SIMPLICITY BIAS LEADS TO AMPLIFIED PERFORMANCE DISPARITIES

Anonymous authors
Paper under double-blind review

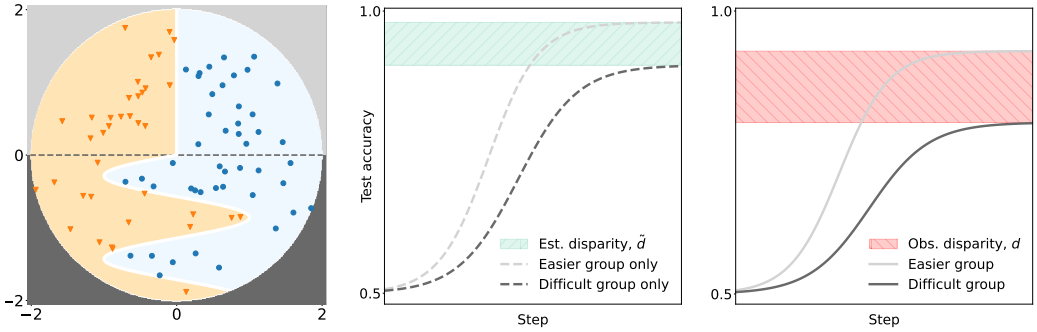
ABSTRACT

The simple idea that not all things are equally difficult has surprising implications when applied in a fairness context. In this work we explore how "difficulty" is model-specific, such that different models find different parts of a dataset challenging. When difficulty correlates with group information, we term this *difficulty disparity*. Drawing a connection with recent work exploring the inductive bias towards simplicity of SGD-trained models, we show that when such a disparity exists, it is further *amplified* by commonly-used models. We quantify this *amplification factor* across a range of settings aiming towards a fuller understanding of the role of model bias. We also present a challenge to the simplifying assumption that "fixing" a dataset is sufficient to ensure unbiased performance.

1 INTRODUCTION

Without actually training, understanding *what* a model will find challenging is far from trivial. A certain dataset may be hard for one model but not for another (Wolpert & Macready, 1997). For a given model, two classes may be easily separable, while for another they may be hard to distinguish. Given this, it follows naturally that "difficulty" is a function of both *data and model*, such that we can't properly account for difficulty by analyzing the dataset alone.

In the context of fairness in machine learning, for a given task, a data-model pair may be more difficult for one social group than another, leading to disparate impact (Barocas & Selbst, 2016). For example, Buolamwini & Gebru's 2018 audit of commercial image recognition systems finds



(a) Example distribution (b) Train each group separately (c) Train both groups together

Figure 1: **What is difficulty amplification?** **(a)** Consider a binary classification of circles and triangles. Above $y = 0$ (light gray background) we have a simple group which is linearly separable. Below $y = 0$ (dark gray) we have a more complex group with a non-linear decision boundary. **(b)** Illustration of test accuracy when training on the simple group only (light gray) and the complex group only (dark gray). As expected we obtain better accuracy on the simple group. **(c)** However, when training on both groups at once, our model exacerbates the difference: the observed accuracy disparity d (height of pink area) exceeds the estimated accuracy disparity from individual group training $\tilde{d}$ (height of green area). When $d > \tilde{d}$, we call this **difficulty amplification**.

they exhibit worse accuracy for darker-skinned women than for any other group. Typically, accuracy disparity of this kind is attributed to either under-representation of certain groups or spurious correlations between group information and the variable of interest. In this work, we further show that—even with perfectly balanced data and in the absence of correlations between group labels and class labels—trained models can find certain groups harder than others. Crucially, group difficulty is not always predictable from a dataset audit, providing key evidence for the necessity of a complementary post-training model audit.

Having identified model-specific disparities in the post-dataset-audit setting, we turn to the role of the model itself. We show that implicit bias of certain model classes towards simple functions (Arpit et al., 2017; Kalimeris et al., 2019; Rahaman et al., 2019; Valle-Perez et al., 2019; Shah et al., 2020) further amplifies disparity: When a model finds one group easier than another, its bias towards the easy group leads to greater than expected performance disparity after training (see fig. 1). We show that *difficulty amplification* is highly sensitive to model architecture, training time, and parameter count. Seemingly innocuous design decisions, such as whether to use early stopping, can have a significant impact on the amount of amplification and consequently the performance disparity.

Contributions

1. We identify **difficulty disparity**, a pervasive phenomenon that persists in the post-dataset-audit setting: using data with perfect representation and without spurious correlations.
2. We introduce **difficulty amplification factor** to quantify how much a model exacerbates difficulty disparity.
3. We **empirically evaluate** how choices including model architecture, training time, and parameter count impact difficulty amplification.

Paper structure. In § 2 we provide background and related work. We begin in § 3 by evaluating variability of dataset difficulty across models. In § 4 we formalize difficulty disparity and amplification and design a synthetic task to isolate it. Subsequently, in § 5 we show how factors such as model architecture, scale, training time and regularization impact amplification. After a real-world example using Dollar Street in § 6, we discuss the fairness implication of our work in § 7 before concluding in § 8.

2 Background and related work

2.1 Biased datasets and biased models

Bias in ML systems arises from many sources. At the most basic level, a dataset itself is biased if certain groups are under-represented (Stock & Cisse, 2018; Hendricks et al., 2018; Yang et al., 2020; Menon et al., 2021). Proposals to rectify under-representation include actively collecting more data for marginalized groups (Dutta et al., 2020), under/oversampling or reweighting (Byrd & Lipton, 2019; Sagawa et al., 2020; Idrissi et al., 2022; Arjovsky et al., 2022) during training, and optimizing for worst-group (as opposed to average) accuracy (Sagawa et al., 2019). Other recent work has suggested fine-tuning on an explicitly balanced set (Kirichenko et al., 2022).

