When do Minimax-fair Learning and Empirical Risk Minimization Coincide?

Harvineet Singh [*1] **Matthäus Kleindessner** [2] **Volkan Cevher** [3 2] **Rumi Chunara** [1] **Chris Russell** [2]

Abstract

Minimax-fair machine learning minimizes the error for the worst-off group. However, empirical evidence suggests that when sophisticated models are trained with standard empirical risk minimization (ERM), they often have the same performance on the worst-off group as a minimax-trained model. Our work makes this counter-intuitive observation concrete. We prove that if the hypothesis class is sufficiently expressive and the group information is recoverable from the features, ERM and minimax-fairness learning formulations indeed have the same performance on the worst-off group. We provide additional empirical evidence of how this observation holds on a wide range of datasets and hypothesis classes. Since ERM is fundamentally easier than minimax optimization, our findings have implications on the practice of fair machine learning.

1. Introduction

There have been many proposals to address systematic differences in model performance among protected groups. While the majority of proposals in algorithmic fairness aim to equalize the performance across groups (Mitchell et al., 2021), more recent work in fairness aims to improve performance on the worst-affected group(s) without needlessly decreasing performance in the other groups. Such approaches are referred to as minimax-fairness (Martinez et al., 2020).

Minimax refers to *min*imizing the *max*imum error across groups. This framing of the fairness objective avoids unintended consequences of the equal-error proposal – achieving equality in all groups may end up increasing error on the well-performing groups *without* any gains for the remaining groups (Zietlow et al., 2022). Thus, minimax-fairness can be a preferable notion to enforce when performance improvements for any group is more desirable than parity.

A wide range of algorithms have been proposed for learning minimax-fair predictors (Martinez et al., 2020; Diana et al., 2021b; Abernethy et al., 2022; Pethick et al., 2023; Shekhar et al., 2021). In principle, these algorithms support learning from arbitrarily complex hypothesis classes ranging from linear models to neural networks. However, they are typically evaluated on linear models. When using neural networks, for example, empirical evidence suggests that minimax predictors do not consistently improve upon ERM predictors in terms of the maximum group error (Gardner et al., 2022; Pfohl et al., 2022). This can be explained in retrospect that highly flexible models can fit each group's data well even under the ERM objective. In this work, we seek to formalize this observation and study the question *how do minimax-fair learning and ERM relate to each other*?

We bridge this gap in the understanding of the minimax predictors both theoretically and empirically. We show that ERM is minimax-fair when trained under a flexible hypothesis class and given access to the group information. We show that this observation holds empirically for different hypothesis classes like decision trees. The results have implications for the practice of learning minimax-fair predictors.

Indeed, simple training paradigms like ERM should be comprehensively tested before going to more involved, and often harder, optimization solutions. Minimax optimization is hard for even convex problems where standard versions of gradient descent solvers might not converge. For non-convex non-concave problems, such as minimax-fair learning with non-convex loss functions, the challenges persist (Hsieh et al., 2021). In contrast, ERM with gradient descent based solvers can avoid saddle points (Lee et al., 2019).

Assuming that the hypothesis class is sufficiently expressive, we claim for a range of minimax-related notions of fairness (see Figure 1):

1. *ERM is minimax-fair given perfect group information.* ERM has the same worst-case risk as the minimax predictor trained on the known group information.

2. *ERM is group-optimal given approximate group information.* If the predictor of group information is

[*]Work done during an internship at Amazon Web Services. [1]New York University, New York, USA [2]Amazon Web Services, Tübingen, Germany [3]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. Correspondence to: Harvineet Singh <hs3673@nyu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

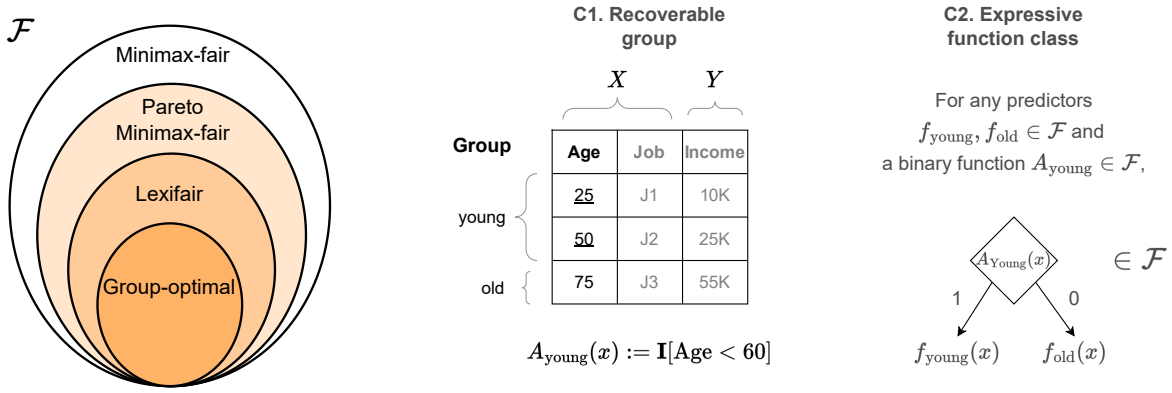

Figure 1: **Summary of results.** (**left**) Nested space of functions satisfying various minimax-fairness definitions – minimax-fair (Diana et al., 2021b), Pareto minimax-fair (Martinez et al., 2020), lexifair (Diana et al., 2021a), and functions which are optimal separately on each group's data. (**right**) Our main result is that ERM satisfies group-optimality (and hence the different minimax-fairness definitions) under *recoverability* (C1) and *expressiveness* (C2) conditions. C1 means that the young/old group membership is recoverable from the features X, here, by $A_{\text{young}}(x)$. C2 means that the function class $\mathcal{F}$ contains the combination of any two functions in a decision tree-like structure, here, $A_{\text{young}}(x)f_{\text{young}}(x) + (1 - A_{\text{young}}(x))f_{\text{old}}(x)$.

sufficiently accurate, then ERM has the same risk as the best predictor for the group on the set of correctly identified data points of the group.

3. *ERM is minimax-fair given expressive features.* If the labels are conditionally independent of the group information given the features, then ERM has the same worst-case risk as the minimax predictor trained on the known group information.

2. Related Work

Minimax-fairness. As an alternative to equalizing errors across groups, recent work proposed to minimize error on the worst-off group (Diana et al., 2021b; Martinez et al., 2020; Abernethy et al., 2022; Pethick et al., 2023; Diana et al., 2021a; Shekhar et al., 2021). The works of (Diana et al., 2021b;a) solved minimax problems by performing alternative updates for the min (learner) and max (group) players. Effectively, it repeatedly performs ERM on reweighted group errors until convergence. As such, the techniques are applicable to any hypothesis class that supports sample weights. Recent methods (Shekhar et al., 2021; Abernethy et al., 2022; Pethick et al., 2023) also support updating the models by actively observing a single or a batch of data points. A relaxation of minimax-fairness was proposed by Williamson & Menon (2019) which replaced the max group error by average over the error for the k worse-off groups. One limitation of these lines of work is that groups must be defined beforehand in order to control their worst-case error. Instead of pre-specified groups, Hashimoto et al. (2018) considered error on worst-case perturbations in the neighbourhood of the training data. Work on multi-group calibration aims to control error for any *computationally-identifiable* group which can be expressed as a function of the features (Hebert-Johnson et al., 2018; Kim et al., 2019).

Empirical evidence. Concurrent work by Gardner et al. (2022) empirically found that tree-based models trained with ERM perform the same or better than minimax-fair methods based on distributional robust optimization (e.g., Sagawa* et al., 2020) in terms of their worst-case performance. However, they did not investigate reasons for the findings theoretically nor empirically. They only used neural network models in the minimax-fair methods, and did not compare ERM and minimax-fairness under the wide range of hypothesis classes (e.g., trees), which we do in our work. Moreover, results in the original minimax-fairness work (Martinez et al., 2020) show minor difference between ERM and their specialized minimax-fair method except in datasets where groups are constructed using the labels (which does not satisfy our Assumptions C1 and 4.10). Notably, the difference is minor even for the less-expressive hypothesis class of linear classifiers. Pfohl et al. (2022) and Zong et al. (2023) made similar observations that minimax-fair methods do not outperform ERM for neural networks. Zietlow et al. (2022) found that generating synthetic data for the worst performing group could slightly improve performance. However, they also found the focus on the worst performing group was not necessary and all groups could be improved simultaneously. Lastly, Pethick et al. (2023) observed that minimax-fair training could improve over ERM in the case of adversarial training (with labels as groups) but the gains can be marginal.

Relation to domain generalization. Minimax-fairness has also been seen as a criterion to help generalize across domains or groups. The formulation of Sagawa* et al. (2020)

named Group-DRO is equivalent to minimax-fairness (Eq. (3)). Zhai et al. (2021) showed that ERM was equivalent to a variant of Group-DRO with adversarially selected groups, namely *CVaR* classification under 0-1 loss and deterministic models. In Hu et al. (2018), the same result was shown for the case of a larger set of robust optimization objectives. When performing minimax in high-dimensional data settings, highly flexible models can overfit the train set to near-zero loss and lead to models that generalize poorly (Sagawa* et al., 2020). However, overfitting might be less severe in the low-dimensional data considered in our work. We see empirically that ERM remains minimax-fair on the held-out test set.

