Learning to Walk Impartially on the Pareto Frontier of Fairness, Privacy, and Utility

Abstract

Deploying machine learning (ML) models often requires both fairness and privacy guarantees. Both objectives often present notable trade-offs with the accuracy of the model—the primary focus of most applications. Thus, utility is prioritized while privacy and fairness constraints are treated as simple hyperparameters. In this work, we argue that by prioritizing one objective over others, we disregard more favorable solutions where at least certain objectives could have been improved without degrading any other. We adopt *impartiality* as a design principle: ML pipelines should not favor one objective over another. We theoretically show that a common ML pipeline design that features an unfairness mitigation step followed by private training is non-impartial. Then, parting from the two most common privacy frameworks for ML, we propose *FairDP-SGD* and *FairPATE* to train impartially specified private *and* fair models. Because impartially specified models recover the Pareto frontiers, i.e., the best trade-offs between different objectives, we show that they yield significantly better trade-offs than models optimized for one objective and hyperparameter-tuned for the others. Thus, our approach allows us to mitigate tensions between objectives previously found incompatible.

1 Introduction

Acknowledging that machine learning (ML) models can pose risks to society, it is a reasonable expectation that a regulatory body should produce technical specifications to curb the corresponding societal risks. We study the "specification problem" for trustworthy ML where the regulator needs to *specify the minimal levels of guarantee for fairness, privacy, and utility* (*e.g.* accuracy). This is crucial when deploying ML models in critical contexts with high-stake decisions—such as medical applications [16] and infrastructure planning with census data [8].

There are two broad approaches to combine fair and private learning: a) making DP-learning algorithms fair [39, 40, 35], or b) integrating privacy constraints into bias mitigation methods [17]. In practical implementations with either approach, however, not every objective receives equal attention: often, one objective is optimized while others are considered as (hyper)parametrized constraints.

The “pre-selection bias” introduced by looking only at certain ranges of hyperparameters during tuning which is necessary for practical reasons[1] puts socially-salient choices at the behest of algorithm designers and engineers. This can create potentially dangerous scenarios, such as introducing additional privacy leakage in the attempt to increase model fairness, or degrading fairness by introducing privacy [10, 31]. To avoid the pre-selection bias, we argue for exposing the inherent trade-off between the objectives by presenting a *Pareto frontier*. A Pareto frontier is the set of achievable guarantees for all objectives with the property that improving any guarantee from this set requires that we degrade the guarantee for another objective.

[1] This phenomenon is known as "omitted pay-off bias" in other contexts [20]

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

The Pareto frontier in Figure 1 helps demonstrate the pre-selection bias: If we starts with specifying arbitrary bounds on privacy budgets ($\varepsilon \leq 5$), and fairness violations ($\gamma \leq 0.1$), it can only recover a small portion of the Pareto frontier (highlighted in pink). Presented with these limited choices, we are forced to make suboptimal decisions: had we explored the full Pareto frontier, we could have achieved much better accuracy, possibly at a modest cost to fairness, or privacy. Nevertheless, this is a trade-off that we cannot observe, let alone choose, if we do not have the full picture at hand.

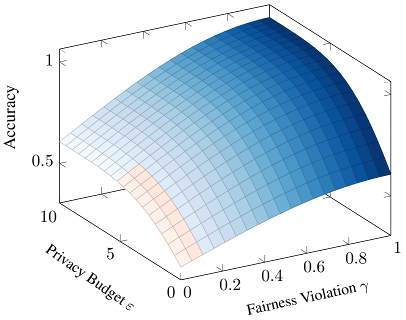


Figure 1: **Pre-selection of trustworthiness parameters only recovers a portion of the Pareto frontier.** The remaining parts of frontier (shaded blue) are never explored.

To address the challenge of calculating an accurate Pareto frontier, we propose to adopt *impartiality* as a principle: our design should not explicitly or implicitly favour one objective to another. Impartiality is easiest to satisfy when our task can be reduced to a single optimization problem: as long as all objectives are similarly weighted, we remain impartial because all objectives are considered *simultaneously*. However, **current methods that incorporate unfairness or privacy mitigations rarely fit into a simultaneous optimization setting. This is because they are implemented at different points in an ML pipeline.**

We provide an example of the drawbacks of a non-impartial design by theoretically showing that adding a fairness pre-processing step before introducing privacy will degrade privacy guarantee. Given that this design will also likely cause utility loss due to the introduction of the unfairness mitigation and that it fails to provide fairness guarantees at inference-time, the design is Pareto-inefficient—a result that we also verify empirically. Since there are currently only two major private learning algorithms, namely DP-SGD and PATE, we leave exploring the Pareto efficiency of other *more general* constructions for impartial private and fair ML pipelines to future work, and instead focus our attention on designing impartial pipelines featuring these two frameworks.

We evaluate our resulting frameworks FairDP-SGD and FairPATE against a suite of non-impartial approaches on multiple datasets and for different tasks. We find that our impartial designs often produce improved results in at least one objective compared to the baselines, and therefore, naturally surface the Pareto frontier—representing the irreconcilable trade-offs between various trustworthiness objectives. Additionally, we show that the Pareto frontiers transfer between different datasets on the same task. This means that producing recommendations for an operating point on the trustworthy Pareto frontier is task and *not* data-dependent. As a consequence, regulators can provide specifications even without access to the private or proprietary datasets they need to provide specifications for.

In summary, our contributions are as follows:

1. We adopt the principle of impartiality between fairness (demographic parity), privacy (differential privacy), and utility. We present a theoretical result showcasing that demographic parity pre-processing followed by private training will degrade the privacy guarantee.
2. We propose two methods (FairDP-SGD and FairPATE) that allow us to train impartially and to recover the Pareto frontier between the objectives. The Pareto frontiers provide richer representations of multi-objective ML trustworthiness.
3. We run an extensive empirical evaluation in several domain and datasets, including a medical Chest Xray disease diagnosis dataset (CheXpert) with known fairness issues [4, 29]. We provide interactive[2] Pareto frontiers for various vision tasks. Our empirical results show that FairPATE can improve accuracy up to 5% over non-impartial models through careful privacy budget consumption.

2 Background

We denote the ML model for classification by θ, the features as $(\mathbf{x}, z) \in \mathcal{X} \times \mathcal{Z}$ where $\mathcal{X}$ is the domain of non-sensitive attributes, $\mathcal{Z}$ is the domain of the sensitive attribute (categorical variable).

[2] Anonymously made available at https://impartiality-ml.github.io/impartiality/

The categorical class-label is denoted by $y \in [1, \ldots, K]$. We refer the interested reader to Appendix G for a more thorough overview.

Fairness: Demographic Parity. We base our work on the fairness metric of *multi-class demographic parity* which requires that ML models produce similar success rates (*i.e.*, rate of predicting a desirable outcome, such as getting a loan) for all sub-populations [9]. More formally, the *demographic disparity* $\Gamma(z,k)$ of subgroup z for class k is the difference between the probability of predicting class k for the subgroup z and the probability of the same event for any other subgroup: $\Gamma(z,k) := \mathbb{P}[\hat{Y} = k \mid Z = z] - \mathbb{P}[\hat{Y} = k \mid Z \neq z]$. In practice, we estimate multi-class demographic disparity for class k and subgroup z with: $\widehat{\Gamma}(z,k) := \frac{|\{\hat{Y}=k, Z=z\}|}{|\{Z=z\}|} - \frac{|\{\hat{Y}=k, Z\neq z\}|}{|\{Z\neq z\}|}$, where $\hat{Y} = \theta(\mathbf{x}, z)$. We define demographic *parity* when the worst-case demographic disparity between members and non-members for any subgroup, and for any class is bounded by γ:

Definition 1 (γ-DemParity). *For predictions Y with corresponding sensitive attributes Z to satisfy γ-bounded demographic parity (γ-DemParity), it must be that for all z in $\mathcal{Z}$ and for all k in $\mathcal{K}$, the demographic disparity is at most γ: $\Gamma(z,k) \leq \gamma$.*

Differential Privacy. Differential privacy [13] is a framework to protect privacy of individuals when analyzing their data. It achieves this by *adding controlled noise* to the algorithm used for analysis, making it difficult to identify individual contributions while still providing useful statistical results. More formally, (ε, δ)-differential privacy can be expressed as follows:

Definition 2 ((ε, δ)-Differential Privacy). *Let $\mathcal{M}\colon \mathcal{D}^* \to \mathcal{R}$ be a randomized algorithm that satisfies (ε, δ)-DP with $\varepsilon \in \mathbb{R}_+$ and $\delta \in [0,1]$ if for all neighboring datasets $D \sim D'$, i.e., datasets that differ in only one data point, and for all possible subsets $R \subseteq \mathcal{R}$ of the result space it must hold that $\mathbb{P}\left[\mathcal{M}(D) \in R\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(D') \in R\right] + \delta$.*

A natural way of integrating DP into the training process of ML models is by adding the controlled noise to the model gradients as done in the Differential Private Stochastic Gradient Descent **(DP-SGD)** algorithm [1]. Prior to noising the gradients, DP-SGD also implements a clipping operation that limits the maximum norm of individual gradients. This ensures that no data point leaves too significant impact on the model. Yet, training with clipping and noise can be challenging [11, 32]. As an alternative, the Private Aggregation of Teacher Ensemble **(PATE)** algorithm [26] considers the training algorithm as a non-private black-box and introduces noise to the model outputs. More concretely, PATE trains an ensemble of so-called *teacher models* on disjoint subsets of the private data without any privacy protection. As a result, these teachers cannot be publicly exposed because they would leak information about the private data. Instead, PATE utilizes them to label a public dataset within an appropriately noised knowledge transfer process. Therefore, the teachers in the ensemble each vote for a label for each public data point, and the final label is determined as a noisy $\arg\max$ over the teachers' vote. A separate *student model* is then trained on the labeled public dataset and can then be deployed publicly while the teachers are never externally exposed. We present more thorough introductions to DP-SGD and PATE in Appendix G.

Pareto Efficiency. Let Θ be the set of all feasible ML models with an element $\theta \in \Theta$, where a feasible model is one that is achievable through learning (optimization) over a given dataset. I is the set of measurable objectives with *loss value* of objective $i \in I$ denoted as $\ell_i(\theta)$. For instance, without loss of generality, $\ell_{\text{priv}} = \varepsilon$ where ε (DP privacy budget), and $\ell_{\text{fair}} = \hat{\Gamma}(z,k)$ (demographic parity loss). Lower loss values are desirable for every objective.

Definition 3 (Pareto Efficiency). *$\theta \in \Theta$ is Pareto-efficient if there exists no $\theta' \in \Theta$ such that (a) $\forall i \in I$ we have $\ell_i(\theta') \leq \ell_i(\theta)$, and that (b) for at least one objective $j \in I$ the inequality is strict $\ell_j(\theta') < \ell_j(\theta)$.*

3 In Search of Impartial Algorithms

We previously motivated that in designing a multi-objective trustworthy model, we should avoid favoring one objective over the other. We called this the *impartiality* principle. We also mentioned that perfect impartiality is achievable when we can optimize objectives *simultaneously*. However, this is challenging in common machine learning pipeline designs where mitigations can be implemented at different points in the ML pipeline.

Next, we make this observation rigorous for at least one such design, namely when an unfairness mitigation is implemented before executing a private training algorithm. Concretely, we show that employing a common unfairness mitigation technique, namely pre-processing the training data to equalize subpopulation rates, will degrade the privacy guarantee of any proceeding private learning algorithm:

Theorem 1. *Assume the training dataset* $D = \{(\mathbf{x}, z, y) \mid \mathbf{x} \in \mathcal{X}, z \in \mathcal{Z}, y \in \mathcal{Y}\}$ *is fed through the demographic parity pre-processor* $\mathcal{P}_{pre}$ *following an ordering defined over the input space* $\mathcal{X}$*. Let* $\mathcal{P}_{pre}$ *enforce a maximum violation* γ*, and* $|Z| = 2$*. Suppose now* $\mathcal{M}$ *is an* (ε, δ) *training mechanism, then* $\mathcal{M} \circ \mathcal{P}_{pre}$ *is* $(K_\gamma \varepsilon, K_\gamma e^{K_\gamma \varepsilon} \delta)$*-DP where* $K_\gamma = 2 + \left\lceil \frac{2\gamma}{1-\gamma} \right\rceil$.