Alternatively, datasets can reinforce harmful associations (Goyal et al., 2022b), both due to sampling error and by inadvertently capturing an undesirable association that is present in society. Bucketed as "spurious correlations" (Muthukumar et al., 2018; Wang et al., 2019; Sagawa et al., 2020) or "shortcuts" (Geirhos et al., 2020), a large body of fairness work seeks to train models that learn some true function invariant to the spuriously correlated feature. Both under-representation and spurious correlations dominate the fairness literature landscape, though in both cases the onus is squarely on the data.

2.2 Beyond spurious correlations

There is increasing focus on the model itself, independent of the role of data (Hooker, 2021), of which *bias amplification* is a perfect example (Zhao et al., 2017; Wang & Russakovsky, 2021). Here, a small correlation in the training set is amplified into a larger correlation at test time. Models don't just replicate the bias in the data, but exacerbate it. In empirical experiments evaluating bias

amplification, Hall et al. (2022) suggest that if group membership is easier to identify than class membership, models prefer to use the spurious correlation. Hall et al. report no bias amplification in the case where there is no spurious correlation present in the data: this is expected given the definition of amplification involves a multiplication of an existing data bias.

Other subtle biases have been identified in the balanced data setting. Leino et al. (2019) show that models trained with SGD overly rely upon moderately spuriously-correlated features if they are sufficiently numerous relative to the size of training set. Khani & Liang (2020) find that adding feature noise equally across groups induces disparity, a fact that can also be attributed to the relative difficulty of group information versus the desired target. Khani & Liang (2021) find removing spurious features can disproportionately lower performance on certain groups, and argue (as do we) that a balanced dataset is not a sufficient guard against biased performance. Mannelli et al. (2022) use teacher-student networks to show subtle properties such as differences in group distance from the overall mean and differences in group variance are sufficient to induce biased outcomes in the absence of spurious correlations.

2.3 Inductive bias towards simplicity

Numerous recent works have identified the tendency for SGD-trained models to prioritize simple data points during training, resulting in simple functions being learned before more complex ones (Arpit et al., 2017; Kalimeris et al., 2019; Valle-Perez et al., 2019). Jo & Bengio (2017) show that convolutional networks are overly dependent on surface-level statistical properties of images, such that applying a Fourier filter to the training set is sufficient to radically degrade test performance. Rahaman et al. (2019) also show a bias towards low-frequency functions, learning these simpler functions before more complex, higher-frequency examples. Though often framed as a positive, allowing neural networks to learn functions that generalize well by applying Occam's razor, Shah et al. (2020) cites this simplicity bias as potentially harmful, at root causing both vulnerability to adversarial attacks and over-reliance on spurious correlations. Dagaev et al. (2021) reserve as much skepticism for overly simple solutions as for overly complicated, arguing that excessively simple solutions are likely to rely on potentially harmful "shortcuts" (Geirhos et al., 2020). Sagawa et al. (2020) find that increasing model size yet further pushes the model to rely upon spuriously-correlated features where they carry more signal than the intended features. In our work, we show how this inductive bias toward simplicity manifests as disparate outcomes when there is a perceived complexity difference between groups.

3 Prelude: Data difficulty is model-specific

This work is predicated on a simple, perhaps even obvious idea: that certain facets of a dataset may be more or less difficult to a given model. Whether individual datapoints, classes, or social groups, identifying what will be difficult to a model isn't always apparent ahead of time, nor need it align with human intuition. As an illustration, we begin by investigating difficulty difference *between classes*, though in following sections we will transition to a focus on social groups.

Class difficulty. We measure *difficulty* using cross-validated test accuracy over the full dataset, i.e. after merging the source train and test sets.[1] We train a classifier on CIFAR-100, after which we evaluate binary classification accuracy on each pair of (coarse-grained) labels and extract pairwise binary accuracy by masking irrelevant outputs before the softmax layer.

Figure 2a shows how a ResNet-18 (He et al., 2016) trained on the 20 coarse-grained classes of CIFAR-100 (Krizhevsky et al., 2009) yields variable test accuracy between classes. For example, this model achieves near perfect test accuracy on *flowers / aquatic mammals*, and substantially lower performance on *non-insect invertebrates / insects*. Crucially, this view of difficulty is just that of this particular ensemble of ResNets. Were we to perform this difficulty evaluation on a different model, our pairwise matrix would differ.

Model specificity. To confirm this, we repeat the experiment using the following models: an SVM with an RBF kernel; LeNet, a simple CNN (Lecun et al., 1998); a 3-layer and a 5-layer fully-connected network; and a fully-connected layer over pre-trained representations extracted from

[1] We merge train and test to maximize available data, and use cross-validation to control for covariate shift.