Advantages of ERM. ERM is a well-motivated learning principle (Vapnik, 1999). In addition to being tractable and conceptually simple, ERM can be used to output models satisfying different fairness definitions. Results in Corbett-Davies et al. (2017) show that definitions such as demographic parity can be achieved starting from a Bayes-optimal classifier and suitably setting the decision thresholds. Under an expressive hypothesis class and approximate group information, ERM produces predictors that calibrated within groups (Definition 4.5) (Liu et al., 2019). In a related work, Globus-Harris et al. (2022) show that minimax-fair classifiers can be obtained using ERM by repeating the process of finding regions of high model risk and appending the region-specific optimal predictors to the overall model. Our sufficiency conditions posit that this process is possible through ERM itself. Lastly, Muandet (2022) shows that ERM is the only admissible method compatible with a set of desired properties expected from a learning method.

Our main contribution to the study of ERM's properties is that it can satisfy minimax-fairness under stated conditions. This explains the empirical observations made in prior work.

3. Preliminaries

We describe the notation, problem setup, and necessary background on ERM and minimax methods.

3.1. Notation

Given a class of functions $\mathcal{F}$ that map from features x to labels y, we want to find a function $f \in \mathcal{F}$ that minimizes some loss defined by $\ell(f(x), y) \in \mathbb{R}$. For doing this we have a training dataset of n (feature, label) tuples $D := \{(x, y)\}$. The dataset D is partitioned (or grouped) into K disjoint groups $\{D_1, \cdots, D_K\}$ such that $\cup_{g\in[K]} D_g = D$ and $D_i \cap D_j = \phi$ for all $i, j \in [K]$. Here, $[K] := \{1, \cdots, K\}$. Empirical risk for a function f on the dataset D is defined as $R(f, D) := 1/|D| \sum_{(x,y)\in D} \ell(f(x), y)$. Similarly, the empirical risk for the group g is $R(f, D_g) = 1/|D_g| \sum_{(x,y)\sim D_g} [\ell(f(x), y)]$. Note that the risk is computed on the training data in all our results.

3.2. Problem setup

ERM solutions are functions that minimize the risk on the full dataset,

$$\mathcal{F}_{\text{ERM}} := \arg\min_{f\in\mathcal{F}} R(f, D). \tag{1}$$

Instead of minimizing risk on the full dataset, we can define functions that minimize the risk separately for each group. A group-optimal predictor $f^*(x) \in \mathcal{F}_g$ with respect to dataset D and loss function ℓ is defined as,

$$R(f^*, D_g) = \arg\min_{f\in\mathcal{F}} R(f, D_g),\ \forall g \in [K]. \tag{2}$$

Minimax-fairness instead advocates to minimize the risk for the worst-off group,

$$\mathcal{F}_{\text{MM}} := \arg\min_{f\in\mathcal{F}} \max_{g\in[K]} R(f, D_g). \tag{3}$$

Our goal is to show that $\mathcal{F}_{\text{ERM}}$ achieves the same risk as the minimizers $\mathcal{F}_g, \mathcal{F}_{\text{MM}}$ on specific groups. Before that, we discuss other notions of minimax-fairness from prior work.

3.3. Minimax Pareto-fair and Lexicographic fairness

Martinez et al. (2020) defined $\mathcal{F}_{\text{MM}}$ by minimizing over the class of *Pareto-optimal* functions (that is, functions for which the risk cannot be unilaterally improved on both the groups without degrading risk for one of the groups). Alternatively, Diana et al. (2021a) defined minimax solution as the predictors which minimize worst-group's risk, the second worst-group's risk, and so on, namely lexicographic minimax-fairness. We show that ERM satisfies both the definitions of minimax-fairness under our stated conditions.

Definition 3.1 (Pareto front)**.** Let the vector $\mathbf{R}(f) := (R(f, D_1), R(f, D_2), \cdots, R(f, D_K))$ denote the group-specific risks of the predictor f. A predictor f is said to Pareto dominate another $f \prec f'$ if it has equal or better risk for all groups $\forall i \in [K], \mathbf{R}(f)_i \leq \mathbf{R}(f')_i$ and strict inequality holds for at least one group $\exists j \in [K], \mathbf{R}(f)_j < \mathbf{R}(f')_j$. Given a set of predictors $\mathcal{F}$, the Pareto front $\mathcal{P}_{\mathcal{F}}$ is defined as the subset such that $\{f \in \mathcal{F} : \nexists f' \in \mathcal{F} | f' \prec f\}$.

For the Pareto front to exist, we need the following technical condition on the function class $\mathcal{F}$.

Assumption 3.2. Consider the partial order induced on functions in $\mathcal{F}$ by the dominated relation $\prec$. Then, every totally ordered subset of $\mathcal{F}$ (that is, an ordered sequence of dominated predictors) has an upper bound in $\mathcal{F}$.

This condition holds, for instance, when we use regularization (like in ridge regression) to restrict the function class.

Definition 3.3 (Minimax Pareto fairness (Martinez et al., 2020))**.** Minimax Pareto fair predictors are defined as predictors among the Pareto front which minimize the worst-off group's risk as follows:

$$\mathcal{F}_{\text{PMM}} := \{\arg\min_{f \in \mathcal{P}_{\mathcal{F}}} \max_{g \in [K]} R(f, D_g)\}. \tag{4}$$

Definition 3.4 (Lexical minimax-fairness (Diana et al., 2021a))**.** Let $\bar{f}(j)$ be the group with the j^{th} highest risk for a predictor f, ties broken arbitrarily. Define the lowest risks γ_j in a set of nested hypothesis classes $\mathcal{F}_{(j)}$ as follows:

$$\begin{aligned} \gamma_j &:= \min_{f \in \mathcal{F}_{(j-1)}} R(f, D_{\bar{f}(j)}), \\ \mathcal{F}_{(j)} &:= \{f \in \mathcal{F}_{(j-1)} : R(f, D_{\bar{f}(j)}) = \gamma_j\}, \end{aligned} \tag{5}$$

where $1 \leq j \leq K$ and $\mathcal{F}_0 := \mathcal{F}$. Then, a lexical minimax or lexifair predictor of level ℓ where $1 \leq \ell \leq K$ is defined as a predictor f such that for all $j \leq \ell, R(f, D_{\bar{f}(j)}) \leq \gamma_j$. A lexifair predictor of level K will have $R(f, D_{\bar{f}(j)}) = \gamma_j$ for all j.

3.4. Algorithm for solving the minimax problem

To find the minimax-fair solution of Eq. (3), we will employ the method proposed by Diana et al. (2021b). The main advantage is that it can be used as a wrapper around arbitrary hypothesis classes. It works by making repeated calls to an ERM solver for the given class after reweighting the data points. In contrast the gradient-based approaches of Abernethy et al. (2022) require computing gradients of the loss function with respect to the model parameters. Thus it is unclear how to use them, for instance, in the case of decision trees which are an important class for our analysis.

4. Main Result – ERM can be Minimax-fair

We first state a restrictive condition for our result which, intuitively, implies that the group information is perfectly recoverable from the features and the function class is sufficiently expressive. We will relax this assumption in Section 4.3.

4.1. Perfectly recoverable groups

We make the following assumption on D and $\mathcal{F}$.

Assumption 4.1 (Sufficiency condition)**.** The groups are *recoverable* from the features that is there exists functions $A_g \in \mathcal{F}$ that can determine group membership for any x,

$$A_g(x) = \begin{cases} 1 & \text{if } x \in D_g \\ 0 & \text{if } x \notin D_g \end{cases}, \ \forall g \in [K] \tag{C1}$$

and the function class $\mathcal{F}$ is *closed* under addition-and-multiplication that is if f, g, h are in $\mathcal{F}$, then

$$e : e(x) = h(x)f(x) + (1 - h(x))g(x) \text{ is also in } \mathcal{F}. \tag{C2}$$

Condition (C1) trivially holds if the features include the group attribute. Otherwise, it requires that the attribute is a function (from $\mathcal{F}$) of the features. Condition (C1) can be reinterpreted as saying that an *interpolating* classifier exists for predicting group labels, where interpolating classifiers are functions that can predict the exact labels for the train set (e.g. see (Wyner et al., 2017, Page 8)).

Condition (C2) (where the function h is fixed to be the group indicator function A_g in the definition) holds for the class of *decoupled* functions defined by Dwork et al. (2018) to be functions which learn separate predictors for each group, given Condition (C1) holds. Decision trees where the first split is on the group attribute, $f(x) = \{f_1(x)$ if $A_g(x) = 1$ else $f_0(x)\}$ is also an example where $f_0(x), f_1(x)$ are any two classifiers from $\mathcal{F}$. More generally, decision tree based predictors with unbounded depth like random forests and boosted trees, and universal approximators like neural networks satisfy the condition.

Remark 4.2 (Decoupled classes). Condition 4.1 implies that $\mathcal{F}$ is a *decoupled class*: *cf.* Section (C) in the appendix.

Remark 4.3 (Overlapping groups). Condition (C1) is not satisfied if the dataset has overlapping groups, that is, same data points belonging to more than one group e.g., race, gender. In such a case, we can redefine the group attribute to consider all intersections of the groups e.g., Asian male, Black female, and so on.

Theorem 4.4. *The structure shown in Figure 1 holds.*

(a) Pareto minimax implies minimax.

(b) Lexical minimax implies Pareto minimax.

(c) Group-optimal implies Lexical minimax.

