Tighter Generalization Bounds for Iterative Differentially Private Learning Algorithms

Fengxiang He * † Bohan Wang * † Dacheng Tao †

Abstract

This paper studies the relationship between generalization and privacy preservation in iterative learning algorithms by two sequential steps. We first establish the generalization-privacy relationship for any learning algorithm. We prove that \((\varepsilon, \delta)\)-differential privacy implies an on-average generalization bound for multi-database learning algorithms which further leads to a high-probability generalization bound. The high-probability generalization bound implies a PAC-learnable guarantee for differentially private algorithms. We then investigate how the iterative nature would influence the generalizability and privacy. Three new composition theorems are proposed to approximate the \((\varepsilon', \delta')\)-differential privacy of any iterative algorithm through the differential privacy of its every iteration. By integrating the above two steps, we deliver two generalization bounds for iterative learning algorithms, which characterize how privacy-preserving ability guarantees generalizability and how the iterative nature contributes to the generalization-privacy relationship. All the theoretical results are strictly tighter than the existing results in the literature and do not explicitly rely on the model size which can be prohibitively large in deep models. The theories directly apply to a wide spectrum of learning algorithms. In this paper, we take stochastic gradient Langevin dynamics and the agnostic federated learning from the client view for examples to show one can simultaneously enhance privacy preservation and generalizability through the proposed theories.

Keywords: learning theory, differential privacy, generalization.

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 training data rather than just memorize the data [71, 53]. 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 in the data while protecting the highly sensitive individual privacy has profound importance [22, 23, 62].

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

We first prove two theorems that upper bound the generalization error via its differential privacy. Specifically, we prove a high-probability generalization bound that characterizes the cumulative distribu-
tion function of the generalization error,
\[ R(A(S)) - \hat{R}_S(A(S)), \]
where \( A(S) \) is the hypothesis learned by algorithm \( A \) on dataset \( S \), \( R(A(S)) \) is the expected risk, and \( \hat{R}_S(A(S)) \) is the empirical risk. This bound is established based on a novel on-average generalization bound for any \((\varepsilon, \delta)\)-differentially private multi-database learning algorithm. Our high-probability generalization bound implies that differentially private machine learning algorithms are probably approximately correct (PAC)-learnable. These two generalization bounds indicate that the algorithms with a good privacy-preserving ability also have good generalizability. We, therefore, can expect to innovate algorithms to approach better generalization performance by enhancing the privacy-preserving ability.

We then studied how the iterative nature influences the generalization-privacy relationship. Generally, the privacy-preserving ability of an iterative algorithm degenerates along with iterations, since the amount of leaked information accumulates 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 iterations. By combining Theorem 1 with the composition theorems, we can establish the generalization-privacy relationship for iterative learning algorithms.

Our results considerably extend the current understanding of the generalization-privacy relationship. Some works [21, 58, 59] prove some high-probability generalization bounds in the form as below,
\[ \mathbb{P}\left[R(A(S)) - \hat{R}_S(A(S)) > a\right] < b. \]
Among them, the current tightest generalization bounds are obtained by Nissim and Stemmer [58]. Our high-probability bound is strictly tighter than the current best results. Besides, we prove the first PAC-learnable guarantee for differentially private machine learning algorithms via our high-probability generalization bound. Nissim and Stemmer also prove an on-average multi-database generalization bound. Our bound also tightens it by a factor of \( e^\varepsilon \), which would be considerably large in practice. Also, the bounds in [58] are only for binary classification, while ours apply to any differentially private learning algorithm. Some works have also proved composition theorems [23, 36]. The approximation of factor \( \delta \) in our composition theorem is tighter than the tightest existing result by the following factor,
\[ \delta e^{\varepsilon} - 1 \left( e^{\varepsilon} - 1 \right), \]
for \( T \) iterations, while the estimate of \( \varepsilon' \) remains the same [36]. Our composition theorems can considerably tighten our generalization bounds for iterative learning algorithms because the term \( b \) is directly proportional to \( \delta \).

Our theories apply to a wide spectrum of machine learning algorithms. This paper applies them to stochastic gradient Langevin dynamics [75] as an example of stochastic gradient Markov chain Monte Carlo [47], and the agnostic federated learning from the client view [28]. Our theories help deliver an on-average generalization bounds and a high-probability generalization bounds for SGLD and agnostic federated learning. Furthermore, 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. These implementations are natural but technically non-trivial.

The rest of this paper is organized as follows. Section 2 reviews related works of generalization, differential privacy, and algorithmic stability. Section 3 defines notations and recall preliminaries. Section 4 provides main results: specifically, Section 4.1 establishes the relationships between generalizability and privacy preservation, sketches the proofs wherein, and compares existing results with ours; and Section 4.2 establishes the degenerative nature of differential privacy in iterative algorithms, sketches the proofs wherein, and compares existing results with ours. Section 5 applies our theories to stochastic gradient Langevin dynamics (Section 5.1) and agnostic federated learning (Section 5.2). Section 6 concludes this paper.


2 Background

Generalization bound is a standard measurement of the generalization ability, which is defined as the upper bound of the difference between the expected risk and the empirical risk [56, 71, 53]. Since the two risks can be treated as the training error and the expectation of the test error, generalization bound expresses the gap between the performance on existing data and the performance on unknown data. Therefore, we can expect an algorithm with a small generalization bound to generalize well. Existing generalization bounds are mainly obtained from three stems: (1) concentration inequalities derive many high-probability generalization bounds based on the hypothesis complexity, such as VC dimension [9, 70], Rademacher complexity [39, 38, 6], and covering number [19, 32]. These generalization bounds suggest implementations consistent with the principle of Occam’s razor that controls the hypothesis complexity to help models generalize better; (2) some on-average and high-probability generalization bounds are proved based on the algorithmic stability to the disturbance in the training sample set [65, 11], following an intuition that an algorithm with good generalization ability is insensitive to the disturbance in individual data points; and (3) under the PAC-Bayes framework [49, 50], generalization bounds are established on information-theoretical distances between the output hypothesis and the prior, such as KL divergence and mutual information. As an over-parameterized model, deep learning demonstrates excellent generalizability, which is somehow beyond the explanation of the existing statistical learning theory and thus attracts the community’s interests. Recent advances in deep learning theory include generalization bounds via VC dimension [5], Rademacher complexity [29, 4], covering number [4], Fisher-Rao norm [44, 68], PAC-Bayesian framework [57], algorithmic stability [30, 42, 72], and via the dynamics of stochastic gradient descent or its variant [48, 55, 33] driven by the loss surface. A major difficulty wherein is the prohibitively complicated nature of the loss surfaces, which, however, exactly characterizes deep learning’s behavior [34].

Differential privacy measures an algorithm according to its privacy-preserving ability [23, 20]. Specifically, $(\varepsilon, \delta)$-differential privacy is defined as the change in output hypothesis when the algorithm $\mathcal{A}$ is exposed to attacks as follows, 

$$
\log \left[ \frac{P_{\mathcal{A}(S)}(\mathcal{A}(S) \in B) - \delta}{P_{\mathcal{A}(S')} (\mathcal{A}(S') \in B)} \right] \leq \varepsilon,
$$

where $B$ is an arbitrary subset of the hypothesis space and $(S, S')$ is a neighboring sample set pair in which $S$ and $S'$ only differ by one example. Therefore, an algorithm with small differential privacy $(\varepsilon, \delta)$ robust to changes in individual training examples. Thus, the magnitude of differential privacy $(\varepsilon, \delta)$ indexes the ability to resist differential attacks that uses fake sample points as probes to attack machine learning algorithms, and then infer the individual privacy via the changes of output hypotheses. Many variants of differential privacy have been designed by modifying the division operation: (1) concentration differential privacy assumes that the privacy loss defined as below, 

$$
\log \left[ \frac{P_{\mathcal{A}(S)}(\mathcal{A}(S) \in B)}{P_{\mathcal{A}(S')} (\mathcal{A}(S') \in B)} \right],
$$

is sub-Gaussian [24, 12]; (2) mutual-information differential privacy and KL differential privacy adapt mutual information and KL divergence, respectively, to measure changes of the hypotheses [15, 73, 45, 13]; (3) Rényi differential privacy further replaces the KL divergence by Rényi divergence [52, 27]; etc. As a game-changer, deep learning has reshaped machine learning and become a major player in many areas. Its privacy-preserving ability is thus of much interest [1, 3].

Algorithmic stability measures how stable an algorithm is when the training sample is exposed to disturbance [65, 37, 11, 76, 69]. It is defined as the change to the output hypothesis when a single training example is disturbed. The change is measured according to the distances or pseudo-distances in the hypothesis space $\mathcal{H}$ or its variants [16]; for example, (1) hypothesis stability measures the changes based on the $L_1$ norm on the composited space $l \circ \mathcal{H}$ ($l$ is the loss function) [11]; (2) uniform stability is defined based on the $L_{\infty}$ norm on $l \circ \mathcal{H}$ [11]; (3) error stability further calculates the expectation of the uniform stability with respect to the randomness in the sample set [11]; (4) TV stability measures the
change according to the difference in probability:

$$\sup_{\mathcal{R}} |\mathbb{P}(A(S) \in \mathcal{R}) - \mathbb{P}(A(S') \in \mathcal{R})|,$$

where $\mathcal{R}$ is a subset of the hypothesis space $\mathcal{H}$ [7]; and (5) KL stability measures the change by the KL divergence [7]. In the context of model over-parameterized machine learning models, algorithmic stability sheds lights to understanding their success [30, 42, 72].

These definitions of algorithmic stability provide a natural way to bridge generalization and differential privacy. Some works have connected generalization with algorithmic stability. For example, [11] presented polynomial high-probability generalization bounds via hypothesis stability and exponential high-probability generalization bounds via uniform stability. Furthermore, algorithmic stability and differential privacy measure almost the same object from different aspects. Specifically, algorithmic stability measures the change in the output hypothesis when the training sample is disturbed by the subtraction operation, while differential privacy measures the same change by the division operation. Given this similarity, some papers have even treated differential privacy as algorithmic stability; for example, [7] called differential privacy and KL differential privacy as max-KL stability and KL stability, respectively.

3 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, $d_X$ and $d_Y$ are the dimensions of the feature $X$ and the label $Y$, respectively. For the brevity, we define $z_i = (x_i, y_i)$. All $z_i$ 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.

A machine learning algorithm $A$ learns a hypothesis,

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

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

$$\mathcal{R}(A(S)) = \mathbb{E}_Z l(A(S), Z),$$

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

where $l : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}^+$ is the loss function. It worths noting that both the algorithm $A$ and the training sample set $S$ can introduce randomness in the expected risk $\mathcal{R}(A(S))$ and empirical risk $\mathcal{R}(A(S))$.

The generalization error is defined as the difference between the expected risk and empirical risk,

$$\hat{\mathcal{R}}_S(A(S)) - \mathcal{R}(A(S)),$$

whose upper bound is called the generalization bound.

Privacy-preserving ability is important to machine learning. Differential privacy measures the ability to preserve privacy. Differential privacy is defined as follows.

**Definition 1** (Differential Privacy; cf. [23]). A stochastic algorithm $A$ is called $(\varepsilon, \delta)$-differentially private if for any subset $B \subset \mathcal{H}$ and any neighboring sample set pair $S$ and $S'$ which are different by only one example, we have

$$\log \left[ \frac{\mathbb{P}_{A(S)}(A(S) \in B)}{\mathbb{P}_{A(S')} (A(S') \in B)} \right] \leq \varepsilon.$$

The algorithm $A$ is also called $\varepsilon$-differentially private, if it is $(\varepsilon, 0)$-differentially private.
Algorithmic stability measures a machine learning algorithm according to the stability of its output hypothesis when the training sample set is disturbed [11]. While algorithmic stability has many different definitions, this paper mainly discusses uniform stability.