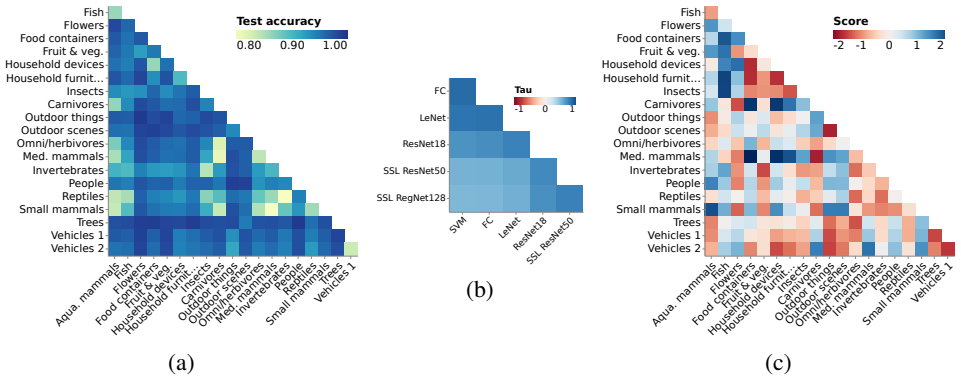


Figure 2: **(a)** ResNet-18 finds certain classes of CIFAR-100 more difficult than others, according to binary test accuracy. **(b)** Rank ordering of difficulty varies across models, measured here by Kendall's τ. **(c)** 1D PLS projection of binary accuracies onto cosine distance between class mean vectors. Red cells—pairs where increasing inter-class distance decreases accuracy—demonstrate **there is no clear way to identify what a model will find difficult without training.**

ResNet-50 trained using SimCLR (Chen et al., 2020) on ImageNet-1K (Russakovsky et al., 2015), and those of RegNet-128Gf (Radosavovic et al., 2020) trained with SwAV (Caron et al., 2020) on 1 billion public images from Instagram (Goyal et al., 2022a) (see appendix B for a full description).

Figure 2b shows how the rank order correlation of pairwise difficulties varies between models, as measured by Kendall's τ. Specifically, between the 3-layer FC model and LeNet, there is high (but not perfect) τ, indicating that broadly what LeNet finds difficult so too does the FC network. In contrast, between the RBF SVM and a linear layer on RegNet-128Gf representations, there is much lower (though not zero) τ, indicating that the pairwise ordering varies considerably.

PLS analysis. Figure 2c further reinforces this notion. We use partial least squares (PLS) to search for a low-dimensional projection from the models' difficulty ranks to a *data-only* measure of difficulty: the rank of the cosine distance between class means (see appendix A). If model difficulty was purely a function of data difficulty, we would expect PLS to find a well fitting linear regression model between model and data. Instead, we find further evidence to the contrary. From the 1D projection, it should be immediately apparent that for some classes (in blue, e.g. *carnivores* / *food containers*), increasing inter-class distance tends to increases binary accuracy, though for many class pairs (in red, *outdoor scenes* / *outdoor things*) the opposite is true. Crucially, PLS struggles to find a good fit between model ranks and the data distance ranks, with a near-zero fit $R^2 = 0.058$.

Summary. *There is no clear difficulty ordering that is consistent between all model classes. What a model finds difficult is not solely a function of the data.*

4 Neural networks prioritize "easy"

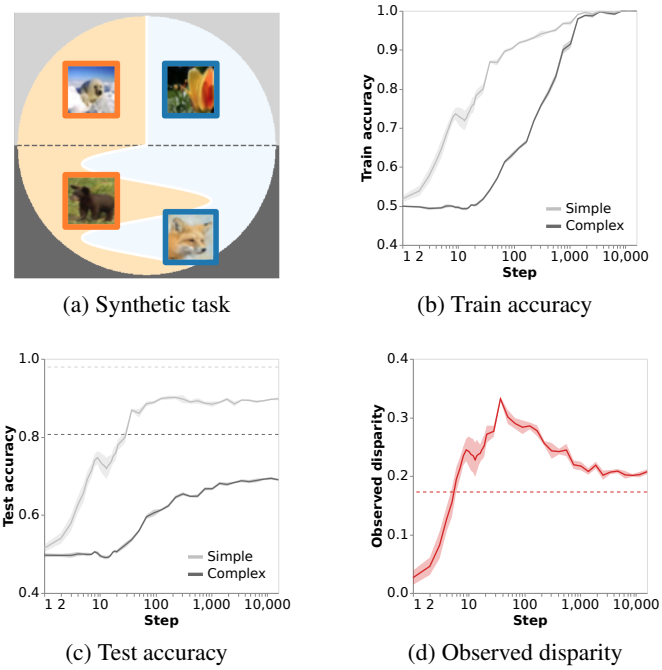
Having established that difficulty is a function of both model and data, we now measure how difficuly affects performance disparity and quantify the effect of the simplicity bias.

4.1 Definitions

Let $\mathrm{acc}(X, \mathbf{y}, \mathcal{M})$ be the cross-validated test accuracy on the classification dataset $(X, \mathbf{y})$ of a model $\mathcal{M}$, and $\mathcal{M}_{X,\mathbf{y}}$ be a model trained (with cross-validation) on $(X, \mathbf{y})$. Given two groups α and β, let $(X_\alpha, \mathbf{y}_\alpha)$ and $(X_\beta, \mathbf{y}_\beta)$ denote corresponding slices of the dataset, such that a model trained only on group α is $\mathcal{M}_{X_\alpha,\mathbf{y}_\alpha}$.

Estimated difficulty disparity. First, we define the estimated difficulty disparity $\tilde{d}$ as the difference in accuracy between a model trained and evaluated on each group in isolation,

$$\tilde{d} = \left|\mathrm{acc}(X_\alpha, \mathbf{y}_\alpha, \mathcal{M}_{X_\alpha,\mathbf{y}_\alpha}) - \mathrm{acc}(X_\beta, \mathbf{y}_\beta, \mathcal{M}_{X_\beta,\mathbf{y}_\beta})\right| . \tag{1}$$



(a) Synthetic task (b) Train accuracy

(c) Test accuracy (d) Observed disparity

Figure 3: Accuracy and disparity for ResNet-18. **(a)** Binary classification (orange / blue) comprising two groups. Easy group (light gray) is high-accuracy pair *flowers* / *aquatic mammal*. Difficult group (dark gray) is low-accuracy pair *medium mammals* / *carnivores*. **(b)** Train accuracy for complex group is learned slower but both groups reach perfect train accuracy. **(c)** However, test accuracy for complex group is persistently lower. Dashed lines are binary accuracy from single-group training. **(d)** Observed disparity d peaks early in training but large gap persists after training. Red dashed line is estimated disparity $\tilde{d}$: observed disparity above this line indicates *amplification*. Shaded area is standard error over 10 runs each with different train/test split.

