Randomized Response Has No Disparate Impact on Model Accuracy

Alycia N. Carey, Karuna Bhaila, Xintao Wu
Department of Electrical Engineering and Computer Science
University of Arkansas
{ancarey, kbhaila, xintaowu}@uark.edu

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386574

Abstract—Differential privacy, the current gold standard for data anonymization and protection, is commonly known to cause degraded utility, and exacerbate unfairness, for different demographic groups when it is used to train a private machine learning model. However, in contrast with this long-held perception, recent work has shown that local differential privacy, a variant of differential privacy where users perturb their data on their device before it is aggregated, can surprisingly lead to improved fairness measures without significantly affecting the utility of the underlying machine learning model. Motivated by this previous work, in this paper we further show that applying randomized response, a popular local differential privacy method, does not incur disparate impact on the private model's accuracy for different demographic groups. Specifically, through conducting thorough empirical analysis in which we perform randomized response on the labels, the features, or on both the features and labels across multiple data modalities and model architectures, we empirically show that the absolute difference in utility loss for different demographic groups is negligible.

Index Terms—randomized response, local differential privacy, fairness, machine learning

I. Introduction

Differential privacy (DP) [1] is currently the gold standard for data anonymization and protection. Due to its formal privacy guarantees, DP is commonly utilized by companies such as Google [2], Apple [3], and Microsoft [4] to collect privatized statistics from their end users. In addition to being used in privacy preserving data collection, there has been considerable research on using DP in the training of private machine learning models [5]. DP can be introduced at any stage of the training pipeline, from data collection to model output, and it provides assurance to data contributors that the probability of their data being leaked is bound by a privacy parameter ϵ. One important research question is to determine whether enforcing DP incurs accuracy disparity between different demographic subgroups. In other words, does making a machine learning model DP cause the accuracy of different demographic subgroups to change in a non-uniform manner?

Why accuracy parity? While making sure that machine learning models perform equally well on different demographics is important for mitigating allocation harms like not awarding jobs to qualified applicants on the basis of gender, or representative harms such as Amazon labeling LGBT+ literature as 'adult content' and removing it from sales rankings [6], it is equally important to ensure that applying privacy preserving mechanisms like randomized response or output perturbation does not cause disparate impact on the accuracy of different demographic subgroups where some subgroups experience a greater change in accuracy than others. In doing so, we prevent reinforcing the social idea that minority groups inherently do not deserve the same privacy affordances that majority groups experience (e.g., monitoring of American Muslims by the New York City Police Department [7] and using machine learning to determine someone's sexuality [8]).

When global DP (a variant of DP in which user data is collected unperturbed and noise is added during training of the model or on the output of the query) is used in the training of a machine learning model, it is common that DP-stochastic gradient descent (DP-SGD) [9] is utilized. However, while DP-SGD has shown success in training DP machine learning models, it has recently been shown that using DP-SGD causes the accuracy of the model to drop more for underrepresented subgroups [10]. Since the publication of [10], several works have been published showing that applying DP methods such as DP-SGD, or post-processing differentially private data [11], when learning a machine learning model can be detrimental to fairness from the accuracy parity perspective [12].

Yet, there has been no study on if local DP (LDP) – a variant of DP which eliminates the requirement of a trusted aggregation server – causes accuracy disparity. In this work, we aim to answer this question and analyze to what extent that accuracy (dis)parity occurs when randomized response (RR), a popular LDP method, is applied. While other advanced LDP methods based on RR, such as Optimal Linear Hashing [13] and RAPPOR [2], have been proposed we leave the analysis of accuracy (dis)parity experienced when these methods are applied to our future work. Further, in our setting we consider a privatized model to be fair if the minority group and the majority group experience similar changes in accuracy (i.e., experience accuracy parity) from the original model. We note that in our setting it does not matter if the original accuracies of the two subgroups differ. As long as they experience the same effect on accuracy after applying RR, then the process of applying RR does not result in disparate impact.

