Adaptive Privacy Preserving Deep Learning Algorithms for Medical Data

Xinyue Zhang
University of Houston
xzhang67@uh.edu

Jiahao Ding
University of Houston
jding7@uh.edu

Maoqiang Wu
Guangdong University of Technology
maoqiang.wu@vip.163.com

Stephen T.C. Wong
Houston Methodist Hospital
STWong@houstonmethodist.org

Hien Van Nguyen
University of Houston
hvnguy35@central.uh.edu

Miao Pan
University of Houston
mpan2@uh.edu

Abstract

Deep learning holds a great promise of revolutionizing healthcare and medicine. Unfortunately, various inference attack models demonstrated that deep learning puts sensitive patient information at risk. The high capacity of deep neural networks is the main reason behind the privacy loss. In particular, patient information in the training data can be unintentionally memorized by a deep network. Adversarial parties can extract that information given the ability to access or query the network. In this paper, we propose a novel privacy-preserving mechanism for training deep neural networks. Our approach adds decaying Gaussian noise to the gradients at every training iteration. This is in contrast to the mainstream approach adopted by Google's TensorFlow Privacy, which employs the same noise scale in each step of the whole training process. Compared to existing methods, our proposed approach provides an explicit closed-form mathematical expression to approximately estimate the privacy loss. It is easy to compute and can be useful when the users would like to decide proper training time, noise scale, and sampling ratio during the planning phase. We provide extensive experimental results using one real-world medical dataset (chest radiographs from the CheXpert dataset) to validate the effectiveness of the proposed approach. The proposed differential privacy based deep learning model achieves significantly higher classification accuracy over the existing methods with the same privacy budget.

1. Introduction

Deep learning holds great promise in improving healthcare and medicine. Examples include but not limited to: i) deep neural networks have exceeded expert performance on referral recommendation of sight-threatening retinal diseases [6]; ii) convolutional neural networks trained with more than 100,000 radiographs have shown competitive diagnostic accuracy compared to six board-certified radiologists while being two orders of magnitude faster [28]. Accenture estimates that artificial intelligence, in which deep learning is a crucial component, could save the healthcare industry $150 billion annually by 2026. For deep neural networks to work well, they need to be trained with a large number of examples. Unfortunately, sensitive information, including patient images and electronic health records, can be reconstructed with high fidelity from deep neural networks using privacy attacks during or after the network training process. To make the situation worse, the common strategy of data anonymization is not safe enough because adversarial parties can re-identify individuals in anonymized datasets by combining the data with background information. A notable experiment shows that combining public anonymized medical records and voter registration records can successfully identify the personal health information of a former Massachusetts governor, which is called linkage attack [32].

There are several popular types of attacks, such as (i) attribute attacks [44] which infer sensitive pieces of information (e.g, whether a patient has cancer) given the patient's public record and the ability to query the machine learning model; (ii) membership inference attacks [30] whose goal is to find out if a patient record is in the pool of the data used to train the machine learning model; and (iii) model inversion attacks [14] which attempt to reconstruct the entire patient data given only access to an intermediate layer of the deep network. As the medical records contain patients' sensitive data, realizing the full potential of deep learning in healthcare requires an innovative approach for building and deploying deep neural networks without sacrificing patients' privacy.

Differential privacy (DP) [12] as a golden standard of privacy provides strong guarantees on the risk of compromising the sensitive users' data in machine learning ap-

plications. Intuitively, it works by adding random noise to the model parameters so that an adversary with arbitrary background knowledge cannot confidently conclude whether a users' data is used in training a model or not. There are many papers focusing on designs for differentially private machine learning algorithms including empirical risk minimization and deep neural networks. The approaches to achieve private empirical risk minimization mainly include: output perturbation [5, 35, 36, 10] (add DP noise to model parameter obtained after the training), objective perturbation (add DP noise to objective function) [5, 17, 11], and gradient perturbation [3, 37, 7, 41] (add DP noise to the gradient). Note that the output and objective perturbation methods require the (strong) convexity of the objective function, which makes them impossible to apply in deep learning problems. Hence, injecting differentially private noise into gradient is a proper way to obtain a private deep learning model. The first work employed gradient perturbation method to achieve differential privacy on deep learning is called differentially private stochastic gradient descent (DPSGD) algorithm [1], which is also adopted by Google's TensorFlow Privacy. Since the gradient norm is usually unbounded in deep network optimization, gradient perturbation can be used after manually clipping the gradients at each iteration. [1] utilized norm gradient clipping to bound the effects of an individual data sample on the gradients, which is required for generating noise in the gradient perturbation method. Then, the differentially private noise is injected into the clipped gradient. As we can update the gradient of each step differentially privately, it is guaranteed that the overall deep learning model is private. Although [1] utilizes the moments accountant method to achieve a tight analysis of the privacy loss over the large number of iterations, the classification performance of DPSGD is still far inferior to the original SGD.

In this paper, we aim to build an accurate deep learning model without compromising medical data privacy. To be specific, we first clip the gradient with l_2 norm and then inject linear decaying Gaussian noise to the gradient of each step. Our salient contributions are summarized as follows.

