# Label differential privacy via clustering

**Hossein Esfandiari** **Vahab Mirrokni** **Umar Syed** **Sergei Vassilvitskii**
Google Research

## Abstract

We present new mechanisms for *label differential privacy*, a relaxation of differentially private machine learning that only protects the privacy of the labels in the training set. Our mechanisms cluster the examples in the training set using their (non-private) feature vectors, randomly re-sample each label from examples in the same cluster, and output a training set with noisy labels as well as a modified version of the true loss function. We prove that when the clusters are both large and homogeneous, the model that minimizes the modified loss on the noisy training set converges to small excess risk at a rate that is comparable to the rate for non-private learning. We also describe a learning problem in which large clusters are necessary to achieve both strong privacy and either good precision or good recall. Our experiments show that randomizing the labels within each cluster significantly improves the privacy vs. accuracy trade-off compared to applying uniform randomized response to the labels, and also compared to learning a model via DP-SGD.

## 1 INTRODUCTION

The goal of *differentially private machine learning* is to train predictive models while preserving the privacy of user data in a training set. Most differentially private learning algorithms protect the privacy of every feature of every training example, and consequently inject so much noise into the learning process that they significantly underperform their non-private counterparts with respect to the utility of the learned model (see for instance the results on CIFAR-10 for DP-SGD (Abadi et al., 2016)). Differentially private learning algorithms also typically need full access to the private training data. These constraints can be a poor fit for many applications.

For example, consider a hospital that wants to use demographic data to train a diagnostic model for a rare illness. The input features to the model (such as a patient's age, sex, and race) may be far less sensitive than the label (whether the patient has the illness). Also, building an accurate predictive model is a hands-on, trial-and-error process that requires technically sophisticated data scientists, and the hospital is likely to achieve better results if it can share the training data with outside experts instead of having to keep all the data in-house.

*Label differential privacy*, introduced by Chaudhuri and Hsu (2011), relaxes the goal of differentially private machine learning so that only the privacy of the training labels is protected, since in many applications that is the only sensitive user attribute. In this paper, we propose differentially private mechanisms that add noise to the labels in a training set, and then output the noisy training set and a modified loss function, where the latter corrects for the noise added by the mechanism. A learner who wants to build a predictive model can use the output of our mechanism to freely experiment with modeling choices without observing any private user data.

Our approach is to use a variant of *randomized response* (Warner, 1965) to achieve label differential privacy. We cluster training examples according to their (non-private) features, and when randomizing an example's label, we choose the replacement label from the label distribution of the example's cluster instead of from the uniform distribution. We show that this improves the privacy vs. utility tradeoff for learning from the noisy training data when the clusters are large and the examples within a cluster have similar conditional label distributions, a prop-

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

erty we call *cluster homogeneity*. In particular, we show that an oracle that minimizes the modified loss function on the noisy training set outputs a model whose excess risk depends on the number of samples and the desired level of privacy, as well as the size and homogeneity of the clusters.

Our approach requires users to be able to privately sample from the label distribution of their example's cluster. We first describe a mechanism which uses a trusted server to perform the sampling and forwards the result to the learner. We also study a peer-to-peer setting where users are able to exchange messages with each other without a server's intervention. For this setting, we describe a distributed mechanism in which each user requests a noisy label from exactly one user in their cluster and then forwards that label to the learner. We prove that, from the learner's perspective, the privacy of this mechanism increases with the number of users per cluster.

**Our contributions:** We present our main results in Table 1.

- In Section 4 we describe a centralized cluster-based randomized response mechanism with excess risk at most $\tilde{O}\left(K\sqrt{\frac{d}{n}} + \frac{K^2\phi}{1+(e^\epsilon-1)\phi}\right)$, where $n$ is the size of the training set, $d$ is the dimension of hypothesis class, $\phi$ is the cluster heterogeneity (the inverse of homogeniety) and $K$ is the number of classes. Note that the privacy parameter $\epsilon$ appears in a separate term as the dimension $d$ in the excess risk bound, and thus the convergence rate of the dimension-dependent term *matches the optimal non-private convergence rate.* Also note that the dimension-free term is small if either $\epsilon$ is large *or* $\phi$ is small, and so there is no cost of privacy if the clustering is good enough.
- In Section 5 we describe a peer-to-peer cluster-based randomized response mechanism that satisfies $\left(\epsilon, \frac{1}{s^2}\right)$-label differential privacy and has excess risk $\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}} + \phi\right)$ for binary classification problems, where $s$ is the minimum cluster size. While this is worse than the best-known convergence rate, our mechanism only involves label flipping and empirical risk minimization, and is therefore significantly more practical than existing mechanisms that run in exponential time, and also does not require a trusted server.
- In Section 6 we present a hardness result relevant to multiclass classification and label differential privacy. Our hardness result suggests that a residual $\frac{K}{s\epsilon}$ term cannot be avoided even when the clustering is pure. To prove the hardness result we develop a probabilistic analysis method that bounds the performance of any differential privacy mechanism.
- Finally, in Section 7 we present experiments showing that our mechanisms can leverage a good clustering to improve the privacy vs. utility trade-off, outperforming both uniform randomized response and DP-SGD on real data.

## 2 RELATED WORK

There is an extensive literature on differentially private machine learning. The most common techniques include output and objective perturbation (Chaudhuri et al., 2011) and gradient perturbation (Abadi et al., 2016). In comparison, label differential privacy has received much less attention. Chaudhuri and Hsu (2011) introduced the concept and proved a lower bound on excess risk. Beimel et al. (2013) proved an upper bound for an inefficient mechanism, while Bassily et al. (2018) described the first efficient mechanism with a non-trivial excess risk bound. Their work is the most closely related to ours, since they use a PAC oracle as a black box to learn a model on a private training set. Most previous work relies on a trusted server to implement the differentially private mechanism, with the notable exception of (Wang and Xu, 2019), who studied sparse linear regression in the local model. We will say more about the connections between previous work and our contributions when presenting our results below.

Our work is also connected to several areas of research in non-private machine learning, including learning from label proportions (Quadrianto et al., 2009) and learning from noisy labels (Natarajan et al., 2013).

## 3 PRELIMINARIES

Let $\mathcal{X}$ be the *example space*. Let $\mathcal{Y}$ be the *label space*, with $K = |\mathcal{Y}| < \infty$. Let $D \in (\mathcal{X} \times \mathcal{Y})^n$ denote a *dataset* of $n$ labeled examples, with $(x_i, y_i)$ denoting the $i$th element of $D$.

For each $x \in \mathcal{X}$ let $c_x \in \mathcal{C}$ be the *cluster* of example $x$, where $\mathcal{C}$ is the set of all clusters and $C = |\mathcal{C}| < \infty$. In our setting, clusters are determined using unlabeled data (*i.e.*, unsupervised clustering), and since unlabeled data is not private and typically very abundant, all of our theoretical analysis will assume that the cluster $c_x$ of each example $x$ is given. Let $n_c(D) = |\{(x_i, y_i) \in D : c_{x_i} = c\}|$ be the size of cluster $c$ in dataset $D$.

| Algorithm | Excess risk | Comments |
|---|---|---|
| Optimal | $\tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ | Not private |
| Beimel *et al* (2013) | $\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}}\right)$ | Binary-labels only. Inefficient algorithm. |
| Bassily *et al* (2018) | $\tilde{O}\left(\frac{d^{3/5}}{(\epsilon n)^{2/5}}\right)$ | Binary labels only. |
| Our centralized mechanism | $\tilde{O}\left(K\sqrt{\frac{d}{n}} + \frac{K^2\phi}{1+(e^\epsilon-1)\phi}\right)$ | Cluster size $s \geq \frac{1}{\phi\epsilon}$ |
| Our peer-to-peer mechanism | $\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}} + \phi\right)$ | Binary labels only, cluster size $s \geq \frac{1}{\phi\epsilon^2}$ $(\epsilon, \delta)$-DP with $\delta = \frac{1}{s^2}$ |

Table 1: Summary of our results. Let $n$ denote the size of the training set, $d$ the dimension of hypothesis class, $K$ the number of classes, $s$ the minimum cluster size, and $\phi$ the cluster heterogeneity.

Let $\mathcal{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Let $(X, Y) \sim \mathcal{P}$ denote that $(X, Y)$ is drawn from $\mathcal{P}$ and let $X \sim \mathcal{P}_{\mathcal{X}}$ denote that $X$ is drawn from the marginal distribution of $\mathcal{P}$ on $\mathcal{X}$. We write $D \sim \mathcal{P}^n$ to indicate that dataset $D$ contains $n$ labeled examples each drawn independently from $\mathcal{P}$. Let

$$p(y|x) = \Pr_{(X,Y)\sim\mathcal{P}}[Y = y \mid X = x]$$
$$\hat{p}_{y|x}(D) = \frac{|\{(x_i, y_i) \in D : x_i = x \wedge y_i = y\}|}{|\{(x_i, y_i) \in D : x_i = x\}|}$$

denote true conditional probability and empirical conditional probability, respectively, of label $y \in \mathcal{Y}$ for example $x \in \mathcal{X}$. With a slight abuse of notation, let

$$p(y|c) = \Pr_{(X,Y)\sim\mathcal{P}}[Y = y \mid c_X = c]$$
$$\hat{p}_{y|c}(D) = \frac{|\{(x_i, y_i) \in D : c_{x_i} = c \wedge y_i = y\}|}{|\{(x_i, y_i) \in D : c_{x_i} = c\}|}$$

denote true conditional probability and empirical conditional probability, respectively, of label $y \in \mathcal{Y}$ in cluster $c \in \mathcal{C}$.

We write $\mathbf{q}$ to denote arbitrary *cluster label distributions*, where $q(y|c)$ is the conditional probability of label $y \in \mathcal{Y}$ in cluster $c \in \mathcal{C}$ according to $\mathbf{q}$.

A pair of datasets $D, D' \in (\mathcal{X} \times \mathcal{Y})^n$ are *label neighbors* if they contain exactly the same labeled examples except that one example's label may differ between $D$ and $D'$. A *mechanism* $M : (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{O}$ is a randomized algorithm that takes as input a dataset and outputs into some set $\mathcal{O}$. Mechanism $M$ satisifies $(\epsilon, \delta)$-*label differential privacy* if for all datasets $D, D'$ that are label neighbors and all subsets $O \subseteq \mathcal{O}$ we have $\Pr[M(D) \in O] \leq e^\epsilon \Pr[M(D') \in O] + \delta$, where the probability is with respect the internal randomization of $M$. Let $\epsilon$-label differential privacy be an abbreviation for $(\epsilon, 0)$-label differential privacy.

Let $\mathcal{H}$ be a *hypothesis class* containing functions with domain $\mathcal{X}$. Let $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \mapsto [0, 1]$ be a *loss function* that maps each hypothesis and labeled example to a non-negative loss value. Let $R(h) = E_{(X,Y)\sim\mathcal{P}}[\ell(h, X, Y)]$ be the *risk* of hypothesis $h \in \mathcal{H}$ with respect to loss function $\ell$. We call $R(h) - \inf_{h\in\mathcal{H}} R(h)$ the *excess risk* of $h$.

