# Tighter Generalization Bounds for Iterative Privacy-Preserving Algorithms

**Fengxiang He**[*1, 2] **Bohan Wang**[*1] **Dacheng Tao**[1, 2]

[1]School of Computer Science, the University of Sydney, Australia
[2]JD Explore Academy, JD.com, China

## Abstract

This paper studies the relationship between generalization and privacy preservation of machine learning in two steps. We first establish an alignment between the two facets for any learning algorithm. We prove that $(\varepsilon, \delta)$-differential privacy implies an on-average generalization bound for a multi-sample-set learning algorithm, which further leads to a high-probability bound for any learning algorithm. We then investigate how the iterative nature shared by most learning algorithms influences privacy preservation and further generalization. Three composition theorems are proved to approximate the differential privacy of an iterative algorithm through the differential privacy of its every iteration. Integrating the above two steps, we eventually deliver generalization bounds for iterative learning algorithms. Our results are strictly tighter than the existing works. Particularly, our generalization bounds do not rely on the model size which is prohibitively large in deep learning. Experiments of MLP, VGG, and ResNet on MNIST, CIFAR-10, and CIFAR-100 are in full agreement with our theory. The theory applies to a wide spectrum of learning algorithms. In this paper, it is applied to the Gaussian mechanism as an example.

## 1 INTRODUCTION

Generalization to unseen data and privacy preservation are two increasingly important facets of machine learning. Specifically, good generalization guarantees that an algorithm learns the underlying patterns in the training data rather than just memorizes the data [Vapnik, 2013, Mohri et al., 2018]. In this way, good generalization abilities provide confidence that the models trained on existing data can be applied to similar but unseen scenarios. Additionally, massive personal data has been collected, such as financial and medical records. How to discover the highly valuable population knowledge carried in the data while protecting the highly sensitive individual privacy has profound importance [Dwork and Roth, 2014, Pittaluga and Koppal, 2016].

This paper investigates the relationship between generalization and privacy preservation in machine learning algorithms by the following two steps: (1) exploring the relationship between generalization and privacy preservation in any learning algorithm; and (2) analyzing how the iterative nature shared by most learning algorithms would influence the privacy-preserving ability and further the generalizability.

We first prove two theorems that upper bound the generalization error of an learning algorithm via its differential privacy. Specifically, we prove a high-probability upper bound for the generalization error,

$$\mathrm{Gen}_{S,\mathcal{A}(S)} = \mathcal{R}_{\mathcal{D}}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)),$$

where $S$ is the training set sampled i.i.d. from some distribution $\mathcal{D}$, $\mathcal{A}(S)$ is the hypothesis learned by algorithm $\mathcal{A}$ on $S$, $\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S))$ is the expected risk, and $\hat{\mathcal{R}}_S(\mathcal{A}(S))$ is the empirical risk. This bound is established based on a novel on-average generalization bound for any $(\varepsilon, \delta)$-differentially private multi-sample-set learning algorithm. These results indicate that the algorithms with a good privacy-preserving ability also have a good generalizability. We, therefore, can expect to design novel learning algorithms for better generalizability by enhancing its privacy-preserving ability.

We then studied how the iterative nature shared by most learning algorithms influences the privacy-preserving ability. Generally, the privacy-preserving ability of an iterative algorithm degenerates along iterations, since the amount of leaked information cumulates when the algorithm is progressing. To capture this degenerative property, we further prove three composition theorems that calculate the differential privacy of any iterative algorithm via the differential privacy of its every iteration. Combining with the

[*]The authors contributed equally.

*Accepted for the 37th Conference on Uncertainty in Artificial Intelligence* (UAI 2021).

established relationship between generalization and privacy preservation, our composition theorems help characterizing the generalizabilities of iterative learning algorithms.

Our results considerably extend the current understanding of the relationship between generalization and privacy preservation in iterative learning algorithms.

Existing works [Dwork et al., 2015, Nissim and Stemmer, 2015, Oneto et al., 2017] have proved some high-probability generalization bounds in the following form,

$$\mathbb{P}\left(\left|\operatorname{Gen}_{S,\mathcal{A}(S)}\right| > a\right) < b, \tag{1}$$

where $a$ and $b$ are two positive constant real numbers. Our high-probability bound is strictly tighter than the current tightest results by [Nissim and Stemmer, 2015] which only holds for $\varepsilon \leq \frac{1}{10}$ from two aspects: (1) our bound tightens the term $a$ from $13\varepsilon$ to $4\varepsilon$ and the term $b$ from $\frac{2\delta}{\varepsilon}\log\left(\frac{2}{\varepsilon}\right)$ to $\frac{2e^{-1.7\varepsilon}\delta}{\varepsilon}\log\left(\frac{2}{\varepsilon}\right)$, and (2) our bound further cover the case when $\varepsilon > \frac{1}{10}$. These improvements are significant in practice because the factor $\varepsilon$ can be as large as 10 in the experiments by [Abadi et al., 2016]. Also, the bounds by [Nissim and Stemmer, 2015] are only for binary classification, while ours apply to any differentially private learning algorithm.

There are also existing literature on the differential privacy composition bound [Dwork and Roth, 2014, Kairouz et al., 2017], and the current state-of-art bound is given by [Kairouz et al., 2015] using "privacy region" technique. The approximation of factor $\delta$ in our composition theorems is tighter than that in [Kairouz et al., 2017] by

$$\delta\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\left(T-\left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil\right),$$

where $T$ is number of iterations, while the estimate of $\varepsilon'$ remains the same. This improvement is significant because the iteration number $T$ can be considerably large in practice. This helps our composition theorems to further tighten our generalization bounds for iterative learning algorithms considerably.

We trained MLPs on the MNIST dataset [LeCun, 1998], and VGG-16 [Simonyan and Zisserman, 2014] and ResNet-18 [He et al., 2016] on the CIFAR-10 and CIFAR-100 datasets [Krizhevsky et al., 2009]. Membership inference attack [Yeom et al., 2018, Shokri et al., 2017] is performed in every epoch. The membership inference attack accuracy and the generalization error (the difference between test error and training error) in every epoch are collected for evaluating the privacy-preserving ability and the generalizability, respectively. The collected (membership inference attack accuracy, generalization error) pairs, (membership inference attack accuracy, training time) pairs, and (generalization error, training time) pairs are divided into groups according to the neural architecture and the dataset. Spearman's rank order correlation test [Spearman, 1987] is performed to all groups. The correlation coefficients and $p$-values show three statistically significant correlations: (1) the positive correlation between the generalization and privacy preservation; (2) the negative correlation between the generalization and training time; and (3) the negative correlation between the privacy preservation and training time, which are in full agreement with out theory.

