

# A FAIRNESS ANALYSIS ON DIFFERENTIALLY PRIVATE AGGREGATION OF TEACHER ENSEMBLES SUPPLEMENTAL MATERIAL

## A MISSING PROOFS

This section contains the missing proofs associated with the theorems and corollaries presented in the main paper. The theorems are restated for completeness.

First we provide the upper bound on the excess risk per group  $a \in \mathcal{A}$  in the following Lemma [1](#). This helps to understand what factors control the excess risk for a particular group.

**Lemma 1.** *The excess risk  $R(\bar{D}_{\leftarrow a})$  of a group  $a \in \mathcal{A}$  is upper bounded as:*

$$R(D_{\leftarrow a}) \leq \|G_a\| \mathbb{E}[\Delta_{\tilde{\theta}}] + 1/2 \beta_a \mathbb{E}[\Delta_{\tilde{\theta}}^2], \quad (10)$$

where  $G_a = \mathbb{E}_{\mathbf{x} \sim \bar{D}_{\leftarrow a}} [\nabla_{\theta^*} \ell(\tilde{f}_{\theta^*}(\mathbf{x}), y)]$  is the gradient of the group loss evaluated at  $\theta^*$ , and  $\Delta_{\tilde{\theta}}$  and  $\Delta_{\tilde{\theta}}^2$  capture the first and second order statistics of the model deviation.

*Proof.* By  $\beta_a$  smoothness assumption on the loss function defined over a group  $a \in \mathcal{A}$  it follows that:

$$\mathcal{L}(\tilde{\theta}; D_{\leftarrow a}, T) \leq \mathcal{L}(\theta^*; D_{\leftarrow a}, T) + (\tilde{\theta} - \theta^*)^T G_a + \frac{\beta_a}{2} \|\tilde{\theta} - \theta^*\|^2. \quad (11)$$

By taking the expectation on both sides of the above equation w.r.t. the randomness of the noise, we obtain:

$$\mathbb{E}[\mathcal{L}(\tilde{\theta}; D_{\leftarrow a}, T)] \leq \mathcal{L}(\theta^*; D_{\leftarrow a}, T) + G_a^T \mathbb{E}[(\tilde{\theta} - \theta^*)] + \frac{\beta_a}{2} \mathbb{E}[\|\tilde{\theta} - \theta^*\|^2] \quad (12)$$

$$\leq \mathcal{L}(\theta^*; D_{\leftarrow a}, T) + \|G_a\| \mathbb{E}[\Delta_{\tilde{\theta}}] + \frac{1}{2} \beta_a \mathbb{E}[\Delta_{\tilde{\theta}}^2], \quad (13)$$

where the last inequality is by Cauchy-Schwarz inequality on vectors. Next, by substituting  $R(\bar{D}_{\leftarrow a}) = \mathbb{E}[\mathcal{L}(\tilde{\theta}; D_{\leftarrow a}, T)] - \mathcal{L}(\theta^*; D_{\leftarrow a}, T)$  into Equation [13](#) we obtain the Lemma statement.  $\square$

**Theorem 1.** *The model fairness is upper bounded as:*

$$\xi(\bar{D}) \leq \max_a 2\|G_a\| \mathbb{E}[\Delta_{\tilde{\theta}}] + \max_a 1/2 \beta_a \mathbb{E}[\Delta_{\tilde{\theta}}^2]. \quad (14)$$

*Proof.* By convexity assumption on the loss function defined over a group  $a \in \mathcal{A}$  it follows that:

$$\mathcal{L}(\theta^*; D_{\leftarrow a}, T) + (\tilde{\theta} - \theta^*)^T G_a \leq \mathcal{L}(\tilde{\theta}; D_{\leftarrow a}, T) \quad (15)$$

By taking the expectation on both sides of the above equation w.r.t. the randomness of the noise, we obtain:

$$\mathbb{E}[\mathcal{L}(\tilde{\theta}; D_{\leftarrow a}, T)] \geq \mathcal{L}(\theta^*; D_{\leftarrow a}, T) + \mathbb{E}[(\tilde{\theta} - \theta^*)^T] G_a \quad (16)$$

By combining Equation [16](#) and Equation [12](#) we obtain the following:

$$\mathbb{E}[(\tilde{\theta} - \theta^*)^T] G_a \leq R(\bar{D}_{\leftarrow a}) \leq \mathbb{E}[(\tilde{\theta} - \theta^*)^T] G_a + \frac{\beta_a}{2} \mathbb{E}[\Delta_{\tilde{\theta}}^2] \quad (17)$$

Based on the definition of fairness in Equation [4](#), it follows that:

$$\xi(\bar{D}) = \max_{a, a' \in \mathcal{A}} R(\bar{D}_{\leftarrow a}) - R(\bar{D}_{\leftarrow a'}) \leq \max_{a, a' \in \mathcal{A}} \mathbb{E}[(\tilde{\theta} - \theta^*)^T] (G_a - G_{a'}) + \max_{a \in \mathcal{A}} \frac{\beta_a}{2} \mathbb{E}[\Delta_{\tilde{\theta}}^2] \quad (18)$$

$$\leq \max_a 2\|G_a\| \mathbb{E}[\|\tilde{\theta} - \theta^*\|] + \max_a \frac{\beta_a}{2} \mathbb{E}[\Delta_{\tilde{\theta}}^2] = 2 \max_a \|G_a\| \mathbb{E}[\Delta_{\tilde{\theta}}] + \max_a \frac{\beta_a}{2} \mathbb{E}[\Delta_{\tilde{\theta}}^2] \quad (19)$$

$\square$

**Theorem 2.** Consider a student model  $\bar{f}_\theta$  trained with a convex and decomposable loss function  $\ell(\cdot)$ . Then, the expected difference between the private and non-private model parameters is upper bounded as follows:

$$\mathbb{E}[\Delta_\theta] \leq \frac{|c|}{m\lambda} \left[ \sum_{x \in \mathcal{D}} p_x^\leftrightarrow \|G_x^{\max}\| \right], \quad (20)$$

where  $c$  is a real constant and  $G_x^{\max} = \max_\theta \|\nabla_\theta h_\theta(x)\|$  represents the maximum gradient norm distortion introduced by a sample  $x$ . Both  $c$  and  $h$  are defined as in Equation equation 6

Proof of Theorem 2 requires the following Lemma 2 from Shalev-Shwartz (2007) on the property of strongly convex functions.

