# The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions

Anna Meyer
apmeyer4@wisc.edu
University of Wisconsin - Madison
Madison, USA

Aws Albarghouthi
aws@cs.wisc.edu
University of Wisconsin - Madison
Madison, USA

Loris D'Antoni
loris@cs.wisc.edu
University of Wisconsin - Madison
Madison, USA

## ABSTRACT

We introduce dataset multiplicity, a way to study how inaccuracies, uncertainty, and social bias in training datasets impact test-time predictions. The dataset multiplicity framework asks a counterfactual question of what the set of resultant models (and associated test-time predictions) would be if we could somehow access *all* hypothetical, unbiased versions of the dataset. We discuss how to use this framework to encapsulate various sources of uncertainty in datasets' factualness, including systemic social bias, data collection practices, and noisy labels or features. We show how to exactly analyze the impacts of dataset multiplicity for a specific model architecture and type of uncertainty: linear models with label errors. Our empirical analysis shows that real-world datasets, under reasonable assumptions, contain many test samples whose predictions are affected by dataset multiplicity. Furthermore, the choice of domain-specific dataset multiplicity definition determines what samples are affected, and whether different demographic groups are disparately impacted. Finally, we discuss implications of dataset multiplicity for machine learning practice and research, including considerations for when model outcomes should not be trusted.

## CCS CONCEPTS

• **Computing methodologies** → *Machine learning approaches*; • **General and reference** → *Evaluation*; • **Social and professional topics** → **Computing / technology policy**; • **Theory of computation** → Machine learning theory.

## KEYWORDS

Dataset multiplicity, procedural fairness, model robustness, data bias, model multiplicity

**ACM Reference Format:**
Anna Meyer, Aws Albarghouthi, and Loris D'Antoni. 2023. The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3593013.3593988

## 1 INTRODUCTION

Datasets that power machine learning algorithms are supposed to be accurate and fully representative of the world, but in practice, this level of precision and representativeness is impossible [27, 44]. Datasets display inaccuracies — which we use as a catch-all term for both errors and nonrepresentativeness — due to sampling bias [10], human errors in label or feature transcription [39, 63], and sometimes deliberate poisoning attacks [3, 52]. Datasets can also reflect undesirable societal inequities. But more broadly, datasets never reflect objective truths because the worldview of their creators is imbued in the data collection and postprocessing [27, 42, 44]. Additionally, seemingly-trivial decisions in the data collection or annotation process influence exactly what data is included, or not [42, 45]. In psychology, these minute decisions have been termed 'researcher degrees of freedom,' i.e., choices that can inadvertently influence conclusions that one ultimately draws from the data analysis [55]. In this paper, we study how unreliable data of all kinds impacts the predictions of the models trained on such data and frame this analysis as a 'multiplicity problem.'

Multiplicity occurs when there are multiple explanations for the same phenomenon. Many recent works in machine learning have studied predictive multiplicity, which occurs when multiple models have equivalent accuracy, but still give different predictions to individual samples [14, 33, 51]. A consequence of predictive multiplicity is procedural unfairness concerns; namely, defending the choice of model may be challenging when there are alternatives that give more favorable predictions to some individuals [6]. But model selection is just one source of multiplicity. In this paper, we argue that it makes sense to consider training datasets through a multiplicity lens, as well. To do so, we will consider a *set of datasets*. Intuitively, this set captures all datasets that could have been collected if the world was slightly different, i.e., if we could correct the unknown inaccuracies in the data. We illustrate this idea through the following example.

*Example dataset multiplicity use case.* Suppose a company wants to deploy a machine learning model to decide what to pay new employees. They have access to current employees' backgrounds, qualifications, and salaries. However, they are aware that in various societies, including the United States, there is a gender wage gap, i.e., systematic disparities in the average pay between men and women [65]. Economists believe that while some of the gap is attributable to factors like choice of job industry, overt discrimination also plays a role [7]. But even though we know that discrimination exists, it is very difficult to adjudicate whether specific compensation decisions are affected by discrimination.

The original dataset is shown, along with the best-fit model $f$, in Figure 1(a). Note that under $f$, the proposed salary for a new

*FAccT '23, June 12–15, 2023, Chicago, IL, USA*

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3593988

employee $\mathbf{x}$ is \$73,000. But an alternate possibility of the 'ground-truth,' debiased dataset is shown in Figure 1(b). In the world that produced this dataset, perhaps there is no gender discrimination, so, ideally, we would learn from this dataset and yield model $g$, which places $\mathbf{x}$'s salary at \$78,000.

The modified dataset in Figure 1(b) is just one example of how different data collection practices — in this case, collecting data from an alternate universe where there is no gender-based discrimination in salaries — can lead to various datasets that produce models that make conflicting predictions for individual test samples.

But what if we could consider the entire range of candidate 'ground-truth' datasets? For example, all datasets where each woman's salary may be increased by up to \$10,000 to account for the impacts of potential gender discrimination. Figure 1(c) shows what we could hypothetically do if that set of datasets were available — and we had unlimited computing power. Rather than outputting a single model, we bound the range of models (the highlighted region) obtainable from alternate-universe training datasets. We could then use this set of models to obtain a confidence interval for $\mathbf{x}$'s prediction - in this case, \$68,000-\$83,000. This range corresponds to the *dataset multiplicity robustness* of $\mathbf{x}$, that is, the sensitivity of the model's prediction on $\mathbf{x}$ given specific types of changes in the training dataset.

Work in algorithmic stability, robust statistics, and distributional robustness has attempted to quantify how varying the training data impacts downstream predictions. However, as illustrated by the example, we aim to find the *pointwise* impact of uncertainty in training data rather than studying robustness purely in aggregate, and we want our analysis to encompass the *worst-case* (i.e., adversarial) reasonable alternate models, unlike the purely statistical approaches.

*Our vision for dataset multiplicity in machine learning.* The proposed solution in the above example suffers from a number of drawbacks. First, is the solution solving the right problem? That is, is dataset multiplicity a better choice for reasoning about unreliable data than existing learning theory techniques? Second, how do we define what is a reasonable alternate-universe dataset to include in the set of datasets? Third, even if we had a set of datasets encompassing all alternate universes, how would we compute the graph in Figure 1(c)? And finally, what are the implications for fair and trustworthy machine learning? How can, and should, we incorporate dataset multiplicity into machine learning research, development and deployment?

We address all of these concerns in this paper through the following contributions:

**Conceptual Contribution 1** We formally define the dataset multiplicity problem and give several example use cases to provide intuition of how to define the set of reasonable alternate-universe datasets (Section 2)

**Theoretical Contribution** We present a novel technique that, for linear models with label errors, can exactly characterize the range of a test sample's prediction. We also show how to over-approximate the range of models (i.e., Figure 1(c)) (Section 3)

**Experimental Contribution** We use our approaches to evaluate the effects of dataset multiplicity on real-world datasets with a particular eye towards how demographic subgroups are differently affected (Section 4)

**Conceptual Contribution 2** We explore the implications of dataset multiplicity (Section 5)

## 2 THE DATASET MULTIPLICITY PROBLEM

We formalize *dataset multiplicity* as a technique that conceptualizes uncertainty and societal bias in training datasets and discuss how to use dataset multiplicity as a tool to critically assess machine learning models' outputs.

We assume the following supervised machine learning setup: we start with a fixed, deterministic learning algorithm $A$ and a training dataset $D = (\mathbf{X}, \mathbf{y})$ with features $\mathbf{X} \in \mathbb{R}^{n \times d}$ and outputs $\mathbf{y} \in \mathbb{R}^n$.[1] We run $A$ on $D$ to get a model $f$, that is, $f = A(D)$. Given a test sample $\mathbf{x}$, we obtain an associated prediction $\hat{y} = f(\mathbf{x})$.

### 2.1 Defining Dataset Multiplicity

We describe dataset multiplicity for a dataset $D$ with a *dataset multiplicity model* $\mathcal{M}(D)$. Intuitively, we want $\mathcal{M}(D)$ to be the smallest set that contains all conceivable alternatives to $D$. We present a few examples of $\mathcal{M}$:

*$\mathcal{M}$ under societally-biased labels.* We continue the example from the introduction. Suppose we believe that women in a dataset are underpaid by up to \$10,000 each. In this case, we define $\mathcal{M}(D)$ as the set of all datasets $D'$ with identical features to $D$, such that all labels for men in $D'$ are identical to the labels in $D$ and all labels for women in $D'$ are the same as in $D$, or increased by up to \$10,000.

*$\mathcal{M}$ under noisy measurement.* Suppose a dataset contains a weight feature and where data was collected using a tool whose measurements may be inaccurate by up to 5 grams. If weight is the $k^{\text{th}}$ feature, then we can represent $\mathcal{M}(D)$ as the set of all datasets $D'$ that are identical to $D$ except that feature $k$ may differ by up to 5 grams.

*$\mathcal{M}$ given unreliable feature data.* People commonly misreport their height on dating profiles [61]: men add 0.5"(±0.88) to their height, on average, while women add 0.17"(±0.98). So, for a given man with reported height $h$ we can be 95% confident that his height is in $[h - 1.26, h + 2.26]$ and for a women with height $h$, $[h - 1.79, h + 2.13]$. If a dataset contains height (as self-reported via dating apps) in position $i$, then $\mathcal{M}(D)$ will contain all datasets $D'$ that are equivalent to $D$, except that each sample's $i^{\text{th}}$ feature may be modified according to the gender-specific 95% confidence intervals.

*$\mathcal{M}$ with missing data.* Getting a representative population sample can be a challenge for many data collection tasks. Suppose we suspect that a dataset underrepresents a specific minority population by up to 20 samples. If the undersampled group has feature $i = k$, then $\mathcal{M}(D)$ will be the set of datasets $D'$ such that $D \subseteq D'$, there are at most 20 additional samples in $D'$, and all new samples have feature $i = k$. (We additionally assume that all new samples are in the proper feature space, i.e., that they are valid data samples.)

[1] $A$ is inclusive of all modeling steps including preprocessing the training data, selecting the model hyperparameters through a holdout validation set (segmented off from $D$), etc.