Our results apply to a wide spectrum of machine learning algorithms. This paper applies them to the iterative Gaussian mechanism with mini-batch (IGMM), which includes stochastic gradient Langevin dynamics [Welling and Teh, 2011] as an example of the stochastic gradient Markov chain Monte Carlo scheme [Ma et al., 2015] and agnostic federated learning [Geyer et al., 2017]. Our results deliver generalization bounds for agnostic federated learning. The obtained generalization bounds do not explicitly rely on the model size, which can be prohibitively large in modern methods, such as deep neural networks.

## 2 NOTATIONS AND PRELIMINARIES

Suppose $S = \{(x_1, y_1), \ldots, (x_N, y_N) | x_i \in \mathcal{X} \subset \mathbb{R}^{d_X}, y_i \in \mathcal{Y} \subset \mathbb{R}^{d_Y}, i = 1, \ldots, N\}$ is a training sample set, where $x_i$ is the $i$-th feature, $y_i$ is the corresponding label, and $d_X$ and $d_Y$ are the dimensions of the feature and the label, respectively. For the brevity, we define $z_i = (x_i, y_i)$. We also define random variables $Z = (X, Y)$, such that all $z_i = (x_i, y_i)$ are independent and identically distributed (i.i.d.) observations of the variable $Z = (X, Y) \in \mathcal{Z}$, $Z \sim \mathcal{D}$, where $\mathcal{D}$ is the data distribution. For any set $U$, we denote its boundary points as $\partial U$. For any function $f(x) : \mathbb{R} \to \mathbb{R}$, we write $g = \tilde{\mathcal{O}}(f)$, if there exists another function $h$, such that $g = fh$, and for any $\alpha > 0$, $\lim_{x\to\infty} \frac{h(x)}{x^{\alpha}} = 0$. We also write $g = \tilde{\Omega}(f)$ if $f = \tilde{\mathcal{O}}(g)$, and $g = \tilde{\Theta}(f)$ if both $f = \tilde{\mathcal{O}}(g)$ and $g = \tilde{\mathcal{O}}(f)$. We also use $\boldsymbol{p}$ as the probability density, with $\boldsymbol{p}^V$ and $\mathbb{P}^V$ respectively the probability density and the probability conditional on any random variable $V$.

A machine learning algorithm $\mathcal{A}$ learns a hypothesis,

$$\mathcal{A}(S) \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathcal{Y}\},$$

from the training sample $S \in \mathcal{Z}^N$. The expected risk $\mathcal{R}_S(\mathcal{A}(S))$ and empirical risk $\hat{\mathcal{R}}_{\mathcal{D}}(\mathcal{A}(S))$ of the algorithm $\mathcal{A}$ are defined as follows,

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S)) = \mathbb{E}_{z\sim\mathcal{D}}\ell(\mathcal{A}(S), z),$$

$$\hat{\mathcal{R}}_S(\mathcal{A}(S)) = \frac{1}{N}\sum_{i=1}^{N}\ell(\mathcal{A}(S), z_i),$$

where $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ is the loss function. It is worth noting that both the algorithm $\mathcal{A}$ and the training sample set $S$ can introduce randomness in the expected risk $\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S))$ and empirical risk $\hat{\mathcal{R}}_S(\mathcal{A}(S))$. The generalization error is defined as the difference between the expected risk and empirical risk,

$$\mathrm{Gen}_{S,\mathcal{A}(S)} \overset{\triangle}{=} \mathcal{R}_{\mathcal{D}}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)),$$

whose upper bound is called the generalization bound.

Differential privacy measures the ability to preserve privacy, which is defined as follows (cf. [Dwork and Roth, 2014]).

**Definition 1** (Differential Privacy)**.** *A stochastic algorithm $\mathcal{A}$ is called $(\varepsilon, \delta)$-differentially private if for any hypothesis subset $\mathcal{H}_0 \subset \mathcal{H}$ and any neighboring sample set pair $S$ and $S'$ which differ by only one example (called $S$ and $S'$ adjacent), we have*

$$\log\left[\frac{\mathbb{P}_{\mathcal{A}(S)}(\mathcal{A}(S) \in \mathcal{H}_0) - \delta}{\mathbb{P}_{\mathcal{A}(S')}(\mathcal{A}(S') \in \mathcal{H}_0)}\right] \le \varepsilon. \qquad (2)$$

*The algorithm $\mathcal{A}$ is also called $\varepsilon$-differentially private, if it is $(\varepsilon, 0)$-differentially private.*

Differential privacy measures the "worst case" distance between the hypothesis distributions, in the sense that for a $(\varepsilon, \delta)$ differential privacy preserving algorithm, $\varepsilon$ needs to be larger than the left hand of eq.(2) for any $\mathcal{H}_0$. We then introduce KL divergence, which measures the "average" distance between two distributions, and are helpful in approximating differential privacy (cf. [Kullback and Leibler, 1951]).

**Definition 2** (KL Divergence)**.** *Suppose two distributions $P(\cdot)$ and $Q(\cdot)$ are defined on the same support $\mathcal{H}$. Then the KL divergence between $Q$ and $P$ is defined as*

$$D_{KL}(Q\|P) = \int_{h\in\mathcal{H}} \left(\log\frac{\mathrm{d}Q(h)}{\mathrm{d}P(h)}\right) \mathrm{d}Q(h).$$

In this paper, we will slightly abuse the notations of distribution and its cumulative distribution function when no ambiguity is introduced, since there is a one-one mapping between them if zero-probability events are ignored.

# 3 GENERALIZATION BOUNDS FOR ITERATIVE DIFFERENTIALLY PRIVATE ALGORITHMS

This section establishes the generalizability of iterative differentially private algorithms. The establishment has two steps. We first establish generalization bounds for any differentially private learning algorithm. Then, we investigate how the iterative nature shared by most learning algorithms would influence the differential privacy and further the generalizability via three composition theorems. We also sketch the proofs for these results and demonstrate their advantages compared with existing literature.

## 3.1 BRIDGING GENERALIZATION AND PRIVACY PRESERVATION

We first prove a high-probability generalization bound for any $(\varepsilon, \delta)$-differentially private machine learning algorithm as follows.

