Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data

Esther Rolf [1] **Theodora Worledge** [1] **Benjamin Recht** [1] **Michael I. Jordan** [1 2]

Abstract

Collecting more diverse and representative training data is often touted as a remedy for the disparate performance of machine learning predictors across subpopulations. However, a precise framework for understanding how dataset properties like diversity affect learning outcomes is largely lacking. By casting data collection as part of the learning process, we demonstrate that diverse representation in training data is key not only to increasing subgroup performances, but also to achieving population-level objectives. Our analysis and experiments describe how dataset compositions influence performance and provide constructive results for using trends in existing data, alongside domain knowledge, to help guide intentional, objective-aware dataset design.

1. Introduction

Datasets play a critical role in shaping the perception of performance and progress in machine learning (ML)—the way we collect, process, and analyze data affects the way we benchmark success and form new research agendas (Paullada et al., 2020; Dotan & Milli, 2020). A growing appreciation of this determinative role of datasets has sparked a concomitant concern that standard datasets used for training and evaluating ML models lack diversity along significant dimensions, for example, geography, gender, and skin type (Shankar et al., 2017; Buolamwini & Gebru, 2018). Lack of diversity in evaluation data can obfuscate disparate performance when evaluating based on aggregate accuracy (Buolamwini & Gebru, 2018). Lack of diversity in training data can limit the extent to which learned models can adequately apply to all portions of a population, a concern highlighted in recent work in the medical domain (Habib et al., 2019; Hofmanninger et al., 2020).

[1]Department of EECS, University of California, Berkeley [2]Department of Statistics, University of California, Berkeley. Correspondence to: Esther Rolf <esther_rolf@berkeley.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Our work aims to develop a general unifying perspective on the way that dataset composition affects outcomes of machine learning systems. We focus on *dataset allocations*: the number of datapoints from predefined subsets of the population. While we acknowledge that numerical inclusion of groups is an imperfect proxy of representation, we believe that allocations provide a useful initial mathematical abstraction for formulating relationships among diversity, data collection, and statistical risk. We discuss broader implications of our formulation in Section 5.

With the implicit assumption that the learning task is well specified and performance evaluation from data is meaningful for all groups, we ask:

Are group allocations in training data pivotal to performance? To what extent can up-weighting underrepresented groups help, and when might it actually hurt performance?

Taking a point of view that data collection is a critical component of the overall machine learning process, we study the effect that dataset composition has on group and population accuracies. This complements work showing that simply gathering more data can mitigate some sources of bias or unfairness in ML outcomes (Chen et al., 2018), a phenomenon which has been observed in practice as well. Indeed, in response to the Gender Shades study (Buolamwini & Gebru, 2018), companies selectively collected additional data to decrease the exposed inaccuracies of their facial recognition models for certain groups, often raising aggregate accuracy in the process (Raji & Buolamwini, 2019). Given the potential for targeted data collection efforts to repair unintended outcomes of ML systems, we next ask:

How can we describe "optimal" data allocations for different learning objectives? Does a lack of diversity in large-scale datasets align with maximizing population accuracy?

We show that purposeful data collection efforts can proactively support intentional objectives of an ML system, and that diversity and population objectives are often aligned. Many datasets have recently been designed or amended to exhibit diversity of the underlying population (Ryu et al., 2017; Tschandl et al., 2018; Yang et al., 2020). These are significant undertakings, as data gathering and annotation must consider consent, privacy, and power concerns in ad-

dition to inclusivity, transparency and reusability (Gebru et al., 2018; Gebru, 2020; Wilkinson et al., 2016). Given the importance of more representative and diverse datasets, and the effort required to create them, our final question asks:

When and how can we leverage existing datasets to help inform better allocations, towards achieving a diverse set of objectives in a subsequent dataset collection effort?

Representation bias, or systematic underrepresentation of subpopulations in data, is one of many forms of bias in ML (Suresh & Guttag, 2019). Our work provides a data-focused perspective on the design and evaluation of ML pipelines. Our main contributions are:

1. We analyze the complementary roles of dataset allocation and algorithmic interventions for achieving per-group and total-population performance (Section 2). Our experiments show that while algorithmically upweighting underrepresented groups can help, dataset composition is the most consistent determinant of performance (Section 4.1).

2. We propose a scaling model that describes the impact of dataset allocations on group accuracies (Section 3). Under this model, when parameters governing the relative values of within-group data are equal for all groups, the allocation that minimizes *population* risk *overrepresents* minority groups.

3. We demonstrate that our proposed scaling model captures major trends of the relationship between dataset allocations and performance (Sections 4.2 and 4.4). We evidence that a small initial sample can be used to inform subsequent data collection efforts to, for example, maximize the minimum accuracy over groups without sacrificing population accuracy (Section 4.3).

Sections 2 and 3 formalize data collection as part of the learning problem and derive results under illustrative settings. Experiments in Section 4 support these results and expose nuances inherent to real-data contexts. Section 5 synthesizes results and delineates future work.

1.1. Additional Related Work

Targeted data collection in ML. Recent research evidences that targeted data collection can be an effective way to reduce disparate performance of ML models evaluated across sub-populations (Raji & Buolamwini, 2019). Chen et al. (2018) present a formal argument that the addition of training data can lessen discriminatory outcomes while improving accuracy of learned models, and Abernethy et al. (2020) show that adaptively collecting data from the lowest-performing sub-population can increase the minimum accuracy over groups. It is important to note, however, there are many complications associated with simply gathering more data as a solution to disparate performance across groups (Jacobs & Wallach, 2019; Paullada et al., 2020).

With these complexities in mind, we study the importance of numerical representation in training datasets in achieving diverse objectives. Optimal allocation of subpopulations in statistical survey designs dates back to at least Neyman (1934), including stratified sampling methods to ensure coverage across sub-populations (Lohr, 2009). For more complex prediction systems, the field of optimal experimental design (Pukelsheim, 2006) studies what inputs are most valuable for reaching a given objective, often focusing on linear prediction functions. We consider a constrained sampling structure and directly model the impact of group allocations on subgroup performance for general model classes.

Valuing data. In economics, allocations indicate a division of goods to various entities (Cole et al., 2013). While we focus on the influence of data allocations on model accuracies across groups, there are many approaches to valuing data. Methods centering on a theory of Shapley valuations (Yona et al., 2019; Ghorbani & Zou, 2019) complement studies of the influence of individual data points on model performance to aid subsampling data (Vodrahalli et al., 2018).

Handling group-imbalanced data. Importance sampling and importance weighting are standard approaches to addressing class imbalance or small groups sizes (Haixiang et al., 2017; Buda et al., 2018), though the effects of importance weighting for deep learning may vary with regularization (Byrd & Lipton, 2019). Other methods specifically address differential performance between groups. Maximizing minimum performance across groups can reduce accuracy disparities (Sagawa et al., 2020) and promote fairer sequential outcomes (Hashimoto et al., 2018). For broader classes of group-aware objectives, techniques exist to mitigate unfairness or disparate performance of black box prediction functions (Dwork et al., 2018; Kim et al., 2019). It might not be clear a priori which subsets need attention; Sohoni et al. (2020) propose a method to identify and account for hidden strata, while other methods are defined for any subsets (Hashimoto et al., 2018; Kim et al., 2019). One can also downsample or augment the input data to match a desired distribution (Chawla et al., 2002; Iosifidis & Ntoutsi, 2018).

Notation. Δ^k denotes the k-dimensional simplex. $\mathbb{Z}^+$ denotes non-negative integers and $\mathbb{R}^+$ non-negative reals.

2. Training Set Allocations and Alternatives

We study settings in which each data instance is associated with a group g_i, so that the training set can be expressed as $\mathcal{S} = \{x_i, y_i, g_i\}_{i=1}^n$ where x_i, y_i denote the features and labels of each instance. We index the discrete **groups** by integers $\mathcal{G} = \{1, .., |\mathcal{G}|\}$, or when we specifically consider

just two groups, we write $\mathcal{G} = \{A, B\}$. We assume that groups are disjoint and cover the entire population, with $\gamma_g = P_{(X,Y,G)\sim\mathcal{D}}[G = g]$ denoting the **population prevalence** of group g, so that $\vec{\gamma} \in \Delta^{|\mathcal{G}|}$. Groups could represent inclusion in one of many binned demographic categories, or simply a general association with latent characteristics that are relevant to prediction.

For a given population with distribution $\mathcal{D}$ over features, labels, and groups, we are interested in the population level risk, $\mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}) := \mathbb{E}_{(X,Y,G)\sim\mathcal{D}}[\ell(\hat{f}(X), Y)]$, of a predictor $\hat{f}$ trained on dataset $\mathcal{S}$, as well as group specific risks. Denoting the **group distributions** by $\mathcal{D}_g$, defined as conditional distributions, via $P_{(X,Y)\sim\mathcal{D}_g}[X = x, Y = y] = P_{(X,Y,G)\sim\mathcal{D}}[X = x, Y = y, G = g]/\gamma_g$, the population risk decomposes as a weighted average over group risks:

$$\mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}) = \sum_{g\in\mathcal{G}} \gamma_g \cdot \mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}_g)\,. \tag{1}$$

In Section 2.2 we will assume that the loss $\ell(\hat{y}, y)$ is a separable function over data instances. While this holds for many common loss functions, some objectives do not decouple in this sense (e.g., group losses and associated classes of fairness-constrained objectives; see Dwork et al., 2018). We revisit this point in Sections 4 and 5.

2.1. Training Set Allocations

In light of the decomposition of the population-level risk as a weighted average over group risks in Eq. (1), we now consider the composition of fixed-size training sets, in terms of how many samples come from each group.

Definition 1 (Allocations)**.** Given a dataset of n triplets, $\{x_i, y_i, g_i\}_{i=1}^n$, the **allocation** $\vec{\alpha} \in \Delta^{|\mathcal{G}|}$ describes the relative proportions of each group in the dataset:

$$\alpha_g := \tfrac{1}{n}\sum_{i=1}^{n} \mathbb{I}[g_i = g], \quad g \in \mathcal{G}. \tag{2}$$

It will be illuminating to consider $\vec{\alpha}$ not only as a property of an existing dataset, but as a parameter governing dataset construction, as captured in the following definition.

Definition 2 (Sampling from allocation $\vec{\alpha}$)**.** Given the sample size n, group distributions $\{\mathcal{D}_g\}_{g\in\mathcal{G}}$, and allocation $\vec{\alpha} \in \Delta^{|\mathcal{G}|}$, such that $n_g := \alpha_g n \in \mathbb{Z}^+, \forall g \in \mathcal{G}$, to **sample from allocation** $\vec{\alpha}$ is procedurally equivalent to independent sampling of $|\mathcal{G}|$ disjoint datasets $\mathcal{S}_g$ and concatenating:

$$\mathcal{S}(\vec{\alpha}, n) = \bigcup_{g\in\mathcal{G}} \mathcal{S}_g \tag{3}$$
$$\mathcal{S}_g = \{x_i, y_i, g\}_{i=1}^{n_g}, \quad (x_i, y_i) \sim_{i.i.d.} \mathcal{D}_g\,.$$

In the following sections we will generally allow allocations with $n_g \notin \mathbb{Z}$, assuming that the effect of up to $|\mathcal{G}|$ fractionally assigned instances is negligible for large n.

The procedure in Definition 2 suggests formalizing data collection as a component of the learning process in the following way: in addition to choosing a loss function and method for minimizing the risk, choose the relative proportions at which to sample the groups in the training set:

$$\vec{\alpha}^* = \operatorname*{argmin}_{\vec{\alpha}\in\Delta^{|\mathcal{G}|}} \min_{\hat{f}\in\mathcal{F}} \mathcal{R}\left(\hat{f}\left(\mathcal{S}(\vec{\alpha}, n)\right); \mathcal{D}\right).$$

In Section 3, we show that when a dataset curator can design dataset allocations in the sense of Definition 2, they have the opportunity to improve accuracy of the trained model. Section 2.2 considers methods for using fixed datasets that have groups with small training set allocation α_g, relative to γ_g, or high risk for some groups relative to the population.

2.2. Accounting for Small Group Allocations

In classical **empirical risk minimization** (ERM), one learns a function from class $\mathcal{F}$ that minimizes average prediction loss over the training instances $(x_i, y_i, g_i) \in \mathcal{S}$ (we also abuse notation and write $i \in \mathcal{S}$) with optional regularization R:

$$\hat{f}(\mathcal{S}) = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i\in\mathcal{S}} \ell(f(x_i), y_i) + R(f, \mathcal{S}).$$

There are many methods for addressing small group allocations in data (see Section 1.1). Of particular relevance to our work are objective functions that minimize group or population risks. In particular, one approach is to use **importance weighting** (IW) to re-weight training samples with respect to a target distribution defined by $\vec{\gamma}$:

$$\hat{f}^{\text{IW}}(\mathcal{S}) = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{g\in\mathcal{G}} \frac{\gamma_g}{\alpha_g}\Big(\sum_{i\in\mathcal{S}_g} \ell(f(x_i), y_i)\Big) + R(f, \mathcal{S}).$$

This empirical risk with instances weighted by $\gamma_g/\alpha_g = \gamma_g n/n_g$ is an unbiased estimate of the population risk, up to regularization. While unbiasedness is often desirable, IW can induce high variance of the estimator when γ_g/α_g is large for some group (Cortes et al., 2010), which happens when group g is severely underrepresented in the training data relative to their population prevalence.

Alternatively, **group distributionally robust optimization** (GDRO) (Hu et al., 2018; Sagawa et al., 2020) minimizes the maximum empirical risk over all groups:

$$\hat{f}^{\text{GDRO}}(\mathcal{S}) = \operatorname*{argmin}_{f\in\mathcal{F}} \max_{g\in\mathcal{G}}\Big(\tfrac{1}{n_g}\sum_{i\in\mathcal{S}_g} \ell(f(x_i), y_i) + R(f, \mathcal{S}_g)\Big).$$

For losses ℓ which are continuous and convex in the parameters of f, the optimal GDRO solution corresponds to the minimizer of a group-weighted objective: $\frac{1}{n}\sum_{i=1}^n w(g_i) \cdot \ell(f(x_i), y_i)$, though this is not in general true for nonconvex losses (see Prop. 1 of Sagawa et al., 2020, and the remark immediately thereafter).

Given the correspondence of GDRO (for convex loss functions) and IW to the optimization of group-weighted ERM objectives, we now investigate the joint roles of sample allocation and group re-weighting for estimating group-weighted risks. For prediction function f, loss function ℓ, and group weights $w : \mathcal{G} \to \mathbb{R}^+$, let $\hat{L}(w, \alpha, n; f, \ell)$ be the random variable defined by:

$$\hat{L}(w, \alpha, n; f, \ell) := \frac{1}{n} \sum_{i \in \mathcal{S}(\vec{\alpha}, n)} w(g_i) \cdot \ell(f(x_i), y_i) \ ,$$

where the randomness in $\hat{L}$ comes from the draws of x_i, y_i from $\mathcal{D}_{g_i}$ according to procedure $\mathcal{S}(\vec{\alpha}, n)$ (Definition 2), as well as any randomness in f.

