Broadening Differential Privacy for Deep Learning Against Model Inversion Attacks

Qiuchen Zhang*, Jing Ma*, Yonghui Xiao†, Jian Lou* and Li Xiong*
*Department of Computer Science, Emory University, Atlanta, GA
Email: {qiuchen.zhang, jing.ma, jian.lou, lxiong}@emory.edu
†Google Inc, Montain View, CA
Email: yohu@google.com

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378274

***Abstract*—Deep learning models have achieved great success in many real-world tasks such as image recognition, machine translation, and self-driving cars. A large amount of data are needed to train a model, and in many cases, the training data are private. Publishing or sharing a deep learning model trained on private datasets could pose privacy concerns. We study the model inversion attacks against deep learning models, which attempt to reconstruct the features of training data corresponding to a given class given access to the model. While deep learning with differential privacy is state-of-the-art for training privacy-preserving models, whether they can provide meaningful protection against model inversion attacks remains an open question. In this paper, we first improve the existing model inversion attacks (MIA) to successfully reconstruct training images from neural network based image recognition models. Then, we demonstrate that deep learning with the standard record-level differential privacy does not provide quantifiable protection against MIA. Subsequently, we propose class-level and subclass-level differential privacy and develop algorithms to provide a quantifiable privacy guarantee against MIA. Experiments on real datasets demonstrate that our proposed privacy notions and mechanisms can effectively defend against MIA while maintaining model accuracy.**

***Index Terms*—Differential Privacy; Deep Learning; Model Inversion Attack**

I. Introduction

Neural networks have achieved great success in many real-world tasks like image recognition and natural language processing [1], [2]. Training neural networks requires a large amount of training data which may contain users' sensitive information such as images, voice, medical histories, and location traces. Publishing or sharing the deep learning model trained on private data directly could pose privacy concerns [3]–[5]. Even though the adversaries do not have access to the original training data, they can use the models to infer or reconstruct (the features of) the training data. Several works demonstrated different attacks aiming to extract information about the private training datasets from the published deep learning models. *Membership inference attacks* [4] attempt to infer whether or not a specific record was in the training dataset given black-box API access to the model. *Model inversion attack* [3] (MIA) attempts to reconstruct a recognizable face image corresponding to a person (a class) from a face recognition model given the name of the person (the class label) and white-box access to the model. The purpose of membership inference attack and MIA are different. The former attempts to recover the "existence" information of a target data point, while the latter attempts to recover the visual property or features of a target class (which can be also private). Therefore, both of them are considered as privacy threats and violations [6]–[8].

Differential privacy (DP) has been widely accepted as a strong and provable privacy framework for statistical data analysis [9]–[12]. Recent works developed deep learning models with DP [13]–[18]. Standard DP requires that the statistical model (parameters) learned from a set of data is indistinguishable regardless of the presence or absence of any record in the dataset. The common way to train a deep learning model with DP is to use *differentially private Stochastic Gradient Descent (DP-SGD)* which injects Gaussian noise to the gradients in each iteration during the SGD based optimization when learning the model parameters [13], [19].

Most works on deep learning with DP focus on improving model accuracy given a privacy requirement or enhancing the privacy and utility tradeoff. There is still a limited demonstration of how effective DP is in protecting against the above mentioned attacks in practice. [20] evaluated DP against membership inference attacks and showed that DP can protect against the attacks successfully only by sacrificing model utility by a considerable margin. This is not surprising as the indistinguishability guarantee of DP with respect to the presence of a record is directly aligned with the goal of preventing the inference of the membership of a record. Injecting noise to the model parameters required by DP naturally degrades the performance of the model.

Whether DP or other mechanisms can provide meaningful privacy protection against model inversion attacks without sacrificing model utility is still an open question. While [3] proposed some preliminary defense measures against MIA, it does not provide a rigorous or quantifiable guarantee against the attacks. Intuitively, if we apply the standard record-level DP, the perturbed model may provide some mitigation to MIA due to the perturbed model parameters. However, since there are typically multiple instances (e.g. face images) corresponding to the same class (e.g. person), record-level DP which only protects the presence of one record may not prevent the reconstruction attack since all the records of the same class are encoded in the model. Another potential solution is

to use group-DP [9] to protect the presence of all records corresponding to one class as a group. However, this will lead to amplified perturbation by the group size which can be determined by the largest class size. Such an application may yield unacceptable model accuracy due to the significantly amplified perturbation while overprotecting certain data since different classes may have varying numbers of records.

In this paper, we focus on the MIA against deep learning models and aim to understand whether existing DP can provide meaningful defense against MIA. Our results show that while it provides some mitigation, it does not provide effective and quantifiable protection. We subsequently propose new DP notions and mechanisms for more effective and quantifiable protection against MIA. The contributions are as follows:

- We first improve the original MIA and demonstrate its success on neural network models (Section III).
- We propose both class-level DP (class-DP) and subclass-level DP (subclass-DP) for deep neural networks as quantifiable privacy notions against MIA (Section IV).
- We propose algorithms for training deep learning models with class-DP and subclass-DP (Section IV). We formally prove the privacy guarantees of the proposed algorithms.
- We evaluate deep learning models with class and subclass-DP against MIA using real datasets (Section V). The results demonstrate that the level of class and subclass-DP directly correlates with the robustness against MIA and hence can provide a quantifiable measure against the risk.

II. Problem Setting and Background

In this section, we first describe our problem setting and the threat model. Then we give the definitions of MIA and DP.

A. Problem Setting

We consider the setting where a model provider trains a neural network classification model $f(\mathbf{x})$ using a private training set D, where $\mathbf{x} \in \mathbb{R}^d$ is an input record in d-dimensional space. The output of $f(\mathbf{x})$ is the prediction vector $\mathbf{y} \in R^k$ where each dimension corresponds to one predefined label or class. The model provider shares the trained model with other parties without sharing the data. We study model inversion attacks where an adversary abuses the shared model by attempting to reconstruct the original (features of) training data corresponding to a target class. Our goal is to develop privacy notions and algorithms that allow a model provider to build a model that is robust against model inversion attacks.

Threat Model. We assume a white-box attack in which an adversary has access to the published model including model structure and parameters, but has no access to the training data, nor back door access [21] to the training process.

B. Model Inversion Attack

Model inversion attack [3] is a reverse engineering attack that attempts to "reconstruct" the training data from a trained neural network model. Given the model parameters and a target $label$, the goal is to find a data point $\mathbf{x}$ corresponding to the label following the same distribution with data points in $\mathcal{D}$ that maximizes $f_{label}(\mathbf{x})$, which is equivalent to minimize the following objective function:

$$c(\mathbf{x}) = 1 - f_{label}(\mathbf{x}), \tag{1}$$

where $f_{label}(\mathbf{x})$ is the confidence score of the target class.

While the reconstructed data point may not correspond to a specific data point in the dataset, it leaks the statistical property or general features of the target class. For example, a face image generated by a successful MIA reveals how the person with the target name (the class label) looks like [3].

C. (ϵ, δ)-Differential Privacy

Differential privacy (DP) [9], [10] is a strong and rigorous privacy guarantee which ensures the output distributions of an algorithm are indistinguishable with a certain probability when the input datasets differ in only one record.

Definition 1. *((ϵ, δ)-Differential Privacy) [9]. Let $\mathcal{D}$ and $\mathcal{D}'$ be two neighboring datasets that differ in at most one entry. A randomized algorithm $\mathcal{A}$ satisfies (ϵ-δ)-differential privacy if for all $\mathcal{S} \subseteq Range(\mathcal{A})$: $Pr\left[\mathcal{A}(\mathcal{D}) \in \mathcal{S}\right] \leq e^{\epsilon} Pr\left[\mathcal{A}(\mathcal{D}') \in \mathcal{S}\right] + \delta$, where $\mathcal{A}(\mathcal{D})$ represents the output of $\mathcal{A}$ with the input $\mathcal{D}$.*

In the definition of (ϵ, δ)-DP, ϵ and δ are the privacy parameters or privacy budget which indicate the privacy loss. A smaller ϵ means a higher level of indistinguishability and hence stronger privacy. A smaller δ means a lower probability that the privacy guarantee provided by ϵ will be broken.