We provide the proof for Theorem 1 in Appendix F. This result highlights that fairness pre-processor in a differentially private pipeline is likely Pareto-inefficient. Without concrete instantiations of the algorithms involved, we cannot make a general claim about the Pareto efficiency of other constructions of the trustworthy ML pipeline design. Therefore, in the rest of the paper, we focus on building bespoke impartial pipelines. More concretely, we show how we can integrate unfairness mitigation with impartiality in commonly used private learning pipelines.

3.1 FairDP-SGD

Optimizing for fairness on the private data during the training process increases the privacy costs [23]: If we assess fairness on the private data, for example to obtain a regularization term that penalizes unfair model predictions between sub-populations, this consumes from the privacy budget. The budget could otherwise be spent, for instance, on more training iterations on the private data to yield higher accuracy. In other words, integrating unfairness mitigations based on the private data obtains fairness at the costs of privacy and utility. Hence, it is a non-impartial design that might degrade the trade-offs between privacy, fairness, and utility.

We use this insight in our first algorithm that impartially integrates unfairness mitigation into a private ML algorithm, namely FairDP-SGD, our fair extension of DP-SGD. FairDP-SGD indeed relies on extending the private optimization process of DP-SGD with a Demographic Parity Fairness Regularizer (DPFR) that depends on the current fairness violation. However, we avoid paying the extra privacy cost for determining the fairness of the model during training, by estimating the fairness violation—in our case, the demographic disparity—over a public *unlabaled* dataset X_{public}. As consuming public data does not incur a privacy cost by principle [32], this allows to assess and implement a fairness regularizer during training without increasing the privacy costs—therefore, following the impartiality principle. The resulting demographic parity loss term, which can be added to the standard loss function used for training, is given by:

$$\text{DPFR}(\theta; X_{\text{public}}) = \max_k \max_z \widehat{\Gamma}(z,k) = \max_k \max_z \left\{ \frac{|\{\hat{Y} = k, Z = z\}|}{|\{Z = z\}|} - \frac{|\{\hat{Y} = k, Z \neq z\}|}{|\{Z \neq z\}|} \right\} \quad (1)$$

where $\widehat{Y} = \theta(X_{\text{public}})$ is the prediction of the privately trained model θ on the features X_{public} of the public dataset D_{public}.

The estimation of fairness violation with the DPFR in Equation (1) relies on the calculation of a maximum over the classes and sensitive attributes. Yet, such maximum calculations are non-differentiable and hence do not yield useful gradients during the optimization with DP-SGD. To overcome this limitation, we propose to use a tempered softmax to approximate the maximum: $\text{softmax}_T(x_i) = \frac{\exp x_i/T}{\sum_{\tilde{i}} \exp x_{\tilde{i}}/T}$, where T is the temperature. With small T (*e.g.* 0.01), this approximation is close to the actual $\max$ but keeps the overall loss differentiable. During training with FairDP-SGD, we add the DPFR to the original loss function with a weight λ. Note that λ is a hyper-parameter. As with other hyper-parameters, to ensure impartiality and an accurate Pareto frontier, a parameter sweep on λ is necessary. We emphasize that FairDP-SGD only requires *unlabeled* public data to assess and mitigate demographic disparity. Since demographic parity only considers the predicted label, having ground truth labels is redundant from a fairness angle. However, access to public data with labels could provide utility gains [32], a discussion of which we defer to Appendix C.

Note that the privacy analysis of our FairDP-SGD entirely follows that of standard DP-SGD [1, 42]. This is because the fairness assessment at training time is performed on the public dataset, and the post-processor operates on the test set.[3] We present our final FairDP-SGD in Algorithm 6.

[3] Private learning exclusively bounds privacy leakage of the training data, as test data is known to an attacker.

3.2 FairPATE

In PATE, similar to DP-SGD, we can obtain impartiality by simultaneously optimizing for privacy and fairness. The privacy of *standard PATE* is introduced at the level of teacher aggregation during the labeling of the public dataset. This labeling process has two components, the first part implements a logic to reject queries if they incur too high privacy costs. The second part implements the privacy guarantees for the queries that are actually answered by computing the labels as a noisy argmax over all teachers' votes on any given public data point.

We present FairPATE which introduces an unfairness mitigation to extend this private aggregation. Therefore, we propose a new aggregation mechanism, namely Confident&Fair-GNMax (**CF-GNMax**, see Algorithm 1) that extends PATE's standard GNMax algorithm (Algorithm 4 in Appendix G.2) with the idea of rejecting queries also due to their disparate impact on fairness.

Concretely, CF-GNMax, integrates an additional demographic parity constraint within the aggregator which allows rejecting queries on the basis of fairness. The algorithm checks potential violations of demographic disparity violations and maintains an upper bound γ on them in the course of answering PATE queries (Line 7 in Algorithm 1). The goal is to bound the actual $\Gamma(z,k)$—here empirically estimated. Concretely, we measure demographic disparity $\widehat{\Gamma}(z,k)$ using the counter $m : \mathcal{Z} \times \mathcal{K} \mapsto \mathbb{Z}_{\geq 0}$ which tracks per-class, per-subgroup decisions.

Care must be taken to produce accurate $\Gamma(z,k)$ estimations: with few samples, $\widehat{\Gamma}(z,k)$ may be a poor estimator of $\Gamma(z,k)$. Therefore, we have a cold-start stage where there are not yet enough samples to estimate $\widehat{\Gamma}(z)$ accurately. We avoid rejecting queries due to the fairness constraint at this stage. Concretely, we require at least, on average, M samples from the query's subgroup before we reject a query on the basis of fairness (Line 3).

Algorithm 1 – Confident&Fair-GNMax Aggregator

```
Input: query data point $x$, sensitive attribute $z$, predicted class label $k$, subpopulation subclass counts $m : \mathcal{Z} \times \mathcal{K} \mapsto \mathbb{Z}_{\geq 0}$
Require: minimum count $M$, threshold $T$, noise parameters $\sigma_1, \sigma_2$, fairness violation margin $\gamma$
 1: if $\max_j\{n_j(x)\} + \mathcal{N}(0,\sigma_1^2) \geq T$ then
 2:     $k \leftarrow \arg\max_j \left\{n_j(x) + \mathcal{N}(0,\sigma_2^2)\right\}$
 3:     if $\sum_{\tilde{k}} m(z,\tilde{k}) < M$ then
 4:         $m(z,k) \leftarrow m(z,k) + 1$
 5:         return $k$
 6:     else
 7:         if $\left(\frac{m(z,k)+1}{\left(\sum_{\tilde{k}} m(z,\tilde{k})\right)+1} - \frac{\sum_{\tilde{z}\neq z} m(\tilde{z},k)}{\sum_{\tilde{z}\neq z,\tilde{k}} m(\tilde{z},\tilde{k})}\right) < \gamma$ then
 8:             $m(z,k) \leftarrow m(z,k) + 1$
 9:             return $k$
10:         else
11:             return $\perp$
12: else
13:     return $\perp$
```

We note that *the fairness mitigation in FairPATE occurs almost exactly at the same point as the privacy-utility balancing mechanism.* This allows us to get as close as possible to the "simultaneous optimization" of objectives which, as we argued in Section 1, is the most impartial design. We extensively discuss our design choice in Appendix A and empirically validate it in Appendix B.

Privacy Analysis. FairPATE's query phase (CF-GNMAX, Algorithm 1) has two main differences to PATE's (C-GNMAX, Algorithm 4). First, FairPATE involves a *cold-start* stage during which fairness violations estimators are inaccurate. During this stage, no fairness-related rejection takes place until all subgroups have at least M samples. Second, in FairPATE, queries can be rejected for two reasons. Reason 1: Similar to standard PATE, queries that incur too high privacy costs are rejected. Reason 2: Additionally, queries whose answer would violate the fairness (γ-DemParity) constraint are rejected, as well. During the cold-start stage (Line 3), FairPATE follows the privacy analysis of PATE's (Appendix E). Afterwards, we adjust the privacy analysis to account for the rejection due to fairness. We can calculate FairPATE's probability of answering query q_i as:

$$\mathbb{P}[\text{answering } q_i(z,k)] = \begin{cases} 0 & \frac{m(z,k)+1}{\left(\sum_{\tilde{k}} m(z,\tilde{k})\right)+1} - \frac{\sum_{\tilde{z}\neq z} m(\tilde{z},k)}{\sum_{\tilde{z}\neq z,\tilde{k}} m(\tilde{z},\tilde{k})} > \gamma \\ \tilde{q} & \text{otherwise} \end{cases}$$ where k is the noisy argmax (Line 2), $\tilde{q}$ is calculated using Proposition 1 in Appendix E (as before), and the left side of the condition is simply calculating the new tentative demographic disparity violation $\Gamma(z,k)$ if the query is accepted. Note that in PATE (and by extension FairPATE) queries come from a public (and therefore non-private) dataset, and are labeled, noised and only then used to increment $m(z,k)$. Therefore,

since the value of the counter $m(z, k)$ is only conditioned on the value of the noisy argmax, by the post-processing property of DP [12], $m(z, k)$ and by extension, Line 7 do not add any additional privacy cost, *i.e.*, rejecting queries on the basis of fairness, does not incur additional privacy cost.

3.3 Improving the accuracy-trustworthy trade-off with IDP^3

Optimizing for fairness during the training process does not guarantee that fairness is obtained at inference time [3]. As we highlight in Section 2, demographic parity requires the same success rates in the predictions of different sub-populations. With adequate training, we can ensure that the model learns to generate similar success rates on different sub-populations. However, this guarantee is not ensured with differential privacy, due to *label shifts* (in Appendix D we present an example of how DP noising causes label shifts) Indeed, prior work using different privacy and fairness notions has found a "post-hoc" correction to be necessary to maintain the fairness guarantee despite privatization [24].

Algorithm 2 Inference-time Demographic Parity Post-Processor (IDP^3)

```
Input: data point $x$, sensitive attribute $z$, predicted label $\hat{y}$, subpopulation-class counts $m : \mathcal{Z} \times \mathcal{Y} \mapsto \mathbb{Z}_{\geq 0}$
Require: minimum count $M$, fairness violation margin $\gamma$
1: if $\sum_{\tilde{y}} m(z, \tilde{y}) < M$ then
2:     $m(z, y) \leftarrow m(z, \hat{y}) + 1$
3:     return $\hat{y}$
4: else
5:     if $\left( \frac{m(z,\hat{y})+1}{\left(\sum_{\tilde{y}} m(z,\tilde{y})\right)+1} - \frac{\sum_{\tilde{z} \neq z} m(\tilde{z},\hat{y})}{\sum_{\tilde{z} \neq z, \tilde{y}} m(\tilde{z},\tilde{y})} \right) < \gamma$ then
6:         $m(z, y) \leftarrow m(z, \hat{y}) + 1$
7:         return $\hat{y}$
8:     else
9:         return $\perp$
```

To guarantee that the model maintains its required degree of fairness at inference-time, we can enforce using our Inference-time Demographic Parity Post-Processor (IDP^3) highlighted in Algorithm 2. Our design is inspired by our fairness mitigation in FairPATE, but is model-agnostic and generally applicable. At its core, IDP^3 transforms the classification task to a one with the reject option (often referred to as *selective classification* [15]). This adds another dimension to the accuracy-fairness-privacy Pareto frontier: coverage, which is the proportion of queries that are answered at inference-time. We consider coverage an independent utility metric[4].

At inference time, our IDP^3 keeps track of the counts of positive predictions per sub-populations. For every query posed to the model, it first calculates the demographic disparity based on the current counters. Then, it returns a label only if the resulting success rate of the current sub-population (in comparison to the other sub-populations) stays within the tolerated fairness violation. Since, at the beginning, the model has not returned enough predictions to reliably estimate the per-sub-population success rates, we propose a cold-start phase (line 1-3) during which all queries are answered.