The following proposition shows that group weights and allocations play complementary roles in risk function estimation. In particular, if $w(g)$ depends on the sampling allocations α_g, then there are alternative group weights w^* and allocation $\vec{\alpha}^*$ such that the alternative estimator has the same expected value but lower variance.

Proposition 1 (Weights and allocations)**.** *For any loss ℓ, prediction function f and group distributions $\mathcal{D}_g$, there exist weights with $w^*(g) \propto \left(Var_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y))]\right)^{-1/2}$ such that for any triplet $(\vec{\alpha}, w, n)$ with $\sum_g \alpha_g w(g) > 0$, if $w \not\propto w^*$,[1] there exists an alternative $\vec{\alpha}^*$ with*

$$\mathbb{E}[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] = \mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)]$$
$$Var[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] < Var[\hat{L}(w, \vec{\alpha}, n; f, \ell)] \ .$$

If $w(g) > w^*(g)$, $\alpha^*_g > \alpha_g$ and if $w(g) < w^*(g)$, $\alpha^*_g < \alpha_g$.

Proof. (Sketch; full proof appears in Appendix A.1). For any deterministic weighting function $w : \mathcal{G} \to \mathbb{R}^+$, there exists a vector $\vec{\gamma}' \in \Delta^{|\mathcal{G}|}$ with $\gamma'_g \propto w(g)\alpha_g$ such that

$$\mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)] = c \cdot \mathbb{E}_g \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y)],$$

where $g \sim$ Multinomial(γ') and $c = \sum_g \alpha_g w(g)$. For any fixed f and any "target distribution" defined by γ', the $(\vec{\alpha}^*, w^*)$ pair which minimizes the variance of the estimator, constrained so that $w(g)\alpha_g = c\gamma'_g \ \forall g$, has weights w^* with form given above. Since the original (α, w) pair satisfies this constraint, the pair (α^*, w^*) must satisfy $Var[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] \leq Var[\hat{L}(w, \vec{\alpha}, n; f, \ell)]$, while the constraint ensures that $\mathbb{E}[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] = \mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)]$. □

Since the estimation of risk functions is a key component of learning, Proposition 1 illuminates an interplay between the roles of sampling allocations and group-weighting schemes like IW and GDRO. When allocations and weights are jointly maximized, the optimal allocation accounts for an implicit target distribution γ' (defined above), which may vary by objective function. The optimal weights account for per-group variability $Var_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y))]$. In Section 4 we find that it can be advantageous to use IW and GDRO when some groups have small α_g/γ_g; though the boost in accuracy is less than having an optimally allocated training set to begin with, and diminishes when all groups are appropriately represented in the training set allocation.

[1] We use the symbol $\not\propto$ to denote "not approximately proportional to." The approximately part of this relation stems from finite and integer sample concerns; for example, the proposition holds if we consider $w \not\propto w^*$ to mean $\exists g \in \mathcal{G} : |1 - \frac{w(g)}{w^*(g)}| > \frac{|\mathcal{G}|}{\alpha_g n}$.

3. Allocating Samples to Minimize Risk

Having motivated the importance of group allocations, we now investigate the direct effects of training set allocations on group and population risks. Using a model of per-group performance as a function of allocations, we study the optimal allocations under a variety of settings.

3.1. A Per-group Power-law Scaling Model

We model the impact of allocations on performance with scaling laws that describe per-group risks as a function of the number of data points from their respective group, as well as the total number of training instances.

Assumption 1 (Group risk scaling with allocation)**.** *The group risks $\mathcal{R}(\hat{f}; \mathcal{D}_g) := \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(\hat{f}(x), y)]$ scale approximately as the sum of inverse power functions on the number of samples from group g and the total number of samples. That is, $\exists\ M_g > 0$, $\sigma_g, \tau_g, \delta_g \geq 0$, and $p, q > 0$ such that for a learning procedure which returns predictor $\hat{f}(\mathcal{S})$, and training set $\mathcal{S}$ with group sizes $n_g \geq M_g$:*

$$\mathcal{R}\big(\hat{f}\big(\mathcal{S}(\vec{\alpha}, n)\big); \mathcal{D}_g\big) \approx r(\alpha_g n, n; \sigma_g, \tau_g, \delta_g, p, q) \quad \forall\, g \in \mathcal{G}$$
$$r(n_g, n; \sigma_g, \tau_g, \delta_g, p, q) := \sigma_g^2 n_g^{-p} + \tau_g^2 n^{-q} + \delta_g \ . \qquad (4)$$

Assumption 1 is similar to the scaling law in Chen et al. (2018), but includes a $\tau_g^2 n^{-q}$ term to allow for data from other groups to influence the risk evaluated on group g. It additionally requires that the same exponents p, q apply to each group, an assumption that underpins our theoretical results in Section 3. We examine the extent to which Assumption 1 holds empirically in Section 4.2, and will modify Eq. (4) to include group-dependent terms p_g, q_g when appropriate. The following examples give intuition into the form of Eq. (4).

Example 1. When separate models are trained for each group, using training data only from that group, we expect Eq. (4) to apply with $\tau_g = 0\ \forall g \in \mathcal{G}$. The parameter p could derived through generalization bounds (Boucheron et al., 2005), or through modeling assumptions (Example 3). ◊

Example 2. When groups are irrelevant for prediction and

the model class $\mathcal{F}$ correctly accounts for this, we expect Eq. (4) to apply with $\sigma_g = 0\ \forall g \in \mathcal{G}$. $\Diamond$

Example 3. Consider a $(d+1)$-dimensional linear model, where two groups, $\{A, B\}$, share a weight vector β and features $x \sim \mathcal{N}(0, \Sigma_x)$, but the intercept varies by group:

$$y_i = \beta^\top x_i + c_A \mathbb{I}[g_i = A] + c_B \mathbb{I}[g_i = B] + \mathcal{N}(0, \sigma^2).$$

As we show in Appendix A.5, the ordinary least squares predictor has group risks $\mathbb{E}_{(x,y)\sim\mathcal{D}_g}[(x^\top\hat{\beta} + \hat{c}_g - y)^2] = \sigma^2\left(1 + 1/n_g + O(d/n)\right)$, where the $1/n_g$ arises because we need samples from group g to estimate the intercept c_g, whereas samples from both groups help us estimate β. $\Diamond$

Example 3 suggests that in some settings, we can relate σ_g and τ_g to 'group specific' and 'group agnostic' model components that affect performance for group g. In general, the relationship between group sizes and group risks can be more nuanced. Data from different groups may be correlated, so that samples from groups similar to or different from g have greater effect on $\mathcal{R}(\hat{f}; \mathcal{D}_g)$ (see Section 4.4). Eq. 4 is meant to capture the dominant effects of training set allocations on group risks and serves as our main structural assumption in the next section, where we study the allocation that minimizes the approximate *population risk*.

3.2. Optimal (w.r.t. Population Risk) Allocations

We now study properties of the allocation that minimizes the approximated population risk:

$$\begin{aligned}\hat{\mathcal{R}}(\vec{\alpha}, n) &:= \sum_{g\in\mathcal{G}} \gamma_g r(\alpha_g n, n; \sigma_g, \tau_g, \delta_g, p, q) \qquad (5)\\ &\approx \sum_{g\in\mathcal{G}} \gamma_g \mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}_g) = \mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}).\end{aligned}$$

The following proposition lays the foundation for two corollaries which show that: (1) when only the population prevalences $\vec{\gamma}$ vary between groups, the allocation that minimizes the approximate *population risk* up-represents groups with small γ_g; (2) for two groups with different scaling parameters σ_g, the optimal allocation of the group with $\gamma_g < \frac{1}{2}$ is bounded by functions of σ_A, σ_B, and $\vec{\gamma}$.

Proposition 2. *Given a population made up of disjoint groups $g \in \mathcal{G}$ with population prevalences γ_g, under the conditions of Assumption 1, the allocation $\vec{\alpha}^* \in \Delta^{|\mathcal{G}|}$ that minimizes the approximated population risk $\hat{\mathcal{R}}$ in eq. (5) has elements:*

$$\alpha_g^* = \frac{\left(\gamma_g \sigma_g^2\right)^{1/(p+1)}}{\sum_{g\in\mathcal{G}}\left(\gamma_g\sigma_g^2\right)^{1/(p+1)}}. \qquad (6)$$

If $\sigma_g = 0\ \forall g \in \mathcal{G}$, then any allocation in $\Delta^{|\mathcal{G}|}$ minimizes $\hat{\mathcal{R}}$.

The proof of Proposition 2 appears in Appendix A.2. Note that $\vec{\alpha}^*$ does not depend on n, $\{\tau_g\}_{g\in\mathcal{G}}$, or q; this will in general not hold if powers p_g differ by group.

We now study the form of $\vec{\alpha}^*$ under illustrative settings. Corollary 1 shows that when the group scaling parameters σ_g in Eq. (4) are equal across groups, the allocation that minimizes the approximate *population risk* allocates samples to minority groups at higher than their *population prevalences*. The proof of Corollary 1 appears in Appendix A.3.

Corollary 1 (Many groups with equal σ_g)**.** *When $\sigma_g = \sigma > 0,\ \forall g \in \mathcal{G}$, the allocation that minimizes $\hat{\mathcal{R}}$ in Eq. (5) satisfies $\alpha_g^* \geq \gamma_g$ for any group with $\gamma_g \leq \frac{1}{|\mathcal{G}|}$.*

This shows that the allocation that minimizes population risk can differ from the actual population prevalences $\vec{\gamma}$. In fact, Corollary 1 asserts that near the allocation $\vec{\alpha} = \vec{\gamma}$, the marginal returns to additional data from group g are largest for groups with small α_g, enough so as to offset the small weight γ_g in Eq. (1). This result provides evidence *against* the idea that small training set allocation to minority groups might comply with minimizing population risk as a result of a small relative contribution to the population risk.

Remark. A counterexample shows that $\alpha_g^* \leq \gamma_g$ does not hold for all g with $\gamma_g > 1/|\mathcal{G}|$. Take $\vec{\gamma} = [.68, .30, .01, .01]$ and $p = 1$; Eq. (6) gives $\alpha_2^* > 0.3 = \gamma_2 > 1/4$. In general, whether group g with $\gamma_g \geq 1/|\mathcal{G}|$ gets up- or down-sampled depends on the distribution of $\vec{\gamma}$ across all groups.

Complementing the results of Corollary 1, the next corollary shows that $\vec{\alpha}^*$ generally depends on the relative values of σ_g between groups. Inspecting Eq. (4) shows that σ_g defines a limit of performance: if σ_g^2 is large, the only way to make the approximate risk for group g small is to make n_g large. From Eq. (6), we know that for two groups, α_A^* is increasing in $\frac{\sigma_A}{\sigma_B}$; Corollary 2 gives upper and lower bounds on α_A^* in terms of σ_A and σ_B. Corollary 2 is proved in Appendix A.4.

Corollary 2 (Unequal per-group constants)**.** *For two groups $\{A, B\} = \mathcal{G}$ with $\gamma_A < \gamma_B$, and parameters $\sigma_A, \sigma_B > 0$ in Eq. (4), the allocation of the smaller group α_A^* that minimizes $\hat{\mathcal{R}}$ in Eq. (5) is upper and lower bounded as*

$$\begin{aligned}\frac{\gamma_A(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + \gamma_B(\sigma_B^2)^{1/(p+1)}} &< \alpha_A^*\\ &< \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}}.\end{aligned}$$

When $\sigma_A \geq \sigma_B$, $\alpha_A^* > \gamma_A$, and when $\sigma_A \leq \sigma_B$, $\alpha_A^* < 1/2$.

Altogether, these results highlight key properties of training set allocations that minimize population risk. Experiments in Section 4 give further insight into the values of weights and allocations for minimizing group and population risks and apply the scaling law model in real data settings.

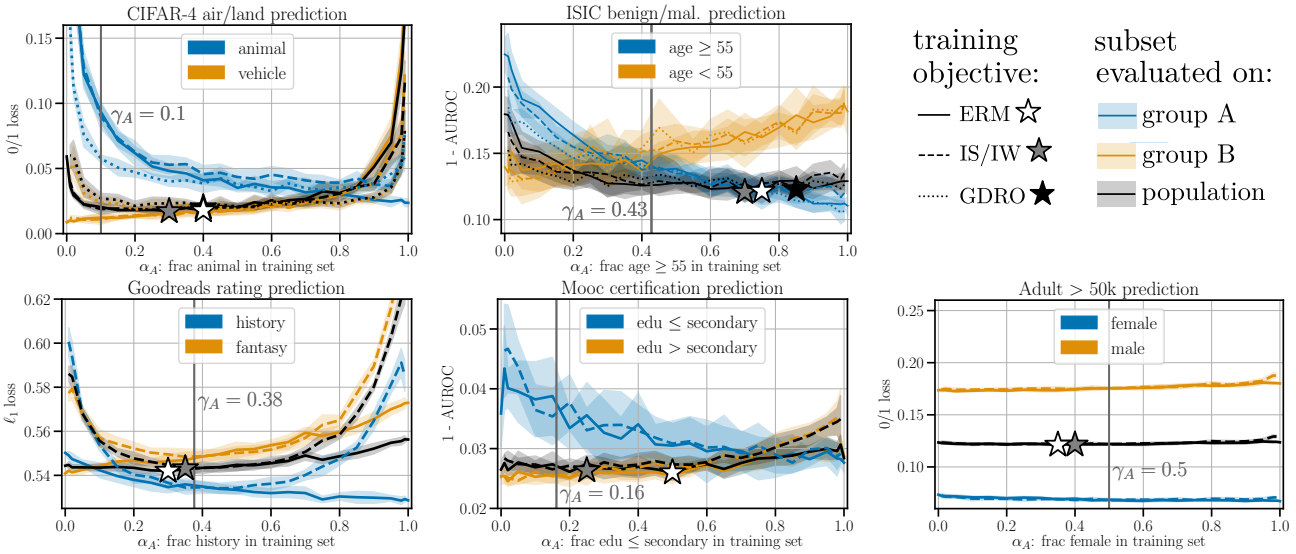


Figure 1. Performance across $\vec{\alpha}$. Shaded regions: one stddev. above/ below mean (10 trials). Stars: population minima for each objective. The loss metrics reported (vertical axes) are the same within panels, while the training objectives differ across solid and dashed lines.