To consider the many ways that RR is applied in practice, we choose to study the effect of applying RR in three main settings: 1) applying RR-based ϵ-LDP to the labels while leaving the features unperturbed, 2) applying RR-based ϵ-LDP

to the features while leaving the labels unperturbed, and 3) applying RR-based ϵ-LDP to both the features and the labels. We note that we only ever apply RR to the training data, while the testing data remains unperturbed as we are most concerned with protecting the privacy of the data used to train the machine learning model. Our empirical evaluations, performed across both image and tabular data modalities, several different model architectures, and a wide range of ϵ values, show that applying RR-based ϵ-LDP to the features, labels, or both the features and labels has no disparate impact on model accuracy.

The rest of this work is as follows. In Section II, we introduce key related works and discuss how our work is different from them. In Section III, we give an overview of RR-based ϵ-LDP and in Section IV we explain the reconstruction methods used (inclusive of why reconstruction methods are used at all). Section V details our experimental setting and provides an analysis of our results of applying RR-based ϵ-LDP on the labels, the features, and on both the features and labels to show that RR has no disparate impact on model accuracy. Finally, in Section VI we offer our concluding remarks.

II. Related Work

Analyzing the effect of DP on the fairness of a machine learning model (and vice versa) has received increasing attention by the research community over the past few years. [14] analyzed the privacy risks of group fairness metrics like equalized odds through the lens of membership inference attacks and show that the information leakage of fair mod-els increases significantly on the unprivileged subgroups. As opposed to analyzing the privacy leakage of fairness metrics, [15]–[18] study how to achieve both privacy and fairness simultaneously and propose methods to ensure that fairness does not come at the cost of privacy. Different from the above works that define fairness as the learned model achieving uniform utility among all demographic groups, our paper's focus is on the difference in accuracy changes experience by different demographic groups when RR-based LDP is applied to train a private machine learning model. In the following paragraphs, we discuss related works along this direction.

In the decision making setting, [19] focuses on decision making tasks (e.g., how to assign voting rights benefits to minority language communities) and shows that under strict privacy constraints, or in decisions involving small populations, adding noise in order to achieve global DP can cause significant inequities in treatment to arise. They further identify multiple causes of these outcome disparities. The decision task is generally defined as a computational formula that can be decomposed into units connected by basic mathematical (e.g., $+, -, \times, \div$) and logical (e.g., AND/OR) operators and in [20] the authors examined how these various operations affect the accuracy of complex DP computations.

In DP preserving machine learning, [10] was the first to show that DP-SGD has disparate impact as the accuracy of a model trained using DP-SGD decreases more on underrepresented and complex classes and subgroups when compared to the original, non-private, model. This work motivated others, such as [12] and [21], who both shed light on why using DP mechanisms leads to disparate impact. But while [12] focused on DP empirical risk minimization through output perturbation and DP-SGD, analyzing the data and model properties that are responsible for causing the disparate impact, [21] studied the conditions in which privacy and fairness have aligned or contrasting goals in both decision and learning tasks. Several research works further study mitigation techniques to remove disparate impact caused by DP-SGD [22], [23].

Despite analyzing the interface of DP and fairness being a popular research field, to our knowledge, there has been no study about accuracy (dis)parity in the LDP setting. [24] studied to train a Naive Bayes classifier over LDP data by estimating probabilities needed by the Naive Bayes classifier using the perturbed data. However, it did not show whether unfairness could incur in the trained classifier. [25] studied how to train a gradient boosting decision tree over LDP data (perturbed by RR and other mechanisms) and showed that using LDP can lead to slightly improved group fairness metrics in learning problems without significantly affecting the performance of machine learning models. While they considered fairness in terms of statistical fairness metrics like equal opportunity difference (EOD) and overall accuracy difference (OAD), our work instead focuses on the study of disparate impact on model accuracy due to RR. Moreover, we show that the accuracy of a model trained using private data via RR, no matter if randomization is performed on the labels, the features, or the features and labels, does not decrease differently for different demographic groups. On the contrary, [25] only focuses on the setting of feature perturbation.

III. Preliminary

DP was originally proposed to privatize information about individuals in the context of databases [5], although it has grown to be used in a myriad of settings such as machine and deep learning. In the global DP setting, noise is often added to the output of a query made on a non-privatized dataset. Since the type of noise that is added is known a priori, statistical queries can still be computed by filtering out the noise without violating any of the users' individual privacy [26].

