\sqrt{1 + n_T/n_F}$ for n_T test and n_F fit rows.

Results. The best signal any threshold rule on the score achieves in the two groups is .633 [.631, .635] and .730 [.727, .733] on Adult, .369 [.362, .374] and .306 [.298, .314] on COMPAS, and .644

Half the median bandwidth



Twice the median bandwidth

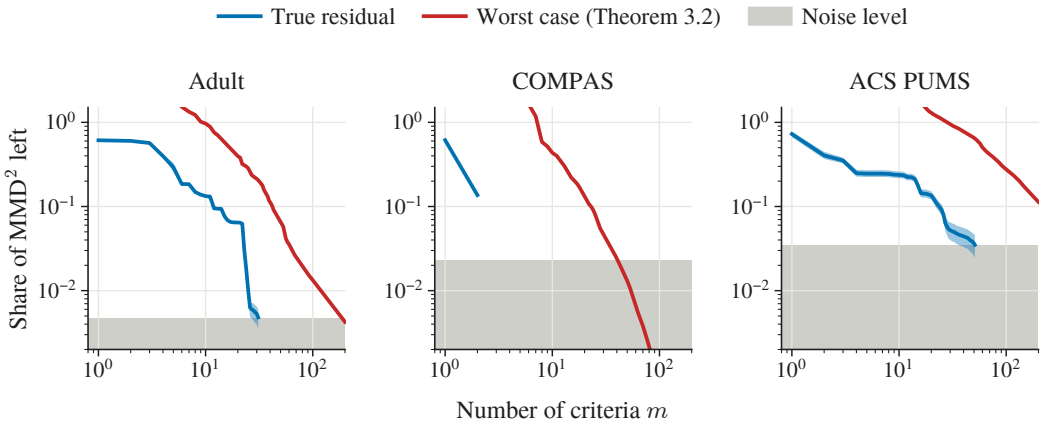


Figure 3: **Other bandwidths.** Figure 1A at $0.5\times$ and $2\times$ the median bandwidth. COMPAS has no band: whole-dataset resamples exist only at $1\times$, and bands over seeds are too narrow there.

[.643, .645] and .636 [.636, .637] on ACS PUMS. On all 30 pairs of dataset and seed, $\|\omega\|_{\text{sep}}$ equals the smaller of the two. The ratio of the smaller to the larger, .868 [.863, .873], .829 [.803, .856] and .989 [.987, .991], is below that of a common ROC curve for both groups ($p = .001$, .007 and .027). It measures how differently predictable the groups are, not the cost of fairness. At $\|\omega\|_{\text{sep}}$, held-out unfairness is 1.09, 1.87 and 1.005 times $|\Delta p|$ times the rule's held-out signal for the mixture, and 1.17, 2.60 and 1.10 times for the mean over seeds. Against the noise band, the one-sided p-values are .11, .11 and .90 for the mixture, and .045, .045 and .72 for the mean over seeds. On COMPAS the test has little power: it would pass a rule at about 2.4 times the efficient unfairness, against about 1.15 times on Adult and 1.2 times on ACS PUMS, so the data cannot say whether COMPAS's rule is efficient.

Methods. Post-processing [Hardt et al., 2016] with a balanced objective solves the same program at $\|\omega\|_{\text{sep}}$. Exponentiated gradient [Agarwal et al., 2018] runs with demographic parity and with equalized odds over a grid of tolerances, with the group among its features. With a balanced objective and tolerances up to .02, it reaches the program's signal at .95 to 1.08 times the program's unfairness on Adult, .97 to 1.45 times on COMPAS and 1.04 to 1.89 times on ACS PUMS; the intervals on COMPAS are wide, from [.49, 1.35] to [.76, 1.71]. Looser tolerances, where active, reach 2.3, 8.3 and 1.8 times. With its default objective, accuracy, it ties a program that minimizes error under equalized odds, and under demographic parity on Adult its error is .014 to .019 higher.