**Lemma 2** (Shalev-Shwartz (2007)). Let  $\mathcal{L}(\theta)$  be a differentiable function. Then  $\mathcal{L}(\theta)$  is  $\lambda$ -strongly convex iff for all vectors  $\theta, \theta'$ :

$$(\nabla_\theta \mathcal{L} - \nabla_{\theta'} \mathcal{L})^T (\theta - \theta') \geq \lambda \|\theta - \theta'\|^2. \quad (21)$$

*Proof of Theorem 2* Let us denote with  $\hat{y}_i = v(T(x_i))$  to indicate the non-private voting label associated with  $x_i$  and  $\tilde{y}_i = \tilde{v}(T(x_i))$  for the private voting label counterpart. The regularized empirical risk function with the non-private voting labels from Equation 1 can be rewritten as follows:

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^m \ell(\bar{f}_\theta(x_i), \hat{y}_i) + \lambda \|\theta\| \quad (22)$$

$$= \frac{1}{m} \sum_{i=1}^m [z(h_\theta(x_i)) + c \hat{y}_i h_\theta(x_i)] + \lambda \|\theta\|^2, \quad (23)$$

where the second equality is due to the decomposable loss assumption. Likewise, define  $\tilde{\mathcal{L}}$  to be the regularized empirical risk function with private voting labels  $\tilde{y}_i$ :

$$\tilde{\mathcal{L}} = \frac{1}{m} \sum_{i=1}^m [z(h_\theta(x_i)) + c \tilde{y}_i h_\theta(x_i)] + \lambda \|\theta\|^2, \quad (24)$$

Based on Equation 23 and Equation 24, it follows that:  $\tilde{\mathcal{L}} = \mathcal{L} + \Delta_\mathcal{L}$  where  $\Delta_\mathcal{L} = \frac{c}{m} \sum_{i=1}^m (\tilde{y}_i - \hat{y}_i) h_\theta(x_i)$ .

Furthermore, since each individual loss function  $\ell(\bar{f}_\theta(x_i), \tilde{y}_i)$  or  $\ell(\bar{f}_\theta(x_i), \hat{y}_i)$  is convex for all  $i$  from the given assumption, then  $\tilde{\mathcal{L}}$  and  $\mathcal{L}$  both are  $\lambda$ -strongly convex.

Next, from the definition of  $\tilde{\theta} = \arg \min_\theta \tilde{\mathcal{L}}$ , and  $\theta^* = \arg \min_\theta \mathcal{L}$  it follows that:

$$\nabla_{\tilde{\theta}} \tilde{\mathcal{L}} = \mathbf{0} \text{ and } \nabla_{\theta^*} \mathcal{L} = \mathbf{0}. \quad (25)$$

By Lemma 2 it follows that:

$$(\nabla_{\tilde{\theta}} \tilde{\mathcal{L}} - \nabla_{\theta^*} \tilde{\mathcal{L}})^T (\tilde{\theta} - \theta^*) \geq \lambda \|\tilde{\theta} - \theta^*\|^2. \quad (26)$$

Now since  $\nabla_{\tilde{\theta}} \tilde{\mathcal{L}} = \mathbf{0}$  by Equation 25, we can rewrite Equation 26 as

$$(-\nabla_{\theta^*} \tilde{\mathcal{L}})^T (\tilde{\theta} - \theta^*) \geq \lambda \|\tilde{\theta} - \theta^*\|^2, \quad (27)$$

since  $\nabla_{\tilde{\theta}} \tilde{\mathcal{L}} = \nabla_{\theta^*} \mathcal{L} + \nabla_{\theta^*} \Delta_\mathcal{L} = \mathbf{0} + \nabla_{\theta^*} \Delta_\mathcal{L} = \nabla_{\theta^*} \Delta_\mathcal{L}$ . In addition, by applying the Cauchy-Schwartz inequality to the L.H.S of Equation 27 we obtain

$$\|\nabla_{\theta^*} \Delta_\mathcal{L}\| \|\tilde{\theta} - \theta^*\| \geq -(\nabla_{\theta^*} \Delta_\mathcal{L})^T (\tilde{\theta} - \theta^*) \geq \lambda \|\tilde{\theta} - \theta^*\|^2, \quad (28)$$

and thus,

$$\|\nabla_{\theta^*} \Delta_\mathcal{L}\| \geq \lambda \|\tilde{\theta} - \theta^*\|. \quad (29)$$

By definition of  $\nabla_{\tilde{\theta}} \Delta_{\mathcal{L}}$  we can rewrite the above inequality as follows:

$$\|\nabla_{\tilde{\theta}} \Delta_{\mathcal{L}}\| = \left\| \frac{c}{m} \sum_{i=1}^m (\tilde{y}_i - \hat{y}_i) \nabla_{\tilde{\theta}} h_{\tilde{\theta}}(x_i) \right\| \geq \lambda \|\tilde{\theta} - \hat{\theta}\|^2. \quad (30)$$

Next, let  $\rho_i = \hat{y}_i - \tilde{y}_i$ , applying this substitution to the above and by triangle inequality it follows that:

$$\frac{|c|}{m} \sum_{i=1}^m |\rho_i| \|g_i\| \geq \frac{|c|}{m} \sum_{i=1}^m |\rho_i| \|\nabla_{\tilde{\theta}} h_{\tilde{\theta}}(x_i)\| \quad (31)$$

$$\geq \left\| \frac{c}{m} \sum_{i=1}^m \rho_i \nabla_{\tilde{\theta}} h_{\tilde{\theta}}(x_i) \right\| \geq \lambda \|\tilde{\theta} - \hat{\theta}\|, \quad (32)$$

where the first inequality is due to definition of  $g_{x_i} = \max_{\theta} \|\nabla_{\theta} h_{\theta}(x_i)\|$  and the second inequality is due to the general triangle inequality. Since  $|\rho_i|$  is a Bernoulli random variable, in which  $|\rho_i| = 1$  w.p.  $p_{x_i}^{\leftrightarrow}$  and  $|\rho_i| = 0$  w.p of  $1 - p_{x_i}^{\leftrightarrow}$ . Therefore  $\mathbb{E}[|\rho_i|] = p_{x_i}^{\leftrightarrow}$ . Thus, it follows that:

$$\mathbb{E}\left[\frac{|c|}{m} \sum_{i=1}^m |\rho_i| \|g_{x_i}\|\right] = \frac{|c|}{m} \sum_{i=1}^m p_{x_i}^{\leftrightarrow} \|g_{x_i}\| \geq \lambda \mathbb{E}[\|\tilde{\theta} - \hat{\theta}\|] = \mathbb{E}[\Delta_{\tilde{\theta}}], \quad (33)$$

which concludes the proof.  $\square$

**Theorem 3.** For a sample  $x \in \bar{D}$  let the teacher models outputs  $f^i(x)$  be in agreement,  $\forall i \in [k]$ . The flipping probability  $p_x^{\leftrightarrow}$  is given by  $p_x^{\leftrightarrow} = 1 - \Phi(\frac{k}{\sqrt{2}\sigma})$ , where  $\Phi(\cdot)$  is the CDF of the std. Normal distribution and  $\sigma$  is the standard deviation in the Gaussian mechanism.

For simplicity of exposition Theorem 3 considers binary classifiers, i.e.,  $\mathcal{Y} = \{0, 1\}$ . The argument, however, can be trivially extended to generic  $C$ -classifiers.

*Proof.* By assumption, for any given sample  $x$ , all teachers agree in their predictions, so w.l.o.g., assume  $k$  teachers output label 0, while none of them outputs label 1. Next, let  $\psi, \psi' \sim \mathcal{N}(0, \sigma^2)$  be two independent Gaussian random variables which are added to true voting counts,  $k$  and 0, respectively. The associated flipping probability is:

$$p_x^{\leftrightarrow} = \Pr[\tilde{v}(T(x)) \neq v(T(x))] = \Pr(k + \psi \leq 0 + \psi') = \Pr(\psi' - \psi \geq k) \quad (34)$$

$$= 1 - \Pr(\psi - \psi' \leq k), \quad (35)$$

since  $\psi, \psi'$  are two independent Gaussian random variable with zero mean and standard deviation of  $\sigma$ . Therefore,  $\psi' - \psi \sim \mathcal{N}(0, 2\sigma^2)$ . Thus:

$$\Pr(\psi - \psi' \leq k) = \Pr(\mathcal{N}(0, 2\sigma^2) \leq k) = \Phi\left(\frac{k}{\sqrt{2}\sigma}\right).$$

Hence, the flipping probability will be:  $p_x^{\leftrightarrow} = 1 - \Phi(\frac{k}{\sqrt{2}\sigma})$ .  $\square$

**Corollary 1** (Theorem 2). Let  $\tilde{f}_{\tilde{\theta}}$  be a logistic regression classifier. Its expected model deviation is upper bounded as:

$$\mathbb{E}[\Delta_{\tilde{\theta}}] \leq \frac{1}{m\lambda} \left[ \sum_{x \in \bar{D}} p_x^{\leftrightarrow} \|x\| \right]. \quad (36)$$

*Proof.* The loss function  $\ell(\tilde{f}_{\tilde{\theta}}(x), y)$  of a logistic regression classifier with binary cross entropy loss can be rewritten as follows:

$$\ell(\tilde{f}_{\tilde{\theta}}(x), y) = -y \log\left(\frac{1}{1 + \exp(-\theta^T x)}\right) - (1 - y) \log\left(\frac{\exp(-\theta^T x)}{1 + \exp(-\theta^T x)}\right) \quad (37)$$

$$= y \log(\exp(-\theta^T x)) - \log\left(\frac{\exp(-\theta^T x)}{1 + \exp(-\theta^T x)}\right) \quad (38)$$

$$= y(-\theta^T x) - \log\left(\frac{\exp(-\theta^T x)}{1 + \exp(-\theta^T x)}\right). \quad (39)$$

Hence,  $\ell(\cdot)$  is decomposable by Definition 3 with  $h_\theta(x) = -\theta^T x$ ,  $c = 1$  and  $z(h) = -\log(\frac{\exp(h)}{1+\exp(h)})$ .

Applying Theorem 2 with  $G_x^{\max} = \max_\theta \|\nabla_\theta h_\theta(x)\| = \max_\theta \|\nabla_\theta - \theta^T x\| = \|x\|$ , and  $c = 1$ , gives the intended result.  $\square$

**Corollary 2** (Theorem 2). *Given the same settings and assumption of Theorem 2 it follows:*

$$\mathbb{E}[\Delta_\theta^2] \leq \frac{|c|^2}{m\lambda^2} \left[ \sum_{x \in \bar{D}} p_x^{\leftrightarrow 2} \|G_x^{\max}\|^2 \right]. \quad (40)$$

*Proof.* First, by Theorem 2 we obtain an upper bound for  $\mathbb{E}[\Delta_\theta^2]$  as follows:

$$\mathbb{E}[\Delta_\theta^2] \leq \frac{c^2}{\lambda^2} \left[ \frac{1}{m} \sum_{x \in \bar{D}} p_x^{\leftrightarrow} \|G_x^{\max}\| \right]^2. \quad (41)$$

Applying the sum of squares inequality on the R.H.S. of Equation 41 we obtain:

$$\frac{c^2}{\lambda^2} \left[ \frac{1}{m} \sum_{x \in \bar{D}} p_x^{\leftrightarrow} \|G_x^{\max}\| \right]^2 \leq \frac{c^2}{\lambda^2} \left[ \frac{1}{m} p_x^{\leftrightarrow 2} \|G_x^{\max}\|^2 \right], \quad (42)$$

which concludes the proof.  $\square$

## B PRIVACY ANALYSIS

This section provides the privacy analysis for the original PATE model and the proposed mitigation solution. In PATE with the noisy-max scheme presented in Equation 2 of the main paper (also called GNMAX), the privacy budget is used for releasing the voting labels  $\tilde{v}(T(x_i))$  (a.k.a. hard labels) for each of the  $m$  public data samples  $x_i \in \bar{D}$  according to:

$$\tilde{v}(T(x_i)) = \arg \max_c \{ \#_c(T(x_i)) + \mathcal{N}(0, \sigma^2) \} \quad (43)$$

The proposed mitigation solution, instead, releases privately the voting counts  $(\#_c(T(x_i)) + \mathcal{N}(0, \sigma^2))_{c=1}^C$  and uses these noisy counts to construct the *soft-labels*, see Equation (11).