4. Empirical Results

Having shown the importance of training set allocations from a theoretical perspective, we now provide a complementary empirical investigation of this phenomenon. See Appendix B for full details on each experimental setup.[2]

Table 1. Brief description of datasets; details in Appendix B.1.

dataset	groups $\{A, B\}$	γ_A	min n_g	target label	loss metric
CIFAR-4	{animal, vehicle}	0.1	10,000	air	0/1 loss
ISIC	{age $\geq$ 55, age $<$ 55}	0.43	4,092	malignant	1 - AUROC
Goodreads	{history, fantasy}	0.38	50,000	book rating	ℓ_1 loss
Mooc	{edu $\leq 2°$, edu $> 2°$}	0.16	3,897	certified	1 - AUROC
Adult	{female, male}	0.5	10,771	income $>$ 50K	0/1 loss

We use a wide range of datasets to give a full empirical characterization of the phenomena of interest (see Table 1). The CIFAR-4 dataset is comprised of bird, car, horse, and plane image instances from CIFAR-10 (Krizhevsky, 2009). The ISIC dataset contains images of skin lesions labelled as benign or malignant (Codella et al., 2019). The Goodreads dataset consists of written book reviews and numerical ratings (Wan & McAuley, 2018). The Mooc dataset contains student demographic and participation data (HarvardX, 2014). The Adult dataset consists of demographic data from the 1994 Census (Dua & Graff, 2017). For Adult we exclude groups from features (Appendix C.1).

In contrast to Section 2, here losses are defined over sets of data; note that AUROC is not separable over groups, and thus Eq. (1) does not apply for this metric.

[2]Code to replicate the experiments is available at https://github.com/estherrolf/representation-matters.

4.1. Allocation-aware Objectives vs. Ideal Allocations

We first investigate (a) the change in group and population performance at different training set allocations, and (b) the extent to which optimizing the three objective functions defined in Section 2.2 decreases average and group errors.

For each dataset, we vary the training set allocations $\vec{\alpha}$ while fixing the training set size as $n = \min_g n_g$ (see Table 1) and evaluate the per-group and population losses on subsets of the heldout test sets.[3] For the image classification tasks, we compare group-agnostic empirical risk minimization (ERM) to importance weighting (implemented via importance sampling (IS) batches following the findings of Buda et al. (2018)) and group distributionally robust optimization (GDRO) with group-dependent regularization as in Sagawa et al. (2020). For the non-image datasets, we implement importance weighting (IW) by weighting instances in the loss function during training, and do not compare to GDRO.[4]

Figure 1 highlights the importance of at least a minimal representation of each group in order to achieve low population loss (black curves) for all objectives. For CIFAR-4, the population loss increases sharply for $\alpha_A < 0.1$ and $\alpha_A > 0.8$, and for ISIC, when $\alpha_A < 0.2$. While not as crucial for achieving low population losses for the remaining datasets, the *optimal* allocations $\vec{\alpha}^*$ (stars) do require a minimal representation of each group. The $\vec{\alpha}^*$ are largely consistent across the training objectives (different star colors). The population

[3]We pick models and parameters via a cross-validation procedure over a coarse grid of $\vec{\alpha}$; details are given in Appendix B.3.

[4]The gradient-based algorithm of Sagawa et al. (2020) is not easily adaptable to the predictors we use for these datasets.

Table 2. Estimated scaling parameters for Eq. (7). Parentheses denote standard deviations estimated by the nonlinear least squares fit. Parameters are constrained so that $\hat{\tau}_g, \hat{\sigma}_g, \hat{\delta}_g \geq 0$ and $\hat{p}_g, \hat{q}_g \in [0, 2]$.

dataset	M_g	group g	$\hat{\sigma}_g$	$\hat{p}_g$	$\hat{\tau}_g$	$\hat{q}_g$	$\hat{\delta}_g$
CIFAR-4	500	animal	1.9 (0.12)	0.47 (9.8e-04)	4.5e-09 (1.8e+06)	2.0 (0.0e+00)	1.1e-03 (8.9e-06)
		vehicle	1.6 (0.19)	0.54 (2.0e-03)	3.2e-12 (1.1e+06)	2.0 (0.0e+00)	1.4e-03 (2.8e-06)
ISIC	200	age ≥ 55	0.61 (1.7e-03)	0.20 (1.1e-03)	1.7e-09 (1.9e+04)	1.9 (0.0e+00)	1.4e-15 (6.1e-04)
		age < 55	0.26 (9.3e-04)	0.13 (0.012)	0.61 (0.044)	0.3 (7.5e-03)	7.5e-11 (7.2e-03)
Goodreads	2500	history	0.16 (1.2e-03)	0.074 (2.5e-03)	2.5 (0.058)	0.37 (2.0e-04)	0.41 (3.0e-03)
		fantasy	0.62 (0.69)	0.020 (1.2e-03)	3.1 (0.093)	0.39 (1.9e-04)	7.2e-21 (0.72)
Mooc	50	edu $\leq 2^\circ$	0.08 (2.6e-05)	0.14 (6.0e-03)	0.73 (0.059)	0.63 (4.8e-03)	1.3e-15 (2.6e-04)
		edu $> 2^\circ$	0.038 (6.2e-04)	0.068 (6.3e-03)	0.54 (6.5e-03)	0.61 (9.8e-04)	2.8e-12 (8.0e-04)
Adult	50	female	0.078 (0.051)	0.018 (3.6e-03)	0.43 (8.3e-03)	0.59 (1.6e-03)	8.0e-16 (0.052)
		male	0.066 (2.6e-05)	0.21 (1.2e-03)	0.47 (6.5e-03)	0.50 (1.1e-03)	0.16 (5.4e-06)

losses (black curves) are largely consistent across mid-range values of α_A for all training objectives. This stands in contrast to the per-group losses (blue and orange curves), which can vary considerably as $\vec{\alpha}$ changes. From the perspective of model evaluation, this reinforces a well-documented need for more comprehensive reporting of performance. From the view of dataset design, this exposes an opportunity to choose allocations which optimize diverse evaluation objectives while maintaining low population loss. Experiments in Section 4.3 investigate this further.

Across the CIFAR-4 and ISIC tasks, GDRO (dotted curves) is more effective than IS (dashed curves) at reducing per-group losses. This is expected, as minimizing the largest loss of any group is the explicit objective of GDRO. Figure 1 shows that GDRO can also improve the *population loss* (see $\alpha_A > 0.7$ for CIFAR-4 and $\alpha_A < 0.2$ for ISIC). IW (dashed curves) has little effect on performance for Mooc and Adult (random forest models), and actually increases the loss for Goodreads (multiclass logistic regression model).

For all the datasets we study, the advantages of using IS or GDRO are greatest when one group has very small training set allocation (α_A near 0 or 1). When allocations are optimized (stars in Figure 1), the boost that these methods give over ERM diminishes. In light of Proposition 1, these results suggest that in practice, part of the value of such methods is in compensating for sub-optimal allocations. We find, however, that explicitly optimizing the maximum per-group loss with GDRO can reduce population loss more effectively than directly accounting for allocations with IS.

Appendix C.2 shows that similar phenomena hold for different loss functions and models on the same dataset, though the exact $\vec{\alpha}^*$ can differ. In Appendix C.1, we show that losses of groups with small α_g can degrade more severely when group attributes are included in the feature matrix, likely a result of the small number of samples from which to learn group-specific model components (see Example 3).

4.2. Assessing the Scaling Law Fits

For each dataset, we combine the results in Figure 1 with extra subsetting runs where we vary both n_g and n. We use nonlinear least squares to estimate the parameters of modified scaling laws, where exponents can differ by group

$$\text{loss}_g \approx \sigma_g^2 n_g^{-p_g} + \tau_g^2 n^{-q_g} + \delta_g \,. \tag{7}$$

The estimated parameters of Eq. (7) given in Table 2 capture different high-level phenomena across the five datasets. For CIFAR-4, $\hat{\tau}_g \approx 0$ for both groups, indicating that most of the group performance is explained by n_g. For Goodreads, both n_g and n have influence in the fitted model, though $\hat{\tau}_g$ and $\hat{q}_g$ are larger than $\hat{\sigma}_g$ and $\hat{p}_g$, respectively. For ISIC, $\hat{\tau}_A \approx 0$ but $\hat{\tau}_B \not\approx 0$, suggesting other-group data has little effect on the first group, but is beneficial to the latter. For the non-image datasets (Goodreads, Mooc, and Adult), $0 < \hat{\sigma}_g < \hat{\tau}_g$ and $\hat{p}_g < \hat{q}_g$ for all groups.

Figure B.2 in Appendix B.5 shows that the fitted curves capture the overall trends of per-group losses as a function of n and n_g. However, the assumptions of Proposition 2 and Corollaries 1 and 2 (e.g., equal p_g for all $g \in \mathcal{G}$) are not always reflected in the empirical fits. Results in Section 3 use Eq. (4) to describe optimal allocations under different hypothetical settings; we find that Eq. (7) is more realistic in empirical settings.

The estimated models describe the overall trends (Figure B.2), but the parameter estimates are variable (Table 2), indicating that a range of parameters can fit the data, a well-known phenomenon in fitting power laws to data (Clauset et al., 2009). While we caution against absolute or prescriptive interpretations based on the estimates given in Table 2, if such interpretations are desired (Chen et al., 2018), we suggest evaluating variation due to subsetting patterns and comparing to alternative models such as log-normal and exponential fits (cf. Clauset et al., 2009).

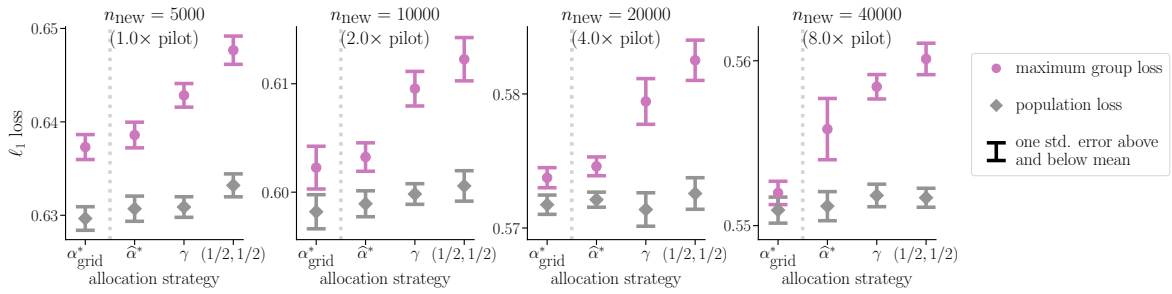


Figure 2. Pilot sample experiment. Panels show the result of the three allocations $\vec{\alpha} \in [\hat{\alpha}^*_{\text{minmax}}, \vec{\gamma}, (1/2, 1/2)]$ for different sizes of the new training sets compared with an α^*_{grid} baseline that minimizes the maximum group loss over a grid of resolution 0.01, averaged over the random trials. Purple circles indicate average maximum error over groups and grey diamonds indicate average population error. Ranges denote standard errors taken over the 10 trials.

4.3. Targeted Data Collection with Fitted Scaling Laws

We now study the use of scaling models fitted on a small pilot dataset to inform a subsequent data collection effort. Given the results of Section 4.1, we aim to collect a training set that minimizes the maximum loss on any group. This procedure goes beyond the descriptive use of the estimated scaling models in Section 4.2; important considerations for operationalizing these findings are discussed below.

We perform this experiment with the Goodreads dataset, the largest of the five we study. The pilot sample contains 2,500 instances from each group, drawn at random from the full training set. We estimate the parameters of Eq. (7) using a procedure similar to that described in Section 4.2. For a new training set of size n_{new}, we suggest an allocation to minimize the maximum forecasted loss of any group:

$$\hat{\alpha}^*_{\text{minmax}} = \underset{\vec{\alpha} \in \Delta^2}{\operatorname{argmin}} \max_{g \in \mathcal{G}} \left(\hat{\sigma}_g^2 (\alpha_g n_{\text{new}})^{-\hat{p}_g} + \hat{\tau}_g^2 n_{\text{new}}^{-\hat{q}_g} + \hat{\delta}_g \right).$$

For $n_{\text{new}} \in [2\times, 4\times, 8\times]$, the pilot sample size, we simulate collecting a new training set by drawing n_{new} fresh samples from the training set with allocation $\vec{\alpha} = \hat{\alpha}^*_{\text{minmax}}(n_{\text{new}})$. We train a model on this sample (ERM objective) and evaluate on the test set. For comparison, we also sample at $\vec{\alpha} = \vec{\gamma}$ (population proportions) and $\vec{\alpha} = (0.5, 0.5)$ (equal allocation to both groups). We repeat the experiment, starting with the random instantiation of the pilot dataset, for ten trials. As a point of comparison, we also compute the results for all α in a grid of resolution 0.01, and denote the allocation value in this grid that minimizes the average maximum group loss over the ten trials as α^*_{grid}.

Among the three allocation strategies we compare, $\hat{\alpha}^*_{\text{minmax}}$ minimizes the average maximum loss over groups, across n_{new} (Figure 2). In contrast, $\hat{\alpha}^*_{\text{minmax}}$ does not increase the population loss (grey bars) over that of the other allocation strategies. This reinforces the finding of Section 4.1 and provides evidence that we can leverage information from a small initial sample to help raise the minimum accuracy over groups, without sacrificing population accuracy.

While the results in Figure 2 are promising, error bars highlight the variation across trials. The variability in performance across trials for allocation baseline α^*_{grid} (which is kept constant across the ten trials) is largely consistent with that of the other allocation sampling strategies examined (standard errors in Figure 2). However, the estimation of $\hat{\alpha}^*$ in each trial does introduce additional variation: across the ten draws of the pilot data, the range of $\hat{\alpha}^*$ values for subsequent dataset size $n_{\text{new}} = 10000$ is [1e-04,0.05], for $n_{\text{new}} = 20000$ it is [5e-05,0.14], and for $n_{\text{new}} = 40000$ it is [2e-05,0.82]. Therefore, the estimated $\hat{\alpha}^*$ should be leveraged with caution, especially if the subsequent sample will be much larger than the pilot sample. Further caution should be taken if there may be distribution shifts between the pilot and subsequent samples. We suggest to interpret estimated $\hat{\alpha}^*$ values as one signal among many that can inform a dataset design in conjunction with current and emerging practices for ethical data collection (see Section 5).

4.4. Interactions Between Groups

We now shift the focus of our analysis to explore potential between- and within- group interactions that are more nuanced than the scaling law in Eq. (4) provides for. The results highlight the need for and encourage future work extending our analysis to more complex notions of groups (e.g., intersectional, continuous, or proxy groups).

As discussed in Section 3, data from groups similar to or different from group g may have greater effect on $\mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}_g)$ compared to data drawn at random from the entire distribution. We examine this possibility on the ISIC dataset, which is aggregated from different studies (Appendix B.1). We measure baseline performance of the model trained on data from all of the studies. We then remove one study at a time from the training set, retrain the model, and evaluate the change in performance for all studies in the test set.