The encoders are adversarial debiasing [Edwards and Storkey, 2016], a variational encoder with an MMD penalty [Louizos et al., 2016], and LFR [Zemel et al., 2013], each trained on training rows over a grid of fairness weights. LFR is fit with exact gradients of the objective of the `aif360` implementation [Bellamy et al., 2019], which appends the group to the features. Run on a score fit to each encoder's features, the program keeps, at each method's best setting, .952, .931 and .978 (adversarial, variational, LFR) of the signal it keeps on the input features on Adult, 1.04, 1.00 and 1.06 on COMPAS, and 1.004, .995 and .977 on ACS PUMS. The COMPAS values are within noise of 1 and carry the upward bias of a maximum over groups. The program uses the group, so these ratios show that the features keep the inputs' signal, not that they remove the group; they are not evidence about Theorem 5.1.

E.5 Choices Made After Seeing Results

- The 10-seed mixture, and its noise band for a rule refit on each seed, were adopted after seeing the figures. This moved two marginal results to not rejected: the test at $\|\omega\|_{\mathrm{sep}}$ gives $p = .045$ on Adult and COMPAS for the mean over seeds, and $p = .11$ for the mixture. The two statistics test different things.
- The program is fit on training and validation rows. The plan fit it on validation rows alone, and we changed this after exponentiated gradient exceeded the validation-only frontier on COMPAS. That frontier, fit on 791 rows, lies outside the noise band on COMPAS by about what its sample size predicts.
- The balanced objective for post-processing and exponentiated gradient, and the program that minimizes error.
- The rule for including encoders, the exact-gradient fit of LFR, and showing only tolerances up to .02 in the figure.
- In E1: the ridge-free residual for input features, after the ridge image failed its control (Section E.7); one-hot ACS PUMS as the main setup; the witness; the whole-dataset resamples on COMPAS; the order with the largest difference first; the floor of the input-feature curve, the larger of the noise level and the witness's residual; reading "far more" as at least twice as much; and moving the input-feature curve to the appendix.

E.6 Predictions That Failed

- *A list of the input features' averages leaves far more than the top m eigendirections at every m.* The planned test could not fail, because the ridge image passes on complete lists. On the ridge-free measure, adopted afterwards, the list leaves less than the top eigendirection at $m = 1$ on COMPAS and ACS PUMS at every bandwidth, by .14 to .42 of MMD^2. On Adult, the list leaves at least twice as much at 35 of 38 resolved values of m.
- *The frontier is a ceiling for held-out methods.* Exponentiated gradient exceeds it on ACS PUMS, with intervals below zero at all four tight tolerances, and on Adult by point estimate. Only the line of Theorem 6.1 is a bound.
- *Standard methods need more unfairness at tight tolerances.* This was mostly a mismatch of objectives.
- *The share of the largest signal kept by $\|\omega\|_{\mathrm{sep}}$ measures the cost of separation.* It measures how differently predictable the groups are. On COMPAS, an earlier run with five seeds gave .87; with ten seeds it is .83 [.80, .86].

E.7 Experiments from the Review

During review we ran four further experiments, each with five seeds.

Finite-dimensional checklists. A linear kernel has a d-dimensional feature map, so d criteria exhaust it (Section B.1). On subsamples of about $5{,}000$ rows, the residual after the d linear criteria was zero to numerical precision on every dataset and seed, while d eigendirections of a Gaussian kernel, chosen greedily to capture as much of the difference as possible, still left .001 to .112 of MMD^2. After the groups were matched so that all d feature averages agreed exactly, a Gaussian MMD test still separated them on all three datasets ($p = .001$ with 999 permutations), with an MMD^2

23 to 88 times the bias that the matching introduces. Satisfying every linear criterion certified that d averages agree, not that the distributions do.

Input features against the best criteria. The comparison of input features with the top eigendirections that we reported during review (.230, .495 and .606 of MMD^2 captured, the last from a list that contained the race column) used the kernel ridge image of the features, which reads .71 to .85 on COMPAS even for a complete list and so cannot show what a checklist leaves; the ridge-free measure of Section E.3 replaces it.

