On the Epistemic Limits of Personalized Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predictive models often include group attributes that encode personal characteristics like sex, blood type, or HIV status. Personalized models must ensure *fair use* – i.e., groups who provide personal data should expect to receive a tailored improvement in accuracy compared to a non-personalized model. In this paper, we derive conditions under which one can detect fair use violations in predictive models and characterize when estimating fair use is impossible. We propose a metric to evaluate the worst-case accuracy gain across groups called the *benefit of personalization* (BoP). We obtain bounds on the error probability for testing if the BoP is above a target threshold given finite samples. Remarkably, our bounds provide an information-theoretic limit on the number of group attributes that a model can use to allow for verification – beyond this limit, it is impossible to reliably detect if personalization harms or benefits all groups. We also derive statistical limits for the minimax mean-square error of estimating the BoP. Our results show that there is no way to reliably determine if a personalized model with $k \geq 19$ attributes benefits every group that provides personal data, even when we are given a dataset with $N = 8 \times 10^9$ samples (i.e., one observation for each person in the world).

1 Introduction

Machine learning models are increasingly personalized[1]. In applications such as healthcare, models often use *group attributes* to encode personal characteristics (e.g., sex, age group), sensitive information (HIV status), or information that must be acquired at prediction time (blood type, malignant biopsy, see e.g., Taylor et al., 2017; Bertsimas et al., 2019). The collection and use of group-denoting features in prediction models is adopted to improve performance (Fan and Poole, 2006; McAuley, 2022). Categorical group features divide the population into subgroups distinguished by the combination of feature values (e.g., women with ages above 70 years). At first, one might assume that a personalized model that uses group data produces predictions that are *at least as accurate* compared to a non-personalized model for all subgroups. Otherwise, the personalization is unfair: a subgroup is harmed when group data is included in the prediction model. Thus, an essential goal for fairness in personalized models is to ensure **fair use** — i.e., to ensure that models using group attributes produce a tailored improvement in accuracy for all subgroups (Ustun et al., 2019; Suriyakumar et al., 2022).

Standard practices in training machine learning models do not ensure fair use in personalization. Personalized models are trained to minimize empirical risk on testing data, which does not necessarily

[1]Personalization here refers to approaches that target *subgroups* of the population rather than *individuals* (i.e., “categorization" rather than “individualization" as per the taxonomy of Fan and Poole, 2006).

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

lead to better prediction for all subgroups. In fact, health technology companies have faced backlash over mounting concerns about the fairness of their algorithms on different racial groups (Vartan, 2019). However, testing personalized models for fair use on all (potentially exponentially) many subgroups induced by different values of group attributes is challenging (Kearns et al., 2018; Hebert-Johnson et al., 2018). It may be the case that, for specific subgroups, a group-blind model that does not use group attributes may perform better than a personalized one (Suriyakumar et al., 2022).

We propose a metric called the *benefit of personalization* (BoP) that measures the minimum gain in accuracy across all subgroups by using a personalized model instead of a group-blind model. BoP must be positive whenever a personalized model is used—otherwise, at least one group is harmed by choosing a personalized model. Estimating the BoP empowers individual stakeholders to ensure their personal group information is used fairly and that they receive a more accurate prediction in return for their data. Moreover, reporting the BoP enables practitioners to justify the collection of group information for personalized models.

The goal of this paper is to understand the conditions under which one can check for fair use of group data in prediction models and, more interestingly, the conditions where estimating fair use is impossible. We consider a setting where only a finite number of samples of the underlying population is available to audit a set of personalized models for a BoP. The number of subgroups grows exponentially with the number of group attributes used for personalization. As we will show, when there are many attributes and not enough samples, reliable estimation and testing for BoP become impossible. This is, of course, expected: an exponential increase in group-denoting attributes inevitably requires more samples per subgroup—in fact, this is the motivation behind auditing for computationally identifiable groups (e.g., Hebert-Johnson et al., 2018). What is *surprising* in our results is that, given a finite number of samples (at most one sample for each person in the world $\approx 8 \times 10^9$), the *information-theoretic* maximum of group attributes that can be collected and still used to verify for a BoP is, in fact, small: with more than 19 binary attributes, it is (statistically) impossible to test a personalized model for a BoP without further assumptions on the underlying data distribution.