- We propose a novel adaptive differentially private deep learning algorithm to protect medical training data. Intuitively, compared with the DPSGD algorithm, the advantages of the proposed algorithm include: a) We carefully adjust the scale of noise in each iteration controlled by a decay rate to reduce the negative noise addition and guarantee the convergence property of deep learning algorithm; b) instead of using the moments accountant applied in DPSGD [1], we adopt the truncated concentrated differential privacy (tCDP), which provides a simple, explicit, and tight privacy bound analysis on adaptive noise injection while avoiding the numerical computation of log moments. Moreover, tCDP can provide privacy amplification via random sampling compared with zero concentrated differential Privacy (zCDP) [45].
- We evaluate the performance of the proposed adaptive DP deep learning algorithm based on real-world chest radiographs. As far as we know, this is the first work focusing on multi-label classification tasks on medical datasets. We compare the performance of the proposed model with DPSGD on the same privacy preserving level. Our extensive experimental results show that the convergence of the proposed model is faster and the accuracy is higher. Moreover, our hyperparameter settings may pave the way for further the application of differentially private deep learning in medical domains.

The rest of paper is organized as follows. In Section 3, we present the differential privacy background. In Section 4, we propose our adaptive differentially private deep networks. In Section 5, we analyze the performance evaluation with CheXpert dataset. We illustrate the privacy threats in machine learning and review the related work of privacy preserving machine learning methods in Section 2. Finally, we draw conclusions in Section 6.

2. Related Work

2.1. Privacy threats in machine learning

Many attack models have been proposed in the literature. The membership inference attack [30] is proposed to infer whether the training dataset consists of a specific data sample. Fredrikson et al. introduced model inversion attack in [14], where the adversary can reconstruct training samples with some known features and the access to the machine learning model. In [40], the authors proposed a power side-channel attack model to recover the input data. Tramer et al. proposed the model stealing attack [33], where the adversary only has the access to a target model but not has any other knowledge of the model, and aims to generate a model that has similar performance of a target model. Moreover, other works focus on inferring the hyperparameters of the learning model [25, 34].

2.2. Privacy preserving empirical risk minimization

Recently, many researchers focus on private empirical risk minimization (ERM) problems [42, 38]. In [2], the authors designed a differentially private algorithm for online linear optimization problems with optimal regret bounds. The authors in [39] investigated the relationships between learnability and stability and privacy and concluded that a problem is privately learnable only when existing a private algorithm that can asymptotically minimize the empirical risk. [20] proposed private incremental regression and a

private incremental ERM problem combining continual release to analyze the utility bound of several algorithms. In [18], the authors provided small excess risk in the generalized linear model with sampling based method for entropy regularized ERM. There are also some papers targeting at private ERM learning on high dimensional datasets. The authors in [21] provided differentially private algorithms for sparse regression problems in high-dimensional settings. Smith et al. [31] used an algorithm based on a sample efficiency test of stability to extend and improve the results. In [19], the authors introduced Gaussian width of the parameter space in the random projection to derive a risk bound by using a private compress learning method in ERM algorithms. In distributed machine learning, [7, 8, 9] proposed differentially private alternating direction method of multipliers (ADMM) algorithms with Gaussian mechanism.

2.3. Privacy preserving deep learning

As differential privacy can provide strong privacy guarantee, differentially private deep learning models have attracted enormous attentions. Abadi et al. [1] proposed the differentially private stochastic gradient descent (DPSGD) algorithm and adopted moments accountant (MA) to calculate the overall privacy budget. However, there is no closed-form mathematical expression to estimate privacy budget. In order to improve the utility of the DPSGD while preserving privacy, the authors in [45, 24] designed several adaptive differentially private deep learning models by allocating different privacy budgets to each iteration and employed zero-concentrated differential privacy (zCDP) to analyse the privacy loss during the training. The difference between this paper and our proposed model is the design of the decay function and the DP definition. In our work, we adopted tCDP for privacy bound analysis. The tCDP is the relaxation of zCDP, which can provide privacy amplification via random sampling compared with zCDP. In [26], the authors trained an ensemble teacher model by combining a set of teacher models, which are trained over disjoint training datasets and the author also trained the differentially private student model by querying the ensemble teacher to label public data. Furthermore, Xie et al. [43] and Zhang et al. [46] focused on achieving differential privacy on Generative Adversarial Nets (GAN). In [10, 46], the authors injected differentially private noises to the loss function based on the functional mechanism. However, none of these works provide utility guarantees for their algorithms.

3. Background on Differential Privacy

Differential privacy (DP) [12] that provides a strong standard privacy guarantee is being widely applied to many research areas. Basically, it is used to protect data providers' privacy when the statistical information of a database is publishing. Its wide acceptance is based on its merits of effectively protecting the data providers' privacy while publishing the statistical information of the databases. DP indicates that the participation of one patient in the training phase has an inconsiderable effect on the final deep network model. In our setting, we consider the medical image dataset can be represented as a dataset $D = \{x_i, y_i\}_{i=1}^m$. The randomized algorithm $\mathcal{M}$ is the algorithm used to train the deep neural network, and the parameter space (also called weights or coefficients) of deep neural network is denoted as $Range(\mathcal{M})$. The definition of DP is described as follows.