4 Empirical Evaluation

We evaluate FairPATE and FairDP-SGD on multiple datasets and derive the *Pareto frontiers* between privacy, utility, and fairness. Our Pareto frontiers represent the set of all Pareto efficient solutions obtained through our methods and characterize the best trade-offs that can be achieved between the three objectives. Based on the Pareto frontiers, we answer the following research questions (**RQ**s): **RQ1**: Can we achieve better trade-offs through impartial design? **RQ2**: How do FairPATE and FairDP-SGD differ in performance? **RQ3**: Can a regulatory body carry out baseline specification without direct access to the private data?

Experimental Setup. We evaluate five datasets, namely ColorMNIST [2], CelebA [22], FairFace [19], UTKFace[41], and CheXpert[16]. See Table 2 in Appendix I for details on the datasets.

We introduce two PATE-based non-impartial baselines to compare FairPATE and FairDP-SGD. Both baselines use the standard PATE's query selection process. **PATE-S$_{\text{pre}}$** incorporates an unfairness pre-processor for the Student model to ensure the maximum fairness gap in student training data is within the constraint. This pre-filters the training data points on which the student will be trained on (see Algorithm 3 for a full description of the pre-processor). **PATE-S$_{\text{in}}$**, on the other hand,

[4] If *any* trade-off with coverage is not an acceptable outcome for a particular application, we introduce an alternative mechanism in Appendix D that takes advantage of public data and calibrates the model post-training but pre-inference. This design maintains the privacy budget while reducing (but not necessarily eliminating) the fairness gap at inference-time albeit with a trade-off with model accuracy, as expected.

Setting	Method	IDP^3	ε-(↓) Budget	Max (↓) Disparity	Acc. (↑)	Cov. (↑)	ε-(↓) Budget	Max (↓) Disparity	Acc. (↑)	Cov. (↑)
			ColorMNIST				**UTKFace**			
Fair	**FairPATE**	✓	2.88	0.01	**85.6**	0.62	8.65	0.01	**83.8**	0.78
	FairDP-SGD	✓	**1.0**	0.01	85.4	0.64	**8.0**	0.01	81.2	0.72
	PATE-S$_{\text{pre}}$	✓	2.88	0.01	80.9	**0.69**	10.0	0.01	82.6	**0.82**
	PATE-S$_{\text{in}}$	✓	2.88	0.01	84.6	0.63	10.0	0.01	81.4	0.74
Private	**FairPATE**	✓	1.0	0.10	73.8	1.00	2.0	**0.13**	74.0	0.98
	FairDP-SGD	✓	1.0	0.10	**88.8**	1.00	2.0	0.15	**75.3**	0.99
	PATE-S$_{\text{pre}}$	✓	1.0	0.10	73.1	1.00	2.0	0.14	72.3	0.99
	PATE-S$_{\text{in}}$	✓	1.0	0.10	74.2	0.98	2.0	0.15	72.3	0.99
	PATE	-	1.0	0.10	73.8	1.00	2.0	0.14	72.5	0.98
	DP-SGD	-	1.0	0.10	**88.8**	1.00	2.0	0.16	**75.3**	1.0
Accurate	**FairPATE**	✓	2.87	0.10	88.5	0.99	10.0	0.2	**82.9**	**0.97**
	FairDP-SGD	✓	**2.0**	0.10	**88.6**	0.99	10.0	0.1	81.3	0.96
	PATE-S$_{\text{pre}}$	✓	2.88	0.10	88.1	1.0	10.0	0.01	82.6	0.82
	PATE-S$_{\text{in}}$	✓	3.0	0.10	88.5	0.99	10.0	0.15	81.6	0.92
	PATE	-	2.88	0.11	88.1	1.0	10.0	0.25	81.4	1.0
	DP-SGD	-	**2.0**	0.10	**88.6**	1.0	10.0	0.14	80.6	1.0

Table 1: **Baseline Comparisons**[6]**.** For a fair comparisons, for both baselines **PATE-S$_{\text{pre}}$**, and **PATE-S$_{\text{in}}$**, we apply our post-processor IDP^3. PATE-S$_{\text{in}}$ additionally includes the regularization for the student training, while PATE-S$_{\text{pre}}$ employs a student pre-processor. We also report results obtained with the standard versions of PATE and DP-SGD without any fairness mitigation. The values reported are obtained by first generating the whole Pareto frontier, then choosing points from the surface that satisfy following criteria: selecting one objective that we want to optimize for (**Setting**) and putting a hard constraint on its achieved value, *e.g.*, $\varepsilon = 1$ for "Private", and then reporting the achieved guarantees for all other objectives. Note that standard PATE and DP-SGD cannot specifically optimize for fairness, hence they are not reported for that subsection. FairPATE and FairDP-SGD achieve the highest accuracy in most settings.

incorporates our fairness regularizer (Equation (1)) as an unfairness in-processor for the Student during training. This implies training the student on all the queries labeled by the teachers, but setting additional constraints during training. Both baselines serve to understand the impact of implementing fairness *after* the noisy aggregation of teachers, *i.e.*, after privacy in the ML pipeline[5].

In FairPATE, we apply the fairness constraint γ, whereas in FairDP-SGD, we have an unconstrained optimization problem and control the regularization factor λ. In the experiments, we use λ between 0 and 10. In our results, we report the achieved privacy budget ε, the fairness gap γ, as well as model accuracy, and coverage, *i.e.*, the fraction of data points at inference that obtains a label by the model.

RQ1: FairPATE Pareto-dominates similar designs in most contexts. To assess the improvement of trade-offs from our impartial design, we compare FairPATE and FairDP-SGD against the two non-impartial PATE-based baselines PATE-S$_{\text{pre}}$ and PATE-S$_{\text{in}}$. We also compare against **standard PATE** and **standard DP-SGD** (*i.e.*, without any unfairness mitigation) to understand the inherent fairness and utility obtained through the private algorithms we build on. Finally, since our algorithms (FairPATE and FairDP-SGD) take advantage of the post-processor introduced in Section 3.3; for a fair comparison, we enforce IDP^3 for all four baselines. In Appendix B, we present an ablation study on the importance of the post-processing which highlights that post-processing helps satisfy small fairness constraints while preserving model accuracy at the cost of answering fewer queries. Our final models' results for both our methods and the baselines are reported in Table 1 and Figure 2.

The values in Table 1 highlight that FairPATE and FairDP-SGD obtain the highest accuracy in most settings. In comparing FairPATE and PATE-S$_{\text{pre}}$ in Figure 2, note that their only difference is that in FairPATE fairness mitigation occurs at the same point as the privacy mechanism whereas in PATE-S$_{\text{pre}}$, it occurs after the privacy mechanism of the PATE. We see that at low privacy and fairness regimes, FairPATE always outperforms PATE-S$_{\text{pre}}$ in terms of utility and has higher coverage in most cases. This highlights the benefit of our impartial design: by rejecting queries because of the fairness constraints at the point where privacy is also implemented, we can save privacy budget. With the saved budget, other queries can be answered that then help improve the student model's

[5] We provide a detailed comparison between more PATE-based baselines in Appendix A

[6] We note that choosing results to tabulate can lead to pre-selection bias, and therefore non-impartiality. Thereforre, our prefered representation of our results is the (interactive) Pareto frontier plots. Nevertheless, we present tabulated data for quick comparisons with a clear selection criteria for the results.

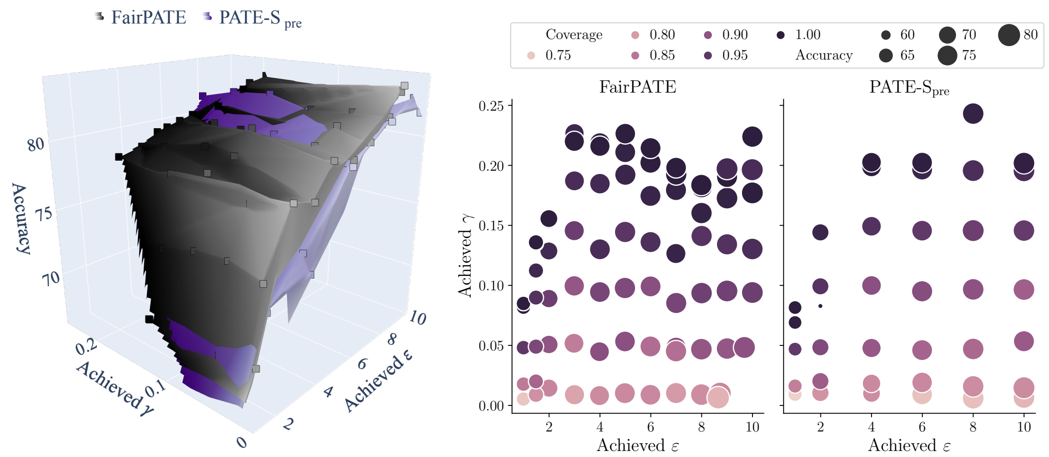
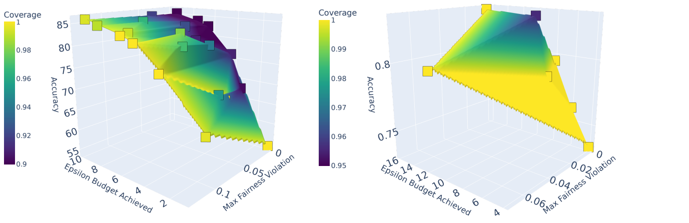


Figure 2: **Achieved trade-off between privacy, fairness, utility, and coverage for FairPATE and PATE-S$_{\text{pre}}$.** *Left*: estimated Pareto frontier with sampled points indicated. The lighter parts of the surfaces indicate lower coverage. *Right*: sampled points from the Pareto frontier in 2d with size indicating accuracy and hue, coverage. FairPATE outperforms PATE-S$_{\text{pre}}$ in terms of accuracy for high privacy budget-low fairness violation regions while they have similar accuracy in other regions.



(a) FairPATE-CelebA

(b) FairPATE-CheXpert

Figure 3: **FairPATE on CelebA and CheXpert.** The figure plots the model results that are on the Pareto frontier. We observe that in both figures, accuracy increases with higher privacy budget ε, and that looser fairness constraints yield higher coverage.

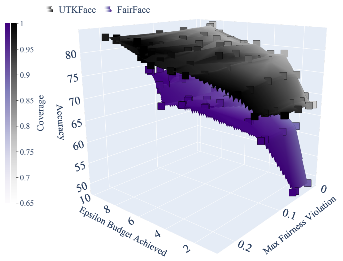


Figure 4: **Pareto Frontier of FairPATE results on UTKFace and FairFace.** The two surface have very similar shapes despite the differences in accuracy.

utility. In higher privacy budget and higher fairness violation regions, both methods achieve similar performances. This is because the fairness constraint is too loose to activate FairPATE's fairness mechanism. From Table 1, we also understand that FairPATE and PATE-S$_{\text{in}}$ perform similarly in larger fairness violation regions but FairPATE performs better in smaller fairness violation regions with better accuracy and coverage. This is again because with larger fairness violation (higher γ), FairPATE's rejection mechanism is not activated. However, in the smaller-violation regions, the impartial design shows its advantage. Overall, FairPATE outperforms PATE-S$_{\text{pre}}$ and PATE-S$_{\text{in}}$ in most regions.

RQ2: FairPATE performs better than FairDP-SGD, especially with higher privacy budgets. We compare our two methods, FairPATE and FairDP-SGD in Table 1. We observe that while they yield similar accuracy in low privacy budget regions, FairPATE provides better accuracy in higher privacy budget regions. Additionally, in low fairness violation regions, FairPATE achieves higher accuracy and higher coverage. In general, we find that tuning the fairness regularizer λ for FairDP-SGD is more difficult than tuning the FairPATE counterpart γ in CF-GNMAX. This leads to a smoother Pareto frontier for FairPATE than FairDP-SGD. Theoretically, the reason for this is two fold: the upperbound γ is enforced as a constraint, and not as a highly non-convex loss that is the demographic parity fairness regularizer (DPFR) in Equation (1). Furthermore, FairPATE's constraint is applied in the query (sample) space–effectively as a sampler–whereas DPFL is applied in (student model's) weight space. Empirically, we attribute this to the fact that performance (accuracy) of PATE

(FairPATE) is meaningfully correlated with the number of answered queries (See Appendix J). This allows for more fine-grained control on privacy costs, and thus a smoother Pareto frontier.