Stability of the witness. The estimated witness moves by at most $2\|\hat{\delta} - \delta\|_{\mathcal{H}}/\|\delta\|_{\mathcal{H}}$, so it is stable when the MMD is large compared with the sampling error. We refit it on resamples of growing size, using the continuous features of each dataset. At the largest sizes ($16{,}000$, $4{,}000$ and $8{,}000$ rows), its cosine with a resampled witness was at least .998, .997 and .985 on Adult, COMPAS and ACS PUMS, against about .7 with randomly permuted group labels. The loss $1 - \cos$ fell roughly as $1/n$ on Adult and COMPAS (log–log slopes within .16 of -1), and more slowly on ACS PUMS, whose groups differ least (slopes $-.66$ to $-.81$); witnesses fit on disjoint samples gave more variable slopes. The witness scores correlate most with hours worked on Adult ($.83$), age on COMPAS ($-.74$) and age on ACS PUMS ($.89$); these are descriptive associations.

Tail bound. Our comparison of the tail bound of Section C.3 with the realized $\Pr[|S - Y| > t]$, on synthetic data and on scores for the three datasets, checked the computation rather than the theorem, since with tolerances estimated on the same rows the bound holds by algebra and Markov's inequality.

E.8 Code and Reproducibility

The pipeline is implemented in Python 3.12 with dependencies pinned by `uv`. It ran on one workstation with four RTX 4090 GPUs and 64 CPU cores, using 64-bit eigendecompositions. Code, protocols and figure scripts will be released under the MIT license on publication.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the contribution list in Section 1 name exactly the three theoretical results proved in the paper: (i) the recovery of the classical impossibility results, including the Kleinberg–Mullainathan–Raghavan theorem, from the same geometry (Theorem 4.1, Section 4); (ii) the qualitative and quantitative Pokémon theorems (Theorem 3.1, Theorem 3.2, Section 3); and (iii) the fair-feature-learning impossibility with its approximate version and binary-classifier corollary (Theorems 5.1 and 6.1 and Corollary D.1, Sections 5 and 6). Each claim is stated in its natural regime (characteristic kernel, binary G, $\Delta p \neq 0$) and we do not claim anything that the theorems do not deliver.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 "Scope and future extensions" enumerates the main restrictions: binary protected attribute $G \in \{a, b\}$, binary or real-valued Y, scalar score S, characteristic-kernel regime, and the base-rate-gap hypothesis $\Delta p \neq 0$. Each restriction is flagged as conventional rather than essential, and the dimensional-transition caveat (Pokémon theorem inverting at finite $\dim \mathcal{H}$), the Indep+Sufficiency resistance to mean-level relaxation, and the universal-kernel sharpening of Theorem 5.1 are named as open items. Individual, causal, and counterfactual fairness are explicitly out of scope.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theorem, corollary, and lemma is numbered and cross-referenced, with its assumptions stated in the theorem body (and, for the quantitative Pokémon theorem, in the labelled assumptions **(A1)** source condition and **(A2)** polynomial eigendecay). Proof sketches are in the main text immediately after the corresponding statements; full proofs are in the appendix: Section B.2 (quantitative Pokémon, Kolmogorov m-width), Section D.1 (fair-feature-learning impossibility, with the characteristic-kernel distributional lift spelled out), Section D.2 (the $(\varepsilon + \rho)/|\Delta p|$ bound), Section D.3 (binary-classifier Pareto bound including the corner-case analysis), and Section C (recovering the classical impossibilities).