**Theorem 1** (High-Probability Generalization Bound)**.** *Suppose algorithm $\mathcal{A}$ is $(\varepsilon, \delta)$-differentially private, the training sample size ($c_1, \cdots, c_7$ are some positive constants)*

$$N \ge \max\left\{\frac{c_1}{\varepsilon^2}\ln\left(\frac{c_2}{e^{-c_3\varepsilon}\delta}\right), \frac{c_4}{c_5(1-c_6e^{-\varepsilon})}\ln c_7e^{-\varepsilon}\delta\right\},$$

*and the loss function $\|\ell\|_\infty \le 1$. Then, for any data distribution $\mathcal{D}$ over data space $\mathcal{Z}$, eq.(1) holds with*

$$\begin{cases} a = 4\varepsilon, b = \dfrac{2e^{-1.7\varepsilon}\delta}{\varepsilon}\ln\left(\dfrac{2}{\varepsilon}\right) & , \varepsilon \le \dfrac{1}{5}; \\ a = 1.2(1-0.9e^{-\varepsilon}), b = \dfrac{18e^{-\varepsilon}\delta}{1-0.9e^{-\varepsilon}}\ln\left(\dfrac{220}{1-0.9e^{-\varepsilon}}\right) & , \varepsilon > \dfrac{1}{5}. \end{cases}$$

Theorem 1 demonstrates that a good privacy-preserving ability implies a good generalizability. Thus, we can unify the algorithm designing for enhancing privacy preservation and for improving generalization.

### 3.1.1 Proof Sketch

We now give the proof sketch for Theorem 1 (the details are deferred to Appendix A). The proofs have three stages: (1) we first prove an on-average generalization bound for multi-sample-set learning algorithms defined as below; (2) we then obtain a high-probability generalization bound for multi-sample-set algorithms; and (3) we eventually prove Theorem 1 by reduction to absurdity.

**Definition 3** (Multi-Sample-Set Learning Algorithms)**.** *Suppose the training sample set $S$ with size $kN$ is separated to $k$ sub-sample-sets $S_1, \ldots, S_k$, each of which has the size of $N$. In another word, $S$ is formed by $k$ sub-sample-sets as*

$$S = (S_1, \ldots, S_k).$$

*The hypothesis $\mathcal{B}(S)$ learned by multi-sample-set algorithm $\mathcal{B}$ on dataset $S$ is defined as follows,*

$$\mathcal{B} : \mathcal{Z}^{k\times N} \mapsto \mathcal{H} \times \{1, \ldots, k\},\ \mathcal{B}(S) = \left(h_{\mathcal{B}(S)}, i_{\mathcal{B}(S)}\right).$$

To obtain the high-probability generalization bound for a differentially private algorithm $\mathcal{A}$ on $N$ examples, we sample $KN$ examples and divide them into $K$ sub-sample-sets. A $(2\varepsilon, \delta)$-differentially private multi-sample-set algorithm $\mathcal{B}$ is constructed as (1) employing an $(\varepsilon, \delta)$-differentially private algorithm $\mathcal{A}$ to learn $K$ hypotheses from the $K$ sub-sample-sets; and (2) performing an $\varepsilon$-differentially private mechanism (e.g., exponential mechanism) to select the final hypothesis and its index (see Sec A.2). The generalization bound of algorithm $\mathcal{A}$ is then obtained from the bound of multi-sample-set algorithm $\mathcal{B}$ (see Lemma 1).

**Stage 1: Prove an on-average generalization bound for multi-sample-set learning algorithms.**

We first prove the following on-average generalization bound for multi-sample-set learning algorithms.

**Theorem 2** (On-Average Multi-Sample-Set Generalization Bound)**.** *Let multi-sample-set algorithm $\mathcal{B} : \mathcal{Z}^{k \times N} \mapsto \mathcal{H} \times \{1, \cdots, k\}$ be $(\varepsilon, \delta)$-differentially private and the loss function $\|\ell\|_\infty \leq 1$. Then, for any data distribution $\mathcal{D}$ over data space $\mathcal{Z}$, we have the following inequality,*

$$\left| \underset{S \sim \mathcal{D}^N, \mathcal{B}(S)}{\mathbb{E}} \left[ \mathrm{Gen}_{S_{i_{\mathcal{B}(S)}}, h_{\mathcal{B}(S)}} \right] \right| \leq e^{-\varepsilon} k \delta + 1 - e^{-\varepsilon}. \quad (3)$$

**Stage 2: Prove a high-probability generalization bound for multi-sample-set algorithms.**

Markov bound (cf. [Mohri et al., 2018], Theorem C.1) is an important concentration inequality in learning theory. Here, we slightly modify the original version as follows,

$$\mathbb{E}_x \left[ h(x) \right] \geq \mathbb{E}_x \left[ h(x) \mathbb{I}_{h(x) \geq g(x)} \right] \geq \mathbb{E}_x \left[ g(x) \mathbb{I}_{h(x) \geq g(x)} \right].$$

Then, combining it with Theorem 2, we derive the following high-probability generalization bound for multi-sample-set algorithms.

**Theorem 3** (High-Probability Multi-Sample-Set Generalization Bound)**.** *Let all the notations be as Theorem 2. Then, if $\varepsilon \leq \frac{17}{50}$, for any $k \leq \frac{\varepsilon}{1.7 e^{-1.7\varepsilon} \delta}$, we have*

$$\mathbb{P} \left( \mathrm{Gen}_{S_{i_{\mathcal{B}(S)}}, h_{\mathcal{B}(S)}} \leq k e^{-1.7\varepsilon} \delta + 1.7\varepsilon \right) \geq \frac{85\varepsilon}{127}.$$

*Otherwise, if $\varepsilon > \frac{17}{50}$, for any $k \leq \frac{1 - e^{-\varepsilon}}{10 e^{-\varepsilon} \delta}$,*

$$\mathbb{P} \left( \mathrm{Gen}_{S_{i_{\mathcal{B}(S)}}, h_{\mathcal{B}(S)}} \leq k e^{-\varepsilon} \delta + 1.1(1 - e^{-\varepsilon}) \right) \geq \frac{1 - e^{-\varepsilon}}{219}.$$

**Stage 3: Prove Theorem 1 by Reduction to Absurdity.**

We eventually prove Theorem 1 by *reduction to absurdity*. Assume there exists an algorithm $\mathcal{A}$ which conflicts with Theorem 1. We can then construct an algorithm $\mathcal{B}$ based on the exponential mechanism which is defined as follows (cf. [McSherry and Talwar, 2007]]).

**Definition 4** (Exponential Mechanism)**.** *Suppose that $S$ is a sample set, $\mathcal{I}$ is an index set, $\varepsilon$ is the privacy parameter, and $u : (S, r) \mapsto \mathbb{R}^+$ is a function with the sensitivity $\Delta u$ defined by*

$$\Delta u \triangleq \max_{i \in \mathcal{I}} \max_{S, S' \text{ adjacent}} |u(S, i) - u(S', i)|.$$

*Then, the exponential mechanism $\mathcal{E}(S, u, \mathcal{I}, \varepsilon)$ outputs an element $i \in \mathcal{I}$ with probability proportional to* $\exp(\frac{\varepsilon u(S,i)}{2\Delta u})$.

Then, we can prove the following lemma.