The granularity of DP is dependent on the definition of neighboring datasets. In the original DP definition, two neighboring datasets differ in one record, which can be considered as record-level DP (record-DP). It hides the presence of any record in the input dataset. The standard hamming distance-based DP can be extended depending on other notions of distance between the neighboring datasets under different situations [22], [23].

D. Deep Learning with Differential Privacy

DP has been applied to deep learning models with DP-SGD algorithms [13], [19] in order to protect the privacy of training datasets. In each SGD iteration, DP-SGD clips the Euclidean norm of the gradient and injects calibrated Gaussian noise to the clipped gradient. Each iteration of the DP-SGD becomes a randomized mechanism with a quantifiable *privacy loss* which is defined as follows:

Definition 2. *(Privacy Loss [13]) For neighboring datasets $\mathcal{D}, \mathcal{D}'$, auxiliary input $\boldsymbol{aux}$ and output $\boldsymbol{o} \in Range(\mathcal{A})$, the privacy loss at a particular output $\boldsymbol{o}$ is defined as,*

$$c(\boldsymbol{o} | \mathcal{A}, \boldsymbol{aux}, \mathcal{D}, \mathcal{D}') := \log \frac{\mathbb{P}[\mathcal{A}(\boldsymbol{aux}, \mathcal{D}) = \boldsymbol{o}]}{\mathbb{P}[\mathcal{A}(\boldsymbol{aux}, \mathcal{D}') = \boldsymbol{o}]}. \tag{2}$$

For DP-SGD, each iteration incurs a privacy loss, where $\mathcal{A}$ represents one iteration of the DP-SGD update procedure, $\boldsymbol{o}$ is the updated parameter vector, $\boldsymbol{aux}$ is all the parameter sequences obtained before this iteration, and $\mathcal{D}$ is the training dataset. Abadi et al. [13] proposed the moments accountant

technique with random sampling that provides a tighter privacy loss composition than the advanced composition theorems [24] for the overall privacy loss of DP-SGD over multiple iterations. We will use it for our privacy analysis in this paper.

Definition 3. *(Moments Accountant [13]) The moments accountant of a randomized mechanism $\mathcal{A}$ with κ-th moment is defined as follows:*

$$\alpha_{\mathcal{A}}(\kappa) := \arg \max_{\boldsymbol{aux},(\mathcal{D},\mathcal{D}')} \log \mathbb{E}[exp(\kappa c(\boldsymbol{o}|\mathcal{A}, \boldsymbol{aux}, \mathcal{D}, \mathcal{D}'))], \quad (3)$$

where the expectation is taken over the output distribution $\boldsymbol{o} \sim \mathcal{A}(\boldsymbol{aux}, \mathcal{D})$ and $c(\boldsymbol{o}|\mathcal{A}, \boldsymbol{aux}, \mathcal{D}, \mathcal{D}')$ is the privacy loss.

III. Improved Model Inversion Attack

While the original MIA has gained success on simple neural networks such as Softmax regression and Multilayer perceptron network (MLP) [3], it has limited success on deep neural networks only with auxiliary training data [6] or with adversarial training [25]. For more complex models, MIA tends to produce images that look unrealistic even with the denoising and sharpening filter [3]. In this section, we propose new regularization terms to enhance the optimization used in MIA to produce more recognizable images. We demonstrate that the enhanced MIA can be effective against deep learning models with more complex network structures.

ℓ_1-Norm Regularization. ℓ_1-norm regularization can be used to enforce sparsity on the solution vector, or reconstructed image $\mathbf{x}$. The sparsity constraint reduces and limits the intensity of pixels which are not important in leading the model to output the target class label. Therefore, it can help with removing noise and enhancing the contrast of the output image $\mathbf{x}$, especially with black and white images. The loss function of MIA with ℓ_1-norm regularization on image $\mathbf{x}$ becomes:

$$c(\mathbf{x}) = 1 - f_{label}(\mathbf{x}) + \lambda \|\mathbf{x}\|_1 \quad (4)$$

where the coefficient λ controls the penalty effect caused by ℓ_1-norm regularization.

BTV Regularization. While ℓ_1-norm regularizer may achieve noise removal and contrast enhancement on black and white images with a clear contrast, this benefit may be limited on gray scale images. For such images, we propose to use the bilateral total variation (BTV) regularization [26]:

$$R_{BTV} = \sum_{l=-p}^{p} \sum_{\substack{m=0 \\ m+l \geq 0}}^{p} \alpha^{m+l} \left\| \mathbf{x} - S_x^l S_y^m \mathbf{x} \right\|_1 \quad (5)$$

The BTV regularizer is essentially the accumulation of differences between central pixels and their neighborhoods within the spatial window size measured by p. It helps to maintain the main image features and preserve sharp edges when performing the super-resolution reconstruction task where the goal is to recover a single high-resolution image from a set of low-resolution images [26], [27]. The loss function of MIA with BTV regularization becomes:

$$c(\mathbf{x}) = 1 - f_{label}(\mathbf{x}) + \lambda R_{BTV} \quad (6)$$

Enhanced MIA Algorithm. Algorithms 1 outlines our enhanced MIA with the new regularizers. Line 4 uses an optional change-of-variable for the optimization by introducing a "box constraint" [28] to ensure that the value of each pixel in the reconstructed image stays in the range $[0, 1]$: $\mathbf{x} = \frac{1}{2}(\tanh(\mathbf{w}) + 1)$. The optimization is then implemented over $\mathbf{w}$. If no change-of-variable is used, we can directly set $\mathbf{x} = \mathbf{w}$. We use Adam optimizer instead of SGD used in the original MIA [3], which uses the moving average of the first and second moments of gradients (line 6 and 7) to scale the learning rate adaptively.

Algorithm 1: Improved MIA Algorithm

Input: *label*, T, β_1, β_2, τ, η, λ, the target model f.

1 Initialize variables $\mathbf{w}$, $\mathbf{m}$, and $\mathbf{v}$ to be zeros with the same size as training images of f.
2 Define $c(\mathbf{x})$ using eq.(4) or (6)
3 **for** $t = 1 \cdots T$ **do**
4 $\quad \mathbf{x}_{t-1} = \frac{1}{2}(\tanh(\mathbf{w}_{t-1}) + 1)$
5 $\quad \mathbf{g}_t = \nabla c(\mathbf{x}_{t-1})$
6 $\quad \mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$
7 $\quad \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$
8 $\quad \mathbf{w}_t = \mathbf{w}_{t-1} - \eta \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t} + \tau}$
9 **end**
10 **return** $\mathbf{x}_T = \frac{1}{2}(\tanh(\mathbf{w}_T) + 1)$

Fig. 1: MIA on MNIST dataset.

Fig. 2: MIA on Faces94 dataset.

Visual Results. Figure 1 shows the reconstructed images of the original and enhanced MIA (using ℓ_1 norm and change-of-variable) against a CNN model trained on the MNIST dataset in comparison to sample original images. We set the parameter values as $T = 5000$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\tau = 10^{-8}$, $\eta = 0.1$, and $\lambda = 0.05$. We can observe that the enhanced MIA generates more realistic and similar images to the original images than the original MIA. Figure 2 shows the reconstructed images of the original and enhanced MIA (using BTV regularization) against a softmax classifier trained on the Faces94 dataset (see Section V for details). We set $p = 2$, $\alpha = 0.9$, and $\lambda = 0.001$ for the BTV regularization. We observe that the reconstructed face images by enhanced MIA preserve sharper edges and corners, and less blur compared to original MIA.