Figure 3 shows the changes in performance due to leaving out studies from the training set. Rows correspond to the

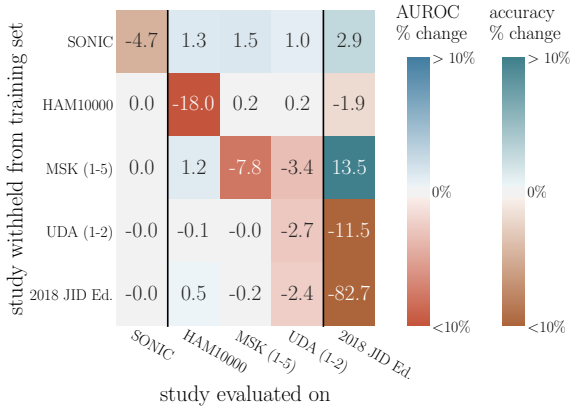


Figure 3. Percent change in performance (AUROC / accuracy) due to withholding a study from the training set. SONIC contains only benign instances and 2018 JID Ed. contains only malignant instances; for these we report % change in binary accuracy.

study withheld from the training set and columns correspond to the study used for evaluation. Rows and columns are ordered by % malignancy. For Figure 3 this is the same as ordering by dataset size, SONIC being the largest study.

Consistent with our modelling assumptions and results so far, accuracies evaluated on group g decrease as a result of removing group g from the training set (diagonal entries of Figure 3). However, additional patterns show more nuanced relationships between groups.

Positive values in the upper right region of Figure 3 show that excluding studies with low malignancy rates can raise performance evaluated on studies with high malignancy rates. This could be partially due to differences in label distributions when removing certain studies from the training data. Importantly, this provides a counterpoint to an assumption implicit in Assumption 1, that group risks decrease in the total training set size n, regardless of the groups these n instances belong to. To study more nuanced interactions between pairs $g' \neq g$, future work could modify Eq. (4) by reparameterizing $r(\cdot)$ to directly account for $n_{g'}$.

Grouping by substudies within the UDA and MSK studies reveals that even within well defined groups, interactions between subgroups can arise. Negative off-diagonal entries in Figure B.5b suggest strong interactions between different groups, underscoring the importance of evaluating results across hierarchies and intersections of groups when feasible.

Of the 16,965 images in the full training set, 7,401 are from the SONIC study. When evaluating on all non-SONIC instances (like the evaluation set from the rest of the paper), withholding the SONIC study from the training set leads to higher AUROC (.905) than training on all studies (0.890). This demonstrates that more data is not always better, especially if the distributional differences between the additional data and the target populations are not well accounted for.

5. Discussion

We study the ways in which group and population performance depend on the numerical allocations of discrete groups in training sets. While focusing on discrete groups allows us to derive meaningful results, understanding similar phenomena for intersectional groups and continuous notions of inclusion is an important next step. Addressing the more nuanced relationships between the allocations of different data sources (Section 4.4) is a first step in this direction.

We find that underrepresentation of groups in training data can limit group and population accuracies. However, naive targeted data collection attempts can present undue burdens of surveillance or skirt consent (Paullada et al., 2020). When ML systems fail subpopulations due to measurement or construct validity issues, more comprehensive interventions are needed (Jacobs & Wallach, 2019).

Our results expose key properties of sub-group representation in training data from a statistical sampling perspective, complementary to current and emerging practices for ethical, contextualized data collection and curation (Gebru et al., 2018; Gebru, 2020; Denton et al., 2020; Abebe et al., 2021). Studying the role of numerical allocation targets within ethical and context-aware data collection practices will be an important step toward operationalizing our findings.

Representation is a broad, often ambiguous concept (Chasalow & Levy, 2021), and numerical allocation is an imperfect proxy of representation or inclusion. That said, if the optimal allocation for a certain group is well beyond that group's population proportion, this may be cause to reflect on why that is the case. Future work could consider allocations as a lens for auditing the limits of prediction models from a data-focused perspective and extend analysis to more objectives and loss functions (e.g. robustness or fairness objectives).

Acknowledgements

We thank Inioluwa Deborah Raji and Ludwig Schmidt for feedback at various stages of this work, and Andrea Bajcsy, Sara Fridovich-Keil, and Max Simchowitz for comments and suggestions during the editing of this manuscript. We thank Nimit Sohoni and Jared Dunnmon for helpful discussions regarding the ISIC dataset.

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE 1752814. ER acknowledges the support of a Google PhD Fellowship. This research is generously supported in part by ONR awards N00014-20-1-2497 and N00014-18-1-2833, NSF CPS award 1931853, and the DARPA Assured Autonomy program (FA8750-18-C-0101).

References

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., and Sadagopan, S. Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 329–341, 2021.

Abernethy, J., Awasthi, P., Kleindessner, M., Morgenstern, J., and Zhang, J. Adaptive sampling to reduce disparate performance. *arXiv preprint arXiv:2006.06879*, 2020.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Byrd, J. and Lipton, Z. C. What is the effect of importance weighting in deep learning? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 872–881. PMLR, 2019. URL http://proceedings.mlr.press/v97/byrd19a.html.

Chasalow, K. and Levy, K. Representativeness in statistics, politics, and machine learning. *arXiv preprint arXiv:2101.03827*, 2021.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Chen, I. Y., Johansson, F. D., and Sontag, D. A. Why is my classifier discriminatory? In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 3543–3554, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1f1baa5b8edac74eb4eaa329f14a0361-Abstract.html.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM Review*, 51(4): 661–703, 2009.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N. K., Kittler, H., and Halpern, A. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). *arXiv preprint*, 2017.

Cole, R., Gkatzelis, V., and Goel, G. Mechanism design for fair division: Allocating divisible items without payments. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pp. 251–268, New York, NY, USA, 2013. Association for Computing Machinery.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Vancouver, British Columbia, Canada*, pp. 442–450. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/59c33016884a62116be975a9bb8257e3-Abstract.html.

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

Dotan, R. and Milli, S. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 294–294, 2020.

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 *Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

Gebru, T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Gupta, R., Liu, Y., Tang, J., and Prakash, B. A. (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 3609. ACM, 2020. URL https://dl.acm.org/doi/10.1145/3394486.3409559.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Ghorbani, A. and Zou, J. Y. Data shapley: Equitable valuation of data for machine learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL http://proceedings.mlr.press/v97/ghorbani19c.html.

Habib, A., Karmakar, C., and Yearwood, J. Impact of ECG dataset diversity on generalization of CNN model for detecting QRS complex. *IEEE Access*, 7:93275–93285, 2019.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

HarvardX. HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0, 2014. URL https://doi.org/10.7910/DVN/26147.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1934–1943. PMLR, 2018. URL http://proceedings.mlr.press/v80/hashimoto18a.html.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., and Langs, G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):1–13, 2020.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2034–2042. PMLR, 2018. URL http://proceedings.mlr.press/v80/hu18a.html.

Iosifidis, V. and Ntoutsi, E. Dealing with bias via data augmentation in supervised learning scenarios. In *Proceedings of the Workshop on Bias in Information, Algorithms*, pp. 24–29, 2018.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. *arXiv preprint arXiv:1912.05511*, 2019.

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*, pp. 247–254, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lohr, S. L. *Sampling: design and analysis*. Nelson Education, 2009.

Neyman, J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.

Pukelsheim, F. *Optimal design of experiments*. Classics in applied mathematics ; 50. Society for Industrial and Applied Mathematics, classic ed. edition, 2006.

Raji, I. D. and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019.

Ryu, H. J., Adam, H., and Mitchell, M. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Suresh, H. and Guttag, J. V. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

Tschandl, P., Rosendahl, C., and Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), 2018. URL https://doi.org/10.1038/sdata.2018.161.

Vodrahalli, K., Li, K., and Malik, J. Are all training examples created equal? An empirical study. *arXiv preprint arXiv:1811.12569*, 2018.

Wan, M. and McAuley, J. J. Item recommendation on monotonic behavior chains. In Pera, S., Ekstrand, M. D., Amatriain, X., and O'Donovan, J. (eds.), *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pp. 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Wright, T. A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics & Probability Letters*, pp. 108829, 2020.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 547–558, 2020.

Yona, G., Ghorbani, A., and Zou, J. Who's responsible? Jointly quantifying the contribution of the learning algorithm and training data. *arXiv preprint arXiv:1910.04214*, 2019.

A. Proofs and Derivations

A.1. Proof of Proposition 1

Recall the random variable $\hat{L}$ defined with respect to function f, loss function ℓ, and group weights $w : \mathcal{G} \to \mathbb{R}^+$:

$$\hat{L}(w, \alpha, n; f, \ell) := \frac{1}{n} \sum_{i \in \mathcal{S}(\vec{\alpha}, n)} w(g_i) \cdot \ell(f(x_i), y_i)$$

where the randomness comes from the draws of x_i, y_i from $\mathcal{D}_{g_i}$ according to procedure $\mathcal{S}(\vec{\alpha}, n)$ (Definition 2), as well as any randomness in f.

Proposition 1 (Weights and allocations)**.** *For any loss ℓ, prediction function f and group distributions $\mathcal{D}_g$, there exist weights with $w^*(g) \propto \left(Var_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y))]\right)^{-1/2}$ such that for any triplet $(\vec{\alpha}, w, n)$ with $\sum_g \alpha_g w(g) > 0$, if $w \not\propto w^*$,[5] there exists an alternative $\vec{\alpha}^*$ with*

$$\mathbb{E}[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] = \mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)]$$
$$Var[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] < Var[\hat{L}(w, \vec{\alpha}, n; f, \ell)] \, .$$

If $w(g) > w^*(g)$, $\alpha^*_g > \alpha_g$ and if $w(g) < w^*(g)$, $\alpha^*_g < \alpha_g$.

Proof. For any n, any (α, w) pair induces a vector $\vec{\gamma}'$ with entries $\gamma'_g(w, \vec{\alpha}) := w(g)\alpha_g / \left(\sum_{g\in\mathcal{G}} w(g)\alpha_g\right)$, where

$$\begin{aligned}
\mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)] &= \frac{1}{n} \sum_{(x_i, y_i, g_i) \in \mathcal{S}(\vec{\alpha}, n)} w(g_i)\mathbb{E}[\ell(f(x_i), y_i)] \\
&= \sum_{g\in\mathcal{G}} \frac{n_g}{n} w(g)\mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y)] \\
&= c \cdot \mathbb{E}_{g\sim\text{Multinomial}(\vec{\gamma}')}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_g}\left[\ell(f(x), y)\right]\right]
\end{aligned}$$

for constant $c = \sum_g \alpha_g w(g)$. The vector $\vec{\gamma}'$ in this sense describes an implicit 'target distribution' induced by the applying weights w after sampling with allocation $\vec{\alpha}$. Note that unless $w_g = 0$ for all g with $\alpha_g > 0$, $\vec{\gamma}'$ has at least one nonzero entry. The constant c re-scales the weighted objective function with original weights w so as to match the expected loss with respect to the group proportions $\vec{\gamma}'$. Stated another way, for any alternative allocation α', we could pick weights $w'(g) = c\gamma'_g/\alpha'_g$ (letting $w'(g) = 0$ if $\alpha'_g = 0$), and satisfy

$$\mathbb{E}[\hat{L}(w', \vec{\alpha}', n; f, \ell)] = \mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)] \, .$$

Given this correspondence, we now find the pair $(\vec{\alpha}^*, w^*)$ which minimizes $Var[\hat{L}(cw', \vec{\alpha}', n; f, \ell)]$, subject to $w'(g)\alpha'_g = c\gamma'_g$. Since the original pair $(\vec{\alpha}, w)$ satisfies this constraint (by construction), we must have

$$\min_{\vec{\alpha}', w' : w'(g)\alpha'_g = c\gamma'_g} Var[\hat{L}(w', \vec{\alpha}', n; f, \ell)] \leq Var[\hat{L}(w, \vec{\alpha}, n; f, \ell)] \, .$$

We first compute $Var[\hat{L}(w', \vec{\alpha}', n; f, \ell)]$. By Definition 2, samples (x_i, y_i) are assumed to be independent draws from distributions $\mathcal{D}_{g_i}$, so that the variance of the estimator can be written as (for convenience we assume here that $n\alpha'_g \in \mathbb{Z}$, see the discussion below):

$$\begin{aligned}
Var\left[\hat{L}(w', \vec{\alpha}', n; f, \ell)\right] &= \frac{1}{n^2} \sum_{g\in\mathcal{G}} w'(g)^2 \sum_{i=1}^{n\alpha'_g} \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_g}\left[\left(\ell(f(x_i), y_i) - \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y)]\right)^2\right] \\
&= \frac{1}{n} \sum_{g\in\mathcal{G}} \alpha'_g w'(g)^2 Var\left[\ell_g^{(i)}\right] \, ,
\end{aligned}$$

[5] We use the symbol $\not\propto$ to denote "not approximately proportional to." The approximately part of this relation stems from finite and integer sample concerns; for example, the proposition holds if we consider $w \not\propto w^*$ to mean $\exists g \in \mathcal{G} : |1 - \frac{w(g)}{w^*(g)}| > \frac{|\mathcal{G}|}{\alpha_g n}$.

where $Var\left[\ell_g^{(i)}\right]$ denotes shorthand for $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_g}\left[\left(\ell(f(x_i),y_i)-\mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x),y)]\right)^2\right]$. Now, to respect the constraint $w'(g)\alpha'_g = c\gamma'_g$ means that for any g with $\gamma'_g > 0$, $w'(g)$ is a deterministic function of α'_g, since c and $\vec{\gamma}'$ are determined by the initial pair $(\vec{\alpha}, w)$. Then it is sufficient to compute

$$\operatorname*{argmin}_{\alpha'\in\Delta^{|\mathcal{G}|}} \frac{1}{n}\sum_{g\in\mathcal{G}:\gamma'_g>0}\alpha'_g\left(\frac{c\gamma'_g}{\alpha'_g}\right)^2 Var\left[\ell_g^{(i)}\right] = \operatorname*{argmin}_{\alpha'\in\Delta^{|\mathcal{G}|}} \frac{c^2}{n}\sum_{g\in\mathcal{G}:\gamma'_g>0}\frac{(\gamma'_g)^2}{\alpha'_g}Var\left[\ell_g^{(i)}\right] .$$

The minimizer $\vec{\alpha}^*$ has entries $\alpha^*_g = \gamma'_g\sqrt{Var[\ell_g^{(i)}]}/\left(\sum_{g\in\mathcal{G}}\gamma'_g\sqrt{Var[\ell_g^{(i)}]}\right)$. Because $\vec{\alpha}^*$ is unique and determines w^*, $Var[\hat{L}(w^*,\vec{\alpha}^*,n;f,\ell)] < Var[\hat{L}(w',\vec{\alpha}',n;f,\ell)]$ with strict inequality unless $(w',\vec{\alpha}') = (w^*,\vec{\alpha}^*)$. The optimal weights are $w^*(g) = c\gamma'_g/\alpha^*_g = c\left(\sum_{g\in\mathcal{G}}\gamma'_g\sqrt{Var[\ell_g^{(i)}]}\right)/\sqrt{Var[\ell_g^{(i)}]}$. Note that for pair of groups (g,g'), the relative weights satisfy $w^*(g)/w^*(g') = \sqrt{Var[\ell_{g'}^{(i)}]/Var[\ell_g^{(i)}]}$, and thus do not depend on $\vec{\gamma}'$.