However, global DP requires each user to have trust that the central server is non-malicious. LDP [27] was proposed to overcome the limitation of requiring each user to trust a centralized authority. In LDP, each user perturbs their own data before transmitting them to the (untrusted) server, which can then compute statistical queries over the randomized client data [26]. More formally, ϵ-LDP is defined as follows:

Definition 1 (ϵ-LDP [27]): A randomized mechanism M satisfies ϵ-local differential privacy if and only if for any pair of input values r, r' in the domain of M, and for any possible output o in the range of M:

$$P[M(r) = o] \leq e^{\epsilon} \cdot P[M(r') = o]$$

Definition 1 states that the probability of outputting o on record r is at most e^{ϵ} times the probability of outputting o

on record r'. Here, ϵ captures the privacy loss of the system. When $\epsilon = 0$, perfect privacy is achieved ($e^0 = 1$). On the other hand, when $\epsilon = \infty$, there are no privacy guarantees on the system. The choice of ϵ is a crucial decision in practice as the increase in privacy risks is proportional to e^{ϵ} [5], [26].

Randomized Response. The main motivation behind LDP is a surveying technique termed randomized response (RR) [28]. Proposed by Warner in 1965 [28], RR allows more accurate statistics to be collected about sensitive topics such as sexual orientation or drug use. To define RR, let u be a private variable that can take one value from $C = \{1, 2, \ldots, C\}$. RR is defined as a $C \times C$ distortion matrix $P = (p_{uv})_{C \times C}$ where $p_{uv} = P[v|u]$ C denotes the probability that the output of the RR process is v C when the real attribute value is u C [29]. When u = v, $p_{uv} = \frac{e^{\epsilon}}{C-1+e^{\epsilon}}$ and when u = v, $p_{uv} = \frac{1}{C-1+e^{\epsilon}}$. We note that each variable (feature or label) being perturbed has its own (possibly non-distinct) distortion matrix P_i. In Section IV, we show how the gathered randomized variables can be reconstructed to form an unbiased representation of the original, non-randomized, population data. Additionally, while numerous LDP protocols have been derived from RR such as RAPPOR [2], Optimal Local Hashing [13], [30], and Thresholding with Histogram Encoding [13], in this work we choose to analyze RR only due to space constraints and leave analysis over other methods to future work. We hypothesize, however, that methods derived from RR should exhibit results similar to those shown in Section V.

IV. Methodology

While performing RR alone protects the privacy of users data, when the randomized data is used directly to answer queries or train a machine learning model, accuracy issues can arise as the training distribution could be different from the testing. For this reason, reconstruction is usually applied to the collected randomized data. If the distortion matrices for each private variable is known, then the true population distribution can be estimated from the noisy collected data.

Let $z = \{z_1, \ldots, z_n\}$ be the collection of randomized data with $z_s = (x_s, y_s)$ and $x = \{x_1, \ldots, x_m\}$. In other words, each record z_s is made up of m features and one label y and it is possible for any combination of the features and label to have been randomized. Let each feature x_u have d_u mutually exclusive and exhaustive possible values while the label y can have d_y mutually exclusive and exhaustive values. Further, let $i_u = 1, \ldots, d_u$ denote the index of each feature x_u's categories and let $i_y = 1, \ldots, d_y$ denote the index of the label y's possible categories. For each feature x_u and label y, assume the distortion matrix P_u or P_y is known. In cases where a feature (or label) is not randomized, the distortion matrix P_u (P_y) is simply the $d_u \times d_u$ ($d_y \times d_y$) identity matrix.