Definition 3.1 (**Differential Privacy).** A randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for any two adjacent datasets D and $\hat{D}$ that differ in only a single record, the absolute value of the privacy loss random variable of an output $o \in Range(\mathcal{M})$, $Z(o) = \log \frac{\Pr[\mathcal{M}(D)=o]}{\Pr[\mathcal{M}(\hat{D})=o]}$ is bounded by ϵ, with probability at least $1 - \delta$.

The privacy budget ϵ controls the privacy preservation level and δ is the broken probability, and if $\delta = 0$, the randomized algorithm $\mathcal{M}$ is said to have ϵ-differential privacy. A larger ϵ means lower privacy level, and implies that there is a higher possibility to distinguish the outputs of the randomized algorithm $\mathcal{M}$ with two different input datasets. Intuitively, smaller ϵ means higher privacy preservation level. δ is the broken probability. Differential privacy indicates that the participation of one patient in the training phase has an inconsiderable effect on the final deep network model.

A generic method of achieving (ϵ, δ)-differential privacy is Gaussian mechanism [13] that adds Gaussian noise, calibrated to the query function's sensitivity, to the output. The sensitivity captures the maximum difference of the query function by a single record in the worst case. We define the sensitivity as follows.

Definition 3.2 (**Sensitivity).** The sensitivity of a query function $f(\cdot)$ that takes as input a dataset D is defined as

$$\Delta_f = \max_{D,\hat{D}} \|f(D) - f(\hat{D})\|_2, \tag{1}$$

where D and $\hat{D}$ are any two neighboring datasets differing in at most one record.

In this paper, we consider the gradient perturbation method to provide privacy guarantee of deep neural network. Thus, the query function f is the gradient of deep neural network. We can easily enforce a specific sensitivity value Δ_f by clipping the L_2-norm of gradient value. Based on the definition of sensitivity, we show the Gaussian mechanism in the following theorem.

Theorem 1 (**Gaussian Mechanism [13]).** *For a query function $f : \mathcal{D} \to \mathcal{R}^d$ with sensitivity Δ_f, the Gaussian Mechanism that adds noise generated from the Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbb{I})$ to the output of f satisfies*

(ϵ, δ)-differential privacy, where $\epsilon, \delta \in (0, 1)$ and $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\Delta_f}{\epsilon}$.

When applying gradient perturbation method in training phase of deep neural network, due to the large number of iterations, the composition property of differential privacy is important to estimate the privacy loss. Hence, we adopt truncated concentrated differential privacy (tCDP) [4], a new relaxation of differential privacy, to provide sharper and tighter analysis on the privacy loss for multiple iterative computations compared to (ϵ, δ)-DP. The definition of tCDP is defined as follows.

Definition 3.3 (tCDP). For all $\tau \in (1, \omega)$, a randomized algorithm $\mathcal{M}$ is (ρ, ω)-tCDP if for any neighboring datasets D and $\hat{D}$ and all $\alpha > 1$, we have

$$D_\tau(\mathcal{M}(x)||\mathcal{M}(x')) \leq \rho\alpha, \tag{2}$$

where $D_\tau(\cdot||\cdot)$ is the the Rényi divergence of order τ defined as follows.

Given two distributions μ and ν on a Banach space $(\mathcal{Z}, \|\cdot\|)$, here, we consider the Rényi divergence distance between them:

Definition 3.4 (Rényi Divergence [29])**.** Let $1 < \alpha < \infty$ and μ, ν be measures with $\mu \ll \nu$. The Rényi divergence of order α between μ and ν is defined as

$$D_\tau(\mu\|\nu) \doteq \frac{1}{\alpha - 1} \ln \int \left(\frac{\mu(z)}{\nu(z)}\right)^\alpha \nu(z)\, \mathrm{d}z.$$

Here we follow the convention that $\frac{0}{0} = 0$. If $\mu \not\ll \nu$, we define the Rényi divergence to be ∞. Rényi divergence of orders $\alpha = 1, \infty$ is defined by continuity.

In this paper, we mainly utilize the following properties of tCDP, shown in [4].

Lemma 1. *The Gaussian mechanism, in Theorem 1, satisfies $(\Delta_f^2/(2\sigma^2), \infty)$-tCDP.*

Lemma 2. *If randomized mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy (ρ_1, ω_1)-tCDP, and (ρ_2, ω_2)-tCDP, their composition defined as $(\mathcal{M}_1, \mathcal{M}_2)$ is $(\rho_1 + \rho_2, \min(\omega_1, \omega_2))$-tCDP.*

Lemma 3. *If a randomized mechanism $\mathcal{M}$ satisfies (ρ, ω)-tCDP, then for any $\delta \geq 1/\exp((\omega - 1)^2\rho)$, $\mathcal{M}$ satisfies $(\rho + 2\sqrt{\rho\ln(1/\delta)}, \delta)$-differential privacy.*

Lemma 4. *If a randomized mechanism $\mathcal{M}$ satisfies (ρ, ω)-tCDP, then for any n-element dataset D, executing $\mathcal{M}$ on uniformly random sn entries ensures $(13s^2\rho, \log(1/s)/(4\rho))$-tCDP, with $\rho, s \in (0, 0.1]$, $\log(1/s) \geq 3\rho(2 + log(1/\rho))$ and $\omega \geq \log(1/s)/(2\rho)$.*

Lemma 1 connects the Gaussian mechanism to the new differential privacy definition, i.e., tCDP. It intuitively shows that by injecting the same Gaussian noise the differences between $(\epsilon, \delta)-$DP and (ρ, ω)-tCDP. Lemma 2 indicates the composition theorem of two randomized mechanisms under tCDP. Lemma 3 establishes the link between two differential privacy criteria, and Lemma 4 provides the privacy amplification via random sampling. These lemmas will serve as the basics for the proof of our adaptive random noise mechanism in the later section.