**Definition 2** (Uniform stability; cf. [11]). A machine learning algorithm $A$ is uniformly stable, if for any neighboring sample pair $S$ and $S'$ which are different by only one example, we have the following inequality,

$$|E_{A(S)}(A(S), Z) - E_{A(S')}(A(S'), Z)| \leq \beta,$$

where $Z$ is an arbitrary example, $A(S)$ and $A(S')$ are the output hypotheses learned on the training sets $S$ and $S'$, respectively, and $\beta$ is a positive real constant. The constant $\beta$ is called the uniform stability of the algorithm $A$.

## 4 Generalization Bounds for Iterative Differentially Private Algorithms

This section analyses the generalizability of iterative differentially private algorithms. The establishment has two steps. We first establish the generalization-privacy relationship for any machine learning algorithm. Then, we investigate how the iterative nature shared by most algorithms would influence the differential privacy by delivering three novel composition theorems. The two stages collectively establish the generalizability of iterative differentially private algorithms.

### 4.1 Bridging Generalization and Privacy

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

**Theorem 1** (High-Probability Generalization Bound via Differential Privacy). Suppose algorithm $A$ is $(\varepsilon, \delta)$-differentially private, the training sample size,

$$N \geq \frac{2}{\varepsilon^2} \ln \left( \frac{16}{e^{-\varepsilon \delta}} \right),$$

and the loss function $\|l\|_{\infty} \leq 1$. Then, for any distribution $D$ over $Z$, we have

$$\mathbb{P} \left[ \left| \hat{R}_S(A(S)) - R(A(S)) \right| < 9 \varepsilon \right] > 1 - \frac{e^{-\varepsilon \delta}}{\varepsilon} \ln \left( \frac{2}{\varepsilon} \right).$$

This high-probability generalization bound characterizes how the cumulative distribution function of the generalization error $\hat{R}_S(A(S)) - R(A(S))$. This bound demonstrates that the privacy-preserving ability implies generalizability. Thus, we can unify the algorithm design for enhancing privacy preservation and for improving generalization.

Theorem 1 implies that $(\varepsilon, \delta)$-differentially private algorithms are also probably approximately correct (PAC)-learnable, which is the first in the literature. PAC-learnability is defined as below,

**Definition 3** (PAC-learnability). A concept class $C$ is said to be PAC-learnable if there exists an algorithm $A$ and a polynomial function $\text{poly}(\cdot, \cdot, \cdot)$ such that for any $s > 0$ and $t > 0$, for all distributions $D$ on the training example $Z$, any target concept $c \in C$, and any sample size

$$m \geq \text{poly}(1/s, 1/t, n, \text{size}(C)),$$

the following holds,

$$\mathbb{P}_{S \sim D^n}(|R(A(S)) < s|) > 1 - t. \quad (1)$$

In Section 5, we prove the PAC-learnability for stochastic gradient Langevin dynamics (SGLD) as an example of the Gaussian mechanism. The PAC-learnable guarantee for the Laplacian mechanism, another major stream of privacy-preserving algorithms, can be similarly obtained.
4.1.1 Proof Skeleton

Theorems 1 is established on a novel on-average generalization bound for multi-database algorithms. Suppose the training sample \( S \) is separated to \( k \) i.i.d. sub-databases \( S_1, \ldots, S_k \). For the brevity, we write the training sample set as below

\[
\vec{S} = (S_1, \ldots, S_k).
\]

Correspondingly, the output of a multi-database algorithm \( \vec{A} \) is a pair as follows,

\[
(h_{A(\vec{S})}, i_{A(\vec{S})}) \in \mathcal{H} \times \{1, \ldots, k\}.
\]

**Theorem 2** (On-Average Multi-Database Generalization Bound). *Let algorithm as follows,

\[
\vec{A} : \vec{S} \to \mathcal{H} \times \{1, \ldots, k\},
\]

is \((\varepsilon, \delta)\)-differentially private and the loss function \( \|l\|_{\infty} \leq 1 \). Then, for any distribution \( \mathcal{D} \) over \( Z \), we have

\[
\mathbb{E}_{\vec{S} \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(\vec{S})} \left[ \hat{R}_{S_{A(\vec{S})}} (h_{A(\vec{S})}) \right] - \mathbb{E}_{A(\vec{S})} \left[ R \left( h_{A(\vec{S})} \right) \right] \right] \leq e^{-\varepsilon}k\delta + 1 - e^{-\varepsilon}. \tag{2}
\]

Since \( 1 - e^{-\varepsilon} \leq \varepsilon \), we have the following corollary:

**Corollary 1**. *Suppose all the condition in Theorem 2 holds, then

\[
\mathbb{E}_{\vec{S} \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(\vec{S})} \left[ \hat{R}_{S_{A(\vec{S})}} (h_{A(\vec{S})}) \right] \right] \leq e^{-\varepsilon}k\delta + \varepsilon \]

Markov bound (cf. [53], Theorem C.1) is an important concentration inequality in learning theory. Here, we slightly modify the original version as follows,

\[
\mathbb{E}_{x} [h(x)] \geq \mathbb{E}_{x} [h(x)1_{h(x) \geq g(x)}] \geq \mathbb{E}_{x} [g(x)1_{h(x) \geq g(x)}].
\]

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

**Lemma 1**. *Let algorithm \( A : \mathcal{Z}^{kN} \to \mathcal{Y} \times \{1, \cdot, k\} \) have \((\varepsilon, \delta)\) differential privacy. Let \( A(\vec{S}) = (h_{A(\vec{S})}, i_{A(\vec{S})}) \). Then, for any distribution \( \mathcal{D} \) over \( Z \), any database set \( \vec{S} = \{S_i\}_{i=1}^{k} \) where \( S_i \) is a database contains \( N \) i.i.d. sample from \( \mathcal{D} \), we have

\[
\mathbb{P} \left[ \hat{R}_{S_{A(\vec{S})}} (h_{A(\vec{S})}) \leq R \left( h_{A(\vec{S})} \right) + ke^{-\varepsilon} \delta + 3\varepsilon \right] \geq \varepsilon.
\]

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.

**Definition 4** (Exponential Mechanism). *The exponential mechanism \( q(S, u, R, \varepsilon) \) selects and outputs an element \( r \in R \) with probability proportional to \( \exp \left( \frac{u(S, r) - u(S', r)}{2\Delta u} \right) \), where \( u(S, r) \) is the utility function and \( \Delta u \) is the sensitivity of \( u \) defined by

\[
\Delta u = \Delta \max_{r \in R} \max_{S, S' \text{ adjacent}} |u(S, r) - u(S', r)|.
\]

Specifically, we let the input be \( \vec{S} = (S_1, \cdots, S_k) \) and \( T \), where \( S_i, T \in \mathcal{Z}^N \). We then run the exponential mechanism with utility function as follows,

\[
u(\vec{S}, T, i) = N\hat{R}_{S_i}(A(S_i)) - N\hat{R}_{T}(A(S_i)).
\]
We can prove $\mathcal{B}$ is $(2\varepsilon, \delta)$-DP and

$$
P \left[ \hat{R}_{\mathcal{B}(\mathcal{S})} \left( h_{\mathcal{B}(\mathcal{S})} \right) \leq R \left( h_{\mathcal{B}(\mathcal{S})} \right) + ke^{-\varepsilon} \delta + 6\varepsilon \right] < \varepsilon,
$$

which then leads to Theorem 1.

This is summarized as the following lemma.

**Lemma 2.** Let algorithm $A : \mathcal{Z}^N \rightarrow \mathbb{R}^Z$, $k = \frac{\varepsilon}{e^{-\varepsilon} \delta}$, and

$$
N \geq \frac{2}{\varepsilon^2} \ln \left( \frac{16}{e^{-\varepsilon} \delta} \right).
$$

If algorithm $A$ and distribution $D$ over sample space $\mathcal{Z}$ satisfy

$$
P \left[ \hat{R}(A(S)) \leq e^{-\varepsilon} k \delta + 8\varepsilon + R(A(S)) \right] < 1 - \frac{e^{-\varepsilon} \delta}{\varepsilon} \ln \left( \frac{2}{\varepsilon} \right),
$$

then there exists an algorithm

$$
\mathcal{B} : \mathcal{Z}^k N \rightarrow \mathbb{R}^Z \times \{1, \cdots, k\},
$$

is $(2\varepsilon, \delta)$-differentially private and

$$
P \left[ \hat{R}_{\mathcal{B}(\mathcal{S})} \left( h_{\mathcal{B}(\mathcal{S})} \right) \leq R \left( h_{\mathcal{B}(\mathcal{S})} \right) + ke^{-\varepsilon} \delta + 6\varepsilon \right] < \varepsilon,
$$

where $\mathcal{S} = \{S_i\}_{i=1}^k$ and $S_i$ is a database contains $N$ i.i.d. sample from $D$.

### 4.1.2 Comparison with Existing Results

**Comparison for Theorem 1.** There have been several high-probability generalization bounds for $(\varepsilon, \delta)$-differentially private machine learning algorithms. Dwork et al. [21] prove that

$$
P \left[ R(A(S)) - \hat{R}_S(A(S)) < 4\varepsilon \right] > 1 - 8\delta^\varepsilon.
$$

Oneto et al. [59] prove that

$$
P \left[ \text{Diff} \ R < \sqrt{6\hat{R}_S(A(S))\hat{\varepsilon} + 6(\varepsilon^2 + 1/N)} \right] > 1 - 3e^{-N\varepsilon^2},
$$

and

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

where

$$
\text{Diff} \ R = R(A(S)) - \hat{R}_S(A(S)),
$$

$$
\hat{\varepsilon} = \varepsilon + \sqrt{1/N},
$$

$$
\hat{V}_S(A(S)) = \frac{1}{2N(N-1)} \sum_{i \neq j} \left[ \ell(A(S), z_i) - \ell(A(S), z_j) \right]^2.
$$

It is worth noting that $\hat{V}_S(A(S))$ is the empirical variance of $l(A(S), \cdot)$ (see [59], Lemmas 3 and 5, respectively).

The existing tightest high-probability generalization bound in the literature is obtained by Nissim and
Stemmer [58] as follows
\[ P \left[ R(A(S)) - \hat{R}_S(A(S)) < 13\epsilon \right] > 1 - \frac{2\delta}{\epsilon} \log \left( \frac{2}{\epsilon} \right). \]

However, this bound only stands for binary classification problem. By contrast, our high-probability generalization bound holds for any machine learning algorithm.

Also, our bound is strictly tighter than the existing results. All the bounds, including ours, are in the following form,
\[ P \left[ R(A(S)) - \hat{R}_S(A(S)) < a \right] > 1 - b. \]

Apparently, a smaller \( a \) and a smaller \( b \) imply a tighter generalization bound. Thus, our bound improves the current tightest result from two aspects: (1) our bound tightens the term \( a \) from \( 13\epsilon \) to \( 9\epsilon \); and (2) it tightens the term \( b \) from \( 2\epsilon \delta \log \left( \frac{2}{\epsilon} \right) \) to \( 2e^{-\epsilon} \delta \log \left( \frac{2}{\epsilon} \right) \).

Recently, Jung et al. [35] recently defined two new measures to evaluate the generalizability, i.e., in-sample accuracy and distributional accuracy, upon which it also developed the privacy-generalization relationship. Specifically, it proves high-probability bounds on the distributional accuracy of any mechanism that is both differentially private and has a bounded in-sample accuracy. It is an interesting open problem to discover the linkage of the conventional generalization error with the in-sample accuracy and the out-of-sample accuracy.