Attack Success Metric. To quantify the results of reconstruction besides visual inspection, we define an attack success metric, *MIA distance*, as the minimum distance between the reconstructed image and all the training images in the target

TABLE I: MIA distance for MNIST dataset

Class (Digit)	0	1	2	3	4	5	6	7	8	9
Original MIA	8.02	8.36	9.14	9.50	8.35	9.76	9.64	8.38	8.79	8.93
Enhanced MIA	**5.93**	**7.89**	**8.51**	**7.92**	**8.2**	**8.41**	**8.51**	**7.76**	**8.06**	**8.76**

TABLE II: MIA distance for Faces94 dataset

Class (Person)	1	2	3	4	5	6
Original MIA	0.707	0.663	0.64	0.681	0.668	0.707
Enhanced MIA	**0.691**	**0.643**	**0.626**	**0.664**	**0.65**	**0.692**

class. In contrast to the average distance which measures the distance between the reconstructed image and the "average" image of the target class, we use minimum distance because it represents the worst case scenario. A smaller distance indicates the recovered image is more similar to the original training data, suggesting a more successful attack. A larger distance means that the attack is less successful and the model is more robust. We will also use this metric to evaluate the model's robustness against MIA.

Different distance metrics can be used depending on the data. For example, for the MNIST dataset which includes simple white and black images, we can use Euclidean distance which is shown in Table I. We observe that the reconstructed images from enhanced MIA have a consistently smaller distance than the original MIA. For the gray-scale face images, we adopt the structural similarity index (SSIM) [29], which is more suitable for measuring the perceptual similarity between two face images than Euclidean distance [30]. The value range of SSIM is [0, 1] where 1 indicates the most similar. Table II shows the minimum distance (1-SSIM) between the reconstructed images and training images in the target class for both original MIA and enhanced MIA. We can observe that enhanced MIA achieves better results.

IV. Class and Subclass Differential Privacy

In this section, we propose class-DP and subclass-DP as quantifiable privacy notions against MIA and corresponding privacy algorithms to achieve them.

A. Class-Level Differential Privacy

Definition of Class-Level DP. Our main goal is to provide a rigorous and strong privacy notion that can quantify the protection against MIA which targets the statistical property of a given class corresponding to a set of records in the training data. Intuitively, our secret to be protected is the statistical properties or features of a target class. Motivated by this, we propose class-level DP which defines the neighboring databases as two datasets differing in one class (i.e. all records that belong to the same class). Class-level DP guarantees that the resulting models are indistinguishable even if all the records in any one class are substituted. Therefore, MIA can not reconstruct a representative image of any target class.

Definition 4. *(Class Neighboring Datasets). Let $\mathcal{D}$ denote a dataset with K classes of records. The class neighboring datasets to $\mathcal{D}$ are the datasets $\mathcal{D}'$ that can be obtained from $\mathcal{D}$ by replacing all the records in an arbitrary class $k \in 1, ..., K$.*

Compared to the definition of neighboring datasets [9] in record-DP, a pair of class neighboring datasets differ in one class of data, which indicates they have the same number of classes and all of those classes are the same (same data and labels) except one. For example, let $\mathcal{D}$ be a hand-written digit dataset containing images of digits from 0 to 9, the class number of $\mathcal{D}$ is 10. Replacing all images of digit 0 with images of letter *a* in $\mathcal{D}$ forms a class neighboring dataset $\mathcal{D}'$.

Definition 5. *(Class-Level Differential Privacy). A randomized algorithm $\mathcal{A}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ satisfies class-level (ϵ, δ)-differential privacy if for all $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{A})$ and for all class neighboring datasets $\mathcal{D}$, $\mathcal{D}' \in \mathbb{N}^{|\mathcal{X}|}$:*

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta.$$

Considering record-DP can be unbounded (neighboring datasets are formed by adding or removing one record) or bounded (replacing one record or removing and then adding one record), our class-DP definition here is bounded, i.e. the class neighboring datasets are formed by replacing an entire class. We can also define the unbounded class-DP which requires the indistinguishability of the resulting model regardless of whether an entire class is present or not in the training data. We will show in our privacy analysis later all proofs can be derived similarly with these two versions of DP with the only difference being a constant factor.

Comparison with Record and Group-DP. Class-DP is a strong privacy guarantee. It protects not only one data record in private datasets like in record-DP [9] but also all other records which share common patterns or follow the same distribution with that record in the same class. Class-DP is different from group-DP [9] which ensures the indistinguishability of the statistical output regardless of the presence or absence of any group of a given size of m. While bearing some similarities, class-DP is not equivalent to group-DP even if we assume all classes have the same size m. This is because the neighboring pairs in class-DP differ in one class, and the classes are only a subset of all possible groups of size m. We can consider class-DP (with the same class size m) as a weaker version of group-DP, but specifically designed to protect against MIA. In addition, class-DP allows groups of different sizes which are determined by the size of each class and hence provide more precise protection against MIA.

We can potentially adopt group-DP to protect against MIA by setting the group size as the largest class size. However, doing so will lead to amplified perturbation by the group size and hence unacceptable model accuracy. In fact, any (ϵ, δ)-DP mechanism *M* satisfies $(m\epsilon, m\delta)$-group-DP for group size m with no necessary change to the private training process. This amplifying factor m can be very large and will render the model not useful with a meaningful privacy guarantee.

B. Algorithm for Class-DP

Algorithm 2 outlines the steps to achieve class-DP for deep learning models based on class-based sampling. Suppose the training dataset $\mathcal{D} = \{C_1, \cdots, C_K\}$ contains K classes of

data. During each step of the SGD, each class is sampled with probability q (line 3). The data of all selected classes will be used in the current step of SGD for calculating gradients and updating parameters. Dividing the noisy sum of the clipped gradient by the number of selected classes for the current SGD step approximates the average update of all classes while preventing the information of a single class from leakage.

Algorithm 2: Class-level differentially private SGD

Input: Training dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$, learning rate η_t, noise scale σ, gradient norm bound S, sampling ratio q.

1 Initialize θ_0 randomly.
2 **for** $t = 1 \cdots T$ **do**
3 Sample each class with probability q.
4 **for** *each selected class* $C_i (i = 1, \cdots, k_t)$ **do**
5 For each $x_j \in C_i$, compute $\mathbf{g}_t(x_j) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_j)$
6 Average gradients within class C_i, $g^{(i)} \leftarrow \frac{1}{|C_i|}\sum_{j=1}^{|C_i|} g_t(x_j)$ **%%Compute gradient**
7 $\overline{g}^{(i)} \leftarrow g^{(i)} / \max\left(1, \frac{\|g^{(i)}\|_2}{S}\right)$ **%%Clip gradient**
8 **end**
9 $\tilde{g}_t \leftarrow \frac{1}{qK}\left(\sum_{i=1}^{k_t}\overline{g}^{(i)} + \mathcal{N}\left(0, \sigma^2 S^2 \mathbf{I}\right)\right)$ **%%Add noise**
10 $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$ **%%Update**
11 **end**
12 **return** θ_T and overall privacy budget (ϵ, δ) computed by moments accountant with sampling.

Privacy Analysis. We analyze the privacy of Algorithm 2 by extending the moments accountant technique in Definition 3 to the class-DP setting. There are two key differences between the setting in [13] and ours: 1) [13] considers record-level DP while ours is class-level; 2) the neighboring concept in [13] is random "in or out" of a record while ours is a random substitution of a class. Due to such discrepancy, our analysis deviates from theirs. To begin with, we recall the following two lemmas from [13]. Lemma 1 facilitates the composition of the moments accountant of an iterative algorithm. Lemma 2 provides the translation of moments accountant to (ϵ, δ)-DP.