4. Adaptive Differentially Private Deep Networks

4.1. Threat model

Before presenting the adaptive differentially private deep learning model, we first describe the threat model. As DP can provide strong privacy guarantee, it is a worst-case notion of privacy. DP ensures that although attackers can have all information from the training dataset except one data sample, they still cannot get this data sample after launching attacks [12]. Specifically, in this work, we consider the white-box attack [15] where the adversary has the full knowledge of our deep networks, including their architectures and parameters. In other words, attackers can access to the published model instead of the training process. The goal of the proposed scheme is that even though the attackers have the ability to obtain other data samples in the training dataset, they cannot infer the target training data sample.

4.2. Privacy preserving deep learning model

We assume there are m training data samples and each data sample is denoted by $\{x_i, y_i\}$, where y_i is the label. The loss function of the training model with parameter w is defined as $L(w, x)$. The gradient of the loss function $\nabla L(w, x)$ is updated by stochastic gradient descent (SGD) during each iteration. In order to preserve privacy of the training data, the differentially private noise is supposed to add to the gradient in each iteration. Based on Theorem 1, when calculating how much noise needs to be injected into the gradient, it is supposed to have the sensitivity of the gradient, which is difficult to characterize. Therefore, we control the sensitivity by clipping the gradient in l_2 norm. With a clipping threshold C, we can replace the gradient g_t of each step by $\frac{1}{s}\sum_{i \in S_t}\left(\nabla L(w_t, x_i, y_i)/\max(1, \frac{||\nabla L(w_t)||_2}{C})\right)$, where s is the batch size. Then, we can add Gaussian noise to the clipped gradient. Consequently, each SGD step is considered as differentially private. Based on the composition theorem of differential privacy, the overall model is supposed to be differentially private with accumulated privacy budget.

When injecting Gaussian noise to the gradient, the privacy budget will be accumulated due to the iterations within

each epoch as described in Lemma 2. If the total privacy budget is certain, we need to allocate it to each training step. The noise scale of Gaussian mechanism is decided by the privacy budget allocated to each epoch, which influences the final model accuracy. Our purpose is to achieve better accuracy of the differentially private training model without compromising data privacy. Therefore, we propose the adaptive differentially private deep learning model, which is inspired by the adaptive learning rate strategy. During the practical training processes, the learning rate is recommended to be decreased instead of fixed, in order to improve the model performance. Hence, in the DP learning model, we propose to reduce the injected noise along with the training iterations. In other words, in order to increase the accuracy, it is intended to add smaller and smaller noise to the gradients through the training time. Therefore, we propose the adaptive differentially private deep networks by injecting linear decaying Gaussian noise to the gradient during the training.

The overall procedure of our mechanism is shown in Algorithm 1. Note that we adopt tCDP in our algorithm instead of approximately differential privacy since its composition property is more straightforward for our adaptive noise addition. In each iteration of our algorithm, a batch of examples S_t with size s is sampled from the training dataset, and the algorithm computes the gradient of the loss on the examples in the batch and uses the average in the gradient descent step. The gradient clipping bounds per-example gradients by l_2 norm clipping with a threshold C. After gradient clipping, the sensitivity of the average gradient is $\frac{2C}{s}$. We next inject *linear decaying Gaussian noise* to the gradients at every training iteration with a decay rate R. This is in contrast to the mainstream approach adopted by Google's TensorFlow Privacy, which employs the same noise scale in each step of the whole training process. Specifically, we apply the Gaussian mechanism to add random noise following $\mathcal{N}(0, \sigma_t^2\mathbb{I})$ distribution to the network's gradients. The noise variance varies with a linear decay model as $\sigma_{t+1}^2 = R\sigma_t^2$, where $R \in (0, 1)$. Moreover, by considering the privacy composition between iterations and privacy amplification by sampling, the privacy guarantee of Algorithm 1 is provided in the next section.

4.3. Privacy guarantee

We employ the composition theorem of Truncated Concentrated Differential Privacy (tCDP) to analyze the cumulative privacy loss of differentially private stochastic gradient descent (DPSGD), which was developed to accommodate a larger number of computations and provides a sharper and tighter analysis of privacy loss than the strong composition theorem of (ϵ, δ)-DP. One popular way to track the privacy loss of DPSGD is the Moments Accountant (MA) method [1], which is adopted by Google's TensorFlow Privacy. As for the proposed approach, a Gaussian mechanism with a linearly decaying variance is applied to DPSGD to improve the model accuracy.