The main contributions of this work include:

1. We propose a metric, the *benefit of personalization*, which quantifies the minimum gain in accuracy by using group attributes across all subgroups for a prediction model.
2. We design a hypothesis test framework to audit if a personalized model violates fair use of group attributes. We pair our test with an information-theoretic limit on the error probability that holds for any hypothesis test that audits a model for fair use. A rule of thumb is that the number of samples per subgroup should be greater than the number of subgroups for testing error to be small.
3. We characterize statistical limits on the Mean Square Error (MSE) of any estimate of BoP. Our results show that we cannot reliably estimate the benefit of personalization across subgroups for models that use more than 25 personal attributes.
4. We conduct experiments to demonstrate the behavior of the probability of error in testing for fair use. Our experiments show that the MSE of a BoP estimation increases exponentially fast when the number of samples per group drops. Finally, we evaluate our theoretical minimax bounds on the MSE of the BoP estimation on real-world datasets.

1.1 Related Work

Algorithmic Fairness. *Benefit of personalization*, our proposed metric, embodies a fairness principle that is distinct from fairness metrics that aim to quantify disparate impact. Fairness metrics can be broadly characterized as individual fairness and group fairness. The motivating philosophy behind individual fairness is to treat similar individuals similarly (Dwork et al., 2012), while group fairness aims to ensure similar performance across different groups (Ustun et al., 2019). An influential principle that captures popular metrics like Equalized Odds and Equal Opportunity (Hardt et al., 2016) is parity, which seeks to achieve proportional allocation of resources conditioned on the group (Dwork and Ilvento, 2018). Although parity is an appropriate notion of fairness for applications in

criminal justice, hiring, and college admissions since it signals disparate treatment among groups (Zafar et al., 2017), it is less suitable for domains with different ethical principles. For example in medical applications, the relevant principles are non-maleficence (do no harm) and beneficence (do good) for all groups rather than achieving parity on a statistical metric (Beauchamp et al., 2001). In such settings, accuracy is paramount. Our proposed metric (*benefit of personalization*) fills the gap. Ensuring a positive BoP is essential for non-maleficence in model personalization.

Moreover, BoP is complementary to existing group-fairness metrics. BoP may mask imbalanced performance across groups, while group-fairness metrics do not capture the gain (or loss) in using group attributes. Hence, there are situations where both metrics are called for.

Auditing for Fairness. Auditing whether an ML model is fair is a challenging problem with important consequences. In criminal justice, for example, Neil and Winship (2019) argue that testing for racial discrimination in policing can yield diverging results using different testing methods. Berk et al. (2021) show that it may be impossible to simultaneously satisfy different fairness constraints. To regulate algorithmic filtering on social media, Cen and Shah (2021) develop a statistical test that enforces a regulation. In greater generality, many recent works explore causal approaches that test for unfairness in models (Kilbertus et al., 2017; Nabi and Shpitser, 2018). Si et al. (2021) propose a testing algorithm via Optimal Transport Projections, only considering one protected attribute at a time without intersectional subgroups. In this work, we prove an information-theoretic impossibility result on testing that works for all tests.

Multicalibration & Fairness Gerrymandering. Testing for fairness is complicated in settings with a richer class of subgroups. Specifically, the framework of our work is similar to the Multi-X approach considered by Hebert-Johnson et al. (2018) and Kearns et al. (2018). Fairness constraints, such as equal opportunity and statistical parity, are imposed on all computationally identifiable – potentially exponentially-many – subgroups within a population. Under this framework, Kearns et al. (2018) show that auditing for fairness is computationally-intractable for worst-case distributions. Our impossibility result in Corollary 1 answers a different question. In particular, we characterize when the probability of error of an arbitrary auditing algorithm is too large to flag a concrete outcome. We derive a fundamental limit on the number of attributes a model can have to reliably test for fairness with any auditing algorithm.

Our findings differ from those of Kearns et al. (2018) resulting from a key difference in the assumption of sample size. Specifically, in (Kearns et al., 2018), the authors assume that it is possible to have more data points than the VC dimension of the model hypothesis class. Note that for models like deep neural networks, the VC dimension approaches infinity (Sontag, 1998). Our work builds on the key observation that with a finite population on earth, the number of available independent samples is finite. Thus, the available data points for each subgroup are limited.

Personalization, Privacy & Data Minimization. Data is collected and used for building personalized algorithms. When training personalized algorithms using personal data, prior work in both Computer Science and Law advocate for including individual rights in data privacy laws (Kaminski, 2022) (Isaak and Hanna, 2018). Ustun et al. (2019) argue that given the rights, a rational individual will only provide information on personal data if the decoupled classifiers lead to higher performance. This concept is called beneficence and is formalized in fair use.