**Lemma 1.** *Suppose a positive integer $N$ satisfies that $N \geq \max\{\frac{2}{0.077\varepsilon^2} \ln\left(\frac{43}{254 e^{-1.7\varepsilon} \delta}\right), \frac{200}{\ln 0.9(1 - 0.9e^{-\varepsilon})} \ln \frac{9e^{-\varepsilon}\delta}{48180}\}$. If for an $(\varepsilon, \delta)$ differential privacy preserving algorithm $\mathcal{A} : \mathcal{Z}^N \to \mathcal{H}$, there is*

$$\mathbb{P} \left( \mathrm{Gen}_{S, \mathcal{A}(S)} \leq a \right) < \frac{b}{2}, \quad (4)$$

*for $a$ and $b$ defined as Theorem 1. Then if $\varepsilon < \frac{1}{5}$, there exists a multi-sample-set algorithm $\mathcal{B} : \mathcal{Z}^{k \times N} \to \mathcal{H} \times \{1, \cdots, k\}$ with $(1.7\varepsilon, \delta)$-differential privacy, $k = \frac{\varepsilon}{e^{-1.7\varepsilon} \delta}$, and*

$$\mathbb{P} \left( \mathrm{Gen}_{\tilde{S}_{i_{\mathcal{B}(\tilde{S})}}, h_{\mathcal{B}(\tilde{S})}} \leq k e^{-1.7\varepsilon} \delta + 1.7 \times 1.7\varepsilon \right) < \frac{85\varepsilon}{127}. \quad (5)$$

*If $\varepsilon \geq \frac{1}{5}$, there exists a multi-sample-set algorithm $\mathcal{B} : \mathcal{Z}^{k \times N} \to \mathcal{H} \times \{1, \cdots, k\}$ with $(\varepsilon - \ln(0.9), \delta)$-differential privacy, $k = \frac{1 - 0.9e^{-\varepsilon}}{9e^{-\varepsilon} \delta}$, and*

$$\mathbb{P} \left( \mathrm{Gen}_{\tilde{S}_{i_{\mathcal{B}(\tilde{S})}}, h_{\mathcal{B}(\tilde{S})}} \leq 0.9 k e^{-\varepsilon} \delta + 1.1(1 - 0.9e^{-\varepsilon}) \right) < \frac{1 - 0.9e^{-\varepsilon}}{219}. \quad (6)$$

Therefore, for any algorithm $\mathcal{A}$ on which Theorem 1 fails to hold, eq. (5) or eq.(6) in Lemma 1 will conflict with Theorem 3, which completes the proof of Theorem 1.

**Remark 1.** *Obtained bounds are all in the same form as eq.(1), where multi-sample-set algorithms help significantly decrease $b$ with a small increase of $a$. In this way, we obtain a tighter generalization bound.*

### 3.1.2 Comparison with Existing Results

This section compares our results with the existing works.

**Comparison of Theorem 1.** There have been several high-probability generalization bounds for $(\varepsilon, \delta)$-differentially private machine learning algorithms.

Dwork et al. [2015] proved that

$$\mathbb{P} \left[ \mathrm{Gen}_{S, \mathcal{A}(S)} < 4\varepsilon \right] > 1 - 8\delta^{\varepsilon}.$$

Oneto et al. [2017] proved that

$$\mathbb{P} \left[ \mathrm{Gen}_{S, \mathcal{A}(S)} < \sqrt{6 \hat{\mathcal{R}}_S(\mathcal{A}(S)) \hat{\varepsilon}} + 6\varepsilon^2 + \frac{6}{N} \right] > 1 - 3e^{-N\varepsilon^2},$$

$$\mathbb{P} \left[ \mathrm{Gen}_{S, \mathcal{A}(S)} < \sqrt{4 \hat{V}_S(\mathcal{A}(S)) \hat{\varepsilon}} + \frac{5N}{N-1} \left( \varepsilon^2 + 1/N \right) \right] > 1 - 3e^{-N\varepsilon^2},$$

where $\hat{\varepsilon} = \varepsilon + \sqrt{1/N}$, and $\hat{V}_S(\mathcal{A}(S))$ is the empirical variance of $\ell(\mathcal{A}(S), \cdot)$:

$$\hat{V}_S(\mathcal{A}(S)) = \mathrm{Cov}_{z \sim S}(\ell(\mathcal{A}(S), z)).$$

Nissim and Stemmer [2015] proved that

$$\mathbb{P}\left[|\operatorname{Gen}_{S,\mathcal{A}(S)}| < 13\varepsilon\right] > 1 - \frac{2\delta}{\varepsilon}\log\left(\frac{2}{\varepsilon}\right).$$

This is the existing tightest high-probability generalization bound in the literature. However, this bound only stands for binary classification problems with $\varepsilon \leq \frac{1}{10}$. By contrast, our high-probability generalization bound holds for any machine learning algorithm and stays valid for any $\varepsilon > 0$.

Also, our bound is strictly tighter. All the bounds, including ours, are in the same form as eq.(1). Apparently, a smaller $a$ and a smaller $b$ imply a tighter generalization bound. Our bound improves the current tightest result from two aspects:

- For $\varepsilon \leq \frac{1}{10}$, our bounds tightens the term $a$ from $13\varepsilon$ to $4\varepsilon$.
- For $\varepsilon \leq \frac{1}{10}$, our bounds tightens the term $b$ from $\frac{2\delta}{\varepsilon}\log\left(\frac{2}{\varepsilon}\right)$ to $\frac{2e^{-1.7\varepsilon}\delta}{\varepsilon}\log\left(\frac{2}{\varepsilon}\right)$.
- Our bounds further cover the case when $\varepsilon > \frac{1}{10}$.

These improvements are significant. Abadi et al. [2016] conducted experiments on the differential privacy in deep learning. Their empirical results demonstrate that the factor $\varepsilon$ can be as large as 10.

**Comparison of Theorem 2.** There is only one related work in the literature that presents an on-average generalization bound for multi-sample-set algorithm. Nissim and Stemmer [2015] proved that,

$$\left|\mathop{\mathbb{E}}_{S \sim \mathcal{D}^N, \mathcal{B}(S)}\left[\operatorname{Gen}_{S_{i_{\mathcal{B}(S)}}, h_{\mathcal{B}(S)}}\right]\right| \leq k\delta + 2\varepsilon.$$

Our bound is tighter by a factor of $\frac{e^{\varepsilon}}{2}$. According to the empirical results by [Abadi et al., 2016], this factor can be as large as $e^{10} \approx 20,000$. It is a significant multiplier for loss function. Furthermore, the result by [Nissim and Stemmer, 2015] stands only for binary classification, while our result applies to all differentially private learning algorithms.

### 3.2 HOW THE ITERATIVE NATURE CONTRIBUTES?

Most machine learning algorithms are iterative, which may degenerate the privacy-preserving ability along with iterations. This section studies the degenerative nature of the privacy preservation in iterative machine learning algorithms and its influence to the generalization.