(d) Every group-optimal solution is an ERM.

(e) Under the sufficiency condition (C1) *and* (C2)*, ERM satisfies group-optimality.*

The result implies that under the sufficiency condition, ERM is group-optimal, and due to the nested structure, it is lexical minimax-fair, Pareto minimax-fair, and minimax-fair.

The proof is deferred to Section A in the appendix. The first four statements follow from the definitions of the minimax-fairness notions and group-optimality. A brief justification for the last statement can be given as follows. Suppose the statement was not true, that is ERM is not group-optimal. Then we can improve the overall risk of the ERM solution by composing it with the group-specific optimal predictor, given that groups are recoverable. This new predictor by our expressiveness condition will still be in the hypothesis class.

This would mean that ERM does not minimize overall risk which is a contradiction.

4.2. ERM can achieve a broader class of fairness notions

A definition of classifier fairness requires that the predictions are *calibrated* for each group (Kleinberg et al., 2017).

Definition 4.5 (Well-calibration across groups)**.** Given a predictor $f : x \mapsto f(x) \in [0, 1]$ which outputs a real value between $[0, 1]$ in a binary classification problem of predicting $y \in \{0, 1\}$ from x, we say that f is well-calibrated across groups if $\mathbb{E}[Y|G = g, f(x) = t] = t$ for all $g \in [K]$ and $t \in [0, 1]$. Here, G is the random variable denoting the known or unknown group attribute.

Proposition 4.6. *If ERM is performed using a calibrated loss*[1] *and Assumption 4.1 holds, then any minimizer of empirical risk is well-calibrated across groups.*

Proof. By Theorem 4.4 (e), any ERM solution is group optimal and a minimizer of the empirical risk over each individual group D_i. It therefore follows from the definition of calibrated loss that it is calibrated for each group, and is consequentially well-calibrated over the entire dataset. □

This result is also shown in Liu et al. (2019) under different assumptions on the loss function which determines the ability to find the conditional expectation $\mathbb{E}[y|x]$ using ERM.

Given access to a well-calibrated model, Corbett-Davies et al. (2017) show that setting group-specific thresholds on the predictions results in classifiers maximizing *utility* (say, accuracy) while satisfying fairness constraints such as (conditional) demographic parity and false positive rate equality. Thus, achieving a calibrated model for each group by ERM we can repurpose the predictors to satisfy different fairness properties based on the application context. This highlights another advantage of ERM over specialized fair learning methods, in addition to minimax-fairness.

4.3. When group information is imperfectly recoverable

Theorem 4.4 requires perfect recovery of the group information by a function $A_g(x)$. We can relax this requirement to datasets where we can only imperfectly recover group information. We define imperfect recovery as the case when we have a classifier for the group (that is, an approximation to $A_g(x)$) which has perfect precision and at least k recall.

Definition 4.7 (k-recoverable group)**.** A group g is k-recoverable in a given dataset D if there exists a group classifier $\tilde{A}_g(x) \in \{0, 1\}$ such that the subset of data with $\tilde{A}_g(x) = 1$, written as $D_{\tilde{A}_g(x)=1}$, is correctly predicted to be group g and contains at least k-fraction of the total data points with attribute g in D. That is, $\frac{|D_{\tilde{A}_g(x)=1}|}{|D_g|} \geq k$.

Next we define a relaxed notion of group-optimality in the case of imperfect recovery which requires that a predictor achieves the same risk as the group-optimal predictor on at least k-fraction of the data points for the group.

Definition 4.8 (k-optimal predictor for a group)**.** A predictor f is k-optimal compared to any group-optimal classifier $f_{g\_} \in \mathcal{F}_g$ on group g if there exists a subset of data points $\tilde{D}_g \subseteq D$ consisting solely of data points with attribute g and recovering at least k-fraction of the total points with attribute g (that is, $A_g(x) = 1\ \forall (x, y) \in \tilde{D}_g$ and $\frac{|\tilde{D}_g|}{|D_g|} \geq k$), and the risk for f on $\tilde{D}_g$ is the same as f_g, that is $R(f, \tilde{D}_g) = R(f_g, \tilde{D}_g)$.

We show that ERM is k-optimal for the recoverable group.

Theorem 4.9. *Given that group g is k-recoverable in D and that the corresponding group classifier $\tilde{A}_g(x)$ exists in the function class $\mathcal{F}$, then ERM outputs k-optimal predictors for group g.*

The proof is included in Section B in the appendix. It follows similar arguments as used for proving Theorem 4.4 (e).

4.4. When labels are independent of the group information given the features

Continuing the scenario where we are not given the group information, we can still perform as well as the minimax predictor when the features are sufficiently expressive as defined below.

Assumption 4.10. Observed features are expressive enough such that the labels are conditionally independent of the group indicator given the features, that is $(Y \perp G)|X$.

Here, X, Y, G denote random variables for features, label, and group. This assumption holds trivially when features contain the group indicator variable. However, we consider settings where X does not explicitly contain G but has all the label-relevant information that knowing G provides.

Definition 4.11 (Bayes optimal predictor)**.** Function $f^*(x)$ is a Bayes optimal predictor for some distribution $\mathcal{D}$ if it minimizes the risk for each data point x sampled from $\mathcal{D}$. That is $f^*(x) \in \arg\min_{\hat{y}} \mathbb{E}_{y \sim \mathcal{D}_{Y|X=x}}[\ell(\hat{y}, y)]$.

Assumption 4.12. $\mathcal{F}_{\text{ERM}}$ are Bayes optimal predictors for the empirical distribution of the dataset D.

For the next claim, we will define the joint distribution $D_{Y,X}$ to be the uniform distribution over the data points in the given dataset D. Definitions of marginal and conditional probabilities yield $D_X, D_{Y|X}$. Distributions involving G

[1] A calibrated loss refers to one such as squared loss or log loss (Banerjee et al., 2005) that results in a calibrated function over any training set.

Table 1: **Datasets.** ACS datasets are curated for 4 US states (NY, CA, TX, IN). Datasets for different domains combined with different group types result in 36 datasets in total. All labels are binary. Dataset sources are given in Section E.2 in the appendix

Dataset	Samples	Features	Groups (levels)
ACS Income (NY, CA, TX, IN)	103021, 195665, 135924, 35022	26	Sex (2), Race (4)
ACS Employment (NY, CA, TX, IN)	196967, 378817, 268100, 67680	36	Sex (2), Race (4)
ACS Health Insurance (NY, CA, TX, IN) (Ding et al., 2021)	67551, 138554, 98928, 24330	38	Sex (2), Race (4)
UCI Adult Income (Dua & Graff, 2017)	39073	14	Sex (2), Race (3)
COMPAS (ProPublica, 2020)	7214	6	Race (4)
Diabetes (Strack et al., 2014)	101766	23	Age (5)
Drug Consumption (Fehrman et al., 2015)	1885	8	Country (3)
eICU (Pollard et al., 2018)	20000	25	Sex (2), Race (2)
Default (Yeh & hui Lien, 2009)	30000	33	Sex (2)
Communities (Redmond & Baveja, 2002)	1994	123	Race (4)
German Credit (Dua & Graff, 2017)	1000	20	Sex & Marital status (4)
Heart (Chicco & Jurman, 2020)	299	12	Sex (2)
Marketing (Moro et al., 2014)	45211	48	Job (12)

such as $D_{X|G}, D_{Y|X,G}$ will refer to a hypothetical dataset where G could be observed.

Proposition 4.13. *Given Assumptions 4.10, 4.12 hold, then ERM predictors $\mathcal{F}_{\text{ERM}}$ achieve the same or better risk than the group-specific optimal predictors $\mathcal{F}_g$ and the minimax-fair predictor $\mathcal{F}_{\text{MM}}$.*

The proof is included in Section D in the appendix. Thus, in the absence of group information, access to sufficiently expressive features and function class means that ERM minimizes both the group-specific risk and the worst group-specific risk. When the group information is not encoded in the features, for example when groups are based on the labels, ERM and minimax-fair methods can differ in performance, as seen in experiments from Martinez et al. (2020).

5. Empirical Study

We perform a large-scale study to test how well do our theoretical results generalize to practical scenarios where the assumptions may not hold.[2] Accordingly, we include multiple real datasets and hypothesis classes with varying levels of expressiveness in the study. Table 1 lists the 36 datasets used in the study. We chose to test on tabular datasets following much of the prior work in fairness (Diana et al., 2021b). Moreover, similar observations have already been made on high-dimensional datasets (Gardner et al., 2022; Pfohl et al., 2022; Zong et al., 2023). We train minimax-fair models for Diana et al. (2021b)'s method using their code[3] with different classifiers.

Classifiers. We use random forest (Breiman, 2001), a multilayer perceptron (MLP) architecture used in a tabular data benchmark (Gorishniy et al., 2021), decision trees, linear support vector classifiers (SVC, Fan et al., 2008), and logistic regression models. We use the default ERM solvers available in the `scikit-learn` Python package (Pedregosa et al., 2011). We use stochastic gradient descent for MLPs.

Metrics. We compare the models in terms of their worst-case accuracy and negative log loss across groups. Higher value of the metrics is better as a convention.

Setup. Datasets are divided randomly into 70-30 train-test split. For the optimization procedure of the minimax method, we set the convergence threshold as 10^{-12} and run at most 10000 iterations, except for MLP we use 200 iterations to reduce compute time. Rest of the hyperparameters are detailed in Table 5 in the appendix. Experiments were run on a compute cluster using 36 nodes with an Intel Xeon 2.9 GHz processor, 1 NVIDIA RTX8000 GPU and 24 GB system memory for each node.