RQ3: Specification without direct data access is possible. Returning to the specification problem where regulators need to specify trustworthiness guarantee as we introduced in the introduction, we explore whether the regulators can produce good recommendations even if they do not have access to the actual private data. We compare the Pareto frontier surfaces obtained on different datasets for multiple tasks. Figure 3a and Figure 3b plots the Pareto frontier surface from FairPATE on CelebA and CheXpert respectively. Figure 4 plots the Pareto frontier from FairPATE on UTKFace and FairFace. Although the Pareto frontier surfaces show similar trends, the shapes in Figure 3 are dataset dependent: different datasets show different trade-offs between the objectives. However, we notice that the Pareto frontier surface shapes on UTKFace and FairFace in Figure 4 are very similar. The classification task on both datasets is gender, with race as the sensitive attribute. This shows that a regulator could use the Pareto frontier from a different dataset (which they have access to) to design baseline specifications—as long as the datasets share the same data domain and task.

5 Related Work

Due to the multiplicity of algorithmic fairness notions, as well as privacy; defining a benchmark to study fairness-privacy-utility trade-offs is difficult. In this paper, we focus on discovering the Pareto frontier between demographic parity fairness (a *group fairness* notion [5]) and (central) differential privacy [13]. While these objectives have a significant impact on each other, each has been defined and developed independently of one another. In contrast, there is a lineage of work that provides new definitions of fairness by characterizing the disparate impact of employing a privacy-aware mechanism [33, 35]. While useful in their own regard, these new definitions do not alleviate the burden of satisfying established notions of fairness, such as demographic parity. Additionally, other works consider these trade-off for particular classes of ML pipeline desings (such as federated learning [27, 21]), that are important in their use-case, but are not generally applicable.

Conceptually, the closest works to our setup are Jagielski et al. [17] and Mozannar et al. [24] which assume different privacy notions. Both works strive to provide differential privacy (DP) with respect to the sensitive attribute. Jagielski et al. [17] assumes a central notion of DP, while Mozannar et al. [24] assume a *local* DP notion [6]. Importantly, neither of the definitions used provide classical (approximate) differential privacy [13] guarantees with respect to *all features*. Furthermore, algorithms provided in these works, consider linear models and are optimized over tabular data. FairPATE and FairDP-SGD, on the other hand, are scalable deep-learning algorithms.

6 Limitations & Conclusions

Ensuring trustworthy machine learning is inherently a multi-objective endeavour. We acknowledge that as algorithm designers, we are only a part of the decision making process which likely occurs before any human judgement is passed. As such, it is imperative that (i) our design decisions should not limit (human) decision maker choices; and (ii) not favour one objective over another. In this paper, we addressed the first challenge by providing a rich trade-off representation between the different objectives (fairness, privacy, and accuracy) in the form of a Pareto frontier. Our answer to the second challenge emerged as a design principle, which we called the *impartiality* principle. We showed that models that break the impartiality principle are likely not on the Pareto frontier.

Moving forward, the intuition behind our framework is pervasive to different formulations of what it means to be trustworthy. However, our current work assumes demographic parity as the fairness notion. We acknowledge that other fairness notions (group-, individual (metric)- and causality-based), as well as other privacy notions are prevalent in the literature. It is important to note that in this paper we argued for impartiality in the *decision-theoretic* sense, that is assuming we already have good metrics on which we can build ML pipelines that do not favour one objective or another. The search for *impartial metrics* is a separate research endeavour. As long as we can formulate a measure of fairness, FairPATE and FairDP-SGD can be adopted to implement them subject to availability of labeled or unlabeled public data. We leave their study to future work.

References

[1] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[2] Martin Arjovsky et al. “Invariant risk minimization”. In: *arXiv preprint arXiv:1907.02893* (2019).

[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. “Differential Privacy Has Disparate Impact on Model Accuracy”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[4] Imon Banerjee et al. “Reading Race: AI Recognises Patient’s Racial Identity In Medical Images”. In: *The Lancet Digital Health* 4.6 (June 2022), e406–e414. ISSN: 25897500. DOI: 10.1016/S2589-7500(22)00063-2. arXiv: 2107.10356[cs,eess]. URL: http://arxiv.org/abs/2107.10356 (visited on 02/07/2023).

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[6] Björn Bebensee. “Local Differential Privacy: a tutorial”. In: (July 27, 2019). DOI: 10.48550/arXiv.1907.11908. URL: https://arxiv.org/abs/1907.11908v1 (visited on 02/03/2023).

[7] Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. “Bounds on the sample complexity for private learning and private data release”. In: *Theory of Cryptography Conference*. Springer. 2010, pp. 437–454.

[8] US Census Bureau. *Formal Privacy Methods for the 2020 Census*. Census.gov. Section: Government. 2020. URL: https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/planning-docs/privacy-methods-2020-census.html (visited on 01/31/2023).

[9] Toon Calders and Sicco Verwer. “Three naive Bayes approaches for discrimination-free classification”. In: *Data mining and knowledge discovery* 21.2 (2010), pp. 277–292.

[10] Hongyan Chang and Reza Shokri. “On the Privacy Risks of Algorithmic Fairness”. In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE Computer Society, Sept. 1, 2021, pp. 292–303. ISBN: 978-1-66541-491-3. DOI: 10.1109/EuroSP51992.2021.00028. URL: https://www.computer.org/csdl/proceedings-article/euros&p/2021/149100a292/1yg1gS8yxq0 (visited on 02/01/2023).

[11] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. “Understanding gradient clipping in private sgd: A geometric perspective”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13773–13782. eprint: 2006.15429.

[12] Cynthia Dwork. “Differential privacy”. In: *Automata, languages and programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 978-3-540-35908-1.

[13] Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3 (2013), pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042. URL: http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042 (visited on 03/12/2021).

[14] Tom Farrand et al. “Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy”. In: *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*. 2020, pp. 15–19.

[15] Yonatan Geifman and Ran El-Yaniv. “Selective Classification for Deep Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html (visited on 05/15/2023).

[16] Jeremy Irvin et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[17] Matthew Jagielski et al. “Differentially private fair learning”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3000–3008.

[18] Peter Kairouz et al. “Practical and private (deep) learning without sampling or shuffling”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5213–5225.

[19] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1548–1558.

[20] Jon Kleinberg et al. "Human Decisions and Machine Predictions*". In: *The Quarterly Journal of Economics* 133.1 (Feb. 1, 2018), pp. 237–293. ISSN: 0033-5533. DOI: `10.1093/qje/qjx032`. URL: `https://doi.org/10.1093/qje/qjx032` (visited on 04/28/2023).

[21] Tian Li et al. "Ditto: Fair and Robust Federated Learning Through Personalization". In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, July 1, 2021, pp. 6357–6368. URL: `https://proceedings.mlr.press/v139/li21h.html` (visited on 05/17/2023).

[22] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.

[23] Ilya Mironov. "Renyi Differential Privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (Aug. 2017), pp. 263–275. DOI: `10.1109/CSF.2017.11`. arXiv: `1702.07476`. URL: `http://arxiv.org/abs/1702.07476` (visited on 12/19/2020).

[24] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. "Fair learning with private demographic data". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7066–7075.

[25] Nicolas Papernot et al. "Scalable private learning with pate". In: *arXiv preprint arXiv:1802.08908* (2018).

[26] Nicolas Papernot et al. "Semi-supervised knowledge transfer for deep learning from private training data". In: *arXiv preprint arXiv:1610.05755* (2016).

[27] Sikha Pentyala et al. "PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning". In: *arXiv preprint arXiv:2205.11584* (2022).

[28] Geoff Pleiss et al. "On Fairness and Calibration". In: *arXiv:1709.02012 [cs, stat]* (Sept. 6, 2017). arXiv: `1709.02012`. URL: `http://arxiv.org/abs/1709.02012` (visited on 03/19/2019).

[29] Laleh Seyyed-Kalantari et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers". In: *Biocomputing 2021*. WORLD SCIENTIFIC, Oct. 30, 2020, pp. 232–243. ISBN: 9789811232695. DOI: `10.1142/9789811232701_0022`. URL: `https://www.worldscientific.com/doi/10.1142/9789811232701_0022` (visited on 02/04/2023).

[30] Laleh Seyyed-Kalantari et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers". In: *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. World Scientific. 2020, pp. 232–243.

[31] Vinith M. Suriyakumar et al. "Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, 723–734. ISBN: 9781450383097. DOI: `10.1145/3442188.3445934`. URL: `https://doi.org/10.1145/3442188.3445934`.

[32] Florian Tramer and Dan Boneh. "Differentially Private Learning Needs Better Features (or Much More Data)". In: International Conference on Learning Representations. Jan. 12, 2021. URL: `https://openreview.net/forum?id=YTWGvpFOQD-` (visited on 05/17/2023).

[33] Cuong Tran, My H Dinh, and Ferdinando Fioretto. "Differentially Private Deep Learning under the Fairness Lens". In: *arXiv preprint arXiv:2106.02674* (2021).

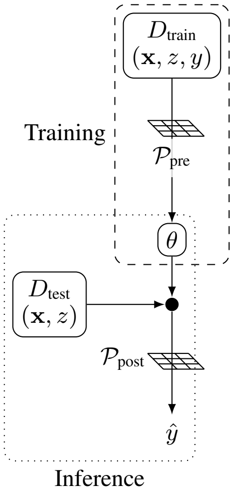
[34] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. "Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021.

[35] Cuong Tran et al. "A Fairness Analysis on Private Aggregation of Teacher Ensembles". In: *arXiv preprint arXiv:2109.08630* (2021).

[36] Archit Uniyal et al. "DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy?" In: *arXiv preprint arXiv:2106.12576* (2021).

[37] Salil Vadhan. "The complexity of differential privacy". In: *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich* (2017), pp. 347–450.

[38] Jiaqi Wang et al. "In Differential Privacy, There is Truth: On Vote Leakage in Ensemble Private Learning". In: *arXiv preprint arXiv:2209.10732* (2022).



Algorithm 3 **Pre-Processor $\mathcal{P}_{\text{pre}}$**

```
Input: data point $x$, sensitive attribute $z$, true label $y$,
    subpopulation-class counts $m : \mathcal{Z} \times \mathcal{Y} \mapsto \mathbb{Z}_{\geq 0}$
Require: minimum count $M$, fairness violation margin $\gamma$
1: if $\sum_{\tilde{y}} m(z, \tilde{y}) < M$ then
2:     $m(z, y) \leftarrow m(z, y) + 1$
3:     return $x$
4: else
5:     if $\left( \frac{m(z,\hat{y})+1}{\left(\sum_{\tilde{y}} m(z,\tilde{y})\right)+1} - \frac{\sum_{\tilde{z} \neq z} m(\tilde{z},\hat{y})}{\sum_{\tilde{z} \neq z, \tilde{y}} m(\tilde{z},\tilde{y})} \right) < \gamma$ then
6:         $m(z, y) \leftarrow m(z, y) + 1$
7:         return $x$
8:     else
9:         return $\perp$
```

Figure 5: **Demographic Parity Mitigation.** We depict the placement of the fairness pre and post-processing (left). The pre-processor $\mathcal{P}_{\text{pre}}$ (Algorithm 3, right) operates before the training of the model θ is started, *i.e.* it takes place in sample space $\mathcal{X}$. A post-processor such as IDP^3 (see Algorithm 2) is applied at inference time and operates in label space $\mathcal{Y}$. In Appendix D we introduce a fairness calibrator that operates in model weight space $\theta \in \Theta$. In Algorithm 3, subpopulation-class count m refers to the number of data points per-class within each of the subpopulation groups. It is used to empirically estimate the demographic disparity $\widehat{\Gamma}(z, k)$, $m : \mathcal{Z} \times \mathcal{K} \mapsto \mathbb{Z}_{\geq 0}$. After a *cold-start phase* (line 1-3), we start rejecting queries for x if we have to few samples from a given class.

[39] Depeng Xu, Wei Du, and Xintao Wu. "Removing disparate impact on model accuracy in differentially private stochastic gradient descent". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 1924–1932.