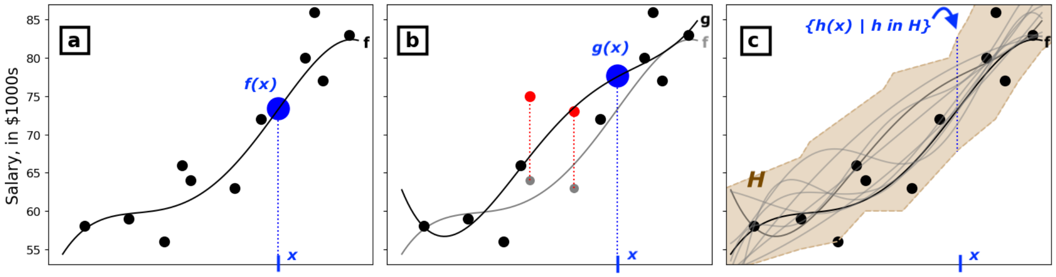

**Figure 1: Salary prediction: (a) Training dataset and resultant model $f$. The prediction for the test sample x is $\$73{,}000$. (b) Training dataset with two label modifications (in red) along with the newly-learned model ($g$). The prediction for the test sample x is now about $\$78{,}000$. (c) $H$ contains the set of models $h$ that we could have obtained based on various small modifications to the provided dataset. We see that x's prediction can be anywhere between $\$68{,}000$ and $\$83{,}000$ (blue dotted line).**

## 2.2 Learning with Dataset Multiplicity

We define $A(\mathcal{M}(D))$ to be the set of all models obtainable by training with some dataset in $\mathcal{M}(D)$, i.e., $A(\mathcal{M}(D)) = \{f \mid \exists D' \in \mathcal{M}(D) \text{ s.t. } A(D') = f\}$. Given this set of models, we can inquire about the range of possible predictions for a test sample $\mathbf{x}$. In particular, we can ask whether $\mathbf{x}$ is *robust* to dataset multiplicity, that is, will it receive a different prediction if we started with any other model in $\mathcal{M}(D)$? Formally, we say that a deterministic algorithm $A$, a dataset $D$, and a dataset multiplicity model $\mathcal{M}(D)$ are $\epsilon$-robust to dataset multiplicity on a sample $\mathbf{x}$ if Equation (1) holds. (Equivalently, we will say that $\mathbf{x}$ is $\epsilon$-robust.)

$$D' \in \mathcal{M}(D) \implies A(D')(\mathbf{x}) \in [A(D)(\mathbf{x}) - \epsilon, A(D)(\mathbf{x}) + \epsilon] \quad (1)$$

Example 2.1. *Returning to the example from the introduction, the test sample* $\mathbf{x}$ *originally receives a prediction of* $\$78{,}000$ *(Figure 1a). Figure 1c shows that* $\mathbf{x}$ *is* not *$\epsilon$-robust for* $\epsilon = \$5{,}000$*, since it can receive any prediction in* $[68{,}000,\ 83{,}000]$*, and* $78{,}000 - 68{,}000 > 5{,}000$*. However,* $\mathbf{x}$ *is $\epsilon$-robust for* $\epsilon = 10{,}000$*.*

If a sample $\mathbf{x}$ is $\epsilon$-robust, then we can be certain that its prediction will not change by more than $\epsilon$ due to dataset multiplicity. In practice, this may mean we can deploy the prediction with greater confidence, or less oversight. Conversely, if $\mathbf{x}$ is not $\epsilon$-robust, then this means there is some plausible alternate training dataset that, when used to train a model, would result in a different prediction for $\mathbf{x}$. In this case, the prediction on $\mathbf{x}$ may be less trustworthy — we discuss options for dealing with non-robustness in Section 5.

## 2.3 Choosing a Dataset Multiplicity Model

We have discussed how to formalize $\mathcal{M}(D)$ given various conceptions of dataset inaccuracy; however, we have not discussed how to determine in what ways a given dataset may be inaccurate. In practice, these judgments should be made in collaboration with domain experts, both within the data science and social science realms. Still, there is no one normative, 'right' answer of how to define $\mathcal{M}$ for a given situation — any judgment will be normative. Furthermore, there may be multiple ways to describe the same social phenomenon, as illustrated by the following example:

Example 2.2. *The first example in Section 2.1 formalizes gender discrimination in salaries. When index 0 is gender and value 1 is woman, we define* $\mathcal{M}$ *as* $\mathcal{M}(D) = \{(\mathbf{X}, \mathbf{y}') \mid (\mathbf{X}_i)_0 = 1 \implies y_i' \in [y_i, y_i + 10{,}000]$ *and* $(\mathbf{X}_i)_0 \neq 1 \implies y_i = y_i'\}$*. However, what if we frame the problem as men are overpaid, rather than women are underpaid? In that case, a more appropriate formalization would be* $\mathcal{M}(D) = \{(\mathbf{X}, \mathbf{y}') \mid (\mathbf{X}_i)_0 = 0 \implies y_i' \in [y_i - 10{,}000, y_i]$ *and* $(\mathbf{X}_i)_0 \neq 0 \implies y_i = y_i'\}$*.*

As we will see in Section 4.2, this variability in framing can affect the conclusions we draw about dataset multiplicity's impacts, highlighting the need for thoughtful reflection and interdisciplinary collaboration when choosing $\mathcal{M}$.

# 3 THE DATASET MULTIPLICITY PROBLEM FOR LINEAR MODELS WITH LABEL ERRORS

We consider a special case of the dataset multiplicity problem introduced in Section 2, namely, linear models given noise or errors in the training data's labels. (We use the term 'label' in the context of both linear regression and classification.) We present this analysis to begin to characterize the impact of dataset multiplicity on real-world datasets and models, and to provide an example for how we envision the study of dataset multiplicity's impacts to continue in future work.

We focus on linear models with label errors for a few reasons. First, linear models are well-studied and used in practice, especially with tabular data, which is common in areas with societal implications. Furthermore, complicated models like neural nets can often be conceived of as encoders followed by a final linear layer, making our results more widely applicable. Second, label errors and noise are common, well-documented realities in many applications [39]. Finally, as we will see, the closed-form solution for linear regression allows us to solve this problem exactly, a challenge that is currently impractical even for other simple, widely-studied model families like decision trees [35].

## 3.1 Formulating the Dataset Label Multiplicity Problem

We assume our dataset is of the form $D = (\mathbf{X}, \mathbf{y})$ with feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and output vector $\mathbf{y} \in \mathbb{R}^n$. (Even though $\mathbf{y}$ is continuous, we borrow terminology from classification to also refer to $y_i$ as the *label* for $\mathbf{X}_i$.) At times, we will reference the interval domain, $\mathbb{IR}$, that is, $\mathbb{IR} = \{[a, b] \mid a, b \in \mathbb{R}, a \leq b\}$.

*Parameterizing $\mathcal{M}$ given label perturbations.* We parameterize $\mathcal{M}$ given label noise with three parameters, $k$, $\Delta$, and $\phi$. First, $k \in \mathbb{N}$ is an upper bound on the number of training samples that have an inaccurate label. Second, $\Delta \in \mathbb{IR}^n$ stores the amount that each label can change. The $i^{\text{th}}$ element of $\Delta$ is $[\delta_i^l, \delta_i^u]$, signifying that the true value of $y_i$ falls in the interval $[y_i + \delta_i^l, y_i + \delta_i^u]$. Finally, $\phi$ is predicate over the feature space specifying whether we can modify a given sample when, for example, label errors are limited to a population subgroup. Given $k$, $\Delta$, and $\phi$ we define $\mathcal{M}$ as

$$\mathcal{M}_{k,\Delta,\phi}((\mathbf{X}, \mathbf{y})) = \{(\mathbf{X}, \mathbf{y}') \mid \|\mathbb{1}[\mathbf{y} \neq \mathbf{y}']\|_1 \leq k \wedge \\ \forall i. y_i \neq y_i' \implies \phi(\mathbf{x}_i) \wedge \forall i. y_i' - y_i \in \delta_i\}$$

We describe $k$, $\Delta$, and $\phi$ for the following examples:

Example 3.1. *We assume that women in a salary dataset are underpaid by up to* \$10,000 *each. Since we place no limit on how many labels may be incorrect — beyond the proportion of women in the dataset — we set $k = n$, the total number of samples. Since labels may be underreported by up to* \$10,000, *we set $\Delta = [0, 10{,}000]^n$. And finally, since $\mathbf{x}_0 = 1$ means that $\mathbf{x}$ is a woman, we define $\phi(\mathbf{x}) = \mathbb{1}[\mathbf{x}_0 = 1]$ since we assume that only women's salaries may change.*

Example 3.2. *(Spam filter) Suppose a dataset $D = (\mathbf{X}, \mathbf{y})$ contains emails $\mathbf{X}$ that are labeled as not spam or spam (i.e., $\mathbf{y} \in \{-1, 1\}^n$). From manual inspection of a small portion of the dataset, we estimate that up to 2% of the emails are mislabeled. Since up to 2% of the labels may be incorrect, we set $k = 0.02n$. As the labels are binary, we can modify each label by $+/-2$, depending on its original label.*[2] *Thus, $\Delta = [a, b]^n$ where $[a, b]_i = [0, 2]$ when $y_i = -1$ and $[a, b]_i = [-2, 0]$ when $y_i = 1$. Finally, $\phi(\mathbf{x}) = 1$ since there are no limitations on which samples have inaccurate labels.*

## 3.2 Linear Regression Overview

Our goal is to find the optimal linear regression parameter $\theta$, i.e.,

$$\theta = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} (\theta^T \mathbf{X} - \mathbf{y})^2 \tag{2}$$

Least-squares regression admits a closed-form solution[3]

$$\theta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{3}$$

for which we will analyze dataset label multiplicity. We work with the closed-form solution, instead of a gradient-based one, as it is deterministic and holistically considers the whole dataset, allowing us to exactly measure dataset multiplicity by exploiting linearity (Rosenfeld et al. make an analogous observation [47]). On medium-sized datasets and modern machines, computing this closed-form solution is efficient.

Given a solution, $\theta$, to Equation (2), we output the prediction $\hat{y} = \theta^\top \mathbf{x}$ for a test point $\mathbf{x}$.

**Algorithm 1** Find the allowable perturbation that makes $\hat{y}$ as large as possible, i.e., $\max_{(\mathbf{X}, \tilde{\mathbf{y}}) \in \mathcal{M}_{k,\Delta,\phi}((\mathbf{X}, \mathbf{y}))} \mathbf{z}\tilde{\mathbf{y}}$

```
Require: $\mathbf{z} \in \mathbb{R}^n$, $(\mathbf{X}, \mathbf{y}) \in (\mathbb{R}^{n \times d}, \mathbb{R}^n)$, $\Delta \in \mathbb{IR}^n$ with $0 \in \Delta$, $k \geq 0$,
    $\phi : \mathbb{R}^d \to \{0, 1\}$
  $\mathbf{y}^u \leftarrow \mathbf{y}$
  if $z_i \geq 0$ then $\rho_i^+ \leftarrow z_i \delta_i^u$ else $\rho_i^+ \leftarrow z_i \delta_i^l$
  if not $\phi(\mathbf{x}_i)$ then $\rho_i^+ \leftarrow 0$
  Let $\rho_{i_1}^+, \ldots, \rho_{i_l}^+$ be the $l$ largest elements of $\rho^+$ by absolute value
  for each $\rho_{i_j}^+$ do
      if $z_{i_j} \geq 0$ then $y_{i_j}^u \leftarrow y_{i_j}^u + \delta_{i_j}^u$ else $y_{i_j}^u \leftarrow y_{i_j}^u + \delta_{i_j}^l$
  return $\mathbf{z}\mathbf{y}^u$
```

*Extension to binary classification.* Given a binary output vector $\mathbf{y} \in \{-1, 1\}^n$, we find $\theta$ in the same way, but when making test-time predictions, use 0 as a cutoff between the two classes, i.e., given parameter vector $\theta$ and test sample $\mathbf{x}$, we return 1 if $\theta^T \mathbf{x} > 0$ and $-1$ otherwise. To evaluate robustness for binary classification, we care about whether the predicted label can change when training with any dataset $D' \in \mathcal{M}(D)$. Thus, if $\theta^T \mathbf{x} \geq 0$, then $\mathbf{x}$ is dataset multiplicity robust if there is no model $\theta'$ such that $(\theta')^T \mathbf{x} < 0$ (and vice-versa when $\theta^T \mathbf{x} < 0$).