Define $\dim(\mathcal{H}, \ell)$ to be the *dimension* of loss function $\ell$ and hypothesis class $\mathcal{H}$: $\dim(\mathcal{H}, \ell) = \frac{\log N_\alpha(\mathcal{F})}{\log(1/\alpha)}$, where $N_\alpha(\mathcal{F})$ is the $\alpha$-covering number of the function class $\mathcal{F} = \{(x, y) \mapsto \ell(h, x, y) : h \in \mathcal{H}\}$. We use covering number as our definition of dimension mostly for convenience, as it applies to any real-valued loss function and simplifies comparisons to previous work. For example, it is known (Mohri et al., 2018) that if $\ell$ is boolean-valued (say $\ell(h, x, y) = \mathbf{1}\{h(x) \neq y\}$ is the zero-one loss) then $\dim(\mathcal{H}, \ell)$ is at most the VC dimension of $\mathcal{H}$ (up to a constant factor), which permits a direct comparison with Beimel et al. (2013) and Bassily et al. (2018). We could substitute another learning-theoretic notion of the complexity of a hypothesis class (such as pseudodimension) without significantly affecting our results.

## 4 CENTRALIZED MECHANISM

In this setting, the dataset is stored by the curator, who applies a privacy mechanism to the dataset and outputs a dataset with noisy labels, as well as a modified loss function.

The centralized mechanism (Algorithm 1) adds noise to the labels as follows: (1) Compute the empirical label distribution in each cluster. (2) Add noise drawn from Laplace$(\sigma/n_c(D))$ to each label probability in each cluster $c$. (3) Truncate the per-cluster label probabilities so they are each in the interval $[\tau, 1]$.

(4) Renormalize the per-cluster label probabilities so that they form distributions. (5) With probability $\lambda$, replace each label with a random label drawn from the example's cluster label distribution.

The modified loss function output by the centralized mechanism reduces the bias that was introduced by adding noise to the labels. The modified loss is constructed by re-weighting the original loss using matrix inverses that essentially 'undo' the randomization of the labels. Setting the bias correction parameter $\beta = \lambda$ in Algorithm 1 completely removes the effect of this randomization, in expectation (see Corollary 1). Natarajan et al. (2013) developed this debiasing technique for the special case of binary labels, which we generalize to $K > 2$ labels.

In the supplement we show that the renormalization procedure in Algorithm 1 keeps each per-cluster label probability above threshold $\tau$, which is key to proving the following privacy guarantee.

**Theorem 1** (Centralized privacy)**.** *The centralized mechanism (Algorithm 1) satisfies $\epsilon$-label differential privacy with $\epsilon = \frac{1}{\sigma} + \log\left(1 + \frac{1-\lambda}{\lambda\tau}\right)$.*

Given a noisy dataset $\tilde{D}$ and a modified loss function $\tilde{\ell}$ output by Algorithm 1, our goal is to upper bound the excess risk (also called the *generalization error*) of the hypothesis $\tilde{h}$ that minimizes the average of $\tilde{\ell}$ on $\tilde{D}$. A key benefit of such a guarantee is that it is agnostic to the internal operation of the learning algorithm, and thus applies to any algorithm for empirical risk minimization.

The minimum excess risk we can achieve, and the rate at which we approach that excess risk, will depend on both the size and quality of the clusters. We measure the quality of the clusters in terms of their *heterogeneity*.

**Definition 1** (Cluster heterogeneity)**.** *Let $\phi = \mathrm{E}_{X\sim\mathcal{P}_\mathcal{X}}\left[\sum_y |p(y|X) - p(y|c_X)|\right]$ be the average total variation distance between the conditional label distribution of an example and its cluster.*

If clusters have low heterogeneity then it should be easier to add privatizing noise to the labels without impacting utility because, intuitively, one can swap labels among examples in the same cluster without badly distorting the original data distribution. Our analysis confirms this intuition.

**Theorem 2** (Centralized utility)**.** *Let $\tilde{D}$ and $\tilde{\ell}$ be the dataset and loss function output by the centralized mechanism (Algorithm 1) with threshold $\tau$, noise scale $\sigma$, resampling probability $\lambda$, and bias correction $\beta$, and dataset $D$ as input, and assume*

**Algorithm 1** Centralized mechanism

**Parameters:** Threshold $\tau \in \left[0, \frac{1}{K}\right]$; noise scale $\sigma \geq 0$; label resampling probability $\lambda \in [0, 1)$; bias correction parameter $\beta \in [0, 1)$.
**Input:** Dataset $D = ((x_1, y_1), \ldots, (x_n, y_n))$ of labeled examples.

*// Add noise to cluster label distributions.*
**for** $c \in \mathcal{C}$ **do**
  *// Add noise to each empirical probability and clip.*
  **for** $y \in \mathcal{Y}$ **do**
    $q(y|c) \leftarrow \max\left\{\tau, \min\left\{1, \hat{p}_{y|c}(D) + z_{y,c}\right\}\right\}$,
      where $z_{y,c} \sim \text{Laplace}\left(\frac{\sigma}{n_c(D)}\right)$.
  **end for**

  *// Renormalize distribution.*
  $\Delta_c \leftarrow 1 - \sum_y q(y|c)$
  **for** $y \in \mathcal{Y}$ **do**
    **if** $\Delta_c < 0$ **then**
      $\xi_{y,c} \leftarrow q(y|c) - \tau$
    **else**
      $\xi_{y,c} \leftarrow 1 - q(y|c)$
    **end if**
  **end for**
  **for** $y \in \mathcal{Y}$ **do**
    $\tilde{q}(y|c) \leftarrow q(y|c) + \frac{\xi_{y,c}}{\sum_{y'} \xi_{y',c}} \Delta_c$
  **end for**
**end for**

*// Randomize labels.*
**for** $(x_i, y_i) \in D$ **do**
  $\mathring{y}_i \leftarrow y$ with probability $\tilde{q}(y|c_{x_i})$.
  $\tilde{y}_i \leftarrow \begin{cases} y_i & \text{with probability } 1 - \lambda \\ \mathring{y}_i & \text{with probability } \lambda \end{cases}$
**end for**

*// Construct noisy dataset.*
$\tilde{D} \leftarrow ((x_1, \tilde{y}_1), \ldots, (x_n, \tilde{y}_n))$.

*// Construct modified loss function.*
For each $x \in \mathcal{X}$ let $\tilde{\mathbf{Q}}_{x,\beta} \in \mathbb{R}^{K\times K}$ be the label randomization matrix defined by

$$\tilde{Q}_{x,\beta}[y', y] = (1 - \beta)\mathbf{1}\{y' = y\} + \beta\tilde{q}(y'|c_x).$$

Define the loss function $\tilde{\ell} : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ as

$$\tilde{\ell}(h, x, y) = \sum_{y'} \tilde{Q}^{-1}_{x,\beta}[y', y]\ell(h, x, y').$$

**return** Dataset $\tilde{D}$ and loss function $\tilde{\ell}$.

*each cluster in $D$ has size at least $s$. Let $\tilde{h} = \arg\min_{h\in\mathcal{H}} \sum_{(x,y)\in\tilde{D}} \tilde{\ell}(h,x,y)$ be the hypothesis in $\mathcal{H}$ that minimizes $\tilde{\ell}$ over $\tilde{D}$. Then with probability $1-\gamma$ over the choice of $D \sim \mathcal{P}^n$*

$$E[R(\tilde{h})] - \inf_{h\in\mathcal{H}} R(h) \leq \frac{CK}{1-\beta}\sqrt{\frac{\dim(\mathcal{H},\ell)\log\frac{1}{\gamma}}{n}} + \frac{CK|\beta-\lambda|}{1-\beta}\left(\phi + \frac{K\sigma}{s} + K\tau\right)$$

*where $C > 0$ is a universal constant and the expectation is with respect to the Laplace random variables (the $z_{y,c}$'s) in Algorithm 1.*

*Proof sketch.* The complete proof is lengthy and is deferred to the supplementary material. Here we present a sketch of the main ideas.

Given access to the original dataset $D$, it is well-known that the excess risk of the empirical risk minimizer is upper bounded with high-probability by $\tilde{O}\left(L\sqrt{\frac{d}{n}}\right)$, where $\max_{h,x,y}|\ell(h,x,y)| \leq L$. We show that the modified loss function $\tilde{\ell}$ satisfies

$$\max_{h,x,y}|\tilde{\ell}(h,x,y)| \leq \frac{K}{1-\beta}, \tag{1}$$

also show that when $\beta = \lambda$ we have

$$\mathrm{E}\left[\sum_{(x,y)\in\tilde{D}} \tilde{\ell}(h,x,y)\right] = \sum_{(x,y)\in D} \ell(h,x,y)$$

for every hypothesis $h$. In other words, the modified loss on the noisy dataset is an unbiased estimate of the true loss on the original dataset. Combining these results enables us to prove that $\tilde{h}$ converges to zero excess risk as $n \to \infty$ if $\beta = \lambda$, at the cost of increasing the convergence rate by a $O(\frac{K}{1-\beta})$ factor compared to non-private learning.

If $\beta \neq \lambda$ then the modified loss function does not completely remove the bias introduced by adding noise to the labels. In this case we prove that the residual excess risk as $n \to \infty$ is upper bounded by

$$|\tilde{\ell}(h,x,y)| \cdot |\beta-\lambda| \cdot (\phi+\psi)$$

where $\phi$ is the cluster heterogeneity from Definition 1 and $\psi$ is a measure of cluster distortion. The proof is completed by using Eq. (1) to bound $|\tilde{\ell}(h,x,y)|$ and also by showing

$$\psi \leq O\left(\frac{K\sigma}{s} + K\tau\right). \qquad \square$$

### 4.1 Discussion

Taken together, Theorems 1 and 2 specify a three-way trade-off between privacy, excess risk and convergence rate. The first term in the upper bound in Theorem 2 is asymptotically zero as $n \to \infty$ and determines the convergence rate, while the remaining terms are asymptotically non-zero when $\beta \neq \lambda$ and represent the residual excess risk when $n \to \infty$. Thus the bias correction parameter $\beta$ of Algorithm 1 trades-off between excess risk and convergence rate, while the label resampling probability $\lambda$, the noise scale $\sigma$, and the threshold $\tau$ trade-off between excess risk and privacy.

To illustrate these trade-offs, we consider some special cases of Theorems 1 and 2, starting with a setting of the parameters in Algorithm 1 that reduces the centralized mechanism to uniform randomized response on the labels (which can of course be implemented as a local mechanism).