[40] Tao Zhang et al. "Balancing Learning Model Privacy, Fairness, and Accuracy With Early Stopping Criteria". In: *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[41] Zhifei Zhang, Yang Song, and Hairong Qi. "Age Progression/Regression by Conditional Adversarial Autoencoder". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017.

[42] Yuqing Zhu and Yu-Xiang Wang. "Poission Subsampled Rényi Differential Privacy". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, May 24, 2019, pp. 7634–7642. URL: https://proceedings.mlr.press/v97/zhu19c.html (visited on 01/30/2023).

7 Appendices

A Integrating Unfairness Mitigation into PATE

In the following, we analyze other joints in private ML pipelines where fairness could be integrated. DP-SGD has only a few degrees of freedom where fairness measures can be implemented, so we turn our study to the more complex PATE framework. The modular design of PATE (see Figure 6) allows us multiple points of fairness integration. Note that these are not necessarily impartial designs.

Teacher-level ❶, ❷, and ❸. All three designs are non-impartial as they place fairness before privacy mitigation. Since in PATE, privacy is ensured at the level of aggregated teachers and not individual teachers, all the three alternatives can be seen as instances of the fairness pre-processor $\mathcal{P}_{\text{pre}}$ in Theorem 1 from the final student model's perspective. As a result, they all suffer from additional privacy leakage; on the level of teacher data ❶, model ❷, or vote ❸.

Student-level ❺, ❻, and ❼. These designs place privacy before fairness, therefore, they are also non-impartial. Thanks to differential privacy post-processing, the privacy budget remains unchanged. However, the drawbacks are in terms of fairness and accuracy and caused by the label shifts. We discussed the former in Section 3.3. Regarding the impact on accuracy, remember that in the query

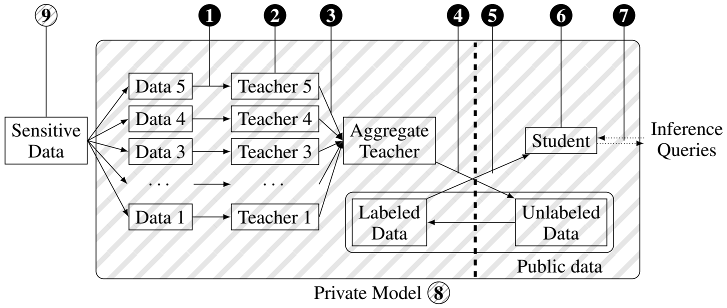


Figure 6: **Various ways to integrate fairness in PATE:** For teachers: Pre- ❶/In- ❷/Post-Processing ❸. For the student: Pre- ❺/In- ❻/Post-Processing ❼. A fair supervised privacy-preserving algorithm (*e.g.*, Our FairDP-SGD) replaces the private model (grey stripes) in-processing ⑧, while a pre-processor applies to sensitive data directly ⑨. Dashed line separates public and private data domains. Our FairPATE's intervention occurs at ❹.

phase of PATE, we incur a much smaller privacy cost for rejecting a query than for answering it (see Section 2). Now consider the scenario in ❺ where a query is labeled but is ultimately rejected due to a fairness violation. In this case, the extra budget incurred for answering the query is wasted. Saving this budget could have allowed us to answer more queries, thus enabling higher student accuracy. Therefore, ❺ is not Pareto-efficient. We note that our demographic parity post-processor in Algorithm 2 is suitable for ❼ but, on its own, still inefficient. We demonstrate the inefficiencies of ❻ and ❼ empirically in Appendix B.

B Ablation: Is post-processing necessary for ensuring tight fairness gaps?

We evaluate FairPATE and FairDP-SGD models without the post-processor and show results in Figure 7 and Figure 8, respectively. In FairPATE, without the post-processor, results span a smaller range of fairness violations. This is expected as FairPATE introduces a label shift in its training data that should be mirrored in the test data by the post-processor. The post-processor, thus, ensures that tighter fairness gaps are feasible. With FairDP-SGD, using the DPFR, we can achieve smaller fairness violations but the model utility decreases accordingly. In order to reach very small fairness gaps, we lose all utility as the model becomes increasingly inaccurate. The post-processor can preserve utility while satisfying tight fairness constraints at the cost of answering slightly fewer queries.

C Pre-Training with Public Data

Access to *labeled* public data can provide utility gains without increasing the privacy budget. This has been shown, for instance, for feature engineering using public data followed by training on private data [32]. Public pre-training is a common technique for training large private language models.

The assumption in all such models, is access to good-quality labeled public training sets. When this assumption holds, as it does often benchmark visiona and language tasks, non-private pre-training with public data followed by private training would likely be the Pareto-efficient solution due to the much improved utility gains from non-private pre-training. These utility gains, in turn, can compensate for a fairness mitigation post-processor either in the model weight space [28] or in the output space (such as IDP^3 in Section 3.3). As a result, these models could achieve a better overal trade-off, and end up being the Pareto-efficient choice.

If the assumption of access to large amounts of labeled data for pre-training does not hold; what is the best course of action? What if we only have access to a relatively small amount of public labeled data? In this case, what we can do is to move the use of public data from the pre-training stage to a *fine-tuning* stage. We call this fairness-focused fine-tuning stage, *fairness calibration* and discuss it in detail in Appendix D.1.

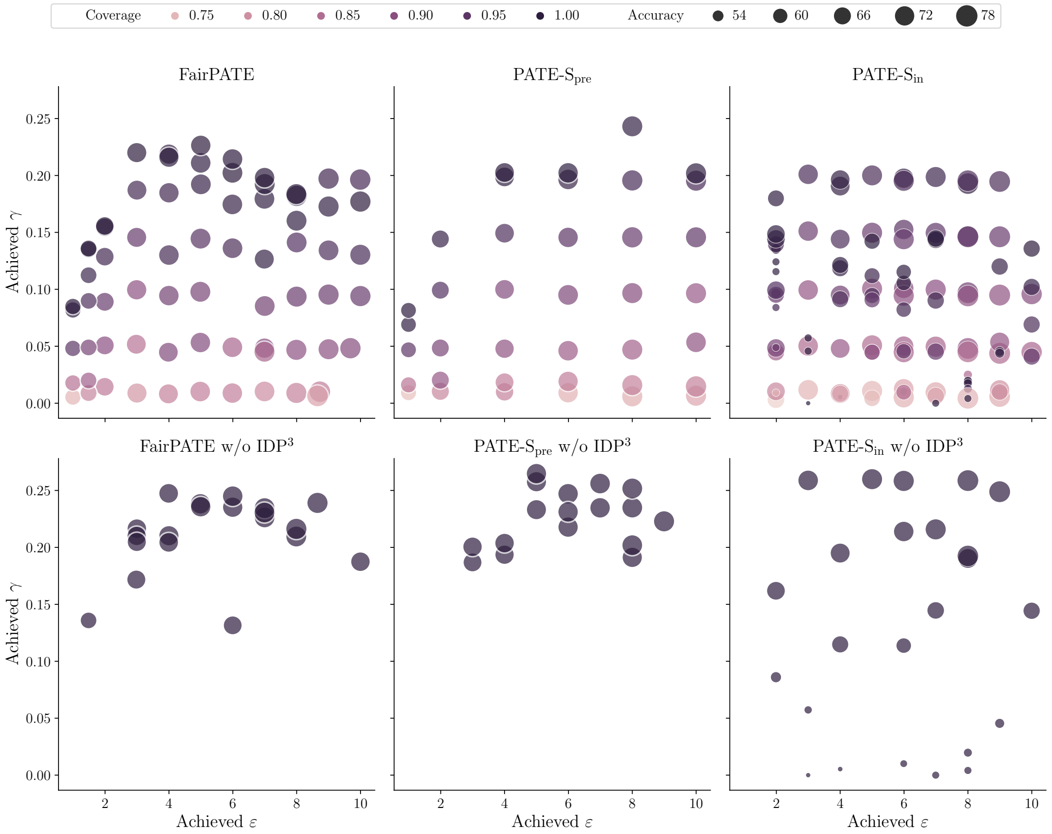


Figure 7: **Effect of post-processing on PATE family of models.**

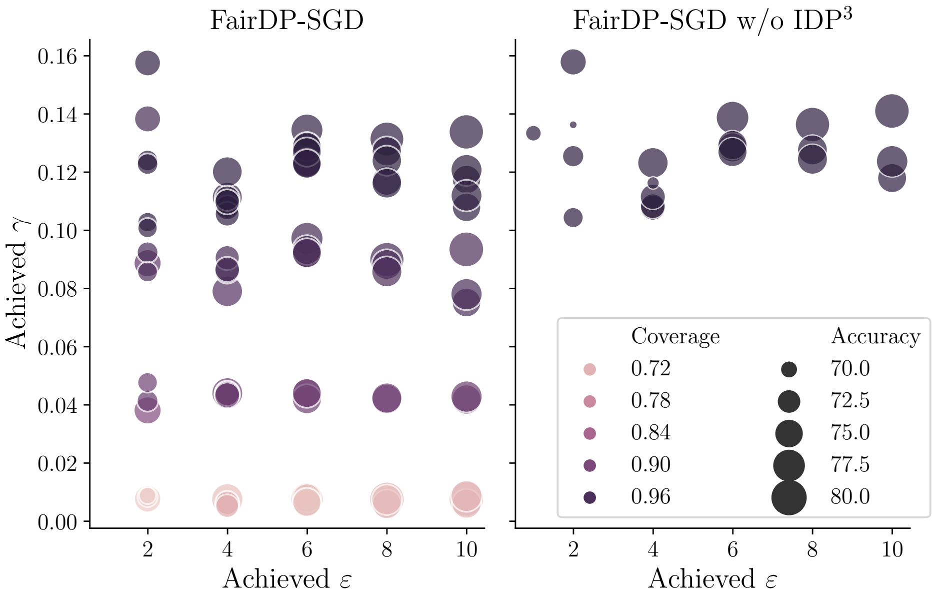


Figure 8: **FairDP-SGD with vs. without post-processor on UTKFace.** Post-processor helps satisfy small fairness constraints while preserving model accuracy at the cost of answering fewer queries.

D Calibrating for Fairness Using Public Data

Label Shift. In Section 3.3, we discussed how by adding noise to make outputs indistinguishable, differential privacy can cause label shifts that can break previously achieved fairness violation guarantees. Figure 9 shows a concrete example of this, where we show the ordained level of fairness gap against the measured gap at the output of a pre-processed PATE model (PATE-S_{pre} in blue). Clearly the model is not meeting it ordained fairness guarantee. A smaller epsilon leads to higher nosing levels; and as a result it can lead to a bigger fairness gap. However, this is not a consistent trend since by nature, DP noising is a probablistic mechanism. In any case, a certain ordained level of fairness violations is not guaranteed. In Figure 9 we also show the same model now with the inference-time demographic parity post-processor (IDP^3 in orange), which manages to keep the fairness gaps at least as low as the ordained level. As discussed in Section 3.3, this is by turning the classifier into a selective classifier and introducing a trade-off with coverage. In the next section, we discuss how this trade-off can be avoided using a calibration scheme.