Observed difficulty disparity. Second, the observed difficulty disparity is the difference in accuracy between groups on a model trained on both groups,

$$d = |\mathrm{acc}(X_\alpha, \mathbf{y}_\alpha, \mathcal{M}_{X,\mathbf{y}}) - \mathrm{acc}(X_\beta, \mathbf{y}_\beta, \mathcal{M}_{X,\mathbf{y}})| \,. \tag{2}$$

Difficulty amplification. Finally, if the model trained on both groups exhibits worse disparity than when trained in isolation, $d > \tilde{d}$, we say that the model exhibits difficulty amplification. Over many model runs, groups, or samples from the dataset we can define an amplification factor $k = d/\tilde{d}$.

In practice, calculating amplification is a two-stage process. First, we train N randomly-initialized models on each group in isolation, and compute the average cross-validated test accuracy. Between each pair of groups, we calculate the estimated difficulty disparity $\tilde{d}$. Second, we train a new set of N models on the full dataset including all groups, and compute average test accuracy broken out by group. For each group pair, we calculate the observed difficulty disparity d.

4.2 Simulating difficulty disparity with CIFAR-100

To measure difficulty disparity in a controlled setting, we design a task based on CIFAR-100 that is group-balanced and absent correlations between group labels and target labels. We extract the binary test accuracies for each pair of coarse classes and treat their pairwise differences as estimated difficulty disparity $\tilde{d}$. We let group α be the class pair y_0^α, y_1^α with the highest accuracy and β be the lowest y_0^β, y_1^β. To simulate a binary classification task with two differently-difficult groups, we stitch these pairs together into a single binary task, where $y_0 = \{y_0^\alpha, y_0^\beta\}, y_1 = \{y_1^\alpha, y_1^\beta\}$. See fig. 3a. Finally, we train N models on this task and calculate observed difficulty disparity d.

4.3 RESULTS

In fig. 3b we see the training accuracy of the simple group α improves much more rapidly than the complex β, though both groups reach perfect train accuracy eventually. In contrast, the test accuracies in fig. 3c remain notably different at convergence, with the model displaying lower accuracy on the complex group. This gap is the observed accuracy disparity shown in Figure 3d. Here, we observe that observed disparity peaks after just a few steps, before a slight decline to a plateau. Evidencing our difficulty amplification claim, observed disparity remains higher than estimated disparity we would have expected from separate training (the red dashed line).

Summary. *This simple experiment shows that models trained across groups with different difficulty do prioritize the simpler group, leading to an outsized observed disparity, primarily driven by underperformance on the more difficult group.*

5 AMPLIFICATION FACTOR VARIES ACROSS MODELS

To compute an amplification factor, we repeat the above experiment using different pairs of classes with different estimated disparity, and compute their observed disparity after combined training. We retrain on 30 sampled pairs of label pairs, recomputing both estimated and observed difficulty disparity, and apply OLS linear regression to estimate the *amplification factor* (see appendix D for details). This method can easily be applied to any dataset annotated with group information, by replacing the sampling of pairs of classes with the sampling of different group combinations. We compute k for each model listed in § 3.

Furthermore, we evaluate the effect of model scale on amplification factor by varying the width of ResNet-18; evaluate various settings of weight decay; and evaluate the role of early stopping by computing amplification through training.

For an additional test of the effect of the simplicity bias, we also try to push the model to choose a more complex solution. We enforce a Lipschitz constraint by applying a penalty on the norm of the gradients, a technique commonly used to stabilize discriminator training in GANs (Gulrajani et al., 2017). We add the following penalty term to our loss function L,

$$L' = L + \lambda(||\nabla_{\mathbf{x}} f(\mathbf{x})||_2 - C)^2 \ , \tag{3}$$

where $\nabla_{\mathbf{x}} f(\mathbf{x})$ is the gradient of the network's outputs with respect to its inputs, the penalty coefficient $\lambda = 10$, and C determines the Lipschitz constraint: a low C pushes the model towards simpler functions, and high C towards more complex.

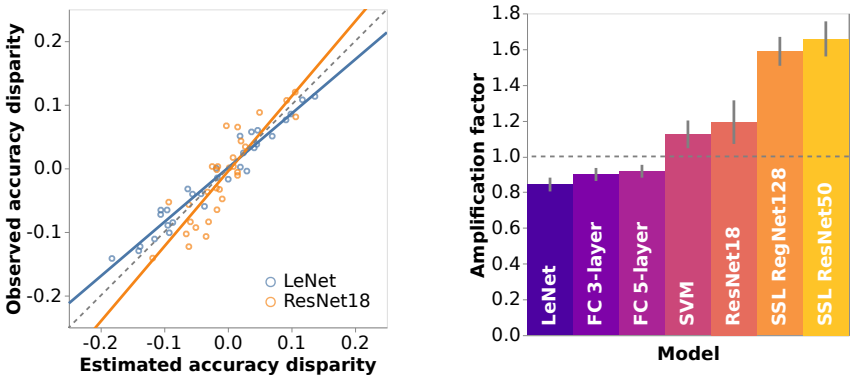
5.1 RESULTS

Model architecture. In fig. 4a, we illustrate the difference in amplification factor between two models, LeNet and ResNet-18. We find that the ResNet-18 amplifies disparity by a factor of $k = 1.19 \pm 0.12$. In contrast, LeNet diminishes disparity ($k = 0.84 \pm 0.04$), resulting in an observed disparity lower than expected. Thus, from this simple example we show that model choice influences difficulty amplification.