**Comparison for Theorem 2.** There is only one related work in the literature that presents an on-average generalization bound for multi-database algorithm [58] as follows,
\[ \left| \mathbb{E}_{S \sim \mathcal{D}^k} \left[ \mathbb{E}_{A(S)} \left[ \hat{R}_{S,A(S)}(h_{A(S)}) \right] - \mathbb{E}_{A(S)} \left[ R(h_{A(S)}) \right] \right] \right| \leq k\delta + 2\epsilon. \]

Our bound is tighter by a factor of \( e\epsilon \). Furthermore, the result in [58] stands only for binary classification, while our result apply to all differentially private learning algorithms.

### 4.2 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 generalization-privacy relationship in iterative machine learning algorithms. Detailed proofs are given in Appendix B.

Suppose an iterative machine learning algorithm \( \mathcal{A} \) has \( n \) steps: \( \{Y_i(S)\}_{i=0}^T \), where \( Y_i \) is the learned hypothesis after the \( i \)-th iteration. Also, we define the \( i \)-th iterator as below,
\[ M_i : (Y_{i-1}(S), S) \mapsto Y_i(S). \]

Additionally, we assume that the initial hypothesis \( h_0 \) is independent with the sample \( S \).

Then, we have the following composition theorem.

**Theorem 3** (Composition Theorem I). Suppose an iterative machine learning algorithm \( \mathcal{A} \) has \( T \) steps: \( \{Y_i(S)\}_{i=0}^T \), where \( Y_i \) is the learned hypothesis after the \( i \)-th iteration. Suppose the \( i \)-th iterator
\[ M_i : (Y_{i-1}, S) \mapsto Y_i \]
is \((\epsilon, \delta)\)-differentially private. Then, the algorithm \( \mathcal{A} \) is \((\epsilon', \delta')\)-differentially private. The factor \( \epsilon' \) is as follows,
\[ \epsilon' = \min \{\epsilon_1', \epsilon_2', \epsilon_3'\}, \quad (3) \]
where

\[
\varepsilon_1' = \sum_{i=1}^{T} \varepsilon_i,
\]

\[
\varepsilon_2' = \sum_{i=1}^{T} \frac{(e^{\varepsilon_i} - 1) \varepsilon_i}{e^{\varepsilon_i} + 1} + \sqrt{2 \sum_{i=1}^{T} \varepsilon_i^2 \log \left( e + \sqrt{\frac{\sum_{i=1}^{T} \varepsilon_i^2}{\delta}} \right)},
\]

\[
\varepsilon_3' = \sum_{i=1}^{T} \frac{(e^{\varepsilon_i} - 1) \varepsilon_i}{e^{\varepsilon_i} + 1} + \sqrt{2 \log \left( \frac{1}{\delta} \right) \sum_{i=1}^{T} \varepsilon_i^2},
\]

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 \( \{\alpha_i\}_{i=1}^{T} \in I \),

\[
1 - \prod_{i=1}^{T} \left( 1 - e^{\alpha_i} \frac{\delta_i}{1 + e^{\varepsilon_i}} \right) + 1 - \prod_{i=1}^{T} \left( 1 - \frac{\delta_i}{1 + e^{\varepsilon_i}} \right) + \tilde{\delta},
\]

where

\[
I = \left\{ \{\alpha_i\}_{i=1}^{T} : \sum_{i=1}^{T} \alpha_i = \varepsilon', |\{i : \alpha_i \neq \varepsilon_i, \alpha_i 
eq 0\}| \leq 1 \right\},
\]

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

\[
\delta' = 1 - \left( 1 - e^{\varepsilon} \frac{\delta}{1 + e^{\varepsilon}} \right)^{\left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil} \left( 1 - \frac{\delta}{1 + e^{\varepsilon}} \right)^T \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil + 1 - \left( 1 - \frac{\delta}{1 + e^{\varepsilon}} \right)^T + \tilde{\delta}
\]

\[
= \left( T - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \right) \frac{2\delta}{1 + e^{\varepsilon}} + \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \delta + \tilde{\delta} + O \left( \left( \frac{\delta}{1 + e^{\varepsilon}} \right)^2 \right).
\]

**Proof.** The maximum of \( \delta' \) is achieved when at most \( T - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \) elements \( \alpha_i \neq 0 \). Since \( (1 - x)^n = 1 - nx + O(x^2) \), the \( \delta' \) in Theorem 3 can be estimated as

\[
\delta' = 1 - \left( 1 - e^{\varepsilon} \frac{\delta}{1 + e^{\varepsilon}} \right)^{\left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil} \left( 1 - \frac{\delta}{1 + e^{\varepsilon}} \right)^T \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil + 1 - \left( 1 - \frac{\delta}{1 + e^{\varepsilon}} \right)^T + \tilde{\delta}
\]

\[
= 1 + T \frac{\delta}{1 + e^{\varepsilon}} + \tilde{\delta} + O \left( \left( \frac{\delta}{1 + e^{\varepsilon}} \right)^2 \right)
\]

\[
- \left( 1 - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \right) \frac{\delta}{1 + e^{\varepsilon}} + O \left( \left( \frac{\delta}{1 + e^{\varepsilon}} \right)^2 \right) \left( 1 - \left( T - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \right) \frac{\delta}{1 + e^{\varepsilon}} + O \left( \left( \frac{\delta}{1 + e^{\varepsilon}} \right)^2 \right) \right)
\]

\[
= \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \frac{\delta}{1 + e^{\varepsilon}} + \left( T - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \right) \frac{\delta}{1 + e^{\varepsilon}} + O \left( \left( \frac{\delta}{1 + e^{\varepsilon}} \right)^2 \right) + T \frac{\delta}{1 + e^{\varepsilon}} + \tilde{\delta} + O \left( \left( \frac{\delta}{1 + e^{\varepsilon}} \right)^2 \right)
\]

\[
\approx \left( T - \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \right) \frac{2\delta}{1 + e^{\varepsilon}} + \left\lceil \frac{\varepsilon'}{\varepsilon} \right\rceil \delta + \tilde{\delta}.
\]

The proof is completed. \( \Box \)

When all the iterators \( M_i \) are \( \varepsilon \)-differentially private, we can further tighten the third estimation of \( \varepsilon' \)
in Theorem 3, eq. (3) as the following composition theorem.

**Corollary 3** (Composition Theorem III). Suppose all the iterators \( M_i \) are \( \varepsilon \)-differentially private and all the other conditions in Theorem 3 hold. Then, the algorithm \( A \) is \((\varepsilon', \delta')\)-differentially private that

\[
\varepsilon' = T\left(\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\right) + \sqrt{2\log\left(\frac{1}{\delta}\right)} T\varepsilon^2,
\]

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

where \( \delta'' \) is defined as follows:

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

**Remark 1.** The three composition theorems extend the developed generalization-privacy relationship to iterative machine learning algorithms. At this point, we establish the theoretical foundation for the generalization-privacy relationship for iterative differentially private learning algorithms.

### 4.2.1 Proof Skeleton

Differential privacy measures the distance between the hypotheses learned from neighboring training sample sets which are different by only one single example. It is worth noting that three additional composition theorems are proposed in this section which is, however, weaker than our main results, the aforementioned three.

To approach the differential privacy of an iterative learning algorithm from the differential privacy of every iteration directly would be technically difficult. In this paper, we employ KL divergence as a bridge to relive the technical difficulty, which is defined as follows.

**Definition 5** (KL Divergence; cf. [41]). Suppose two distributions \( P \) and \( Q \) are defined on the same support. Then the KL divergence between \( Q \) and \( P \) is defined by

\[
D_{KL}(P \parallel Q) = \mathbb{E}_P \left( \log \frac{dP}{dQ} \right).
\]

Here, we confuse the notations of distribution and its cumulative distribution function.

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 3.** If \( \mathcal{A} : \mathcal{X} \rightarrow \mathcal{Y} \) is a \( \varepsilon \)-differentially private algorithm, then for every neighbor database (or sample set) \( S \) and \( S' \), the KL divergence of \( \mathcal{A}(S) \) and \( \mathcal{A}(S') \) satisfies

\[
D_{KL}(\mathcal{A}(S) \parallel \mathcal{A}(S')) \leq \frac{e^\varepsilon - 1}{e^\varepsilon + 1}.
\]

This lemma is novel and the proof is technically non-trivial. There are two related results in the literature, which are however looser than ours. Dwork et al. [25] proved an inequality of the KL divergence as follows,

\[
D_{KL}(\mathcal{A}(S) \parallel \mathcal{A}(S')) \leq \varepsilon(e^\varepsilon - 1).
\]

Then, Dwork and Rothblum [24] pushed the upper bound to

\[
D_{KL}(\mathcal{A}(S) \parallel \mathcal{A}(S')) \leq \frac{1}{2}\varepsilon(e^\varepsilon - 1).
\]

Compared with ours, eq. (7) is larger by a factor \((1 + e^\varepsilon)/2\), which can be very large in practice.

We then recall several lemmas necessary to the proof of our composition theorems.
Azuma Lemma [10] which derives an upper bound based on a concentration probability of martingales.

**Lemma 4** (Azuma Lemma; cf. [53], p. 371). Suppose \( \{Y_i\}_{i=1}^T \) is a sequence of random variables that \( Y_i \in [-a_i, a_i] \). Also, suppose \( \{X_i\}_{i=1}^T \) is a sequence of random variables such that

\[
E(Y_i|X_{i-1}, \ldots, X_1) \leq C_i,
\]

where \( \{C_i\}_{i=1}^T \) is a sequence of constant reals. Then, we have the following upper bound of the sum of \( Y_i \)

\[
P \left( \sum_{i=1}^T Y_i \geq \sum_{i=1}^T C_i + t \sqrt{\sum_{i=1}^T a_i^2} \right) \leq e^{-\frac{t^2}{2}}
\]

**Lemma 5** (cf. [24], Lemmas 3.9 and 3.10). For any two distributions \( D \) and \( D' \), there exist distributions \( M \) and \( M' \) such that

\[
\max \{D_\infty(M\parallel M'), D_\infty(M'\parallel M)\} = \max \{D_\infty(D\parallel D'), D_\infty(D'\parallel D)\}
\]

and

\[
D_{KL}(D\parallel D') \leq D_{KL}(M\parallel M') = D_{KL}(M'\parallel M).
\]

**Lemma 6** (cf. [23], Theorem 3.17). We have that

\[
D^\delta_\infty(Y\parallel Z) \leq \varepsilon, \quad D^\delta_\infty(Z\parallel Y) \leq \varepsilon,
\]

if and only if there exist random variables \( Y', Z' \) such that

\[
\Delta(Y\parallel Y') \leq \frac{\delta}{e^\varepsilon + 1}, \quad \Delta(Z\parallel Z') \leq \frac{\delta}{1 + e^\varepsilon},
\]

\[
D_\infty(Y'\parallel Z') \leq \varepsilon, \quad D_\infty(Z'\parallel Y') \leq \varepsilon.
\]

Based on Lemma 5, we proved the following lemma that illustrates that the maximum of \( f \) is achieved when \( \{\alpha_i\}_{i=1}^T \) are at the boundary and will be used in the rest of this section.

**Lemma 7.** The maximum of

\[
f \left( \{\alpha_i\}_{i=1}^T \right) = 1 - \prod_{i=1}^T (1 - \alpha_i A_i),
\]

where \( A_i \) is positive real, subject to

\[
1 \leq \alpha_i \leq c_i, \quad \text{(here } c_i A_i \leq 1), \quad \prod_{i=1}^T \alpha_i = c,
\]

is achieved at the point when the cardinality follows the inequality below:

\[
|\{i : \alpha_i \neq c_i \text{ and } \alpha_i \neq 1\}| \leq 1.
\]  

(8)

Based on Lemma 3, we can prove the following composition theorem for \( \varepsilon \)-differential privacy as a preparation theorem of the general case.