We have the following composition theorem.

**Theorem 4** (Composition Theorem I). *Suppose an iterative machine learning algorithm $\mathcal{A}$ has $T$ steps: $\{W_i(S)\}_{i=0}^{T}$, where $W_i$ is the learned hypothesis after the $i$-th iteration. Suppose the $i$-th iterator*

$$\mathcal{M}_i : (W_{i-1}, S) \mapsto W_i$$

*is $(\varepsilon, \delta)$-differentially private. Then, the algorithm $\mathcal{A}$ is $(\varepsilon', \delta')$-differentially private, where $\varepsilon' = \min\{\varepsilon_1', \varepsilon_2', \varepsilon_3'\}$ with*

$$\begin{aligned}
\varepsilon_1' &= \sum_{i=1}^{T} \varepsilon_i, \\
\varepsilon_2' &= \sqrt{2\sum_{i=1}^{T}\varepsilon_i^2 \log\left(e + \frac{\sqrt{\sum_{i=1}^{T}\varepsilon_i^2}}{\tilde{\delta}}\right)} \\
&\quad + \sum_{i=1}^{T}\frac{(e^{\varepsilon_i}-1)\varepsilon_i}{e^{\varepsilon_i}+1}, \\
\varepsilon_3' &= \sum_{i=1}^{T}\frac{(e^{\varepsilon_i}-1)\varepsilon_i}{e^{\varepsilon_i}+1} + \sqrt{2\log\left(\frac{1}{\tilde{\delta}}\right)\sum_{i=1}^{T}\varepsilon_i^2},
\end{aligned}$$

*and $\tilde{\delta}$ is an arbitrary positive real constant.*

*Correspondingly, the factor $\delta'$ is defined as the maximal value of the following equation with respect to $\{\alpha_i\}_{i=1}^{T} \in I$,*

$$2 - \prod_{i=1}^{T}\left(1 - e^{\alpha_i}\frac{\delta_i}{1+e^{\varepsilon_i}}\right) - \prod_{i=1}^{T}\left(1 - \frac{\delta_i}{1+e^{\varepsilon_i}}\right) + \tilde{\delta}, \quad (7)$$

*where $I = \partial\{\sum_{i=1}^{T}\alpha_i = \varepsilon', \varepsilon_i \geq \alpha_i \geq 0\}$, and $\tilde{\delta}$ is the same real constant mentioned above.*

When all the iterations have the same privacy-preserving ability, we can tighten the approximation of the factor $\delta'$ as the following corollary.

**Corollary 1** (Composition Theorem II). *When all the iterations are $(\varepsilon, \delta)$-differential private, $\delta'$ is*

$$\begin{aligned}
\delta' =& 1 - \left(1 - e^{\varepsilon}\frac{\delta}{1+e^{\varepsilon}}\right)^{\left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil}\left(1 - \frac{\delta}{1+e^{\varepsilon}}\right)^{T-\left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil} \\
&+ 1 - \left(1 - \frac{\delta}{1+e^{\varepsilon}}\right)^{T} + \tilde{\delta} \\
=& \left(T - \left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil\right)\frac{2\delta}{1+e^{\varepsilon}} + \left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil\delta + \tilde{\delta} + \mathcal{O}\left(\left(\frac{\delta}{1+e^{\varepsilon}}\right)^2\right).
\end{aligned}$$

Here we make some explanation about the above corollary. The maximum of $\delta'$ is achieved when at most $T - \left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil$ elements $\alpha_i \neq 0$. We note that

$$(1-x)^n = 1 - nx + \mathcal{O}(x^2).$$

Then, the $\delta'$ in Theorem 4 can be estimated as

$$
\begin{aligned}
\delta' =& 1 - \left(1 - e^{\varepsilon}\frac{\delta}{1+e^{\varepsilon}}\right)^{\left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil}\left(1 - \frac{\delta}{1+e^{\varepsilon}}\right)^{T-\left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil} \\
&+ 1 - \left(1 - \frac{\delta}{1+e^{\varepsilon}}\right)^{T} + \tilde{\delta} \\
\approx& \left(T - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil\right)\frac{2\delta}{1+e^{\varepsilon}} + \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \delta + \tilde{\delta}.
\end{aligned}
$$

When all the iterators $\mathcal{M}_i$ satisfy $\varepsilon_i \equiv \varepsilon$, we can further tighten the estimation of $\delta'$ for $\varepsilon'_3$ in Theorem 4.

**Corollary 2** (Composition Theorem III)**.** *Suppose the iterators $\mathcal{M}_i$ are $(\varepsilon, \delta_i)$-differentially private and all the other conditions in Theorem 4 hold. Then, algorithm $\mathcal{A}$ is $(\varepsilon', \delta')$-differentially private, where $\varepsilon' = \min\{\varepsilon'_1, \varepsilon'_2, \varepsilon'_3\}$, and*

$$
\begin{aligned}
\delta' =& e^{-\frac{\varepsilon'+T\varepsilon}{2}}\left(\frac{1}{1+e^{\varepsilon}}\left(\frac{2T\varepsilon}{T\varepsilon - \varepsilon'}\right)\right)^{T}\left(\frac{T\varepsilon+\varepsilon'}{T\varepsilon-\varepsilon'}\right)^{-\frac{\varepsilon'+T\varepsilon}{2\varepsilon}} \\
&+ 2 - \left(1 - e^{\varepsilon}\frac{\delta}{1+e^{\varepsilon}}\right)^{\left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil}\left(1 - \frac{\delta}{1+e^{\varepsilon}}\right)^{T-\left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil} \\
&- \left(1 - \frac{\delta}{1+e^{\varepsilon}}\right)^{T}.
\end{aligned}
$$

*Furthermore, for the case $\varepsilon' = \varepsilon'_3$, $\delta'$ is strictly tighter than that in Theorem 4.*

The three composition theorems extend the developed relationship between generalization and privacy preservation to iterative machine learning algorithms. At this point, we establish the theoretical foundation for the generalizability of iterative differentially private machine learning algorithms.

### 3.2.1 Proof Sketch

We now sketch the proofs for Theorem 4 (for more details, please refer to Appendix B). The proofs have four stages: (1) we first approximate the KL-divergence between hypotheses learned on neighboring training sample sets; (2) we then prove a composition bound for algorithms with step $i$ $\varepsilon_i$-differentially private ; (3) this composition theorem is extended to for algorithms with step $i$ $(\varepsilon_i, \delta_i)$-differentially private; and (4) we eventually tighten the result in (3) to obtain Theorem 4. We also proved two additional composition theorems as by-products. The two composition theorems are weaker than Theorem 4 but play essential roles in the proofs.