Across the full suite of models (fig. 4b) we again see significant variation in amplification factors across the different models, with the simpler models (FC, LeNet) attenuating and the larger models (ResNet, etc.) increasing amplification. *However*, the simpler models all exhibit poor test accuracy averaged over the entire dataset (see fig. S1), offering a candidate explanation for the lack of amplification. These models may be too simple to learn the dataset at all, resulting in equally poor performance across all groups.

Width. However, within a specific architecture, increasing width seems to *reduce* amplification, in contrast with our expectations. Figure 5a shows the amplification factor for ResNet-18 rapidly decreasing to almost 1 (no amplification) as network width increases.

Early stopping. As training proceeds (fig. 5a), k increases to a peak around 1.2 early in training, before decreasing to a plateau a little over 1. This highlights the important role of early stopping in amplifying disparity, particularly in light of prior work arguing that models learn more simple functions earlier in training Kalimeris et al. (2019).



(a) Observed vs. estimated disparity

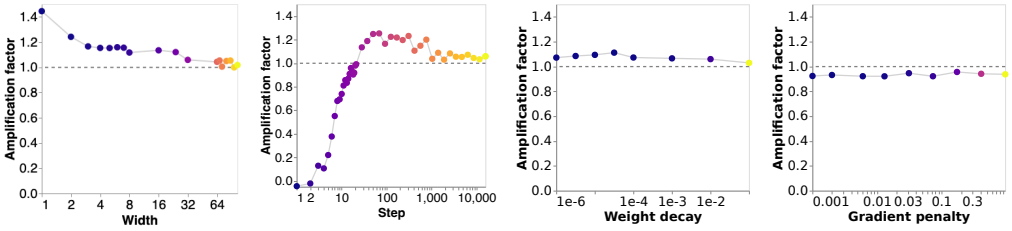
(b) Amplification factor by model

Figure 4: Observed difficulty disparity as a function of estimated difficulty disparity. **(a) ResNet-18 amplifies difficulty disparity but LeNet attenuates.** Points are sampled tasks with varying estimated difficulty disparities. Solid line is linear regression. Dashed grey line is amplification factor $k = 1$. **(b) Model architecture impacts difficulty amplification factor.** See fig. S2 for R^2 and fig. S1 for raw test accuracies.

Weight decay. Figure 5c shows next to no effect of weight decay across all values considered.

Gradient penalty. In contrast, applying a penalty to the norm of the gradients, rather than the parameters, is sufficient to lower k to below 1 for all values of C considered here. This suggests that applying a gradient penalty to balance out the implicit bias towards simplicity may be a helpful strategy in combating difficulty disparity.

Summary. *High-performing models—those optimized for average test accuracy—consistently display difficulty amplification. This phenomenon is exacerbated by early stopping, but may be reduced using a gradient penalty.*



(a) Width (b) Step (c) Weight decay (d) Gradient penalty

Figure 5: Design decisions influence amplification factor. Solid colored circles are amplification factor k. All panels are a ResNet-18 on 30 sampled synthetic tasks based on CIFAR-100 (see fig. 3a). **(a)** Network width has a strong negative effect on amplification, reducing it to near 1. **(b)** Through training, k peaks early in training before plateauing just above 1. **(c)** Weight decay has no effect on amplification. **(d)** Applying a gradient penalty to bias the model toward a C-Lipschitz functions works, across all values of C. See fig. S3 for R^2.

6 DIFFICULTY AMPLIFICATION HAS REAL-WORLD IMPACT

We now evaluate difficulty disparity and amplification in a real-world setting. Dollar Street[2] is a dataset of geographically-diverse images spanning a broad range of household incomes. In our setting, we use the labels associated with each image in a 138-class object classification task, where group information is household income quartile. As above, we evaluate a model trained on the lowest income quartile only, the highest income quartile only, and a model trained on both lowest and highest together. We train a single-layer FC network on representations extracted from ResNet-18 pretrained on ImageNet-1K. See appendix C for further training details.

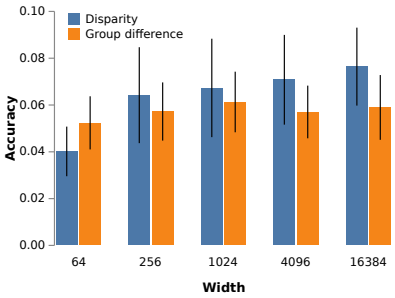


Figure 6: Estimated and observed disparity between lowest and highest income quartiles on the Dollar Street dataset of geographically-diverse images. Except for the smallest network, observed disparity d is higher than estimated disparity $\tilde{d}$, indicating amplification. Error bars are standard error.

Throughout this work, we have endeavored to remove bias resulting from group imbalance and label/group correlation. However, in this example we introduce an additional possible source of bias via ImageNet-1K pretraining. ImageNet-1K significantly under-represents many social groups (Dulhanty & Wong, 2019) and geographies (Shankar et al., 2017; de Vries et al., 2019), and exhibits harmful associations between race and certain class labels (Stock & Cisse, 2018). Geographic under-representation is a plausible reason for income-quartile difficulty disparity. This, however, cannot explain difficulty amplification.

In fig. 6 we show both estimated disparity and observed disparity on Dollar Street, for models with various widths. From this figure it is clear that observed disparity exceeds estimated disparity in all but the smallest network: Pre-trained representations are not necessarily immune from difficulty amplification.