**Corollary 1** (Uniform randomized response)**.** *If $\epsilon > 0$, $\tau = \frac{1}{K}$, $\beta = \lambda = \frac{K}{K-1+e^{\epsilon}}$ and $\sigma = \infty$ then the centralized mechanism (Algorithm 1) replaces each label with a uniform random label and satisfies $\epsilon$-label differential privacy. If in addition $\epsilon < 1$ and $D \sim \mathcal{P}^n$ then with probability $1-\gamma$ over the choice of $D$ and the randomness in the mechanism the hypothesis $\tilde{h}$ from Theorem 2 satisfies*

$$R(\tilde{h}) - \inf_{h\in\mathcal{H}} R(h) = O\left(\frac{K}{\epsilon}\sqrt{\frac{\dim(\mathcal{H},\ell)\log\frac{1}{\gamma}}{n}}\right)$$

Despite its extreme simplicity, to the best of our knowledge the excess risk of uniform randomized response for label differential privacy has not previously been analyzed. For binary classification (*i.e.*, $K = 2$) we know that $\dim(\mathcal{H},\ell) = O(d)$, where $d$ is the VC dimension of hypothesis class $\mathcal{H}$, and thus the excess risk converges asymptotically to zero at a rate $\tilde{O}\left(\frac{1}{\epsilon}\sqrt{\frac{d}{n}}\right)$. By comparison, the convergence rate of the mechanism due to Beimel et al. (2013) is $\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}}\right)$. However, their mechanism is significantly less practical than empirical risk minimization, as it involves running the exponential mechanism on $\Omega(2^d)$ hypotheses. Bassily et al. (2018) give an efficient algorithm that obtains a rate of $\tilde{O}\left(\frac{d^{3/5}}{(\epsilon n)^{2/5}}\right)$, which is a worse dependence on both $d$ and $n$. Also, previous work was limited to binary classification, while our analysis applies to multi-class classification.

We now show that the convergence rate can be sig-

nificantly improved when the clusters are both large and have low heterogeneity.

**Corollary 2** (Cluster-based randomized response)**.** *If $\epsilon > 0$, $\tau = \phi$, $\beta = 0$, $\lambda = \frac{1}{1+(e^\epsilon-1)\phi}$ and $\sigma = \frac{1}{\epsilon}$ then the centralized mechanism (Algorithm 1) satisfies $O(\epsilon)$-label differential privacy. If in addition each cluster has size at least $s \geq \frac{1}{\epsilon\phi}$ and $D \sim \mathcal{P}^n$ then with probability $1-\gamma$ over the choice of $D$ the hypothesis $\tilde{h}$ from Theorem 2 satisfies*

$$E[R(\tilde{h})] - \inf_{h\in\mathcal{H}} R(h) \leq O\left(K\sqrt{\frac{\dim(\mathcal{H},\ell)\log\frac{1}{\gamma}}{n}} + \frac{K^2\phi}{1+(e^\epsilon-1)\phi}\right)$$

*where the expectation is with respect to the Laplace random variables (the $z_{y,c}$'s) in Algorithm 1.*

If we let $K = 2$ then the dimension-dependent term in the convergence rate in Corollary 2 is $\tilde{O}\left(\sqrt{\frac{d}{n}}\right)$, where $d$ is the VC dimension of hypothesis class $\mathcal{H}$, and this is the optimal rate for non-private learning. However, instead of converging to zero, the excess risk converges to $O\left(\frac{\phi}{1+(e^\epsilon-1)\phi}\right)$ when the minimum cluster size $s \geq \frac{1}{\epsilon\phi}$. Note that this residual excess risk is small when the privacy parameter $\epsilon$ is large *or* the cluster heterogeneity $\phi$ (see Definition 1) is small. Thus there is not necessarily any cost of privacy if the clustering is good enough.

## 5 PEER-TO-PEER MECHANISM

In the peer-to-peer setting, the dataset is stored in a distributed manner, with each user $i$ storing labeled example $(x_i, y_i)$. Instead of communicating with a central curator, each user sends and receives messages directly to other users. In this section we assume the labels are binary, so that each $y_i \in \{0, 1\}$.

In the peer-to-peer mechanism (Algorithm 2), each user $i$ first adds noise to her own label, and then replaces her label with the noisy label of a user $j$. User $j$ is selected uniformly at random from among all the users in user $i$'s cluster. Note that this means we may have $i = j$, and also that user $j$ may be selected by other users besides user $i$. In other words, the mechanism is based on resampling, not permuting, the labels within a cluster.

An alternative approach would be for users to communicate with a server that randomly permutes the labels within each cluster before forwarding the data to the learner. We could analyze such a mechanism via the technique of privacy amplification by shuffling (Cheu et al., 2019; Erlingsson et al., 2019). But this approach would require a shuffling server that is trusted by all users.

**Theorem 3** (Peer-to-peer privacy)**.** *There exists a constant $C > 0$ such that if each cluster in $D$ has size at least $s$ and $\alpha = \frac{C\log s}{\sqrt{\theta s}}$ then the peer-to-peer mechanism (Algorithm 2) satisfies $(\epsilon, \delta)$-label differential privacy with*

$$\epsilon \leq O\left(\theta + \frac{\theta^{3/2}}{\sqrt{s}\log s} + \frac{\theta^{3/4}}{s^{1/4}}\right) \text{ and } \delta \leq \frac{1}{s^2}.$$

**Algorithm 2** Peer-to-peer mechanism

**Parameters:** Label flipping probability $\alpha \in [0, 1]$; subsampling rate $\theta \in [0, 1]$
**Assume:** Label set $\mathcal{Y} = \{0, 1\}$.
**Input:** Dataset $D = ((x_1, y_1), \ldots, (x_n, y_n))$, where each labeled example $(x_i, y_i)$ is stored by user $i$.

**for** user $i$ **do**
  *// Add noise to own label.*
  $\mathring{y}_i \leftarrow \begin{cases} y_i & \text{with probability } 1-\alpha \\ 1-y_i & \text{with probability } \alpha \end{cases}$
  *// Select a random user in the same cluster.*
  Select user $j$ uniformly at random from the set $\{j' : c_{x_{j'}} = c_{x_i}\}$.
  *// Replace own label with other user's noisy label.*
  $\tilde{y}_i \leftarrow \mathring{y}_j$
  *// Subsample.*
  Add $i$ to $I$ with probability $\theta$.
**end for**
*// Construct noisy dataset.*
$\tilde{D} \leftarrow ((x_{i_1}, \tilde{y}_{i_1}), \ldots, (x_{i_m}, \tilde{y}_{i_m}))$, where each $i_j \in I$.
**return** Dataset $\tilde{D}$.

While the centralized mechanism outputs a modified loss function that corrects for the bias introduced by adding noise to the labels, the peer-to-peer mechanism does not output a modified loss function, since there is no single party with knowledge of how the labels were randomized. As a result, our upper bound on excess risk (Theorem 4) does not converge asymptotically to zero, although it does converge to small excess risk when the clusters have low heterogeneity.

**Theorem 4** (Peer-to-peer utility)**.** *Let $\tilde{D}$ be the dataset output by the peer-to-peer mechanism (Algorithm 2) when given dataset $D$ as input. Let $\tilde{h} = \arg\min_{h\in\mathcal{H}} \sum_{(x,y)\in\tilde{D}} \ell(h,x,y)$ be the hypothesis in $\mathcal{H}$ that minimizes the true loss function $\ell$ over $\tilde{D}$. Then with probability $1-\gamma$ over the choice of $D \sim \mathcal{P}^n$ and the randomness in the mechanism*

$$R(\tilde{h}) - \inf_{h\in\mathcal{H}} R(h) \leq O\left(\sqrt{\frac{\dim(\mathcal{H},\ell)\log\frac{1}{\gamma}}{\theta n}} + \phi + \alpha\right)$$

Combining Theorems 3 and 4 shows that if each cluster has minimum size $s \geq \frac{1}{\phi\epsilon^2}$ and $\epsilon < 1$ then the peer-to-peer mechanism satisfies $(\epsilon, \frac{1}{s^2})$-label differential privacy and has excess risk $\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}} + \phi\right)$. This is worse than the $\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}}\right)$ convergence rate obtained by Beimel et al. (2013), but our peer-to-peer mechanism is significantly more practical, since it only consists of label flipping and empirical risk minimization, instead of requiring the exponential mechanism to be run on $\Omega(2^d)$ hypotheses. Our mechanism also does not require a central curator.

### 5.1 Comparison to the shuffle model

The shuffle model (Cheu et al., 2019; Erlingsson et al., 2019) involves (at least) two servers: a curator and a shuffler. Typically, each user applies a local randomizer to her data, encrypts the noisy data using the curator's public key, and sends the encrypted data to the shuffler. The shuffler strips identifiers from the messages it receives and randomly permutes them, then forwards the messages to the curator, who decrypts them.

The major benefit of the shuffle model is that the privacy provided by the local randomizers is amplified by the shuffling procedure and increases with the number of users. However, if the curator and shuffler collude with one another, then this privacy amplification property is invalidated. In real-world implementations of the shuffle model (*e.g.*, RAPPOR (Erlingsson et al., 2014)) both servers are operated by the same entity thus limiting the privacy benefits.

By contrast, in our peer-to-peer model, each user receives an unencryted message from exactly one other user, and the learner need not be trusted by any user for privacy amplification to be achieved.

Of course, the peer-to-peer model has its own limitations. Unlike in the shuffle model, we have not shown a privacy amplification result that applies to any local randomizer, but only to simple label flipping. Also, each user observes the noisy label of another user, and the privacy of this label is not amplified. Indeed, it is straightforward to show that, from the perspective of each user, Algorithm 2 only satisfies $\log(\frac{1-\alpha}{\alpha})$-label differential privacy, as well as only $(0, \frac{1}{s})$-label differential privacy. However, the amount of data observed by any single user is minuscule (*i.e.*, a single bit).

One could implement shuffling in our peer-to-peer model by having all users in each cluster agree on a random permutation of the users, and then have each user request the noisy label of the user they are mapped to by the permutation. However, agreeing on a random permutation (say, by agreeing on a pseudorandom seed) would itself require a cryptographic protocol (such as key-agreement protocol (Merkle, 1978)), since the permutation must be kept secret from the learner.

Since Algorithm 2 involves a subsampling step, it is tempting to ask whether we could achieve privacy amplification in the peer-to-peer model by label flipping and subsampling alone, without exchanging messages among users. It is straightforward to see that this will not work. Since only the labels of the dataset are private, any subsampling applies to the labels only, so the learner can always construct a complete dataset in which some of the labels are replaced with $\perp$, indicating that a label was not provided by the user. So a mechanism in which some users drop their label, but do not communicate with other users, is equivalent to randomized response on the set $\{0, 1, \perp\}$. Essentially, amplification by subsampling is only effective when users can completely remove themselves from the dataset, but this isn't possible when only the users' labels are private.

## 6 LOWER BOUND

Note that, in Theorem 2, even if we assume that we have perfect clusters (i.e. $\phi = 0$) of size $s$, then by setting $\beta = 0$, $\lambda = 1$, $\tau = 0$, and $\sigma = \frac{1}{\epsilon}$, we have an $\epsilon$-label differentially private mechanism with an excess risk of $\tilde{O}\left(\sqrt{\frac{d}{n}} + \frac{K}{s\epsilon}\right)$. In other words, we have the optimal non-private convergence rate plus a residual term $\frac{K}{s\epsilon}$, which is $\Theta(1)$ for $s = K$. In this section we motivate this relationship between the size of clusters and the number of labels. We fix a basic learning task and show that, for any constant $\epsilon$ it is not possible to learn a nontrivial $\epsilon$-label differentially private model, when the size of high quality clusters are small. In

fact, our result holds in a simpler yet relevant setting where we have access to the whole label distribution statically. This motivates the necessity of having large high quality clusters in our positive result when the number of labels is large.

We first define our learning task.