Lemma 1. *(Composibility [13]) Let mechanism $\mathcal{A}$ be a composition of a sequence of adaptive mechanisms $\mathcal{A}_1, ..., \mathcal{A}_T$, where $\mathcal{A}_t : \prod_{i=1}^{t-1} Range(\mathcal{A}_i) \times \mathcal{D} \to Range(\mathcal{A}_t)$. For any κ, it gives $\alpha_A(\kappa) \le \sum_{t=1}^{T} \alpha_{\mathcal{A}_t}(\kappa)$.*

Lemma 2. *(Tail Bound [13]) For any $\epsilon > 0$, the mechanism $\mathcal{A}$ is (ϵ, δ)-DP for $\delta = \min_\kappa exp(\alpha_\mathcal{A} - \kappa\epsilon)$.*

We also develop the following Lemma 3 that will be used in our main DP result in Theorem 1. It adapts Theorem 2 in [13], where the main difference is to replace $\boldsymbol{\mu_1} \sim \mathcal{N}(1, \sigma^2)$ there to $\boldsymbol{\mu_2} \sim \mathcal{N}(2, \sigma^2)$ here, and quantify the new $\alpha_{\boldsymbol{\mu}_0, \boldsymbol{\mu}}(\kappa)$ accordingly. This is because for class-DP and subclass-DP, we prefer the neighboring dataset notion to be a random substitution of class/subclass rather than "in or out" of an arbitrary record.

Lemma 3. *Let $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_2}$ denote the probability density function of $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(2, \sigma^2)$ respectively. Let $\boldsymbol{\mu}$ be the mixture of $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_2}$: $\boldsymbol{\mu} = (1-q)\boldsymbol{\mu_0} + q\boldsymbol{\mu_2}$. Let $\alpha_{\boldsymbol{\mu}_0, \boldsymbol{\mu}}(\kappa) = \log\max(E_1, E_2)$, where $E_1 = \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu_0}}[(\frac{\boldsymbol{\mu_0}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^\kappa]$ and $E_2 = \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu}}[(\frac{\boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu_0}(\boldsymbol{z})})^\kappa]$. Suppose $q < \frac{1}{16\sigma}$ and $\kappa \le \sigma^2 \ln\frac{1}{q\sigma}$, then it gives $\alpha_{\boldsymbol{\mu}_0, \boldsymbol{\mu}}(\kappa) \le \frac{4q^2\kappa(\kappa+1)}{(1-q)\sigma^2} + O(q^3/\sigma^3)$.*

Proof: We follow the proof of Theorem 2 in [13] with an emphasize on the difference part with [13]. Let $\alpha = \frac{4q^2\kappa(\kappa+1)}{(1-q)\sigma^2} + O(q^3/\sigma^3)$. To prove $\alpha_{\boldsymbol{\mu}_0, \boldsymbol{\mu}}(\kappa) = \log\max(E_1, E_2) \le \alpha$, we need to prove $E_1 = \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu_0}}[(\frac{\boldsymbol{\mu_0}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^\kappa] \le 1+\alpha$; $E_2 = \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu}}[(\frac{\boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu_0}(\boldsymbol{z})})^\kappa] \le 1+\alpha$, so that $\alpha_{\boldsymbol{\mu}_0, \boldsymbol{\mu}}(\kappa) \le \log(1+\alpha) \le \alpha$. Following [13], both inequalities can be proved by the same method. For any distributions ν_a and ν_b, $\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\nu_a}}[(\frac{\boldsymbol{\nu_a}(\boldsymbol{z})}{\boldsymbol{\nu_b}(\boldsymbol{z})})^\kappa] = \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\nu_b}}[(\frac{\boldsymbol{\nu_a}(\boldsymbol{z})}{\boldsymbol{\nu_b}(\boldsymbol{z})})^{\kappa+1}]$, where the latter can be expanded using binomial expansion,

$$\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\nu_b}}[(\frac{\boldsymbol{\nu_a}(\boldsymbol{z})}{\boldsymbol{\nu_b}(\boldsymbol{z})})^{\kappa+1}] = \sum_{i=0}^{\kappa+1}\binom{\kappa+1}{i}\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\nu_b}}[(\frac{\boldsymbol{\nu_a}(\boldsymbol{z}) - \boldsymbol{\nu_b}(\boldsymbol{z})}{\boldsymbol{\nu_b}(\boldsymbol{z})})^i].$$

Substituting $(\nu_a, \nu_b) = (\mu, \mu_0)$ and $(\nu_a, \nu_b) = (\mu_0, \mu)$ in, when $i = 0$, the first term is 1; when $i = 1$, the second term is 0. In the following, we calculate the third term with the more difficult case $(\nu_a, \nu_b) = (\mu_0, \mu)$, i.e., $i = 2$, which starts to deviate from [13]:

$$\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu}}[(\frac{\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^2] = q^2\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu}}[(\frac{\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu_2}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^2] \overset{(i)}{\le} \frac{q^2}{1-q}$$
$$\int_{-\infty}^{+\infty}\frac{(\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu_2}(\boldsymbol{z}))^2}{\boldsymbol{\mu_0}(\boldsymbol{z})}d\boldsymbol{z} = \frac{q^2}{1-q}\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu_0}}[\frac{(\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu_2}(\boldsymbol{z}))^2}{\boldsymbol{\mu_0}(\boldsymbol{z})}],$$

where (i) is by $\boldsymbol{\mu} \ge (1-q)\boldsymbol{\mu_0}$ and the above can be further bounded by calculating the last expectation:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu_0}}[\frac{(\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu_2}(\boldsymbol{z}))^2}{\boldsymbol{\mu_0}(\boldsymbol{z})}] &= \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu_0}}[(1 - exp(\frac{4\boldsymbol{z}-4}{2\sigma^2}))^2] \\ &= exp(\frac{4}{\sigma^2}) - 1 \le \frac{4}{\sigma^2}.\end{aligned} \tag{7}$$

The third term can be bounded as

$$\binom{\kappa+1}{2}\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu}}[(\frac{\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^2] \le \frac{4q^2\kappa(\kappa+1)}{(1-q)\sigma^2}. \tag{8}$$

In the following, we show the terms from $i = 3, ...$ are dominated by $i = 3$ term which is of order $O(\frac{q^3\kappa^3}{\sigma^3})$.

$$\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{\mu}}[(\frac{\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^i] \le \overbrace{\int_{-\infty}^{0}\boldsymbol{\mu}(\boldsymbol{z})|(\frac{\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^i|d\boldsymbol{z}}^{(I)} +$$
$$\overbrace{\int_{0}^{2}\boldsymbol{\mu}(\boldsymbol{z})|\frac{(\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu}(\boldsymbol{z}))^i}{\boldsymbol{\mu}(\boldsymbol{z})}|d\boldsymbol{z}}^{(II)} + \overbrace{\int_{2}^{+\infty}\boldsymbol{\mu}(\boldsymbol{z})|(\frac{\boldsymbol{\mu_0}(\boldsymbol{z}) - \boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\mu}(\boldsymbol{z})})^i|d\boldsymbol{z}}^{(III)}.$$

$$(I) \le \frac{2^i q^i}{(1-q)^{i-1}\sigma^{2i}}\int_{-\infty}^{0}\boldsymbol{\mu_0}(\boldsymbol{z})|\boldsymbol{z}-1|^i d\boldsymbol{z} \le \frac{(4q)^i(i-1)!!}{2(1-q)^{i-1}\sigma^i}.$$

$$(II) \le \frac{q^i}{(1-q)^i}\int_{0}^{2}\boldsymbol{\mu}(\boldsymbol{z})\frac{4^i}{\sigma^{2i}}d\boldsymbol{z} \le \frac{(4q)^i}{(1-q)^i\sigma^{2i}}$$

$$(III) \le \frac{q^i}{(1-q)^{i-1}\sigma^{2i}}\int_{2}^{+\infty}\boldsymbol{\mu_0}(\boldsymbol{z})(\frac{2\boldsymbol{\mu_2}(\boldsymbol{z})}{\boldsymbol{\mu_0}(\boldsymbol{z})})^i d\boldsymbol{z},$$

which is 2^i factor larger the estimation in [13]. Together, the $i \ge 3$ terms are dominated by $i = 3$ term with

order $O(\frac{q^3\kappa^3}{\sigma^3})$. In sum, we have proved that $\alpha_{\boldsymbol{\mu}_0,\boldsymbol{\mu}}(\kappa) \leq \frac{4q^2\kappa(\kappa+1)}{(1-q)\sigma^2} + O(q^3/\sigma^3)$. ∎