If we consider finite sample concerns, the minimizer α^* must satisfy integer values $n\alpha^*_g \in \mathbb{Z}^+\ \forall g\in\mathcal{G}$. In this case, efficient algorithms exist for finding the integral solution to allocating n_g (Wright, 2020). However, the non-integer restricted solution α^* has a closed form solution, and we will use the fact that for any group g, α^*_g as defined above and its variant with the additional constraint that $n\alpha^*_g\in\mathbb{Z}^+$ can differ by at most $\frac{|\mathcal{G}|}{n}$. This means that any $\vec{\alpha}$ with $|\alpha_g-\alpha^*_g| > \frac{|\mathcal{G}|}{n}$ cannot be a minimizer of the objective function, even constrained to $n\alpha^*_g\in\mathbb{Z}$. Since $w(g)\alpha_g = c\gamma'_g$, an equivalent statement in terms of w is $|1-\frac{w(g)}{w^*(g)}| > \frac{w(g)|\mathcal{G}|}{nc\gamma'(g)} = \frac{|\mathcal{G}|}{n\alpha_g} = \frac{|\mathcal{G}|}{n_g}$.

Finally, we show that if $w^*(g) < w(g)$, $\alpha^*_g > \alpha_g$. This follows from our definition of $\vec{\gamma}'$ such that $w(g) = c\gamma'_g/\alpha_g$, and our constraint, such that $w^*(g) = c\gamma'_g/\alpha^*_g$. From these, we must have that $w^*(g)\alpha^*_g = w(g)\alpha_g$, from which the claim and its reverse follow. □

A.2. Proof of Proposition 2

Proposition 2. *Given a population made up of disjoint groups $g\in\mathcal{G}$ with population prevalences γ_g, under the conditions of Assumption 1, the allocation $\vec{\alpha}^*\in\Delta^{|\mathcal{G}|}$ that minimizes the approximated population risk $\hat{\mathcal{R}}$ in eq. (5) has elements:*

$$\alpha^*_g = \frac{\left(\gamma_g\sigma_g^2\right)^{1/(p+1)}}{\sum_{g\in\mathcal{G}}\left(\gamma_g\sigma_g^2\right)^{1/(p+1)}} . \tag{6}$$

If $\sigma_g = 0\ \forall g\in\mathcal{G}$, then any allocation in $\Delta^{|\mathcal{G}|}$ minimizes $\hat{\mathcal{R}}$.

Proof. Recall the decomposition of the estimated population risk:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(\hat{f}_{\mathcal{S}}(x),y)\right] = \sum_{g\in\mathcal{G}}\gamma_g\cdot\mathbb{E}_{(x,y)\sim\mathcal{D}_g}\left[\ell(\hat{f}(x),y)\right] \approx \sum_{g\in\mathcal{G}}\gamma_g\left(\sigma_g^2(n\alpha_g)^{-p}+\tau_g^2 n^{-q}+\delta_g\right) .$$

Now we find

$$\vec{\alpha}^* = \operatorname*{argmin}_{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}\sum_{g\in\mathcal{G}}\gamma_g\left(\sigma_g^2(n\alpha_g)^{-p}+\tau_g^2n^{-q}+\delta_g\right) = \operatorname*{argmin}_{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}(n^{-p})\sum_{g\in\mathcal{G}}\gamma_g\left(\sigma_g^2(\alpha_g)^{-p}\right) = \operatorname*{argmin}_{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}\sum_{g\in\mathcal{G}}\gamma_g\sigma_g^2\alpha_g^{-p} .$$

If $\sigma_g = 0\ \forall g\in\mathcal{G}$, then any allocation $\vec{\alpha}^*\in\Delta^{|\mathcal{G}|}$ minimizes the approximated population loss. Otherwise, $\alpha^*_g = 0$ will be 0 for any group with $\sigma_g = 0$; what follows describes the solution α^*_g for g with $\sigma_g > 0$. If any $\alpha_g = 0$, then the objective is unbounded above, so we can restrict our constraints to $\vec{\alpha}\in(0,1)^{|\mathcal{G}|}$. As the sum of convex functions, the objective function is convex in $\vec{\alpha}$. It is continuously differentiable when $\alpha_g > 0$, $\forall g\in\mathcal{G}$. The KKT conditions are satisfied when

$$p\gamma_g\sigma_g^2\alpha_g^{-(p+1)} = \lambda \quad \forall g$$
$$\sum_g\alpha_g = 1 .$$

Solving this system of equations yields that the KKT conditions are satisfied when $\alpha^*_g = \left(\gamma_g\sigma_g^2\right)^{1/(p+1)}/\sum_g\left(\gamma_g\sigma_g^2\right)^{1/(p+1)}$. Since this is the only solution to the KKT conditions, it is the unique minimizer. □

A.3. Proof of Corollary 1

Corollary 1 (Many groups with equal σ_g)**.** *When $\sigma_g = \sigma > 0,\ \forall g \in \mathcal{G}$, the allocation that minimizes $\hat{\mathcal{R}}$ in Eq. (5) satisfies $\alpha_g^* \geq \gamma_g$ for any group with $\gamma_g \leq \frac{1}{|\mathcal{G}|}$.*

Proof. Let $m = |\mathcal{G}|$ denote the number of groups. When $\sigma_g = \sigma\ \forall g \in \mathcal{G}$,

$$\alpha_g^* = \frac{\gamma_g^{1/(p+1)}}{\sum_g \gamma_g^{1/(p+1)}} = \gamma_g \cdot \frac{1}{\gamma_g + \gamma_g^{p/(p+1)} \sum_{g' \neq g} \gamma_{g'}^{1/(p+1)}} \,.$$

Since $p > 0$ by Assumption 1, we have that $\sum_{i=1}^n {\gamma_i}^{1/(p+1)}$ with γ_i subject to (a) $\sum_{i=1}^n \gamma_i = s$ and (b) $\gamma_i > 0$ is maximized when all γ_i are equal. Then, since $\sum_{g' \neq g} \gamma_g' = 1 - \gamma_g$,

$$\begin{aligned}
\gamma_g \cdot \frac{1}{\gamma_g + \gamma_g^{p/(p+1)} \sum_{g' \neq g} \gamma_{g'}^{1/(p+1)}} &\geq \gamma_g \cdot \frac{1}{\gamma_g + \gamma_g^{p/(p+1)} \sum_{g' \neq g} \left(\frac{1-\gamma_g}{m-1}\right)^{1/(p+1)}} \\
&= \gamma_g \cdot \frac{1}{\gamma_g + ((m-1)\gamma_g)^{p/(p+1)} \, (1-\gamma_g)^{1/(p+1)}} \,.
\end{aligned}$$

When $\gamma_g \leq 1/m$, $\gamma_g/(1-\gamma_g) \leq \frac{1}{m-1}$, so that

$$\alpha_g^* \geq \gamma_g \cdot \frac{1}{\gamma_g + (1-\gamma_g)\,(m-1)^{p/(p+1)}\,(\gamma_g/(1-\gamma_g))^{p/(p+1)}} \geq \gamma_g \cdot \frac{1}{\gamma_g + (1-\gamma_g)} = \gamma_g \,. \qquad \square$$

A.4. Proof of Corollary 2

Corollary 2 (Unequal per-group constants)**.** *For two groups $\{A, B\} = \mathcal{G}$ with $\gamma_A < \gamma_B$, and parameters $\sigma_A, \sigma_B > 0$ in Eq. (4), the allocation of the smaller group α_A^* that minimizes $\hat{\mathcal{R}}$ in Eq. (5) is upper and lower bounded as*

$$\begin{aligned}
\frac{\gamma_A(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + \gamma_B(\sigma_B^2)^{1/(p+1)}} &< \alpha_A^* \\
&< \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}} \,.
\end{aligned}$$

When $\sigma_A \geq \sigma_B$, $\alpha_A^* > \gamma_A$, and when $\sigma_A \leq \sigma_B$, $\alpha_A^* < 1/2$.

Proof. For the setting of two groups, we can express the optimal allocations as:

$$\alpha_A^* = \frac{\left(\gamma_A \sigma_A^2\right)^{1/(p+1)}}{\left(\gamma_A \sigma_A^2\right)^{1/(p+1)} + \left((1-\gamma_A)\sigma_B^2\right)^{1/(p+1)}}, \quad \alpha_B^* = 1 - \alpha_A^*$$

Rearranging,

$$\alpha_A^* = \gamma_A \frac{1}{\gamma_A + (\sigma_B^2/\sigma_A^2)^{1/(p+1)}(1-\gamma_A)^{1/(p+1)}(\gamma_A)^{p/(p+1)}} \,.$$

For $p > 0$, it holds that $0 < \frac{1}{p+1} < 1$. Therefore, for any $p > 0$ and $\gamma < 0.5$,

$$\gamma < (\gamma)^{\frac{1}{p+1}}(1-\gamma)^{\frac{p}{p+1}} < (1-\gamma) \,.$$

From this, we derive the upper bound

$$\alpha_A^* < \gamma_A \frac{1}{\gamma_A + (\sigma_B^2/\sigma_A^2)^{1/(p+1)}(\gamma_A)} = \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}},$$

and the lower bound

$$\alpha_A^* > \gamma_A \frac{1}{\gamma_A + (\sigma_B^2/\sigma_A^2)^{1/(p+1)}(1-\gamma_A)} = \gamma_A \frac{(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}(1-\gamma_A)} .$$

When $\sigma_A \geq \sigma_B$,

$$\alpha_A^* > \gamma_A \frac{(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}(1-\gamma_A)} > \gamma_A \frac{(\sigma_A^2)^{1/(p+1)}}{(\gamma_A + 1 - \gamma_A)(\sigma_A^2)^{1/(p+1)}} = \gamma_A ,$$

and when $\sigma_A \leq \sigma_B$,

$$\alpha_A^* < \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}} < \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_A^2)^{1/(p+1)}} = 1/2 .$$

□

A.5. Additional Derivations

In Example 3, we consider the model $y_i = x_i^\top \beta + \alpha_1 \mathbb{I}[g_i = A] + \alpha_2 \mathbb{I}[g_i = B] + \mathcal{N}(0, \sigma^2)$ where $x_i \sim \mathcal{N}(0, \Sigma_x)$ and denote $\theta = [\alpha_1, \alpha_2, \beta^\top]$ (note: here α_i denote the model coefficients, not allocations). We want to compute

$$\begin{aligned}\mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[(\hat{f}(x) - y)^2 | g = A\right] &= \sigma^2 + \mathbb{E}\left[\|x^\top(\hat{\beta} - \beta) + (\hat{\alpha}_1 - \alpha_1)\|^2\right] \\ &= \sigma^2 + \mathbb{E}\left[x^\top(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top x\right] + 2\mathbb{E}\left[(\hat{\alpha}_1 - \alpha_1)(\hat{\beta} - \beta)^\top x\right] + \mathbb{E}\left[(\hat{\alpha}_1 - \alpha_1)^2\right].\end{aligned}$$

Since the draw $(x, y, g) \sim \mathcal{D}$ is independent of the data from which the ordinary least squares solution $\hat{\theta}$ is predicted, we can write out each of these terms in terms of the dependence on n, the total number of data points, as well as n_A and n_B (where $n_A + n_B = n$), the total number of datapoints for each group, from which $\hat{\theta}$ is estimated. To do this, we'll solve for the entries of the covariance matrix:

$$\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top] = \begin{bmatrix} \mathbb{E}\left[(\hat{\alpha_1} - \alpha_1)^2\right] & \mathbb{E}\left[(\hat{\alpha_1} - \alpha_1)(\hat{\alpha_2} - \alpha_2)\right] & \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\alpha_1} - \alpha_1)\right]^\top \\ \mathbb{E}\left[(\hat{\alpha_1} - \alpha_1)(\hat{\alpha_2} - \alpha_2)\right] & \mathbb{E}\left[(\hat{\alpha_2} - \alpha_2)^2\right] & \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\alpha_2} - \alpha_2)\right]^\top \\ \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\alpha_1} - \alpha_1)\right] & \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\alpha_2} - \alpha_2)\right] & \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top\right] \end{bmatrix} = \sigma^2(Z^\top Z)^{-1}$$

where Z is the $n \times (d+2)$ design matrix with rows $\{(\mathbb{I}[g_i = A], \mathbb{I}[g_i = B], x_i^\top)\}_{i=1}^n$. Next, we find the block entries of the matrix $(Z^\top Z)^{-1}$. We first interrogate the term within the inverse:

$$Z^\top Z = \begin{bmatrix} n_A & 0 & n_A \bar{x}_A^\top \\ 0 & n_B & n_B \bar{x}_B^\top \\ n_A \bar{x}_A & n_B \bar{x}_B & X^\top X \end{bmatrix}$$

where $\bar{x}_A = \frac{1}{n_A} \sum_{i=1}^n x_i \cdot \mathbb{I}[g_i = A]$, and similarly for $\bar{x}_B$. We'll now use the Schur complement to compute the desired blocks of Σ. The Schur complement is

$$S = X^\top X - \begin{bmatrix} n_A \bar{x}_A & n_B \bar{x}_B \end{bmatrix} \begin{bmatrix} n_A^{-1} & 0 \\ 0 & n_B^{-1} \end{bmatrix} \begin{bmatrix} n_A \bar{x}_A^\top \\ n_B \bar{x}_B^\top \end{bmatrix} = X^\top X - n \bar{x} \bar{x}^\top ,$$

which we simplify to $S = X^\top X$ by assuming that we zero-mean the sample feature matrix X before calculating the least squares solution. Using the Schur complement, the covariance matrix in block form is

$$(Z^\top Z)^{-1} = \begin{bmatrix} \frac{1}{n_A} + \bar{x}_A^\top S^{-1} \bar{x}_A & \bar{x}_A^\top S^{-1} \bar{x}_B & -\bar{x}_A^\top S^{-1} \\ \bar{x}_B^\top S^{-1} \bar{x}_A & \frac{1}{n_B} + \bar{x}_B^\top S^{-1} \bar{x}_B & -\bar{x}_B^\top S^{-1} \\ -S^{-1} \bar{x}_A & -S^{-1} \bar{x}_B & S^{-1} \end{bmatrix} .$$

Plugging in the appropriate blocks to our original equation, we get:

$$\mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[(\hat{f}(x)-y)^2|g=A\right] = \sigma^2 + \mathbb{E}_{(x,y)\sim\mathcal{D}_A}\left[x^\top(\hat{\beta}-\beta)(\hat{\beta}-\beta)^\top x + 2(\hat{\alpha}_1-\alpha_1)(\hat{\beta}-\beta)^\top x + (\hat{\alpha}_1-\alpha_1)^2\right]$$
$$= \sigma^2(1+\frac{1}{n_A}+\mathbb{E}_{g=A}\left[x^\top S^{-1}x + \bar{x}_A^\top S^{-1}\bar{x}_A\right]),$$

where the middle term cancels since $\mathbb{E}[x]=0$. Note that S is the scaled sample covariance matrix. The vectors x_i are drawn i.i.d. from $\mathcal{N}(0,\Sigma_x)$ so that S^{-1} follows an inverse Wishart distribution with parameters n, d, Σ_x. For the fresh sample x,

$$\mathbb{E}[x^\top S^{-1}x] = \text{Trace}(\mathbb{E}[S^{-1}]\mathbb{E}[xx^\top]) = \frac{d}{n-d-1}Tr(\Sigma_x^{-1}\Sigma_x) = \frac{d}{n-d-1}\,.$$