**Definition 2.** *Label Association Problem (LAP):* ***Setup:*** *We have a dataset $D \in (\mathcal{X} \times \mathcal{Y})^n$, where each example $x$ appears with only one label $y$. $\mathcal{C}$ is a partitioning of the data in $D$ and size of each cluster $c \in \mathcal{C}$ is exactly $s$.* ***Task:*** *For each cluster $c \in \mathcal{C}$ we intend to learn the set of labels that are associated with examples in $c$, denoted as $\mathcal{Y}_c = \{y \mid \exists x \in \mathcal{X} \text{ s.t. } c_x = c \wedge (x, y) \in D\}$.*

Let $M$ be a label differentially private mechanism for LAP. $\tilde{D} = M(D)$ is a set of pairs $(c, y)$. We interpret $\tilde{D}$ as a binary classification, where the input is $(c, y)$ and the output is 1 if $(c, y) \in \tilde{D}$. We use precision and recall defined as follows to measure the accuracy of model $\tilde{D}$. We have

$$\text{Precision} = \frac{\sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}_c} \mathrm{E}_M[\tilde{D}(c, y)]}{\mathrm{E}_M[|\tilde{D}|]}$$

$$\text{Recall} = \frac{\sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}_c} \mathrm{E}_M[\tilde{D}(c, y)]}{\sum_{c \in \mathcal{C}} |\mathcal{Y}_c|},$$

where the expectations are over the randomness of the mechanism $M$. Note that without differential privacy, this problem can be learned with precision 1 and recall 1.

The next theorem states our main hardness result, and the proof is in the supplementary material. Our proof defines a randomized process that generates two neighboring datasets $D$ and $D'$. Then we fix an arbitrary $\epsilon$-label differentially private mechanism $M$ and show that if $\epsilon$ is a constant either recall of $M$ on $D'$ is sub-constant or precision of $M$ on $D$ is sub-constant. We do this by analysing the probability that, the label in $D'$ that is not in $D$, is preserved by mechanism $M$. If such probability is small then the the recall of $M(D')$ is small, if it is large the precision of $M(D)$ is small.

**Theorem 5.** *When $s \leq o(K)$, it is impossible to have an $\epsilon$-differential privacy mechanism $M$ for LAP with a constant $\epsilon$, that guarantees a constant precision and a constant recall.*

# 7 EXPERIMENTS

We evaluated the following mechanisms on the MNIST (LeCun and Cortes, 2010), Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky, 2009) datasets:

- `UniformRR`: Algorithm 1 with parameters set according to Corollary 1.
- `ClusterRR`: Algorithm 1 with parameters set according to Corollary 2.
- `DP-SGD`: Differentially-private variant of SGD (Abadi et al., 2016).

For the `ClusterRR` mechanism we learned 100 clusters on each unlabeled training set using the `sklearn.cluster.KMeans` package (with default parameters). For both randomized response mechanisms we used the `sklearn.linear_model.LogisticRegression` package (set to 'multinomial' and using the 'SAGA' solver) to learn a classifer on the noisy training set output by the mechanism. For `DP-SGD` we learned a logistic regression model by adapting the implementation from the TensorFlow Privacy library (TFP, 2019). We varied the noise added to the gradients, and for each noise level computed $\epsilon$ using the privacy-by-iteration method (Feldman et al., 2018) with $\delta = 1/n$, where $n$ is the training set size.

For each mechanism and each dataset we evaluated the learned classifer's accuracy on the test set. See the first three panels of Figure 1 for results, where each data point is the average of 5 trials, and each y-axis is normalized, *i.e.*, divided by the accuracy of the non-private classifier that is learned on the original training set. Observe that `ClusterRR` outperforms both `UniformRR` and `DP-SGD` on each dataset for a wide range of the privacy parameter $\epsilon$.

We also assessed the importance of a good clustering for `ClusterRR` by fixing the privacy parameter $\epsilon = 0.5$ and varying the number of clusters. See the bottom panel of Figure 1, which shows that the performance of `ClusterRR` degrades sharply when there are too few clusters (since the clusters are too heterogeneous) or too many (since the clusters are too small). This empirical finding matches our theoretical analysis (see Theorem 2).

## 7.1 Comparsion to `LP-2ST` mechanism

We also compared `ClusterRR` to `LP-2ST`, a mechanism proposed in recent and concurrent work by Ghazi et al. (2021). Similar to `ClusterRR`, the `LP-2ST` mechanism generates a label-private training set by using a variant of randomized response. The mechanism proceeds in two stages. In the first stage,

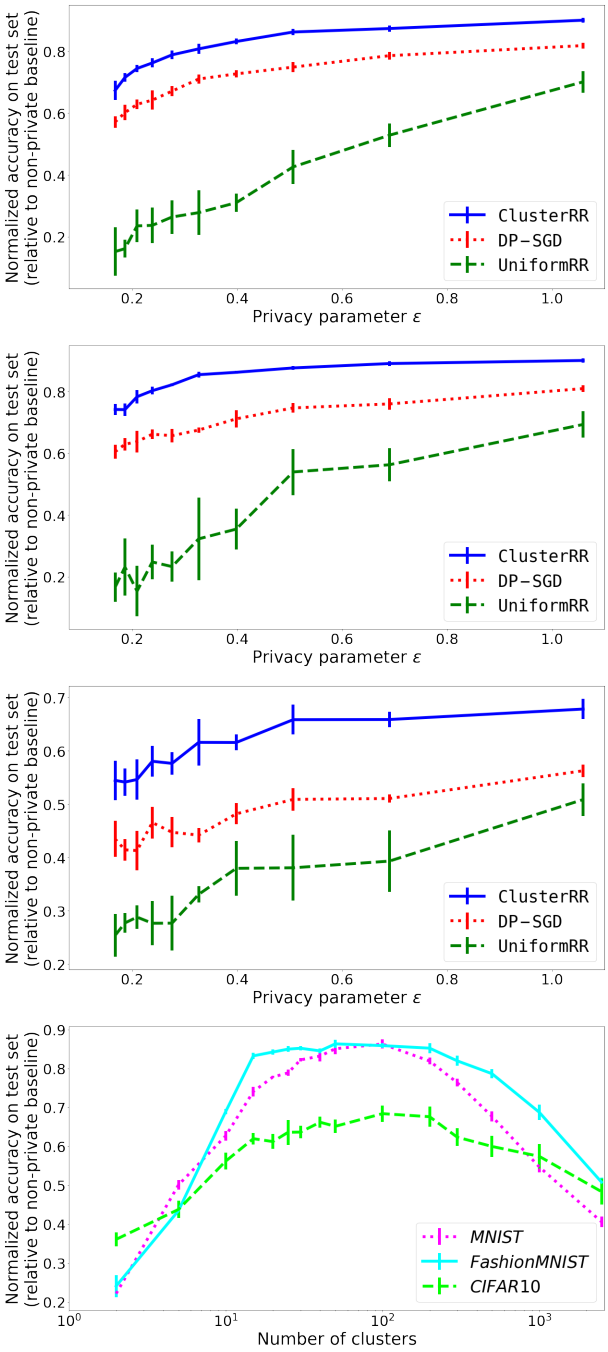


Figure 1: Performace of each mechanism on the MNIST (*first panel*), Fashion-MNIST (*second panel*) and CIFAR-10 (*third panel*) datasets. Performance of the `ClusterRR` mechanism (for $\epsilon = 0.5$) on each dataset when varying the number of clusters (*bottom panel*).

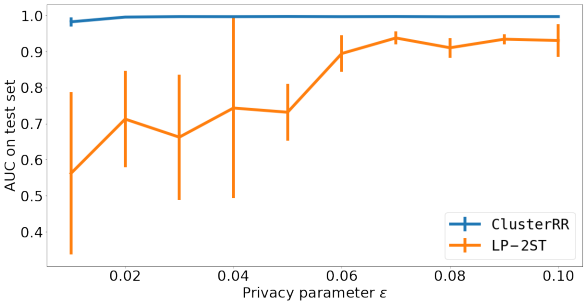


Figure 2: Performance of each mechanism on binarized MNIST dataset.

the mechanism applies uniform randomized response to a portion of the training set, which is then used to learn a model. In the second stage, the mechanism applies uniform randomized response to the remainder of the training set, but only after using the model's predictions to prune the label set to the most likely labels. While effective when the number of labels is large, this approach can lead to degenerate prunings consisting of only a single label in the second stage, especially for binary classification problems and for small values of the privacy parameter $\epsilon$. Figure 2 compares `ClusterRR` to `LP-2ST` on a binarized version of MNIST (*i.e.*, one digit is the positive label and all other digits are the negative label). Each data point in the figure is the average of 5 trials.

## 8 CONCLUSION

In this work we presented centralized and distributed label differential privacy mechanisms. Our mechanisms are based on a clustering of examples in the training set. We upper bound the excess risk of our mechanisms by a rate comparable to that of non-private learning, especially when the clusters are both large and homogeneous. We complement our results with a lower bound that illustrates why it is hard to learn privately when we do not have large homogeneous clusters. We also present experimental results on real data showing that our proposed mechanisms outperform existing mechanisms for differentially private learning.

### References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC confer-*

*ence on computer and communications security*, pages 308–318, 2016.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186, 2011.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60 (309):63–69, 1965.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 2018.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, pages 6628–6637. PMLR, 2019.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(10), 2009.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, volume 26, pages 1196–1204, 2013.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 375–403. Springer, 2019.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

Ralph C Merkle. Secure communications over insecure channels. *Communications of the ACM*, 21 (4):294–299, 1978.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, 2017. URL `http://arxiv.org/abs/1708.07747`. cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

DP Logisitic Regression on MNIST. `https://github.com/tensorflow/privacy/blob/master/tutorials/mnist_lr_tutorial.py`, 2019. Copyright 2019, The TensorFlow Authors. Licensed under the Apache License, Version 2.0.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34, 2021.

# Supplementary Material: Label differential privacy via clustering

## A Analysis of the centralized mechanism

### A.1 Label randomization matrices

We first establish properties of the label randomization matrices $\tilde{\mathbf{Q}}_{x,\beta} \in \mathbb{R}^{K\times K}$ defined by the centralized mechanism (see line 29 of Algorithm 1).

**Lemma 1.** *The minimum singular value of any label randomization matrix $\tilde{\mathbf{Q}}_{x,\beta}$ is at least $\frac{1-\beta}{\sqrt{2K}}$.*

*Proof.* For brevity, we drop subscripts and conditioning on $x$, letting $\mathbf{Q} = \tilde{\mathbf{Q}}_{x,\beta}$ and $\mathbf{q} = \tilde{\mathbf{q}}(\cdot|c_x)$ Let $\mathbf{v} =< v_1, \ldots, v_K >$ be an arbitrary vector such that $\|\mathbf{v}\|_2 = 1$. Let $\mathbf{u} =< u_1, \ldots, u_K >= \mathbf{Q}\mathbf{v}$. Note that for all $i$ we have $u_i = (1-\beta)v_i + \beta\mathbf{q}^\top\mathbf{v}$. We prove this lemma in two cases: First, all $v_i$s have the same sign. Second, there exists a $v_i$ which is negative and a $v_j$ which is positive.