Theorem 1. *Let* $\sigma^2 = \frac{16q^2T\ln(\frac{1}{\delta})}{\epsilon^2}$ *and* $q < \sqrt{\frac{\epsilon}{64\sqrt{T\ln(1/\delta)}}}$. *Algorithm 2 satisfies* (ϵ, δ)*-class-DP.*

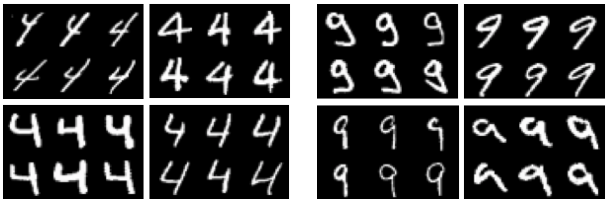
Proof: Let $\mathcal{A}_t(\mathcal{D}) := \sum_{i\in[k_t]} \frac{1}{qK}\left(\sum_{i=1}^{k_t} \overline{g}^{(i)} + \mathcal{N}\left(0, \sigma^2 S^2 \mathbf{I}\right)\right)$, where each $\overline{g}^{(i)}$ is the S-clipped gradient computed based on the sampled class C_i and satisfies $\|\overline{g}^{(i)}\|_2 \leq S$. First, we upper bound $\alpha_{\mathcal{A}_t}(\kappa)$. For class neighboring datasets $(\mathcal{D}, \mathcal{D}')$, Without loss of generality, let $\mathcal{D} = \{C_1, ..., C_{K-1}, C_K\}$ and $\mathcal{D}' = \{C_1, ..., C_{K-1}, C'_K\}$, where each C_k, $k = 1, ..., K$, denotes all the data (records and label) in class k. The distribution of $\mathcal{A}_t(\mathcal{D}') \sim \mathcal{N}(\frac{1}{qK}\sum_{i=1}^{k_t} \overline{g}^{(i)}_{|\mathcal{D}'}, \frac{1}{(qK)^2}\sum_{i=1}^{k_t}\sigma^2 S^2 \boldsymbol{I})$, where $\overline{g}^{(i)}_{|\mathcal{D}'}$ denotes the stochastic gradient computed on $\mathcal{D}'$. It is equivalent to $\mathcal{A}_t(\mathcal{D}') \sim \sum_{i=1}^{k_t}\frac{1}{qK}\left(\overline{g}^{(i)}_{|\mathcal{D}'} + S\cdot\boldsymbol{\mu}_0\right)$, with $\boldsymbol{\mu}_0 \sim \mathcal{N}(0, \sigma^2)$, where $\overline{g}^{(i)}_{|\mathcal{D}'}$ is the clipped gradient computed based on $\mathcal{D}'$. For $\mathcal{A}_t(\mathcal{D})$, depending on whether the K-th class is sampled or not, the mean of $\mathcal{A}_t(\mathcal{D})$ is $\{(1-q)\sum_{i=1}^{k_t}\overline{g}^{(i)}_{|\mathcal{D}'}\} + \{q(\sum_{i=1}^{k_t}\overline{g}^{(i)}_{|\mathcal{D}'} - \overline{g}^{(K)}_{|C'_K} + \overline{g}^{(K)}_{|C_K})\}$. The $\arg\max$ in eq.(3) is achieved when $\|\overline{g}^{(K)}_{|C^{K'}} - \overline{g}^{(K)}_{|C^K}\|_2 = 2S$, which gives $\mathcal{A}_t(\mathcal{D}) \sim \frac{1}{qK}\sum_{i=1}^{k_t}\overline{g}^{(i)}_{|\mathcal{D}'} + S\cdot((1-q)\boldsymbol{\mu}_0 + q\boldsymbol{\mu}_2)$, with $\boldsymbol{\mu}_2 \sim \mathcal{N}(2, \sigma^2)$. Thus, to bound $\alpha_{\mathcal{A}_t}(\kappa)$, it suffices to estimate $\alpha_{\boldsymbol{\mu}_0,\boldsymbol{\mu}}(\kappa)$ which is given in Lemma 3. With the composition property in Lemma 1, we have $\alpha_{\mathcal{A}}(\kappa) \leq \sum_{t=1}^{T}\alpha_{\mathcal{A}_t}(\kappa) \leq \frac{4Tq^2\kappa^2}{\sigma^2}$. By Lemma 2, to ensure (ϵ, δ)-class-DP, it suffices to ensure $\frac{4Tq^2\kappa^2}{\sigma^2} \leq \frac{\kappa\epsilon}{2}$, $exp(-\frac{\kappa\epsilon}{2}) \leq \delta$. In addition, since having used Lemma 3, we need to satisfy its constraints: $q < \frac{1}{16\sigma}$, $\kappa \leq \sigma^2\log(\frac{1}{q\sigma})$. With our choice of q and σ, we can verify that the above constraints hold. Finally, Algorithm 2 is (ϵ, δ)-class-DP. ∎

Remark 1. *For unbounded class-DP (i.e. random deletion of a class), we can still provide similar privacy guarantee by following similar procedure as the proof for the bounded class-DP case above: we ensure that Algorithm 2 is* (ϵ, δ)*-class DP if* $\sigma^2 = \frac{4q^2T\ln(\frac{1}{\delta})}{\epsilon^2}$ *and* $q < \sqrt{\frac{\epsilon}{32\sqrt{T\ln(1/\delta)}}}$.

C. Subclass-Level Differential Privacy

While class-DP provides strong protection against MIA, it may require a large amount of noise when the number of classes is small, i.e. class size is large. Recall that in our privacy analysis we utilize moments accountant to provide tight privacy loss analysis for our class-DP-SGD algorithm. Moments accountant itself depends heavily on the privacy amplification via random sampling with a sampling ratio of q. The smaller the q, the better the amplification and the smaller the privacy loss. For datasets where the number of classes is small comparing to the number of data records, i.e. q will be large, achieving meaningful class-DP while preserving model accuracy may not be feasible.

Definition of Subclass-DP. To address this, we propose subclass-DP that defines the neighboring databases based on a subclass, a predefined subset of records within a single class. In many practical applications, there exist natural subclasses within a large class. Subclass-DP ensures the indistinguishability of the output model with respect to any subclass. It can be considered as a weaker version of class-DP. We show that it will allow better and customizable privacy and utility tradeoff.



(a) Class of digit 4 (b) Class of digit 9

Fig. 3: Examples of Subclasses from the MNIST dataset.

Consider the image classification tasks that we focus on in this paper, images with the same label in the dataset often exhibit different sub-patterns. For example, as shown in Figure 3 where images are from the MNIST dataset, each class of digit images can be naturally divided into different groups, and images within a group are more similar to each other than those from other groups.