For the $\bar{x}_A^\top S^{-1}\bar{x}_A$ term we invoke the matrix inversion lemma. For a single row x_i of X, let X_{-i} denote the $(n-1)\times d$ matrix comprised of all rows of X except X_i. Then

$$x_i^\top(X^\top X)^{-1}x_i = x_i^\top(X_{-i}^\top X_{-i} + x_i x_i^\top)^{-1}x_i$$
$$= x_i^\top\left((X_{-i}^\top X_{-i})^{-1} - (X_{-i}^\top X_{-i})^{-1}x_i(1+x_i^\top(X_{-i}^\top X_{-i})^{-1}x_i)^{-1}x_i^\top(X_{-i}^\top X_{-i})^{-1}\right)x_i\,.$$

Letting $a_i = x_i^\top(X_{-i}^\top X_{-i})^{-1}x_i \geq 0$, we rewrite the above as

$$x_i^\top(X^\top X)^{-1}x_i = a_i - \frac{a_i^2}{1+a_i} = \frac{a_i}{1+a_i} \leq a_i\,.$$

Since the x_i are independent and zero mean, $\mathbb{E}\left[x_i^\top(X_{-i}^\top X_{-i})^{-1}x_j\right] = \mathbb{E}[x_i^\top]\mathbb{E}[(X_{-i}^\top X_{-i})^{-1}x_j] = 0\,\forall i\neq j$. From a similar argument to that given above, we derive that $\mathbb{E}[a_i] = d/(n-d-2)$, so that

$$\mathbb{E}\left[\bar{x}_A^\top S^{-1}\bar{x}_A\right] = \mathbb{E}\left[\left(\frac{1}{n_A}\sum_i^{n_A}x_i\right)^\top S^{-1}\left(\frac{1}{n_A}\sum_i^{n_A}x_i\right)\right] = \frac{1}{n_A^2}\mathbb{E}\left[\sum_i^{n_A}x_i^\top S^{-1}x_i\right] \leq \frac{1}{n_A}\cdot\frac{d}{n-d-2}\,.$$

Putting this all together, we conclude that for $n \gg d$,

$$\mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[(\hat{f}(x)-y)^2|g=A\right] = \sigma^2\left(1+\frac{1}{n_A}+\frac{d}{n-d-1}+\mathbb{E}[\bar{x}_A^\top S^{-1}\bar{x}_A]\right) = \sigma^2\left(1+\frac{1}{n_A}+O\left(\frac{d}{n}\right)\right)\,.$$

B. Experiment Details

Appendix B.1 details the datasets used including preprocessing steps and any data excluded from the experiments; the remainder of Appendix B provides a detailed explanation of each experiment described in the main text. Appendix C.2 describes additional experiments to complement the findings of the main experiments. Code to process data and replicate the experiments detailed here is available at `https://github.com/estherrolf/representation-matters`.

All image prediction models were run on a machine with 8 nVidia 2080 GPUs, 384GB RAM, and and Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz, 12C/24HT, with the exception of the experiments detailed in Appendix B.7, which was run on a machine with 1 nVidia Titan-V GPU, 256GB RAM, Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz, 20C/40HT. The non-image models were run on the latter machine..

B.1. Dataset Descriptions

We use and modify benchmark machine learning and fairness datasets, as well as more recent datasets on image diagnosis and student performance to study the effect of training set allocations on group and population accuracies in a systematic fashion. Each of the datasets we use is described below, with download links given at the end of this section.

Modified CIFAR-4. We modify an existing machine learning benchmark dataset, CIFAR-10 (Krizhevsky, 2009), to instantiate binary classification tasks with binary groups, where groups g are statistically independent of the labels y. We take four of the ten CIFAR-10 classes: {plane, car, bird, horse}, and sort them into binary categories of {land/air, animal/vehicle} as in Fig. B.1. This instantiation balances the classes labels among the two groups g, so that no matter the

allocation of groups to the training set, the label distribution will remain balanced. This will eliminate class imbalance as the cause of changes in accuracy due to training set composition or up-weighting methods.

There are $5,000$ training and $1,000$ test instances of each class in the CIFAR-10 dataset, resulting in $10,000$ training instances of each group in our modified CIFAR-4 dataset ($20,000$ instances total), and $2,000$ instances of each group in the test set ($4,000$ instances total). By construction, the average label in the test and train sets is 0.5. Since this dataset is designed to assess the main questions of our study under a controlled setting and there is not a natural setting of population rates of animal and vehicle photos, we set the population prevalence parameter of group A (images of animals) as $\gamma_A = 0.1$.

Figure B.1. Modified CIFAR-4 dataset setup.

Goodreads ratings prediction. The Goodreads dataset (Wan & McAuley, 2018) contains a large corpus of book reviews and ratings, collected from public book reviews online. From the text of the reviews, we predict the numerical rating (integers 1-5) a reviewer assigns to a book. From the larger corpus, we consider reviews for two genres: history/biography (henceforth "history") and fantasy/paranormal (henceforth "fantasy"). We calculate the population prevalences from the total number of reviews in the corpus for each genre. As of the writing of this document, there are $2,066,193$ reviews for history and biography books, and $3,424,641$ for fantasy and paranormal books, so that $\vec{\gamma} = [0.376, 0.624]$.

After dropping entries with no valid review or rating, we have 1,985,464 reviews from the history genre, and 3,309,417 reviews from the fantasy genre, with no books assigned to both genres. To reduce dataset size and increase performance of the basic prediction task, we further reduce each dataset to only the 100 most frequently reviewed books of each genre (following a procedure similar to (Chen et al., 2018)). To instantiate the dataset we use in our analysis, we draw 62,500 review/rating pairs uniformly at random from each of these pools. The mean review for fantasy instances is 4.146, for history it is 4.103. We allocate 20% of the data to a test set, and the remaining 80% to the training set, with an equal number of instances of each genre in each set.

We use tf-idf vectors of the 2,000 most frequently occurring reviews from the entire dataset of 125,000 instances as features (a similar featurization to (Chen et al., 2018), with fewer total features). We note that this is a different version of the goodreads dataset from that used in (Chen et al., 2018); the updated dataset we use has more reviews, and we use different group variables.

Adult. The Adult dataset, originally extracted from the 1994 Census database, is downloaded from the UCI Machine Learning Repository (Dua & Graff, 2017). The dataset contains demographic features and the classification task is to predict whether an individual makes more than 50K per year. We drop the final-weight feature, a measure of the proportion of the population represented by that profile. There are $d = 13$ remaining features; we one-hot encode non-binary discontinuous features including work class, education, marital status, occupation, relationship, race, and native country, resulting in 103 feature columns.

For the main analysis, we exclude features for sex, husband, and wife as they indicate group status but do not affect predictive ability (see Appendix C.1), for a total of 10 features and 100 columns in the feature matrix. We keep the original train and test splits, with $32,561$ ($21,790$ from group male and $10,771$ from group female) and $16,281$ ($10,860$ from group male and $5,421$ from group female) instances respectively. We group instances based on gender (encoded as binary male/female).

While the training and tests sets have roughly a $2/3$ male group, $1/3$ female group balance, we set $\vec{\gamma} = [0.5, 0.5]$ to more adequately reflect the true proportions of men and women in the world. In the test set, the average label for the male group is 0.300, for the female group it is 0.109 (for the train set, the numbers are similar; 0.306 for the male group and 0.109 for the female group).

Mooc dataset. The dataset we refer to as the MOOC (Massive Open Online Course) dataset is the HarvardX-MITx Person-Course Dataset (HarvardX, 2014). This dataset contains anonymized data from the 2013 academic year offerings of MITx and HarvardX edX classes. Each row in the dataset corresponds to a student-course pairing; students taking more than one course may correspond to multiple rows. We keep the following demographic, participation, and administrative features: gender, LoE_DI (highest completed level of education), final_cc_cname_DI (country), YoB, ndays_act (number of active days on the edX platform), nplay_video (number of video plays on edX), nforum_posts (number of discussion forum posts on edX), nevents (number of interactions on edX), course_id, certified (whether the student achieved a high enough grade for certification). After dropping instances without valid entries for these features, we have 25,213 instances. We one-hot encode non-binary discontinuous features including course_id, LoE_DI, and final_cc_cname_DI, resulting in an expanded 47 feature columns from the original 9 features. In Appendix C.1 we exclude demographic features. We partition the data into a train set of size 24,130 and a test set of size 6,032, with the same proportion of groups in each set.

We group instances by the highest level of education self-reported by the person taking the class; we bin this into those who have completed at most secondary education, and those who have completed more than secondary education. For reference, in the USA, completing secondary education corresponds to completing high school. The train and test sets contain an equal fraction of each group, about 16% instances where the person taking the course had no recorded education higher than secondary. Of all training and test instances, those from group A (student had at most secondary education) have a 5.2% certification rate, and those from group B (student had more than secondary education) have an 8.2% certification rate.

ISIC skin cancer detection dataset. We download the dataset from the ISIC archive (Codella et al., 2017; Tschandl et al., 2018; Codella et al., 2019) using the repository `https://github.com/GalAvineri/ISIC-Archive-Downloader` to assist in downloading. To match the dataset used in Sohoni et al. (2020), we use only images added or modified prior to 2019, which gives us 23,906 instances. From this set we include only images that are explicitly labeled as benign or malignant. We additionally remove any data points from the 'SONIC' study, as these are all benign cases, and are easily identified via colorful dots on the images (Sohoni et al., 2020). The remaining 11,952 instances are to our knowledge identical to the "Non-patch" dataset in (Sohoni et al., 2020), up to the random splits into training/validation and test sets.

As groups, we subset based on the approximate age of the patient that is the subject of the photo. Approximate age is binned to the nearest 5 years in the original dataset; we design groups as A $= \{i : \texttt{age\_approx[i]} \geq 55\}$ and B $= \{i : \texttt{age\_approx[i]} < 55\}$. Of the group A instances in the training and test sets, 31.4% are malignant; of the group B instances, 8.37% are malignant. We set $\vec{\gamma}$ to match the distribution in the 11,952 data points in our preprocessed set, so that $\vec{\gamma} = [0.43, 0.57]$. We split 80% of the data into a train set and 20% into a test set, with the same proportion of groups in each set.

The ISIC archive is a compilation of medical images aggregated from many individual studies. Experiments detailed in Section 4.4 and Appendix B.7 use the study from which the image originates as groups to investigate the interactions between groups when $|\mathcal{G}| > 2$. These results also motivate our choice to exclude the SONIC study from the dataset we use for the main analysis.

Download links. Code for processing these datasets as described above can be found in the code repository accompanying this paper. The datasets we use (before our subsetting/preprocessing) can be downloaded at the following links:

- CIFAR-10: `https://www.cs.toronto.edu/~kriz/cifar.html`
- Goodreads: `https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home`
- Adult: `https://archive.ics.uci.edu/ml/datasets/adult`
- Mooc: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147`
- ISIC: `https://www.isic-archive.com`

dataset name	metric	model used	objective	parameters
CIFAR-4	0/1 loss	pretrained resnet-18 (fine-tuned)	ERM	num_epochs = 20, lr = 1e-3 , wd = 1e-4, momentum = 0.9
			IS	num_epochs = 20, lr = 1e-3, momentum = 0.9 wd = 1e-2 if $\alpha_A \geq 0.98$ or $\alpha_A \leq 0.02$; wd = 1e-3 otherwise
			GDRO	num_epochs = 20, lr = 1e-3, wd = 1e-3, momentum = 0.9, gdro_ss = 1e-2, group_adj = 4.0
ISIC	1 - AUROC	pretrained resnet-18 (fine-tuned)	ERM	num_epochs = 20, lr = 1e-3, wd = 1e-4, momentum = 0.9
			IS	num_epochs = 20, lr = 1e-3, wd = 1e-3, momentum = 0.9
			GDRO	num_epochs = 20, lr = 1e-3, wd = 1e-4, momentum = 0.9, gdro_ss = 1e-1, group_adj = 1.0
Goodreads*	ℓ_1 loss (MAE)	multiclass logistic regression	ERM	C = 1.0, penalty = ℓ_2
			IW	C = 1.0, penalty = ℓ_2
Goodreads	ℓ_2 loss (MSE)	multiclass logistic regression	ERM	$\lambda = 0.1$
			IW	$\lambda = 1.0$ if $\alpha_A \geq 0.95$ or $\alpha_A \leq 0.05$; $\lambda = 0.1$ otherwise
Mooc* (with dem. features)	1 - AUROC	random forest classifier	ERM	num_trees = 400, max_depth = 16
			IW	num_trees = 400, max_depth = 16
Mooc (no dem. features)	1 - AUROC	random forest classifier	ERM	num_trees = 400, max_depth = 8
			IW	num_trees = 400, max_depth = 8
Adult (with group features)	0/1 loss	random forest classifier	ERM	num_trees = 200, max_depth = 16
			IW	num_trees = 200, max_depth = 16
Adult* (without group features)	0/1 loss	random forest classifier	ERM	num_trees = 400, max_depth = 16
			IW	num_trees = 400, max_depth = 16

Table B.3. Models and hyperparameters used to generate Figure 1. Asterisks denote the Goodreads, Mooc and Adult setting that appear in Figure 1, the other settings are shown in Figure C.1.

Loss metrics. For binary prediction problems, we report the the 0/1 loss when there is not significant class imbalance (modified CIFAR-4, Adult). For the ISIC and Mooc tasks, we report 1 - AUROC, where AUROC is the area under the reciever operating curve. AUROC is a standard metric for medical image prediction (Sohoni et al., 2020), and for Mooc we choose this metric due to the label imbalance (low certification rates). Since AUROC is constrained to be between 0 and 1, the loss metric 1 - AUROC will also be between 0 and 1. For the Goodreads dataset, we optimize the ℓ_1 loss (mean absolute error, MAE); in Appendix C.2 we compare this to minimizing the ℓ_2 loss (mean squared error, MSE) of the predictions.

B.2. Models Details

For the neural net classifiers, we compare the empirical risk minimization (ERM) objective with an importance weighted objective, implemented via importance sampling (IS) following results in (Buda et al., 2018), and a group distributionally robust (GDRO) objective (Sagawa et al., 2020). We adapt our group distributionally robust training procedure from `https://github.com/kohpangwei/group_DRO`, the code repository accompanying (Sagawa et al., 2020). For the non-image datasets, we choose the model class from a set of common prediction functions (Appendix B.3). We use implementations of these algorithms from `https://scikit-learn.org`, using the built in sample weight parameters to implement importance weighting (IW). Since many of the prediction functions we consider are not gradient-based, we cannot apply the algorithm from (Sagawa et al., 2020), thus we do not compare to GDRO for the non-image datasets.