Using an analogous analysis as that provided in Papernot et al. (2018), adding or removing one individual sample  $x$  from any disjoint partition  $D_i$  of  $D$  can change the voting count vector by at most two. This value of the query deviation is obtained by GNMAX Papernot et al. (2018). Therefore the privacy cost for releasing hard labels or soft-labels is equivalent.

Next, this section provides the privacy computation  $\epsilon$  given by Gaussian mechanism which adds Gaussian noise with standard deviation  $\sigma$  to the voting counts.

The privacy analysis of PATE with hard or soft-labels is based on the concept of Renyi differential privacy (RDP) Mironov (2017). In either implementations, the process uses the Gaussian mechanism to add independent Gaussian noise to the voting counts. The following Proposition 1 from Papernot et al. (2018) derives the privacy guarantee for GNMAX.

**Proposition 1.** *The GNMAX aggregator with private Gaussian noise  $\mathcal{N}(0, \sigma^2)$  satisfies  $(\gamma, \gamma/\sigma^2)$ -RDP for all  $\gamma \geq 1$ .*

Since the GNMAX mechanism is applied on  $m$  public data samples from  $\bar{D}$ , the total privacy loss spent to provide the private labels is derived by the following composition theorem.

**Theorem 4** (Composition for RDP). *If a mechanism  $\mathcal{M}$  consists of a sequence of adaptive mechanisms  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_m$  such that for any  $i \in [m]$ ,  $\mathcal{M}_i$  guarantees  $(\gamma, \epsilon_i)$ -RDP, then  $\mathcal{M}$  guarantees  $(\gamma, \sum_{i=1}^m \epsilon_i)$ -RDP.*

Based on Theorem 4 and Proposition 1, PATE satisfies  $(\gamma, m\gamma/\sigma^2)$ -RDP. PATE also satisfies  $(\epsilon, \delta)$ -DP by the following theorem.

**Theorem 5** (From RDP to DP). *If a mechanism  $\mathcal{M}$  guarantees  $(\gamma, \epsilon)$ -RDP, then  $\mathcal{M}$  guarantees  $(\epsilon + \frac{\log 1/\delta}{\gamma-1}, \delta)$ -DP for any  $\delta \in (0, 1)$ .*

As a result of Theorem 5, PATE (with either hard or soft labels) satisfies  $(m\gamma/\sigma^2 + \frac{\log 1/\delta}{\gamma-1}, \delta)$ -DP.

## C EXPERIMENTAL ANALYSIS (EXT)

This section reports detailed information about the experimental setting as well as additional results conducted on the Income, Bank, Parkinsons, Credit Card and UTKFace datasets.

### C.1 SETTING AND DATASETS

**Computing Infrastructure** All of our experiments are performed on a distributed cluster equipped with Intel(R) Xeon(R) Platinum 8260 CPU @ 2.40GHz and 8GB of RAM.

**Software and Libraries** All models and experiments were written in Python 3.7. All neural network classifier models in our paper were implemented in Pytorch 1.5.0.

The Tensorflow Privacy package was also employed for computing the privacy loss.

**Datasets** This paper evaluates the fairness analysis of PATE on the following four UCI datasets: *Bank, Income, Parkinsons, Credit card* and UTKFace dataset. A descriptions of each dataset is reported as follows:

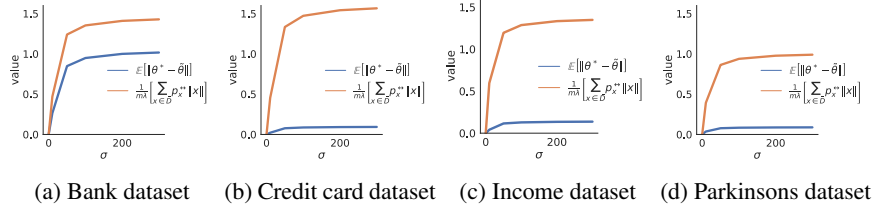
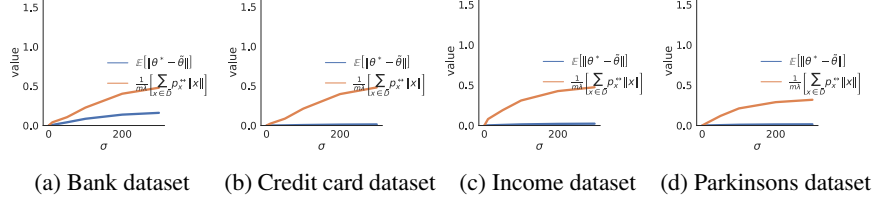
1. **Income** (Adult) dataset, where the task is to predict if an individual has low or high income, and the group labels are defined by race: *White vs Non-White* Blake & Merz (1988).
2. **Bank** dataset, where the task is to predict if a user subscribes a term deposit or not and the group labels are defined by age: *people whose age is less than vs greater than 60 years old* Blake & Merz (1988).
3. **Parkinsons** dataset, where the task is to predict if a patient has total UPDRS score that exceeds the median value, and the group labels are defined by gender: *female vs male* Little et al. (2007).
4. **Credit Card** dataset, where the task is to predict if a customer defaults a loan or not. The group labels are defined by gender: *female vs male* Carcillo et al. (2019).
5. **UTKFace** dataset, where the task is to predict the gender of a given facial image. The group labels are defined based on the following 9 age ranges: 0-10, 10-15, 15-20, 20-25, 25-30, 30-40, 40-50, 50-60, 60-120. Hwang et al. (2020)

On each dataset we perform standardization to render all input features with zero mean and unit standard deviation. Each dataset was partitioned into three disjoint subsets: private set, public train, and test set, as follows. We randomly select 75% of the dataset to use as private data and the rest for public data. For the public data,  $m = 200$  samples are randomly selected to train the student model, and the rest of the data is used as a test set to evaluate that model.