Let $\pi_{i_1,\ldots,i_m,i_y}$ denote the true proportion corresponding to the categorical combination of m variables and one label y $(x_{1i_1}, \ldots, x_{mi_m}, y_{i_y})$ in the original, non-randomized, data. Here, x_{1i_1} denotes the i_1th category of feature x_1 and y_{i_y} denotes the i_yth category of label y. Let π be a vector with the elements $\pi_{i_1,\ldots,i_m,i_y}$ listed in order (e.g., $\pi = \{\pi_{1,\ldots,1,1}, \ldots, \pi_{d_1,\ldots,d_m,d_y}\}$). Similarly, we denote $\lambda_{i_1,\ldots,i_m,i_y}$ as the expected proportion of categorical combinations in the randomized data and λ is a vector with the elements $\lambda_{i_1,\ldots,i_m,i_y}$ listed in order. For a concrete example, consider a dataset where each record contains two features and one label. E.g., x = {race, gender} and y {1, 2} where race {Black, White, Asian} and gender {male, female}. Here, $d_1 = 3$ and $d_2 = 2$ while $d_y = 2$. In this example the vector $\pi = \{\pi_{111}, \pi_{112}, \pi_{121}, \pi_{122}, \pi_{211}, \pi_{212}, \pi_{221}, \pi_{222}, \pi_{311}, \pi_{312}, \pi_{321}, \pi_{322}\}$ lists all proportions of categorical combinations of the race and gender features with the label. Note, π_{312} denotes the proportion of records that are Asian males with a label of 2.

Letting $P = P_1 \cdots P_m P_y$ we can obtain an unbiased representation of π as:

$$\hat{\pi} = P^{-1}\lambda = (P_1^{-1} \cdots P_m^{-1} P_y^{-1})\lambda \quad (1)$$

where stands for the Kronecker product and P_u^{-1} (P_y^{-1}) denotes the inverse of the distortion matrix P_u (P_y).

This reconstruction method works well for tabular datasets that contain few features with limited categories. However, when the feature space is large (e.g., images), reconstruction can be computationally expensive, or even computationally infeasible, to compute. For this reason, other techniques to correct noisy data have been proposed. In this work, we focus on a correction method called forward loss correction proposed in [31] to correct for noisy labels and leave other methods to correct for noisy data in general as future work.

Forward Loss Correction. FLC [31] is an approach to train machine learning models robust to class-dependent label noise (not necessarily noise crafted by RR). In [31], the authors note that a model trained on noisy label data without using a loss correction method would result in the model being tailored to predict noisy labels instead of the true labels. To perform FLC, the authors correct the model predictions using a probability matrix that defines the noisy data distribution (in our case, P) before calculating the loss between the model prediction and the noisy labels $\tilde{y}$ $\tilde{Y}$. FLC is defined as:

$$L^{FLC}(\tilde{Z}, \theta) = \frac{1}{n}\sum_{s=1}^{n} \ell(P^T f(x_s; \theta), \tilde{y}_s) \quad (2)$$

where $\tilde{Z} = \{\tilde{z}_s = (x_s, \tilde{y}_s)\}_{s=1}^{n}$ and ℓ is a proper composite loss[1] [32] such as cross-entropy or square loss. Here, $f(x; \theta) = P(y \mid x)$. In other words, the output of f is the predicted probability of the class being y when the input is x. We can rewrite Eq. 2 to explicitly show that ℓ is a proper composite loss:

$$L_{\psi}^{FLC}(\tilde{Z}, \theta) = \frac{1}{n}\sum_{s=1}^{n} \ell(P^T \psi^{-1}(f(x_s; \theta)), \tilde{y}_s) \quad (3)$$

Here, ψ is the link function associated with a particular proper loss. For example, softmax is the inverse link function for cross-entropy. When FLC is applied while minimizing a proper

[1]A proper loss is a loss function that both predicts the binary classification label as well as provides an estimate of the probability that an example will have positive label. Proper losses are called proper composite losses when a link function is used to map the output of the predictor to the interval [0, 1] in order for the output to be interpreted as a probability [32].

composite loss function, [31] notes that the minimizer of the corrected loss under the noisy distribution is the same as the minimizer of the original loss under the clean distribution:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} L_{\psi}^{FLC}(\tilde{Z}, \theta) = \arg\min_{\theta \in \Theta} L_{\psi}(Z, \theta) \quad (4)$$

In other words, the learned model will make correct predictions on future non-randomized test data. For brevity, we refer readers to [31] for an in-depth discussion of FLC. We note that in Section V we use FLC when image data is being tested and the reconstruction method of Eq. 1 in all other cases.

V. Experimentation

In this section, we empirically validate our claim that applying RR to the labels, the features, or the features and labels has no disparate impact on the accuracy of the resulting privatized model. We test several different ϵ values, specifically $\epsilon = \{0.001, .01, .1, .25, .5, 1, 2, 5\}$, and we run each image (tabular) experiment 5 (100) times, reporting the average results. We provide the code used in our experimentation at https://tinyurl.com/3jhwd5cc.