**Theorem 4** (Composition Theorem IV). Suppose an iterative machine learning algorithm \( \mathcal{A} \) has \( T \) steps: \( \{Y_i(S)\}_{i=1}^T \). Specifically, \( \mathcal{M}_i : (Y_{i-1}(S), S) \mapsto Y_i(S) \) is the \( i \)-th iterator. Assume that \( Y_0(S) \) is the initial hypothesis (which does not depend on \( S \)). If for any fixed observation \( y_{i-1} \) of the variable \( Y_{i-1} \),
\[ M_i(y_{i-1}, S) \text{ is } \varepsilon_i\text{-differentially private, then } \{Y_i(S)\}_{i=0}^T \text{ is } (\varepsilon', \delta')\text{-differentially private that} \]
\[ \varepsilon' = \sqrt{2 \log \left( \frac{1}{\delta} \right) \left( \sum_{i=1}^T \varepsilon_i^2 \right) + \sum_{i=1}^T \varepsilon_i \varepsilon_i' - 1} + 1. \]

Based on Lemmas 6 and 7, we can prove our composition theorems for \((\varepsilon, \delta)\)-differential privacy. We first prove a composition algorithm of \((\varepsilon, \delta)\)-differential privacy whose estimate of \(\varepsilon'\) is somewhat looser than the existing results.

**Theorem 5** (Composition Theorem V). Suppose an iterative machine learning algorithm \(A\) has \(T\) steps: \(\{Y_i(S)\}_{i=1}^T\). Specifically, let the \(i\)-th iterator be as follows,
\[ M_i : (Y_{i-1}(S), S) \rightarrow Y_i(S). \tag{9} \]
Assume that \(Y_0\) is the initial hypothesis (which does not depend on \(S\)). If for any fixed observation \(y_{i-1}\) of the variable \(Y_{i-1}\), \(M_i(y_{i-1}, S)\) is \((\varepsilon_i, \delta_i)\)-differentially private \((i \geq 1)\), then \(\{Y_i(S)\}_{i=0}^T\) is \((\varepsilon', \delta')\) differentially private with
\[ \varepsilon' = \sqrt{2 \log \left( \frac{1}{\delta} \right) \left( \sum_{i=1}^T \varepsilon_i^2 \right) + \sum_{i=1}^T \varepsilon_i \varepsilon_i' - 1} + 1, \]
\[ \delta' = \max_{\{\alpha_i\}_{i=1}^T \in I} \left[ 1 - \prod_{i=1}^T \left( 1 - e^{\alpha_i} \frac{\delta_i}{1 + \varepsilon_i} \right) + 1 - \prod_{i=1}^T \left( 1 - \frac{\delta_i}{1 + \varepsilon_i} \right) + \delta, \right] \]
where \(I\) is defined as the set of \(\{\alpha_i\}_{i=1}^T\) such that
\[ \sum_{i=1}^T \alpha_i = \varepsilon', \{i : \alpha_i \neq \varepsilon_i \text{ and } \alpha_i \neq 0\} \leq 1. \]

Then, for any iterative learning algorithm whose \(i\)-th iteration is \((\varepsilon_i, \delta_i)\)-differentially private, we prove there exists an iterative learning algorithm \(\tilde{A}\) whose \(i\)-th iteration is \(\varepsilon'_i\)-differentially private and the distance between them is controlled by constants relying on \(\varepsilon_i\) and \(\delta_i\) measured by \(D_\infty(X\|Y)\), \(D_\delta(X\|Y)\), and \(\Delta(X\|Y)\) defined as follows.\(^1\)

**Definition 6** (Max Divergence; cf. [23], Definition 3.6). For any random variables \(X\) and \(Y\), the max divergence between \(X\) and \(Y\) is defined as
\[ D_\infty(X\|Y) = \max_{S \subseteq \text{Supp}(X)} \left[ \log \frac{\mathbb{P}(X \in S)}{\mathbb{P}(Y \in S)} \right]. \]
Furthermore, the \(\delta\)-approximate max divergence from \(X\) to \(Y\) is defined as
\[ D_\delta(X\|Y) = \max_{S \subseteq \text{Supp}(X) : \mathbb{P}(Y \in S) \geq \delta} \left[ \log \frac{\mathbb{P}(X \in S) - \delta}{\mathbb{P}(Y \in S)} \right]. \]

**Definition 7** (Statistical Distance; cf. [23]). For any random variables \(X\) and \(Y\), the statistical distance between \(X\) and \(Y\) is defined as
\[ \Delta(X\|Y) = \max_{S} |\mathbb{P}(X \in S) - \mathbb{P}(Y \in S)|. \]

Based on Theorem 5, we can calculate the \((\varepsilon', \delta')\)-differential privacy of algorithm \(\tilde{A}\). Eventually, we can calculate the \((\varepsilon, \delta)\)-differential privacy that algorithm \(A\) as a weaker version of Theorem 3 in which
\[^1\]These two pseudo-distances have implicitly close relationships with the "differential loss" (see xxx) and thus helpful in approximating differential privacy.
\( \varepsilon' \) is replaced by \( \varepsilon_2 \) wherein as follows.

Applying the Chernoff concentration inequality [53] and Theorem 3.5 in [36], we eventually extend the weaker version to Theorem 3. Theorem 3.5 in [36] relies on a term privacy area defined wherein. A larger privacy area corresponds to a worse privacy-preservation. In this paper, we make a novel contribution that proves the moment generating function of the following random variable represents the worst case,

\[
\log \left( \frac{P(\cap_i Y_i(S) \in B_i)}{P(\cap_i Y_i(S') \in B_i)} \right),
\]

where \( Y_i(S) \) and \( Y_i(S') \) are the mechanisms achieving the largest privacy area.

### 4.2.2 Comparison with Existing Results

Our composition theorem is strictly tighter than the existing results. A classic composition theorem addressing the \((\varepsilon', \delta')\)-differential privacy of an iterative algorithm is as follows (see [23], Theorem 3.20 and Corollary 3.21, pp. 49-52),

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

where \( \tilde{\delta} \) is an arbitrary positive real, \((\varepsilon', \delta')\) is the differential privacy of the whole algorithm, and \((\varepsilon_i, \delta_i)\) is the differential privacy of the \(i\)-th iteration. Both the estimates of \(\varepsilon'\) and \(\delta'\) are strictly looser than ours: (1) the estimate of \(\varepsilon'\) (eq. 3) is strictly smaller than the \(\varepsilon_3'\) in our estimate (see eq. 5). Specifically, the \(i\)-th term,

\[
\varepsilon_i \left( \frac{e^{\varepsilon_i} - 1}{e^{\varepsilon_i} + 1} \right),
\]

of the series,

\[
\sum_{i=1}^{T} \varepsilon_i \left( \frac{e^{\varepsilon_i} - 1}{e^{\varepsilon_i} + 1} \right),
\]

in our estimate is smaller than the corresponding term by \( (e^{\varepsilon_i} + 1) \); and (2) the difference between our estimate of \(\delta'\) and the corresponding term is as follows,

\[
\delta e^{\varepsilon - 1} \left( T - \left\lceil \frac{\varepsilon'}{\tilde{\varepsilon}} \right\rceil \right).
\]

Currently, the tightest approximation is as follows [36],

\[
\varepsilon' = \min \{ \varepsilon_1', \varepsilon_2', \varepsilon_3' \}, \quad \delta' = 1 - (1 - \tilde{\delta})^T (1 - \tilde{\delta}).
\]

where

\[
\varepsilon_1' = T \varepsilon,
\]

\[
\varepsilon_2' = \frac{(e^\varepsilon - 1) e^T}{e^\varepsilon + 1} + \varepsilon \sqrt{2T \log \left( e + \frac{\sqrt{T e^2}}{\delta} \right)},
\]

\[
\varepsilon_3' = \frac{(e^\varepsilon - 1) e^T}{e^\varepsilon + 1} + \varepsilon \sqrt{2T \log \left( \frac{1}{\delta} \right)}.
\]

The estimate of the \(\varepsilon'\) is the same as ours, while their \(\delta'\) is also larger than ours approximately by

\[
\delta \frac{e^\varepsilon - 1}{e^\varepsilon + 1} \left( T - \left\lceil \frac{\varepsilon'}{\tilde{\varepsilon}} \right\rceil \right).
\]
In many situations, the number of iterations $T$ is overwhelmingly larger than $\lceil \epsilon' / \epsilon \rceil$, which guarantees that our advantage is significant.

5 Applications

Our theories apply to a wide spectrum of machine learning algorithms. This section implements them to two popular regimes as examples: (1) stochastic gradient Langevin dynamics as an example of the stochastic gradient Monte Carlo scheme [75, 47, 78]; and (2) agnostic federated learning [28, 54] from the client view. The theories help deliver $O(\sqrt{\log N/N})$ on-average generalization bounds and $O(\sqrt{\log N/N})$ high-probability generalization bounds for the two schemes. Detailed proofs are given in Section C.

5.1 Application in SGLD

Bayesian learning aims to obtain the posterior of model parameters of parametric machine learning models and then approach the best parameter. However, the analytic expression of the posterior is in most real-world cases. To solve this problem, Markov chain Monte Carlo (MCMC) methods are employed to infer the posterior [31, 26, 18]. However, MCMC can be prohibitively time-consuming on large-scale data. To address this issue, stochastic gradient Markov chain Monte Carlo (SGMCMC; [47]) introduces stochastic gradient estimate [64] into Bayesian learning. The family of SGMCMC algorithms includes stochastic gradient Langevin dynamics (SGLD; [75]), stochastic gradient Riemannian Langevin dynamics (SGRLD; [60]), stochastic gradient Fisher scoring (SGFS; [2]), stochastic gradient Hamiltonian Monte Carlo (SGHMC; [14]), stochastic gradient Nosé-Hoover Thermostat (SGNHT; [17]), etc. SGMCMC has been applied to many areas, including topic model [43, 79] and Bayesian neural network [46, 66]. This paper analyses SGLD as an example of the SGMCMC scheme.

An example of SGLD is shown by the following chart.

Algorithm 1: Stochastic Gradient Langevin Dynamics

**Require:** Samples $S = \{z_1, ..., z_N\}$, Gauss noise variance $\sigma$, size of mini-batch $\tau$, iteration steps $T$, learning rate $\{\eta_1, ..., \eta_T\}$, Regularization function $r$, Lipschitz constant $L$ of loss $l$.

1: Initialize $\theta_0$ randomly.
2: For $t = 1$ to $T$ do:
3: Randomly sample a mini-batch $B$ of size $\tau$;
4: Sample $g_t$ from $\mathcal{N}(0, \sigma^2 I)$;
5: Update $\theta_t \leftarrow \theta_{t-1} - \eta_t \left[ \frac{1}{\tau} \nabla r(\theta_{t-1}) + \frac{1}{\tau} \sum_{z \in B} \nabla l(z|\theta_{t-1}) + g_t \right]$.

The following theorem provides an estimation of the differential privacy and generalization bounds of SGLD.