2 Framework

In this section, we introduce notations and definitions used in this paper. In particular, we define a new metric to evaluate fair use called the *benefit of personalization* (BOP).

2.1 Preliminaries

Consider a binary classification task where categorical attributes represent group membership. We aim to predict the label Y by training a classifier on the dataset $\mathcal{D} \triangleq \{(X_i, S_i, Y_i)\}_{i=1}^N$, where $X_i \in \mathcal{X}$ is a vector of non-group features, and $S_i \in \mathcal{S} \triangleq \{0,1\}^k$ is a vector of k group attributes (e.g. sex, blood_type, HIV_status), and $Y_i \in \{0,1\}$ is a binary label. The elements of $\mathcal{S}$ are binary k-dimensional vectors, which means that there are $d \triangleq 2^k$ groups that provide personal data. Hence, we denote $\mathcal{S}$ as a one-hot encoding of the binary group attributes, with $d = |S| = 2^k$. To simplify the analysis, we consider a scenario where each subgroup has the same number of data points. Let N be the total number of samples, then the number of samples for each group is $m \triangleq \lfloor N/d \rfloor$ data points. We argue that this is the best-case scenario when prior knowledge of the dataset distribution is not known: intuitively, the group with fewer samples will dominate testing or estimation performance (see supplementary material (SM) for a more detailed discussion). We frame our impossibility results for this best-case setting; an imbalance in samples will only render more difficult auditing for the benefit of personalization.

Let the distribution of the dataset be $P_{\mathcal{D}}$. At the group level, suppose that the distributions for each group are independent, and we draw m i.i.d. data points for each subgroup from the corresponding distribution P_j:

$$(X_i, Y_i, S_i = j) \sim P_j \triangleq P_{X,Y|S=j}, \forall j \in [d]. \tag{1}$$

In this paper, we compare two classifiers, a *group-blind classifier* and a *personalized classifier*. Specifically, the group-blind classifier h_0 only takes non-group attributes as input, while the personalized classifier h_g takes both non-group and group attributes, i.e., $h_0 : \mathcal{X} \to \{0,1\}$ and $h_g : \mathcal{X} \times \mathcal{S} \to \{0,1\}$. h_0 is taken to be a baseline model trained in good faith, e.g. by empirical risk minimization on a dataset with group attributes removed.

A core principle that motivates our work is the Fair Use of group attributes (Ustun et al., 2019), we make this definition mathematically precise next.

Definition 1 (Fair Use)**.** Let h_0 be a group-blind classifier, h_g a personalized classifier, and P_j as in (1). We say that h_g follows the fair use principle if

$$P_j\left(Y = h_0(X)\right) \leq P_j(Y = h_g(X, S)), \;\; \forall j \in \mathcal{S}, (X, Y) \in P_{\mathcal{D}} \tag{2}$$

2.2 The Benefit of Personalization

A personalized classifier outperforms a group-blind classifier when the metric is the overall accuracy of a model, i.e., average accuracy across all data points. In contrast, a group-blind classifier can have higher accuracy for specific subgroups in comparison with a personalized classifier (Suriyakumar et al., 2022).

We next introduce a metric called the *benefit of personalization*. This metric captures the minimum gain in accuracy across all intersectional groups by using a personalized classifier and can be used to identify fair use violations.

Definition 2. (Benefit of Personalization). Given a fixed data distribution $P_{S,X,Y}$. The *benefit of personalization* is given by:

$$\gamma(h_0, h_g, P_{\mathcal{D}}) \triangleq \min_{j \in [d]} P_j(Y = h_g(X, S)) - P_j\left(Y = h_0(X)\right). \tag{3}$$

In addition, we call $P_j(Y = h_g(X, S)) - P_j\left(Y = h_0(X)\right)$ the per-group benefit. The benefit of personalization is a function of both classifiers and the data distribution. When the parameters are clear from the context, we denote $\gamma(h_0, h_g, P_{\mathcal{D}}) = \gamma$.

In this paper we choose accuracy as a benchmark metric for the BoP. Our results still hold if accuracy is replaced with False Positive Rates (FPR) or False Negative Rates (FNR) (see supplementary material for a more detailed discussion).

By the definition in (2), the BoP is the gain in accuracy by using a personalized classifier. For example, $\gamma = 0.4$ means that, for every subgroup, a personalized classifier is at least 40% more accurate than the group-blind classifier.