A. Datasets

We use the UTKFace [33], CI-MNIST [34], and two datasets from the Folktables repository [35] (ACSIncome and ACSEmployment) in our experimentation. The UTKFace dataset consists of 23,705 face images with annotations of age, gender, and ethnicity. In experiments using the UTKFace dataset, we choose gender as the label and ethnicity as the sensitive attribute. The CI-MNIST dataset is a variant of MNIST where the authors introduced different types of correlations between dataset features and eligibility criterion. For an input image x, the label y $\{0, 1\}$ indicates eligibility or ineligibility, respectively, given that x is even or odd. The dataset defines the background colors as the protected or sensitive attributes where blue denotes the underprivileged group and red denotes the privileged group. In this work, we let 40% of the 50,000 images have a blue background (20% of the even images and 60% of the odd) while 60% of the images have red backgrounds (80% of even, 40% of odd). The two Folktables datasets ACSIncome (Income) and ACSEmployment (Employment) are based on the 1-year American Community Survey Public Use Microdata Sample (ACS PUMS) from California in 2018. Income is a replacement for the Adult dataset and the task is to predict whether an individual's income is above $50,000 after filtering the ACS PUMS data sample to only include records of those over the age of 16, who worked at least 1 hour per week, and have an income of at least $100. After cleaning, the income dataset has 189,954 data points and we considered gender as the sensitive feature. The main task of Employment is to predict whether an individual is employed, after the ACS PUMS data sample is filtered to only include records of people between the ages of 16 and 90. After cleaning, the employment dataset has 112,569 data points and we consider race as the sensitive feature. For all datasets, we used a train/test split of 80/20. We note that we only apply RR on the training dataset (either on the labels, features, or features and labels) while the testing set is left unperturbed and that we only consider binary labels (and features on tabular data). For brevity, we leave details of the exact pre-processing and cleaning steps performed on the datasets to our GitHub repository.

TABLE I
Accuracy per demographic subgroup without randomized response

	Income		Employment	
	Male	Female	White	Non-White
LR	.722	.726	.585	.626
NB	.722	.727	.590	.622
LGBM	.723	.727	.599	.630

	UTKFace		CI-MNIST	
	White	Non-White	Red	Blue
ResNet-18	.719	.734	.779	.670
VGG-16	.770	.781	.849	.794
DenseNet-169	.768	.768	.862	.761

B. Architecture

We test several architectures to see if the model type has an effect on the degree of accuracy disparity that occurs between demographic groups when RR is applied. Specifically, we use ResNet-18, VGG-16, and DenseNet-169 for images and logistic regression (LR), Naive Bayes (NB), and LightGBM (LGBM) for tabular data. We employ the scikit-learn Python package and for simplicity, all hyperparameters and settings were set to the default scikit-learn values and each image model was trained for 100 epochs. Future experimentation could consider the effect of hyperparameters on accuracy disparity when RR is applied. Table I lists the average accuracy for each dataset using each model type.

C. Metrics

In this work, we aim to understand the accuracy disparity between different demographic groups when LDP via RR is applied. Specifically, we use the absolute difference between the original accuracy (Acc_{orig}) and the accuracy of the model under RR (Acc_{rr}):

$$\Delta_{acc} = |Acc_{orig} - Acc_{rr}| \quad (5)$$

where denotes the demographic group being considered. For example, Δ_{acc}^{f} denotes the change in accuracy of the female subgroup when RR is applied. We calculate Δ_{acc} for each demographic subgroup being considered (e.g., female and male) and compare the values to analyze the overall accuracy disparity. Specifically, we define accuracy disparity as:

$$Disp = |\Delta_{acc}^{A} - \Delta_{acc}^{B}| \quad (6)$$

where A, B are the two demographic groups being compared (e.g., A: male, B: female or A: white, B: non-white). While not a perfect measure for determining disparate impact, we choose to use the 80% rule from US discrimination law as our cut-off (i.e., $Disp \leq 0.20$) for determining if disparate impact has occurred on the accuracy or not.