**Theorem 6.** SGLD is $(\epsilon', \delta')$-differentially private. The factor $\epsilon'$ is as follows,

$$\epsilon' = \sqrt{8 \log \left( \frac{1}{\delta} \right) \left( \frac{T^2 \bar{\epsilon}^2}{N^2} \right) + 2T \frac{\tau}{N} \frac{e^{2\bar{\epsilon} \bar{\epsilon}}}{\epsilon^2 \bar{\epsilon}} - 1},$$

and the factor $\delta'$ is as follows,

$$\delta' = 1 - \left( 1 - e^{2\bar{\epsilon} \bar{\epsilon}} \frac{\tau \delta}{1 + e^{2\bar{\epsilon} \bar{\epsilon}}} \right)^{\frac{\bar{\epsilon}'}{\bar{\epsilon} T}} \left( 1 - \frac{\tau \delta}{1 + e^{2\bar{\epsilon} \bar{\epsilon}}} \right)^{T - \frac{\bar{\epsilon}'}{\bar{\epsilon} T}} + 1 - \left( 1 - \frac{\tau \delta}{1 + e^{2\bar{\epsilon} \bar{\epsilon}}} \right)^T$$

$$+ e^{-\frac{\epsilon' + \frac{\epsilon}{2} \bar{\epsilon}}{\bar{\epsilon}}} \left( \frac{1}{1 + e^{\frac{\epsilon}{2} \bar{\epsilon}}} \left( \frac{2 \bar{\epsilon} T \bar{\epsilon}}{N T \bar{\epsilon} - \bar{\epsilon}'} \right)^T \left( \frac{2 \bar{\epsilon} T \bar{\epsilon} + \epsilon'}{N T \bar{\epsilon} - \bar{\epsilon}'} \right)^{\frac{\epsilon'}{2} \bar{\epsilon} + \frac{T}{2}} \right),$$

14
where
\[
\tilde{\varepsilon} = \frac{2\sqrt{2}L\sigma}{\tau} \sqrt{\log \frac{1}{\delta}} + \frac{\frac{1}{\tau} L^2}{2\sigma^2}.
\]
Additionally, an on-average generalization bound and a high-probability bound are delivered combined with Theorem 1.

**Remark 2.** Regarding the dependence on the training sample size \( N \), we have
\[
P \left[ \hat{R}_S(A(S)) \leq O \left( \frac{T}{\sqrt{N}} \right) \right] \geq 1 - O \left( \frac{T}{\sqrt{N}} \right).
\]

We further prove that SGLD is PAC-learnable under the following assumption.

**Assumption 1.** There exist constants \( K_1 > 0, K_2, T_0, \) and \( N_0 \), such that, for \( T > T_0 \) and any \( N > N_0 \), we have
\[
\hat{R}_S(A(S)) \leq \exp(-K_1T + K_2).
\]

This assumption is justified as the training loss can almost surely achieve near-0 in modern machine learning.

Some existing works have also studied the generalizability and privacy-preservation of SGLD. Wang et al. [74] prove that SGLD has "privacy for free" without injecting noise. Specifically, the authors proved that (1) SGLD is \((\varepsilon, \delta)\)-differentially private if
\[
T > \frac{\varepsilon^2 N}{32\tau \log(2/\delta)};
\]
(2) SGHMC is \((\varepsilon, \delta)\)-differentially private if
\[
2(A - \hat{B})/h_t > \frac{128nTL^2}{\tau \varepsilon^2} \log(2/\delta)I_N;
\]
and (3) SGHMC is \((\varepsilon, \delta)\)-differentially private if
\[
\frac{2a}{\eta_t} \geq \frac{128NTL^2}{\tau \varepsilon^2} \log \left( \frac{2NT}{\tau \delta} \right) \log \left( \frac{1}{\delta} \right).
\]

Pensia et al. [61] analyzes the generalizability of SGLD via information theory. Some works also deliver generalization bounds via algorithmic stability or the PAC-Bayesian framework [30, 63, 55].

### 5.2 Application in Agnostic Federated Learning

Smart mobile phones continuously collect massive amounts of information including text messages, GPS records, pictures, etc. The data is highly valuable to the industry but also highly private. This phenomenon creates the dilemma of how to extract population knowledge while protecting individual privacy. Federated learning [67, 40, 51, 77] adapts a decentralized regime that does not collect the raw information collected in the personal terminals to a centralized cloud. Instead, it trains models on the edge (e.g., mobile phones) or just collect processed data (e.g., noised data). This mechanism sheds light on solving the dilemma. The following algorithm designed by [28, 54] further enhances the privacy preservation to protect client identity from differential attacks.
Algorithm 2 Differentially Private Federated Learning

**Require:** Clients \(\{c_1, \ldots, c_N\}\), Gaussian noise variance \(\sigma\), size of mini-batch \(\tau\), iteration steps \(T\), upper bound \(L\) of the step size.

1. Initialize \(\theta_0\) randomly.
2. For \(t = 1\) to \(T\) do:
   3. Randomly sample a mini-batch of clients of size \(\tau\);
   4. Randomly sample \(b_t\) from \(\mathcal{N}(0, L^2 \sigma^2 I)\);
   5. Central curator distributes \(\theta_{t-1}\) to the clients in the mini-batch \(B\);
   6. Update \(\theta_{t+1} \leftarrow \theta_t + \left(\frac{1}{B} \sum_{i \in B} \text{ClientUpdate}(c_i, \theta_t) + b_t\right)\).

The following theorem provides an estimation of the differential privacy and generalization bounds of agnostic federated learning.

**Theorem 7** (Differential Privacy and Generalization Bounds of Differentially Private Federated Learning). Agnostic federated learning is \((\varepsilon', \delta')\)-differentially private. The factor \(\varepsilon'\) is as follows,

\[
\varepsilon' = \sqrt{8 \log \left(\frac{1}{\delta} \right)} \left(\frac{\tau^2}{N^2} T \varepsilon^2\right) + 2T \frac{\tau}{N} \frac{e^{2 \tau \varepsilon} - 1}{e^{2 \tau \varepsilon} + 1},
\]

and the factor \(\delta'\) is defined as follows,

\[
\delta' = 1 - \left(1 - e^{2 \tau \varepsilon} \frac{\frac{\tau}{N} \delta}{1 + e^{2 \tau \varepsilon}} \right)^{\frac{\tau T}{N \varepsilon}} \left(1 - \frac{\frac{\tau}{N} \delta}{1 + e^{2 \tau \varepsilon}} \right)^T + 1 - \left(1 - \frac{\tau}{N} \delta\right)^T
\]

\[
+ e^{-\frac{T - \varepsilon T}{2}} \frac{1}{1 + e^{2 \tau \varepsilon}} \left(\frac{1}{\frac{T}{N} \varepsilon T} - \varepsilon'\right)^T \left(\frac{1}{\frac{T}{N} \varepsilon T} + \varepsilon'\right)^T e^{-\frac{T - \varepsilon T}{2}}\frac{\varepsilon'}{2 \sigma^2},
\]

where

\[
\tilde{\varepsilon} = \sqrt{\frac{4 \sigma^2}{\tau} \log \frac{1}{\eta} + \frac{1}{\tau^2}}.
\]

Additionally, an on-average generalization bound and a high-probability bound are delivered combined with Theorem 1.

**Remark 3.** The on-average generalization bound is \(O(\sqrt{\log N/N})\) and the high-probability generalization bound is \(O(\sqrt{\log N/N})\).

## 6 Conclusion

This paper studies the relationships between generalization and privacy preservation in two steps. We first establish the generalization-privacy relationship 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-database algorithms. This high-probability generalization bound delivers a PAC-learnable guarantee for differentially private learning 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 the generalization-privacy relationship of iterative learning algorithms. Compared with existing works, our theoretical results are strictly tighter and apply to a wider application domain. We then use them to study the generalization-privacy relationship in stochastic gradient Langevin dynamics (SGLD), as an example of the stochastic gradient Markov chain Monte Carlo, and agnostic federated learning from the client view. We obtain the approximation of differential privacy of SGLD and agnostic federated learning which further leads to high-probability bounds that do not explicitly rely on the model size which would be prohibitively large in many deep models.
References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In Proceedings of ACM SIGSAC Conference on Computer and Communications Security, pages 308–318, 2016.

[2] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. arXiv preprint arXiv:1206.6380, 2012.

[3] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman. Local differential privacy for deep learning. IEEE Internet of Things Journal, 2019.

[4] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In Advances in Neural Information Processing Systems, pages 6240–6249, 2017.

[5] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. The Journal of Machine Learning Research, 20(63):1–17, 2019.

[6] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. The Journal of Machine Learning Research, 3(Nov):463–482, 2002.

[7] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In Proceedings of Annual ACM Symposium on Theory of Computing, pages 1046–1059, 2016.

[8] A. Beimel, S. P. Kasiviswanathan, and K. Nissim. Bounds on the sample complexity for private learning and private data release. In Proceedings of Theory of Cryptography Conference, pages 437–454, 2010.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. Journal of the ACM, 36(4):929–965, 1989.

[10] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities: A nonasymptotic theory of independence. Oxford University Press, 2013.

[11] O. Bousquet and A. Elisseeff. Stability and generalization. The Journal of Machine Learning Research, 2(Mar):499–526, 2002.

[12] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Proceedings of Theory of Cryptography Conference, pages 635–658, 2016.

[13] K. Chaudhuri, J. Imola, and A. Machanavajjhala. Capacity bounded differential privacy. arXiv preprint arXiv:1907.02159, 2019.

[14] T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In Proceedings of International Conference on Machine Learning, pages 1683–1691, 2014.

[15] P. Cuff and L. Yu. Differential privacy as a mutual information constraint. In Proceedings of ACM SIGSAC Conference on Computer and Communications Security, pages 43–54, 2016.

[16] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. IEEE Transactions on Information Theory, 25(5):601–604, 1979.

[17] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In Advances in Neural Information Processing Systems, pages 3203–3211, 2014.

[18] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. Physics Letters B, 195(2):216–222, 1987.

[19] R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. Journal of Functional Analysis, 1(3):290–330, 1967.

[20] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, Automata, Languages and Programming, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[21] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In Proceedings of Annual ACM Symposium on Theory of Computing, pages 117–126, 2015.

[22] C. Dwork and D. K. Mulligan. It’s not privacy, and it’s not fair. Stanford Law Review Online, 66:35, 2013.
[23] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[24] C. Dwork and G. N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[25] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Proceedings of IEEE Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

[26] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[27] J. Geumlek, S. Song, and K. Chaudhuri. Renyi differential privacy mechanisms for posterior sampling. In *Advances in Neural Information Processing Systems*, pages 5289–5298, 2017.

[28] R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. In *Advances in Neural Information Processing Systems*, 2017.

[29] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Proceedings of Annual Conference on Learning Theory*, pages 297–299, 2018.

[30] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of International Conference on Machine Learning*, pages 1225–1234, 2016.

[31] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[32] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[33] F. He, T. Liu, and D. Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, pages 1143–1152, 2019.

[34] F. He, B. Wang, and D. Tao. Piecewise linear activations substantially shape the loss surfaces of neural networks. In *International Conference on Learning Representations*, 2020.

[35] C. Jung, K. Ligett, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and M. Shenfeld. A new analysis of differential privacy’s generalization guarantees. *arXiv preprint arXiv:1909.03577*, 2019.

[36] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[37] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

[38] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[39] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457. Springer, 2000.

[40] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. In *Advances in Neural Information Processing Systems Workshop on Private Multi-Party Machine Learning*, 2016.

[41] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[42] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *Proceedings of International Conference on Machine Learning*, pages 2815–2824, 2018.

[43] H. Larochelle and S. Lauly. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems*, pages 2708–2716, 2012.

[44] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 888–896, 2019.

[45] J. Liao, L. Sankar, V. Y. Tan, and F. du Pin Calmon. Hypothesis testing under mutual information privacy constraints in the high privacy regime. *IEEE Transactions on Information Forensics and Security*, 13(4):1058–1071, 2017.

[46] C. Louizos and M. Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of International Conference on Machine Learning*, pages 2218–2227, 2017.
[47] Y.-A. Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient mcmc. In Advances in Neural Information Processing Systems, pages 2917–2925, 2015.

[48] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate bayesian inference. The Journal of Machine Learning Research, 18(1):4873–4907, 2017.

[49] D. A. McAllester. Pac-bayesian model averaging. In Annual Conference of Learning Theory, volume 99, pages 164–170, 1999.

[50] D. A. McAllester. Some pac-bayesian theorems. Machine Learning, 37(3):355–363, 1999.