Summary. *Example of difficulty disparity in social context: Preference for the simpler group—in this case the high-income quartile—is further amplified by most models.*

7 DISCUSSION

Auditing for bias. At its heart, our work presents yet another way in which models exhibit bias and performance disparities across demographics. A frequent refrain in the ML community is that such disparities are the fault of the data, rather than algorithmic bias (Hooker, 2021). Indeed, a series of thorough audits have revealed that popular datasets under-represent minoritized groups (Shankar et al., 2017; Stock & Cisse, 2018; Buolamwini & Gebru, 2018; de Vries et al., 2019; Dulhanty & Wong, 2019; Wilson et al., 2019); reify harmful associations and perpetuate stereotypes (Bolukbasi et al., 2016; van Miltenburg, 2016; Garg et al., 2018; Dixon et al., 2018; Birhane & Prabhu, 2021; Raji & Fried, 2021); and operationalize concepts such as gender and race in a way that applies a veneer of "objectivity" to socially-constructed and culturally specific concepts (Keyes, 2018; Paullada et al., 2021; Denton et al., 2021; Raji et al., 2021). Fixing these issues at the level of the data may not even be possible, for example it is often undesirable to collect the demographic information needed to ensure balance in the first place (Veale & Binns, 2017; Andrus et al., 2021; Hooker, 2021). That being said, acknowledging issues with our use of data does not absolve all that comes after, as exemplified by bias amplification (Zhao et al., 2017; Wang & Russakovsky, 2021; Hall et al., 2022). Here, in support of the role of *post-training* audit, we choose the setting where the data is "perfect", in that it is both explicitly balanced, and groups and labels are decorrelated. The variability of both difficulty disparity and amplification from model to model is a strong reminder that both those who develop and deploy ML systems must take action to ensure their fairness.

Fairness definitions. By discussing issues of bias and disparity, we naturally enter into a broader discussion about fairness in ML systems. Here, we follow others in focusing on the performance

[2] https://www.gapminder.org/dollar-street

gap between groups (Dwork et al., 2012; Hardt et al., 2016; Woodworth et al., 2017; Agarwal et al., 2018; Khani et al., 2019; Goyal et al., 2022b), though an alternative approach would be to focus explicitly on worst-group performance instead (Mohri et al., 2019; Sagawa et al., 2019; Zhang et al., 2020). Others rely upon counterfactual fairness (Kusner et al., 2017; Kilbertus et al., 2017; Loftus et al., 2018), according to which a "fair" system reaches the same decision on two otherwise identical individuals belonging to different protected groups, though this draws increasing criticism due to its requirement that concepts such as race or gender both be well-defined (Benthall & Haynes, 2019) and can be changed while only *minimally* impacting other attributes (Hu & Kohler-Hausmann, 2020; Hanna et al., 2020). Our aim in this work is not to use a metric by which to deem systems fair or unfair, but to highlight the possible role of model bias—in this case, due to preference for simplicity—that will have subsequent fairness impacts. That being said, it could be suggested that we believe a system absent of this disparity would be somehow fair. This is *not at all* the case: disparity between groups is only one possible metric among many alternatives, each of which may bring new issues to light (Lum et al., 2022). Even assuming a satisfactory yardstick by which to measure, and a model accordingly deemed fair, fairness is of course not necessarily implied. When situated within a broader societal context, any model can be put to harmful use, and it is a common pitfall of the ML community to narrowly situate our work inside neatly-defined abstractions (Selbst et al., 2019).

Spurious correlations. Similarly, a key ambition of our work is to push research into sources of bias outside of the typical characterizations: spurious correlations and under-representation. Indeed we suggest that reducing the study of model bias to these two dimensions is an instance of excessive abstraction through formalization (Selbst et al., 2019). By focusing on the settings where these issues are resolved, we hope that future research can take a more nuanced look at the biased behavior of models where not *obviously* the result of a data issue. A plausible outcome of this kind of research could be that in certain situations ML might not be appropriate *at all*, if we can't guarantee that the system won't develop unpredictable and hidden biases.

Measuring difficulty. In this work we *choose* to measure difficulty using cross-validated test accuracy, averaged over all samples in a group or class. Recent works have investigated alternative methods for quantifying model-specific example difficulty, including loss (Arazo et al., 2019; Han et al., 2018) and prediction disagreement between models (Simsek et al., 2022), mini-batches (Chang et al., 2017), and throughout training (Toneva et al., 2019; Swayamdipta et al., 2020). Hooker (2021) identifies samples that are often forgotten after compression. Applying these sample-level measures to evaluating group-level difficulty disparity remains an interesting future direction.

8 Conclusion

We have argued that what a model finds difficult is not simply a function of the data, but a function of both model and dataset. This is particularly a problem in a fairness context if difficulty is correlated with group information. We have found that certain models further amplify difficulty disparity, resulting in observed difficulty disparity over and above estimated difficulty disparity, as a result of the bias of certain models towards easy examples. Difficulty amplification varies with model architecture, model scale, training time and regularization strategy, and seemingly innocuous design decisions can have a substantial and counter-intuitive impact. Finally, we have shown how difficulty disparity and amplification take place in the Dollar Street setting, where our simple model is biased against images in the low income quartile. Taken together, our results highlight the key role of the model—above and beyond the dataset—in creating group disparities.

Limitations. Our primary aim is to further highlight the key role of the model in accuracy disparity. To simplify our setting we rely on binary group information (α/β, low-/high-income), though we fully expect the idea of difficulty disparity to generalize to other settings. We do however assume access to group information for audit purposes, which may not be available in many realistic scenarios, nor desirable to obtain. We intentionally choose to explore the balanced dataset setting, though separating difficulty disparity from other sources of bias may be difficult in practice. Our work also focuses on neural networks (with the exception of the SVM), such that extrapolation to other classes of models is not yet possible. Future work may seek to explore a broader array of model families. Finally, our synthetic experiments rely primarily on CIFAR-100 and evaluating our methodology on a larger datasets would be a practical extension.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 60–69, 2018.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 249–260, 2021. doi: 10.1145/3442188.3445888.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 312–321, 2019.