5.1. Main questions to test Theorem 4.4 in practice

We design the experiments to test the following questions.

Q1. Does ERM perform differently than minimax model on the train set i.e. does Theorem 4.4 hold in practice?

[2] Code to reproduce the experiments is available at https://github.com/amazon-science/rethinking-minimax-fairness

[3] https://github.com/amazon-science/minimax-fair

Model type	ERM v/s GROUP-OPTIMAL		ERM v/s MINIMAX	
	p-value equivalence	p-value non-inferior	p-value equivalence	p-value non-inferior
Logistic Regression	0.9955	**0.0039**	0.7324	0.7324
Linear SVC	0.1795	0.0562	1.0000	**0.0000**
Decision Tree depth 8	0.9984	0.9984	0.9858	**0.0106**
Random Forest	**0.0000**	**0.0000**	0.9801	**0.0010**
MLP	0.0694	0.0694	0.9946	**0.0050**

Table 2: Q1. **ERM ≈ group-optimal or minimax-fair on train set (via hypothesis tests).** p-values from the two hypothesis tests of ERM against group-optimal and minimax-fair models. Metric is negative logloss and threshold is $t = 0.01$. Significance level is set to p-value<0.05, values in bold. We reject the hypothesis that ERM is not equivalent (or is inferior) to Group-optimal for random forest models. The p-value for MLP models is small as well. We reject the hypothesis that ERM is inferior to minimax in all model classes except logistic regression models.

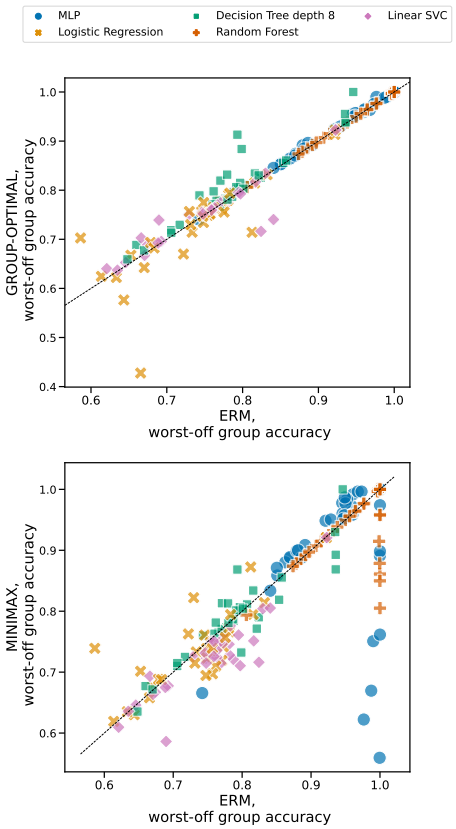

Figure 2: Q1. **ERM ≈ group-optimal or minimax-fair on train set (visually).** Worst-case accuracy of ERM vs group-optimal (top) and minimax models (bottom) on train set. Each point on the plot corresponds to a model type trained on a dataset for one of the three methods. Dotted line is $y = x$. Visually, ERM matches the worst-case performance for group-optimal and minimax models for most classes. Figure 3 in the appendix makes the same observation when evaluating models on the test set.

Q2. How does the result vary with model expressiveness?

Q3. Does Theorem 4.4 generalize to the test set?

Owing to the optimization challenges of the minimax problem, we also compare ERM with group-optimal models, trained via ERM on the group's data. Theorem 4.4 (a-c) implies that group-optimal models will be minimax-fair.

Hypothesis tests. To answer the above questions, we adopt the hypothesis testing framework which is prevalent in the physical sciences. We test whether the behavior predicted by the theoretical claim (ERM ≈ minimax models) matches the empirical results. If so, then we conclude that the claim is *likely* to hold. This, in general, does not imply that the claim is true in practical scenarios not covered by the theory.

Accordingly, our null hypothesis is that ERM behaves differently from minimax (or group-optimal) models, in order to reject the null. We cannot use the standard null hypothesis testing framework since it assumes that the null is the no difference case. Therefore, we use hypothesis tests from the equivalence testing literature (Wellek, 2002, Chapter 5) that flip the null and the alternative hypothesis.

Let ERM denote a random variable for the worst-case risk of ERM. Similarly, MINIMAX denotes a random variable for the worst-case risk of minimax-fair predictor. Each dataset provides us with a pair of samples for ERM and MINIMAX. We need a threshold t to say that the difference between the two methods is practically significant. Then, we test for equivalence of ERM and minimax via the hypotheses,

$$\mathrm{H}_0 : \text{ERM} \leq \text{MINIMAX} - t \text{ or ERM} \geq \text{MINIMAX} + t$$
$$\mathrm{H}_{\text{alt}} : |\text{ERM} - \text{MINIMAX}| < t$$

The composite null hypothesis is that ERM and minimax differ by at least t, while the alternate is that they do not.

Additionally, we can also test if ERM is better than minimax using non-inferiority tests via,

$$\mathrm{H}_0 : \text{ERM} \leq \textsc{Minimax} - t$$
$$\mathrm{H}_{\text{alt}} : \text{ERM} > \textsc{Minimax} - t$$

Table 3: Q2. **Ablation study by decreasing model expressiveness.** We train decision trees with three values of maximum depth. The p-values is shown for accuracy metric, threshold t=0.01. Significance level is taken as p-value<0.05, values in bold. We observe that p-values increase as max depth (model expressiveness) is decreased.

Model type	ERM v/s MINIMAX p-value equivalence	p-value non-inferior
Decision Tree depth 8	**0.0281**	**0.0281**
Decision Tree depth 4	**0.0075**	**0.0075**
Decision Tree depth 2	0.1399	0.1399

Here, the null hypothesis is that ERM is worse than minimax by at least t. We set t as 0.01 which for accuracy means a 1 percentage point difference. We conduct two one-sided t-tests which are commonly used for equivalence and non-inferiority testing (Schuirmann, 1987), implemented in the Python package `statsmodels` by the function `stats.weightstats.ttost_paired`. This is a parametric test and makes the standard assumptions for t-tests including that the variables are approximately normally distributed.

5.2. Results and discussion

We include group information as a feature in all the experiments. Thus, Condition (C1) holds by design. According to Theorem 4.4, we expect that ERM should perform similarly to minimax (and group-optimal) models for expressive model classes that satisfy Condition (C2). That is, we expect to reject the nulls for equivalence and/or non-inferiority tests. Figure 2 plots the minimum accuracy across groups for different classes. This data is analyzed using hypothesis tests in Table 2.

The main takeaways are as follows.

- **Theorem 4.4 is likely to hold for random forests and MLPs.** Results for random forests in Table 2 favor the hypothesis that ERM is equivalent to group-optimal models and is not inferior to minimax (since p-value < 0.05). Table 6 shows the results for accuracy metric where our theoretical result is validated for both random forest and MLP classes.
- **Theorem 4.4 is unlikely to hold as model expressiveness is decreased** When we make the hypothesis classes less expressive by decreasing the maximum depth of the decision trees, we observe that p-values are higher (result unlikely to hold) for depth 2 than for depth 8 in Table 3.
- **Result is likely to generalize to the test set.** For random forests and MLPs, Table 4 shows that Theorem 4.4 is likely to hold (ERM is group-optimal and minimax) even when models are evaluated on the held-out test sets. This differs from our setup as we only analyze train set risk. This suggests that the result may hold for population risk which is desirable for practice.

In addition to testing Theorem 4.4, we present preliminary evidence for Theorem 4.9 in Section E.5 in the appendix. To simulate the imperfectly recoverable groups, we omit the group information from the features given to the models, and repeat the experiments comparing ERM with group-optimal and minimax-fair models. Results suggest that Theorem 4.9 is likely to hold for random forests and MLPs.

Lastly, we remark the inconsistencies in the results for MLPs. We expect to reject both the hypotheses for MLPs since these are a flexible model class. However, Table 2 shows that we cannot reject both the null hypotheses for group-optimal models. We believe that this behavior can be caused by difficulties with stochastically optimizing non-convex objectives. In particular, the iterative (re)training that is part of the minimax optimization is similar to an annealing strategy, and may make the optimization more likely to stop in minima with different properties. Note that the minimax-fair learning method we use (Diana et al., 2021b) does not have convergence guarantees in the case of a non-convex classification loss. This may also explain the significantly low accuracy for minimax models trained with MLP in the bottom plot of Figure 2. We note that necessary changes in optimization from the minimax method of Diana et al. (2021b) to other approaches could lead to apparently different behavior.

5.3. Limitations

A major limitation of the work is that testing whether Condition (C2) holds for any given hypothesis class is difficult. This assumption was critical in Theorem 4.4 to show that ERM can be minimax-fair. Our emphasis here was solely on accuracy-based measures of fairness where a better accuracy for any group is desirable. Alternatively, notions of parity can be the preferred fairness goals under some contexts. An important limitation of the work is that we assume that training data perfectly represents the world as it should be, that is, there are no distribution shifts in features or labels. Relatedly, we ignore estimation and optimization errors due to small sample size during minimax learning. Finally, the preference for using flexible hypothesis classes

Table 4: Q3. **ERM ≈ group-optimal or minimax-fair on test.** p-values from the two hypothesis tests of ERM against group-optimal and minimax-fair models. Significance level is set to p-value<0.05, values in bold. For the case of negative logloss metric, we reject the hypothesis that ERM is inferior to Group-optimal for all models. We reject the hypothesis that ERM is inferior to minimax in all model classes except logistic regression and decision tree models.