3 Testing for Fair Use

In this section, we formulate fair use as a hypothesis test on the benefit of personalization. We provide an information-theoretic bound on the reliability of testing in terms of a trade-off between the number of group attributes and the probability of error of the test for a fixed sample size. Our bound leads to the maximum number of group features that ensures reliable testing for fair use of group attributes and an interesting phase transition phenomenon on the probability of error in the hypothesis test. All proofs are provided in the supplementary material.

3.1 Hypothesis Test Famework

The disclosure of group information to a model should incur a gain in accuracy. It is critical to reliably test if, on average, a personalized model yields a gain in accuracy across all subgroups. The following hypothesis test framework verifies if the benefit of personalization given in Def. 2 is above a certain threshold ϵ.

- H_0: $\gamma < 0 \Leftrightarrow \min_{j\in[d]} P_j(Y = h_g(X,S)) - P_j\left(Y = h_0(X)\right) < 0$
- H_1: $\gamma \geq \epsilon \Leftrightarrow \min_{j\in[d]} P_j(Y = h_g(X,S)) - P_j\left(Y = h_0(X)\right) \geq \epsilon$

In practice, a personalized model should be deployed if there is evidence from data to reject the null hypothesis (H_0). Otherwise, a personalized model *should not* be deployed, since there is not enough evidence to conclude that all groups receive an average gain in exchange for revealing their personal group attributes. Note that, when $\epsilon = 0$, the above hypothesis test is equivalent to verifying fair use of group attributes according to Defn. 1.

3.2 Limits of Testing

In real-world applications, the reliability of testing for fair use is governed by sample size: reliable testing requires a minimal number of samples per group. In the extreme case where each group has only one data point, for example, it is impossible to reliably test for fair use of group attributes.

Suppose that a practitioner has access to a dataset $\mathcal{D}$ with only $N \in \mathbb{N}$ samples, and uses it to train a classifier with k group attributes. In this case, the number of data points per group is given by $m = N/2^k$. In Theorem 1, we characterize how the exponential reduction of data at the group level affects the precision of the hypothesis test in the minimax sense.

To characterize the trade-off between error probability, define V_h as the indicator function that h does not satisfy the null hypothesis, i.e., for each classifier h, $V_h = 0$ if and only if $h \in H_0$.

Theorem 1 (Error in Hypothesis Test)**.** *Let $\epsilon \in [-1/2, 1/2]$, h_0 be a group-blind classifier, and $P_{\mathcal{D}}$ be the distribution of the dataset $\mathcal{D}$. Also, suppose that the dataset is described in Sec. 2.1. Consider the hypothesis regions $H_0 = \{h \mid \mathbb{E}_{\mathcal{D}}\left[\gamma(h,h_0)\right] > \epsilon\}$, and $H_1 = \{h \mid \mathbb{E}_{\mathcal{D}}\left[\gamma(h,h_0)\right] \leq 0\}$. Let $\Psi : \mathcal{H} \to \{0,1\}$ be a decision function, and define the* error probability *in the hypothesis test to be P_e given by:*

$$P_e = \Pr\left(\Psi(h) = 1 \mid V_h = 0\right) + \Pr\left(\Psi(h) = 0 \mid V_h = 1\right) \tag{4}$$

Then,

$$\min_{\Psi} \max_{\mathcal{D}} P_e \geq 1 - \frac{1}{\sqrt{d}}\left(1 + 4\epsilon^2\right)^{m/2} \tag{5}$$

Given that practitioners do not have access to the dataset distribution, Theorem 1 presents a minimax lower bound on the probability of error in a hypothesis test. The minimax lower bound implies that for every decision function Ψ, there exists a dataset distribution such that the probability of error for the hypothesis test is greater than the lower bound in (5).

To prove this Theorem, we lower bound the probability of error of any decision function by the total variation distance of two distributions (belonging to different hypotheses) that are "difficult" to distinguish. We use a generalized version of Le Cam's Lemma stated in Lemma 1 that might be of independent interest. The original Le Cam Lemma (Le Cam, 1973) states that (6) holds when θ_1, θ_2 are elements of Θ. We generalize this lemma for the case where θ_1, θ_2 are elements of Θ chosen at random.