It is worth noting that the proofs of composition theorems are significantly different from [Kairouz et al., 2017] which calculate the composition combining $\varepsilon$ and $\delta$. This combination causes difficulties in obtaining tight estimation. In contrast, our Theorem 4 and Corollary 2 separate the considerations on $\varepsilon$ and $\delta$ by coupling technique and momentum method, which considerably improve the composition theorem.

**Stage 1: Approximate the KL-divergence between hypotheses learned on neighboring training sample sets.**

It would be technically difficult to approach directly the differential privacy of an iterative learning algorithm from the differential privacy of every iteration. To relieve the technical difficulty, we employ KL divergence as a bridge in this paper. For any $\varepsilon$-differentially private learning algorithm, we prove the following lemma to approximate the KL-divergence between hypotheses learned on neighboring training sample sets.

**Lemma 2.** *If $\mathcal{A}$ is an $\varepsilon$-differentially private algorithm , then for every neighbor database pair $S$ and $S'$, the KL divergence between hypotheses $\mathcal{A}(S)$ and $\mathcal{A}(S')$ satisfies the following inequality,*

$$
D_{KL}(\mathcal{A}(S)\|\mathcal{A}(S')) \le \varepsilon\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}.
$$

To the best of our knowledge, Lemma 2 is the current tightest bound of KL divergence by differential privacy parameter $\varepsilon$. There are two related results in the literature, which are considerably looser than ours. Dwork et al. [2010] proved an inequality of the KL divergence as follows,

$$
D_{KL}(\mathcal{A}(S)\|\mathcal{A}(S')) \le \varepsilon(e^{\varepsilon}-1).
$$

Then, Dwork and Rothblum [2016] further improved it to

$$
D_{KL}(\mathcal{A}(S)\|\mathcal{A}(S')) \le \frac{1}{2}\varepsilon(e^{\varepsilon}-1). \tag{8}
$$

Compared with ours, eq. (8) is larger by a factor $(1+e^{\varepsilon})/2$, which can be very large in practice.

**Stage 2: Prove a weaker composition theorem where the $i$th iteration is $\varepsilon_i$-differential private.**

Based on Lemma 2, we can prove the following composition theorem as a preparation theorem.

**Theorem 5** (Composition Theorem IV)**.** *Suppose an iterative machine learning algorithm $\mathcal{A}$ has $T$ steps: $\{W_i(S)\}_{i=1}^{T}$. Specifically, we define the $i$-th iterator as follows,*

$$
\mathcal{M}_i : (W_{i-1}(S), S) \mapsto W_i(S). \tag{9}
$$

*Assume that $W_0$ is the initial hypothesis (which does not depend on $S$). If for any fixed $W_{i-1}$, $\mathcal{M}_i(W_{i-1}, S)$ is $\varepsilon_i$-differentially private, then $\{W_i\}_{i=0}^{T}$ is ($\varepsilon'$, $\delta'$)-differentially private that*

$$
\varepsilon' = \sqrt{2\log\left(\frac{1}{\delta'}\right)\left(\sum_{i=1}^{T}\varepsilon_i^2\right)} + \sum_{i=1}^{T}\varepsilon_i\frac{e^{\varepsilon_i}-1}{e^{\varepsilon_i}+1}.
$$

**Stage 2: Prove a weaker composition theorem where the $i$th iteration is $(\varepsilon_i, \delta_i)$-differentially private.** The following technical lemma is adopted to derive the generalization bound:

**Lemma 3** (cf. [Dwork and Roth, 2014], Theorem 3.17). *For any random variables $Y$ and $Z$, we have that*

$$D_{\infty}^{\delta}(Y\|Z) \leq \varepsilon,\ D_{\infty}^{\delta}(Z\|Y) \leq \varepsilon,$$

*if and only if there exist random variables $Y'$, $Z'$ such that*

$$\Delta(Y\|Y') \leq \frac{\delta}{e^{\varepsilon}+1},\ \Delta(Z\|Z') \leq \frac{\delta}{e^{\varepsilon}+1},$$
$$D_{\infty}(Y'\|Z') \leq \varepsilon,\ D_{\infty}(Z'\|Y') \leq \varepsilon.$$

Here $D_{\infty}^{\delta}$ is the $\delta$-approximate max divergence, $D_{\infty}$ is the max divergence, and $\Delta$ is the statistical distance, and we defer the formal definitions to Appendix B.1.

Intuitively, lemma 3 allows us to separate $\varepsilon$ and $\delta$ for variables with $\delta$ max divergence $\varepsilon$ and derive coupled variables with max divergence $\varepsilon$. Based on Lemmas 3, we can prove the following composition theorem whose estimate of $\varepsilon'$ is somewhat looser than our main results.

**Theorem 6** (Composition Theorem V). *Let all the conditions and notations in Theorem 5 hold except that the ith step is $(\varepsilon_i, \delta_i)$ differentially private. Then, $\{W_i\}_{i=0}^{T}$ is $(\varepsilon', \delta')$-differentially private where*

$$\varepsilon' = \sqrt{2\log\left(\frac{1}{\tilde{\delta}}\right)\left(\sum_{i=1}^{T}\varepsilon_i^2\right)} + \sum_{i=1}^{T}\varepsilon_i\frac{e^{\varepsilon_i}-1}{e^{\varepsilon_i}+1},$$
$$\delta' = \max_{\{\alpha_i\}_{i=1}^{T}\in\partial I} \tilde{\delta} + 2 - \prod_{i=1}^{T}\left(1 - e^{\alpha_i}\frac{\delta_i}{1+e^{\varepsilon_i}}\right)$$
$$- \prod_{i=1}^{T}\left(1 - \frac{\delta_i}{1+e^{\varepsilon_i}}\right),$$

*and $I = \{\sum_{i=1}^{T}\alpha_i = \varepsilon', \alpha_i \geq 0\}$.*

**Stage 4: Prove Theorem 4.**

By the same routine of deriving Theorem 6 together with Theorem 3.5 in [Kairouz et al., 2017], we eventually extend the weaker versions to Theorem 4.

### 3.2.2 Comparison with Existing Results

Our composition theorem is strictly tighter than the existing results: a classic composition theorem is as follows (see [Dwork and Roth, 2014], Theorem 3.20 and Corollary 3.21, pp. 49-52),

$$\varepsilon' = \sum_{i=1}^{T}\varepsilon_i(e^{\varepsilon_i}-1) + \sqrt{2\log\left(\frac{1}{\delta}\right)\sum_{i=1}^{T}\varepsilon_i^2},$$
$$\delta' = \tilde{\delta} + \sum_{i=1}^{T}\delta_i,$$

where $\tilde{\delta}$ is an arbitrary positive real number, $(\varepsilon', \delta')$ is the differential privacy of the whole algorithm, and $(\varepsilon_i, \delta_i)$ is the differential privacy of the $i$-th iteration. Currently, the tightest approximation is given by [Kairouz et al., 2017] as follows,

$$\varepsilon' = \min\left\{\varepsilon_1', \varepsilon_2', \varepsilon_3'\right\},$$
$$\delta' = 1 - (1-\delta)^T(1-\tilde{\delta}),$$

where $\varepsilon_1'$, $\varepsilon_2'$, and $\varepsilon_3'$ are the same as those in Theorem 4. Therefore, their estimate of the $\varepsilon'$ is the same as ours, while their $\delta'$ is also larger than ours approximately by

$$\delta\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\left(T - \left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil\right).$$

The iteration number $T$ is usually overwhelmingly large, which guarantees our advantage is significant.