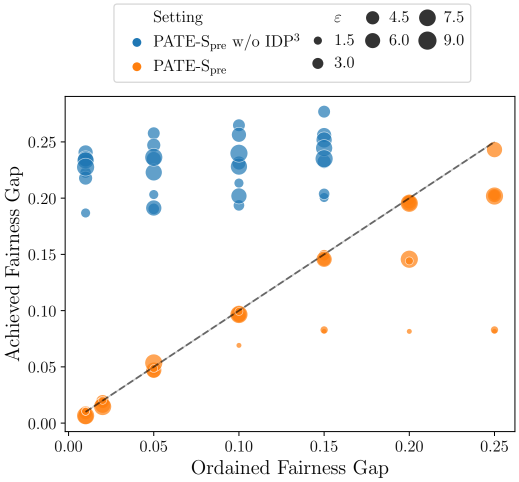


Figure 9: **Despite pre-processing, differential privacy causes label shifts that may break the pre-specified fairness constraint.**

D.1 Public Data Calibration

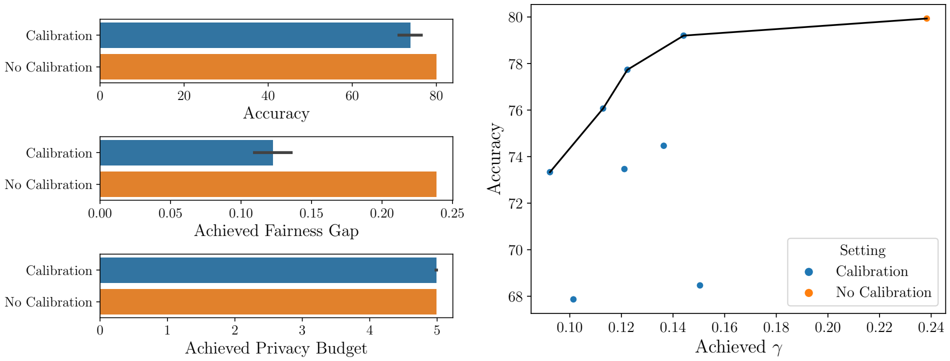
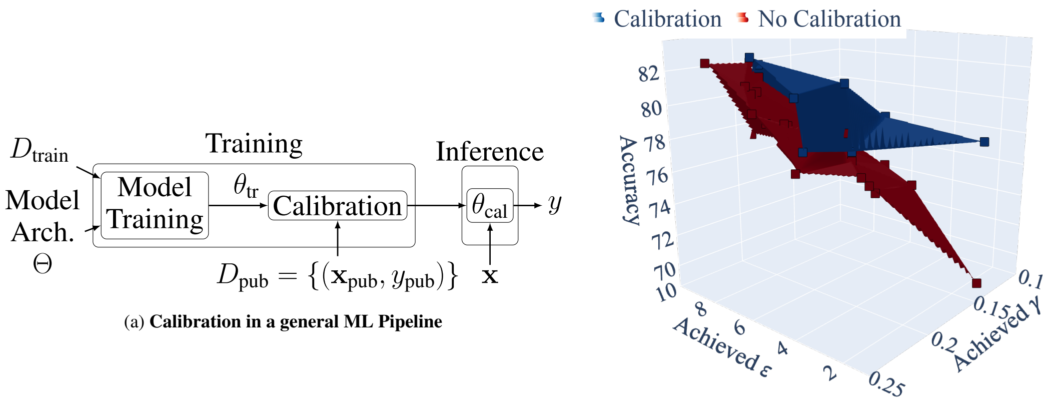


Figure 10: **Public Data Calibration trade-off fairness for accuracy without impacting the privacy budget**

Our goal is to reduce the fairness gap post differential privacy noising, but avoid a trade-off with coverage. We start by making two key observations: **(a)** in private training, we care about the privacy of training data but not inference data (since inference data features are already known to users of the models), and **(b)** in providing fairness guarantees for models, we care about the fairness gaps at inference-time—indeed, the training fairness-gaps are taken only as a proxy for test-time gaps, with an expectation that fairness properties of the trained model generalizes at test time.

In Figure 11a, we present a fairness calibration mechanism that takes advantage of public data to adjust the model post-training but pre-inference. The effect is that no privacy budget is spent as data used to calibrate the model is public, and on the other hand, the model exhibits better fairness gaps as a result.

Our empirical result in Figure 10 on UTKFace using FairPATE without IDP^3, shows that calibration helps reduce the fairness gap by about 9% (from 24% to 15%) with less than 2% drop in accuracy, for the model on the Pareto frontier shown in Figure 10 (right).



(a) **Calibration in a general ML Pipeline**

(b) **FairPATE (w/o IDP^3) With and Without Calibration on UTKFace**

Figure 11: **Fairness Calibration using Public Data.** Note that model $\omega \in \mathcal{W}$ can be any model, for instance a FairPATE student, or FairDP-SGD or, any other trained model

Our more extensive Pareto frontier in Figure 11b confirms our findings. In low privacy and low-fairness violation regions in particular, we see the largest gains from calibration. Calibration allows us to achieve better utility for the same privacy and fairness level of guarantees. In one case, for $\varepsilon = 1.48$ and $\gamma = 0.14$ we achieve a 9% improvement to accuracy (from 69% to 78%).

E Standard PATE Privacy Analysis

Papernot et al. [25] use Rényi differential privacy (RDP) [23] for accounting of the privacy budget expanded in answering each query. While the true privacy cost for each query is not known, an upporbound is estimated and summed over the course of the query phase. Answering queries stop when a pre-defined budget is exhausted. A student model is then trained on the answered queries.

Theorem 2 establishes that the upperbound is a function of the probability of *not* answering a query i with the plurality vote i^*. Unsurprisingly, this privacy cost function must tends to zero when the said event is very unlikely (*i.e.*, strong consensus):

Theorem 2 (From [25])**.** *Let $\mathcal{M}$ be a randomized algorithm with $(\mu_1, \varepsilon_1) - RDP$ and $(\mu_2, \varepsilon_2) - RDP$ guarantees and suppose that given a dataset D, there exists a likely outcome i^* such that $\Pr[\mathcal{M}(D) \neq i^*] \leq \tilde{q}$. Then the data-dependent Rényi differential privacy for $\mathcal{M}$ of order $\lambda \leq \mu_1, \mu_2$ at D is bounded by a function of $\tilde{q}, \mu_1, \varepsilon_1, \mu_2, \varepsilon_2$, which approaches 0 as $\tilde{q} \to 0$.*

In practice, Proposition 1 is used to find $\tilde{q}_i$ in Theorem 2, and μ_1, μ_2 are optimized to achieve the lowest upperbound on the privacy cost of each query for every order λ of RDP.

Proposition 1 (From [25])**.** *For any $i^* \in [m]$, we have $\Pr[\mathcal{M}_\sigma(D) \neq i^*] \leq \frac{1}{2} \sum_{i \neq i^*} \operatorname{erfc}\left(\frac{n_{i^*} - n_i}{2\sigma}\right)$, where erfc is the complementary error function.*

F Privacy Cost of Pre-Processing

Fairness pre-processing can lead to increased privacy costs during private training. A consequence of differential privacy is the privacy consumption regime [23]: just by observing the data for the purposes of equalizing a fairness measure between subpopulations, we may consume from the privacy budget.[7] This budget could otherwise be spent, for instance, on more training passes on data to yield higher accuracy. We formalize this observation in Theorem 1 for the case when a universal ordering exists.

[7] Note that this disadvantage does not hold for fairness post-processing which does not incur additional privacy costs due to the differential privacy post-processing property.

Proof. We will proceed to show that using a pre-processing that sorts through data following some ordering defined over the whole input space [8] and, for any given label y, removes the last datapoints (following the ordering) in the majority subclass until it satisfies the γ-constraint will produce datasets at most $2 + K_\gamma = 2 + \left\lceil \frac{2\gamma}{1-\gamma} \right\rceil$ apart. One then applies group privacy to obtain the final claim of the theorem.

Let $D' = D \cup x^*$, and the label of x^* is y^*. We now proceed to analyze how far apart $\mathcal{P}_{\text{pre}}(D)$ and $\mathcal{P}_{\text{pre}}(D')$ can be. First note, they are the same on all labels not y^*, so we need only consider the difference on this label. First, let m be the size of the minority subclass for label y^* and let $m + c$ be the admissible size of the majority class. That is, we have $\frac{m}{2m+c} - \frac{m+c}{2m+c} < \gamma$. From this we can conclude $c = \lfloor \frac{\gamma}{1-\gamma} 2m \rfloor$. Given this relation between the size of majority class a function of the minority class, we proceed to go through all logical cases to show the maximum difference is as claimed above.

Suppose x^* belongs to the minority subclass for y^* in D. Then we have $m \rightarrow m + 1$ and hence $c \rightarrow \lfloor \frac{\gamma}{1-\gamma} 2(m + 1) \rfloor$. Thus we see $\mathcal{P}_{\text{pre}}(D')$ now admits one more point in the minority class of y^* and at most $1 + \lceil \frac{2\gamma}{1-\gamma} \rceil$ more points to the the majority subclass (note we do not replace existing points as we follow the ordering on the input space). Thus the max change between $\mathcal{P}_{\text{pre}}(D)$ and $\mathcal{P}_{\text{pre}}(D')$ is $2 + \lceil \frac{2\gamma}{1-\gamma} \rceil$

Now suppose x^* belongs to majority subclass for y^* in D. In this case we have either x^* appears early enough in the ordering that it now replaces another point in the majority class when applying P, or it is not added. In the former case, this mean we have changed $\mathcal{P}_{\text{pre}}(D)$ by 2: we first removed a point and then added x^*. In the latter case, x^* did not get added into the dataset, more so because of the ordering, $\mathcal{P}_{\text{pre}}(D') = \mathcal{P}_{\text{pre}}(D)$ as the order of points before x^* is still the same. So in this case, once again, the change between $\mathcal{P}_{\text{pre}}(D)$ and $\mathcal{P}_{\text{pre}}(D')$ is less than $2 + \lceil \frac{2\gamma}{1-\gamma} \rceil$.

Thus we have by group privacy (see lemma 2.2 in [37]) that $\mathcal{M} \circ \mathcal{P}_{\text{pre}}$ gives the claimed DP-guarantee, as we set $K_\gamma = 2 + \lceil \frac{2\gamma}{1-\gamma} \rceil$

□

G Extended Background

In the following, we assume a classification task where a model $\theta : \mathcal{X} \times \mathcal{Z} \mapsto \mathcal{K}$ maps the features $(\mathbf{x}, z) \in \mathcal{X} \times \mathcal{Z}$ to a label $y \in \mathcal{K}$, where: $\mathcal{X}$ is the domain of non-sensitive attributes, $\mathcal{Z}$ is the domain of the sensitive attribute (as a categorical variable), and $\mathcal{K}$ is the domain of the output label (also categorical). Without loss of generality, we will assume $\mathcal{Z} = [Z]$ (*i.e.* $\mathcal{Z} = \{1, \ldots, Z\}$) and $\mathcal{K} = [K]$.

G.1 Fairness Notion: Demographic Parity

We note that in a multi-class setting (*i.e.*, $K > 2$), and even in the binary-class settings where the problem does not admit a reasonable notion of the "desirable outcome", there can be multiple formulations of the notion of demographic parity (Appendix H). We adopt a natural extension of the well-known binary notion that requires equal rates for any class. Let us first define demographic disparity:

The *demographic disparity* $\Gamma(z, k)$ of subgroup z for class k is the difference between the probability of predicting class k for the subgroup z and the probability of the same event for any other subgroup: $\Gamma(z, k) := \mathbb{P}[\hat{Y} = k \mid Z = z] - \mathbb{P}[\hat{Y} = k \mid Z \neq z]$.

G.2 Privacy Notion: Differential Privacy

In (ε, δ)-DP, the parameter ε bounds the maximal difference between the analysis results on the neighboring datasets while the second parameter δ represents a relaxation of the bound by allowing

[8] An example of such ordering would be to order images based on their pixel values in some specified order of height, width and channel starting by checking the first pixel, then the second pixel, and so on.

the results to vary more than the factor e^{ε}. Hence, the total privacy loss is bounded by ε with a probability of at least $1-\delta$ [13]. Note that smaller ε correspond to better privacy guarantees for the data.

PATE. (Figure 6), takes advantage of an unlabeled public data set D_{public} to conserve the privacy of sensitive data D_{private}. Therefore, an ensemble of B *teacher* models $\{\theta_i\}_{i=1}^{B}$ is trained using disjoint subsets of D_{private} and their knowledge is transferred to a separate *student* model that can be publicly released. For the knowledge transfer, trained teachers label query data points from D_{public}. The final label of the query is a noisy argmax over the vote counts as $N(\mathbf{x}) = \arg\max\left([n_{i,j}]_{B\times K} + \mathcal{N}(0,\sigma_1^2)\right)$, where K is the number of classes (see aggregation in Algorithm 4). Noising the argmax enables to implement the privacy guarantees according to DP.