## 3.3 Exact Pointwise Solution

Given a model $\theta$ and test sample $\mathbf{x}$ we can expand and rearrange $\theta^\top \mathbf{x}$ as follows:

$$\hat{y} = \theta^\top \mathbf{x} = ((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})^\top \mathbf{x} = \underbrace{(\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)}_{\mathbf{z}} \mathbf{y}$$

This form is useful since it isolates $\mathbf{y}$, which under our dataset multiplicity assumption contains all of the dataset's uncertainty. We will use $\mathbf{z}$ to denote $\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. Thus, our goal is to find

$$(\_\,, \mathbf{y}') = \underset{(\_\,, \tilde{\mathbf{y}}) \in \mathcal{M}_{k,\Delta,\phi}(D)}{\operatorname{argmax}} |\mathbf{z}\tilde{\mathbf{y}} - \mathbf{z}\mathbf{y}| \tag{4}$$

and then to check whether $\|\hat{y} - \mathbf{z}\mathbf{y}'\| < \epsilon$. If so, then we will have proved that the prediction for $\mathbf{x}$ is $\epsilon$-robust under $\mathcal{M}_{k,\Delta,\phi}(D)$. (Conversely, if $\|\hat{y} - \mathbf{z}\mathbf{y}'\| \geq \epsilon$, then $\mathbf{y}'$ is a counterexample proving that $\mathbf{x}$ is not $\epsilon$-robust under $\mathcal{M}_{k,\Delta,\phi}(D)$.)

*Solving for Equation* (4). One option is to formulate Equation (4) as a mixed-integer linear program. However, due to the vast number of possible label perturbation combinations, this approach is prohibitively slow (we provide a runtime comparison with our approach in the appendix). Instead, we use the algorithmic technique presented in Algorithm 1. Intuitively, one iteration of the algorithm's inner loop identifies what output $y_i \in \mathbf{y}$ to modify so that we maximally increase $\mathbf{z}\mathbf{y}$. After $k$ output modifications — or once all outputs eligible for modification under $\phi$ have been modified — we check whether the new prediction, $\hat{y}'$, has $\hat{y}' > \theta^T \mathbf{x} + \epsilon$. If this is the case, we stop because we have shown that $\mathbf{x}$ is not $\epsilon$-dataset multiplicity robust. Otherwise, we repeat a variation of the algorithm (see the appendix) to maximally decrease $\mathbf{z}\mathbf{y}$ and check whether we can achieve $\hat{y}' < \theta^T \mathbf{x} - \epsilon$.

[2] The linearity of the algorithm ensures that all labels will remain valid, i.e., either -1 or 1.

[3] In practice, we implement ridge regression, $\theta = (\mathbf{X}^T \mathbf{X} - \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$, for greater stability.

*Extension to binary classification.* The binary classification version of the algorithm works identically, except we check whether $\hat{y}$ rounds to a different class than $\theta^T\mathbf{x}$ to ascertain $\mathbf{x}$'s robustness.

## 3.4 Over-Approximate Global Solution

In Section 3.3, we described a procedure to find the exact dataset multiplicity range for a test point $\mathbf{x}$. For every input $\mathbf{x}$ for which we want to know the dataset multiplicity, the procedure effectively relearns the worst-case linear regression model for $\mathbf{x}$. In practice, we may want to explore the dataset multiplicity of a large number of samples, e.g., a whole test dataset, or we may need to perform online analysis.

We would like to understand the dataset multiplicity range for a large number of test points without performing linear regression for every input. We formalize capturing all linear regression models we may obtain as follows:

$$\Theta = \{\theta \mid \theta = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{y}} \text{ for some } (\mathbf{X}, \tilde{\mathbf{y}}) \in \mathcal{M}_{k,\Delta,\phi}(D)\}$$

To see whether $\mathbf{x}$ is $\epsilon$-robust, we check whether $\tilde{\theta}^\top\mathbf{x} \in [\theta^\top\mathbf{x} - \epsilon, \theta^\top\mathbf{x} + \epsilon]$ for all $\tilde{\theta} \in \Theta$. For ease of notation, let $\mathbf{C} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Note that $\mathbf{C} \in \mathbb{R}^{m\times n}$, while $\mathbf{z} \in \mathbb{R}^{1\times n}$.

*Challenges.* The set of weights $\Theta$ is not enumerable and is non-convex (proof in the appendix). Our goal is to represent $\Theta$ efficiently so that we can simultaneously apply all weights $\theta \in \Theta$ to a point $\mathbf{x}$. Our key observation is that we can easily compute a hyperrectangular over-approximation of $\Theta$. In other words, we want to compute a set $\Theta^a$ such that $\Theta \subseteq \Theta^a$. Note that the set $\Theta^a$ is an interval vector in $\mathbb{IR}^n$, since interval vectors represent hyperrectangles in Euclidean space.

This approach results in an overapproximation of the true dataset multiplicity range for a test sample $\mathbf{x}$ — that is, some values in the range may not be attainable via any allowable training label modification.

*Approximation approach.* We will iteratively compute components of the vector $\Theta^a$ by finding each coordinate $i$ as the following interval, where $\mathbf{c}_i$ are the column vectors of $\mathbf{C}$:

$$\Theta_i^a = \left[\min_{(\mathbf{X},\mathbf{y}')\in\mathcal{M}_{k,\Delta,\phi}(D)} \mathbf{c}_i\mathbf{y}', \max_{(\mathbf{X},\mathbf{y}')\in\mathcal{M}_{k,\Delta,\phi}(D)} \mathbf{c}_i\mathbf{y}'\right]$$

To find $\min_{\mathbf{y}'\in\mathcal{M}_{k,\Delta}(\mathbf{y})} \mathbf{c}_i\mathbf{y}'$, we use the same process as in Section 3.3. Specifically, we use Algorithm 1 to compute the lower and upper bounds of each $\Theta_i^a$. We show in the appendix that the interval matrix $\Theta^a$ is the tightest possible hyperrectangular over-approximation of the set $\theta$.

*Evaluating the impact of dataset multiplicity on predictions.* Given $\Theta^a$ as described above, the output for an input $\mathbf{x}$ is provably robust to dataset multiplicity if

$$(\Theta^a)^\top\mathbf{x} \subseteq [\theta^\top\mathbf{x} - \epsilon, \theta^\top\mathbf{x} + \epsilon] \qquad (5)$$

Note that $(\Theta^a)^\top\mathbf{x}$ is computed using standard interval arithmetic, e.g., $[a, b] + [a', b'] = [a + a', b + b']$. Also note that the above is a one-sided check: we can only say that the model's output given $\mathbf{x}$ is robust to dataset multiplicity, but because $\Theta^a$ is an overapproximation, if Equation (5) does not hold, we cannot conclusively say that the model's prediction on $\mathbf{x}$ is subject to dataset multiplicity.

# 4 EMPIRICAL EVALUATION

We use Python to implement the algorithms from Sections 3.3 and 3.4 for measuring label-error multiplicity in linear models.[4] To speed up the evaluation, we use a high-throughput computing cluster. (We request 8GB memory and 8GB disk, but all experiments are feasible to run on a standard laptop.) This approach does not have a direct baseline with which to compare, as ours is the first work to propose and analyze the dataset-multiplicity problem.

*Datasets and tasks.* We analyze our approach on three datasets: the Income prediction task from the FolkTables project [20], the Loan Application Register (LAR) from the Home Mortgage Disclosure Act publication materials [22], and MNIST 1/7 (i.e., the MNIST dataset limited to 1's and 7's) [32]. We divide each dataset into train (80%), test (10%), and validation (10%) datasets and repeat all experiments across 10 folds, except when a standard train/test split is provided, as with MNIST. We perform classification on the Income dataset (whether or not an individual earned over $50,000), on LAR (whether or not a home mortgage loan was approved), and on MNIST (binary classification limited to 1's and 7's). Additionally, in the appendix we evaluate the regression version of Income by predicting an individual's exact income. For all of the Income experiments, we limit the dataset to only include data from a single U.S. state to speed computations. In Sections 4.1 and 4.3 we present results from a single state, Wisconsin, while in Section 4.2 we compare results across five different US states.

*Accuracy-Robustness Tradeoff.* There is a tradeoff between accuracy and robustness to dataset multiplicity that is controlled by the regularization parameter $\lambda$ in the ridge regression formula $\theta = (\mathbf{X}^\top\mathbf{X} - \lambda I)^{-1}\mathbf{X}^\top\mathbf{y}$. Larger values of $\lambda$ improve robustness at the expense of accuracy. Figure 2 illustrates this tradeoff. All results below, unless otherwise stated, use a value of $\lambda$ that maximizes accuracy.

*Experiment goals.* Our core objective is to see how robust linear models are to dataset label multiplicity. We measure this sensitivity with the *robustness rate*, that is, the fraction of test points that receive invariant predictions (within a radius of $\epsilon$) given a certain level of label inaccuracies. The robustness rate is a proxy for the stability of a modeling process under dataset multiplicity, so knowing this rate — and comparing it across various datasets, demographic groups, and algorithms — can help ML practitioners analyze the trustworthiness of their models' outputs. In Section 4.1, we describe the overall robustness results for each dataset. Then, in Section 4.2, we perform a stratified analysis across demographic groups and show how varying the dataset multiplicity model definition can significantly change data's vulnerability to dataset multiplicity. Finally, in Section 4.3, we discuss results of the over-approximate approach and how it can be used to evaluate dataset multiplicity robustness.

## 4.1 Robustness to Dataset Multiplicity

**Key insights:** When a small percentage (e.g., 1%) of labels are incorrect, a significant minority of test samples are not robust to

[4] Our code is available at https://github.com/annapmeyer/linear-bias-certification.

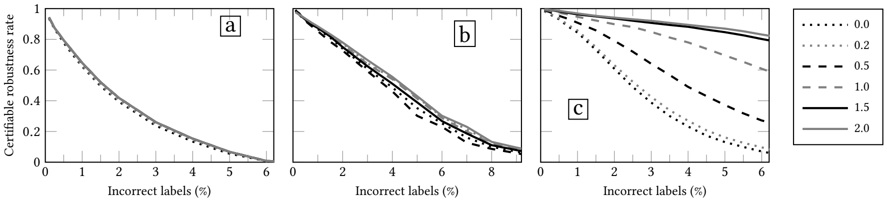

**Figure 2: Fraction of test samples whose predictions are robust to dataset multiplicity for different accuracy/robustness trade-offs as controlled by the ridge regression parameter $\lambda$. The different lines within each graph plot robustness for different accuracy/robustness trade-offs. The line labeled 0.0 corresponds to the value of $\lambda$ that achieves maximal accuracy, the line labeled 0.2 corresponds to the value of $\lambda$ that sacrifices a 0.2 percentage-point drop in accuracy for more robustness, etc. We include the specific accuracy values in the appendix. The datasets for each graph are (a) LAR, (b) Income, and (c) MNIST 1/7.**

dataset multiplicity, raising questions about the reliability of the models' predictions.

Table 1 shows the fraction of test points that are dataset multiplicity robust for classification datasets at various levels of label inaccuracies. For each dataset, the robustness rates are relatively high (> 80%) when fewer than 0.25% of the labels can be modified, and stay above 50% for 1% label error.