Definition 6. *(Subclass Neighboring Datasets). Let* $\mathcal{D}$ *denote a dataset with* K *subclasses of records. The subclass neighboring datasets to* $\mathcal{D}$ *are datasets* $\mathcal{D}'$ *that can be obtained from* $\mathcal{D}$ *by replacing all the records in an arbitrary subclass* $k \in 1, ..., K$.

Definition 7. *(Subclass-Level Differential Privacy). A randomized algorithm* $\mathcal{A}$ *with domain* $\mathbb{N}^{|\mathcal{X}|}$ *satisfies subclass level* (ϵ, δ)*-differential privacy if for all* $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{A})$ *and for all subclass neighboring datasets* $\mathcal{D}$, $\mathcal{D}' \in \mathbb{N}^{|\mathcal{X}|}$*:*

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon}\Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta.$$

Subclass-DP can be considered as a generalized notion of record-DP and class-DP. When the number of subclasses in each class k_{sub} is 1, it is equivalent to class-DP. When k_{sub} is the number of records in each class, it is equivalent to record-DP. The relationship between subclass-DP and group-DP is the same as that between class-DP and group-DP, except that subclass-DP corresponds to smaller group size.

Algorithm for Subclass-DP. The algorithm for Subclass-DP is the same as the class-DP algorithm (Algorithm 2) except that we sample random subclasses instead of random class (line 3) and the average gradient is computed within each sampled subclass (line 7). This additional subclass-based sampling provides additional privacy amplification which promises better privacy and utility tradeoff. In this paper, we form subclasses using k-means clustering algorithm with a predefined number of clusters k_{sub} to mimic the natural subclasses.

Privacy Analysis. The privacy analysis for the subclass-DP algorithm is inherited from Theorem 1 and it's proof by switching the class-level related notion to the subclass-level ones. We summarize the subclass-DP guarantee in the following corollary while omitting its detailed proof.

Corollary 1. *Let* $\sigma^2 = \frac{16q^2T\ln(\frac{1}{\delta})}{\epsilon^2}$ *and* $q < \sqrt{\frac{\epsilon}{64\sqrt{T\ln(1/\delta)}}}$. *Algorithm 2 with subclass sampling is* (ϵ, δ)*-subclass differentially private.*

V. Experiments

We evaluate the proposed class and subclass DP-SGD algorithms on MNIST [31] and Faces94[1] to demonstrate their effectiveness in defending against MIA while preserving good model utility. MNIST contains 60,000/10,000 training/test examples which are gray-scale handwritten digit images with the size 28×28. Faces94 is a facial image dataset with 153 individuals and each has 20 color facial images with the size 180×200. We convert color images into grayscale and rescale them to 60×70. We use image augmentation techniques[2] to create additional facial images such that each individual has 220 images. We then randomly divide the training/test set into 190/30. We use a vanilla model without privacy protection and a model with record-DP as baseline comparisons.

A. MNIST

A convolutional neural network (CNN) with two convolution layers followed by two fully connected layers is used. We train it without DP protection as the **vanilla model** and the test accuracy reaches 98.9%. We train the same CNN models with record-DP using the DP-SGD method proposed in [13] as the **record-DP model**. We use three choices of noise scale for the Gaussian noise which are $\sigma = 0.65, 1.0, 1.8$, and obtain three models with test accuracy of 96%, 93%, and 91% and corresponding privacy loss of $(6, 10^{-5})$, $(1.6, 10^{-5})$, and $(0.5, 10^{-5})$-DP respectively. Finally, we train the same CNN model using the subclass-DP-SGD algorithm (Algorithm 2 with random subclass sampling) as the **subclass-DP model**. The reason we use subclass-DP instead of class-DP is that the number of classes is small for the MNIST dataset which will make class-DP not meaningful. We will evaluate class-DP on the Faces94 dataset later in the section. Each class of digits in the MNIST dataset is divided into 50 subclasses using the k-means clustering algorithm (k=50). We set the subclass sampling ratio of q to be 0.2. We also use three choices of noise scale which are $\sigma = 1.6, 3.0, 3.65$. We choose a fixed gradient norm bound 3.0. We obtain three models with test accuracy of 96%, 93%, and 91% and corresponding privacy loss of $(18.6, 10^{-3})$, $(8.2, 10^{-3})$, and $(7, 10^{-3})$-subclass-DP respectively. Note that our criteria for the three subclass-DP models are to have matching accuracy with the three record-DP models. This way, we can have a fair comparison for each pair of record-level DP model and subclass-DP model at the same level of model accuracy in terms of their robustness to MIA.

MIA. We implement the improved MIA in Algorithm 1 with the same parameter setting and evaluate it against all models. The parameters of MIA are set as: $T = 5000$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\tau = 10^{-8}$, $\eta = 0.1$, and $\lambda = 0.05$. Figure 4 and 5 demonstrate the MIA results on the record-DP models and subclass-DP models respectively. The first row of each figure shows the ground truth training image samples of digit 0 to 9, and the second row shows the reconstructed images of MIA on the vanilla model. The third/fourth/fifth row of Figure 4 and 5 show the reconstructed images of MIA on the record-DP model and subclass-DP model for model accuracy at 96%, 93%, and 91% respectively. Notice that the record-DP model and the subclass-DP model at the same row in Figure 4 and 5 have the same model accuracy so we can have a fair comparison of their robustness. Comparing Figure 4 and Figure 5, we can see that record-DP can not defend against MIA. Even with small ϵ (row 5) which corresponds to $(0.5, 10^{-5})$-DP, MIA can still reconstruct the corresponding digits. On the other hand, subclass-DP provides strong protection against MIA which fails to reconstruct original training data representatives.

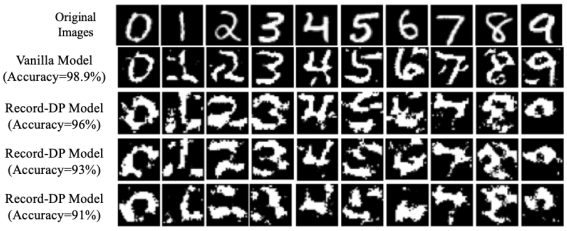


Fig. 4: MIA results on vanilla model and record-DP models trained on MNIST.

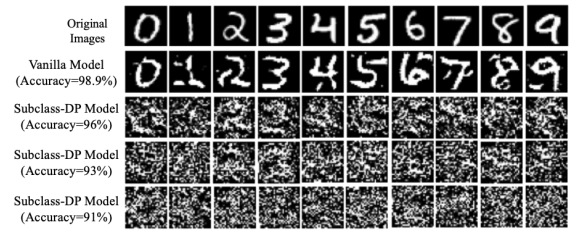
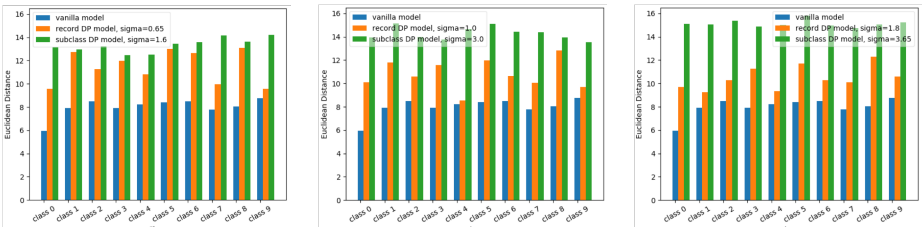


Fig. 5: MIA results on vanilla model and subclass-DP models trained on MNIST.

MIA Robustness. Figure 6 shows MIA robustness of subclass DP in comparison with vanilla model and record-level DP model in terms of the MIA distance (minimum Euclidean distance between the reconstructed image and training images in the target class) as defined in Section 3. We can see that record-DP models provide some protection against MIA compared to the vanilla model. Comparing the three figures, we observe that the subclass-DP models have stronger MIA robustness (larger distance) than the record-DP models at the same level of model accuracy for all classes, providing more effective protection against MIA (a better MIA robustness and accuracy tradeoff).