(a) Metric = negative logloss, threshold $t = 0.01$

Model type	ERM v/s GROUP-OPTIMAL p-value equivalence	ERM v/s GROUP-OPTIMAL p-value non-inferior	ERM v/s MINIMAX p-value equivalence	ERM v/s MINIMAX p-value non-inferior
Logistic Regression	0.9938	**0.0057**	0.7877	0.7877
Linear SVC	0.0671	**0.0239**	1.0000	**0.0000**
Decision Tree depth 8	0.9992	**0.0007**	0.9245	0.0605
Random Forest	0.6403	**0.0082**	0.2191	**0.0025**
MLP	0.9685	**0.0216**	1.0000	**0.0000**

(b) Metric = accuracy, threshold $t = 0.01$

Model type	ERM v/s GROUP-OPTIMAL p-value equivalence	ERM v/s GROUP-OPTIMAL p-value non-inferior	ERM v/s MINIMAX p-value equivalence	ERM v/s MINIMAX p-value non-inferior
Logistic Regression	0.9155	**0.0007**	0.3654	**0.0004**
Linear SVC	**0.0007**	**0.0007**	0.9935	**0.0000**
Decision Tree depth 8	**0.0220**	**0.0000**	0.4671	**0.0000**
Random Forest	0.2855	**0.0001**	0.1968	**0.0000**
MLP	**0.0406**	**0.0000**	0.9183	**0.0000**

to be minimax-fair has to be carefully considered along with the need for interpretability.

6. Conclusion

Our work shows that ERM can satisfy minimax notions of fairness given that (1) the hypothesis classes are sufficiently expressive and (2) group information can be predicted from the features. This explains the overwhelming evidence from recent work that finds ERM is rarely outperformed by more sophisticated minimax learning methods on its performance on worst-off group (Gardner et al., 2022; Pfohl et al., 2022; Zong et al., 2023; Martinez et al., 2020). We provide more comprehensive evidence for the same on multiple tabular datasets from different application domains.

An important direction of further work is to verify and ensure that the sufficiency condition is satisfied while performing ERM for a given model class. This may include designing model architectures that are decoupled, that is, the model has a dedicated function for each group to predict their labels. Verifying this condition for the model classes and datasets used in our experiments will provide more conclusive evidence that the theoretical results continue to hold in practice. Future work should study better ways to learn minimax-fair models (via ERM or otherwise) when the groups are not encoded in the features and yet are correlated with the labels. Another interesting open question is to study whether the results generalize to unseen test sets with possibly different distributions from the train sets. In essence, our findings suggest including ERM as a potential solution when optimizing for worst-case performance.

Acknowledgements

We thank the anonymous reviewers for their thoughtful feedback which improved the paper. We are grateful to the members of the Amazon Tübingen lab for helpful discussions. RC and HS were partially supported by National Science Foundation award 1845487, HS also acknowledges support from the National Science Foundation under NSF Award 1922658.

References

U. S. Department of Commerce, Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a (Computer Files), 1990.

U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan,

1992a.

U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan, 1992b.

U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File), 1995.

Abernethy, J. D., Awasthi, P., Kleindessner, M., Morgenstern, J., Russell, C., and Zhang, J. Active sampling for min-max fairness. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 53–65. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/abernethy22a.html.

Banerjee, A., Guo, X., and Wang, H. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005. doi: 10.1109/TIT.2005.850145. URL https://ieeexplore.ieee.org/document/1459065.

Breiman, L. Random forests. *Machine learning*, 45: 5–32, 2001. URL https://doi.org/10.1023/A:1010933404324.

Chicco, D. and Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):1–16, 2020. Dataset available at https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098095. URL https://doi.org/10.1145/3097983.3098095.

Diana, E., Gill, W., Globus-Harris, I., Kearns, M., Roth, A., and Sharifi-Malvajerdi, S. Lexicographically fair learning: Algorithms and generalization. In *2nd Symposium on Foundations of Responsible Computing*, pp. 1, 2021a. URL https://drops.dagstuhl.de/opus/volltexte/2021/13874/pdf/LIPIcs-FORC-2021-6.pdf.

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 66–76, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462523. URL https://doi.org/10.1145/3461702.3462523.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bYi_2708mKK. Datasets accessible by Python package https://github.com/zykls/folktables.

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

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 119–133. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/dwork18a.html.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(61):1871–1874, 2008. URL http://jmlr.org/papers/v9/fan08a.html.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N. The five factor model of personality and evaluation of drug consumption risk, 2015. URL https://arxiv.org/abs/1506.06297. Dataset available at https://archive.ics.uci.edu/ml/datasets/Drug%20consumption+(quantified).

Gardner, J. P., Popovi, Z., and Schmidt, L. Subgroup robustness grows on trees: An empirical baseline investigation. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=6QvmtRjWNRy.

Globus-Harris, I., Kearns, M., and Roth, A. An algorithmic framework for bias bounties. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1106–1124, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi:

10.1145/3531146.3533172. URL https://doi.org/10.1145/3531146.3533172.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=i_Q1yrOegLY.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Hofmann, H. German credit data. Dataset available at https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4337–4348. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hsieh21a.html.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2029–2037. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hu18a.html.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 247–254, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314287. URL https://doi.org/10.1145/3306618.3314287.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pp. 43. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. URL https://drops.dagstuhl.de/opus/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43.pdf.

Kohavi, R. and Becker, B. Adult income dataset. Dataset available at https://archive.ics.uci.edu/ml/datasets/adult.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1–2):311–337, jul 2019. ISSN 0025-5610. doi: 10.1007/s10107-019-01374-3. URL https://doi.org/10.1007/s10107-019-01374-3.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4051–4060. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/liu19f.html.

Martinez, N., Bertran, M., and Sapiro, G. Minimax pareto fairness: A multi objective perspective. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6755–6764. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/martinez20a.html.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, 2021. doi: 10.1146/annurev-statistics-042720-125902. URL https://doi.org/10.1146/annurev-statistics-042720-125902.

Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S016792361400061X. Dataset available at https://archive.ics.uci.edu/ml/datasets/bank+marketing.

Muandet, K. Impossibility of collective intelligence, 2022. URL https://arxiv.org/abs/2206.02786.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Pethick, T., Chrysos, G., and Cevher, V. Revisiting adversarial training for the worst-performing class. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=wkecshlYxI.

Pfohl, S. R., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., and Shah, N. H. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports*, 12(1):1–13, 2022. URL https://www.nature.com/articles/s41598-022-07167-7.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5:180178, 2018. Dataset available from https://physionet.org/content/eicu-crd/2.0/.

ProPublica. Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis, 2020. Broward County Clerk's Office, Broward County Sherrif's Office, Florida Department of Corrections, ProPublica. Accessed: 2022-09-30.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002. ISSN 0377-2217. doi: https://doi.org/10.1016/S0377-2217(01)00264-8. URL https://www.sciencedirect.com/science/article/pii/S0377221701002648. Dataset available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime.

Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680, 1987. URL https://link.springer.com/article/10.1007/BF01068419.

Shekhar, S., Fields, G., Ghavamzadeh, M., and Javidi, T. Adaptive sampling for minimax fair classification. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24535–24544. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/cd7c230fc5deb01ff5f7b1be1acef9cf-Paper.pdf.

Singh, H., Mhasawade, V., and Chunara, R. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, 1(4):1–17, 04 2022. doi: 10.1371/journal.pdig.0000023. URL https://doi.org/10.1371/journal.pdig.0000023.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, 2014. Dataset available at https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.

Vapnik, V. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. doi: 10.1109/72.788640. URL https://ieeexplore.ieee.org/document/788640.

Wellek, S. *Testing statistical hypotheses of equivalence*. Chapman and Hall/CRC, 2002. URL https://doi.org/10.1201/EBK1439808184.

Williamson, R. and Menon, A. Fairness risk measures. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/williamson19a.html.

Wyner, A. J., Olson, M., Bleich, J., and Mease, D. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017. URL http://jmlr.org/papers/v18/15-240.html.

Yeh, I.-C. and hui Lien, C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2007.12.020. URL https://www.sciencedirect.com/science/article/pii/S0957417407006719. Dataset available at https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Zhai, R., Dan, C., Suggala, A., Kolter, J. Z., and Ravikumar, P. K. Boosted CVar classification. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=INsYqFjBWnF.

Zietlow, D., Lohaus, M., Balakrishnan, G., Kleindessner, M., Locatello, F., Scholkopf, B., and Russell, C. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10400–10411, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01016. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01016.

Zong, Y., Yang, Y., and Hospedales, T. MEDFAIR: Benchmarking fairness for medical imaging. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6ve2CkeQe5S.

A. Proof of the Main Result in Theorem 4.4

Theorem 4.4 (Nested structure)**.** *The structure shown in Figure 1 holds.*

(a) Pareto minimax implies minimax under Condition (3.2).

(b) Lexical minimax implies Pareto minimax.

(c) Group-optimal implies Lexical minimax.

(d) Every group-optimal solution is an ERM.

(e) Under the sufficiency condition (C1) and (C2), ERM satisfies group-optimality.