TABLE II

CHANGE IN ACCURACY WHEN RANDOMIZED RESPONSE ϵ-LDP IS PERFORMED ON THE LABELS. BLUE: $.12 > |\Delta^A_{acc} - \Delta^B_{acc}| \geq .05$.

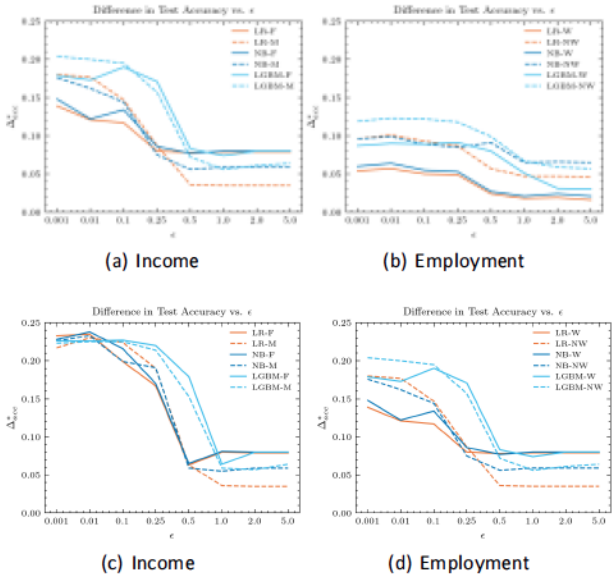
		ϵ															
		.001		.01		.1		.25		.5		1		2		5	
Dataset	Model	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}	Δ^A_{acc}	Δ^B_{acc}
UTKFace	ResNet-18	.208	.228	.202	.214	.172	.183	.127	.122	.057	.058	.015	.021	.005	.005	.009	.003
	VGG-16	.181	.203	.165	.185	.161	.171	.107	.113	.043	.046	.010	.010	.002	.004	.004	.001
	DenseNet-169	.254	.254	.238	.253	.187	.187	.112	.108	.045	.040	.014	.006	.007	.004	.009	.003
CI-MNIST	ResNet-18	.645	.540	.640	.543	.627	.513	.566	.489	.396	.359	.092	.084	.018	.012	.016	.025
	VGG-16	.723	.657	.671	.592	.274	.299	.129	.163	.077	.103	.033	.047	.011	.013	.0004	.001
	DenseNet-169	.741	.627	.730	.637	.709	.620	.629	.539	.362	.329	.055	.051	.006	.005	.007	.012
ACSIncome	LR	.216	.214	.170	.151	.049	.065	.038	.075	.035	.079	.035	.079	.035	.080	.035	.058
	NB	.219	.217	.162	.146	.048	.063	.050	.071	.056	.080	.058	.080	.059	.081	.058	.080
	LGBM	.216	.226	.178	.203	.066	.072	.056	.065	.057	.070	.059	.079	.063	.081	.064	.080
ACSEmployment	LR	.088	.123	.088	.118	.042	.075	.024	.059	.022	.053	.018	.050	.018	.047	.018	.046
	NB	.094	.118	.093	.114	.044	.078	.029	.067	.025	.065	.023	.065	.023	.066	.023	.065
	LGBM	.100	.130	.100	.128	.083	.105	.061	.076	.045	.062	.035	.062	.029	.059	.032	.058

D. Applying RR on the Labels

First, we examine the setting where RR ϵ-LDP is performed on the labels only while the features of the training dataset are left un-perturbed. We note that we perform reconstruction on the noisy data using Eq. 1 on the two tabular datasets and by using FLC on the two image datasets. The results are shown in Table II. Here, the results listed in blue mean $0.12 > |\Delta^A_{acc} - \Delta^B_{acc}| \geq 0.05$. In other words, the difference between the two demographic groups' changes in accuracy is above 0.05 but below 0.12. All other results have $|\Delta^A_{acc} - \Delta^B_{acc}| < 0.05$. We find that when RR is performed on the labels only, no absolute difference is above 0.12. Further, most of the experimental results fall below 0.05 meaning that little disparity occurred between the two demographics' changes in accuracy, and according to the 80% rule, disparate impact did not occur. In fact, the only time the accuracy disparity was above 0.05 was when the labels of the CI-MNIST dataset were perturbed with small ϵ values ($\epsilon \leq 0.25$). This higher accuracy disparity on the CI-MNIST dataset is not surprising. The CI-MNIST dataset has a total of 10 classes, and the distribution of the minority class to the majority class is non-uniform over all 10 classes. As $\epsilon \to 0$, the model simulates random guessing and since the red class started with higher accuracy, there was a bigger difference between the random guessing accuracy (10%) and the starting accuracy. Overall, it is clear that performing RR ϵ-LDP on the labels does not cause disparate impact on the privatized model's accuracy for different demographic groups.