Martin Arjovsky, Kamalika Chaudhuri, and David Lopez-Paz. Throwing away data improves worst-class error in imbalanced classification, 2022. arXiv:2205.11672.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 233–242, 2017.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Sebastian Benthall and Bruce D. Haynes. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 289–298, 2019. doi: 10.1145/3287560.3287575.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546, 2021. doi: 10.1109/WACV48630.2021.00158.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pp. 77–91, 2018.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 872–881, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607, 2020.

Nikolay Dagaev, Brett D. Roads, Xiaoliang Luo, Daniel N. Barry, Kaustubh R. Patil, and Bradley C. Love. A too-good-to-be-true prior to reduce shortcut reliance, 2021. arXiv:2102.06406.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2): 20539517211035955, 2021. doi: 10.1177/20539517211035955.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, 2018. doi: 10.1145/3278721.3278729.

Chris Dulhanty and Alexander Wong. Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets, 2019. arXiv:1905.01347.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 2803–2813, 2020.

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

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115 (16):E3635–E3644, 2018. doi: 10.1073/pnas.1720347115.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. `https://github.com/facebookresearch/vissl`, 2021.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. 2022a. arXiv:2202.08360.

Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors, 2022b. arXiv:2202.07603.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification, 2022. arXiv:2201.11706.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 501–512, 2020. doi: 10.1145/3351095.3372826.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021. doi: https://doi.org/10.1016/j.patter.2021.100241.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 513, 2020. ISBN 9781450369367. doi: 10.1145/3351095.3375674.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177, pp. 336–351, 2022.

Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities, 2017. arXiv:1711.11561.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018. doi: 10.1145/3274357.

Fereshte Khani and Percy Liang. Feature noise induces loss discrepancy across groups. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5209–5219, 2020.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 196–205, 2021. doi: 10.1145/3442188.3445883.

Fereshte Khani, Aditi Raghunathan, and Percy Liang. Maximum weighted loss discrepancy, 2019. arXiv:1906.03518.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2022. arXiv:2204.02937.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1ecm2C9K7`.

Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness, 2018. arXiv:1805.05859.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 379–389, 2022. doi: 10.1145/3531146.3533105.

Stefano Sarao Mannelli, Federica Gerace, Negar Rostamzadeh, and Luca Saglietti. Inducing bias is simpler than you think, 2022. arXiv:2205.15935.

Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jphnJNOwe36`.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4615–4625, 2019.

Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. Understanding unequal gender classification accuracy from face images, 2018. arXiv:1812.00099.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. 2019.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. doi: https://doi.org/10.1016/j.patter.2021.100336.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5301–5310, 2019.

Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation, 2021. arXiv:2102.00813.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=j6NxpQbREA1`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2019. arXiv:1911.08731.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8346–8356, 2020.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 59–68, 2019. doi: 10.1145/3287560.3287598.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9573–9585, 2020.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017. arXiv:1711.08536.

Berfin Simsek, Melissa Hall, and Levent Sagun. Understanding out-of-distribution accuracies through quantifying difficulty of test samples, 2022. arXiv:2203.15100.

Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics, 2020. arXiv:2009.10795.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BJlxm30cKm`.

Guillermo Valle-Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rye4g3AqFm`.

Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset, 2016. arXiv:1605.06083.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017. doi: 10.1177/2053951717743530.

Angelina Wang and Olga Russakovsky. Directional bias amplification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 10882–10893, 2021.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection, 2019. arXiv:1902.11097.

D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1920–1953, 2017.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 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*, FAT* '20, pp. 547–558, 2020. doi: 10.1145/3351095.3375709.

Jingzhao Zhang, Aditya Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation, 2020. arXiv:2010.12230.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1323.

A PLS METHODOLOGY

Given m classes, we extract $d = m\frac{m-1}{2}$ binary accuracies corresponding to each possible pairing, and convert them to rank orders. We repeat this for each of the n models under consideration, resulting in a $d \times n$ matrix of difficulty ranks. We construct a $d \times 1$ data difficulty matrix from the cosine distance between the mean of each class.

We fit a partial least squares regression model to both the model difficulties and the data difficulties using *scikit-learn* (Pedregosa et al., 2011). For visualization purposes we use a single component, though in subsequent tests we find no difference in model fit when increasing the number of components. We evaluate how well the data difficulty is explained by the model difficulty using R^2.

B MODEL ARCHITECTURE AND HYPERPARAMETERS

B.1 SVM

For the SVM we use an RBF kernel with hyperparameters $C = 1.0$ and $\gamma = \frac{1}{3072}$, using the *scikit-learn* implementation.

B.2 NEURAL NETWORKS

All models are implemented in *PyTorch* (Paszke et al., 2019) with *TorchVision*. Models are trained to minimize cross-entropy loss using SGD with learning rate 0.01, momentum 0.9, weight decay of 0.0001 for 500 epochs with batch size 128.

FC. The fully-connected networks are either 3 or 5 hidden layers with 256 units and ReLU activation. Batch normalization is applied to the inputs.

LeNet. LeNet (Lecun et al., 1998) is a simple CNN, with two convolutional layers interleaved with max pooling, three fully-connected layers, and ReLU activation function.

ResNet-18. We use the variable-width ResNet (He et al., 2016) implementation of (Sagawa et al., 2020).