Lemma 1 (Generalized Le Cam)**.** *Let $\{\Lambda, d\}$ be a metric space, and $\mathcal{D}$ be a dataset. Let $\mathcal{P}$ be a family of probability distributions with parameter $\theta \in \Theta$. Suppose that f is a statistics with image in Λ ($f(\theta)$ for $\theta \in \Theta$ means the statistics calculated for P_θ), and that there exists Λ_1, Λ_2 subsets of Λ that are 2δ-separated. Also, suppose that Θ_1 and Θ_2 are subsets of Θ such that $f(\theta) \in \Lambda_1$ for all $\theta \in \Theta_1$ and $f(\theta) \in \Lambda_2$ for all $\theta \in \Lambda_2$. Then we have that:*

$$\min_{f^* \in \sigma(\mathcal{D})} \max_{\theta \in \Theta} \mathbb{E}\left[d(f(\theta), f^*)\right] \geq 2\delta(1 - \mathsf{TV}(P_{\theta^1} || P_{\theta^2})), \tag{6}$$

where $\theta^i \in \sigma(\Lambda)$ with image in Θ_i, and P_{θ^i} indicates a probability distribution in $\mathcal{P}$ with parameter θ^i for $i \in [2]$.

Threshold Test. Given sufficient samples per group as outlined in Theorem 1, how should testing be done in practice? The most natural choice is a threshold test: Given h_g and h_0, we simply evaluate the performance of each classifier using an auditing dataset and check if there is an ϵ gain in accuracy across all the groups. Intuitively, the threshold test consists in estimating the BoP (e.g., using empirical mean in the group or bootstrapping) and then checking if it is greater than the threshold ϵ. Note that the theorem holds for *all* possible tests (not only the threshold test).

3.3 Implications of the Limits of Testing

Given that the size of the dataset is smaller or equal to 8×10^9 (number of people on Earth), theorem 1 yields a fundamental limit on the number of group attributes in order for the probability of error in the hypothesis test to be smaller than $1/2$ (a random choice). Corollary 1 formalizes this limit.

Corollary 1 (Limit for Testing)**.** *With the same setup as in Theorem 1. Take $N = 8 \times 10^9$, and let $\epsilon = 0.01$. If $\min \max P_e \leq 1/2$ then $k \leq 19$.*

Phase Transition. The lower bound in Theorem 1 gives a characterization for a phase transition in the hypothesis test error. Specifically, consider that $N = 8 \times 10^9$, $\epsilon = 0.01$, $d = 2^k$, and $m = N/d$. For $k = 18$ attributes, the minimax lower bound for the probability of error is 0.13, in contrast for $k = 19$ attributes the minimax lower bound is 0.97. Hence, there exists a notable phase transition in the probability of error in hypothesis test between $k = 18$ and $k = 19$. Figure 1 shows that, for a fixed number of groups $d = 141$ and varying the number of samples per group, there is also a phase transition for the minimax error probability in a hypothesis test. This phase transition can also be verified in the semi-synthetic data case in Figure 2 **Left**.

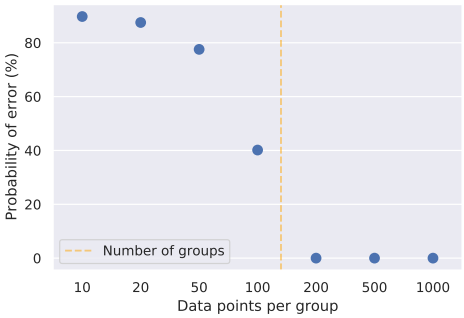

Figure 1: Maximum of 0 and lower bound in (5) for $d = 141$, $\epsilon = 0.1$ is in the y-axis, and m is in the x-axis. The orange line represents when the number of data points per group is equal to the number of groups.