**Case 1: All $v_i$s have the same sign.** Note that we have

$$\begin{aligned}
\|\mathbf{u}\|_2 &= \sqrt{\sum_{i=1}^{K}\big((1-\beta)v_i + \beta\mathbf{q}^\top\mathbf{v}\big)^2} \\
&\geq \sqrt{\sum_{i=1}^{K}\big((1-\beta)v_i\big)^2} && \text{same sign} \\
&\geq \sqrt{\max_{i=1}^{K}\big((1-\beta)v_i\big)^2} \\
&\geq \sqrt{\max_{i=1}^{K}\big((1-\beta)\frac{1}{\sqrt{K}}\big)^2} && \text{since } \sum_{i=1}^{K} v_i^2 = 1 \\
&= \frac{1-\beta}{\sqrt{K}}.
\end{aligned}$$

**Case 2: There exists a $v_i$ which is negative and a $v_j$ which is positive.** Note that we have

$$\begin{aligned}
&\|\mathbf{u}\|_2 = \\
&\sqrt{\sum_{i=1}^{K}\big((1-\beta)v_i + \beta\mathbf{q}^\top\mathbf{v}\big)^2} \geq \\
&\sqrt{\big((1-\beta)\min_i v_i + \beta\mathbf{q}^\top\mathbf{v}\big)^2 + \big((1-\beta)\max_i v_i + \beta\mathbf{q}^\top\mathbf{v}\big)^2} \\
&= (1-\beta)\sqrt{\big(\min_i v_i + \frac{\beta\mathbf{q}^\top\mathbf{v}}{1-\beta}\big)^2 + \big(\max_i v_i + \frac{\beta\mathbf{q}^\top\mathbf{v}}{1-\beta}\big)^2} \\
&\geq (1-\beta)\min_x \sqrt{\big(\min_i v_i + x\big)^2 + \big(\max_i v_i + x\big)^2} = \\
&(1-\beta)\sqrt{\big(\frac{\min_i v_i - \max_i v_i}{2}\big)^2 + \big(\frac{\max_i v_i - \min_i v_i}{2}\big)^2}
\end{aligned}$$

$$= (1-\beta)\sqrt{\left(\frac{1}{2\sqrt{K}}\right)^2 + \left(\frac{1}{2\sqrt{K}}\right)^2}$$
$$= \frac{1-\beta}{\sqrt{2K}}.$$

□

**Lemma 2.** *Each label randomization matrix* $\tilde{\mathbf{Q}}_{x,\beta}$ *satisfies* $\max_y \sum_{y'} \left|\tilde{Q}^{-1}_{x,\beta}[y',y]\right| \leq \frac{\sqrt{2}K}{1-\beta}$.

*Proof.* By properties of matrix norms we have

$$\max_y \sum_{y'} \left|\tilde{Q}^{-1}_{x,\beta}[y',y]\right| = \left\|\tilde{\mathbf{Q}}^{-1}_{x,\beta}\right\|_1 \leq \sqrt{K}\left\|\tilde{\mathbf{Q}}^{-1}_{x,\beta}\right\|_2 \leq \frac{\sqrt{2}K}{1-\beta}$$

where the last inequality follows from Lemma 1. □

### A.2 Well-definedness of centralized mechanism

**Theorem 6** (Well-definedness)**.** *In Algorithm 1, the cluster label distributions* $\tilde{\mathbf{q}}$ *satisfy* $\tilde{q}(y|c) \in [\tau, 1]$ *and* $\sum_{y' \in \mathcal{Y}} \tilde{q}(y'|c) = 1$ *for every label* $y \in \mathcal{Y}$ *and cluster* $c \in \mathcal{C}$*. Also, each label randomization matrix* $\mathbf{Q}_{x,\beta}$ *is invertible.*

*Proof.* To show that $\tilde{q}(y|c) \in [\tau, 1]$, first note that clearly $q(y|c) \in [\tau, 1]$, and therefore $\xi_{y,c} \geq 0$. So if $\Delta_c < 0$ then $\tilde{q}(y|c) \leq 1$ and

$$\begin{aligned}
\tilde{q}(y|c) &= q(y|c) + \frac{\xi_{y,c}}{\sum_{y'} \xi_{y',c}}\Delta_c \\
&= \tau + q(y|c) - \tau + (q(y|c) - \tau)\frac{\Delta_c}{\sum_{y'}(q(y'|c) - \tau)} \\
&= \tau + q(y|c) - \tau + (q(y|c) - \tau)\frac{\Delta_c}{1 - \Delta_c - K\tau} \\
&= \tau + q(y|c) - \tau + (\tau - q(y|c))\frac{-\Delta_c}{-\Delta_c + 1 - K\tau} \\
&\geq \tau + q(y|c) - \tau + \tau - q(y|c) \\
&= \tau,
\end{aligned}$$

where we used $1 - K\tau \geq 0$. Similarly, if $\Delta_c \geq 0$ then $\tilde{q}(y|c) \geq \tau$ and

$$\begin{aligned}
\tilde{q}(y|c) &= q(y|c) + \frac{\xi_{y,c}}{\sum_{y'} \xi_{y',c}}\Delta_c \\
&= 1 + q(y|c) - 1 + (1 - q(y|c))\frac{\Delta_c}{\sum_{y'}(1 - q(y'|c))} \\
&= 1 + q(y|c) - 1 + (1 - q(y|c))\frac{\Delta_c}{\Delta_c + K - 1} \\
&\leq 1 + q(y|c) - 1 + 1 - q(y|c) \\
&= 1,
\end{aligned}$$

where we used $K - 1 \geq 0$. Thus $\tilde{q}(y|c) \in [\tau, 1]$. Also we have $\sum_{y'} \tilde{q}(y'|c) = 1$ because

$$\sum_{y'} \tilde{q}(y'|c) = \sum_{y'} q(y'|c) + \sum_{y'} \frac{\xi_{y',c}}{\sum_{y''} \xi_{y'',c}}\Delta_c = 1 - \Delta_c + \Delta_c = 1.$$

Finally, the invertibility of each label randomization matrix $\tilde{\mathbf{Q}}_{x,\beta}$ is immediate from Lemma 1 and the fact that $\beta < 1$. □

### A.3 Proof of Theorem 1

*Proof.* Let $M$ be the mechanism in Algorithm 1. We can write $M$ as the composition of two mechanisms, $M_1$ and $M_2$, with $M(D) = M_2(D, M_1(D))$, where $M_1(D)$ outputs the noisy cluster label distributions $\tilde{\mathbf{q}}$, and $M_2(D, \tilde{\mathbf{q}})$ uses $\tilde{\mathbf{q}}$ to resample the labels in $D$ to form $\tilde{D}$ and constructs the modified loss function $\tilde{\ell}$. By sequential composition and post-processing, if $M_1$ and $M_2$ are $\epsilon_1$- and $\epsilon_2$-differentially private, respectively, then $M$ is $(\epsilon_1 + \epsilon_2)$-differentially private.

Note that after adding $z_{y,c}$ to each $\hat{p}_{y|c}(D)$, mechanism $M_1$ does not access dataset $D$ again. Since each $\hat{p}_{y|c}(D)$ is computed using a disjoint subset of the dataset and has sensitivity $1/n_c(D)$, and the scale of Laplace random variable $z_{y,c}$ is $\sigma/n_c(D)$, mechanism $M_1$ is $(1/\sigma)$-differentially private.

Mechanism $M_2$ is just randomized response per label, using $q(\cdot|c_{x_i})$ as the random label distribution for each labeled example $(x_i, y_i)$, followed by post-processing. Thus $M_2$ is $\log(1 + (1-\lambda)/\lambda\tau)$-differentially private, since for all $y \in \mathcal{Y}$ we have

$$\frac{\Pr[\tilde{y}_i = y \mid y_i = y]}{\Pr[\tilde{y}_i = y \mid y_i \neq y]} = \frac{1 - \lambda + \lambda\tilde{q}(y|c_{x_i}))}{\lambda\tilde{q}(y|c_{x_i})} \leq 1 + \frac{1-\lambda}{\lambda\tau}.$$

□

### A.4 Proof of Theorem 2

Fix threshold $\tau$ and noise scale $\sigma$. We write $(\mathbf{x}, \mathbf{y}, \mathring{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{z}) \sim \mathcal{P}^n_{\beta,\lambda}$ to denote the following joint distribution: Draw $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{P}^n$, run Algorithm 1 on input dataset $D = (\mathbf{x}, \mathbf{y})$ with bias correction parameter $\beta$ and label flipping probability $\lambda$, and let $\mathring{\mathbf{y}}$, $\tilde{\mathbf{y}}$, $\mathbf{z}$ be the vectors of variables $z_{y,c}$, $\mathring{y}_i$, $\tilde{y}_i$, respectively, defined in the algorithm. Note that Algorithm 1 is deterministic if $(\mathbf{x}, \mathbf{y}, \mathring{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{z})$ is fixed.

Let $\hat{R}_{\beta,\lambda}(h) = \frac{1}{n}\sum_{(x,y)\in\tilde{D}} \tilde{\ell}(h, x, y)$ be the empirical loss of $h$ with respect to the loss function and dataset output by Algorithm 1 when $(\mathbf{x}, \mathbf{y}, \mathring{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{z}) \sim \mathcal{P}^n_{\beta,\lambda}$, and let $R_{\beta,\lambda}(h) = \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h)]$.

**Lemma 3** (Unbiasedness of modified loss)**.** *$R_{\beta,\beta}(h) = R(h)$ for any hypothesis $h \in \mathcal{H}$.*

*Proof.* Let $(\mathbf{x}, \mathbf{y}, \mathring{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{z}) \sim \mathcal{P}^n_{\beta,\beta}$. We have

$$\begin{aligned}
R_{\beta,\beta}(h) &= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\beta}(h)] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \tilde{\ell}(h, x_i, \tilde{y}_i)\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_{y'} \tilde{Q}^{-1}_{x_i,\beta}[y', \tilde{y}_i]\ell(h, x_i, y')\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_{y'} \sum_y \mathbf{1}\{\tilde{y}_i = y\}\, \tilde{Q}^{-1}_{x_i,\beta}[y', y]\ell(h, x_i, y')\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_{y'} \sum_y \tilde{Q}_{x_i,\beta}[y, y_i]\tilde{Q}^{-1}_{x_i,\beta}[y', y]\ell(h, x_i, y')\right] \qquad (2) \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_{y'} \ell(h, x_i, y') \sum_y \tilde{Q}^{-1}_{x_i,\beta}[y', y]\tilde{Q}_{x_i,\beta}[y, y_i]\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y}}\left[\frac{1}{n}\sum_i \ell(h, x_i, y_i)\right] \qquad (3) \\
&= R(h)
\end{aligned}$$

where Eq. (2) follows from the definition of $\tilde{\mathbf{Q}}_{x,\beta}$ in Algorithm 1 (see line 29, and recall that in this case $\beta = \lambda$). We establish Eq. (3) by letting $\mathbf{M} = \tilde{\mathbf{Q}}_{x_i,\beta}^{-1}\tilde{\mathbf{Q}}_{x_i,\beta}$ and noting

$$M[y', y_i] = \sum_y \tilde{Q}_{x_i,\beta}^{-1}[y', y]\tilde{Q}_{x_i,\beta}[y, y_i]$$

and $\mathbf{M} = \mathbf{I}$, and therefore $M[y', y_i] = 1$ if $y' = y_i$ and $M[y', y_i] = 0$ otherwise. □

**Lemma 4** (Boundedness of modified loss)**.** $\max_{h,x,y} \left|\tilde{\ell}(h, x, y)\right| \leq \frac{\sqrt{2}K}{1-\beta}$.