Theorem 2. *Algorithm 1 provides (ϵ, δ)-differential privacy.*

Proof. Since the utilization of Gaussian Mechanism, each iteration is $\rho_t = 2C^2/(s^2\sigma_t^2)$-tCDP (Lemma 1). By Lemma 2 and Lemma 4[1], and the decay rate R of noise scale, we derive that the total privacy loss is $(\rho_{total}, \omega_{total})$-tCDP with

$$\rho_{total} = \frac{13(s/m)^2C^2(1-R^T)}{2\sigma_0^2(R^{T-1}-R^T)}, \tag{3}$$

$$\omega_{total} = \frac{\log(m/s)\sigma_0^2R^T}{2C^2}, \tag{4}$$

where s is the batch size, m is the total number of private training dataset, C is the clipping threshold. By utilizing Lemma 3, we can say that our algorithm satisfies $(\rho_{total} + 2\sqrt{\rho_{total}\log(1/\delta)}, \delta)$-DP, which means $\epsilon = \rho_{total} + 2\sqrt{\rho_{total}\log(1/\delta)}$. □

Algorithm 1 Adaptive Differentially Private Deep Learning

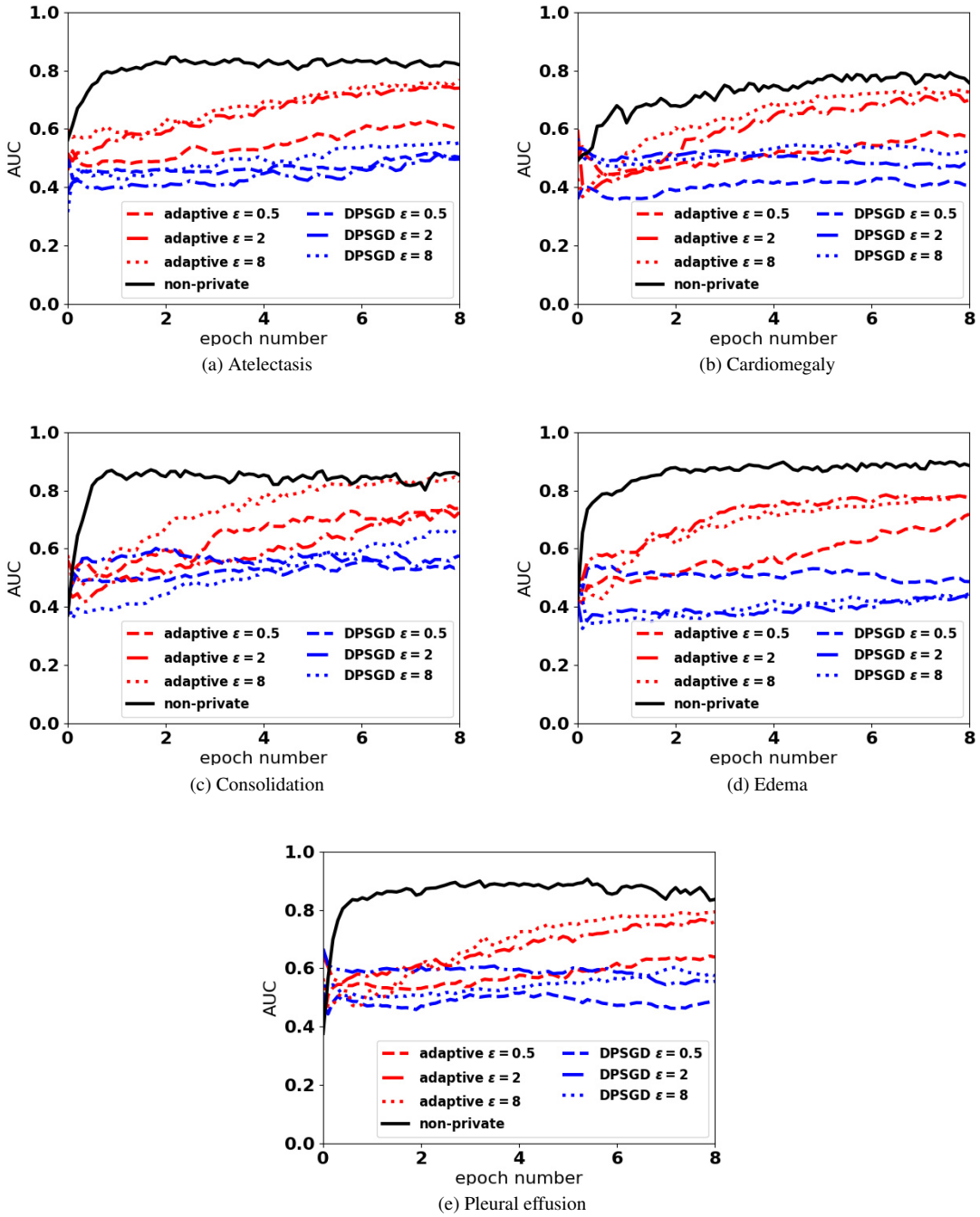
1: **Input:** Private training dataset $\{x_i, y_i\}_{i=1}^m$, loss function L, learning rate η, gradient norm bound C, decay rate $R \in (0, 1)$, batch size s.
2: **Output:** Differentially private model w_T.
3: Initialize w_0, σ_0^2.
4: $t = 0$.
5: **for** $t = 0, \cdots, T-1$ **do**
6: Randomly take a batch of data samples S_t from the training dataset with $|S_t| = s$.
7: Compute the gradient with gradient clipping $g_t = \frac{1}{s}\sum_{i\in S_t}\Big(\nabla L(w_t, x_i, y_i)/\max(1, \frac{||\nabla L(w_t)||_2}{C})\Big)$.
8: Add adaptive Gaussian noise $g_t = g_t + \mathcal{N}(0, \sigma_t^2\mathbb{I})$ with $\sigma_{t+1}^2 = R\sigma_t^2$.
9: Update the model parameter $w_{t+1} = w_t - \eta g_t$.
10: **end for**

Compared with MA, our proposed approach provides an explicit closed-form mathematical expression to approximately estimate the privacy loss. It is easy to compute and can be useful when the users would like to decide proper training time, noise scale, and sampling ratio during the planning phase.