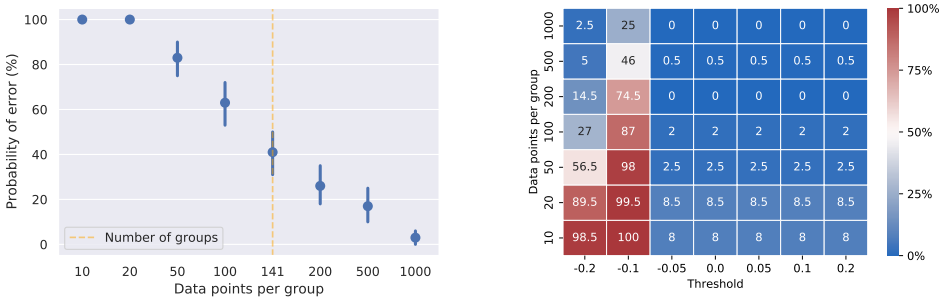

Figure 2: Semi-synthetic dataset created using Adult (Dua and Graff, 2017). The group attributes are working class, education, marital status, occupation, relationship, race, gender, and country. Only groups with more than 100 data points are considered. A dataset with m points per group is generated via sampling replacement from each group in the Adult dataset. A personalized classifier and a group-blind classifier are deployed using logistic regression to predict the income level using the semi-synthetic dataset. The benefit of personalization for these classifiers is approximately -10%. **Left:** Error probability, with 95% confidence interval bars, for a hypothesis test using the threshold test with $\epsilon = -0.15$. **Right:** Heat map of the probability of error in hypothesis test using the threshold test with ϵ given on the x-axis and with the number of data points per group given on the y-axis.

Error Probability in Semi-Synthetic Data. Figure 2 (**Left**) displays the error probability in a hypothesis test using the threshold test. The error probability decreases when the number of data points per group increases. Moreover, the figure shows that when the number of samples per group is greater than the number of groups, the error probability is smaller or equal to 50% for the first time. This is, indeed, a simple rule of thumb as a consequence of our results: the number of groups should never be greater than the number of samples per group; otherwise, testing for fair use is information-theoretically infeasible without prior assumptions on the underlying distributions. In Figure 2 (**Right**), we plot a heat map for the error probability in a hypothesis test using as threshold the values on the x-axis and as the number of data points per group the values of in the y-axis. For higher values of ϵ, there is more "tolerance" for mistakes in the test, and the phase transition happens for a smaller number of samples per group.

4 Estimating the Benefit of Personalization

When the benefit of personalization is positive but small, a user may opt-out of providing group features to a personalized model. If such features are sensitive or costly to produce, their use may only be justified when a personalized model provides a significant accuracy gain. Thus, testing for the benefit of personalization (e.g., if $\gamma > \epsilon$) is, on its own, insufficient: stakeholders must be made aware of the *estimated* increase in accuracy from sharing personal data and only provide their personal information when there is a sufficient gain in accuracy (Ustun et al., 2019).

In this section, we consider the problem of estimating the benefit of personalization. We derive information-theoretic limits for estimation and show that, for a finite dataset, there is a threshold for the number of group attributes required to reliably estimate the BoP with a small error.

4.1 Estimation Framework

We consider the problem of designing an estimator γ^* for the BoP given a dataset $\mathcal{D}$ with n samples. Our metric of choice is the mean-square error (MSE) $\mathbb{E}\left[(\gamma - \gamma^*)^2\right]$. Naturally, the value of the minimum MSE (MMSE) depends on the dataset distribution and the estimator of choice. Instead of restricting our analysis to specific estimators and datasets, we aim to understand the minimax estimation error, i.e., the best MMSE achievable across *all* dataset distribution, defined as

$$R^* \triangleq \inf_{\gamma^* \in \sigma(\mathcal{D})} \sup_{P_{X,S,Y}} \mathbb{E}\left[(\gamma - \gamma^*)^2\right]. \tag{7}$$

To motivate the definition of R^*, suppose that there exist functions $g_1(d,m), g_2(d,m) > 0$ such that $R^* \leq g_1(d,m)$, and $R^* \geq g_2(d,m)$. Recall that d is the number of subgroups and m is the number of samples per group. This implies that for every estimator γ^*, there exists a dataset distribution $P_{X,S,Y}$ such that $g_2(d,m) \leq \mathbb{E}\left[(\gamma - \gamma^*)^2\right] \leq g_1(d,m)$. In practice, this means that for *any* estimator there is a distribution such that the MSE of the resulting BoP estimate will be bounded away from zero by $g_2(d,m)$. The main result in this paper is Theorem 2, which gives explicit expressions for g_1, g_2.

Theorem 2 (Minimax bounds for the Estimation MSE)**.** *Given a dataset $\mathcal{D}$, constants $d, m \in \mathbb{N}$ such that $d \geq 4$, $d \leq 2^{m+2}$, a group-blind and a personalized classifier, the minimax estimation error for the benefit of personalization in* (7) *is such that:*

$$\frac{\log(d)}{16m} - \frac{\log(4)}{16m} \leq \inf_{\gamma^* \in \sigma(\mathcal{D})} \sup_{P_{X,S,Y}} \mathbb{E}\left[(\gamma - \gamma^*)^2\right] \leq \frac{\log(d)}{m} + \frac{\log(m)}{m} + \frac{2 + \log(2)}{m}, \quad (8)$$

Detailed proof for this Theorem is given in the supplementary material. The proof relies on two-point methods for minimax lower bounds; specifically, we use the Generalized Le Cam (Lemma 1) together with standard concentration inequalities. We discuss the implications of this Theorem next.