*Proof.* By the definition of $\tilde{\ell}$ in Algorithm 1 (see line 30)

$$\max_{h,x,y} \left|\tilde{\ell}(h, x, y)\right| = \max_{h,x,y} \left|\sum_{y'} \tilde{Q}_{x,\beta}^{-1}[y', y]\ell(h, x, y')\right| \leq \left(\max_y \sum_{y'} \left|\tilde{Q}_{x,\beta}^{-1}[y', y]\right|\right)\left(\max_{h,x,y} |\ell(h, x, y)|\right) \leq \frac{\sqrt{2}K}{1-\beta}$$

where the last inequality follows from Lemma 2 and the fact that $\ell(h, x, y) \in [0, 1]$. □

**Definition 3** (Cluster distortion)**.** *For any mechanism that takes as input a dataset $D$ and defines cluster label distributions $\tilde{\mathbf{q}}_D$ let*

$$\psi = \mathrm{E}_{D\sim\mathcal{P}^n}\left[\max_c \mathrm{E}\left[\sum_y \left|\tilde{q}_D(y|c) - \hat{p}_{y|c}(D)\right|\right]\right]$$

*be the expected maximum total variation between the empirical cluster label distributions and $\tilde{\mathbf{q}}_D$.*

**Lemma 5** (Boundedness of cluster distortion)**.** *If $n_c(D) \geq s$ with probability 1 then*

$$\psi \leq 2K\tau + \frac{2\sqrt{2}K\sigma}{s}.$$

*Proof.* Let $[z]_+ = \max\{0, z\}$ for all $z \in \mathbb{R}$. For any label $y$ and cluster $c$

$$\begin{aligned} q(y|c) - \hat{p}_{y|c}(D) &= \max\left\{\tau, \min\left\{1, \hat{p}_{y|c}(D) + z_{y,c}\right\}\right\} - \hat{p}_{y|c}(D) \\ &\leq \max\left\{\tau, \hat{p}_{y|c}(D) + z_{y,c}\right\} - \hat{p}_{y|c}(D) \\ &\leq \tau + \hat{p}_{y|c}(D) + [z_{y,c}]_+ - \hat{p}_{y|c}(D) \\ &= \tau + [z_{y,c}]_+ \end{aligned}$$

and

$$\begin{aligned} \hat{p}_{y|c}(D) - q(y|c) &= \hat{p}_{y|c}(D) - \max\left\{\tau, \min\left\{1, \hat{p}_{y|c}(D) + z_{y,c}\right\}\right\} \\ &\leq \hat{p}_{y|c}(D) - \min\left\{1, \hat{p}_{y|c}(D) + z_{y,c}\right\} \\ &= \max\left\{\hat{p}_{y|c}(D) - 1, -z_{y,c}\right\} \\ &\leq [-z_{y,c}]_+ \end{aligned}$$

which implies

$$|q(y|c) - \hat{p}_{y|c}(D)| = \max\left\{q(y|c) - \hat{p}_{y|c}(D), \hat{p}_{y|c}(D) - q(y|c)\right\} \leq \tau + [z_{y,c}]_+ + [-z_{y,c}]_+ = \tau + |z_{y,c}| \tag{4}$$

We also have

$$-\Delta_c = \sum_y q(y|c) - 1 \leq \sum_y \left(\hat{p}_{y|c}(D) + \tau + [z_{y,c}]_+\right) - 1 = K\tau + \sum_y [z_{y,c}]_+$$

and

$$\Delta_c = 1 - \sum_y q(y|c) \leq 1 - \sum_y \left(\hat{p}_{y|c}(D) - [-z_{y,c}]_+\right) = \sum_y [-z_{y,c}]_+$$

which implies

$$|\Delta_c| = \max\{-\Delta_c, \Delta_c\} \le K\tau + \sum_y [z_{y,c}]_+ + \sum_y [-z_{y,c}]_+ = K\tau + \sum_y |z_{y,c}| \tag{5}$$

Therefore

$$\begin{aligned}
\left|\tilde{q}(y|c) - \hat{p}_{y|c}(D)\right| &= \left|q(y|c) + \frac{\xi_{y,c}}{\sum_{y'} \xi_{y',c}}\Delta_c - \hat{p}_{y|c}(D)\right| \\
&\le \left|q(y|c) - \hat{p}_{y|c}(D)\right| + \frac{\xi_{y,c}}{\sum_{y'} \xi_{y',c}} |\Delta_c| \\
&\le \tau + |z_{y,c}| + \frac{\xi_{y,c}}{\sum_{y'} \xi_{y',c}} \left(K\tau + \sum_{y'} |z_{y',c}|\right) \qquad (6)
\end{aligned}$$

where Eq. (6) follows from Eq. (4) and Eq. (5). Therefore for any cluster $c$

$$\begin{aligned}
\mathrm{E}_{\mathbf{z}}\left[\sum_y \left|\tilde{q}(y|c) - \hat{p}_{y|c}(D)\right|\right] &\le K\tau + \sum_y \mathrm{E}_{\mathbf{z}}\left[|z_{y,c}|\right] + \mathrm{E}_{\mathbf{z}}\left[\frac{\sum_y \xi_{y,c}}{\sum_{y'} \xi_{y',c}}\left(K\tau + \sum_{y'} |z_{y',c}|\right)\right] \\
&= 2K\tau + 2\sum_y \mathrm{E}_{\mathbf{z}}\left[|z_{y,c}|\right] \qquad (7)
\end{aligned}$$

Recall that each $z_{y,c}$ has mean zero and standard deviation $\frac{\sqrt{2}\sigma}{n_c(D)}$. Continuing from Eq. (7) we have

$$2K\tau + 2\sum_y \mathrm{E}_{\mathbf{z}}\left[|z_{y,c}|\right] \le 2K\tau + 2\sum_y \sqrt{\mathrm{E}_{\mathbf{z}}\left[z_{y,c}^2\right]} = 2K\tau + \frac{2\sqrt{2}K\sigma}{n_c(D)} \le 2K\tau + \frac{2\sqrt{2}K\sigma}{s}$$

where we used Jensen's inequality and $n_c(D) \ge s$. □

**Lemma 6** (Excess risk)**.** *If* $\max_{h,x,y} |\tilde{\ell}(h,x,y)| \le \tilde{L}$ *then for any hypothesis* $h$

$$|R_{\beta,\beta}(h) - R_{\beta,\lambda}(h)| \le \tilde{L}|\beta - \lambda|\,(\phi + \psi)$$

*Proof.* Let $(\mathbf{x}, \mathbf{y}, \mathring{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{z}) \sim \mathcal{P}^n_{\beta,\beta}$ and $(\mathbf{x}', \mathbf{y}', \mathring{\mathbf{y}}', \tilde{\mathbf{y}}', \mathbf{z}') \sim \mathcal{P}^n_{\beta,\lambda}$. Note that between each corresponding pair of variables only $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$ can have different distributions. Therefore

$$\begin{aligned}
R_{\beta,\beta}(h) - R_{\beta,\lambda}(h) &= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}\left[\hat{R}_{\beta,\beta}(h)\right] - \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}}',\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \tilde{\ell}(h, x_i, \tilde{y}_i)\right] - \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}}',\mathbf{z}}\left[\frac{1}{n}\sum_i \tilde{\ell}(h, x_i, \tilde{y}'_i)\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_y \left((1-\beta)\mathbf{1}\{y_i = y\} + \beta\mathbf{1}\{\mathring{y}_i = y\}\right)\tilde{\ell}(h, x_i, y)\right] \\
&\quad - \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_y \left((1-\lambda)\mathbf{1}\{y_i = y\} + \lambda\mathbf{1}\{\mathring{y}_i = y\}\right)\tilde{\ell}(h, x_i, y)\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\frac{1}{n}\sum_i \sum_y \left((\lambda-\beta)\mathbf{1}\{y_i = y\} + (\beta-\lambda)\mathbf{1}\{\mathring{y}_i = y\}\right)\tilde{\ell}(h, x_i, y)\right] \\
&= (\beta - \lambda)\frac{1}{n}\sum_i \mathrm{E}_{\mathbf{x},\mathbf{y},\mathbf{z}}\left[\sum_y \left(p_{y|c_{x_i}}(D) - \mathbf{1}\{y_i = y\}\right)\tilde{\ell}(h, x_i, y)\right] \qquad (8)
\end{aligned}$$

$$+ (\beta - \lambda)\frac{1}{n}\sum_i \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\sum_y \left(\mathbf{1}\left\{\mathring{y}_i = y\right\} - p_{y|c_{x_i}}(D)\right)\tilde{\ell}(h, x_i, y)\right] \tag{9}$$

Each term in Eq. (8) is

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x},\mathbf{y},\mathbf{z}}\left[\sum_y \left(p_{y|c_{x_i}}(D) - \mathbf{1}\left\{y_i = y\right\}\right)\tilde{\ell}(h, x_i, y)\right] \\
&= \mathrm{E}_{\mathbf{x}}\left[\sum_y \left(\frac{p(y|x_i)}{n_{c_{x_i}}(D)} + \frac{(n_{c_{x_i}}(D) - 1)p(y|c_{x_i})}{n_{c_{x_i}}(D)} - p(y|x_i)\right)\mathrm{E}_{\mathbf{z}}\left[\tilde{\ell}(h, x_i, y)\right]\right] \\
&\leq \mathrm{E}_{\mathbf{x}}\left[\sum_y |p(y|c_{x_i}) - p(y|x_i)|\left|\mathrm{E}_{\mathbf{z}}\left[\tilde{\ell}(h, x_i, y)\right]\right|\right] \\
&\leq \tilde{L}\phi
\end{aligned} \tag{10}$$

where Eq. (10) follows from our assumption about $\tilde{L}$ and the definition of cluster heterogeneity in Definition 1. Each term in Eq. (9) is

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\mathbf{z}}\left[\sum_y \left(\mathbf{1}\left\{\mathring{y}_i = y\right\} - p_{y|c_{x_i}}(D)\right)\tilde{\ell}(h, x_i, y)\right] \\
&= \mathrm{E}_{\mathbf{x},\mathbf{y},\mathbf{z}}\left[\sum_y \left(\tilde{q}(y|c_{x_i}) - p_{y|c_{x_i}}(D)\right)\tilde{\ell}(h, x_i, y)\right] \\
&\leq \mathrm{E}_{\mathbf{x},\mathbf{y}}\left[\mathrm{E}_{\mathbf{z}}\left[\sum_y \left|\tilde{q}(y|c_{x_i}) - p_{y|c_{x_i}}(D)\right|\left|\tilde{\ell}(h, x_i, y)\right|\right]\right] \\
&\leq \tilde{L}\psi
\end{aligned} \tag{11}$$

where Eq. (11) follows from our assumption about $\tilde{L}$ and the definition of cluster distortion in Definition 3. Combining Eq. (8), (9), (10) and (11) proves the lemma. □

**Lemma 7** (Complexity bound)**.** *There exists a universal constant $C > 0$ such that*

$$\max_{h\in\mathcal{H}}\left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - R_{\beta,\lambda}(h)\right| \leq \frac{CK}{1-\beta}\sqrt{\frac{\dim(\mathcal{H}, \ell)\log\frac{1}{\gamma}}{n}}$$