Despite globally high robustness rates, we must also consider the non-robust data points. In particular, we want to emphasize that for Income and LAR, each non-robust point represents an individual whose classification hinges on the labels of only a small number of training samples. That is, given the assumed uncertainty about the labels' accuracy, it is plausible that a 'clean' dataset would output different test-time predictions for these samples. Some data points will almost surely fall into this category — if not, that would mean the model was independent from the training data, which is not our goal! However, if a sample is not robust to a small number of label modifications, perhaps the model should not be deployed on that sample. Instead, if the domain is high-impact, the sample could be evaluated by a human or auxiliary model (see Section 5 for more discussions on how to handle non-robust test samples).

Returning to Table 1, many data points are not robust at low label error rates, e.g., when 1% of labels may be wrong, 49.3% of Income test samples are not robust. Likewise, 38.7% of LAR test samples can receive the opposite classification if the correct subset of 1% of labels change. These low robustness rates call into question the advisability of using linear classifiers on these datasets unless one is confident that label accuracy is very high.

## 4.2 Disparate Impacts of Dataset Multiplicity

In Section 4.1, we showed dataset multiplicity robustness results given the assumption that all labels in the training dataset were potentially inaccurate. However, in practice, label errors may be systemic. In particular, two of the datasets we analyzed in Section 4.1 contain data that may display racial or gender bias. We hypothesize that the Income dataset likely reflects trends where women and people of color are underpaid relative to white men in the United States, and that the LAR dataset may similarly reflect racial and gender biases on the part of mortgage lending decision makers. To leverage this refined understanding of potential inaccuracies in the labels, in this section we evaluate test data robustness under the following two *targeted* dataset multiplicity paradigms:

- *'Promoting' the disadvantaged group*: We restrict label modification to members of the disadvantaged group (i.e., Black people or women); furthermore, we only change labels from the negative class to the positive class.
- *'Demoting' the advantaged group*: We restrict label modification to the advantaged group (i.e., White people or men); furthermore, we only allow change labels from the positive class to the negative class. This setup is compatible with the worldview (for example) that men are overpaid.

Before delving into the results, we want to acknowledge that this analysis has a few shortcomings. First, for simplicity we use binary gender (male/female) and racial (White/Black) breakdowns. Clearly, this dichotomy fails to capture complexities in both gender and racial identification and perceptions. Second, the targeted dataset multiplicity models that we use also over-simplify both how discrimination manifests and how it can interact with other identities not captured by the data. Finally, we are not social scientists or domain experts and it is possible that the folk wisdom we rely on to propose data biases does not fully capture the patterns in the world. Rather, readers should treat this section as an analysis of 'toy phenomena' meant to illustrate how our technique can be used for real-world tasks.

*Basic results.* First, we present the overall dataset multiplicity robustness rates for the various multiplicity definitions.

**Key insights:** Limiting label errors to a subset of the training dataset (i.e., refining $\mathcal{M}$) yields higher dataset multiplicity robustness rates. However, the exact choice of $\mathcal{M}$ is significant.

Figure 3 shows that in all cases the targeted multiplicity definition yields significantly higher overall robustness rates than a broad multiplicity definition does. Notably, limiting all label perturbations to one racial group for Income greatly affects robustness: using the original multiplicity definition (no targeting), no test samples are robust when 12% of labels can be modified. However, when limiting label errors to Black people with the negative label, the robustness rate remains over 95% — it turns out this is not surprising, since fewer than 2% of the data points have race=Black. However, more than 80% of samples have race=White, and limiting label changes to White people with the positive label still yields over 60% robustness

**Table 1: Robustness rates (percentage of test dataset whose predicted label cannot change under dataset label multiplicity) for classification datasets given different rates of inaccurate labels.**

| Dataset | Inaccurate labels as a percentage of training dataset size | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.1% | 0.25% | 0.5% | 0.75% | 1.0% | 1.5% | 2.0% | 3.0% | 4.0% | 5.0% | 6.0% |
| LAR | 93.9 | 88.1 | 79.4 | 69.2 | 61.3 | 45.9 | 33.7 | 16.2 | 5.2 | 0.4 | 0.0 |
| Income | 91.1 | 81.4 | 67.8 | 58.4 | 50.7 | 37.2 | 23.3 | 12.1 | 4.8 | 1.7 | 0.7 |
| MNIST 1/7 | 98.3 | 96.3 | 93.1 | 88.3 | 84.4 | 73.1 | 60.8 | 38.8 | 23.3 | 13.1 | 7.0 |

when 12% of labels can be changed. Similarly, using targeted dataset multiplicity definitions for LAR can increase overall robustness by up to 30 percentage points.

*Demographic group robustness rates.* We also investigate the robustness rates for different demographic groups, both under the original, untargeted multiplicity assumptions and under the targeted versions.

**Key insights:** Different demographic groups do not exhibit the same dataset multiplicity robustness rates and targeting $\mathcal{M}$ to reflect real-world uncertainty can exacerbate discrepancies.

Each row of Figure 4 compares baseline (untargeted) robustness rates, stratified by demographic groups, with targeted versions of $\mathcal{M}$ for five US states. We observe two trends: first, there are commonly racial and gender discrepancies (see the pairs of dotted lines). E.g., for all states except Wisconsin, men consistently have higher baseline robustness rates than women (sometimes by a margin of over 20%). Second, using various targeted versions of $\mathcal{M}$ has unequal impacts across demographic groups. The top row of Figure 4 shows that targeting on race=Black (i.e., allowable label perturbations can change Black people's labels from $-1$ to 1) modestly improves dataset multiplicity robustness rates for Black people, but massively improves them for White people. We see similar trends, namely, that the non-targeted group sees higher robustness rate gains than the targeted group, across the other versions of $\mathcal{M}$, as well.

On LAR, similar results hold. (Graphs and discussion are in the appendix.)

### 4.3 Approximate Approach

**Key insights:** Using the approximate approach greatly reduces precision in proving dataset-multiplicity robustness, but still shows promise for understanding dataset multiplicity's impact given low levels of label errors.

As expected, the approximate approach from Section 3.4 is less precise than the exact one. The loss in precision depends highly on the dataset and the level of label uncertainty, as shown in Figure 5. For Income and MNIST, there are very large gaps: for example, given 2% label error, 80% of test samples are robust to dataset multiplicity, but the approximate version cannot prove robustness for any samples. However, there are some bright points: at 1% label error, we can still prove robustness for 90% of MNIST-1/7 samples, and over 60% of Income samples. In situations where label error is expected to be relatively small, the approximate approach can still be useful for gauging the relative dataset multiplicity robustness of a dataset.

We also measured the time complexity of each approach. To check the robustness of 1,000 Income samples, it takes 37.2 seconds for the exact approach and 6.8 seconds for the approximate approach. For 10,000 samples, it takes 383.5 seconds and 30.7 seconds, respectively. I.e., the exact approach scales linearly with the number of samples, but the approximate approach stays within a single order of magnitude. See appendix for more details and discussion.

## 5 IMPLICATIONS AND ETHICAL CHALLENGES OF DATASET MULTIPLICITY

For the conclusions we draw from machine learning to be robust and generalizable, we need to *understand* dataset multiplicity, *reduce* its impacts on predictions, and *adapt* machine learning practices to consider dataset multiplicity.

*Understanding dataset multiplicity.* Adopting standardized data documentation practices will likely aid in identifying potential inaccuracies or biases in datasets [23, 43]. Further work surrounding *how* and *why* datasets are created with particular worldviews (e.g., [26, 50]) will assist in identifying blind spots in existing datasets and help further the push for more robust dataset curation and documentation. But even if specific shortcomings in data collection are addressed through better curation and documentation practices, unavoidable variations in data collection will still contribute to dataset multiplicity [45], making modeling and model deployment interventions important, too.

*Reducing the impacts of dataset multiplicity.* An advantage of predictive multiplicity (i.e., multiplicity in the model selection process given a fixed training dataset) is that the wide range of equally-accurate models allows model developers to choose a model based on criteria like fairness or robustness without sacrificing predictive accuracy [6, 67]. Likewise, we know that there exist datasets — typically de-biased or more representative than a naïvely-collected baseline dataset — that yield models that are both fair and accurate [11, 21, 64]. If there exists a dataset in the dataset multiplicity set that yields a fairer (or more robust, more interpretable, etc.) model, then we should consider whether it is more appropriate to use that dataset to train the deployed model. (This may or may not be appropriate, depending on the domain, and is a decision that should be considered in conjunction with stakeholders.) Another option is to find learning algorithms or model classes that are inherently more robust to dataset multiplicity. In the context of ridge regression, we found that using a larger regularization parameter (see Figure 2) increases dataset-multiplicity robustness. Ensemble

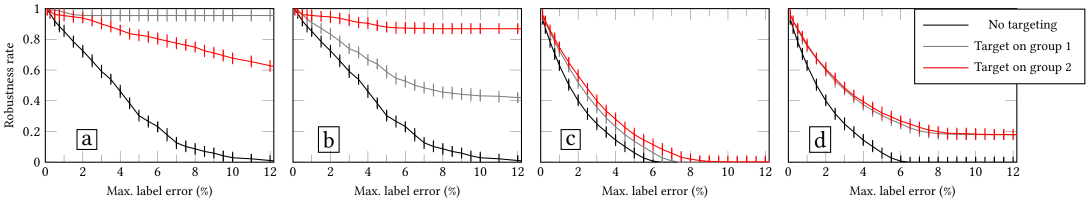

**Figure 3: Dataset multiplicity robustness rates for (a) Income stratified by race, (b) Income stratified by gender, (c) LAR stratified by race, (d) LAR stratified by gender. "No targeting" means that we place no restrictions on which labels can be modified. "Target on group 1" indicates that we can modify labels for group 1 (the minority/disadvantaged group) from the negative to positive class, while "target on group 2" indicates that we can modify labels for group 2 (the majority/advantaged group) from the positive to the negative class. For the race plots, Group 1 is Black people and Group 2 is White people. For the gender plots, Group 1 is women and Group 2 is men.**

