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.

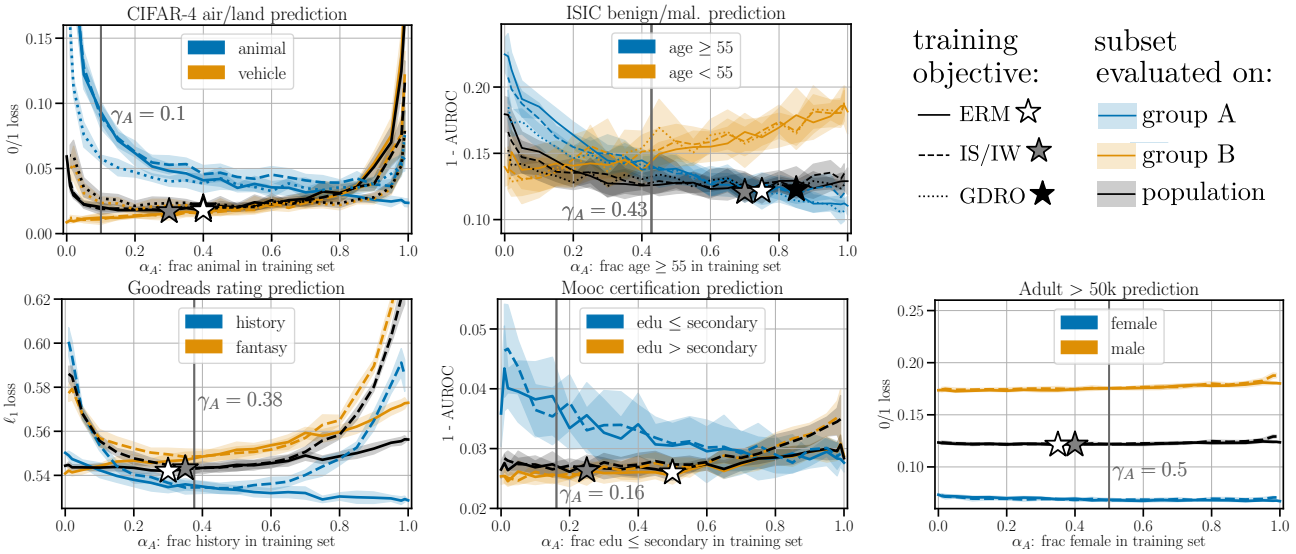

Figure 1. Performance across $\vec{\alpha}$. Shaded regions: one stddev. above/ below mean (10 trials). Stars: population minima for each objective. The loss metrics reported (vertical axes) are the same within panels, while the training objectives differ across solid and dashed lines.

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).

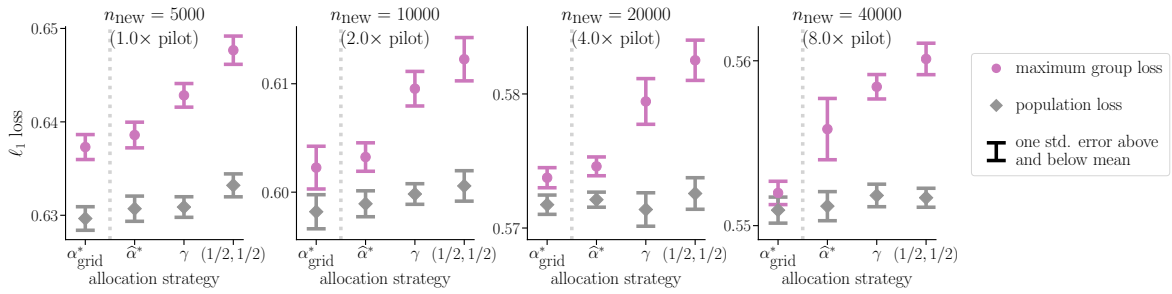

Figure 2. Pilot sample experiment. Panels show the result of the three allocations $\vec{\alpha} \in [\hat{\alpha}^*_{\text{minmax}}, \vec{\gamma}, (1/2, 1/2)]$ for different sizes of the new training sets compared with an α^*_{grid} baseline that minimizes the maximum group loss over a grid of resolution 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

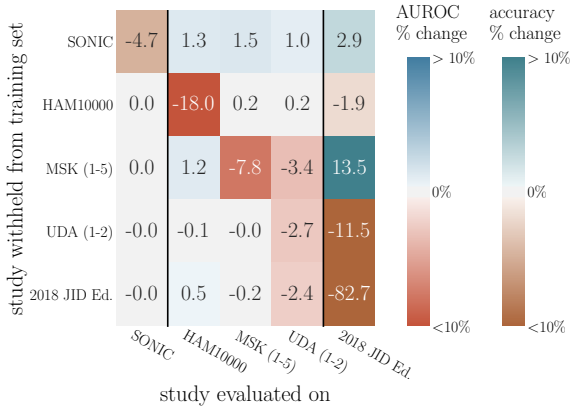

Figure 3. Percent change in performance (AUROC / accuracy) due to withholding a study from the training set. SONIC contains only benign instances and 2018 JID Ed. contains only malignant instances; for these we report % change in binary accuracy.

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.