*with probability $1 - \gamma$.*

*Proof.* We first review some results from statistical learning theory Mohri et al. (2018). Let $\mathbf{a} = (a_1, \ldots, a_n) \in \mathcal{A}^n$ be a vector of independent random variables, and let $\mathcal{F}$ be a class of real-valued functions with domain $\mathcal{A}^n$. We say $\mathcal{F}$ has $b$-bounded differences if $|f(\mathbf{a}) - f((\mathbf{a}_{-i}, a_i))| \leq \frac{b}{n}$ for all $f \in \mathcal{F}$ and $a_i \in \mathcal{A}$. If $\mathcal{F}$ has $b$-bounded differences then with probability $1 - \gamma$

$$\max_{f\in\mathcal{F}}|f(\mathbf{a}) - \mathrm{E}_{\mathbf{a}}[f(\mathbf{a})]| \leq 2\mathfrak{R}(\mathcal{F}, \mathbf{a}) + \sqrt{\frac{b\log\frac{1}{\gamma}}{n}} \tag{12}$$

where $\mathfrak{R}(\mathcal{F}, \mathbf{a})$ is the Rademacher complexity of $\mathcal{F}$ for random variable $\mathbf{a}$. For any $b \geq 0$ let

$$\mathrm{absconv}_b(\mathcal{F}) = \left\{\sum_{i=1}^N w_i f_i : N \in \mathbb{N}, \sum_{i=1}^N |w_i| \leq b, f_i \in \mathcal{F}\right\} \tag{13}$$

be the absolute convex hull of $\mathcal{F}$ scaled by $b$. We have

$$\mathfrak{R}(\mathrm{absconv}_b(\mathcal{F}), \mathbf{a}) = b \cdot \mathfrak{R}(\mathcal{F}, \mathbf{a}) \tag{14}$$

Finally, if $\mathcal{F} = \{(x, y) \mapsto \ell(h, x, y)\}$ then

$$\mathfrak{R}(\mathcal{F}, \mathbf{a}) \leq C\sqrt{\frac{\dim(\mathcal{H}, \ell)}{n}} \tag{15}$$

for a universal constant $C > 0$.

We now proceed to prove the lemma. We have

$$\begin{aligned}
\left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - R_{\beta,\lambda}(h)\right| &= \left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h)]\right| \\
&\leq \left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - \mathrm{E}_{\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h) \mid \mathbf{x}, \mathbf{y}]\right| \\
&\quad + \left|\mathrm{E}_{\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h) \mid \mathbf{x}, \mathbf{y}] - \mathrm{E}_{\mathbf{x},\mathbf{y},\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h)]\right| \\
&= \left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - \mathrm{E}_{\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h) \mid \mathbf{x}, \mathbf{y}]\right| \\
&\quad + \left|\mathrm{E}_{\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h) \mid \mathbf{x}, \mathbf{y}] - \mathrm{E}_{\mathbf{x},\mathbf{y}}\left[\mathrm{E}_{\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}[\hat{R}_{\beta,\lambda}(h) \mid \mathbf{x}, \mathbf{y}]\right]\right|
\end{aligned} \tag{16}$$

which follows from definitions. Let $\mathcal{F}', \mathcal{F}''$ be the function classes

$$\begin{aligned}
\mathcal{F}' &= \left\{(\mathring{\mathbf{y}}, \tilde{\mathbf{y}}) \mapsto \mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right]\right\} \\
\mathcal{F}'' &= \left\{(\mathbf{x}, \mathbf{y}) \mapsto \mathrm{E}_{\mathring{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h) \mid \mathbf{x}, \mathbf{y}\right]\right\}
\end{aligned}$$

Recalling that $(\mathbf{x}, \mathbf{y}, \mathring{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{z}) \sim \mathcal{P}^n_{\beta,\lambda}$, note that each $(x_i, y_i)$ is independent and each $(\mathring{y}_i, \tilde{y}_i)$ is independent given $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Continuing from Eq. (16), we have with probability $1 - \gamma$

$$\left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - R_{\beta,\lambda}(h)\right| \leq 2\mathfrak{R}(\mathcal{F}', (\mathring{\mathbf{y}}, \tilde{\mathbf{y}})) + 2\mathfrak{R}(\mathcal{F}'', (\mathbf{x}, \mathbf{y})) + 2\sqrt{\max_{h,x,y}|\tilde{\ell}(h, x, y)|}\sqrt{\frac{\log\frac{1}{\gamma}}{n}} \tag{17}$$

$$\leq 2\mathfrak{R}(\mathcal{F}', (\mathring{\mathbf{y}}, \tilde{\mathbf{y}})) + 2\mathfrak{R}(\mathcal{F}'', (\mathbf{x}, \mathbf{y})) + \sqrt{\frac{2K}{1-\beta}}\sqrt{\frac{\log\frac{1}{\gamma}}{n}} \tag{18}$$

$$\leq \frac{8CK}{1-\beta}\sqrt{\frac{\dim(\mathcal{H}, \ell)}{n}} + \sqrt{\frac{2K}{1-\beta}}\sqrt{\frac{\log\frac{1}{\gamma}}{n}} \tag{19}$$

where Eq. (17) follows from Eq. (12), Eq. (18) follows from Lemma 4, and Eq. (19) follows from the definition of $\tilde{\ell}$ in Algorithm 1 (see line 30), Lemma 2, Eq. (15) and Eq. (15). Combining terms proves the lemma. □

We are now ready to prove Theorem 2.

*Proof of Theorem 2.* Combining Lemma 4, Lemma 5, Lemma 6 and our assumption that $n_c(D) \geq s$ with probability 1 we have

$$\max_{h\in\mathcal{H}}|R_{\beta,\beta}(h) - R_{\beta,\lambda}(h)| \leq \frac{4K}{1-\beta}\left(\phi + K\tau + \frac{K\sigma}{s}\right) \tag{20}$$

Therefore

$$\begin{aligned}
R(\tilde{h}) - R(h^*) &= R_{\beta,\beta}(\tilde{h}) - R_{\beta,\beta}(h^*) \\
&= \hat{R}_{\beta,\lambda}(\tilde{h}) - \hat{R}_{\beta,\lambda}(h^*) + (R_{\beta,\beta}(\tilde{h}) - R_{\beta,\lambda}(\tilde{h})) + (R_{\beta,\lambda}(h^*) - R_{\beta,\beta}(h^*)) \\
&\quad + (R_{\beta,\lambda}(\tilde{h}) - \hat{R}_{\beta,\lambda}(\tilde{h})) + (\hat{R}_{\beta,\lambda}(h^*) - R_{\beta,\lambda}(h^*)) \\
&\leq 0 + \frac{8K}{1-\beta}\left(\phi + K\tau + \frac{K\sigma}{s}\right) \\
&\quad + (R_{\beta,\lambda}(\tilde{h}) - \hat{R}_{\beta,\lambda}(\tilde{h})) + (\hat{R}_{\beta,\lambda}(h^*) - R_{\beta,\lambda}(h^*))
\end{aligned} \tag{21}$$

where Eq. (21) follows from the choice of $\tilde{h}$ and Eq. (20). Taking the expectation of both sides over $\mathbf{z}$ and continuing from Eq. (21) we have

$$\begin{aligned}\mathrm{E}_{\mathbf{z}}[R(\tilde{h})] - R(h^*) &\leq \frac{8K}{1-\beta}\left(\phi + K\tau + \frac{K\sigma}{s}\right) + 2\max_{h\in\mathcal{H}}\left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - R_{\beta,\lambda}(h)\right| \\ &\leq \frac{8K}{1-\beta}\left(\phi + K\tau + \frac{K\sigma}{s}\right) + 2\max_{h\in\mathcal{H}}\left|\mathrm{E}_{\mathbf{z}}\left[\hat{R}_{\beta,\lambda}(h)\right] - R_{\beta,\lambda}(h)\right| \qquad (22)\end{aligned}$$

where Eq. (22) follows from Lemma 7. □

# B Analysis of peer-to-peer mechanism

## B.1 Proof of Theorem 3

First we need a technical lemma.

**Lemma 8.** $\left(1+\frac{s}{x}\right)^{sx^a} \leq e^{2a-1} + \frac{3}{x^a}$ *for all* $x \geq 2, s \in \{-1,1\}, a \in \{\frac{1}{2}, 1\}$

We next state and prove a more general version of Theorem 3.

**Lemma 9** (Peer-to-peer privacy, general version)**.** *If non-empty clusters in $D$ have size at least $\frac{2}{\alpha}$ then the peer-to-peer mechanism (Algorithm 2) satisfies $(\epsilon,\delta)$-label differential privacy with*

$$\epsilon = \theta\log\left(e + \frac{3}{s\alpha}\right) + \sqrt{\theta}\xi\log\left(1 + \frac{3}{\sqrt{s\alpha}}\right)$$
$$\delta = \exp\left(-\frac{\alpha\xi^2}{4\left(\alpha+\frac{1}{s}\right)(1-\alpha)}\right)$$

*for all* $\xi \in \left[0, 3\alpha\sqrt{\theta s(1-\alpha)}\right]$.

*Proof.* Consider two neighboring datasets $D$ and $D'$ such that there is an example with label 0 in $D$ but label 1 in $D'$, and let $c$ be the cluster containing this example. Let $\tilde{D}$ and $\tilde{D}'$ be the output of the peer-to-peer mechanism when given $D$ and $D'$, respectively, as input. Since the labels in $\tilde{D}$ and $\tilde{D}'$ are chosen independently per cluster, the label distribution in all clusters other than $c$ is identical in both $\tilde{D}$ and $\tilde{D}'$.

Let $n = n_c(D) = n_c(D') \geq s$ be the number of examples in cluster $c$, and let $p$ be the fraction of examples in cluster $c$ with a positive label in $D$. Also let $t = \theta n$ be the fraction of users in cluster $c$ who send an example to the learner. Observe that the label distribution in cluster $c$ in $\tilde{D}$ is completely characterized by the binomial density function $f(k;t,p)$, which gives the probability of $k$ successes in $t$ trials that each have success probability $p$. Similarly, the label distribution in cluster $c$ in $\tilde{D}'$ is completely charaterized by $f(k;t,p')$, where $p' = p + \frac{1}{n}$.