B.3. Hyperparameter Selection

We evaluate hyperparameters for each prediction model using 5-fold cross validation on the training sets (see Appendix B.1), stratifying folds to maintain group proportions. We evaluate the cross-validation across a sparse grid of $\vec{\alpha} \in [0.02, 0.05, 0.2, 0.5, 0.8, 0.95, 0.98]$, allowing us to determine if hyperparameters should be set differently for different values of $\vec{\alpha}$. Table B.3 describes the model and parameters which are chosen as a result of this process.

Image Datasets. For the results shown in Figure 1 with image datasets (CIFAR-4 and ISIC), we train a pretrained resnet-18 by running SGD with momentum for 20 epochs for each dataset. We did not find significant improvements from training for more epochs for either dataset. Sohoni et al. (2020) use a pretrained resnset-50 for the ISIC prediction task; we use a

resnet-18 since we are mostly working with smaller training set sizes due to subsetting. We did not find major differences in performance for ERM for ISIC between the resnset-18 and resnet-50 for the dataset sizes we considered. We use a resnset-18 for all models for the CIFAR-4 and ISIC tasks.

In the 5-fold cross validation procedure, we search over learning rate in $[0.01, 0.001, 0.0001]$ and weight decay in $[0.01, 0.001, 0.0001]$, keeping the momentum at 0.9 for both the importance sampling and ERM procedures. Given these results, for GDRO, we search over group-adjustment in $[1.0, 4.0, 8.0]$, gdro step size in $[0.1, 0.01, 0.001]$, and weight decay in $[0.001, 0.0001]$, fixing momentum at 0.9 and learning rate at $[0.001]$.

The optimal hyperparameter configurations for the coarse grid of $\vec{\alpha}$ are in Table B.3. Across $\vec{\alpha}$ values from the coarse grid, either the optimal parameters for each objective were largely consistent, or performance did not vary greatly between hyperparameter configurations for almost all dataset/objective configurations. As a result, for both the modified CIFAR and ISIC datasets, we keep the hyperparameters fixed across $\vec{\alpha}$, with the exception of IS for CIFAR-4, where we increase weight decay for extreme $\vec{\alpha}$ (see Table B.3).

Non-Image Datasets. For the Goodreads dataset, we evaluate the following models and parameters: ridge regression model with $\lambda \in [0.01, 0.1, 1.0, 10.0, 100.0]$, random forest regressor with splits determined by MSE, and with number of trees and maximum depth of trees $\in [100, 200, 400] \times [32, 64, 128]$, and a multiclass logistic regression classifier with ℓ_2 inverse regularization strength parameter $C \in [0.01, 0.1, 1.0, 10.0]$.

The multiclass logistic regression model minimized mean absolute error (MAE) over the models we considered. For ERM, the optimal regularization parameter was $C = 1.0$ for all $\vec{\alpha}$ in our sparse grid. For IW, the optimal regularization value was $C = 1.0$ for all α_A other than 0.98, where the optimal for history MAE was $C = 10.0$. Since this only effected one group, and was not symmetric across $\vec{\alpha}$, for the evaluation results, we set $C = 1.0$ for all $\vec{\alpha}$.

For the Mooc dataset, we evaluate a binary logistic regression model with ℓ_2 penalty and inverse regularization parameter $C \in [0.001, 0.01, 0.1, 1.0, 10.0]$, a random forest classifier with number of tress and maximum depth of trees $\in [100, 200, 400] \times [8, 16, 32]$, and ridge regression model with $\lambda \in [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0]$ and threshold for binary classification decision 0.5. The best model across both group and population accuracies was a random forest model; the best maximum depth was 16 for both ERM and IW, and the results were robust to the number of estimators, so we chose 200 for both. The best hyperparameters were largely consistent for all $\vec{\alpha}$ in the sparse grid.

For the Adult dataset, we evaluate the same models and parameter configurations as the MOOC dataset. The best model across both ERM and IW was the random forest model, with optimal values given in Table B.3.

B.4. Navigating Tradeoffs

Using the selected hyperparameters from the procedure described in Appendix B.3, we evaluate performance on the heldout test sets (see Appendix B.1). For the final evaluation, we evaluate $\alpha \in [0.0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, .3, 0.35, .4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.98, 0.99, 1.0]$, skipping the extremes for GDRO and IS/IW.

The training set size is determined by the smaller of the training set sizes for each group, denoted as $\min_g n_g$ in Table 1. This results in $n = 10,000$ for the modified CIFAR-4, $n = 4,092$ for ISIC, $n = 50,000$ for goodreads, $n = 3,897$ for Mooc, and $10,771$ for adult. Note that the number of test set instances does not necessarily have the proportions of instances indicated by γ_A in Table 1. For the CIFAR-4 and Goodreads dataset, the test set is constructed to have 50% instances from group A, and 50% instances from group B; in reporting performance, we take a weighted average over the errors from each group, as in Eq. (1). For adult, we re-weight the test set instances to reflect $\vec{\gamma} = (0.5, 0.5)$. For the remaining two datasets, the test set compositions align with $\vec{\gamma}$. We assess variability in the performance of each method under each setting by examining results over 10 random seeds, which apply to both the random sampling in the subsetting procedure and the randomness in the training procedure.

B.5. Assessing scaling law fit

In addition to the results on the holdout set in the previous section, where we vary $\vec{\alpha}$ and keep the size of the training set fixed at $n = \min_g n_g$, we conduct subsetting experiments to assess the affect of n, as well as n_g, on the group accuracies as in Eq. (7). Specifically, for groups A and B, we vary the relative number of data points from group A and group B, as well as the total number of datapoints n. We define our subsetting grid by subsampling twice for each

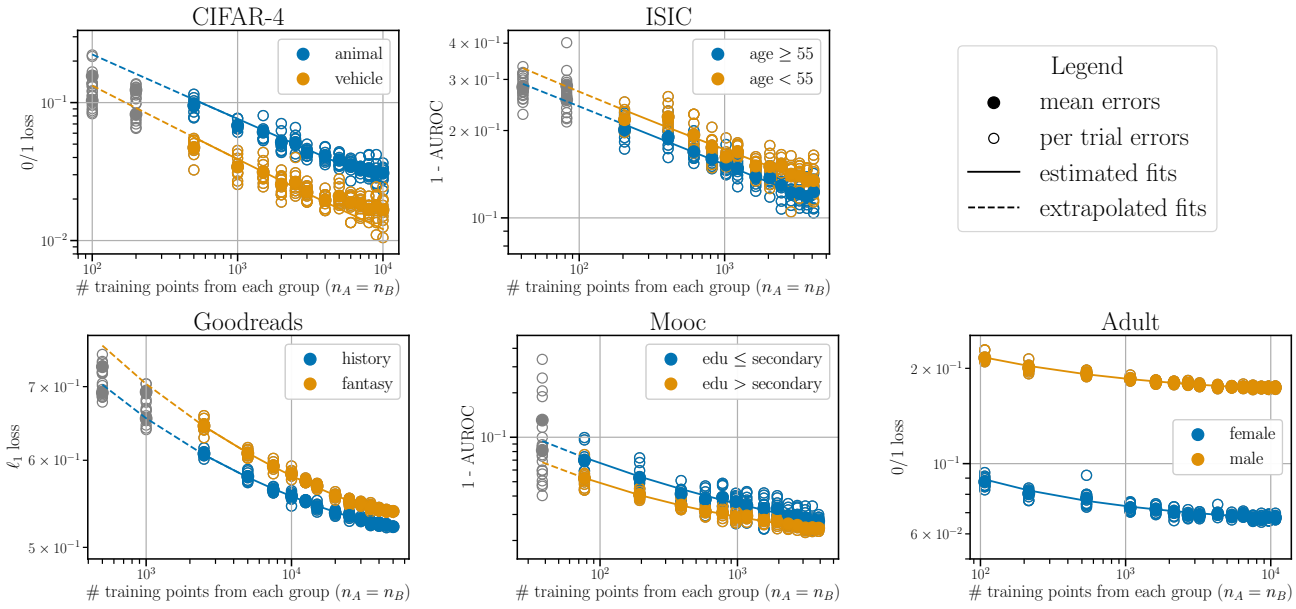


Figure B.2. Estimated scaling law fits describe observed trends of group errors as a function of (n_g, n). Grey points are not included in the scaling law fit, as $n_g < M_g$ (see Table 2).

configuration of (n_g, n). First, we choose an allocation ratio of group A to group B (from options $[0.125 : 1, 0.25 : 1, 0.5 : 1, 1 : 1]$). Then, we subsample again to x fraction of the largest subset size for this allocation ratio, for an $x \in [0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, .3, .4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ (we skip $x \in [0.01]$ for allocation ratios < 1). We repeat the sampling pattern for all pairs of allocation ratio and x, and again switching the roles of groups A and B in this procedure. This results in a set of 99 unique n_A, n_B pairs, which we combine with the 25 pairs from the previous experiment, where we fix $n = n_A + n_B$ (see Figure B.3). We run ten random seeds of (n_A, n_B) configuration, for a total of 1240 points from which we estimate the scaling law fits.

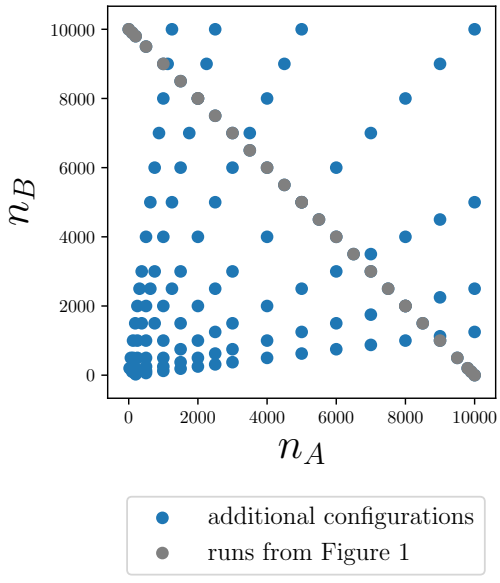


Figure B.3. Sample configurations for (n_g, n) described in Appendix B.5, shown for the CIFAR-4 dataset, where $\min_g n_g = 10,000$. Note that there are two groups, so $n = n_A + n_B$.

Since our main point of comparison is across training set sizes and allocations, rather than optimizing hyperparameters for each new sample size, we use the same hyperparameter configurations as for first set of experiments (ERM rows in Table B.3). For the neural net classifiers, we decrease the number of epochs when the total training set size n increases, so as to keep nearly the same number of gradient steps per training procedure. Specifically, we set the number of epochs as the nearest integer to (# epochs for first set of experiments) $\times$ (n for first set of experiments)/(n for current run) , where the first set of experiments corresponds to those in Figure 1.

Together with the results from the previous experiment shown in Figure 1, we use these values of (n, n_A, n_B) and the accuracies evaluated on groups A and B to estimate the parameters of the scaling laws in Eq. (7). We use the nonlinear least squares implementation of `scipy.optimize.curve_fit`. The estimated parameters are given in Table 2. The standard deviations reported in Table 2 are the estimated standard deviations resulting from the nonlinear least squares fits. Figure B.2 plots the fitted model over a subset of the data used to fit it, showing that the modified scaling model in Eq. (7) can express the trends in per-group losses as a function of n_g and n.

The parameter estimates sometimes have large standard deviations; we also found the parameter estimates to vary with M_g, the minimum number of points required to include a data point in the fitting procedure, as well as the overall subsetting design. When seeking exact and robust estimates of the parameters in Eq. (7), we suggest following the experimental procedures outlined in (Clauset et al., 2009), and additionally accounting for potential variation due to sampling patterns.

B.6. Pilot Sample Experiment

In this experiment, we take a small random sample from the Goodreads training set to instantiate a "pilot sample" from which to estimate scaling law parameters and suggest an $\hat{\alpha}^*$ at which to sample a larger training sample in a subsequent sampling effort. The pilot sample contains 5000 total instances, with $n_A = n_B = 2500$. With the pilot sample, we conduct a series of subsetting pairs for (n_A, n_B), similar to the procedure outlined in Appendix B.5, using relative allocations in $[0.0625{:}1, 0.125{:}1, 0.25{:}1, 0.5{:}1, 1{:}1]$ and $x \in [0.02, 0.05, 0.1, 0.15, 0.2, 0.25, .3, .4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. For each subset configuration, we sample 5 random seeds. From these results, we estimate the scaling parameters of Eq. (7) according to the performance of each subset configuration on the heldout evaluation set. We set the minimum number of points from which to fit the scaling parameters to $M_g = 250$.

Next, we use these estimated fits to extrapolate predicted per-group losses with more data points. For a given sample size, we suggest the $\hat{\alpha}^*$ that would minimize the the maximum of the estimated per-group losses. We use the training data held separate from the pilot sample to sample a training set at $\vec{\alpha} = \hat{\alpha}^*$ allocation, and evaluate performance on the original held out evaluation set. We follow the same procedure, sampling the new training set from $\vec{\alpha} = \vec{\gamma}$ and at $\vec{\alpha} = (0.5, 0.5)$.

As a point of comparison we also compute the results for all α in a grid of resolution 0.01; i.e., $\alpha \in [0.0, 0.01, 0.02, \ldots 0.99, 1.0]$ and denote the allocation value in this grid that minimizes the average maximum group loss over trials as α^*_{grid}. The best allocation possible (up to the 0.01 grid resolution) were: $\alpha^*_{\text{grid}} = 0.01$ for $n_{\text{new}} = 5000$, $\alpha^*_{\text{grid}} = 0.06$ for $n_{\text{new}} = 10000$, $\alpha^*_{\text{grid}} = 0.03$ for $n_{\text{new}} = 20000$, and $\alpha^*_{\text{grid}} = 0.07$ for $n_{\text{new}} = 40000$. The α^*_{grid} baseline helps contextualize performance of other allocation strategies with the optimal-in-hindsight α^*_{minmax}. Furthermore, the variability across trials due to subsetting around $\alpha = \alpha^*_{\text{grid}}$ is largely consistent with that of the other allocation sampling strategies examined.

We repeat this entire procedure (starting with generation of the pilot sample) for 10 random seeds, and the results are reported in Figure 2. Since we have enough training data in the Goodreads dataset outside of the pilot sample to simulate gathering larger datasets of up to $40,000$, we simulate collecting a new training dataset of size $n_{new} \in [5000, 10000, 20000, 40000]$, which are $[1\times, 2\times, 4\times, 8\times]$ the size of the pilot training dataset, respectively.

The error bars in Figure 2 are similarly large for all three allocation strategies we compare. The allocation $\vec{\alpha} = \hat{\alpha}^*_{\text{minmax}}$ when $n_{\text{new}} = 40000$ is an exception, indicating that high variation in $\hat{\alpha}^*_{\text{minmax}}$ may be an issue when n_{new} is large relative to the pilot training sample size.