### Models' Setting

To visually show how tight the upper bound from Corollary 1 is, the paper uses a logistic regression model with 1000 runs to estimate the expected model deviation  $\mathbb{E}[\Delta_\theta] = \mathbb{E}[\|\tilde{\theta} - \hat{\theta}\|]$ .

For other experiments, the paper uses a neural network with two hidden layers and nonlinear ReLU activations for both the ensemble and student models. All reported metrics are an average of 100 repetitions, used to compute the empirical expectations. The batch size for stochastic gradient descent is fixed to 32 and the learning rate is  $\eta = 1e - 4$ .

Figure 8: Upper bound of the expected model deviation on 4 datasets with  $\lambda = 20, k = 20$ .Figure 9: Upper bound of the expected model deviation on 4 datasets with  $\lambda = 100, k = 200$ .

## C.2 UPPER BOUND OF THE EXPECTED MODEL DEVIATION

The following provides empirical results on Corollary 1 on four benchmark datasets. As indicated in this corollary, the expected model deviation is bounded by  $\frac{1}{m\lambda} [\sum_{x \in \bar{D}} p_x^{\leftrightarrow} \|x\|]$ . To visualize how tight the bounds are we report the RHS and LHS values of Equation 8 on different datasets. We run with two settings:  $k = 20, \lambda = 20$  in Figure 8 and  $k = 200, \lambda = 100$  in Figure 9.

## C.3 THE IMPACT OF REGULARIZATION PARAMETER

This section provides further empirical supports regarding impact of the regularization parameter  $\lambda$  to the accuracy and fairness trade-off. As seen from Theorem 2, increasing  $\lambda$  reduces the model deviation which in turns decreases the group excessive risk  $R(\bar{D}_{\leftarrow a})$  by Theorem 2 from the main text. On the other hand, large regularization can intuitively impacts negatively to the model accuracy. This was verified empirically in Figure 10 which shows how model deviation(left), excessive risk difference between two groups (middle) and utility(right) vary according to  $\lambda$ .

## C.4 THE IMPACT OF TEACHERS ENSEMBLE SIZE K

This section illustrates the effect of teacher ensemble sizes  $k$  to: 1) flipping probability  $p_x^{\leftrightarrow}$ , and 2) the trade-offs among model deviation  $\mathbb{E}[\Delta_{\bar{\theta}}]$ , model’s fairness and utilities.

First, Theorem 3 from the main text shows that larger  $k$  values correspond to smaller flipping probability  $p_x^{\leftrightarrow}$ . We provide more empirical evidence on other datasets and report the dependency between flipping probability with number of teachers  $k$  in Figure 11. It can be observed consistently on all datasets, the more number of teachers  $k$ , the smaller the flipping probability  $p_x^{\leftrightarrow}$  over all samples  $x$  is.

Second, regarding to the fairness analysis, similar to the previous subsection, we provide additional empirical supports on the effects of  $k$  on the model deviation, the difference between the group excessive risk, and the utility of the PATE models. We report these metrics on the other three benchmarks datasets in Figure 12. A similar trend with the regularization parameter  $\lambda$  also holds for the parameter  $k$  here. When the parameter  $k$  is increased to a large enough value, both model deviation and accuracy decreases, but the unfairness measured by the excessive risk difference between two groups reduces. This can be explained by looking again Figure 11 and Theorem 3 from the main text. A large number of teachers  $k$  results to a smaller flipping probability which in turns reduces the model deviation. By Theorem 2 a small model deviation can reduce the level of unfairness.

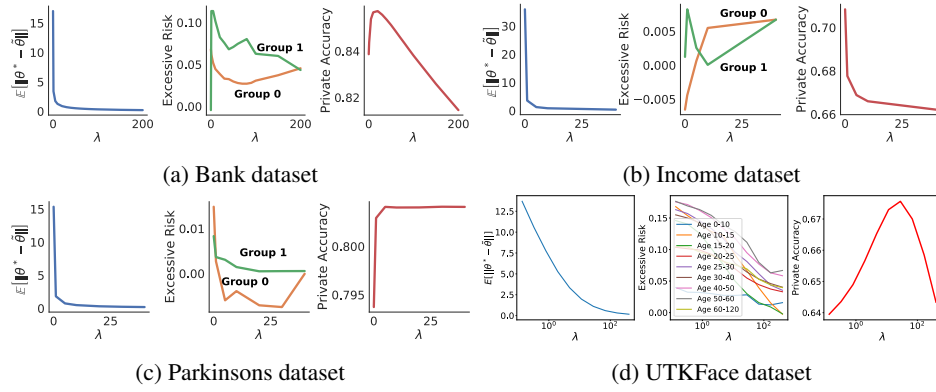


Figure 10: Expected model deviation (left), empirical risk (middle), and model accuracy (right) as a function of the regularization. The experiments are performed with the following settings:  $k = 150, \sigma = 50$ .

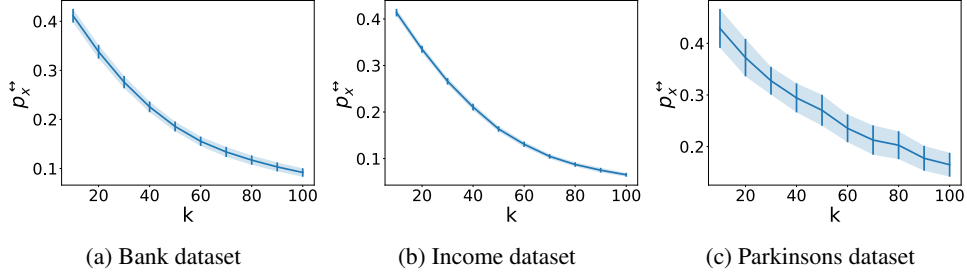


Figure 11: Average flipping probability  $p_x^*$  for samples  $x \in \bar{D}$  as a function of the ensemble size  $k$ .