4.2 Implications of The Limits of Estimating the BoP

For a dataset with a fixed amount of data points (N), given two personalized models with k_1 and k_2 group attributes ($k_1 > k_2$), then in the best case scenario, the number of data points per group will be, respectively, $N/2^{k_1} < N/2^{k_2}$, hence the MSE for estimation will be higher on the "more" personalized classifier. Theorem 2 gives a characterization for the order of the estimation error for the BoP. The bounds in Theorem 2 are illustrated in Figure 3 (**Right**) in $\log$ scale. This figure suggests that even knowing that the bounds in Theorem 2 are in the minimax sense, they are close to the estimated MSE, especially the lower bound.

We apply Theorem 2 to derive a fundamental limit on the number of binary group attributes to reliably estimate the BoP γ. This lower (converse) bound in (8) captures the trade-off between the minimax MSE of an estimator, the number of sensitive attributes, and the dataset size. This trade-off is instantiated in Corollary 2, resulting in an upper bound in the number of group attributes required to achieve a small minimax MSE in estimating the BoP given finite samples.

Corollary 2 (Limits of Estimation)**.** *Given a dataset $\mathcal{D}$ with N data points, $d \geq 4$ groups, $m = N/d$ data points per group, an $0 < \delta < 1/16$, a group-blind and a personalized classifier. If $d \leq 2^{m+2}$, and*

$$\inf_{\gamma^* \in \sigma(\mathcal{D})} \sup_{P_{X,S,Y}} \mathbb{E}\left[(\gamma - \gamma^*)^2\right] \leq \delta,$$

then, the number of binary group attributes (k) satisfies

$$(k-2)2^k \leq \frac{16}{\ln 2} N\delta. \quad (9)$$

Specifically, for $N = 8 \times 10^9$ (approximately the number of humans on the Earth) and $\delta = 1/400 = (5\%)^2$, then the maximum number of binary group attributes is 25, i.e, $k \leq 25$ in order to ensure reliable estimation.

The proof of Corollary 2 is a direct application of the Theorem 2. A more detailed proof is given in the SM. By choosing $\delta = 1/400$ the square root of the MSE is at most $1/20 = 5\%$. Hence, when the estimator γ^* is unbiased, Corollary 2 gives an upper bound of 5% in the estimation variance.

Corollary 2 gives a rule for practitioners to design algorithms that reliably determine the minimum gain from personalization. During a decision-making process, such as deciding to take an exam for a medical diagnosis, the user should take advantage of an accurate estimate of the benefit of personalization and make the best-informed decision.

In general, one would expect that the maximum number of group attributes for a reliable test should exceed the maximum number of group attributes for reliable estimation. Our bounds (5), (8) show otherwise. This is an artifact of the fact that $|\gamma| \in [0,1]$. Easier estimation for statistics with norm smaller than 1 is simply an artifact of scale and follows from the Le Cam Lemma.

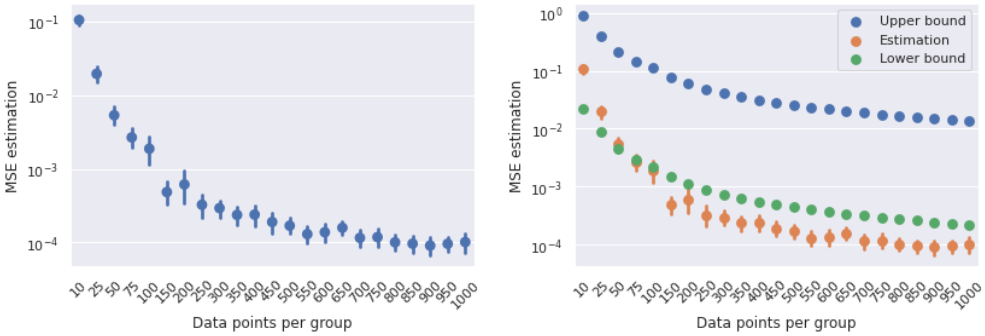

Figure 3: Plots generated using the same semi-synthetic data as in Figure 2. The estimation of the BoP was made via 100 Monte Carlo iterations, and the 95% confidence intervals were estimated via bootstrap. **Left:** MSE of the BoP estimation for different values of samples per group in log scale. **Right:** Also in log scale, MSE of the BoP estimation too, but additionally, it shows the lower and upper bound in (8).