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 7 gives a compact protocol and Section E gives the full specification. Datasets (Adult Income, COMPAS, ACS PUMS) are named with their sources and SHA-256-verified download paths; the 70/15/15 train/val/test split is stratified on (Y, G) and is deterministic given the seed. Kernels (Gaussian RBF with median heuristic on a 5,000-point subsample), the $n = 10{,}000$ kernel-matrix subsampling cap, the top-200 eigendecomposition, the five fixed seeds $\{42, 123, 456, 789, 1011\}$, the $B = 1000$ bootstrap resamples, the 999-permutation tests, and the BH-FDR correction at $q = 0.05$ are all documented. All hyperparameter sweeps (ExpGrad ϵ grid, Fair-VAE λ_{MMD} grid, adversarial λ_{adv} grid, LFR A_z grid), optimiser (Adam, learning rate 10^{-3}), training schedule (100 epochs, 10-epoch warmup), gates (mode-collapse filter, ρ-based near-separation gate), and aggregation choices are stated in Sections E.3 to E.4.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All three datasets are publicly available and are fetched programmatically from their canonical sources (UCI, ProPublica, the `folktables` package [Ding et al., 2021]) with SHA-256 verification, as documented in Section E.1. The full experimental pipeline is implemented in Python 3.12 under a `uv`-pinned lockfile (Section E.8); code, preprocessing scripts, and figure-generation scripts will be released publicly on GitHub under an MIT licence upon acceptance.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 7 summarises the protocol inline with the three illustrations. Section E spells out the details: dataset preprocessing (standardisation of continuous features, one-hot encoding of categorical features, filters, subsample sizes), train/val/test split fractions and stratification, empirical base rates, the median-heuristic kernel-bandwidth computation and its robustness to $0.5\times, 1\times, 2\times$ multipliers, bootstrap and permutation protocols, and per-method hyperparameters (sweep grids, optimiser, learning rates, epochs, architecture widths for Fair-VAE and adversarial debiasing). Hyperparameters are swept rather than tuned, so no validation-based tuning enters the main figures.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The top row of Figure 1 shows mean $\pm$ one standard deviation over five random seeds for the m-width residual decay. The middle row carries percentile bootstrap 95% confidence intervals ($B = 1000$) on the test error for each near-separation classifier, and the 1 of 46 CI violation on ACS PUMS is reported explicitly. The bottom row plots each encoder's point estimate with the fixed ρ-gate counted in the legend. MMD and HSIC comparisons in the appendix use 999-permutation tests with Benjamini–Hochberg FDR correction [Benjamini and Hochberg, 1995] at $q = 0.05$; methods and parameters are detailed in Section E.1.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section E.8 describes the compute: all experiments ran on a workstation with four RTX 4090 GPUs, with kernel eigendecompositions in 64-bit precision.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We conduct no human-subjects research and collect no new data; we use public benchmarks (Adult, COMPAS, ACS PUMS) under their licenses.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 8 discusses the impact: the results show that no checklist certifies fairness, so designers must choose which criteria matter for the people affected. A possible negative reading is that fairness is hopeless; Section 6 counters it by showing that approximate fairness is attainable and how to compute the best trade-off.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no pretrained models, no scraped data, and no content-generation system. The released artefacts are reproduction code and preprocessing scripts for the three public benchmarks; these carry no dual-use or misuse risk.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Every dataset, algorithm, and software library used is cited at its point of first appearance: Adult Income is fetched from the UCI ML Repository; COMPAS from the ProPublica public release (`propublica/compas-analysis`); ACS PUMS via the `folktables` package [Ding et al., 2021]. Method citations include Kleinberg et al. [2017], Chouldechova [2017], Pleiss et al. [2017], Hardt et al. [2016], Dwork et al. [2012], Kamiran and Calders [2012], Agarwal et al. [2018], Platt [2000], Zemel et al. [2013], Louizos et al. [2016], Edwards and Storkey [2016]. Software citations include Pedregosa et al. [2011] (BSD-3), Chen and Guestrin [2016] (Apache-2.0), Paszke et al. [2019] (BSD-style), Weerts et al. [2023] (MIT), and Bellamy et al. [2019] (Apache-2.0). All assets are used under their declared licences.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is the reproduction code, including dataset loaders, kernel utilities, and figure-generation scripts. The code is documented inline; Section E gives all settings needed to reproduce the main experiments. Code will be released publicly on GitHub upon acceptance. No new datasets or pretrained models are introduced.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no research with human subjects. It analyses pre-existing, publicly released tabular benchmarks.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human-subjects research. No IRB approval is required; no individual participants are studied.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as a component of the core methodology. All theoretical results, proofs, and experimental protocols were designed independently of LLM assistance.