E. Applying RR on the Features

In this section, we test the effect of applying RR ϵ-LDP with reconstruction by Eq. 1 on the features of the training data while leaving the labels unperturbed. Note, in this section as well as the following we only perform experimentation on the tabular datasets. We select three features to perturb on each of the tabular datasets: age, sex, and race. We note that each feature is perturbed with ϵ-LDP making the total privacy budget 3ϵ. The results are shown on the top row of Fig. 1. For the Income dataset, the highest difference in accuracy



(a) Income (b) Employment

(c) Income (d) Employment

Fig. 1. Change in accuracy when randomized response ϵ-LDP is performed on the features (top) or features and labels (bottom). LR: logistic regression, NB: Naive Bayes, LGBM: LightGBM. From left to right: income (F: female, M: male) and employment (W: white, NW: non-white).

occurred when $\epsilon = 0.01$ on the logistic regression model. In this setting Disp = 0.056, which is still fairly minimal and falls well below threshold suggested by the 80% rule. The next highest Disp value, 0.044, also occurred on the logistic regression model when $\epsilon \geq 1$. On the other two models, the highest Disp was 0.028 and, overall, the smallest Disp value was 0.005 which occurred at $\epsilon = 0.1$ on the LGBM model. The Employment dataset had slightly higher overall Disp values. The highest Disp value was 0.064 at $\epsilon = 0.5$ on the Naive Bayes model. The lowest Disp values was 0.016 at $\epsilon = 1$ on the LGBM model. In general, neither the model used, or the value of ϵ, had large effect on the disparity experienced between the change in accuracy of the two demographic groups and we conclude that performing RR ϵ-LDP on the features only does not have disparate impact on

the privatized model's accuracy.

F. Applying RR on the Labels and Features

In this section, we combine the settings of the previous two experiments and test the effect of applying RR ϵ-LDP with reconstruction by Eq. 1 on the features and the labels of the training data. The same three features as denoted in Section V-E are used and again, RR-based ϵ-LDP is applied independently on each feature as well as the label making the total privacy budget 4ϵ. We display the results of this experiment on the bottom row of Fig. 1. For the Income dataset, the highest Disp value was 0.044 and occurred when $\epsilon \geq 1$ on the logistic regression model. Additionally, the smallest Disp value, which was 0, also occurred on the logistic regression model at $\epsilon = 0.5$. The Employment dataset, as in the previous experiment, again experienced higher Disp values. Specifically, the highest Disp value was 0.084 at $\epsilon = 0.5$ on the logistic regression model and the smallest Disp value was 0.028 at $\epsilon = 0.25$ on the LGBM model. In general, these values are still well below the 0.2 threshold, and we conclude that performing RR on both the features and labels does not cause disparate impact on the privatized model's accuracy for different demographic groups.

VI. Conclusion

In this work, we empirically showed that performing randomized response does not cause disparate impact on the differentially private model's accuracy for different demographic subgroups. Specifically, for both tabular and image datasets, several different model architectures, and a variety of ϵ values, we showed that the absolute difference in utility loss for different demographic groups is negligible when randomized response is applied to the labels, the features, or the features and labels. In future work, we will explore the accuracy disparity experienced between different subgroups when advanced local differential privacy techniques are applied.