# 4 EXPERIMENTS

We conducted a systematic experiment to verify three relationships suggested by our theory: (1) a positive correlation between privacy preservation and generalization; (2) a negative correlation between privacy preservation and training time; and (3) a negative correlation between generalization and training time. The empirical results are in full agreement with the theoretical results.

## 4.1 IMPLEMENTATION DETAILS

Three benchmark image datasets, MNIST [LeCun, 1998], CIFAR-10 and CIFAR-100 [Krizhevsky et al., 2009], are used in our experiments. The separations of the training sets and the test sets are the same as the official versions. For MNIST, we trained one-hidden-layer MLP, in which the number of the hidden neurons is 100. For CIFAR-10 and CIFAR-100, we trained VGG-16 [Simonyan and Zisserman, 2014] and ResNet-18 [He et al., 2016].

SGD is employed to train the model. The batch size is set as 128, the momentum factor is set as 0.9, and the weight decay factor is set as 0.0005. For the experiments on MNIST, every model is trained for 50 epochs. The learning rate is initialized as 0.1 and decayed by 0.1 every 20 epochs. For the experiments on CIFAR-10 and CIFAR-100, every model is trained for 100 epochs. The learning rate is initialized as 0.1 and decays by 0.1 every 40 epochs.

Membership inference attack [Yeom et al., 2018, Shokri et al., 2017] is performed to the model in every epoch in order to evaluate the privacy-preserving ability. Usually, a larger membership inference attack indicates a weaker privacy-preserving ability. The implementation of membership inference attack is from JD Explore Academy Trustworthy AI Toolkit. We also calculate the generalization error ( the difference between training and test errors) of the model in every epoch. Similarly, a larger generalization error indicates a weaker generalizability.

Table 1: Spearman's rank order correlation coefficient (SCC) and $p$ value of: (1) membership inference attack accuracy and generalization error; (2) membership inference attack accuracy and training time; and (3) generalization error and training time.

| | MLP - MNIST | | VGG-16 - CIFAR-10 | | ResNet-18 - CIFAR-10 | | VGG-16 - CIFAR-100 | | ResNet-18 - CIFAR-100 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SCC | $p$ | SCC | $p$ | SCC | $p$ | SCC | $p$ | SCC | $p$ |
| Privacy - Gen. Err. | 0.92 | $\mathbf{1.2\times10^{-21}}$ | 0.92 | $\mathbf{1.8\times10^{-41}}$ | 0.80 | $\mathbf{1.1\times10^{-23}}$ | 1.00 | $\mathbf{6.4\times10^{-107}}$ | 0.82 | $\mathbf{4.2\times10^{-25}}$ |
| Privacy - Time | 0.96 | $\mathbf{1.7\times10^{-27}}$ | 0.98 | $\mathbf{9.9\times10^{-73}}$ | 0.97 | $\mathbf{1.5\times10^{-64}}$ | 0.99 | $\mathbf{5.0\times10^{-91}}$ | 0.98 | $\mathbf{8.8\times10^{-72}}$ |
| Gen. Err. - Time | 0.91 | $\mathbf{8.6\times10^{-20}}$ | 0.91 | $\mathbf{2.6\times10^{-40}}$ | 0.80 | $\mathbf{2.1\times10^{-23}}$ | 0.99 | $\mathbf{5.5\times10^{-87}}$ | 0.82 | $\mathbf{7.0\times10^{-26}}$ |

**Algorithm 1** Iterative Gaussian Mechanism with Mini-batch (IGMM)

```
Require: Sample $S = \{z_1, ...z_N\}$, Gaussian noise variance $\sigma$, size of mini-batch $\tau$, iteration steps $T$, learning rate $\{\eta_1, ...\eta_T\}$, update function $g(z, W)$, and its diameter $D \triangleq \max_{W,z,z'} \|g(z, W) - g(z', W)\|$.
Initialize $W_0$ randomly.
For $t = 1$ to $T$ do:
    Uniformly sample a mini-batch $\mathcal{B}_t$ of size $\tau$ from $S$ without replacement;
    Sample $g_t$ from $\sigma\mathcal{N}(0, \mathbb{I})$;
    Update
    $W_t \leftarrow W_{t-1} - \eta_t \left[\frac{1}{\tau}\sum_{z\in\mathcal{B}_t} g(z, W_{t-1}) + g_t\right]$.
```

The collected (membership inference attack accuracy, training time) pairs, (generalization error, training time) pairs, and (membership inference attack accuracy, generalization error) pairs are divided into groups according to neural network architectures and datasets. Spearman's rank-order correlation test [Spearman, 1987] is performed on every group.

### 4.2 EMPIRICAL RESULTS

The Spearman's rank-order correlation coefficients and the $p$ values are collected in Table 1. The $p$ values are all smaller than $0.05$, which show the following three correlations are statistically significant: (1) a positive correlation between privacy preservation and generalization; (2) a negative correlation between privacy preservation and training time; and (3) a negative correlation between generalization and training time. These results are in full agreement with our theory.

## 5 APPLICATIONS

Our theories apply to a wide spectrum of machine learning algorithms. This paper implements them to the popular iterative Gaussian mechanisms with mini-batch (1) as an example. Gaussian mechanism are widely adopted in privacy-preserving machine learning. In this paper, we show the application in agnostic federated learning.

We first formally describe iterative Gaussian mechanism with mini-batch in Algorithm 4.1. We then apply Theorem 1 and Corollary 2 to estimate the differential privacy and deliver a generalization bound for SGLD.