5. Performance Evaluation

In this section, we demonstrate the experimental results of our proposed scheme on one medical dataset CheXpert, and two popular image datasets MNIST, and CIFAR10.

[1] Several conditions for privacy amplification via sampling (Lemma 4) are required.



(a) Atelectasis

(b) Cardiomegaly

(c) Consolidation

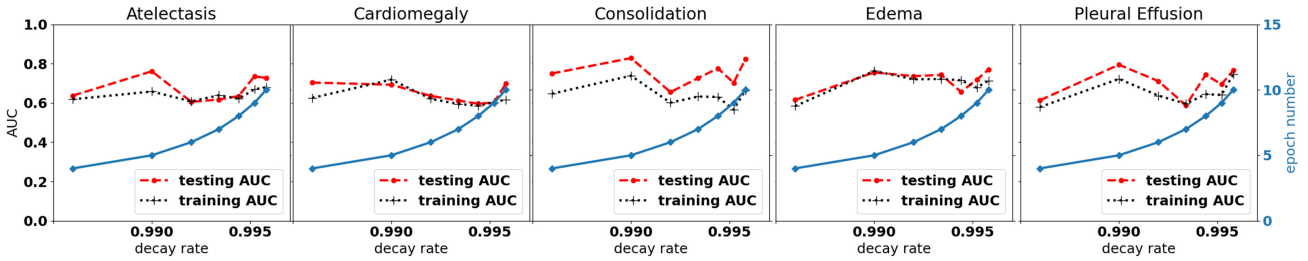
(d) Edema

(e) Pleural effusion

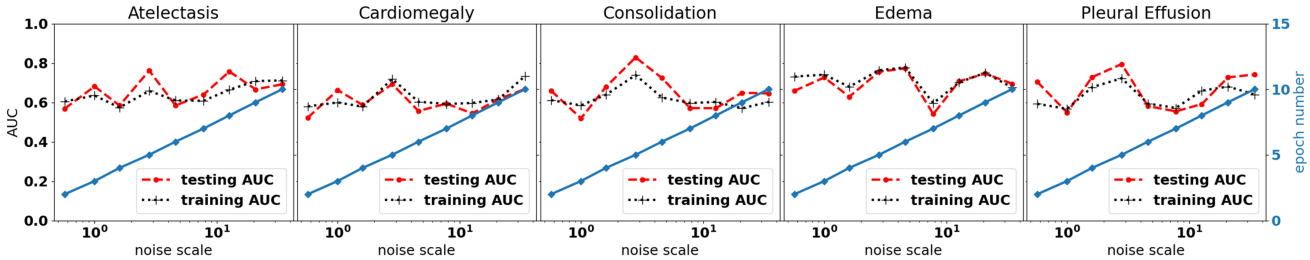
Figure 1: Comparison between non-private model, DPSGD and our adaptive DP model.

5.1. Experiment settings

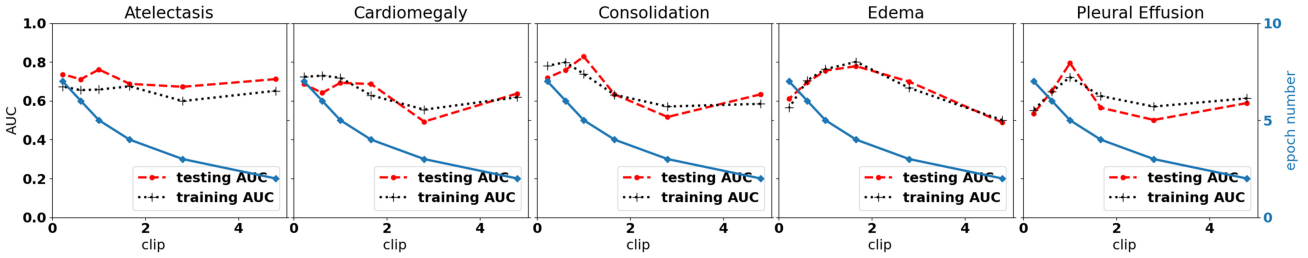
CheXpert. We conduct experiments on the CheXpert dataset [16], which is a large dataset containing 224,316 chest X-rays of 65,240 patients. There are 5 classes corresponding to different thoracic pathologies: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural Effusion. The images with size 320×320 pixels are fed into the pre-trained DenseNet-121. We only re-initialize the fully connected layer and fix the other convolutional layers, which will not have influences on the privacy leakage [1]. For illustrative purposes, we use 10,000 radiographs for training and 234 for testing. We use the test set provided by Stanford CheXpert to make our results com-



(a) decay rate (Fixed parameters: $\sigma = 2.8, C = 1, \delta = 10^{-4}$).