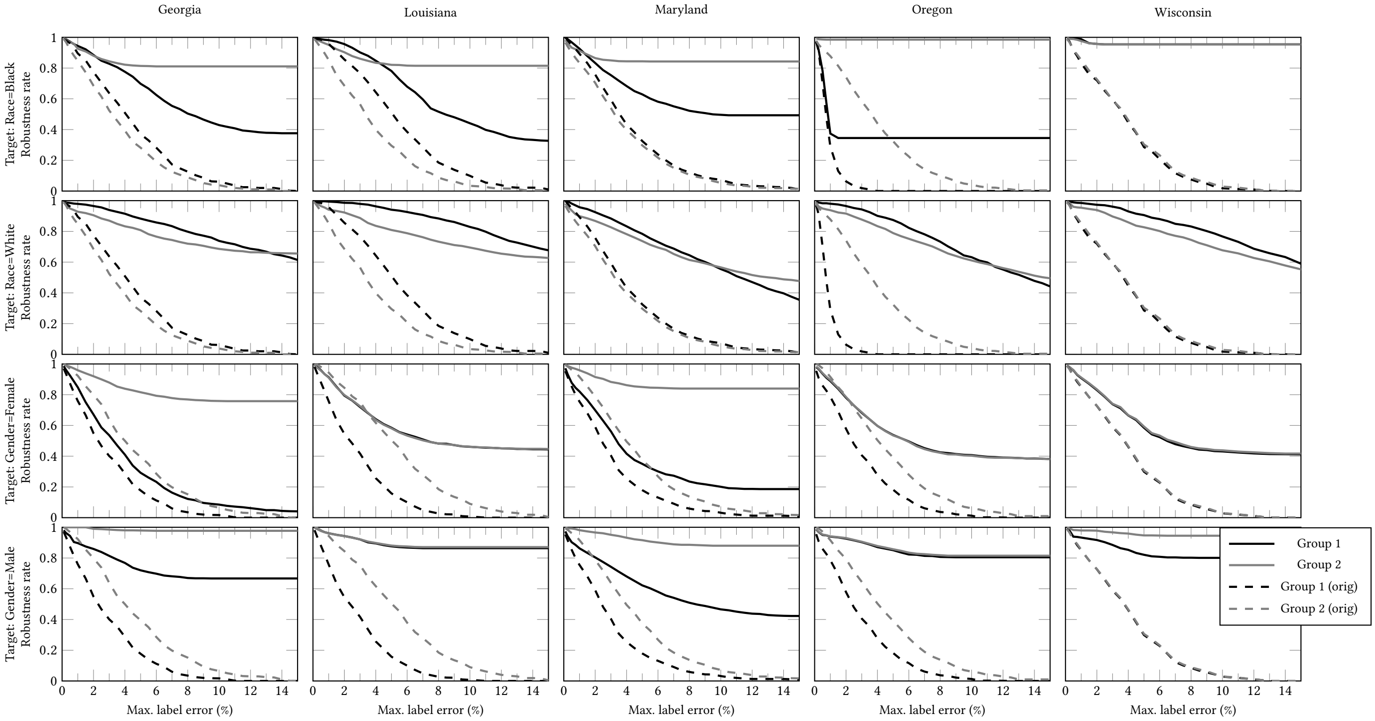

**Figure 4: Dataset multiplicity robustness rates given targeted label errors across several states' Income datasets. Row 1: target on race=Black, flipping labels to -1 to 1. Row 2: target on race=White, flipping labels from 1 to -1. Row 3: target on gender=female, flipping labels from -1 to 1. Row 4: target on gender=male, flipping labels from 1 to -1. For the graphs targeting on race, Group 1 is Black people and group 2 is White people. For the graphs targeting on gender, Group 1 is women and Group 2 is men. "Orig" refers to the baseline, untargeted models.**

learning is another promising direction, as it has been shown to decrease predictive multiplicity [5].

*Adapting ML to handle non-robustness to dataset multiplicity.* If a model has low dataset multiplicity robustness in aggregate across a test dataset, then confidence may be too low to deploy the model because of procedural fairness concerns [6]. It is also important to consider how robustness to dataset multiplicity varies across different population subgroups. As we saw in Section 4.2, different multiplicity definitions can yield disparate multiplicity-robustness rates across populations. The approximate approach from Section 3.4 is well-suited to these aggregate analyses. If we find that dataset multiplicity rates are low (either overall or for some demographic groups), it may be more appropriate to train with a different algorithm, refine the training dataset so that multiplicity is lessened, or avoid machine learning for the task at hand.

Dataset multiplicity robustness should also be considered on an individual level, as in the exact approach from Section 3.3. If

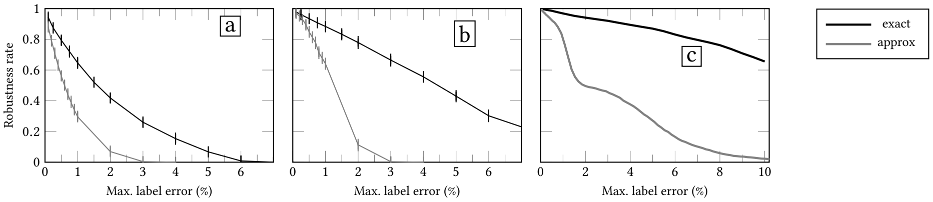

**Figure 5: Left to right: Robustness rates (fraction of the test set) for the exact and approximate techniques on (a) LAR, (b) Income, and (c) MNIST 1/7. For all examples, $\lambda$ was chosen to obtain results within 2% of the optimal accuracy. Error bars (for Income and LAR) are the median 50% for 10-fold cross validation.**

a given dataset and algorithm are not dataset multiplicity-robust on a test sample $\mathbf{x}$, options include abstaining on $\mathbf{x}$ or using the most favorable outcome in its dataset multiplicity range. But in many cases, an algorithm is deployed to allocate a finite resource — thus, returning the best-case label for all samples is likely infeasible. Instead, the chosen model is a function of the arbitrary nature of the provided dataset. Creel and Hellman explain that arbitrary models are not necessarily cause for concern, however, algorithmic monoculture becomes a concern when the same arbitrary outcomes are used widely, thereby broadly excluding otherwise-qualified people from opportunities [15, 31]. Given machine learning's reliance on benchmark datasets, it seems plausible that there is algorithmic monoculture stemming from the choice of arbitrary training dataset. To avoid algorithmic monoculture effects in the modeling process, scholars have proposed randomizing over model selection or outcomes [24, 31] — we suspect that a similar approach would make sense in the context of dataset multiplicity.

## 6 RELATED WORK

Dataset multiplicity robustness can be used either to *certify* (i.e., prove) that individual predictions are stable given uncertainty in the training data, or to characterize the overall stability of a model. Other approaches in the realms of adversarial ML, uncertainty quantification, and learning theory aim to answer similar questions.

*Comparison with other sources of multiplicity.* Predictive multiplicity and underspecification show that there are often many models that fit a given dataset equally well [9, 16, 33, 60]. Because of this multiplicity, models can often be selected to simultaneously achieve accuracy and also fairness, robustness, or other desirable model-level properties [6, 14, 51, 60, 67]. The extent of predictive multiplicity can be lessened by constructing more sophisticated models (e.g., [48]), however, this type of intervention only reduces algorithmic multiplicity and, furthermore, does not address the underlying procedural fairness concern that individuals can justifiably receive different decisions. Multiplicity also arises when modifying training parameters like random seed, data ordering, and hyperparameters [8, 13, 34, 54, 56], but most of the works on this topic focus on either attack vulnerability or the reproducibility and generalizability of the training process. In a notable exception, Bell et al. explore the 'multiverse' of models by characterizing what hyperparameter values correspond with what conclusions [1], but their analysis does not account for uncertainty in the training dataset, nor does the predictive multiplicity literature. There is, however, a line of work that aims to increase the fairness of a model by debiasing or augmenting a dataset [11, 21, 64]. Our dataset multiplicity framework is more broad than those approaches since we aim to understand the entire range of feasible datasets and models, rather than identify a single fair alternative.

*Other approaches to bounding uncertainty.* Approaches from causal inference, uncertainty quantification, and learning theory aim to measure and reduce uncertainty in machine learning. One major concern in this realm is the role that researcher decisions can play in reproducibility [12, 55]. Coker et al. propose 'hacking intervals' to capture the range of outcomes that any realistic researcher could obtain through different analysis choices or datasets [12]. Our dataset multiplicity framework can be viewed as extending their prescriptively constrained hacking-interval concept to allow for arbitrarily-defined changes to the training dataset. However, their results employ causality to make a stronger case for defining reasonable dataset modifications. Likewise, partial identification in economics uses domain knowledge and statistical tools to bound the range of possible outcomes in a data analysis [59].

The methods described above — 'hacking intervals' and partial identification — are special formulations of uncertainty quantification (UQ), which aims to understand the range of predictions that a model may output. UQ can occur either through Bayesian methods that treat model weights as random variables, or through ensembling or bootstrapping [58]. While UQ shares a common goal with dataset multiplicity — i.e., understanding the range of outcomes — the assumptions about where the multiplicity arises are different. UQ typically assumes that uncertainty stems from either insufficient data or noisy data, and does not account for the systemic errors that dataset multiplicity can encompass. Work on selection bias aims to learn in the presence of missing data, feature, or labels. For example, multiple imputation fills the missing data in multiple ways and aggregates the results [49, 62], similar to how dataset multiplicity considers all alternative models. The main difference between multiple imputation as a selection bias intervention and dataset multiplicity is that given selection bias, it is easy to identify where the inaccuracies are, and multiple imputation only considers a small number of dataset options, rather than all options as dataset multiplicity aims to do.

Within learning theory, distributional robustness studies how to find models that perform well across a family of distributions [2, 37, 53], robust statistics shows how algorithms can be adapted to account for outliers or other errors in the data [18, 19], and various works focus on robustifying training algorithms to label noise [38, 40, 46]. However, these works all (a) provide *statistical* global robustness guarantees, rather than the provable *exact* robustness guarantees that we make, (b) try to find a single good classifier, rather than understand the range of possible outcomes, and (c) typically require strong assumptions about the data distribution and the types of noise or errors. Algorithmic stability [9, 17] and sensitivity analysis [25] aim to quantify how sensitive algorithms are to small perturbations in the training data. However, they both typically make strict assumptions about the perturbation's form, either as a leave-one-out perturbation model in algorithmic stability [4, 30], or as random noise in sensitivity analysis.

*Robustness in adversarial ML.* Checking robustness to dataset multiplicity has parallels to adversarial machine learning, especially data poisoning, where an attacker modifies a small portion of the training dataset to reduce test-time accuracy [3, 52, 66, 68]. Various defenses counteract these attacks [29, 41, 47, 57, 69], including ones that focus on attacking and defending linear regression models [28, 36]. Some of these works (e.g., [47] for label-flipping) additionally try to *certify* robustness. Our exact solution to dataset multiplicity in linear models with label errors functions as a certificate, since we prove robustness to all allowable label perturbations, including adversarial ones. Three major differences from Rosenfeld et al. [47] are that we do not modify the underlying algorithm to achieve a certificate, the certification process is deterministic, not probabilistic, and we allow the label perturbations to be targeted towards a particular subgroup. Meyer et al. use a similar targeted view on data modifications, but their approach is strictly overapproximate and is limited to decision trees [35]. Furthermore, our definition of dataset multiplicity is distinct from the *defense* papers in that we aim to study dataset multiplicity robustness of existing algorithms; however, an interesting direction for future work would be to improve dataset multiplicity robustness via algorithmic modifications.

## 7 CONCLUSIONS

We defined the dataset multiplicity problem, showed how to evaluate the impacts of dataset multiplicity on linear models in the presence of label noise, and presented thoughts for how dataset multiplicity should be considered as part of the machine learning pipeline. Notably, we find that we can certify robustness to dataset multiplicity for some test samples, indicating that we can deploy these predictions with confidence. By contrast, we show that other test samples are not robust to low levels of dataset multiplicity, meaning that unless labels are very accurate, these test samples may receive predictions that are artifacts of the random nature of data collection, rather than real-world patterns.

Future work in the area of dataset multiplicity abounds, and many connections with other areas are mentioned throughout Sections 5 and 6. The most important technical direction for future exploration, in our opinion, is extending the dataset multiplicity framework to probabilistic settings, e.g., by asking what proportion of reasonable datasets yield a different prediction for a given test sample. This inquiry is likely to be more fruitful than finding exact solutions for more complicated model classes, and it may open opportunities to leverage techniques from areas like distributional robustness or uncertainty quantification. Making direct connections with areas like causal inference and partial identification in economics should also be a priority for future work, as these topics have similar goals and have been studied more broadly. In a social-science realm, dataset multiplicity could benefit from more work on what features and labels in a dataset are most likely to be inaccurate or affected by social biases, since this will allow us to bound dataset multiplicity more precisely. Similarly, it is difficult to separate out instances of direct bias (e.g., salary disparities due to gender discrimination) and indirect bias (e.g., salary disparities due to women feeling unwelcome in STEM careers), and more research is needed into how that distinction should affect dataset multiplicity definitions.