SSL. For both SSL models, we extract final-layer representations for each data point from an SSL-pretrained model. We pass these representations through a 1-layer FC network as described above. Representations are extracted from one two models. The first is from a ResNet-50, pretrained with SimCLR (Chen et al., 2020) on ImageNet-1K (Russakovsky et al., 2015). The second is from a RegNet-128Gf model (Radosavovic et al., 2020) trained with SwAV (Caron et al., 2020) on 1 billion public images from Instagram (Goyal et al., 2022a). Representations were extracted using VISSL (Goyal et al., 2021) from models publicly available in the model zoo.

C DOLLAR STREET EXPERIMENT

C.1 DATASET

Dollar Street[3] is a dataset of geographically-diverse images spanning a broad range of household incomes. Dollar Street comprises 23724 RGB 480×480 images of objects and people in everyday environments around the world, each associated with one of 138 class labels. For our purposes, we discard geographic information and use income quartiles as group label. We use only the lowest and highest income quartiles to create binary groups.

C.2 MODEL

We follow the representation extraction method outlined in appendix B.2, though we use the representations from supervised learning models rather than SSL. Specifically, we extract from a ResNet-50 trained with supervised learning on ImageNet-1K (Russakovsky et al., 2015). We train a single

[3] `https://www.gapminder.org/dollar-street`

layer fully-connected network of varying width, following the standard SGD training regime specified above.

D CALCULATING AMPLIFICATION FACTOR WITH LINEAR REGRESSION

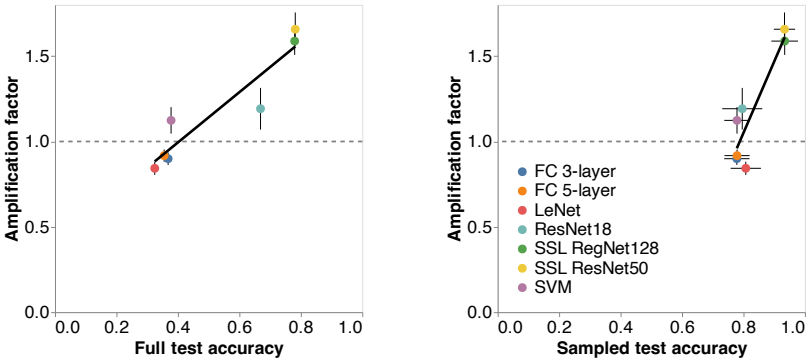
Given a vector of estimated accuracy disparities $\tilde{\mathbf{d}}$, and a vector of observed accuracy disparities $\mathbf{d}$, we estimate the amplification factor using OLS linear regression.

In our synthetic setup, we additionally control for the effect of confounds including within-class separability (e.g. *aquatic mammal / medium mammal* in fig. 3a), and diagonal separability (e.g. *medium mammal / flower*), by including them as nuissance regressors. Our full model is of the form:

$$\mathbf{d} = X\beta + \epsilon, \quad X = \begin{pmatrix} \tilde{\mathbf{d}} \\ \mathbf{s}_{a,0}^{a,1} \\ \mathbf{s}_{b,0}^{b,1} \\ \mathbf{s}_{a,0}^{b,1} \\ \mathbf{s}_{a,1}^{b,0} \end{pmatrix}, \tag{4}$$

where amplification factor is the first parameter, $k = \beta_0$, and $\mathbf{s}_{a,0}^{b,1}$ is the separability (i.e. accuracy) between group a, label 0 and group b, label 1. We use Python *statsmodels* (Seabold & Perktold, 2010) to fit the model.

E SUPPLEMENTARY FIGURES



(a) Full CIFAR-100 coarse (b) Sampled tasks

Figure S1: Amplification factor k as function of average test accuracy on **(a)** the full CIFAR-100 coarse dataset, and **(b)** the 30 sampled tasks used for computing amplification factor. **Choosing a higher accuracy model, e.g./ an SSL model, would increase amplification.** Vertical bars are standard error of the coefficient k. Horizontal bars (barely visible in left panel) are standard deviation of test accuracy over (a) 10 seeds and (b) 10 seeds and 30 tasks. Solid black line fit with linear regression. Dashed gray line is $k = 1$.

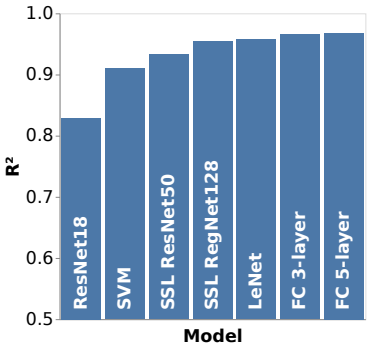
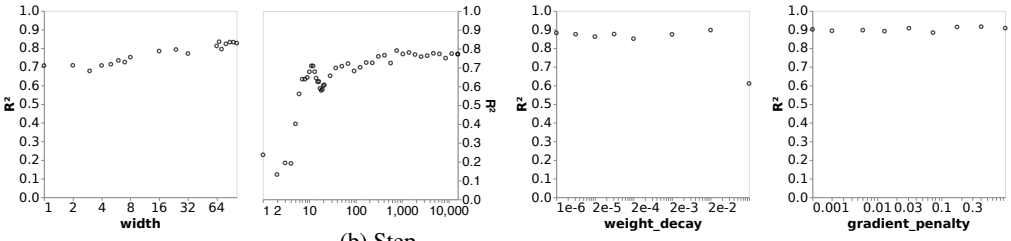


Figure S2: R^2 values for linear regression calculation of amplification factor, for various models, corresponding to fig. 4



(a) Width (b) Step (c) Weight decay (d) Gradient penalty

Figure S3: R^2 values for linear regression calculation of amplification factor, corresponding to fig. 5.