[51] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al. Communication-efficient learning of deep networks from decentralized data. In Proceedings of International Conference on Artificial Intelligence and Statistics, 2017.

[52] I. Mironov. Rényi differential privacy. In Proceedings of IEEE Computer Security Foundations Symposium, pages 263–275, 2017.

[53] M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of machine learning. MIT Press, 2018.

[54] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In Proceedings of International Conference on Machine Learning, pages 4615–4625, 2019.

[55] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In Proceedings of Annual Conference On Learning Theory, pages 605–638, 2018.

[56] M. T. Musavi, K. H. Chan, D. M. Hummels, and K. Kalantri. On the generalization ability of neural network classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(6):659–663, 1994.

[57] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In International Conference on Learning Representations, 2018.

[58] K. Nissim and U. Stemmer. On the generalization properties of differential privacy. CoRR, abs/1504.05800, 2015.

[59] L. Oneto, S. Ridella, and D. Anguita. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters, 89:31–38, 2017.

[60] S. Patterson and Y. W. Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In Advances in Neural Information Processing Systems, pages 3102–3110, 2013.

[61] A. Pensia, V. Jog, and P.-L. Loh. Generalization error bounds for noisy, iterative algorithms. In Proceedings of IEEE International Symposium on Information Theory, pages 546–550, 2018.

[62] F. Pittaluga and S. J. Koppal. Pre-capture privacy for small vision sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11):2215–2226, 2016.

[63] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In Proceedings of Annual Conference on Learning Theory, pages 1674–1703, 2017.

[64] H. Robbins and S. Monro. A stochastic approximation method. The Annals of Mathematical Statistics, pages 400–407, 1951.

[65] W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. The Annals of Statistics, pages 506–514, 1978.

[66] W. Roth and F. Pernkopf. Bayesian neural networks with weight sharing using dirichlet processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(1):246–252, 2018.

[67] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In Proceedings of ACM SIGSAC Conference on Computer and Communications Security, pages 1310–1321, 2015.

[68] Z. Tu, F. He, and D. Tao. Understanding generalization in recurrent neural networks. In International Conference on Learning Representations, 2020.

[69] K. Uematsu and Y. Lee. Statistical optimality in multipartite ranking and ordinal regression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(5):1080–1094, 2014.

[70] V. Vapnik. Estimation of dependences based on empirical data. Springer Science & Business Media, 2006.

[71] V. Vapnik. The nature of statistical learning theory. Springer Science & Business Media, 2013.
[72] S. Verma and Z.-L. Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1539–1548, 2019.

[73] W. Wang, L. Ying, and J. Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, 2016.

[74] Y.-X. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *Proceedings of International Conference on Machine Learning*, pages 2493–2502, 2015.

[75] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of International Conference on Machine Learning*, pages 681–688, 2011.

[76] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):187–193, 2011.

[77] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):12:1–12:19, 2019.

[78] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2018.

[79] H. Zhang, B. Chen, Y. Cong, D. Guo, H. Liu, and M. Zhou. Deep autoencoding topic model with scalable hybrid bayesian inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
A Proofs of Generalization Bounds via Differential Privacy

This appendix collects all the proofs of the generalization bounds. It is organized as follows: (1) Appendix A.1 proves Theorem 2; and (2) Appendix A.2 proves Theorem 1.

A.1 Proof of Theorem 2

Proof of Theorem 2. The left side of eq. (2) can be rewritten as

\[
\begin{align*}
\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{\mathcal{A}(\mathcal{S})} \left[ \mathcal{R}_{\mathcal{S}, \mathcal{A}(\mathcal{S})} \left( h_{\mathcal{A}(\mathcal{S})} \right) \right] \right] &= \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{\mathcal{A}(\mathcal{S})} \left[ \mathbb{E}_{\mathcal{z} \sim \mathcal{D}^N} \left[ \mathbb{E}_{\mathcal{z} \sim \mathcal{S}, \mathcal{A}(\mathcal{S})} \left[ l \left( h_{\mathcal{A}(\mathcal{S}), \mathcal{z}_1} \right) \right] \right] \right] \right] \\
&\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\ quad
Therefore, we have
\[
E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] \leq k\delta + e^{\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) \right] \right].
\] (12)

Rearranging eq. (12), we have
\[
e^{-\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] \leq e^{-\varepsilon} k\delta + e^{\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) \right] \right]
\]
\[
- E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) \right] \right] \leq e^{-\varepsilon} k\delta - e^{-\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right],
\]
which leads to
\[
E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] \geq e^{-\varepsilon} k\delta - e^{-\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] - e^{-\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) \right] \right] \leq 1 - e^{-\varepsilon} + e^{-\varepsilon} k\delta.
\]
The other side of the inequality can be similarly obtained. The proof is completed. □

A.2 Proofs of Theorem 1

We then prove Lemmas 1 and 2 to prove Theorem 1. The proofs are inspired by [58].

Proof of Lemma 1. By Corollary 1, we have that
\[
E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] \leq e^{-\varepsilon} k\delta + e^{\varepsilon} E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) \right] \right].
\]

Since \( \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \geq 0 \), we have that for any \( \alpha > 0 \),
\[
E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] \geq E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \right] \right] I_{\hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \geq \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) + \alpha}
\]
\[
\geq E_{\tilde{S} \sim D^{kN}} \left[ E_{\mathcal{A}(\tilde{S})} \left[ \alpha + \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) \right] \right] I_{\hat{R}_{S_{\mathcal{A}(\hat{S})}} (h_{\mathcal{A}(\hat{S})}) \geq \mathcal{R} (h_{\mathcal{A}(\tilde{S})}) + \alpha}
\].
Furthermore, by splitting $\mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ \mathcal{R} \left( h_{A(S)} \right) \right] \right]$ into two parts, we have

$$
\mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ \hat{R}_{S_{A(S)}} \left( h_{A(S)} \right) \right] \right] - \mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ \mathcal{R} \left( h_{A(S)} \right) \right] \right] \geq \mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ (\alpha + \mathcal{R} \left( h_{A(S)} \right)) \mathbb{I}_{h_{A(S)} \leq h_{A(S)} + \alpha} \right] \right] - \mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ \mathcal{R} \left( h_{A(S)} \right) \mathbb{I}_{h_{A(S)} \leq h_{A(S)} + \alpha} \right] \right] + \mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ \mathcal{R} \left( h_{A(S)} \right) \mathbb{I}_{h_{A(S)} \geq h_{A(S)} + \alpha} \right] \right] - \mathbb{E}_{S \sim \mathcal{D}^{kN}} \left[ \mathbb{E}_{A(S)} \left[ \mathcal{R} \left( h_{A(S)} \right) \mathbb{I}_{h_{A(S)} \geq h_{A(S)} + \alpha} \right] \right] \geq \alpha \mathcal{P} \left( \hat{R}_{S_{A(S)}} \left( h_{A(S)} \right) > \mathcal{R} \left( h_{A(S)} \right) + \alpha \right) - \mathcal{P} \left( \hat{R}_{S_{A(S)}} \left( h_{A(S)} \right) \leq \mathcal{R} \left( h_{A(S)} \right) + \alpha \right).
$$

Let $\alpha = e^{-\varepsilon}k\delta + 3\varepsilon$, we have

$$
\mathcal{P} \left( \hat{R}_{S_{A(S)}} \left( h_{A(S)} \right) \leq \mathcal{R} \left( h_{A(S)} \right) + \alpha \right) \leq \frac{\alpha - (e^{-\varepsilon}k\delta + \varepsilon)}{1 + \alpha} \leq \varepsilon.
$$

The proof is completed. \( \square \)

**Proof of Lemma 2.** Construct algorithm $\mathcal{B}$ with input $S = \{S_i\}_{i=1}^k$ and $T$ (where $S_i, T \in \mathcal{Z}^N$) as follows:

**Step 1.** Run $\mathcal{A}$ on $S_i, i = 1, \cdots, k$. Denote the output as $h_i = \mathcal{A}(S_i)$.

**Step 2.** Let utility function as $q(S, T, i) = N \left( \hat{R}_{S_i}(h_i) - \mathcal{R}(h_i) \right)$. Apply exponential mechanism $\mathcal{M}(h_i, S, T)$ with differential privacy $\varepsilon$ to $q$ and return the output.

We then prove that $\mathcal{B}$ satisfies

$$
\mathbb{P} \left[ l \left( h_{B(S)}, S_{B(S)} \right) \leq \mathcal{R} \left( h_{B(S)} \right) + ke^{-\varepsilon}\delta + 3\varepsilon \right] < \varepsilon.
$$

By eq. (??), we have that

$$
\mathbb{P} \left( \forall i, \hat{R}(\mathcal{A}(S_i)) \leq e^{-\varepsilon}k\delta + 8\varepsilon + \mathcal{R}(\mathcal{A}(S_i)) \right) \leq \left( 1 - \frac{e^{-\varepsilon}\delta}{\varepsilon} \ln \left( \frac{2}{\varepsilon} \right) \right)^k,
$$

which leads to

$$
\mathbb{P} \left( \exists i, \hat{R}(\mathcal{A}(S_i)) > e^{-\varepsilon}k\delta + 8\varepsilon + \mathcal{R}(\mathcal{A}(S_i)) \right) > 1 - \left( 1 - \frac{1}{k} \ln \left( \frac{2}{\varepsilon} \right) \right)^k \geq 1 - \frac{\varepsilon}{2}. \quad (13)
$$

Furthermore, since $T$ is independent with $\hat{S}$, by Hoeffding inequality, we have that

$$
\mathbb{P} \left( \forall i, |l(h_i, T) - \mathcal{R}(h_i)| \leq \frac{\varepsilon}{2} \right) \geq (1 - e^{-\varepsilon^2/2N})^k \geq 1 - \frac{\varepsilon}{8}. \quad (14)
$$
Therefore, combining eq. (13) and eq. (14),

\[ P \left( \exists i, \hat{R}(A(S_i)) > e^{-\varepsilon}k\delta + \frac{15}{2}\varepsilon + l(h_i, T) \right) > 1 - \frac{5\varepsilon}{8}. \]

Since \( q \) has sensitivity 1, we have that fixed \( h_i \)

\[ P \left( M(h_i, \tilde{S}, T) \leq \text{OPT}(q(\tilde{S}, T, i)) - N\varepsilon \right) \geq 1 - \frac{\varepsilon}{4}, \]

which leads to

\[ P \left( \hat{R}_{S_i, \hat{S}(i)}(h_{\hat{S}(i)}) > e^{-\varepsilon}k\delta + 13\varepsilon + \hat{R}_T(h_{\hat{S}(i)}) \right) > 1 - \frac{7\varepsilon}{8}. \]

Then, using eq. (14) again, we have

\[ P \left( \hat{R}_{S_i, \hat{S}(i)}(h_{\hat{S}(i)}) > e^{-\varepsilon}k\delta + 6\varepsilon + R(h_{\hat{S}(i)}) \right) > 1 - \varepsilon. \]

\[ \square \]

## B Proofs of Composition Theorems

This section proves the composition theorems. It is organized as follows: Section B.1 proves a preparation lemma on the KL divergence \( D_{KL}(A(S)||A(S')) \) between the hypotheses \( A(S) \) and \( A(S') \); based on this lemma Section B.2 proves a composition theorem of \( \varepsilon \)-differential privacy; Section B.3 extends the composition theorem to \((\varepsilon, \delta)\)-differential privacy; Section B.4 further tightens the estimate of \( \delta' \) under some assumptions; and Section B.5 analyses the tightness of this estimation.

### B.1 Proof of Lemma 3

**Proof of Lemma 3.** By Lemma 5, we have a random variable \( M(S) \) and \( M(S') \), which satisfies

\[ D_{\infty}(M(S)||M(S')) \leq \varepsilon, \ D_{\infty}(M(S')||M(S)) \leq \varepsilon, \]

and

\[ D_{KL}(A(S)||A(S')) \leq D_{KL}(M(S)||M(S')) = D_{KL}(M(S')||M(S)). \]  \hspace{1cm} (15)

Therefore, we only need to derive a bound for \( D_{KL}(M(S)||M(S')) \).