We prove the five statements below in the order (d), (e), (c), (b), and then (a).

Proof of (d). **Every member of the set of group-optimal solutions minimizes the empirical risk.**

We assume a group optimal function $f^*(x)$ exists, then:

$$\min_f \sum_{g\in[K]} \sum_{(x,y)\in D_g} \ell(f(x),y) \tag{6}$$

$$\geq \sum_{g\in[K]} \min_f \sum_{(x,y)\in D_g} \ell(f(x),y) \tag{7}$$

$$= \sum_{g\in[K]} \sum_{x\in D_g} \ell(f^*(x),y) \tag{8}$$

$$\geq \min_f \sum_{g\in[K]} \sum_{x\in D_g} \ell(f(x),y) \tag{9}$$

and

$$\min_f \sum_{g\in[K]} \sum_{(x,y)\in D_g} \ell(f(x),y) = \sum_{g\in[K]} \sum_{(x,y)\in D_g} \ell(f^*(x),y), \tag{10}$$

as required. □

Remark A.1 (Existence of a group-optimal predictor). The function $f^* \in \mathcal{F}$ which is group-optimal simultaneously for each group exists if the sufficiency condition C1 and C2 is satisfied. It can be constructed by composing the group-specific optimal predictors as follows, $f^*(x) = \sum_{g\in G} A_g(x) f_g^*(x)$, where $f_g^*(x) \in \arg\min_{f\in\mathcal{F}} \sum_{(x,y)\in g} \ell(f(x),y)$.

Proof of (e). **ERM satisfies group-optimality under sufficiency condition.**

We use $f^\dagger$ to indicate an empirical risk minimizer (ERM), and f^g to indicate a function that minimizes only over group g. Then for every $g \in [K]$

$$\sum_{(x,y)\in D} \ell(f^\dagger(x),y) \tag{11}$$

$$\text{[Condition (C1)]} \quad = \sum_{(x,y)\in D} A_g(x)\ell(f^\dagger(x),y) + (1-A_g(x))\ell(f^\dagger(x),y) \tag{12}$$

$$\text{[definition of group-optimal]} \quad \geq \sum_{(x,y)\in D} A_g(x)\ell(f^g(x),y) + (1-A_g(x))\ell(f^\dagger(x),y) \tag{13}$$

$$= \sum_{(x,y)\in D} \ell(A_g(x)f^g(x) + (1-A_g(x)f^\dagger(x),y) \tag{14}$$

$$\text{[Condition (C2) and optimality of ERM]} \quad \geq \sum_{(x,y)\in D} \ell(f^\dagger(x),y) \tag{15}$$

and subtracting $\sum_{(x,y)\notin D_g} \ell(f^\dagger(x),y)$ from equations (11), (13), and (15) we have:

$$\sum_{(x,y)\in D_g} \ell(f^\dagger(x),y) = \sum_{(x,y)\in D_g} \ell(f^g(x),y)\, \forall g \in [K] \tag{16}$$

□

To simplify notation, denote the empirical risk of a predictor f on group g as $R(f, D_g) := 1/|D_g| \sum_{(x,y)\in D_g} \ell(f(x), y)$.

Proof of (c). **Group-optimal implies lexical minimax.**

Consider a group optimal predictor $f^* \in \mathcal{F}$.

From Definition 3.4, recall that a leximax predictor f of level $1 \leq l \leq K$ is a function that satisfies,

$$R(f, D_{\bar{f}(j)}) \leq \gamma_j, \forall\, 1 \leq j \leq l,$$

where $\gamma_j := \min_{f\in\mathcal{F}_{(j-1)}} R(f, D_{\bar{f}(j)})$. Here, $\bar{f}(j)$ refers to the group with the j^{th} highest group risk for predictor f, as defined earlier.

Since $\mathcal{F}_{(j-1)} \subseteq \mathcal{F}$, we observe that

$$\gamma_j \geq \min_{f\in\mathcal{F}} R(f, D_{\bar{f}(j)}) =: R(f^*, D_{\bar{f}(j)}). \tag{17}$$

Thus, the group optimal predictor satisfies lexicographic fairness. □

Proof of (b). **Lexical minimax (of level K) implies Pareto minimax.**

Consider a lexical minimax predictor $f_L^* \in \mathcal{F}$ of level K.

From Definition 3.3, recall that a Pareto minimax predictor is an element of the Pareto front $\mathcal{P}_\mathcal{F}$ which minimizes the worst-case group risk.

First, we show that $f_L^* \in \mathcal{P}_\mathcal{F}$.

If possible, suppose $f_L^* \notin \mathcal{P}_\mathcal{F}$. This implies that there exists a predictor $f' \in \mathcal{F}$ that Pareto dominates f_L^*. That is,

$$\exists g, R(f', D_g) < R(f_L^*, D_g) \tag{18}$$

and the inequality $\leq$ holds for all groups.

Let $j \in [K]$ be the highest index in the ordering of groups by $\bar{f}'(.)$ for which f' has strictly lower risk than f_L^*. That is, $R(f', D_{\bar{f}'(i)}) = R(f_L^*, D_{\bar{f}_L^*(i)}), \forall i < j$ and $R(f', D_{\bar{f}'(j)}) < R(f_L^*, D_{\bar{f}_L^*(j)})$. Such an index j always exists due to (18). By definition of lexical minimax-fairness, we know that for j,

$$R(f', D_{\bar{f}'(j)}) < R(f_L^*, D_{\bar{f}_L^*(j)}) = \min_{f\in\mathcal{F}_{(j-1)}} R(f, D_{\bar{f}(j)}).$$

which will be a contradiction if $f' \in \mathcal{F}_{(j-1)}$. We can prove this by induction from $1 \leq i \leq j$. For the base case, $f' \in \mathcal{F}$, by assumption.

For the $i = 1$ case, consider $\mathcal{F}_{(1)}$ defined as $\{f \in \mathcal{F} : R(f, D_{\bar{f}(1)}) = \gamma_1\}$.

Since $R(f', D_{\bar{f}'(1)}) = R(f_L^*, D_{\bar{f}_L^*(1)}) = \gamma_1$, we know that $f' \in \mathcal{F}_{(1)}$.

Continuing the argument till $i = j - 1$, we can show that $f' \in \mathcal{F}_{(j-1)}$.

Having shown that $f_L^* \in \mathcal{P}_\mathcal{F}$, we now show that f_L^* minimizes the worst-case group risk among predictors on the Pareto front, thus, implying it is Pareto minimax.

That is, we need to show that for all $f \in \mathcal{P}_\mathcal{F}$,

$$\max_{g\in G} R(f_L^*, D_g) \leq \max_{g'\in G} R(f, D_{g'}).$$

If possible, suppose this is not true. Thus, there exists a $f' \in \mathcal{P}_\mathcal{F}$ for which

$$\max_{g} R(f_L^*, D_g) > \max_{g'} R(f', D_{g'}).$$

By the definition of lexical minimax-fairness,

$$\min_{f\in\mathcal{F}} R(f, D_{\bar{f}(1)}) \geq R(f^*_L, D_{\bar{f}^*_L(1)}) \tag{19}$$

$$> \max_{g'} R(f', D_{g'}) \tag{20}$$

$$\geq R(f', D_{\bar{f}'(1)}) \tag{21}$$

Thus, $\min_{f\in\mathcal{F}} R(f, D_{\bar{f}(1)}) > R(f', D_{\bar{f}'(1)})$.

This is a contradiction since $f' \in \mathcal{P}_{\mathcal{F}} \subseteq \mathcal{F}$. □

Proof of (a). **Pareto minimax implies minimax.**

Consider a minimax predictor $f^*_{\text{MM}} \in \mathcal{F}$. We want to show that a solution of the same minimax risk lies in the Pareto front, and therefore, a minimizer of the minimax risk subject to the additional constraint of Pareto efficiency is also generally minimax-fair.

To see this, we note that either f^*_{MM} is in the Pareto front, or it is dominated by another solution f^d that lies in the front. This follows from the technical condition 3.2 which assumes that each ordered sequence of successively dominated predictors in $\mathcal{F}$ terminates in a member of $\mathcal{F}$. Thus, by Zorn's lemma, we know that a maximal element f^d of the set $\mathcal{F}$ exists.

For f^d to dominate f^*_{MM} it must have no higher risk for any group and a strictly lower risk for one group. Hence, f^d must have lower or the same max group risk as f^*_{MM}. As $f^d \in \mathcal{F}$ and f^*_{MM} minimize the max group risk, they must have the same risk. □

B. Proof of Theorem 4.9

Theorem 4.9 states that when only a fraction of a group can be recovered from the dataset, ERM has the same risk as group-specific predictors on that recovered fraction of data points.

Theorem 4.9. *Given that group g is k-recoverable in D and that the corresponding group classifier $\tilde{A}_g(x)$ exists in the function class $\mathcal{F}$, then all $f_{\text{ERM}} \in \mathcal{F}_{\text{ERM}}$ are k-optimal predictors for group g.*

Proof. Since group g is k-recoverable, suppose the function $\tilde{A}_g(x)$ is such that it recovers at least k-fraction of D with attribute g $\frac{|D_{\tilde{A}_g(x)=1}|}{|D_g|} \geq k$ and $D_{\tilde{A}_g(x)=1}$ only contains data points with attribute g.