## ACKNOWLEDGMENTS

This research was supported by NSF grants 1750965, 1763871, 2211968, 1918211, 2023222 and a Microsoft Faculty Fellowship. We thank the anonymous reviewers for their detailed and thoughtful reviews.

## REFERENCES

[1] Samuel Bell, Onno Kampman, Jesse Dodge, and Neil D Lawrence. 2022. Modeling the Machine Learning Multiverse. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=8OH6t0YQGPJ

[2] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science* 59, 2 (2013), 341–357. https://doi.org/10.1287/mnsc.1120.1641 arXiv:https://doi.org/10.1287/mnsc.1120.1641

[3] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning Attacks against Support Vector Machines. In *Proceedings of the 29th International Conference on Machine Learning* (Edinburgh, Scotland) *(ICML'12)*. Omnipress, 1467–1474.

[4] Emily Black and Matt Fredrikson. 2021. Leave-One-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, 285–295. https://doi.org/10.1145/3442188.3445894

[5] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HfUyCRBeQc

[6] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. https://doi.org/10.1145/3531146.3533149

[7] Francine D. Blau and Lawrence M. Kahn. 2017. The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature* 55, 3 (September 2017), 789–865. https://doi.org/10.1257/jel.20160995

[8] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. 2021. Accounting for Variance in Machine Learning Benchmarks. In *Proceedings of Machine Learning and Systems*, A. Smola, A. Dimakis, and I. Stoica (Eds.), Vol. 3. 747–769. https://proceedings.mlsys.org/paper/2021/file/cfecdb276f634854f3ef915e2e980c31-Paper.pdf

[9] Leo Breiman. 1996. Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics* 24, 6 (1996), 2350 – 2383. https://doi.org/10.1214/aos/1032181158

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[11] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is My Classifier Discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3543–3554.

[12] Beau Coker, Cynthia Rudin, and Gary King. 2021. A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results. *Management Science* 67, 10 (2021), 6174–6197.

[13] A. Feder Cooper, Yucheng Lu, Jessica Forde, and Christopher M De Sa. 2021. Hyperparameter Optimization Is Deceiving Us, and How to Stop It. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 3081–3095. https://proceedings.neurips.cc/paper/2021/file/17fafe5f6ce2f1904eb09d2e80a4cbf6-Paper.pdf

[14] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 2144–2155. https://proceedings.mlr.press/v139/coston21a.html

[15] Kathleen Creel and Deborah Hellman. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy* 52, 1 (2022), 26–43. https://doi.org/10.1017/can.2022.3

[16] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. https://doi.org/10.48550/ARXIV.2011.03395

[17] L. Devroye and T. Wagner. 1979. Distribution-Free Performance Bounds for Potential Function Rules. *IEEE Transactions on Information Theory* 25, 5 (1979), 601–604. https://doi.org/10.1109/TIT.1979.1056087

[18] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2021. Robustness Meets Algorithms. *Commun. ACM* 64, 5 (2021), 107–115. https://doi.org/10.1145/3453935

[19] Ilias Diakonikolas and Daniel M. Kane. 2019. Recent Advances in Algorithmic High-Dimensional Robust Statistics. arXiv:1911.05911 [cs.DS]

[20] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *arXiv preprint arXiv:2108.04884* (2021).

[21] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. 2020. 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 (ICML'20)*. JMLR.org, Article 263, 11 pages.

[22] Federal Financial Institutions Examination Council. 2019. One Year National Loan-Level Dataset. https://ffiec.cfpb.gov/data-publication/2019 Accessed 5 Jan. 2023.

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (nov 2021), 86–92. https://doi.org/10.1145/3458723

[24] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2017. On Fairness, Diversity and Randomness in Algorithmic Decision Making. https://doi.org/10.48550/ARXIV.1706.10208

[25] Ali S Hadi and Samprit Chatterjee. 2009. *Sensitivity Analysis in Linear Regression.* John Wiley & Sons.

[26] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[27] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[28] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In *2018 IEEE Symposium on Security and Privacy (SP)*. 19–35. https://doi.org/10.1109/SP.2018.00057

[29] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. 2021. Intrinsic Certified Robustness of Bagging against Data Poisoning Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 9 (May 2021), 7961–7969. https://ojs.aaai.org/index.php/AAAI/article/view/16971

[30] Michael Kearns and Dana Ron. 1999. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation* 11, 6 (1999), 1427–1453. https://doi.org/10.1162/089976699300016304

[31] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic Monoculture and Social Welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118. https://doi.org/10.1073/pnas.2018340118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2018340118

[32] Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. [n.d.]. The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/

[33] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 6765–6774. https://proceedings.mlr.press/v119/marx20a.html

[34] Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. 2020. Individual Differences among Deep Neural Network Models. *Nature Communications* 11 (2020). Issue 1. https://doi.org/10.1038/s41467-020-19632-w

[35] Anna P. Meyer, Aws Albarghouthi, and Loris D' Antoni. 2021. Certifying Robustness to Programmable Data Bias in Decision Trees. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 26276–26288. https://proceedings.neurips.cc/paper/2021/file/dcf531edc9b229acfe0f4b87e1e278dd-Paper.pdf

[36] Nicolas Müller, Daniel Kowatsch, and Konstantin Böttinger. 2020. Data Poisoning Attacks on Regression Learning and Corresponding Defenses. In *2020 IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC)*. 80–89. https://doi.org/10.1109/PRDC50213.2020.00019

[37] Hongseok Namkoong and John C Duchi. 2016. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/4588e674d3f0faf985047d4c3f13ed0d-Paper.pdf

[38] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf

[39] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. https://openreview.net/forum?id=XccDXrDNLek

[40] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2233–2241. https://doi.org/10.1109/CVPR.2017.240

[41] Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. 2019. Label Sanitization Against Label Flipping Poisoning Attacks. In *ECML PKDD 2018 Workshops*, Carlos Alzate, Anna Monreale, Haytham Assem, Albert Bifet, Teodora Sandra Buda, Bora Caglayan, Brett Drury, Eva García-Martín, Ricard Gavaldà, Irena Koprinska, Stefan Kramer, Niklas Lavesson, Michael Madden, Ian Molloy, Maria-Irina Nicolae, and Mathieu Sinn (Eds.). Springer International Publishing, Cham, 5–15.

[42] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2 (11 2021). Issue 11. https://doi.org/10.1016/j.patter.2021.100336

[43] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1776–1826. https://doi.org/10.1145/3531146.3533231

[44] Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. 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)*. https://openreview.net/forum?id=j6NxpQbREA1

[45] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5389–5400. https://proceedings.mlr.press/v97/recht19a.html

[46] David Rolnick, Andreas Veit, Serge J. Belongie, and Nir Shavit. 2017. Deep Learning is Robust to Massive Label Noise. *CoRR* abs/1705.10694 (2017). arXiv:1705.10694 http://arxiv.org/abs/1705.10694

[47] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. 2020. Certified Robustness to Label-Flipping Attacks via Randomized Smoothing. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 8230–8241. http://proceedings.mlr.press/v119/rosenfeld20b.html

[48] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2022. Reconciling Individual Probability Forecasts. *arXiv preprint arXiv:2209.01687* (2022).

[49] Donald B. Rubin. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

[50] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, Article 317 (oct 2021), 37 pages. Issue

CSCW2. https://doi.org/10.1145/3476058

[51] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1827–1858. https://doi.org/10.1145/3531146.3533232

[52] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., 6106–6116.

[53] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. 2015. Distributionally Robust Logistic Regression. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, 1576–1584.

[54] I Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A Erdogdu, and Ross J Anderson. 2021. Manipulating SGD with Data Ordering Attacks. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 18021–18032. https://proceedings.neurips.cc/paper/2021/file/959ab9a0695c467e7caf75431a872e5c-Paper.pdf

[55] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (2011), 1359–1366. https://doi.org/10.1177/0956797611417632 arXiv:https://doi.org/10.1177/0956797611417632 PMID: 22006061.

[56] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems* 25 (2012).

[57] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. 2017. Certified Defenses for Data Poisoning Attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., 3520–3532.

[58] Timothy John Sullivan. 2015. *Introduction to Uncertainty Quantification*. Vol. 63. Springer.

[59] Elie Tamer. 2010. Partial Identification in Econometrics. *Annu. Rev. Econ.* 2, 1 (2010), 167–195.

[60] Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. 2022. Predicting Is Not Understanding: Recognizing and Addressing Underspecification in Machine Learning. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 458–476.

[61] Catalina Toma, Jeffrey Hancock, and Nicole Ellison. 2008. Separating Fact From Fiction: An Examination of Deceptive Self-Presentation in Online Dating Profiles. *Personality & Social Psychology Bulletin* 34 (09 2008), 1023–36. https://doi.org/10.1177/0146167208318067

[62] Stef Van Buuren. 2018. *Flexible Imputation of Missing Data*. CRC press.

[63] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. 2022. When Does Dough Become a Bagel? Analyzing the Remaining Mistakes on ImageNet. https://doi.org/10.48550/ARXIV.2205.04596

[64] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf

[65] Megan Wisniewski. 2021. *In Puerto Rico, No Gap in Median Earnings between Men and Women.* Technical Report. United States Census Bureau. https://www.census.gov/library/stories/2022/03/what-is-the-gender-wage-gap-in-your-state.html

[66] Han Xiao, Huang Xiao, and Claudia Eckert. 2012. Adversarial Label Flips Attack on Support Vector Machines. In *Proceedings of the 20th European Conference on Artificial Intelligence* (Montpellier, France) *(ECAI'12)*. IOS Press, 870–875.

[67] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=WHqVVk3UHr

[68] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM* 64, 3 (Feb. 2021), 107–115. https://doi.org/10.1145/3446776

[69] Xuezhou Zhang, Xiaojin Zhu, and Stephen Wright. 2018. Training Set Debugging Using Trusted Items. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://ojs.aaai.org/index.php/AAAI/article/view/11610

# A ADDITIONAL DETAILS FROM SECTION 3

*A note on the relationship between $\epsilon$ and $\Delta$.* Given a fixed $\epsilon$, if $\delta_i$ is the same for all $i$, the ratio between $\delta$ and $\epsilon$ uniquely determines robustness to dataset multiplicity (we make use of this fact for computing Table 2).

## A.1 Details about Algorithm 1

Algorithm 2 is the complete algorithm for the approach described in Section 3.3. Note that this algorithm supersedes Algorithm 1.

We define the *positive potential impact* $\rho_i^+$ as the maximal positive change that perturbing $\tilde{\mathbf{y}}_i$ can have on $\mathbf{zy}$, likewise, $\rho_i^-$ is the *negative potential impact*, that is, the maximal negative change that perturbing $\tilde{\mathbf{y}}_i$ can have on $\mathbf{zy}$.