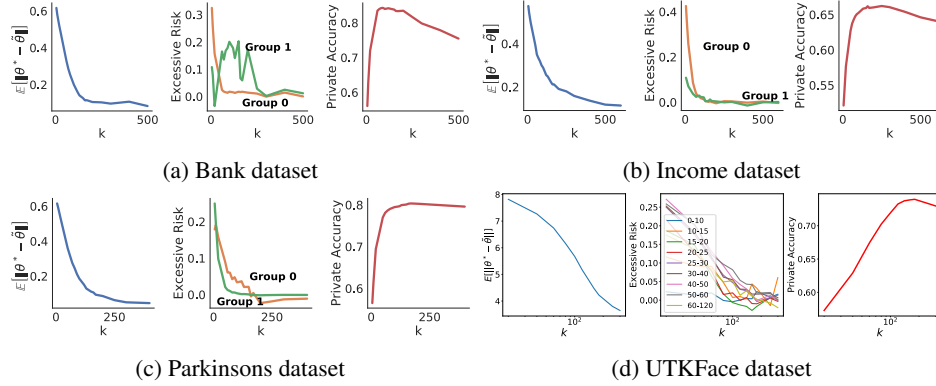


Figure 12: Expected model deviation (left), empirical risk (middle), and model accuracy (right) as a function of the ensemble size. The experiments are performed with the following settings:  $\lambda = 100, \sigma = 50$ .

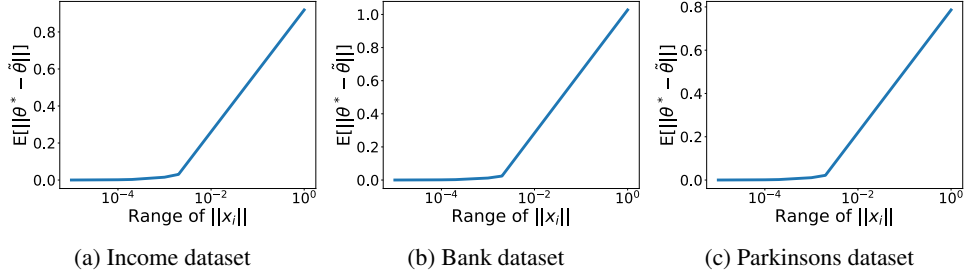
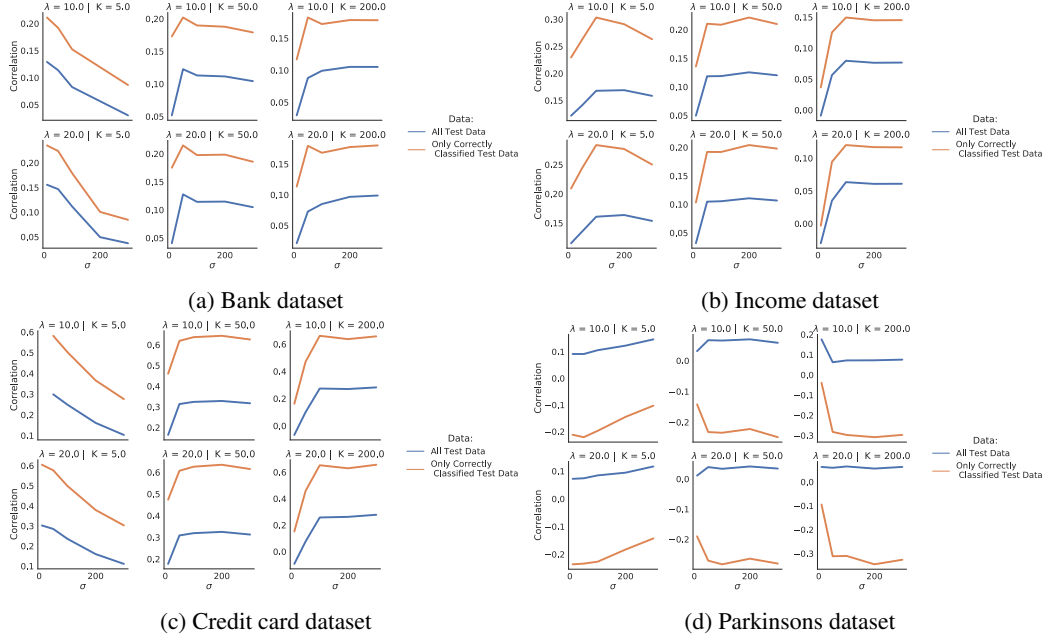


Figure 13: Relation between input norm and model deviation.

Figure 14: Correlation between the excessive risk and input norm on 5 datasets. The experiments are performed with the following settings:  $\lambda = 100$ ,  $\sigma = 50$ ,  $k = 150$ .

### C.5 THE IMPACT OF THE DATA INPUT NORM

This section provides further experimental results regarding relation between (1) the input norm with the private model deviation and (2) the input norm with its excessive risk.

Regarding the first relation, Corollary 1 from the main text implies that the smaller the input norm  $||x||$  is the smaller the model deviation is. For each dataset, we then vary the range of the input norm  $||x||$  and report the associated values of the expected model deviation in Figure 13. It can be seen clearly from the Figure 13, a monotone connection between input norm and the model deviation which verifies the statement from Corollary 1.

On the other hand, the input norm can affect the excessive risk by Lemma 1 of the Appendix, the individuals or group of individuals of large gradient norm can suffer from large excessive risk. In other words, the individuals of large data norm which are often observed at the tail of data can loose more accuracy. To confirm such claims, we report in Figure 14 the Spearman correlation between input norm and the excessive risk at individual levels. On all datasets, we can see obviously a positive relationship between data input norm and the excessive risk.



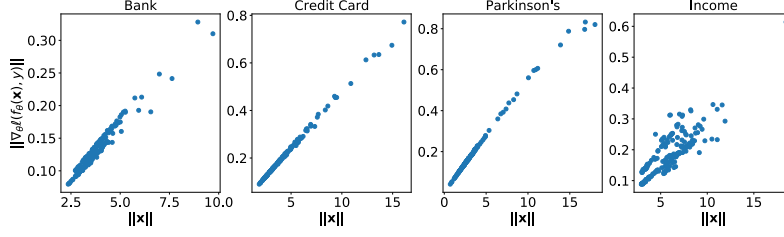


Figure 15: Relation Between Gradient Norm and Input Norm on all datasets.