(b) initial noise scale (Fixed parameters: $R = 0.99, C = 1, \delta = 10^{-4}$).



(c) clip threshold (Fixed parameters: $R = 0.99, \sigma = 2.8, \delta = 10^{-4}$).

Figure 2: The impact of parameters on the training model performance (Red curve is tesing AUC, black curve is training AUC and the blue curve is epoch number).

parable to those in the literature. The test set is small because each sample requires manual annotations by 3 board-certified radiologists to create the ground truth label. For this reason, we cannot mix up noisy labels from the training set with well-validated labels from the test set [16].

Here, we introduce the default values of different parameters in the proposed adaptive differentially private deep learning model. We set the batch size s as 100, and the sample rate $\frac{s}{m} = 0.01$. We assume that the gradient norm clipping threshold C is 1, the initial noise scale σ_0 is 2.8, the noise decay rate R is 0.99, and the broken probability δ is 10^{-4}. Recall the analysis in Section 4.3, we can obtain the relationship between these parameters and the privacy with equations (3) and (4). As long as we have fixed the privacy budget ϵ, we can easily calculate the other parameters with these equations. We employ the area under the curve (AUC) to evaluate the non-private and private deep learning models.

MNIST The handwritten digits dataset [23] consists of 60000 training images and 10000 testing images, which are 28×28 gray scale images. We stack two convolutional layers with max-pooling and two fully-connected layers. Instead of using ReLU as the activation functions, we use tanh in the MNIST model as suggested in [27], which can provide better performance.

CIFAR10 The CIFAR10 [22] dataset consists of 10 classes 32×32 color images. There are 50000 training examples and 10000 testing examples. We use the pretrained ResNet18 as the training model for this dataset and re-initialize the fully connected layer. Then, we train the model with the proposed mechanism.

5.2. Experiment results

CheXpert. As introduced in section 5.1, there are 5 labels of each data sample in the dataset. Hence, we show five figures for each experiment. Firstly, we compare the test-

Dataset	Privacy Budget (ϵ, δ)	Accuracy	
		DPSGD	Proposed
MNIST	Non-Private	99%	
	(1.19, 10^{-5})	96.61%	97.7%
	(3.01, 10^{-5})	97.82%	98.07%
	(7.1, 10^{-5})	97.97%	98.17%
CIFAR10	Non-Private	88.67%	
	(3.02, 10^{-5})	77.16%	83.15%
	(7.03, 10^{-5})	81.42%	84.3%

Table 1: Summary results on MNIST and CIFAR10.

ing AUC of the proposed adaptive model with DPSGD and non-private model in Figure 1. In the experiment, we set the epoch number as 8. The privacy budget ϵ varies according to different initial noise scales in the DPSGD and the proposed adaptive DP model. The figure shows that with a higher epsilon value, the model accuracy is lower, since a larger epsilon means less noise injected to the gradient. We can also observe that our proposed adaptive DP model outperforms the DPSGD model for all thoracic conditions across different privacy budgets. More specifically, with the same privacy budget $\epsilon = 2$ and 8, the adaptive DP model can reach the average of 80% testing AUC for all of the five labels, while the DPSGD only can achieve approximately 60% testing AUC. We can conclude that with adaptive DP model, the performance will not drop too much and the patient's privacy is preserved.

In Figure 2, we demonstrate the impact of different parameter settings of the proposed adaptive differentially private deep learning model. We explore the influences of four parameters, noise decay rate R, initial noise scale σ_0, and clipping threshold C on the performance of the adaptive DP model. In the experiments, we keep the privacy budget fixed as $\epsilon = 8$. When the experiment is focusing on a specific parameter, we only vary values of this parameter and adjust the number of epochs to maintain the fixed privacy budget. In other words, only the discussed parameter and epoch number change in each experiment. With a larger noise decay rate, a higher initial noise scale, or a lower clipping threshold, it costs less privacy budget during each iteration. Therefore, we can achieve more epochs during training as shown as the blue solid curve in Figure 2. Moreover, we can observe that the model performance is better with a higher decay rate, a higher initial noise scale, and a smaller clip threshold, since the privacy budget spends slower as the number of training epochs increases.

MNIST and CIFAR10. We repeat the experiments on MNIST and CIFAR10 datasets and the experimental results are shown in Table 1. We compare the test accuracy by applying DPSGD and the proposed adaptive model. We first train the DPSGD model under a desired epoch number, keep the privacy budget ϵ value and calculate the parameters of the adaptive model with equations (3) and (4). For MNIST dataset, the test accuracy of non-private model can reach 99%. With the privacy budget ϵ equal to 1.19, 3.01, and 7.1, test accuracy of the proposed adaptive model is 1.09%, 0.25% and 0.2% higher than that of DPSGD. For CIFAR10 dataset, the non-private model can get to 88.67% test accuracy in 90 epochs. Compared with the DPSGD model, with the privacy budget ϵ of 3.02 and 7.03, test accuracy of the proposed adaptive model is promoted by 5.99% and 2.88%.