Figure 7 (8) shows (a) the relationship between ϵ and model utility measured by test accuracy, and (b) the relationship between ϵ and MIA robustness for record-DP (subclass-DP) models trained on MNIST with different noise scales and all other hyperparameters fixed. We note that the absolute value of epsilon and their comparison between record-DP and subclass-DP are not very meaningful. Instead, our goal is to adjust the epsilon for the two models to achieve the same range

[1]https://cswww.essex.ac.uk/mv/allfaces/faces94.html
[2]https://github.com/aleju/imgaug



(a) test accuracy=96% (b) test accuracy=93% (c) test accuracy=91%

Fig. 6: MIA Robustness of record-DP and subclass-DP models trained on MNIST with different model utility.

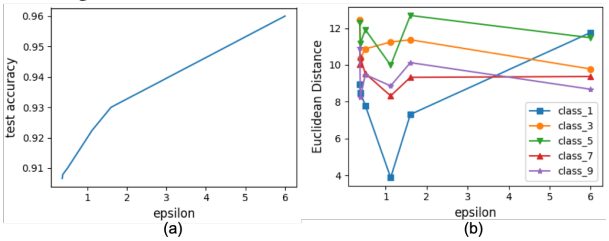


Fig. 7: Record-DP models trained on MNIST: (a) ϵ vs test accuracy (b) ϵ vs MIA robustness

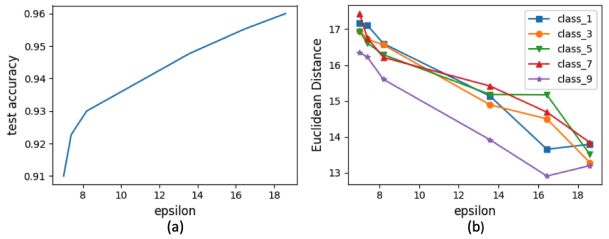


Fig. 8: Subclass-DP models trained on MNIST: (a) ϵ vs test accuracy (b) ϵ vs MIA robustness

of accuracy so we can have a fair comparison of their MIA robustness (i.e. the tradeoff of accuracy and MIA robustness). In addition, what is important is whether the epsilon value correlates with the MIA robustness (i.e. provides quantifiable protection against MIA). By comparing Figure 7(b) and Figure 8(b), we make two observations. First, subclass-DP has a much larger MIA distance than record-DP at the same accuracy level, indicating a much stronger MIA robustness and accuracy trade-off. Second, the level of ϵ in record-DP models does not have any significant correlation with MIA robustness. On the other hand, the ϵ of subclass-DP models directly correlates with their MIA robustness, i.e. a smaller epsilon corresponds to more robustness (larger distance). Hence it validates our hypothesis that subclass-DP can provide a more effective and quantifiable measure against the model inversion risk.

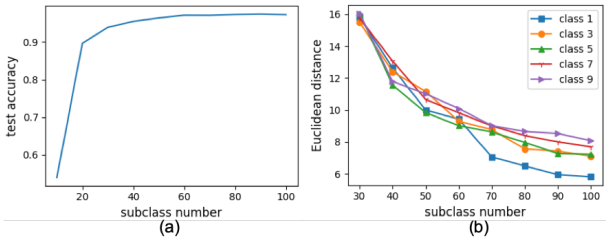


Fig. 9: (a) Model test accuracy vs number of subclasses k_{sub}. (b) MIA robustness vs number of subclasses k_{sub}

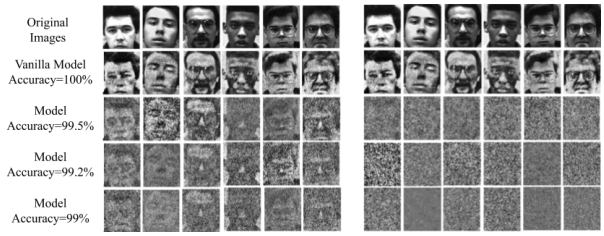


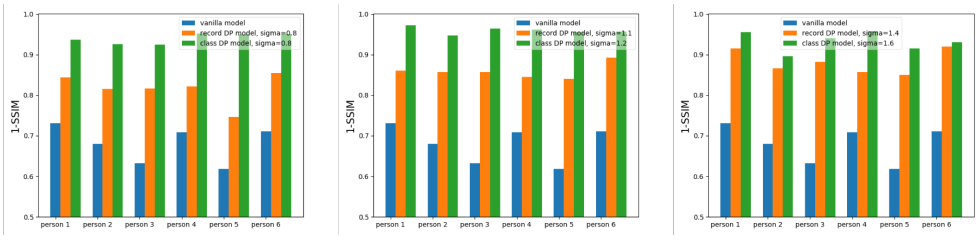
Fig. 10: MIA results on vanilla model and record-DP (left) and class-DP (right) models trained on Faces94 dataset.

Varying Number of Subclasses k_{sub}. Next, we study the impact of the number of subclasses k_{sub} within each class on subclass-DP models in terms of model utility and model robustness against MIA. Figure 9(a) shows the model test accuracy under different k_{sub} and (b) shows the MIA robustness under different k_{sub}. For both figures, the noise scale is fixed as 1.6 and the subclass sampling ratio is fixed as 0.2. We can observe that under the same noise scale and subclass sampling ratio, increasing k_{sub} will increase the model accuracy (becomes flat after k_{sub} is large enough), and decrease the MIA robustness. This utility and robustness trade-off is consistent with our definition of subclass-DP. When the k_{sub} is large enough, it will degrade to record-DP (when k_{sub} equals to the class size), and the model will be under higher risk of MIA.

B. Faces94

We use the softmax regression model as in [3] and train the **vanilla model** without privacy protecting with 100.0% test accuracy. We train **record-DP models** with the same structure using the DP-SGD algorithm [13]. We set three noise scales which are $\sigma = 0.8/1.1/1.4$ and obtain three models with 99.5%, 99.2%, and 99% test accuracy and corresponding $(9.1, 10^{-4})$, $(4.2, 10^{-4})$, and $(2.7, 10^{-4})$-DP respectively. Finally, we train **class-DP models** with the same architecture as the vanilla model using the class-DP-SGD method in Algorithm 2. The class sampling rate q is set to be 0.33. The gradient norm bound is 10. We choose three noise scales which are $\sigma = 0.8/1.2/1.6$ and obtain three models with 99.5%, 99.2%, and 99% test accuracy and corresponding $(62, 10^{-2})$, $(45.5, 10^{-2})$, and $(40.5, 10^{-2})$-class DP respectively.

MIA. We evaluate the improved MIA using Algorithm 1 with the loss function (6) where $\alpha = 0.9$ and $p = 2$. The parameter settings are the same for all the models to recover face images of each class (person). The parameters of MIA are set as: $T = 100$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\tau = 10^{-8}$, $\eta = 0.05$,



(a) test accuracy=99.5% (b) test accuracy=99.2% (c) test accuracy=99%

Fig. 11: MIA Robustness of record-DP and class-DP models trained on Faces94 dataset.