Then, use $\tilde{A}_g(x)$ to subset D into $D_{\tilde{A}_g(x)=1}$. We want to prove that $R(f_{\text{ERM}}, D_{\tilde{A}_g(x)=1}) = R(f_g, D_{\tilde{A}_g(x)=1})$ for all f_{ERM}, f_g.

Suppose there exists f_{ERM}, f_g such that $R(f_{\text{ERM}}, D_{\tilde{A}_g(x)=1}) \neq R(f_g, D_{\tilde{A}_g(x)=1})$.

Take any optimal predictor on $D_{\tilde{A}_g(x)=1}$ as $f_{\tilde{A}_g(x)=1} \in \arg\min_{f\in\mathcal{F}} R(f, D_{\tilde{A}_g(x)=1})$.

Construct a new predictor $\tilde{f}$ on D_g by composing $f_{\tilde{A}_g(x)=1}$ and f_g as follows

$$\tilde{f}(x) = \begin{cases} f_{\tilde{A}_g(x)=1}(x) & \text{if } \tilde{A}_g(x) = 1 \wedge A_g(x) = 1 \text{ (which is same as } \tilde{A}_g(x) = 1 \text{ since by assumption } D_{\tilde{A}_g(x)=1} \subseteq D_g) \\ f_g(x) & \text{else if } \tilde{A}_g(x) = 0 \wedge A_g(x) = 1. \end{cases}$$

Observe that since $f_{\tilde{A}_g(x)=1}$ is a minimizer, $R(f_{\tilde{A}_g(x)=1}, D_{\tilde{A}_g(x)=1}) \leq R(f_a, D_{\tilde{A}_g(x)=1})$. Following the proof for Theorem 4.4 (e), we can show that $R(f_{\text{ERM}}, D_{\tilde{A}_g(x)=1}) = R(f_{\tilde{A}_g(x)=1}, D_{\tilde{A}_g(x)=1})$. Since we assumed that $R(f_{\text{ERM}}, D_{\tilde{A}_g(x)=1}) \neq R(f_{A_i=a}, D_{\tilde{A}_g(x)=1})$, the inequality above is strict, $R(f_{\tilde{A}_g(x)=1}, D_{\tilde{A}_g(x)=1}) < R(f_a, D_{\tilde{A}_g(x)=1})$.

By the definition of k-recoverable, we can write the set $\{(x, y) \in D | \tilde{A}_g(x) = 1 \wedge A_g(x) = 1\} \equiv \{(x, y) \in D | \tilde{A}_g(x) = 1\}$ since $\tilde{A}_g(x)$ has perfect precision.

We can show that the risk of $\tilde{f}$ is lower than the risk of f_g on D_g.

$$\begin{aligned}
R(\tilde{f}, D_g) &= \frac{1}{|D_g|}\left(\sum_{(x,y)\in D_{\tilde{A}_g(x)=1}} \ell(\tilde{f}(x), y) + \sum_{(x,y)\in D_{\tilde{A}_g(x)=0 \wedge A_g(x)=1}} \ell(\tilde{f}(x), y)\right) \\
&= \frac{1}{|D_g|}\left(\sum_{D_{\tilde{A}_g(x)=1}} \ell(f_{\tilde{A}_g(x)=1}(x), y) + \sum_{D_{\tilde{A}_g(x)=0 \wedge A_g(x)=1}} \ell(f_g(x), y)\right) \\
&< \frac{1}{|D_g|}\left(\sum_{D_{A_g(x)=a}} \ell(f_g(x), y) + \sum_{D_{\tilde{A}(x)=0 \wedge A_g(x)=1}} \ell(f_g(x), y)\right) \\
&= R(f_g, D_g)
\end{aligned}$$

which contradicts the fact that f_g is a minimizer for group g. □

C. Decoupled Classifiers and Sufficiency Condition

Definition C.1 (Decoupled class)**.** Given K classifiers, $\vec{f} := (f_1, f_2, \ldots, f_K)$, a decoupled classifier is denoted by $\delta_{\vec{f}}(x) := f_g(x) \in \{0, 1\}$ where $A_g(x) = 1$. A decoupled class is a set of decoupled classifiers defined as $\delta(\mathcal{F}) := \{\delta_{\vec{f}} \mid \vec{f} \in \mathcal{F}^K\}$.

Remark 4.2. Condition 4.1 implies that $\mathcal{F}$ is a decoupled class.

Proof. We want to show that $\delta(\mathcal{F}) = \mathcal{F}$ when Condition 4.1 holds. First, $\delta(\mathcal{F}) \supseteq \mathcal{F}$ since any $f \in \mathcal{F}$ can be written as $\delta_{(f,f,\ldots,f)} \in \delta(\mathcal{F})$. Next we need to show $\delta(\mathcal{F}) \subseteq \mathcal{F}$. Let $\vec{f} = (f_1, f_2, \ldots, f_K)$ with $f_i \in \mathcal{F}$. We recursively define functions $h_2, \ldots, h_K \in \mathcal{F}$ with $h_K = \delta_{\vec{f}}$, which shows $\delta(\mathcal{F}) \subseteq \mathcal{F}$, as follows:

$$h_2 = (1 - A_2)f_1 + A_2 f_2 \quad \text{and} \quad h_i = (1 - A_i)h_{i-1} + A_i f_i, \;\; i = 3, \ldots, K.$$

Due to Condition 4.1, we have $h_2, \ldots, h_K \in \mathcal{F}$. Furthermore, for all $i = 2, \ldots, K$, it is $h_i(x) = f_g(x)$ on group g for $g = 1, \ldots, i$.

□

D. Proof of Proposition 4.13

Proposition 4.13. *Given Assumptions 4.10, 4.12 hold, then ERM predictors $\mathcal{F}_{\text{ERM}}$ achieve the same or better risk than the group-specific optimal predictors $\mathcal{F}_g$ and the minimax-fair predictor $\mathcal{F}_{\text{MM}}$.*

Proof. A Bayes optimal solution for the observed $D_{Y,X}$ is given by $f^*(x) \in \{\arg\min_{\hat{y}} \mathbb{E}_{y\sim D_{Y|X=x}}[\ell(\hat{y}, y)]\}$ which is the same set as $\mathcal{F}_{\text{ERM}}$ by Assumption 4.12.

Consider the group-specific risk,

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim D_{Y,X|G}}[\ell(f(x), y)] = \mathbb{E}_{x\sim D_{X|G}}\mathbb{E}_{y\sim D_{Y|X,G}}[\ell(f(x), y)] \\
&= \mathbb{E}_{x\sim D_{X|G}}\mathbb{E}_{y\sim D_{Y|X}}[\ell(f(x), y)],
\end{aligned}$$

since $D_{Y|X,G} = D_{Y|X}$ by Assumption 4.10.

Denote the inner expectation as $R(f, D_{Y|X=x}) := \mathbb{E}_{y\sim D_{Y|X}}[\ell(f(x), y)]$.

The Bayes optimal predictors minimize the inner expectation. Thus, they are also minimizers of the group-specific risk $\mathcal{F}_g$ for any $g \in [K]$.

$$\begin{aligned}
&\arg\min_f \mathbb{E}_{x\sim D_{X|G=g}}[R(f, D_{Y|X=x})] \\
&\geq \mathbb{E}_{x\sim D_{X|G}}[\arg\min_f R(f, D_{Y|X=x})] \\
&= \mathbb{E}_{x\sim D_{X|G}}[R(f^*, D_{Y|X=x})],
\end{aligned}$$

for any $f^* \in \mathcal{F}_{\text{ERM}}$. When the function class $\mathcal{F}$ used to find $\mathcal{F}_g$ is sufficiently large such that $f^* \in \mathcal{F}_g$, then $\mathcal{F}_{\text{ERM}} \subseteq \mathcal{F}_g$. Otherwise, ERM achieves better risk than $\mathcal{F}_g$.

Similarly rewriting the minimax risk using Assumption 4.10,

$$\begin{aligned}
&\max_a \mathbb{E}_{(x,y)\sim D_{Y,X|G=g}}[\ell(f(x), y)] \\
&\qquad = \max_g \mathbb{E}_{x\sim D_{X|G}}[R(f, D_{Y|X=x})].
\end{aligned}$$

By the max-min inequality, the Bayes optimal predictors minimize the worst group-specific risk,

$$\begin{aligned}
&\min_f \max_g \mathbb{E}_{x\sim D_{X|G}}[R(f, D_{Y|X=x})] \\
&\geq \max_g \min_f \mathbb{E}_{x\sim D_{X|G}}[R(f, D_{Y|X=x})] \\
&\geq \max_g \mathbb{E}_{x\sim D_{X|G}}[R(f^*, D_{Y|X=x})] \text{ for any } f^* \in \mathcal{F}_{\text{ERM}}.
\end{aligned}$$

When the function class $\mathcal{F}$ used to find $\mathcal{F}_{\text{MM}}$ is sufficiently large such that $f^* \in \mathcal{F}_{\text{MM}}$, then $\mathcal{F}_{\text{ERM}} \subseteq \mathcal{F}_{\text{MM}}$. Otherwise, ERM achieves better risk than $\mathcal{F}_{\text{MM}}$. □

E. Experiments

We first motivate the experimentation framework used in our study. Then, we provide the sources of the datasets used and the hyperparameter settings.