PATE estimates the privacy cost of answering queries (*i.e.* labeling data) through *teachers consensus* with higher consensus revealing less information about individual teachers, and, thereby, consuming less privacy costs. To take advantage of the fact that estimating consensus is less privacy-costly than answering queries, PATE rejects high-cost queries to save on the privacy budget (see Algorithm 4). Both consensus estimation and vote aggregation (answering the query) are noised with $\mathcal{N}(0,\sigma_1^2)$ and $\mathcal{N}(0,\sigma_2^2)$, respectively; where σ_1, σ_2 are tuned for better student accuracy.

We include the standard Confident-GNMax Aggregator Algorithm from [25] below.

Algorithm 4 – Confident-GNMax Aggregator (from [25]) given a query, consensus among teachers is first estimated in a privacy-preserving way to then only reveal confident teacher predictions.

```
Require: input $x$, threshold $T$, noise parameters $\sigma_1$ and $\sigma_2$
if $\max_j\{\sum_{i\in[B]} n_{i,j}(x)\} + \mathcal{N}(0,\sigma_1^2) \geq T$ then
    return $\arg\max_j\{\sum_{i\in[B]} n_{i,j}(\mathbf{x}) + \mathcal{N}(0,\sigma_2^2)\}$
else
    return $\perp$
```

DP-SGD extends standard stochastic gradient descent (SGD) with two additional steps to implement privacy guarantees. First, the individual data points' gradients are clipped to a maximum gradient norm bound C. This bounds the gradients' sensitivity, which ensures that no data points can incur changes to the model above magnitude C. After clipping, Gaussian noise with scale $\mathcal{N}(0,\sigma^2 C^2)$ is added to mini-batches of clipped gradients. The noise distribution has zero mean and standard deviation proportional to a pre-defined noise multiplier σ and the clipping norm C. We detail the DP-SGD algorithm in Algorithm 5.

To yield tighter privacy bounds, DP-SGD implements a privacy amplification through subsampling [7]: Training data points are sampled into mini-batches with a Poisson sampling per training iteration, in contrast to grouping the entire training data into mini-batches prior to every epoch as done in standard SGD. Hence, the traditional concept of an epoch (as a full training on the entire training data) does not exist in DP-SGD. Instead, each data point is sampled in every iteration according to a given sampling probability. Privacy amplification through subsampling allows to scale down the noise σ by the factor L/N (with L being the expected mini-batch size, N the total number of data points, and $L \ll N$) while still ensuring the same ε as with σ [18] which is crucial to the practical performance (privacy-utility trade-offs) of DP-SGD.

We include the standard DP-SGD algorithm (Algorithm 5) and FairDP-SGD (Algorithm 6) here for comparison. Details of the FairDP-SGD algorithm is discussed in Section 3.1.

H Fairness Metrics and Evaluations

We evaluate and compare different ways to measure the demographic parity gap, $\Gamma(z,k)$. We then select one method to use in our implementations. We explore three different methods that compare the ratio between different sensitive groups to evaluate the chosen fairness metric.

Algorithm 5 Standard DP-SGD, adapted from [1].

```
Require: Private training set $D_{\text{prv}} = \{(x_i, y_i) \mid i \in [N_{\text{prv}}]\}$, loss function $\mathcal{L}(\theta, x_i)$, Parameters: learning rate $\eta_t$, noise scale $\sigma$, group size $L$, gradient norm bound $C$.
Initialize $\theta_0$ randomly
for $t \in [T]$ do
    Sample mini-batch $L_t$ with sampling probability $L/N$    ▷ Poisson sampling
    For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$    ▷ Compute gradient
    $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$    ▷ Clip gradient
    $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{|L_t|}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}\left(0, \sigma^2 C^2 \mathbf{I}\right)\right)$    ▷ Add noise
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$    ▷ Descent
Output $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$ using a privacy accounting method.
```

Algorithm 6 FairDP-SGD

```
Require: Private training set $D_{\text{prv}} = \{(x_i, y_i) \mid i \in [N_{\text{prv}}]\}$, Public calibration set $D_{\text{pub}} = \{(\tilde{x}_i, z_i) \mid i \in [N_{\text{pub}}]\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$, Demographic Parity loss $\text{DPFR}(\theta; D_{\text{pub}})$. Parameters: learning rate $\eta_t$, noise scale $\sigma$, group size $L$, gradient norm bound $C$.
Initialize $\theta_0$ randomly
for $t \in [T]$ do
    Sample mini-batch $L_t$ with sampling probability $L/N$    ▷ Poisson sampling
    For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\left(\mathcal{L}(\theta_t, x_i) + \lambda\,\text{DPFR}(\theta_t; D_{pub})\right)$    ▷ Compute gradient
    $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$    ▷ Clip gradient
    $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{|L_t|}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}\left(0, \sigma^2 C^2 \mathbf{I}\right)\right)$    ▷ Add noise
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$    ▷ Descent
Output $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$ using a privacy accounting method.
```

H.1 Demographic Parity Gap Measurements

1. Between Groups: This method computes and bounds the maximum difference between two pairs of sensitive groups.

$$\Gamma(z, k) := max_{\tilde{z}} |\mathbb{P}[\hat{Y} = k | Z = z] - \mathbb{P}[\hat{Y} = k \mid Z = \tilde{z}]|. \tag{2}$$

2. To Overall: This method computes and bounds the difference between each sensitive group and the total of all groups.

$$\Gamma(z, k) := \mathbb{P}[\hat{Y} = k | Z = z] - \mathbb{P}[\hat{Y} = k]. \tag{3}$$

3. To Overall Without Double Counting: This method computes and bounds the difference between each sensitive group and the total of all other groups.

$$\Gamma(z, k) := \mathbb{P}[\hat{Y} = k | Z = z] - \mathbb{P}[\hat{Y} = k \mid Z \neq z]. \tag{4}$$

To compare the three methods, we generate some synthetic data and run queries on them using each method to compare the results.

H.2 Evaluation Results

We first generate synthetic data with two classes and three sensitive groups. The distribution of the generated data is shown below.

Class/Sensitive Group	0	1	2
0	324	420	445
1	287	274	250

H.2.1 By Group

Total number of queries answered = 1661

Class/Sensitive Group	0	1	2
0	315	364	361
1	191	213	217

H.2.2 To Overall

Total number of queries answered = 1832

Class/Sensitive Group	0	1	2
0	324	395	361
1	234	271	247

H.2.3 To Overall Without Double Counting

Total number of queries answered = 1772

Class/Sensitive Group	0	1	2
0	318	371	342
1	218	244	229

H.3 Conclusion

We decide to use the third method, to overall without double counting, as the comparison method. It is a balance between the by group method and the to overall method. We do not want the comparison method to be too strict, because then our algorithm would reject most queries due to fairness. On the other hand, we also do not want it to be too lenient that the fairness constraint is not enforced. One major drawback of the to overall method is that if most of the data is from one sensitive group, then that sensitive group would have too much influence over the overall class label distribution.

I Experimental Setup

We split each dataset into a training set, an unlabeled set, and a test set. The sizes of these three datasets are determined based on the dataset sizes specified in original PATE [26, 25], and adapted to the difficulties of the prediction tasks. For CheXpert, we only use the data from two races that have the most data. The other groups have too few data points for our fairness intervention to perform effectively.

In FairPATE, the training set is further split into equal partitions to train the teacher models. We train as many teachers as possible while still achieving good ensemble accuracy overall. In FairDP-SGD, the whole training set is used to train the private model. The test set is used to evaluate the performance of the final model.

I.1 FairPATE

For FairPATE, we first train the teacher ensemble models, then query them with the public dataset, and aggregate their predictions using the FairPATE algorithm. The student model is trained on the public dataset with obtained labels. The model architectures, as well as the parameters used in querying the teacher models are detailed in Table 2 for each dataset, respectively. The model architectures are chosen by referencing what is used in related works for each dataset.

We tune the amount of noise injected into the aggregation mechanism of FairPATE by varying the standard deviation of the Gaussian distribution while ensuring the accuracy of the labels produced by the teacher ensemble models to maximize the accuracy of student models. We used a small validation set taken from the dataset to tune the FairPATE hyperparameters by training student models with different combinations of the hyperparameters and selecting the values that lead to the highest student model accuracy. The validation set is taken from the original training set and the size is half the size of the unlabeled set. When tuning, we vary the threshold T between 0.5 to 1.5 multiplied by the number of teacher models, σ_1 between 0 to the number of teacher models, and σ_2 between 0 to 0.5 multipled by the number of teacher models.

We train the models using Adam optimizer. We use cross entropy loss function when training on ColorMNIST and binary cross entropy with logits on all the other datasets.

I.2 FairDP-SGD

FairDP-SGD models are trained with the same model architecture as indicated in the Table 2.

We train the models with different λ values defined in the fairness mechanism to reflect different levels of fairness interventions. The range we use is between 0 and 10, with 0 being completely turning off the fairness mechanism.

We train the models using SGD optimizer. We use cross entropy loss function combined with Demographic Parity Loss (DPL) when training on ColorMNIST and binary cross entropy with logits with DPL on all the other datasets.

Dataset	Prediction Task	C	Sens. Attr.	SG	Total	U	Model	Number of Teachers	T	σ_1	σ_2
ColorMNIST [2]	Digit	10	Color	2	60 000	1 000	Convolutional Network (Table 3)	200	120	110	20
CelebA [22]	Smiling	2	Gender	2	202 599	9 000	Convolutional Network (Table 4)	150	130	110	10
FairFace [19]	Gender	2	Race	7	97 698	5 000	Pretrained ResNet50	50	30	30	10
UTKFace [41]	Gender	2	Race	5	23 705	1 500	Pretrained ResNet50	100	50	40	15
CheXpert[16] [30]	Disease	2	Race	3	152 847	4 000	Pretrained DenseNet121	50	30	20	10

Table 2: Datasets used for evaluation. Abbreviations: **C**: number of classes in the main task; **SG**: number of sensitive groups; **U**: number of unlabeled samples for the student training . Summary of parameters used in training and querying the teacher models for each dataset. The selection of σ_1 is in accordance with the threshold T. The selection process of σ_2, is shown in the Appendix I.The pre-trained models are all pre-trained on ImageNet. We use the most recent versions from PyTorch.

Layer	Description
Conv2D with ReLU	(3, 20, 5, 1)
Max Pooling	(2, 2)
Conv2D with ReLU	(20, 50, 5, 1)
MaxPool	(2, 2)
Fully Connected 1	(4*4*50, 500)
Fully Connected 2	(500, 10)

Table 3: Convolutional network architecture used in ColorMNIST experiments.

Layer	Description
Conv2D	(3, 64, 3, 1)
Max Pooling	(2, 2)
ReLUS	
Conv2D	(64, 128, 3, 1)
Max Pooling	(2, 2)
ReLUS	
Conv2D	(128, 256, 3, 1)
Max Pooling	(2, 2)
ReLUS	
Conv2D	(256, 512, 3, 1)
Max Pooling	(2, 2)
ReLUS	
Fully Connected 1	(14 * 14 * 512, 1024)
Fully Connected 2	(1024, 256)
Fully Connected 2	(256, 2)

Table 4: Convolutional network architecture used in CelebA experiments.

I.3 Wall Time Measurements

We measure the wall time of running FairPATE and FairDP-SGD compared to PATE and DP-SGD. The setting we use and the results are shown in Table 5.

J Relationship between Number of Queries Answered and Student Accuracy

We run a set of query experiments to investigate the trade-offs between privacy, fairness, and model utility. We do not train the student model for these querying experiments, but they will be trained later

Method	ϵ	γ	Fairness Factor	Batch Size	Number of Epochs	Time
FairPATE	4	0.01	N/A	100	30	6min 14sec
PATE	4	N/A	N/A	100	30	6min 43sec
FairDP-SGD	4	N/A	5	80	30	2h 27min 26sec
DP-SGD	4	N/A	N/A	80	30	1h 18min 28sec

Table 5: Wall time measurements of different methods. All experiments use UTKFace dataset.

on. Instead, we use the number of queries answered as an estimate of the student model utility since an adequate number of queries needs to be answered to train a student model with good accuracy.In the first set of experiments, we run queries with varying consensus threshold T and fairness violation threshold ρ_{fair} at fixed privacy budget ε, and record the number of queries answered. We query the teacher ensemble models with varying privacy budget ε and fairness violation threshold ρ_{fair}. For these queries, we measure the maximum fairness violation γ, the achieved ε, and the number of queries answered. Using these query results, we also select and plot the points on the Pareto frontier. The results for the UTKFace dataset are shown in Figure 12. The results on the other datasets are found in Figure 14.