Let $q = 1-p$. Also let $S_- = \{k \in \mathbb{N} : k \geq tp - \xi\sqrt{tq}\}$ and $S_+ = \{k \in \mathbb{N} : k \leq tp + \xi\sqrt{tp}\}$. Thus to prove the theorem it suffices to show

$$\frac{f(k;t,p)}{f(k;t,p')} \leq e^{\epsilon} \text{ if } k \in S_- \text{ and } \frac{f(k;t,p')}{f(k;t,p)} \leq e^{\epsilon} \text{ if } k \in S_+ \qquad (23)$$

and

$$\sum_{k\notin S_-} f(k;t,p) \leq \delta \text{ and } \sum_{k\notin S_+} f(k;t,p') \leq \delta. \qquad (24)$$

To prove the first part of Eq. (23) we can simplify

$$\frac{f(k;t,p)}{f(k;t,p')} = \frac{p^k(1-p)^{t-k}}{(p')^k(1-p')^{t-k}} = \left(\frac{np}{np+1}\right)^k\left(\frac{nq}{nq-1}\right)^{t-k} = \left(1+\frac{1}{np}\right)^{-k}\left(1-\frac{1}{nq}\right)^{k-t}, \qquad (25)$$

and if $k \in S_-$ then $k \geq tp - \xi\sqrt{tq}$ which implies

$$\left(1+\frac{1}{np}\right)^{-k}\left(1-\frac{1}{nq}\right)^{k-t} \leq \left(1-\frac{1}{nq}\right)^{-tq-\xi\sqrt{tq}} = \left(\left(1-\frac{1}{nq}\right)^{-nq}\right)^{\theta}\left(\left(1-\frac{1}{nq}\right)^{-\sqrt{nq}}\right)^{\sqrt{\theta}\xi}, \tag{26}$$

and by applying Lemma 8 and $nq \geq s\alpha \geq 2$ we have

$$\left(\left(1-\frac{1}{nq}\right)^{-nq}\right)^{\theta}\left(\left(1-\frac{1}{nq}\right)^{-\sqrt{nq}}\right)^{\sqrt{\theta}\xi} \leq \left(e+\frac{3}{nq}\right)^{\theta}\left(1+\frac{3}{\sqrt{nq}}\right)^{\sqrt{\theta}\xi} \tag{27}$$

$$\leq \left(e+\frac{3}{s\alpha}\right)^{\theta}\left(1+\frac{3}{\sqrt{s\alpha}}\right)^{\xi\sqrt{\theta}} = e^{\epsilon}. \tag{28}$$

Similarly, to prove the second part of Eq. (23) we can simplify

$$\frac{f(k;t,p')}{f(k;t,p)} = \frac{(p')^k(1-p')^{t-k}}{p^k(1-p)^{t-k}} = \left(\frac{np+1}{np}\right)^k\left(\frac{nq-1}{nq}\right)^{t-k} = \left(1+\frac{1}{np}\right)^k\left(1-\frac{1}{nq}\right)^{t-k}, \tag{29}$$

and if $k \in S_+$ then $k \leq tp + \xi\sqrt{tp}$ which implies

$$\left(1+\frac{1}{np}\right)^k\left(1-\frac{1}{nq}\right)^{t-k} \leq \left(1+\frac{1}{np}\right)^{tp+\xi\sqrt{tp}} = \left(\left(1+\frac{1}{np}\right)^{np}\right)^{\theta}\left(\left(1+\frac{1}{np}\right)^{\sqrt{np}}\right)^{\sqrt{\theta}\xi}, \tag{30}$$

and by applying Lemma 8 and $np \geq s\alpha \geq 2$ we have

$$\left(\left(1+\frac{1}{np}\right)^{np}\right)^{\theta}\left(\left(1+\frac{1}{np}\right)^{\sqrt{np}}\right)^{\sqrt{\theta}\xi} \leq \left(e+\frac{3}{np}\right)^{\theta}\left(1+\frac{3}{\sqrt{np}}\right)^{\sqrt{\theta}\xi} \tag{31}$$

$$\leq \left(e+\frac{3}{s\alpha}\right)^{\theta}\left(1+\frac{3}{\sqrt{s\alpha}}\right)^{\sqrt{\theta}\xi} = e^{\epsilon}. \tag{32}$$

To prove the first part of Eq. (24) define the binomial cumulative distribution function $F(k;t,p) = \sum_{k' \leq k} f(k;t,p)$. By Bernstein's inequality

$$F(tp - t\gamma; t, p) \leq \exp\left(-\frac{\gamma^2 t}{2pq + 2\gamma/3}\right)$$

for all $\gamma > 0$. Let $\gamma = \xi\sqrt{\frac{q}{t}}$ and note that

$$\frac{2}{3}\gamma = \frac{2}{3}\xi\sqrt{\frac{q}{t}} \leq 2\alpha\sqrt{\theta s(1-\alpha)}\sqrt{\frac{q}{t}} = 2\alpha\sqrt{1-\alpha}\sqrt{q}\sqrt{\frac{s}{n}} \leq 2(1-q)q\sqrt{\frac{s}{n}} \leq 2pq,$$

and thus

$$\sum_{k\in S_-} f(k;t,p) = F(tp - t\gamma; t, p) \leq \exp\left(-\frac{\gamma^2 t}{4pq}\right) \leq \exp\left(-\frac{\xi^2}{4p}\right) \leq \exp\left(-\frac{\xi^2}{4(1-\alpha)}\right) \leq \delta.$$

To prove the second part of Eq. (24) let $\gamma = \xi\sqrt{\frac{p}{t}}$ and $q' = 1 - p'$. We have

$$\frac{2}{3}\gamma = \frac{2}{3}\xi\sqrt{\frac{p}{t}} \leq 2\alpha\sqrt{\theta s(1-\alpha)}\sqrt{\frac{p}{t}} = 2\alpha\sqrt{1-\alpha}\sqrt{p}\sqrt{\frac{s}{n}} \leq 2(1-p)p\sqrt{\frac{s}{n}} \leq 2(1-p')p' = 2p'q',$$

and since $1 - F(tp' + t\gamma; t, p') = F(tq' - t\gamma; t; q')$ we have

$$\sum_{k\in S_+} f(k;t,p') = F(tq' - t\gamma; t; q') \leq \exp\left(-\frac{\gamma^2 t}{4p'q'}\right) \leq \exp\left(-\frac{\alpha\xi^2}{4\left(\alpha+\frac{1}{s}\right)(1-\alpha)}\right) = \delta. \qquad \square$$

We are now ready to prove the Theorem 3.

*Proof of Theorem 3.* Let $\xi = 4\sqrt{\log s}$. Since $\alpha = \frac{4\sqrt{2}\log s}{\sqrt{\theta s}} \leq \frac{1}{2}$

$$\xi = 4\sqrt{\log s} \leq 4\log s \leq \alpha\sqrt{\frac{\theta s}{2}} \leq \alpha\sqrt{\theta s(1-\alpha)}$$

where the last inequality uses $\alpha \leq \frac{1}{2}$. Therefore the conditions of Lemma 9 hold. Also

$$\frac{\alpha}{(\alpha+\frac{1}{s})(1-\alpha)} \geq \frac{\alpha}{\alpha+\frac{1}{s}} = \frac{1}{1+\frac{1}{s\alpha}} \geq \frac{1}{2}$$

since $s \geq \frac{1}{\alpha}$. Therefore by Lemma 9

$$\delta = \exp\left(-\frac{\alpha\xi^2}{4\left(\alpha+\frac{1}{s}\right)(1-\alpha)}\right) \leq \exp\left(-\frac{\xi^2}{8}\right) = e^{-2\log s} = \frac{1}{s^2},$$

and since $\alpha = \frac{\xi^2}{2\sqrt{2}\sqrt{\theta s}}$ it follows from Lemma 9 that

$$\begin{aligned}\epsilon = \theta\log\left(e+\frac{3}{s\alpha}\right) + \sqrt{\theta}\xi\log\left(1+\frac{3}{\sqrt{s\alpha}}\right) &= \theta\log\left(e+\frac{\sqrt{72\theta}}{\sqrt{s}\log s}\right) + \sqrt{\theta}\xi\log\left(1+\frac{2^{3/4}3\theta^{1/4}}{s^{1/4}\xi}\right) \qquad (33)\\ &\leq \theta + \frac{3\theta^{3/2}}{e\sqrt{s}\log s} + \frac{2^{3/4}3\theta^{3/4}}{s^{1/4}} \qquad \square\end{aligned}$$

# C Hardness result

## C.1 Proof of Theorem 5

*Proof.* Fix a set of examples $\mathcal{X}$. Select a pair of neighboring datasets $D \in (\mathcal{X}\times\mathcal{Y})^n$ and $D' \in (\mathcal{X}\times\mathcal{Y}')^n$ as follows.

To construct $D$, for each cluster $c \in \mathcal{C}$, select $s$ labels uniformly at random without replacement and assign them to the examples in $c$. Examples in different clusters may have the similar labels. To construct $D'$ from $D$, select an example $x_i$ uniformly at random form $\mathcal{X}$ and redraw its label uniformly at random from $\mathcal{Y}\setminus\mathcal{Y}_{c_{x_i}}$. We use $i$ to denote the index of the data that differs between $D$ and $D'$, with the datapoints being $(x_i, y_i)$ and $(x_i, y_i')$ respectively. Let mechanism $M$ be an $\epsilon$-differential privacy mechanism for LAP that guarantees a $\phi$ precision and an $\eta$ recall. Denote $\tilde{D} = M(D)$ and $\tilde{D}' = M(D')$.

Note that by construction of $\tilde{D}'$, for each cluster $c$ we have $|\mathcal{Y}_c'| = s$. Hence, we have

$$\begin{aligned}\eta &\leq \frac{\sum_{c\in\mathcal{C}}\sum_{y'\in\mathcal{Y}_c'}\mathrm{E}_M[\tilde{D}'(c,y')]}{\sum_{c\in\mathcal{C}}|\mathcal{Y}_c'|}\\ &= \frac{\sum_{(x,y')\in D'}\mathrm{E}_M[\tilde{D}'(c_x,y')]}{n}.\end{aligned}$$

This means that for a [uniformly] random $(x, y') \in D'$ we have $\tilde{D}'(c_x, y') = 1$ with probability at least $\eta$. Recall that index $i$ that indicates the difference of $D$ and $D'$ is chosen uniformly at random from $\{1,\ldots,n\}$. Let $O$ be the set of all possible models that can be generated by $M(.)$ that contains $(c_{x_i}, y_i')$. By definition of differential privacy we have $\Pr[\tilde{D}' \in O] \leq e^{\epsilon}\Pr[\tilde{D} \in O]$. This implies that

$$\Pr[\tilde{D} \in O] \geq e^{-\epsilon}\Pr[\tilde{D}' \in O] \geq e^{-\epsilon}\eta.$$

Hence, with probability at least $e^{-\epsilon}\eta$, we have $\tilde{D}(c_{x_i}, y_i') = 1$. Recall that by construction $y_i'$ is a label chosen uniformly at random from $\mathcal{Y} \setminus \mathcal{Y}_{c_{x_i}}$. Hence each any cluster $c$ is associated with any label $y_i' \notin \mathcal{Y}_c$ with probability $e^{-\epsilon}\eta$. Therefore, we have

$$\mathrm{E}[|\tilde{D}|] \geq \frac{n}{s} \times (K - s) \times e^{-\epsilon}\eta.$$

Hence, we can bound the precision of $\tilde{D}$ by

$$\begin{aligned}
\phi &\leq \frac{\sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}_c} \mathrm{E}_M[\tilde{D}(c, y)]}{\mathrm{E}_M[|\tilde{D}|]} \\
&\leq \frac{\sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}_c} \mathrm{E}_M[\tilde{D}(c, y)]}{\frac{n}{s}(K - s)e^{-\epsilon}\eta} \\
&\leq \frac{n}{\frac{n}{s}(K - s)e^{-\epsilon}\eta} \\
&= \frac{s}{(K - s)e^{-\epsilon}\eta}.
\end{aligned}$$

This gives us $\phi\eta e^{-\epsilon} \leq \frac{s}{(K-s)} \in o(1)$. Hence, for a constant $\epsilon$, either precision $\phi$ is sub-constant or recall $\eta$ is sub-constant. □