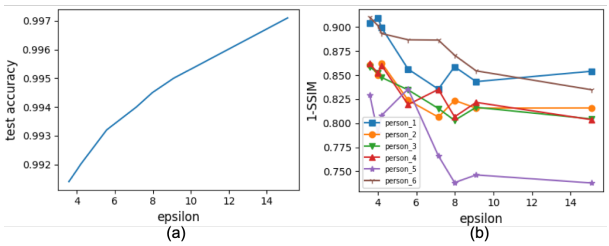


Fig. 12: Record-DP models trained on Faces94 dataset: (a) ϵ vs test accuracy (b) ϵ vs MIA robustness

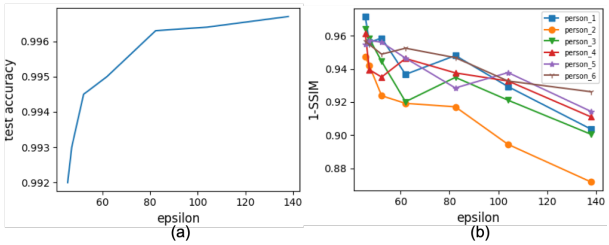


Fig. 13: Class-DP models trained on Faces94 dataset: (a) ϵ vs test accuracy (b) ϵ vs MIA robustness

and $\lambda = 0.001$. Figure 10 demonstrates the MIA results on the record-DP models and class-DP models respectively. Again, the record-DP model and the class-DP model in the same row have the same model accuracy so we can have a fair comparison. We can draw a similar conclusion to the MNIST dataset that the record-DP model can not prevent the reconstruction of the images but class-DP does while achieving the same model accuracy.

MIA Robustness. Figure 11 shows MIA robustness of class DP in comparison with the vanilla model and record-level DP model in terms of the MIA distance. We observe that the class-DP models have stronger MIA robustness than the record-DP models at the same level of model accuracy for all classes, providing more effective protection against MIA.

Figure 12 and 13 show model accuracy and MIA robustness with respect to varying ϵ for record-DP and class-DP respectively. Again, we emphasize that the ϵ value for class-DP may seem significantly large, the absolute value is not very meaningful. What is important is that at the same model accuracy level, class-DP achieves significantly larger MIA distance than record-DP, which means a much more effective MIA protection. Similar to MNIST, we also observe that the level of ϵ in record-DP models do not have a significant correlation with the robustness of the model against MIA, while the ϵ of class-DP models directly correlates with their robustness against MIA.

VI. Related Work

[3] first studied MIA targeting neural network models to recover recognizable facial images of individual's portrait by their names and white-box access to the model parameters. [5] trained an inversion model using an auxiliary dataset composed of the adversary's background knowledge to recover the private dataset. Their attack is different from our improved model inversion attack since they require an auxiliary dataset in order to train an additional model to implement the inversion attack. Our MIA requires no additional datasets and models. However, how to use auxiliary datasets to further improve MIA to generate more recognizable images close to the original one is an interesting topic for further work. There are also works exploring MIA under the distributing setting [6]–[8], [32] which are orthogonal to our improved MIA in a centralized setting. [33] proposed to protect MIA by introducing a regularizer into the training loss to mitigate the mutual information between the model input and the prediction. However, their method does not provide a rigorous or quantifiable guarantee against MIA as we do in this paper.

Existing works on privacy-preserving deep learning mainly focus on achieving better privacy and model utility tradeoff and tighter privacy loss quantification. Various techniques have been developed for this purpose, including moments accountant [13], gradient perturbation with adaptive budget allocation [17], objective function perturbation [14], and private teacher-student knowledge transfer [15]. Despite better privacy-utility tradeoff achieved by the above works under the record-DP notion, there is limited understanding of how effective DP is in protecting against various privacy threats empirically [18], [20]. Whether DP or other privacy notions can provide meaningful protection against MIA without sacrificing model utility is still an open problem, which we focus on in this paper.

As for broadened DP notions, [34] protected "user-level" DP for user-partitioned data when training an LSTM language model with a strong DP guarantee. [35] proposed "client-level" DP, which can be achieved in the federated setting along with good model utility when the number of clients is large enough. The class-level DP coincides with user-level and client-level DP [34], [35] when one class corresponds to one user (client). However, class and user are two orthogonal concepts, e.g. one class may not directly correspond to one

user and each user can have data of multiple classes. In this sense, the class-level DP and user-level DP are not the same for most of the cases. In terms of the targeted problems, our purpose of defending against MIA is very different from [34], [35], which considers the privacy issues for federated training. An important contribution of our paper is to show that DP-based techniques can be applied against MIA with proper development, in spite of the previous unsuccessful attempt [6] which applied record-level DP straightforwardly.

VII. Conclusion

We study the problem of protecting deep learning models against MIA. We show that traditional record-DP for building private deep learning models does not provide effective and quantifiable protection against MIA. Further, we propose two new DP notions, class-DP and subclass-DP, and algorithms for protecting deep learning models against MIA. Experiments on two real datasets show that class or subclass-DP can effectively defend against MIA while preserving good model utility. While we focus on the centralized setting and neural networks in this paper, the class-DP and subclass-DP notions are generally applicable to other machine learning settings (e.g. collaborative setting) and models (e.g. decision trees) to protect against MIA, and we leave the evaluation of them as future works.

Acknowledgments

This work is partially supported by National Science Foundation under CNS-1952192 and IIS-1838200, National Institutes of Health (NIH) under UL1TR002378 and R01GM118609, and Air Force Office of Scientific Research (AFOSR) DDDAS Program under FA9550-12-1-0240.

References

[1] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, 2020.

[2] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM SIGSAC CCS*, 2015.

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (SP)*, 2017.

[5] Z. Yang, E.-C. Chang, and Z. Liang, "Adversarial neural network inversion via auxiliary knowledge alignment," *arXiv preprint arXiv:1902.08552*, 2019.

[6] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *ACM SIGSAC CCS*, 2017.

[7] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2512–2520.

[8] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *USENIX Security Symposium, 2020*.

[9] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, 2014.

[10] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *JMLR*, 2011.

[11] J. Lou and Y.-m. Cheung, "Uplink communication efficient differentially private sparse optimization with feature-wise distributed data," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] ——, "An uplink communication-efficient approach to featurewise distributed sparse optimization with differential privacy," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[13] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC CCS*, 2016.

[14] N. Phan, Y. Wang, X. Wu, and D. Dou, "Differential privacy preservation for deep auto-encoders: an application of human behavior prediction," in *AAAI*, 2016.

[15] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *ICLR*, 2017.

[16] J. Wang, J. Zhang, W. Bao, X. Zhu, B. Cao, and P. S. Yu, "Not just privacy: Improving performance of private deep learning in mobile cloud," in *KDD*, 2018.

[17] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 332–349.

[18] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 1895–1912.

[19] J. Lee and D. Kifer, "Concentrated differentially private gradient descent with adaptive per-iteration privacy budget," in *KDD*, 2018.

[20] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model." *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[21] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *ACM SIGSAC CCS*, 2017.

[22] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC CCS*, 2013.

[23] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *International Symposium on Privacy Enhancing Technologies Symposium*, 2013.

[24] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *FOCS*, 2010.

[25] F. A. Mejia, P. Gamble, Z. Hampel-Arias, M. Lomnitz, N. Lopatina, L. Tindall, and M. A. Barrios, "Robust or private? adversarial training makes models more vulnerable to privacy attacks," *arXiv preprint arXiv:1906.06449*, 2019.

[26] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, 2004.

[27] A. Laghrib, A. Hakim, and S. Raghay, "A combined total variation and bilateral filter approach for image robust super resolution," *EURASIP Journal on Image and Video Processing*, 2015.

[28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[30] Y. Kim, J.-H. Yoo, and K. Choi, "A motion and similarity-based fake detection method for biometric face recognition systems," *IEEE Transactions on Consumer Electronics*, 2011.

[31] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[32] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *NIPS*, 2019, pp. 14 774–14 784.

[33] T. Wang, Y. Zhang, and R. Jia, "Improving robustness to model inversion attacks via mutual information regularization," *arXiv preprint arXiv:2009.05241*, 2020.

[34] M. Brendan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *International conference on learning representations, Vancouver, BC, Canada*, vol. 30, 2018.

[35] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," in *NIPS Workshop*, 2017.