6. Conclusion

In this paper, we propose the adaptive differentially private deep learning model. Intuitively, we first clip the gradient to bound the sensitivity, inject differentially private noise with a specific decay rate based on the Gaussian mechanism into the clipped gradient, and update the gradient with SGD. The proposed algorithm is easy to implement and significantly improve the performances on various well-known datasets. Because of the large number of iterations in deep learning model, we adopt tCDP to obtain a tight bound of privacy leakage, since tCDP can provide a tighter and closed-form mathematical expression to estimate privacy budget compared with MA. Furthermore, tCDP can provide privacy amplification via random sampling compared with zCDP. We also conduct experiments on the public CheXpert dataset to verify the effectiveness of our adaptive differentially private deep learning model. We aim to explore the potential of adaptive differentially private deep learning applications in medicine. Moreover, we used the CheXpert that is a multi-label classification task. As far as we know, there are no works focusing on medical datasets with multi-label classification tasks.

Acknowledgement

The work of Xinyue Zhang, Jiahao Ding and Miao Pan is partly supported by the National Science Foundation (CNS-2029569). The work of Stephen Wong is partly supported by T.T. and W.F. Chao Foundation and John S. Dunn Research Foundation. The work of Hien Van Nguyen is partly supported by the National Science Foundation (1910973).

References

[1] 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 Conference on Computer and Communications Security*, Vienna, Austria, October 2016.

[2] Naman Agarwal and Karan Singh. The price of differential privacy for online learning. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, August 2017.

[3] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, Philadelphia, PA, October 2014.

[4] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.

[5] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, Mar 2011.

[6] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, and Daniel Visentin. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[7] Jiahao Ding, Sai Mounika Errapotu, Haijun Zhang, Miao Pan, and Zhu Han. Stochastic admm based distributed machine learning with differential privacy. In *International conference on security and privacy in communication systems*, Orlando, FL, October 2019.

[8] Jiahao Ding, Yanmin Gong, Chi Zhang, Miao Pan, and Zhu Han. Optimal differentially private ADMM for distributed machine learning. *CoRR*, abs/1901.02094, February 2019.

[9] Jiahao Ding, Jingyi Wang, Guannan Liang, Jinbo Bi, and Miao Pan. Towards plausible differentially private admm based distributed machine learning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 285–294, 2020.

[10] Jiahao Ding, Xinyue Zhang, Mingsong Chen, Kaiping Xue, Chi Zhang, and Miao Pan. Differentially private robust admm for distributed machine learning. In *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, December 2019.

[11] Jiahao Ding, Xinyue Zhang, Xiaohuan Li, Junyi Wang, Rong Yu, and Miao Pan. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, February 2020.

[12] Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, Xi'an, China, April 2008.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, New York, NY, March 2006.

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[15] Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 148–162, New York, NY, December 2019.

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, Honolulu, HI, January 2019.

[17] Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. In *International Conference on Machine Learning*, pages 118–126, 2013.

[18] Prateek Jain and Abhradeep Guha Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, June 2014.

[19] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *Proceedings of The 33rd International Conference on Machine Learning*, New York, NY, June 2016.

[20] Shiva Prasad Kasiviswanathan, Kobbi Nissim, and Hongxia Jin. Private incremental regression. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, Chicago, IL, May 2017.

[21] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, Edinburgh, Scotland, June 2012.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[23] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[24] Jaewoo Lee and Daniel Kifer. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, United Kingdom, July 2018.

[25] Seong Joon Oh, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144. Springer, 2019.

[26] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *The 5th International Conference on Learning Representations*, Toulon, France, April 2017.

[27] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *arXiv preprint arXiv:2007.14191*, 2020.

[28] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, and Curtis P Langlotz. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

[29] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions*

to the Theory of Statistics. The Regents of the University of California, 1961.

[30] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[31] Adam Smith and Abhradeep Guha Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Proceedings of the 26th Annual Conference on Learning Theory*, Princeton, NJ, June 2013.

[32] Latanya Sweeney. Only you, your doctor, and many others may know. `https://techscience.org/a/2015092903/`. Accessed: 2019-12-04.

[33] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

[34] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 36–52. IEEE, 2018.

[35] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535, Long Beach, CA, June 2019.

[36] Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1182–1189, Honolulu, HI, January 2019.

[37] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

[38] Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. Efficient privacy-preserving nonconvex optimization. *arXiv preprint arXiv:1910.13659*, 2019.

[39] Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, 17(183):1–40, 2016.

[40] Lingxiao Wei, Bo Luo, Yu Li, Yannan Liu, and Qiang Xu. I know what you see: Power side-channel attack on convolutional neural network accelerators. In *Proceedings of the 34th Annual Computer Security Applications Conference*, pages 393–406, San Juan, Puerto Rico, December 2018.

[41] Bingzhe Wu, Shiwan Zhao, Guangyu Sun, Xiaolu Zhang, Zhong Su, Caihong Zeng, and Zhihong Liu. P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019.

[42] Liyao Xiang, Jingbo Yang, and Baochun Li. Differentially-private deep learning from an optimization perspective. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 559–567, Paris, France, April 2019.

[43] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[44] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, Oxford, UK, July 2018.

[45] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 332–349, San Francisco, CA, May 2019.

[46] Xinyue Zhang, Jiahao Ding, Sai Mounika Errapotu, Xiaoxia Huang, Pan Li, and Miao Pan. Differentially private functional mechanism for generative adversarial networks. In *2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, December 2019.