### C.6 CONNECTION BETWEEN INPUT NORM AND SMOOTHNESS PARAMETER $\beta_a$

It is noted that the smoothness parameter  $\beta_a$  captures the local flatness of the loss function of a particular group  $a$ . Consider the logistic regression classifier, then the smoothness parameter  $\ell(f_\theta(x), y)$  for one particular data point is given by  $\beta_x = 0.25\|x\|$  [Shi et al. \(2021\)](#). Recall the following important property of the smooth function: If  $L = \sum_i \ell_i$  and each  $\ell_i$  is  $\beta_i$ -smooth then  $L$  is  $\max_i \beta_i$ -smooth. Because of that, the smoothness parameter  $\beta_a$  for one particular group  $a$  is given by:  $\beta_a = 0.25 \max_{x \in D_a} \|x\|$

The above clearly illustrates the relationship between input norms  $\|x\|$  and the smoothness parameters  $\beta_a$ .

### C.7 CONNECTION BETWEEN INPUT NORM AND GRADIENT NORM

In the main text, we have described, for logistic regression classifiers, there is a strong relation between the individual input norm  $\|x\|$  and their gradient norm at optimal parameter  $\|\nabla_{\theta^*} \ell(f_{\theta^*}(x), y)\|$ . In this subsection, we extend the analysis for non-linear model. In particular, we show a similar connection between the gradient norm and the input norm for a neural network with a single hidden layer. We start by considering the following settings:

**Settings** Consider a neural network model  $\tilde{f}_{\tilde{\theta}}(x) \stackrel{\text{def}}{=} \text{softmax}(\tilde{\theta}_1^T \tau(\tilde{\theta}_2^T x))$  where  $x = (x^i)_{i=1}^d$  is a  $d$  dimensional input vector, the parameters  $\tilde{\theta}_2 \in \mathbb{R}^{d \times H}$ ,  $\tilde{\theta}_1 \in \mathbb{R}^{H \times C}$  and the cross entropy loss  $\ell(f_{\tilde{\theta}}(x), y) = -\sum_{c=1}^C y_c \log \tilde{f}_{\tilde{\theta},c}(x)$  where  $\tau(\cdot)$  is a proper activation function, e.g., a sigmoid function. Let  $O = \tau(\tilde{\theta}_2^T x) \in \mathbb{R}^H$  be the vector  $(O_1, \dots, O_H)$  of  $H$  hidden nodes of the network. Denote the variables  $h_j = \sum_{i=1}^d \tilde{\theta}_{2,ji} x^i$  as the  $j$ -th hidden unit before the activation function. Next, denote  $\tilde{\theta}_{1,jk} \in \mathbb{R}$  as the weight parameter that connects the  $j$ -th hidden unit  $h_j$  with the  $c$ -th output unit  $\tilde{f}_c$  and  $\tilde{\theta}_{2,ij} \in \mathbb{R}$  as the weight parameter that connects the  $i$ -th input unit  $x^i$  with the  $j$ -th hidden unit  $h_j$ .

Given the settings above, we now show the dependency between gradient norm and input norm. First notice that we can decompose the gradients norm of this neural network into two layers as follows:

$$\|\nabla_{\tilde{\theta}} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\|^2 = \|\nabla_{\tilde{\theta}_1} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\|^2 + \|\nabla_{\tilde{\theta}_2} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\|^2. \quad (44)$$

We will show that  $\|\nabla_{\tilde{\theta}_2} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\| \propto \|x\|$ .

Notice that:

$$\|\nabla_{\tilde{\theta}_2} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\|^2 = \sum_{i,j} \|\nabla_{\tilde{\theta}_{2,ij}} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\|^2.$$

Applying, Equation (14) from [Sadowski \(2021\)](#), it follows that:

$$\nabla_{\tilde{\theta}_{2,ij}} \ell(\tilde{f}_{\tilde{\theta}}(x), y) = \sum_{c=1}^C (y_c - \tilde{f}_{\tilde{\theta},c}(x)) \tilde{\theta}_{1,jc} (O_j(1 - O_j)) x^i, \quad (45)$$

which highlights the dependency of the gradient norm  $\|\nabla_{\tilde{\theta}_2} \ell(\tilde{f}_{\tilde{\theta}}(x), y)\|$  and the input norm  $\|x\|$ . Figure [15](#) provides an empirical evidence for this dependency on all four datasets used in our analysis.

It can be seen clearly a strong positive correlation between input norm and the gradient norm at individual levels on all datasets from Figure 15.

#### C.8 EFFECTIVENESS OF MITIGATION SOLUTION

This subsection provides extended empirical results regarding the effectiveness of our proposed mitigation solution which was presented in Section 8.

We report the comparison between training PATE with hard and soft labels when  $k = 20$  in Figure 16 and when  $k = 150$  in Figure 17. These figures again illustrate the effects of the proposed mitigating solution in terms of utility/fairness tradeoff on the private student model. The top subplots of each figure show the group excessive risks  $R(\bar{D}_{\leftarrow 0})$  and  $R(\bar{D}_{\leftarrow 1})$  associated with two groups while the bottom subplots illustrate the accuracy of the model, at increasing of the privacy loss  $\epsilon$ . Recall that our mitigation solution does not require the availability of group labels during training. This challenging settings are of importance under the scenario when it is not feasible to collect or use protected features (e.g., under GDPR).