5 Final Remarks and Limitations

Personalized ML models should provide performance benefits to all individuals who provide personal data. We call this gain the *benefit of personalization* (BoP). In this work, we have proposed tools to estimate this quantity and determine if it is large enough to warrant the collection of personal data. Our results suggest that we are limited in our ability to determine if personalization leads to equitable gains in finite-sample settings. In particular, we show that there is no way to reliably determine if a personalized model with $k \geq 19$ attributes benefits all groups that provide personal data, even when we are given a dataset composed of $N = 8 \times 10^9$ samples (i.e., one observation per person in the world).

Limitations. The minimax bounds hold for all decision functions and estimators and inherently consider the worst-case distribution for each case. However, its generality implies that some knowledge of the underlying data distribution may result in a more favorable trade-off between sample size, estimation (or testing) error, and the number of group attributes. In particular, in Figure 3, the MSE in estimating BoP falls below the minimax lower bound, suggesting that the model's underlying distributions for each subgroup is not quite the worst-case scenario. Regardless, the upper bound of the MSE estimation and the impossibility result always hold in the general case when no prior distribution knowledge is known, as usual for minimax results. A seeming constraint in our work is that we consider every group to have the same number of data points, but in Sec. 2.1 we argue that this is the best-case scenario for testing and estimation. Finally, fair use is a complement to group fairness metrics, *not* a substitute. If the base and personalized model are trained in bad-faith to underperform on a target group, then a positive BoP may mask a loss of accuracy on this group. Conversely, group fairness does not capture the gain (or loss) in using group attributes.

Recommendations. At the very least, ML practitioners should use the rules of thumb to control the number of group features to ensure reliable estimation and testing. In addition to not exceeding 19 attributes to ensure reliable testing, there should be more samples per group than the number of subgroups. Failing to do so will require additional assumptions on the data to reliably estimate the BoP or ensure fair use. When one group has very few data points and estimating the BoP is unreliable, practitioners should reduce the number of groups or use the baseline model only for some groups. Our bounds can guide on how the number of groups can be reduced to enable reliable estimation of the BoP. We encourage the report of estimated per-group gains in accuracy for all groups. This enables stakeholders to make an informed decision on whether to share their personal information, particularly if this may incur a data collection or privacy cost. Finally, if the BoP is accurately estimated and is negative, groups without a gain in personalization should revert to group-blind models. A per-group estimate of the gain in accuracy can also be used to guide feature selection (Miao and Niu, 2016).

References

Beauchamp, T. L., Childress, J. F., et al. (2001). *Principles of biomedical ethics*. Oxford University Press, USA.

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

Bertsimas, D., Dunn, J., and Mundru, N. (2019). Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183.

Cen, S. and Shah, D. (2021). Regulating algorithmic filtering on social media. *Advances in Neural Information Processing Systems*, 34.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dwork, C. and Ilvento, C. (2018). Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*.

Fan, H. and Poole, M. S. (2006). What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*.

Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59.

Kaminski, M. E. (2022). The case for data privacy rights (or'please, a little optimism'). *Notre Dame Law Review online, Forthcoming*.

Kearns, M., Neel, S., Roth, A., and Steven Wu, Z. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1.

McAuley, J. (2022). *Personalized Machine Learning*. Cambridge University Press.

Miao, J. and Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, 91:919–926.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Neil, R. and Winship, C. (2019). Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology*, 2:73–98.

Si, N., Murthy, K., Blanchet, J., and Nguyen, V. A. (2021). Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, pages 9649–9659. PMLR.

Sontag, E. (1998). Vc dimension of neural networks.

Suriyakumar, V. M., Ghassemi, M., and Ustun, B. (2022). When personalization harms: Reconsidering the use of group attributes in prediction.

Taylor, S., Jaques, N., Nosakhare, E., Sano, A., and Picard, R. (2017). Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213.

Ustun, B., Liu, Y., and Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR.

Vartan, S. (2019). Racial bias found in a major health care risk algorithm.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [No] The purpose of the work is to create positive societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Theorem 1 and 2
   (b) Did you include complete proofs of all theoretical results? [Yes] Check supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Check supplementary material. All codes are provided.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Check supplementary material. All training details are specified in the code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Check supplementary material. All plots with randomness include confidence interval bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Check supplementary material. All experiments run on Google Collab
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We are using standard and public dataset

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We are using standard and public dataset

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]