By direct calculation,

\[
D_{KL}(M(S)||M(S')) =
\begin{align*}
\frac{1}{2} & \left[ D_{KL}(M(S)||M(S')) + D_{KL}(M(S')||M(S)) \right] \\
& = \frac{1}{2} \int \log \frac{dP(M(S))}{dP(M(S'))} dP(M(S)) + \frac{1}{2} \int \log \frac{dP(M(S'))}{dP(M(S))} dP(M(S')) \\
& = \frac{1}{2} \int \log \frac{dP(M(S))}{dP(M(S'))} d[P(M(S)) - P(M(S'))] \\
& \quad + \frac{1}{2} \int \left( \log \frac{dP(M(S'))}{dP(M(S))} + \log \frac{dP(M(S))}{dP(M(S'))} \right) dP(M(S')) \\
& = \frac{1}{2} \int \log \frac{dP(M(S))}{dP(M(S'))} d[P(M(S)) - P(M(S'))] + \frac{1}{2} \int \log 1 dP(M(S')) \\
& = \frac{1}{2} \int \log \frac{dP(M(S))}{dP(M(S'))} d[P(M(S)) - P(M(S'))] \hspace{1cm} (16)
\end{align*}
\]

where eq. (*) comes from eq. (15).
We argue that it is the distribution that maximizes $k(M(S))$ almost surely. When $k$ becomes as following, corresponds the distribution $P_k$.

We now analyse the last integration in eq. (16). Define

$$k(y) \triangleq \frac{d\mathbb{P}(M(S) = y)}{d\mathbb{P}(M(S') = y)} - 1.$$  \hspace{1cm} (17)

Therefore,

$$k(y)d\mathbb{P}(M(S') = y) = d\mathbb{P}(M(S) = y) - d\mathbb{P}(M(S') = y).$$ \hspace{1cm} (18)

Additionally,

$$\mathbb{E}_{M(S')}k(M(S')) = \int_{y \in \mathcal{H}} k(y)d\mathbb{P}(M(S') = y)$$
$$= \int_{y \in \mathcal{H}} d(\mathbb{P}(M(S) = y) - d\mathbb{P}(M(S') = y))$$
$$= 0.$$ \hspace{1cm} (19)

By calculating the integration of the both sides of eq. (18), we have

$$\int k(y)d\mathbb{P}(M(S') = y) = 0.$$

Also, combined with the definition of $k(y)$ (see eq. 17), the right-hand side (RHS) of eq. (16) becomes

$$\text{RHS} = \mathbb{E}_{M(S')}k(M(S')) \log(k(M(S')) + 1).$$ \hspace{1cm} (20)

Since $M$ is $\varepsilon$-differentially private, $k(y)$ is bounded from both sides as follows,

$$e^{-\varepsilon} - 1 \leq k(y) \leq e^\varepsilon - 1.$$ \hspace{1cm} (21)

We now calculate the maximum of eq. (20) subject to eqs. (19) and (21).

First, we argue that the maximum is achieved when $k(M(S')) \in \{e^{-\varepsilon} - 1, e^\varepsilon - 1\}$ with probability 1 (almost surely). When $k(M(S')) \in \{e^{-\varepsilon} - 1, e^\varepsilon - 1\}$, almost surely, the distribution for $k(M(S'))$ is as following,

$$\mathbb{P}^*(k(M(S')) = e^\varepsilon - 1) = \frac{1}{1 + e^\varepsilon},$$
$$\mathbb{P}^*(k(M(S')) = e^{-\varepsilon} - 1) = \frac{e^\varepsilon}{1 + e^\varepsilon}.$$  

We argue that it is the distribution that maximizes $k(M(S'))$.

For the brevity, we denote the probability measure for a given distribution $Q$ as $\mathbb{P}_Q$. Similarly, $\mathbb{P}^*$ corresponds the distribution $Q^*$. We prove that $Q^*$ maximizes eq. (20) in the following two cases: (1) $\mathbb{P}_Q(k(M(S')) \geq 0) \leq \mathbb{P}^*(k(M(S')) = e^\varepsilon - 1)$, and (2) $\mathbb{P}_Q(k(M(S')) \geq 0) > \mathbb{P}^*(k(M(S')) = e^{-\varepsilon} - 1)$.

**Case 1:** $\mathbb{P}_Q(k(M(S')) \geq 0) \leq \mathbb{P}^*(k(M(S')) = e^\varepsilon - 1)$

We have

$$\mathbb{E}_{M(S') \sim Q^*}(k(M(S')) \log(k(M(S')) + 1))$$
$$= \mathbb{P}^*(k(M(S')) = e^\varepsilon - 1) \cdot \varepsilon(e^\varepsilon - 1) - \mathbb{P}^*(k(M(S')) = e^{-\varepsilon} - 1) \cdot \varepsilon(e^{-\varepsilon} - 1)$$
$$= (\mathbb{P}^*(k(M(S')) = e^\varepsilon - 1) - \mathbb{P}_Q(k(M(S')) \geq 0)) \cdot \varepsilon(e^\varepsilon - 1)$$
$$+ \mathbb{P}_Q(k(M(S')) \geq 0) \cdot \varepsilon(e^\varepsilon - 1) - \mathbb{P}^*(k(M(S')) = e^{-\varepsilon} - 1) \cdot \varepsilon(e^{-\varepsilon} - 1)$$
$$\geq \mathbb{P}_Q(k(M(S')) \geq 0) \cdot \varepsilon(1 - e^{-\varepsilon}) - \mathbb{P}^*(k(M(S')) = e^{-\varepsilon} - 1) \cdot \varepsilon(1 - e^{-\varepsilon})$$
$$+ \mathbb{P}_Q(k(M(S')) \geq 0) \cdot \varepsilon(e^\varepsilon - 1) - \mathbb{P}^*(k(M(S')) = e^\varepsilon - 1) \cdot \varepsilon(e^\varepsilon - 1).$$
Note that

\[
\mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}(k(M(S')) = e^\varepsilon - 1) - \mathbb{P}_Q(k(M(S')) \geq 0) \\
+ \mathbb{P}(k(M(S')) = e^{-\varepsilon} - 1).
\]

Therefore, together with the condition eq. (21),

\[
\mathbb{E}_{M(S') \sim Q}(k(M(S')) \log(k(M(S')) + 1) I_{k(M(S')) \leq 0}) \\
\leq (\mathbb{P}(k(M(S')) = e^\varepsilon - 1) - \mathbb{P}_Q(k(M(S')) \geq 0)) \cdot \varepsilon(1 - e^{-\varepsilon}) \\
+ \mathbb{P}(k(M(S')) = e^{-\varepsilon} - 1) \cdot \varepsilon(1 - e^{-\varepsilon}).
\]

(22)

Also,

\[
\mathbb{E}_{M(S') \sim Q}(k(M(S')) \log(k(M(S')) + 1) I_{k(M(S')) > 0}) \leq \mathbb{P}_Q(k(M(S')) \geq 0) \cdot \varepsilon(e^\varepsilon - 1).
\]

(23)

Therefore, combined inequalities eqs. (22) and (23), we have

\[
\mathbb{E}_{M(S') \sim Q}(k(M(S')) \log(k(M(S')) + 1)) \leq \mathbb{E}_{M(S') \sim Q^*}(k(M(S')) \log(k(M(S')) + 1)).
\]

Since the distribution \(Q\) is arbitrary, the distribution \(Q^*\) maximizes the \(k(M(S')) \log(k(M(S')) + 1)\).

**Case 2:** \(\mathbb{P}_Q(k(M(S')) \geq 0) > \mathbb{P}(k(M(S')) = e^\varepsilon - 1)\)

We first prove that if \(\mathbb{P}_Q(1 - e^{-\varepsilon} < k(M(S')) < 0) \neq 0\), there exists a distribution \(Q'\) such that

\[
\mathbb{P}_Q(k(M(S')) \geq 0) = \mathbb{P}_Q(k(M(S')) \geq 0), \\
\mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}_Q(k(M(S')) < 0), \\
\mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}_Q(k(M(S')) = e^{-\varepsilon} - 1), \\
\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)) > \mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)),
\]

while the two conditions (eqs. 19, 21) still hold.

Additionally, we have assumed that

\[
\mathbb{P}_Q(k(M(S')) \geq 0) > \mathbb{P}(k(M(S')) = e^\varepsilon - 1).
\]

Therefore,

\[
\mathbb{P}_Q(k(M(S')) \leq 0) < \mathbb{P}(k(M(S')) = e^{-\varepsilon} - 1).
\]

Also, since the distribution \(Q'\) is arbitrary, let it satisfy

\[
\mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}_Q(k(M(S')) = e^{-\varepsilon} - 1).
\]

Then, in order to meet the condition eq. (19), let

\[
\mathbb{P}_Q(k(M(S')) = e^\varepsilon - 1) > \mathbb{P}_Q(k(M(S')) = e^\varepsilon - 1),
\]

and

\[
\mathbb{P}_Q(0 < k(M(S')) < e^\varepsilon - 1) \leq \mathbb{P}_Q(0 < k(M(S')) < e^\varepsilon - 1),
\]

Since \(x \log(x + 1)\) increases when \(x > 0\) and decreases when \(x < 0\), we have

\[
\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)) > \mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)).
\]

Therefore, we have proved that the argument when \(\mathbb{P}_Q(k(M(S')) < 0) \neq \mathbb{P}_Q(k(M(S')) = e^{-\varepsilon} - 1)\).

We now prove the case that

\[
\mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}_Q(k(M(S')) = e^{-\varepsilon} - 1),
\]

26
where
\[
\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)I_{k(M(S')) < 0}) = \epsilon(1 - e^{-\epsilon})\mathbb{P}_Q(k(M(S')) < 0).
\]

Applying Jensen’s inequality to bound the \(\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)I_{k(M(S')) < 0})\), we have
\[
\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)I_{k(M(S')) < 0}) = \mathbb{P}_Q(M(S') \geq 0)\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1)I_{k(M(S'))} \geq 0) \leq \mathbb{P}_Q(M(S') \geq 0)\mathbb{E}_Q(k(M(S'))I_{k(M(S'))} \geq 0) \cdot (1 - \mathbb{E}_Q(k(M(S'))I_{k(M(S'))} \geq 0) + 1),
\]
where the upper bound \((*)\) uses Jensen’s inequality \((x \log(1 + x)\) is convex with respect to \(x\) when \(x > 0\). The upper bound in \eqref{eq:24} is achieved as long as
\[
\mathbb{P}_Q(k(M(S')) \geq 0) = \mathbb{P}_Q(k(M(S')) = \mathbb{E}_Q(k(M(S'))I_{k(M(S'))} \geq 0)).
\]

Furthermore,
\[
\mathbb{P}_Q(k(M(S')) < 0) = \mathbb{P}_Q(k(M(S')) = e^{-\epsilon} - 1).
\]

Therefore, the distribution \(Q\) is determined by the cumulative density functions \(\mathbb{P}_Q(k(M(S')) < 0)\) and \(\mathbb{P}_Q(k(M(S')) \geq 0)\).