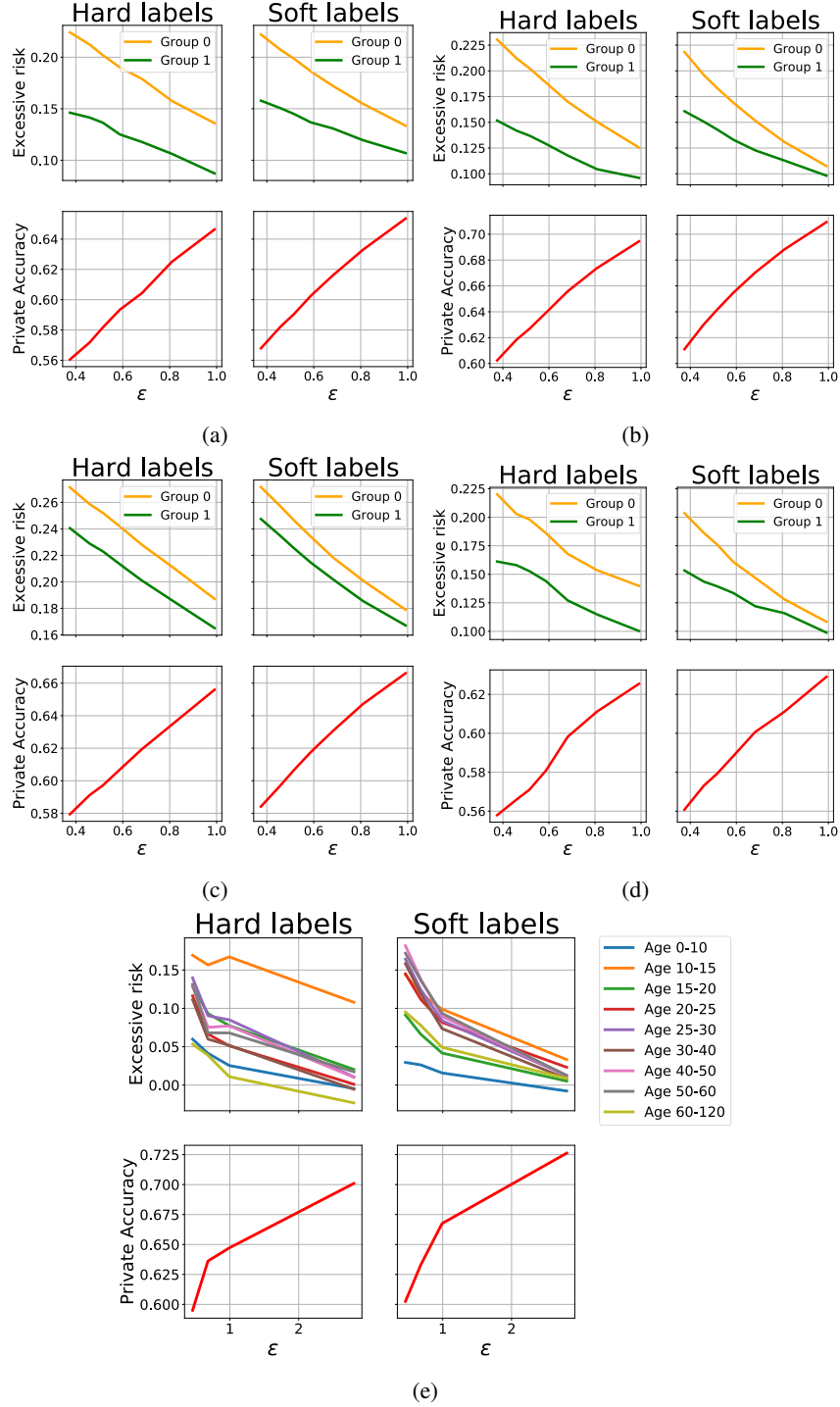


Figure 16: Comparison between training privately PATE with hard labels and soft labels in term of fairness (top subfigures) and utility (bottom subfigures) on (a) Bank, (b) Credit card, (c) Income (d) Parkinsons, (e) UTKFace dataset. Here for each dataset, the number of teachers  $k = 20$ .

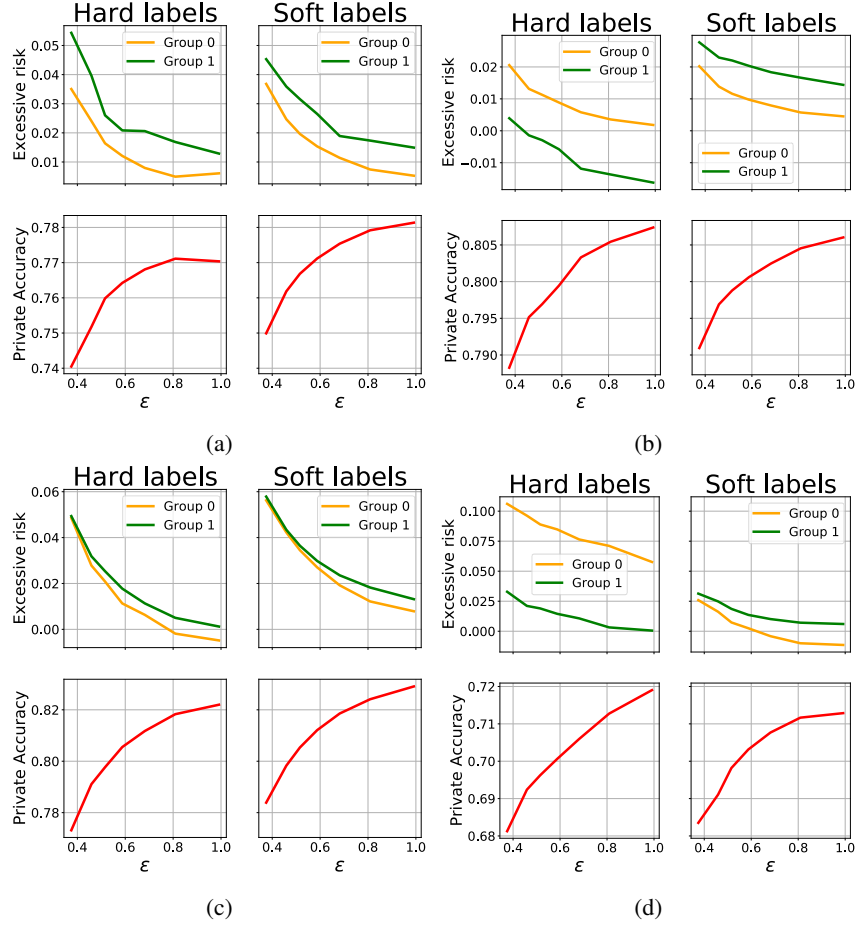


Figure 17: Comparison between training privately PATE with hard labels and soft labels in term of fairness (top subfigures) and utility (bottom subfigures) on (a) Bank, (b) Credit card, (c) Income, and (d) Parkinsons. Here for each dataset, the number of teachers  $k = 150$ .