Algorithm 2 finds the minimal label perturbation necessary to change the label of test point $\mathbf{x}$, a fact we formalize in the following theorem:

Theorem A.1. *Suppose we have a deterministic learning algorithm $A$, a training dataset $D = (\mathbf{X}, \mathbf{y})$ where up to $k$ labels $\mathbf{y}_i$ corresponding to data points $\mathbf{X}_i$ that satisfy $\phi(\mathbf{X}_i)$ are inaccurate by $+/-\Delta$. Let $\mathcal{M}(D)$ be the set of all datasets that can be constructed by modifying $D$ according to $k$, $\phi$, and $\Delta$. Let $F = A(\mathcal{M}(D))$ be the set of models $f$ obtainable by training using $A$ on any dataset $D' \in \mathcal{M}(D)$. Given a test point $\mathbf{x}$, (i) Algorithm 2 outputs an interval containing all values $f(\mathbf{x})$ for all $f \in F$ (i.e., the algorithm is sound) and (ii) there is some $f_1, f_2 \in F$ such that $f_1(\mathbf{x})$ is equal to the upper bound of the output and $f_2(\mathbf{x})$ is equal to the lower bound of the output (i.e., the algorithm is tight).*

Proof. We will prove (i) that the upper bound Theorem A.1's output is an upper bound on the value of $f(\mathbf{x})$ for any $f \in F = A(\mathcal{M}(D))$ and (ii) that this upper bound is achieved by some $f_1 \in F$. The proofs for the lower bounds are analogous.

(i) Let $u$ be the upper bound of Theorem A.1's output. We want to show that $f(\mathbf{x}) \leq u$ for all $f \in F = A(\mathcal{M}(D))$, where

$$\mathcal{M}(D = (\mathbf{X}, \mathbf{y})) = \{(\mathbf{X}, \mathbf{y}') \mid \|\mathbf{y}' - \mathbf{y}\|_1 \leq k \wedge \\ \mathbf{y}_i \neq \mathbf{y}'_i \implies \phi(\mathbf{X}_i) \wedge \|\mathbf{y}_i - \mathbf{y}'_i\| \leq \Delta\}$$

Suppose, towards contraction, that there is some $f' \in A(\mathcal{M}(D))$ such that $f'(\mathbf{x}) = u' > u$. So, there is a set of labels $y_{i_1}, y_{i_2}, \ldots, y_{i_k}$ that can be modified to create a $\mathbf{y}'$ such that $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}'\mathbf{x} = u'$. Recall that $z = \mathbf{x}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. So, either (a), we modify the same set of labels, but modify at least one in a different magnitude or direction, or (b), there must be some $i_j$ that we modify $y_{i_j}$ to find $u'$, but the algorithm does not identify this index in line 8 of the algorithm.

First, suppose (a) occurred. WLOG, suppose the if case on line 2 is satisfied, i.e., $z_{i_j} \geq 0$. Then we hypothetically modified $y_{i_j}$ by $a \neq \delta^u_{i_j}$ in place of line 11. We know $a < \delta^u_{i_j}$ since $\delta$ is an upper bound on how much we can change each label. We have $\delta^u_{i_j} \geq 0$ and $z_{i_j} \geq 0$, so their product is also greater than 0, so $az_{i_j} < \delta^u_{i_j} z_{i_j}$. So, the final product $\mathbf{zy}'$ cannot be larger than had we followed the algorithm.

Now, suppose (b) occurred. Suppose $y_{i_j}$ is modified to yield $u'$, but is not modified in the algorithm. Then, there must be some $y_{i'}$ such that (assume WLOG that $\mathbf{z}_{i'} \geq 0$ and $\mathbf{z}_{i_j} \geq 0$) $\mathbf{z}_{i'}\delta^u_{i'} \geq \mathbf{z}_{i_j}\delta^u_{i_j}$. If equality holds, we will have $u = u'$. So, assume $\mathbf{z}_{i'}\delta^u_{i'} > \mathbf{z}_{i_j}\delta^u_{i_j}$. But then, modifying $\mathbf{y}_{i'}$ by $\delta^u_{i'}$ will result in a greater increase to $mathbfz\mathbf{y}$ than increasing $\mathbf{y}_{i_j}$ by $\delta^u_{i_j}$ will. So, changing $y_{i_j}$ cannot result in an output $u' > u$.

(ii) We need to construct the function $f \in F = A(\mathcal{M}(D))$ such that $f(\mathbf{x}) = u$, where $u$ is the upper bound of Theorem A.1's output. Let $y_{i_1}, y_{i_2}, \ldots, y_{i_k}$ be the labels modified by lines 15-19 of the algorithm to yield some $\mathbf{y}'$. Then, let $f(x) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}'$. □

**Algorithm 2** Solve for $V = [\min_{(\mathbf{X},\tilde{\mathbf{y}})\in\mathcal{M}_{k,\Delta,\phi}((\mathbf{X},\mathbf{y}))} \mathbf{z}\tilde{\mathbf{y}}, \max_{(\mathbf{X},\tilde{\mathbf{y}})\in\mathcal{M}_{k,\Delta,\phi}((\mathbf{X},\mathbf{y}))} \mathbf{z}\tilde{\mathbf{y}}]$ by finding perturbations of $\mathbf{y}$ that maximally decrease/increase $\mathbf{zy}$.

```
Require: $\mathbf{z} \in \mathbb{R}^n$, $(\mathbf{X}, \mathbf{y}) \in (\mathbb{R}^{n\times d}, \mathbb{R}^n)$, $\Delta \in \mathbb{IR}^n$ with $0 \in \Delta$, $k \geq 0$,
    $\phi : \mathbb{R}^d \to \{0, 1\}$
1:  $\mathbf{y}^l \leftarrow \mathbf{y}$ and $\mathbf{y}^u \leftarrow \mathbf{y}$
2:  if $z_i \geq 0$ then
3:      $\rho_i^+ \leftarrow \mathbf{z}_i\delta_i^u$ ,  $\rho_i^- \leftarrow \mathbf{z}_i\delta_i^l$
4:  else
5:      $\rho_i^+ \leftarrow \mathbf{z}_i\delta_i^l$ ,  $\rho_i^- \leftarrow \mathbf{z}_i\delta_i^l$
6:  if not $\phi(\mathbf{x}_i)$ then
7:      $\rho_i^+ \leftarrow 0$,  $\rho_i^- \leftarrow 0$
8:  Let $\rho^+_{i_1}, \ldots, \rho^+_{i_l}$ be the $k$ largest elements of $\rho^+$ by absolute value
9:  for each $\rho^+_{i_j}$ do
10:     if $z_{i_j} \geq 0$ then
11:         $(y^u)_{i_j} \leftarrow (y^u)_{i_j} + \delta^u_{i_j}$
12:     else
13:         $(y^u)_{i_j} \leftarrow (y^u)_{i_j} + \delta^l_{i_j}$
14: Let $\rho^-_{i_1}, \ldots, \rho^-_{i_l}$ be the $k$ largest elements of $\rho^-$ by absolute
    value
15: for each $\rho^-_{i_j}$ do
16:     if $z_{i_j} \geq 0$ then
17:         $(y^l)_{i_j} \leftarrow (y^l)_{i_j} + \delta^l_{i_j}$
18:     else
19:         $(y^l)_{i_j} \leftarrow (y^l)_{i_j} + \delta^u_{i_j}$
20: $V = [\mathbf{zy}^l, \mathbf{zy}^u]$
```

## A.2 Details on the Approximate Approach

We will next present an example to show that $\Theta$ can be non-convex.

Example A.1. *Suppose* $\mathbf{y} = (1, -1, 2)$, $\Delta = [-1, 1]^3$, *and* $k = 2$. *Given* $\mathbf{C} = \begin{pmatrix} 1 & 2 & 1 \\ -1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix}$, *we have*

$$\Theta = \mathbf{Cy} \cup \left\{\mathbf{C}\begin{pmatrix} a \\ b \\ 2 \end{pmatrix}\right\} \cup \left\{\mathbf{C}\begin{pmatrix} a \\ -1 \\ c \end{pmatrix}\right\} \cup \left\{\mathbf{C}\begin{pmatrix} 1 \\ b \\ c \end{pmatrix}\right\}$$

*for* $a \in [0, 2]$, $b \in [-2, 0]$, *and* $c \in [1, 3]$.

*Note that* $(3, 6, 3)^\top \in \Theta$ *and* $(4, 5, 2)^\top \in \Theta$, *but their midpoint* $(3.5, 5.5, 3.5)^\top \notin \Theta$, *thus,* $\Theta$ *is non-convex.*

We present the complete algorithm for procedure described in Section 3.4 Algorithm 3. Next, we will show that this algorithm outputs the tightest hyperrectangular (i.e., box) enclosure of $\mathcal{M}$.

Theorem A.2. *Algorithm 3 computes the tightest hyperrectangular enclosure of* $\mathcal{M}$.

Proof. First, we will show that Algorithm 3 computes an enclosure of $\mathcal{M}$, and next we will show that this output is the tightest hyperrectangular enclosure of $\mathcal{M}$.

By construction of the algorithm, we see that the output will be an enclosure of $\mathcal{M}$. The algorithm constructs the maximal way to increase/decrease each coordinate.

Now, suppose there is another hyperrectangle $\Theta^{a\prime}$ that with $\mathcal{M}(D) \subseteq \Theta^{a\prime}$ and ${\Theta^{a}}'_{i} \subset \Theta^{a}_{i}$. WLOG, assume that the upper bound of ${\Theta^{a}}'_{i}$ is strictly less than the upper bound of $\Theta^{a}_{i}$. But $\Theta^{a}_{i} = \max_{(\mathbf{X},\mathbf{y}') \in \mathcal{M}(D)} (\mathbf{z}_i \mathbf{y}')$, which means that $\mathbf{y}^* = \max_{(\mathbf{X},\mathbf{y}') \in \mathcal{M}(D)} \mathbf{C}\mathbf{y}'$ has $\mathbf{C}_i \mathbf{y}^*$ greater than the upper bound of ${\Theta^{a}}'_{i}$, and thus $\Theta^{a\prime}$ is not a sound enclosure of $\Theta$. □

**Algorithm 3** Computing $\Theta^a$

```
Require: $C \in \mathbb{R}^{d \times n}, (\mathbf{X}, \mathbf{y}) \in (\mathbb{R}^{n \times d}, \mathbb{R}^n), \Delta \in \mathbb{IR}^n$ with $0 \in \Delta$,
    $k \geq 0, \phi : \mathbb{R}^d \rightarrow \{0, 1\}$
$\mathbf{y}^l \leftarrow \mathbf{y}$ and $\mathbf{y}^u \leftarrow \mathbf{y}$
$\Theta^a \leftarrow [0, 0]^d$
for $i$ in range $d$ do:
    $\mathbf{y}^l \leftarrow \mathbf{y}, \ \mathbf{y}^u \leftarrow \mathbf{y}$
    if $c_{ij} < 0$ then
        $\rho^+_j \leftarrow \mathrm{C}_{ij}\delta^l_j , \ \rho^-_j \leftarrow \mathrm{C}_{ij}\delta^u_j$
    else
        $\rho^+_j \leftarrow \mathrm{C}_{ij}\delta^u_j , \ \rho^-_j \leftarrow \mathrm{C}_{ij}\delta^l_j$
    if not $\phi(\mathbf{x}_i)$ then
        $\rho^+_j \leftarrow 0, \ \rho^-_j \leftarrow 0$
    Let $\rho^+_{k_1}, \ldots, \rho^+_{k_l}$ be the $k$ largest elements of $\rho^+$ by absolute value.
    for each $\rho^+_{k_j}$ do
        if $c_{ik_j} \geq 0$ then
            $(y^u)_{k_j} \leftarrow (y^u)_{k_j} + \delta^u_{k_j}$
        else
            $(y^u)_{k_j} \leftarrow (y^u)_{k_j} + \delta^l_{k_j}$
    Let $\rho^-_{k_1}, \ldots, \rho^-_{k_l}$ be the $k$ largest elements of $\rho^-$ by absolute value.
    for each $\rho^-_{k_j}$ do
        if $c_{ik_j} \geq 0$ then
            $(y^l)_{k_j} \leftarrow (y^l)_{k_j} + \delta^l_{k_j}$
        else
            $(y^l)_{k_j} \leftarrow (y^l)_{k_j} + \delta^u_{k_j}$
    $\Theta^a_i \leftarrow [\mathbf{c}_i \mathbf{y}^l, \mathbf{c}_i \mathbf{y}^u]$
```

# B ADDITIONAL EXPERIMENTS

We present additional tables, graphs, and discussion about the experimental results.