**Theorem 7.** *IGMM described as Algorithm 1 is $(\varepsilon', \delta')$-differentially private, where*

$$\begin{aligned}\varepsilon' =& \sqrt{2T\log\left(\frac{1}{\tilde{\delta}}\right)\tilde{\varepsilon}^2 + T\tilde{\varepsilon}\frac{e^{\tilde{\varepsilon}}-1}{e^{\tilde{\varepsilon}}+1}},\\ \delta' =& 2 - \left(1 - \frac{\delta}{1+e^{\tilde{\varepsilon}}}\right)^T + \tilde{\delta}'\\ &- \left(1 - \frac{\delta e^{\tilde{\varepsilon}}}{1+e^{\tilde{\varepsilon}}}\right)^{\left\lceil\frac{N\varepsilon'}{\tau\tilde{\varepsilon}}\right\rceil}\left(1 - \frac{\delta}{1+e^{\tilde{\varepsilon}}}\right)^{T-\left\lceil\frac{N\varepsilon'}{\tau\tilde{\varepsilon}}\right\rceil},\end{aligned}$$

*$\tilde{\varepsilon}$ is defined as*

$$\tilde{\varepsilon} = \log\left(\frac{N-\tau}{N} + \frac{\tau}{N}\exp\left(\frac{\sqrt{2}D\sigma\frac{1}{\tau}\sqrt{\log\frac{1}{\delta}} + \frac{1}{\tau^2}D^2}{2\sigma^2}\right)\right),$$

*and $\tilde{\delta}'$ is given as*

$$\tilde{\delta}' = e^{-\frac{\varepsilon'+T\tilde{\varepsilon}}{2}}\left(\frac{T\tilde{\varepsilon}+\varepsilon'}{T\tilde{\varepsilon}-\varepsilon'}\right)^{-\frac{\varepsilon'+T\tilde{\varepsilon}}{2\tilde{\varepsilon}}}\left(\frac{1}{1+e^{\tilde{\varepsilon}}}\left(\frac{2T\tilde{\varepsilon}}{T\tilde{\varepsilon}-\varepsilon'}\right)\right)^T.$$

*Additionally, a generalization bound is delivered by combining with Theorem 1.*

**Remark 2.** *The generalization bound in Theorem 7 does not explicitly rely on the model size. One may argue the diameter $R$ of update function $g$ implicitly relies on the model size. However, in deep learning, $g$ usually stays in a low rank manifold. As an example, Yu et al. [2021] empirically show that the linear space formed by gradients in deep learning is of low dimension, largely independent of the model size. This dismisses the possibility of introducing dependence on model size by the gradient distance.*

Agnostic federated learning is a special case of IGMM. Therefore, by Theorem 7, we have the following asymptotic high probability generalization bound for agnostic federated learning in terms of client number $C$.

**Corollary 3** (Asymptotic generalization bound for agnostic federated learning)**.** *For agnostic federated learning algorithm $\mathcal{A}^{fed}$ with client set $S$ with size C, the generalization error is bounded as*

$$\mathbb{P}\left(|\operatorname{Gen}_{S,\mathcal{A}_S^{fed}}| > \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{C}}\right)\right) < \tilde{\mathcal{O}}\left(\frac{1}{C^{\frac{3}{2}}}\right). \quad (10)$$

**Remark 3.** *The high probability generalization bound eq.(10) also holds for SGLD with $N$ the sample size.*

As an over-parameterized model, deep learning has demonstrated excellent generalizability, which is somehow beyond the explanation of the existing statistical learning theory and thus attracts the community's interest [E et al., 2020, He and Tao, 2020]. Recent advances include generalization bounds via hypothesis complexity [Bartlett et al., 2019, Golowich et al., 2018, Bartlett et al., 2017, Liang et al., 2019, Tu et al., 2020], PAC-Bayesian framework [Neyshabur et al., 2018], algorithmic stability [Hardt et al., 2016], and stochastic gradient descent or its variant [Mandt et al., 2017, Mou et al., 2018] on the loss surface [Kawaguchi, 2016, Yun et al., 2019, He et al., 2020]. A major difficulty in explaining deep learning's excellent generalizability is that deep learning models usually has prohibitively large parameter size which make many generalization bounds vacuous.

Additionally, deep learning has become a promising player in many real-world applications, including financial services Fischer and Krauss [2018], healthcare Wang et al. [2012], and biometric authentication Snelick et al. [2005], in which the privacy-preserving ability is of vital importance. Several works have also studied the privacy preservation of deep learning and how to improve it further Abadi et al. [2016], Arachchige et al. [2019].

This work establishes generalization bounds for iterative learning algorithms via differential privacy, which do not explicitly rely on the model size. Our results also shed light to understanding the generalizability of deep learning from the privacy-preserving view.

## 6 CONCLUSION

This paper studies the relationships between generalization and privacy preservation in two steps. We first establish the relationship between generalization and privacy preservation for any machine learning algorithm. Specifically, we prove a high-probability bound for differentially private learning algorithms based on a novel on-average generalization bound for multi-sample-set algorithms. Then, we prove three composition theorems that calculate the $(\varepsilon', \delta')$-differential privacy of an iterative algorithm. By integrating the two steps, we establish generalization guarantees for iterative differentially private machine learning algorithms. Compared with existing works, our theoretical results are strictly tighter and apply to a wider application domain. We then use them to the Gaussian mechanism.The obtained generalization bounds do not explicitly rely on the model size which would be prohibitively large in many modern models, such deep neural networks. The empirical results of MLP, VGG, and ResNet on the MNIST, CIFAR-10, and CIFAR-100 datasets are in full agreement with our theory.

**Acknowledgements**

The theoretical part was supported by Australian Research Council Project FL-170100117. The empirical study was completed in JD Explore Academy. The authors appreciate the anonymous UAI 2021 reviewers for the helpful suggestions.

**References**

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 2019.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *NeurIPS*, 2018.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(63):1–17, 2019.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of IEEE Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of Annual ACM Symposium on Theory of Computing*, pages 117–126, 2015.

Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020.

Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. In *Advances in Neural Information Processing Systems*, 2017.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of Annual Conference on Learning Theory*, pages 297–299, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of International Conference on Machine Learning*, pages 1225–1234, 2016.

Fengxiang He and Dacheng Tao. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020.

Fengxiang He, Bohan Wang, and Dacheng Tao. Piecewise linear activations substantially shape the loss surfaces of neural networks. In *International Conference on Learning Representations*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385, 2015.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

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

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 888–896, 2019.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18 (1):4873–4907, 2017.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, pages 94–103, 2007.

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

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Proceedings of Annual Conference On Learning Theory*, pages 605–638, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Kobbi Nissim and Uri Stemmer. On the generalization properties of differential privacy. *CoRR, abs/1504.05800*, 2015.

Luca Oneto, Sandro Ridella, and Davide Anguita. Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognition Letters*, 89:31–38, 2017.

Francesco Pittaluga and Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (11):2215–2226, 2016.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Robert Snelick, Umut Uludag, Alan Mink, Mike Indovina, and Anil Jain. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):450–455, 2005.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. Understanding generalization in recurrent neural networks. In *International Conference on Learning Representations*, 2020.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, Shahram Ebadollahi, and Andrew F Laine. A framework for mining signatures from event sequences and its applications in healthcare data. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):272–285, 2012.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of International Conference on Machine Learning*, pages 681–688, 2011.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations*, 2019.