B.7. Interactions Between Groups

While the main experiments exclude the SONIC study from the ISIC dataset, consistent with the 'non-patch' dataset in (Sohoni et al., 2020), this experiment utilizes the larger corpus of labeled images from the ISIC repository. The difference is the inclusion of the SONIC sub-study. This larger dataset has 21,203 instances after subsetting to precisely benign/malignant cases. The number of data instances coming from each dataset are given by the grey bars (black annotated numbers) in

Figure B.4; purple bars denote the number of malignant instances. Note the logarithmic scale.

We separate 16,965 images to the training set and the remaining 4,238 to the test set, with equal ratios of each sub-study represented in the train and test sets. The train and test splits differ for the study and sub-study analysis, accounting for differing values in corresponding cells of Figure B.5a and Figure B.5, particularly apparent when evaluating on JID Editorial Images, the study with the smallest number of data instances (see Figure B.4). We choose hyperparameters based off a 5 fold cross-validation procedure on the 16,965 training instances, searching over the same hyperparameter options as in Appendix B.3, ultimately using 20 epochs, momentum 0.9, weight decay 0.001 and learning rate 0.001 to tune the resnet-18 model (with ERM objective). For each of the 5 studies, and 10 sub-studies that comprise the larger datasets, we retrain the resnet-18 model on the training set with that sub-study excluded from the training data, and compare to performance of the model trained on the entire training set. As in Appendix B.5, we keep the number of gradient steps for each training process roughly equal. We run 10 random seeds for all conditions and report the differences in the mean performance metrics across groups. Figure B.5a (same as Figure 3) shows the results of subsetting by each of the five studies; Figure B.5b shows the results of subsetting by each of the ten sub-studies. We note that when the evaluation set is very small (as for the 2018 JID Ed. study), there can be high variation in estimated accuracy values due to randomization in the train/evaluation splits, highlighting the need for adequate representation across groups in evaluation data as well as in training data.

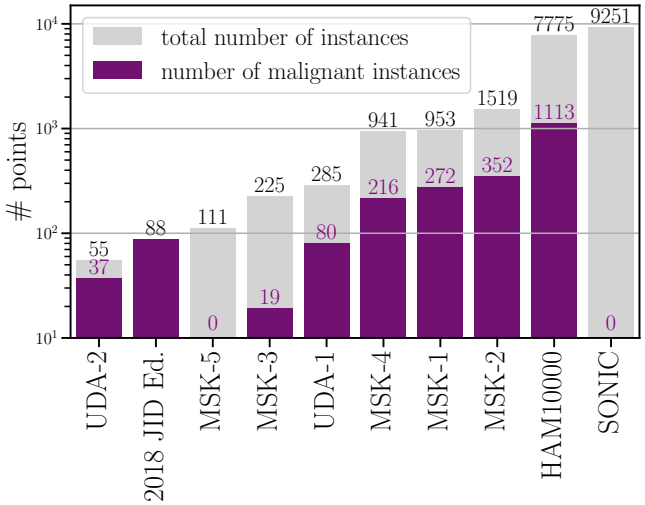


Figure B.4. Sub-studies that comprise the ISIC dataset. The SONIC study is excluded in the main analysis.

C. Supplementary Experiments

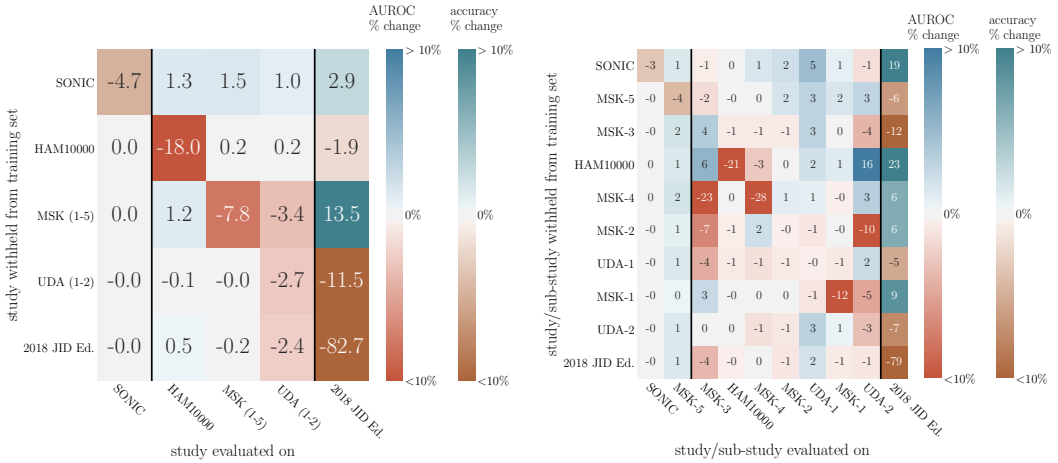
In this appendix we detail additional experiments to supplement the findings of Section 4. These experiments investigate the robustness of our findings to different problem formulations and data pre-processing.

C.1. The Effects of Including Groups as Features or Not

Here we compare models that use group or demographic information as features, and models that do not. We examine differences in group performance across $\vec{\alpha}$ and the effect of importance weighting in both cases.

For Mooc, we compare the model used in the main analysis with removing all demographic information including education level, which we group data instances by. There are five remaining features: number of active days, number of video plays, number of forum posts, and number of events, and course-id, which we one-hot encode. The random forest model is still the best performing of those we considered, though the optimal hyperparameters after a 5 fold cross validation search shifted to `num_trees` $= 400$ and `max_depth` $= 8$. For the Adult dataset, we add back in the features for sex, wife, and husband. This results in 13 unique features, and after one-hot encoding, 103 feature columns. After running a 5 fold cross validation search with the new feature matrix, the optimal model and parameters were `num_trees` $= 400$ and `max_depth` $= 16$.

Figure C.1 compares the results of after excluding (left column) and including (right column) this information in the training feature matrix. Models with the group features excluded generally exhibit less degradation in performance for α_A near 0 or



(a) Grouped by study.

(b) Grouped by sub-study.

Figure B.5. Performance changes due to holding out a study from the training set. Studies are ordered by % malignancy of the data in the evaluation set. SONIC and MSK-5 contain all benign instances and the 2018 JID Editorial Images dataset contains all malignant instances; for these we report % change in binary accuracy. For the remaining groups, we report % change in AUROC. Note that the random training/test splits differ between (a) and (b), accounting for the differences in values for corresponding cells between the two figures.

dataset	M_g	group g	$\hat{\sigma}_g$	$\hat{p}_g$	$\hat{\tau}_g$	$\hat{q}_g$	$\hat{\delta}_g$
Mooc (with dem. features)	50	edu $\leq 2°$	0.08 (2.6e-05)	0.14 (6.0e-03)	0.73 (0.059)	0.63 (4.8e-03)	1.3e-15 (2.6e-04)
		edu $> 2°$	0.038 (6.2e-04)	0.068 (6.3e-03)	0.54 (6.5e-03)	0.61 (9.8e-04)	2.8e-12 (8.0e-04)
Mooc (without dem. features)	50	edu $\leq 2°$	0.41 (0.86)	1.0 (0.26)	0.6 (0.028)	0.56 (3.5e-03)	0.029 (1.4e-06)
		edu $> 2°$	0.029 (1.3e-03)	0.055 (0.011)	0.33 (2.2e-03)	0.54 (9.5e-04)	1.9e-14 (1.5e-03)
Adult (with group features)	50	female	0.036 (3.2e-06)	0.14 (2.5e-03)	0.23 (3.3e-03)	0.47 (2.3e-03)	0.055 (2.1e-05)
		male	0.12 (6.8e-05)	0.24 (5.8e-04)	0.3 (6.2e-03)	0.47 (2.7e-03)	0.16 (3.5e-06)
Adult (without group features)	50	female	0.078 (0.051)	0.018 (3.6e-03)	0.43 (8.3e-03)	0.59 (1.6e-03)	8.0e-16 (0.052)
		male	0.066 (2.6e-05)	0.21 (1.2e-03)	0.47 (6.5e-03)	0.50 (1.1e-03)	0.16 (5.4e-06)

Table C.4. Estimated scaling parameters for eq. (7) with and without demographic features included. Parentheses denote standard deviations estimated by the nonlinear least squares fit. Parameters are constrained so that $\hat{\tau}_g, \hat{\sigma}_g \geq 0$ and $\hat{p}_g, \hat{q}_g \in [0, 2]$.

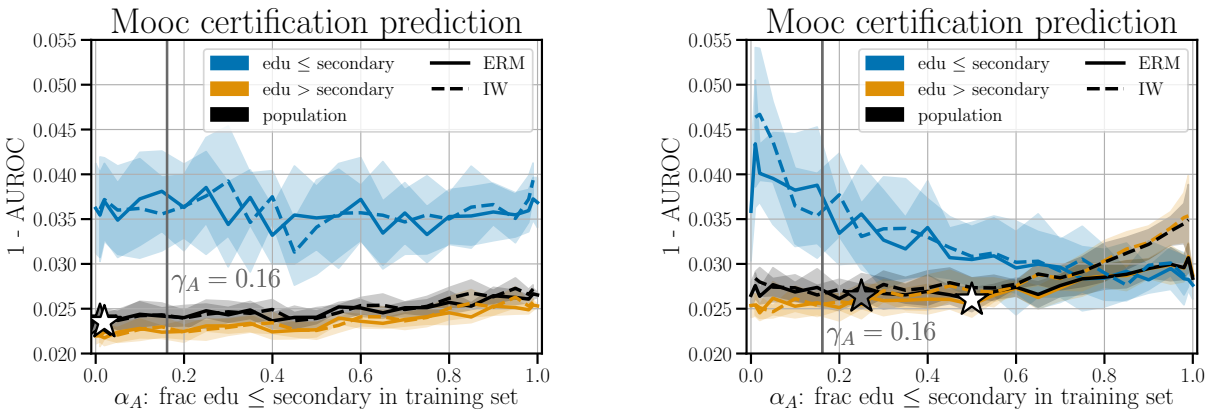
1. The negative impacts of IW are lessened slightly when the models do not incorporate groups as features.

The different scaling law fits are shown in Table C.4. For the Adult task without group features, the estimated exponent and scaling constant on group-agnostic data, $\hat{q}_g$ and $\hat{\tau}_g$, are larger than than for the model with demographic features. For the Mooc task, the effect is the reverse. In all settings, the fitted parameters show that n influences group errors more than n_g, so long as n_g is at least $M_g = 50$ (the region for which we estimate the fit).

In the main analysis, we include demographic features for the Mooc certification prediction task, as this increases accuracy of the "edu $\leq$ secondary group," and exclude gender features from the Adult income prediction, as it does not not greatly impact performance for α_A near 0.5, and improves performance for more extreme settings of α_A (Figure C.1). For the other three datasets we study, there is no singular feature corresponding to the group that we can hold out in a similar fashion.

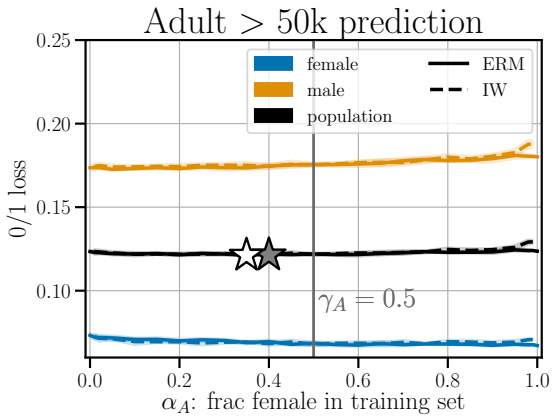
C.2. Trends across models and performance metrics

Figure C.2 contrasts the trade offs we can navigate with different allocations and re-weighting schemes for different models and accuracy metrics. The left hand side of Figure C.2 shows the result of the model fitting and evaluation procedure to minimize ℓ_1 loss (MAE) of the Goodreads predictions (same as the results shown in Figure 1). The right hand side shows the result of the same model fitting and evaluation procedure applied to minimizing the ℓ_2 loss (MSE), where ridge regression is the best performing model of those we consider. The ridge model evaluated with MSE exhibits similar trends across $\vec{\alpha}$

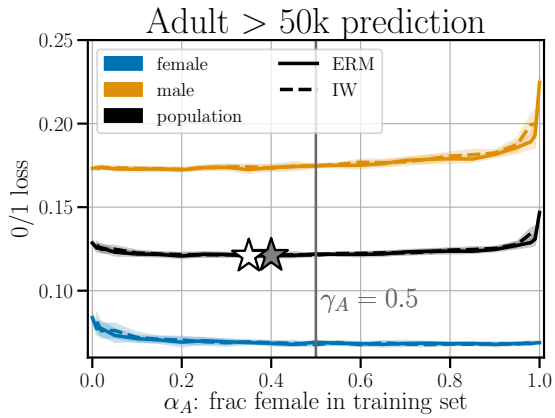


(a) Mooc dataset with demographic features excluded.

(b) Mooc dataset with demographic features included.



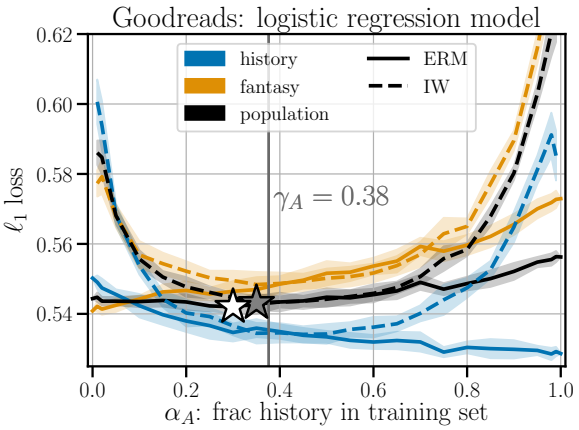
(c) Adult dataset with demographic features excluded.



(d) Adult dataset with demographic features included.

Figure C.1. Comparison of phenomena with group information included or excluded from training set.

values as the multiclass logistic regression model evaluated on MAE. The value α_A^* that minimizes population loss is similar for both methods – near to γ_A, though on opposite sides of γ_A. The degradation in population loss for α_A close to 0 or 1 is also apparent for the ridge model, and the range of $\vec{\alpha}$ for which importance weighting (IW) increases per-group and population losses compared to empirical risk minimization (ERM) is similar across models.



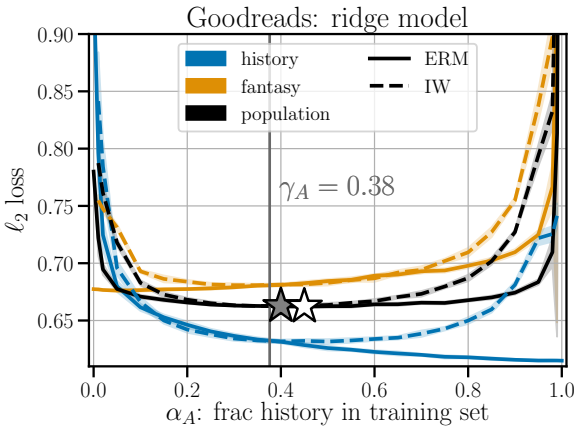


Figure C.2. Comparison of phenomena for different accuracy metrics and models on the Goodreads dataset.