*Accuracy.* The maximal accuracy for each dataset (i.e., in Figure 2, the accuracy for the 0.0 line) is 76.5% for Income, 61.9% for LAR, and 98.9% for MNIST. The exact values we used for $\lambda$ (as well as the procedure to obtain $\lambda$) are in the code.

*Additional LAR data.* Figure 6 shows demographic-stratified dataset-multiplicity robustness rates for LAR under different ways of defining targeted dataset multiplicity. To varying extents, the majority/advantaged group sees higher robustness rates as compared with the disadvantaged group across all versions of $\mathcal{M}$.

*Regression dataset results.* Table 2 presents results on Income-Reg for the fixed robustness radius $\epsilon$ = \$2,000, which we chose as a challenging, but reasonable, definition for two incomes being close. We also empirically validated that the ratio between $\Delta$ and $\epsilon$ uniquely determines robustness for a fixed multiplicity definition. Notably, for small $\Delta$ to $\epsilon$ ratios (i.e., when we can modify labels by small amounts, but predictions can be far apart and still considered robust), many test points are dataset-multiplicity robust, even when the number of untrustworthy training labels is relatively large (up to 10%). However, when this ratio is large, e.g., when $\Delta = 5\epsilon$, we are still able certify a majority of test points as robust up to 1% bias. For example, if we can modify 1% of labels by up to \$10,000, then 69.6% of test samples' predictions cannot be modified by more than \$2,000. By contrast, only allowing 1% of labels to be modified by up to \$4,000 yields a dataset-multiplicity robustness rate of 91.2% (again, within a radius of \$2,000).

*Income demographics.* Table 3 shows the demographic make-up of various states' data from the Income dataset. Notice that Oregon has the lowest percentage of Black people in the dataset - we suspect that this is why the robustness rates between White and Black people is so large.

Figure 7 shows robustness rates, stratified by race or gender, for five U.S. states on the Income dataset. We see that for most states, there is a significant gap in robustness rates across with race and gender. In particular, Georgia and Louisiana have higher robustness rates for Black people, and Oregon has *drastically* higher robustness rates for White people. All states, except Wisconsin, have higher robustness rates for men than women.

Figure 8 shows the demographic group-level robustness rates under the over-approximate technique.

## B.1 Running time

Table 4 shows the running time of our techniques, as evaluated on a 2020 MacBook Pro with 16GB memory and 8 cores. These times should be interpreted as upper bounds; in practice, both approaches are amenable to parallelization, which would yield faster performance. We notice that both the exact approach scales linearly with the number of samples, while the approximate approach stays within a single order of magnitude as the number of samples grows.Clearly, the approximate approach is more scalable for checking the robustness of large numbers of data points.

Table 5 shows the running time of the MILP solver for Income and LAR. We see that the times scale linearly (as with our exact approach), but are much worse. To check robustness for 100 samples, it is over 80% slower to use MILP.

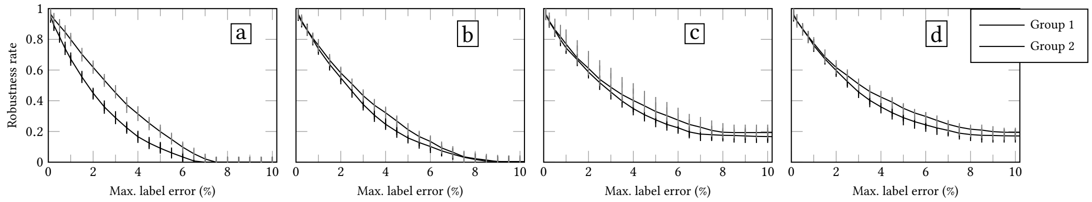

**Figure 6: Targeted label multiplicity for LAR. (a) only allows errors in labels for Black people (flip labels from -1 to 1), (b) only allows errors in labels for White people (flip labels from 1 to -1), (c) only allows errors in labels for women (flip labels from -1 to 1), and (d) only allows errors in labels for men (flip labels from 1 to -1). In graphs (a) and (b), group 1 is Black people while group 2 is White people. In graphs (c) and (d), group 1 is women and group 2 is men. In all graphs, the error bars represent the middle 50% of values across 10-fold cross validation.**

**Table 2: Robustness rates (percentage of test dataset whose prediction cannot change by more than $\epsilon$) for Income-Reg given various $\Delta$ and $k$ values. $\epsilon = 2{,}000$ in all experiments. Note that the shorthand $\Delta = a$ means $\Delta = [-a, a]^n$. Column 2 gives the ratio between the maximum label perturbation ($\Delta$) and the robustness radius ($\epsilon$), which uniquely determines robustness for a given $k$.**

| $\Delta$ | Ratio $\frac{\Delta}{\epsilon}$ | Maximum label error $k$ as a percentage of training dataset size | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1.0% | 2.0% | 3.0% | 4.0% | 5.0% | 6.0% | 7.0% | 8.0% | 9.0% | 10.0% |
| 1,000 | 0.5 | 100.0 | 100.0 | 100.0 | 100.0 | 99.8 | 99.4 | 99.1 | 98.6 | 98.0 | 97.1 |
| 2,000 | 1 | 100.0 | 96.6 | 91.1 | 85.4 | 84.1 | 76.4 | 73.3 | 64.0 | 49.9 | 35.2 |
| 4,000 | 2 | 91.2 | 84.2 | 69.2 | 35.7 | 14.2 | 2.0 | 0 | 0 | 0 | 0 |
| 6,000 | 3 | 85.9 | 61.1 | 15.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8,000 | 4 | 80.1 | 20.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10,000 | 5 | 69.6 | 3.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

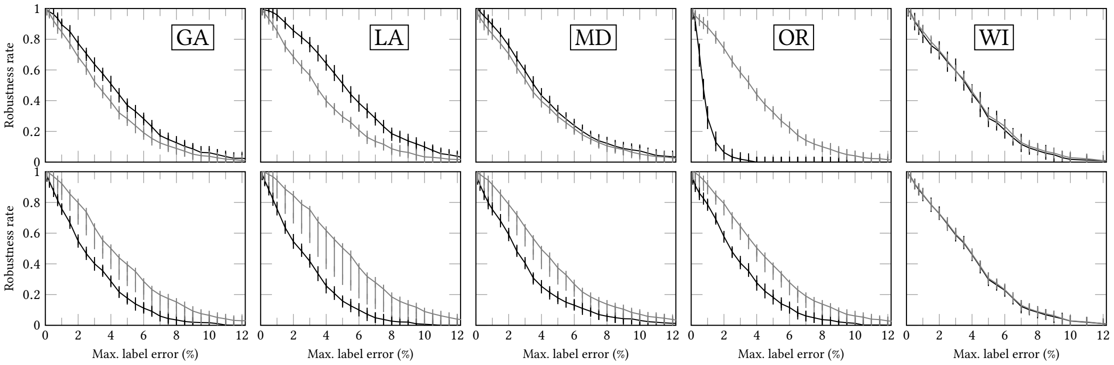

**Figure 7: Row 1: Robustness rate for White people (gray lines) and Black people (black lines) across 5 different states of the Income dataset. Row 2: Robustness rates for men (gray lines) and women (black lines) across 5 different states of the Income dataset. Error bars show averages across 10 folds.**

**Table 4: Running time, in seconds, for exact and approximate experiments. The exact experiments flip labels for each data point until the sample is no longer robust. The approximate experiments flip 10% of the labels (which is enough to bring the robustness to 0%).**

| Dataset | 100 samples | | 1,000 samples | | 10,000 samples | |
|---|---|---|---|---|---|---|
| | Exact | Approx. | Exact | Approx. | Exact | Approx. |
| Income | 2.5 | 4.4 | 37.2 | 6.8 | 383.5 | 30.7 |
| LAR | 7.5 | 1.8 | 73.5 | 2.6 | 730.2 | 10.6 |
| MNIST 1/7 | 4.1 | 3.8 | 24.8 | 6.0 | 448.5 | 24.3 |

**Table 5: Running time, in seconds, for the MILP solver. We modify 10% of the labels.**

| Dataset | 10 samples | 100 samples |
|---|---|---|
| Income | 49.1 | 498.8 |
| LAR | 58.1 | 616.6 |

**Table 3: Summary of data download by state from the Folktables Income task.**

| State | training $n$ | % White | % Black |
|---|---|---|---|
| Georgia | 40731 | 67.6 | 23.9 |
| Louisiana | 16533 | 70.9 | 23.5 |
| Maryland | 26433 | 63.6 | 23.5 |
| Oregon | 17537 | 86.4 | 1.4 |
| Wisconsin | 26153 | 92.7 | 2.6 |

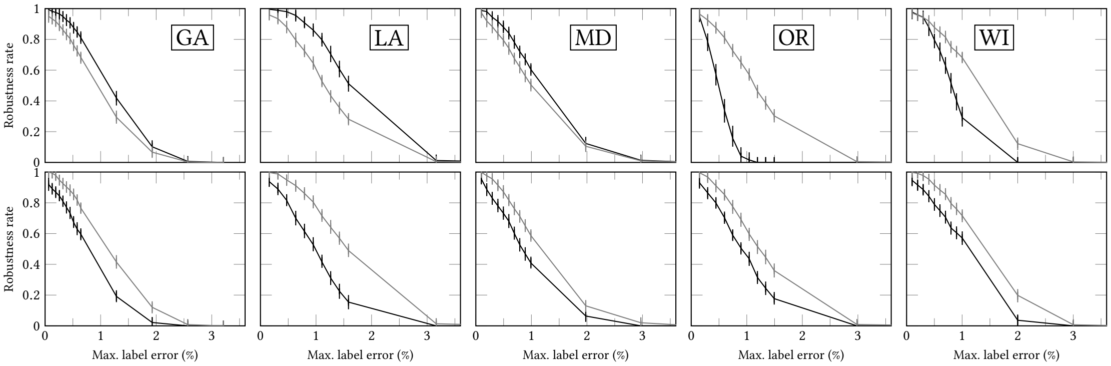

**Figure 8: Row 1: Robustness rate under the over-approximate approach for White people (gray lines) and Black people (black lines) across 5 different states of the Income dataset. Row 2: Robustness rates for men (gray lines) and women (black lines) across 5 different states of the Income dataset. Error bars show averages across 10 folds.**