Figure 12 (left) plots the the trade-offs between the maximum fairness violation γ, the achieved ε, and the number of queries answered. As expected, we observe that increasing ε allows more queries to be answered. Relaxing ρ_{fair} at fixed ε also leads to more queries being answered, although the effect is not as apparent. Additionally, when ε is very low, smaller γ is not achievable due to having too few queries answered and the fairness regulation mechanism not being activated as a result.

Figure 12 (right) plots the Pareto frontier of the query results. We plot the privacy constraint, fairness constraint, and the number of queries answered as a 3D plot to better visualize the tension between these different objectives. The figure gives similar insights as the other figure. Another observation is that although smaller γ is achievable when a higher number of queries are answered, at some point the fairness constraint needs to be relaxed in order to answer more queries.

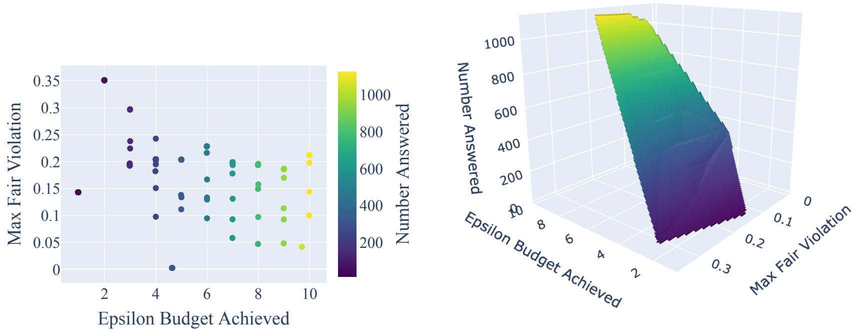


Figure 12: **Query Experiment Results on UTKFace**. Experimental setup from Table 2. The left figure shows the the trade-offs between the maximum ensemble query fairness violation γ_{ens}, the achieved ε, and the number of queries answered. The right figure plots the Pareto-frontier. With increasing privacy budget, more queries can be answered. The same holds when loosening the fairness constraint. At small privacy budgets, small fairness constraint might not be achievable.

We run an additional set of experiments of querying the teacher ensemble models to investigate the effect of different parameters on the number of queries answered. For these experiments, we run queries with varying consensus threshold T and fairness violation threshold ρ_{fair} at fixed privacy budget ε, and record the number of queries answered. Appendix J plots the results on UTKFace, and the results on other datasets are in Figure 15. The graph shows the effect of varying the consensus threshold T and fairness violation threshold ρ_{fair} on the number of queries answered. We observe that decreasing T leads to a higher number of queries answered. Similarly, increasing ρ_{fair} to a certain extent also leads to more queries being answered. Once the fairness violation threshold is too large, further relaxing the constraint would not lead to answering more queries, at which point no more queries are rejected due to the fairness constraint. Furthermore, at a fairness constraint of 0, there is a sharp decrease in the number of queries answered. The reason behind this is that if no fairness violation is allowed, no more queries can be answered after the fairness gap reaches 0, as any additional query would break the balance and increase the gap.

K Extended Related Work: Integrating Fairness into Private Learning

In the literature, different fairness notions have been implemented within DP-SGD and PATE frameworks.

Fairness and DP-SGD. It has been shown that training with DP-SGD leads to disparate accuracy decrease over different data sub-groups [31, 14]. In particular, model accuracy decreases more for underrepresented data from the tails of the distribution [31]. Farrand et al. [14] presented similar findings and observed that privacy can even have a negative impact on the model fairness when the training data is only slightly imbalanced. As potential reasons for this, the authors identified the clipping operation in DP-SGD. Since underrepresented data has larger gradients, these gradients are more effected by the clipping operation, and thereby, this data experiences a higher information loss [14]. To limit this effect, Xu et al. [39] proposed adapting the clipping threshold in DP-SGD individually for each sensitive group. They showed how their approach limits the disparate impact of DP-SGD on different groups. However, due to higher information leakage form larger gradients, their method requires larger perturbations. In a similar vein, Zhang et al. [40] propose early stopping to mitigate the negative impact of DP-SGD on model fairness. The authors observe that DP-SGD makes ML model training less stable which they leverage to interrupt training once high-enough fairness is achieved, without a significant loss in accuracy. However, all these methods solely manage fairness as an indirect byproduct of adapting the private training mechanism. Neither of them integrates explicit fairness constraints to yield formal guarantees, such as done in this work.

Tran et al. [34] proposed applying a Lagrangian dual approach for solving the joint optimization of fairness and privacy in ML. Therefore, they rely on a fairness constraint plus adaptive clipping and make the computations of the primal and dual update steps differentially private w.r.t. the considered sensitive attributes. However, their method adds a significant computational overhead, especially for larger ML models and mini-batch sizes (increase of up to factor 100).

Fairness and PATE. When comparing the fairness impact of DP-SGD and PATE, Uniyal et al. [36] observed that PATE induces lower accuracy parity. The authors reason that this might be because the diversity among the teachers allows to cancel out their individual fairness issues. However, their observations only hold for very small numbers of teachers (10, in contrast to 250 proposed for MNIST in the original PATE paper [26]). This however yields sub-optimal privacy-utility trade-offs since in PATE, stronger privacy guarantees can be obtained when using more teachers which allows for the injection of more noise. In the work closest to ours, Tran et al. [35] study fairness properties of PATE and identified both algorithmic properties of the training (number of teachers, regularizer, privacy noise), and properties of the student data (magnitude of the input norm, and distance to the decision boundary) as factors influencing prediction fairness. To mitigate tensions, they proposed releasing

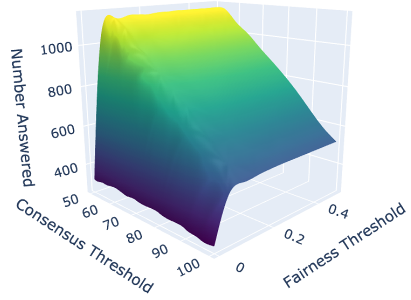
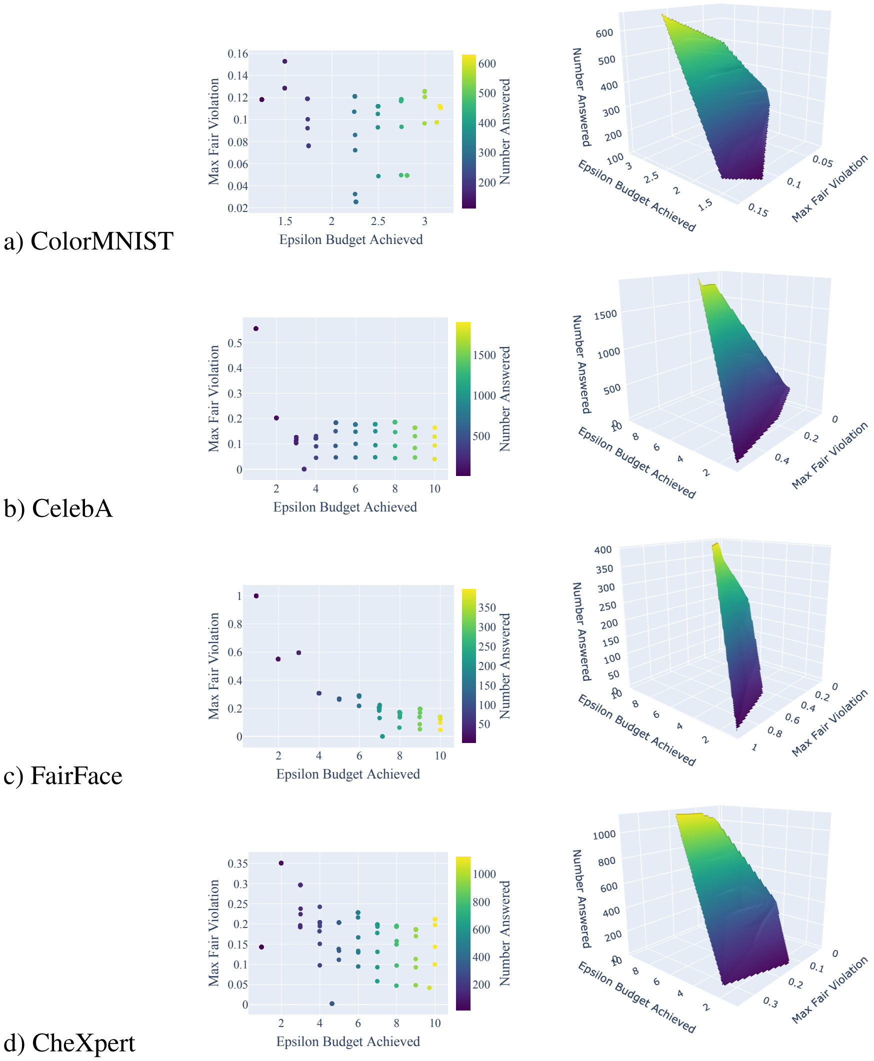


Figure 13: **Query Experiment Results on UTKFace**. The figure plots the effect of consensus threshold T and fairness threshold $\gamma_{threshold}$ on the number of queries answered. We observe that the number of queries answered increases with smaller T and larger $\gamma_{threshold}$.



a) ColorMNIST

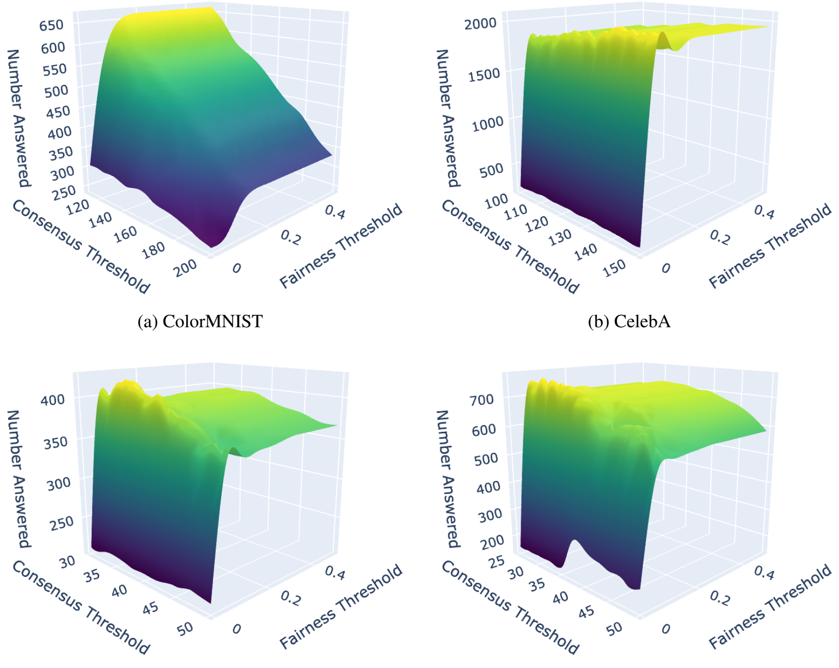
b) CelebA

c) FairFace

d) CheXpert

Figure 14: Query experiments on other datasets. Setup described in Table 2 and discussion is in Appendix J.

the teacher models' prediction histogram as *soft labels* to train the student model. However, it has been shown that releasing the histograms leaks significant amounts of private information [38], which makes their method leaks privacy above the promised DP guarantees. In contrast, in this work, we integrate fairness in the aggregation process while keeping the teachers' votes private, and, thereby providing the promised privacy guarantees.



(a) ColorMNIST

(b) CelebA

(c) FairFace

(d) CheXpert

Figure 15: Query experiments on other datasets. Setup described in Table 2 and discussion is in Appendix J. We found that in order to obtain the best results on student accuracy, some datasets require addition of significant noise σ_1, which leads to differences in surfaces' shapes.