Acknowledgements

This work was supported in part by NSF 1920920, 1946391, and 2137335.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in TCC. Springer, 2006, pp. 265–284.
[2] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in SIGSAC, 2014, pp. 1054–1067.
[3] D. P. T. Apple Inc., "Learning with privacy at scale," Apple MLR, 7.
[4] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," NeurIPS, vol. 30, 2017.
[5] N. Ponomareva, H. Hazimeh, A. Kurakin, Z. Xu, C. Denison, H. B. McMahan, S. Vassilvitskii, S. Chien, and A. G. Thakurta, "How to dp-fy ml: A practical guide to machine learning with differential privacy," JAIR, vol. 77, pp. 1113–1201, 2023.
[6] K. Crawford. The trouble with bias. Keynote at NeurIPS 2017.
[7] D. Shamas and N. Arastu, "Mapping muslims: Nypd spying and its impact on american muslims," MACLC and CLEAR Project, 2013.
[8] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." J. Pers. Soc. Psychol., vol. 114, no. 2, p. 246, 2018.
[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in SIGSAC, 2016, pp. 308–318.
[10] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," NIPS, vol. 32, 2019.
[11] K. Zhu, F. Fioretto, and P. Van Hentenryck, "Post-processing of differentially private data: A fairness perspective," arXiv:2201.09425, 2022.
[12] C. Tran, M. Dinh, and F. Fioretto, "Differentially private empirical risk minimization under the fairness lens," NIPS, vol. 34, pp. 27 555–27 565, 2021.
[13] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in USENIX Security, 2017, pp. 729–745.
[14] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in EuroS&P. IEEE, 2021, pp. 292–303.
[15] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman, "Differentially private fair learning," in ICML. PMLR, 2019, pp. 3000–3008.
[16] H. Mozannar, M. Ohannessian, and N. Srebro, "Fair learning with private demographic data," in ICML. PMLR, 2020, pp. 7066–7075.
[17] D. Xu, S. Yuan, and X. Wu, "Achieving differential privacy and fairness in logistic regression," in Companion proceedings of The 2019 world wide web conference, 2019, pp. 594–599.
[18] C. Tran, F. Fioretto, and P. Van Hentenryck, "Differentially private and fair deep learning: A lagrangian dual approach," in AAAI, vol. 35, no. 11, 2021, pp. 9932–9939.
[19] D. Pujol, R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala, and G. Miklau, "Fair decision making using privacy-protected data," in FAccT, 2020, pp. 189–199.
[20] Y. Wang, X. Wu, J. Zhu, and Y. Xiang, "On learning cluster coefficient of private networks," Social network analysis and mining, vol. 3, pp. 925–938, 2013.
[21] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu, "Differen-tial privacy and fairness in decisions and learning tasks: A survey," arXiv:2202.08187, 2022.
[22] D. Xu, W. Du, and X. Wu, "Removing disparate impact on model accuracy in differentially private stochastic gradient descent," in SIGKDD, 2021, pp. 1924–1932.
[23] M. S. Esipova, A. A. Ghomi, Y. Luo, and J. C. Cresswell, "Disparate impact in differential privacy from gradient misalignment," arXiv preprint arXiv:2206.07737, 2022.
[24] E. Yilmaz, M. Al-Rubaie, and J. M. Chang, "Naive bayes classification under local differential privacy," in 2020 IEEE 7th International Conference On Data Science And Advanced Analytics (DSAA). IEEE, 2020, pp. 709–718.
[25] H. H. Arcolezi, K. Makhlouf, and C. Palamidessi, "(local) differential privacy has no disparate impact on fairness," arXiv:2304.12845, 2023.
[26] B. Bebensee, "Local differential privacy: a tutorial," arXiv:1907.11908, 2019.
[27] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" SIAM Journal on Computing, vol. 40, no. 3, pp. 793–826, 2011.
[28] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," JASA, vol. 60, no. 309, pp. 63–69, 1965.
[29] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection." in EDBT/ICDT Workshops, vol. 1558, 2016, pp. 0090–6778.
[30] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in ToC, 2015, pp. 127–135.
[31] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1944–1952.
[32] M. D. Reid and R. C. Williamson, "Composite binary losses," Journal of Machine Learning Research, vol. 11, pp. 2387–2422, 2010.
[33] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in IEEE CVPR, 2017, pp. 5810–5818.
[34] C. Reddy, D. Sharma, S. Mehri, A. Romero-Soriano, S. Shabanian, and S. Honari, "Benchmarking bias mitigation algorithms in representation learning through fairness metrics," https://github.com/charan223/FairDeepLearning, 2021.
[35] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," Advances in neural information processing systems, vol. 34, pp. 6478–6490, 2021.