E.1. Motivation for hypothesis testing

The purpose of the experiments is to test how well do the theoretical results generalize in practical scenarios when assumptions may not necessarily hold. Such relevant scenarios are not amenable to direct theoretical analysis, thus, we use the hypothesis testing framework to gather evidence where deviations can occur. For practical scenarios not covered by the theory, we are essentially following the experimentation framework used in the physical sciences. Propose a law, match its consequences with the real world, falsify the law if contradictory otherwise assume that the law is true. Through the hypothesis tests for our experiments, we fail to falsify Theorem 4.4 for at least MLP and Random Forest classes, since we observe the behavior predicted by the theorem. Had we failed to see the observed behavior this would have provided direct evidence that assumptions did not hold. However, we note that the negation is not necessarily the case, and a consistent outcome does not guarantee that the assumptions hold. Thus, verifying the assumption made in Condition (C2) is an important direction for further work.

E.2. Sources for datasets

1. ACS Income, ACS Employment, ACS Health Insurance (Ding et al., 2021). Accessed using `folktables` package https://github.com/zykls/folktables from https://www.census.gov/programs-surveys/acs.
2. UCI Adult Income (Kohavi & Becker; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/adult. We follow the preprocessing steps given in https://auto.gluon.ai/stable/tutorials/tabular_prediction/tabular-custom-model.html.
3. COMPAS (ProPublica, 2020). Accessed from propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

4. Diabetes (Strack et al., 2014; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.

5. Drug Consumption (Fehrman et al., 2015; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/Drug%20consumption+(quantified).

6. eICU (Pollard et al., 2018). Accessed from https://physionet.org/content/eicu-crd/2.0/. Dataset preprocessing is the same as done in Singh et al. (2022).

7. Default (Yeh & hui Lien, 2009; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

8. Communities and Crime (Redmond & Baveja, 2002; red, 1990; 1992a;b; 1995; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/communities+and+crime.

9. German Credit (Hofmann; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

10. Heart (Chicco & Jurman, 2020). Accessed from https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data.

11. Marketing (Moro et al., 2014; Dua & Graff, 2017). Accessed from https://archive.ics.uci.edu/ml/datasets/bank+marketing

E.3. Hyperparameter settings

All model implementations, except for MLPs, are from `scikit-learn`. MLP is implemented using code from `rtdl` package[4].

Table 5: Hyperparameters used for the model types. Unless specified we use the default settings in `scikit-learn`.

Model type	Settings
Logistic regression, `SGDClassifier`	loss=log_loss, penalty=none
Random Forest, `RandomForestClassifier`	criterion=log_loss
Decision tree, `DecisionTreeClassifier`	criterion=log_loss, max_depth$\in \{2, 4, 8\}$
Linear SVC, `CalibratedClassifierCV(LinearSVC(),2)`	default values
MLP, `MLP` in Gorishniy et al. (2021)	hidden layer sizes=[1024,1024], dropout=0.1, lr=0.001, batch_size=2048, `AdamW` optimizer in PyTorch (Paszke et al., 2019), ERM iterations=2000

Gradient boosting machines did not work with the minimax solver, possibly due to numerical instabilities in the way sample weights are handled. We found that the solver stopped after less than 10 iterations as the sample weights did not change significantly. Therefore, we did not include results for boosting machines although this hypothesis class is likely to satisfy the sufficiency condition.

E.4. Additional results to validate Theorem 4.4

Figure 3 compares ERM with group-optimal or minimax-fair on test set. Table 6 shows the results of hypothesis tests for the accuracy metric instead of the negative logloss metric used for Table 2 in the main text.

[4] https://github.com/Yura52/rtdl

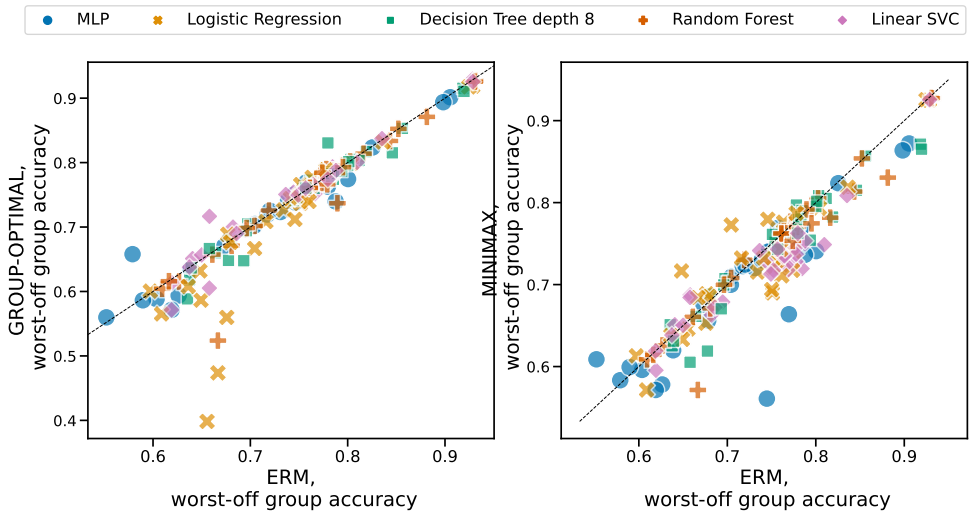

Figure 3: Q1. **ERM ≈ group-optimal or minimax-fair on test (visually).** Worst-case accuracy of ERM vs group-optimal (top) and minimax models (bottom) on test set. Each point on the plot corresponds to a model type trained on a dataset for one of the three methods. Dotted line is $y = x$. Visually, ERM matches the worst-case performance for group-optimal and minimax models for most classes.

Table 6: Q1. **ERM ≈ minimax-fair or group-optimal on train.** p-values from the two one-sided test for equivalence or non-inferiority of ERM against group-optimal and minimax-fair models. We show p-values for negative logloss from results on 36 datasets. Significance level is taken as p-value<0.05, highlighted in bold. We reject the hypothesis that ERM is not equivalent to Group optimal for random forest, MLP, and linear SVC models. For the same models, we reject the hypothesis that ERM is inferior to minimax.

(a) Metric = accuracy, threshold $t = 0.01$

Model type	ERM v/s GROUP-OPTIMAL p-value equivalence	ERM v/s GROUP-OPTIMAL p-value non-inferior	ERM v/s MINIMAX p-value equivalence	ERM v/s MINIMAX p-value non-inferior
Logistic Regression	0.4732	**0.0154**	0.0853	0.0706
Linear SVC	**0.0272**	**0.0173**	0.9992	**0.0000**
Random Forest	**0.0000**	**0.0000**	0.9204	**0.0003**
MLP	**0.0000**	**0.0000**	0.9320	**0.0083**

E.5. Additional results to validate Theorem 4.9

An ideal experimental design for Theorem 4.9 would control for different levels of k-recoverability (Definition (4.7)) across datasets by performing evaluation only on the successfully recovered points from a group. Note that a set of points recovered could be found by training a classifier per group and adjusting the thresholds until the precision is 1. As a more straightforward proxy, we use the same design as used to test the fully-recoverable case in Tables 2 and 6. This is reasonable since we observe high precision for group classifiers on the train set. Thus, we use the same evaluation metrics and training setup except we remove the group attribute from the list of features. Results are in the Tables 7 and 8 for train and test set accuracy, respectively. To reduce computation time, we ran experiments for only 23 datasets for MLP (reducing the iterations of minimax-fair learning to 100, results in the main text are for 200) and 33 datasets for other hypothesis classes.

We find that, in the case of train accuracy, we can reject the hypothesis that ERM is inferior to minimax-fair models for both MLP and random forest classes. However, we cannot reject any hypothesis for group-optimal models for these classes. When comparing test accuracy, we can reject inferior and non-equivalent hypotheses for random forest for minimax-fair models. For MLP class, we can reject inferior hypotheses for both types of models. This is an intriguing observation that results seem to hold better on the test set than on the train set, which we see in Table 4 as well. A generalization analysis will help to study this observation.

Table 7: **ERM $\approx$ minimax-fair on train set when group information is not given to models.** p-values from the two one-sided test for equivalence or non-inferiority of ERM against minimax-fair models' risk.

(a) Metric = accuracy, threshold $t = 0.01$

Model type	ERM v/s GROUP-OPTIMAL p-value equivalence	ERM v/s GROUP-OPTIMAL p-value non-inferior	ERM v/s MINIMAX p-value equivalence	ERM v/s MINIMAX p-value non-inferior
Logistic Regression	0.2431	**0.0235**	0.2793	**0.0014**
Linear SVC	0.1453	0.1453	0.9977	**0.0000**
Random Forest	0.3265	0.3265	0.8313	**0.0038**
MLP	0.7957	0.7957	0.9664	**0.0090**

Table 8: **ERM $\approx$ minimax-fair on test set when group information is not given to the models.** p-values from the two one-sided test for equivalence or non-inferiority of ERM against minimax-fair models' risk.

(a) Metric = accuracy, threshold $t = 0.01$

Model type	ERM v/s GROUP-OPTIMAL p-value equivalence	ERM v/s GROUP-OPTIMAL p-value non-inferior	ERM v/s MINIMAX p-value equivalence	ERM v/s MINIMAX p-value non-inferior
Logistic Regression	0.7829	**0.0007**	0.6008	0.0001
Linear SVC	**0.0095**	**0.0095**	0.9797	**0.0000**
Random Forest	0.0651	**0.0008**	**0.0061**	**0.0032**
MLP	0.1400	**0.0094**	0.8484	**0.0049**