Hence, maximizing \(\mathbb{E}_Q(k(M(S')) \log(k(M(S')) + 1))\) is equivalent to maximizing the following object function,
\[
g(q) = q(1 - e^{-\epsilon}) \log e^\epsilon + (1 - q)\frac{q}{1 - q}(1 - e^{-\epsilon}) \log \left(\frac{q}{1 - q}(1 - e^{-\epsilon}) + 1\right),
\]
subject to
\[
\frac{q}{1 - q} \leq e^\epsilon,
\]
\[
\frac{q}{1 - q} \leq e^\epsilon
\]
where \(g(q)\) is the maximum of \eqref{eq:20} subject to \(\mathbb{P}_Q(k(M(S')) < 0) = q\), and the condition \eqref{eq:25} comes from the \(\mathbb{P}_Q(k(M(S')) \geq 0) > \mathbb{P}_Q(k(M(S')) = e^\epsilon - 1)\) (the assumption of Case 2).

Additionally, \(g(q)\) can be represented as follows,
\[
q(1 - e^{-\epsilon}) \log \left(\frac{q}{1 - q}(e^\epsilon - 1) + \epsilon\right).
\]

Since both \(q\) and \(\frac{q}{1 - q}\) monotonously increase, \(g(q)\) monotonously increases. Therefore, \(\mathbb{E}_Q\) maximize \eqref{eq:20}, which finishes the proof.

\section*{B.2 Proof of Theorem 4}

Based on Lemma 3, we can prove the following composition theorem for \(\epsilon\)-differential privacy as a preparation theorem of the general case.
Proof of Theorem 4. We begin by calculating \( \frac{\mathbb{P}(Y_i(S) = y_i)^T}{\mathbb{P}(Y_i(S') = y_i)^T} \) as follows,

\[
\begin{align*}
&\log \frac{\mathbb{P}(\{Y_i(S) = y_i\}^T_{i=0})}{\mathbb{P}(\{Y_i(S') = y_i\}^T_{i=0})} \\
= &\log \left( \prod_{i=0}^{T} \frac{\mathbb{P}(Y_i(S) = y_i|Y_{i-1}(S) = y_{i-1}, \ldots, Y_0(S) = y_0)}{\mathbb{P}(Y_i(S') = y_i|Y_{i-1}(S') = y_{i-1}, \ldots, Y_0(S') = y_0)} \right) \\
= &\sum_{i=1}^{T} \log \left( \frac{\mathbb{P}(Y_i(S) = y_i|Y_{i-1}(S) = y_{i-1}, \ldots, Y_0(S) = y_0)}{\mathbb{P}(Y_i(S') = y_i|Y_{i-1}(S') = y_{i-1}, \ldots, Y_0(S') = y_0)} \right) \\
= &\sum_{i=1}^{T} \log \left( \frac{\mathbb{P}(M_i(y_{i-1}, S) = y_i|Y_{i-1}(S) = y_{i-1}, \ldots, Y_0(S) = y_0)}{\mathbb{P}(M_i(y_{i-1}, S') = y_i|Y_{i-1}(S') = y_{i-1}, \ldots, Y_0(S') = y_0)} \right)
\end{align*}
\]

(*) where eq. (*) comes from the independence of \( Y_0 \) with respect to \( S \) and eq. (**) is because the independence of \( M_i \) to \( Y_k \) (\( k < i \)) when the \( Y_{i-1} \) is fixed.

By the definition of \( \varepsilon \)-differential privacy, one has for arbitrary \( y_{i-1} \) as the observation of \( Y_{i-1} \),

\[
\begin{align*}
D_{\infty}(M_i(y_{i-1}, S)||M_i(y_{i-1}, S')) < \varepsilon_i, \\
D_{\infty}(M_i(y_{i-1}, S')||M_i(y_{i-1}, S)) < \varepsilon_i.
\end{align*}
\]

Thus, by Lemma 3, we have that

\[
\begin{align*}
&\mathbb{E} \left( \log \left( \frac{\mathbb{P}(M_i(y_{i-1}, S) = Y_i)}{\mathbb{P}(M_i(y_{i-1}, S') = Y_i)} \right) \middle| \ Y_{i-1}(S) = y_{i-1}, \ldots, Y_0(S) = y_0 \right) \\
= &D_{KL}(M_i(y_{i-1}, S)||M_i(y_{i-1}, S')) \\
\leq &\varepsilon_i - \frac{1}{\varepsilon_i + 1}.
\end{align*}
\]

(26)

Combining Azuma Lemma (Lemma 4), eq. (26) derives the following equation

\[
\mathbb{P}(\{Y_i(S) = y_i\}^T_{i=0}) > e^{-\varepsilon_i} \quad \mathbb{P}(\{Y_i(S') = y_i\}^T_{i=0}) > e^{-\varepsilon_i} < \delta',
\]

where \( S \) and \( S' \) are neighbour sample sets.

Therefore, the algorithm \( \mathcal{A} \) is \( \varepsilon' \)-differentially private.

The proof is completed.

\[ \square \]

**B.3 Proof of Theorem 5**

Now, we can prove our composition theorems for \( (\varepsilon, \delta) \)-differential privacy. We first prove a composition algorithm of \( (\varepsilon, \delta) \)-differential privacy whose estimate of \( \varepsilon' \) is somewhat looser than the existing results. Then, we tighten the results and obtain a composition theorem that strictly tighter than the current estimate.
Proof of Theorem 5. It has been proved that the optimal privacy preservation can be achieved by a sequence of independent iterations (see [36], Theorem 3.5). Therefore, without loss of generality, we assume that the iterations in our theorem are independent with each other.

Rewrite \( Y_i(S) \) as \( Y_i^0 \), and \( Y_i(S') \) as \( Y_i^1 \) \((i \geq 1)\). Then, by Lemma 6, there exist random variables \( \tilde{Y}_i^0 \) and \( \tilde{Y}_i^1 \), such that

\[
\Delta \left( Y_i^0 \| \tilde{Y}_i^0 \right) \leq \frac{\delta_i}{1 + e^{\varepsilon_i}}. \tag{27}
\]

\[
\Delta \left( Y_i^1 \| \tilde{Y}_i^1 \right) \leq \frac{\delta_i}{1 + e^{\varepsilon_i}}. \tag{28}
\]

\[
D_\infty \left( \tilde{Y}_i^0 \| \tilde{Y}_i^1 \right) \leq \varepsilon_i. \tag{29}
\]

\[
D_\infty \left( \tilde{Y}_i^1 \| \tilde{Y}_i^0 \right) \leq \varepsilon_i. \tag{30}
\]

Applying Theorem 5 (here, \( \delta = \tilde{\delta} \)), we have that

\[
D_\infty^\tilde{\delta} \left( \{ \tilde{Y}_i^0 \}_{i=0}^T \| \{ \tilde{Y}_i^1 \}_{i=0}^T \right) \leq \varepsilon',
\]

\[
D_\infty^\tilde{\delta} \left( \{ \tilde{Y}_i^1 \}_{i=0}^T \| \{ \tilde{Y}_i^0 \}_{i=0}^T \right) \leq \varepsilon'.
\]

Apparently,

\[
\mathbb{P}(Y_i^0 \in B_i) - \min \left\{ \frac{\delta_i}{1 + e^{\varepsilon_i}}, \mathbb{P}(Y_i^0 \in B_i) \right\} \geq 0.
\]

Therefore, for any sequence of hypothesis sets \( B_0, \ldots, B_T \),

\[
\mathbb{P}(Y_0^0 \in B_0) \left( \mathbb{P}(Y_1^0 \in B_1) - \min \left\{ \frac{\delta}{1 + e^{\varepsilon}}, \mathbb{P}(Y_1^0 \in B_1) \right\} \right)
\]

\[
\cdots \left( \mathbb{P}(Y_T^0 \in B_T) - \min \left\{ \frac{\delta_T}{1 + e^{\varepsilon_T}}, \mathbb{P}(Y_T^0 \in B_T) \right\} \right)
\]

\[
\leq \mathbb{P}(\tilde{Y}_0^0 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^0 \in B_T)
\]

\[
\leq e^{\varepsilon'} \mathbb{P}(\tilde{Y}_1^0 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^1 \in B_T) + \tilde{\delta}.
\] \tag{31}

Furthermore, by eq. (30), we also have that

\[
\mathbb{P}(\tilde{Y}_i^0 \in B_i) \leq \min \left\{ \varepsilon_i, \frac{1}{\mathbb{P}(\tilde{Y}_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_i^1 \in B_i).
\]

Therefore,

\[
\mathbb{P}(\tilde{Y}_0^0 \in B_0) \cdots \mathbb{P}(\tilde{Y}_n^0 \in B_T) \leq \prod_{i=1}^T \min \left\{ \varepsilon_i, \frac{1}{\mathbb{P}(\tilde{Y}_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^1 \in B_T) + \tilde{\delta}.
\]

Then, we prove this theorem in two cases: (1) \( \prod_{i=1}^T \min \left\{ \varepsilon_i, \frac{1}{\mathbb{P}(\tilde{Y}_i^1 \in B_i)} \right\} \leq e^{\varepsilon'} \); and (2)

\[
\prod_{i=1}^T \min \left\{ \varepsilon_i, \frac{1}{\mathbb{P}(\tilde{Y}_i^1 \in B_i)} \right\} > e^{\varepsilon'}.
\]

**Case 1.** \( \prod_{i=1}^T \min \left\{ \varepsilon_i, \frac{1}{\mathbb{P}(\tilde{Y}_i^1 \in B_i)} \right\} \leq e^{\varepsilon'} \).

We have that

\[
\mathbb{P}(\tilde{Y}_0^1 \in B_0) \left( \mathbb{P}(\tilde{Y}_1^1 \in B_1) - \frac{\delta}{1 + e^{\varepsilon_1}} \right) \cdots \left( \mathbb{P}(\tilde{Y}_T^1 \in B_T) - \frac{\delta_T}{1 + e^{\varepsilon_T}} \right)
\]

\[
\leq \mathbb{P}(Y_0^1 \in B_0) \cdots \mathbb{P}(Y_T^1 \in B_T).
\]

\[
\prod_{i=1}^T \min \left\{ \varepsilon_i, \frac{1}{\mathbb{P}(\tilde{Y}_i^1 \in B_i)} \right\} > e^{\varepsilon'}.
\]
By simple calculation, we have that

\[
\prod_{i=1}^{T} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^1 \in B_T) \\
\leq \prod_{i=1}^{T} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^1 \in B_T) \\
+ \prod_{i=1}^{T} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^1 \in B_T) \\
- \prod_{i=1}^{n} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \\
\left( \mathbb{P}(\tilde{Y}_1^1 \in B_1) - \frac{\delta_1}{1 + e^{\varepsilon_1}} \right) \cdots \left( \mathbb{P}(\tilde{Y}_T^1 \in B_T) - \frac{\delta_T}{1 + e^{\varepsilon_T}} \right).
\]

Apparently,

\[
\min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_i) \leq 1,
\]

and when \( A > B \), \( f(x) = Ax - (x - a)B \) increases when \( x \) increases.

Therefore, we have that

\[
\prod_{i=1}^{T} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \cdots \mathbb{P}(\tilde{Y}_T^1 \in B_T) \\
- \prod_{i=1}^{T} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \mathbb{P}(\tilde{Y}_0^1 \in B_0) \\
\left( \mathbb{P}(\tilde{Y}_1^1 \in B_1) - \frac{\delta_1}{1 + e^{\varepsilon_1}} \right) \cdots \left( \mathbb{P}(\tilde{Y}_T^1 \in B_T) - \frac{\delta_T}{1 + e^{\varepsilon_T}} \right) \\
\leq 1 - \prod_{i=1}^{T} \left( 1 - \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \frac{\delta_i}{1 + e^{\varepsilon_i}} \right).
\]

Combining with eq. (31), we have that

\[
\delta' \leq 1 - \prod_{i=1}^{T} \left( 1 - \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} \frac{\delta_i}{1 + e^{\varepsilon_i}} \right) + 1 - \prod_{i=1}^{T} \left( 1 - \frac{\delta_i}{1 + e^{\varepsilon_i}} \right) + \delta.
\]

Case 2: \( \prod_{i=1}^{T} \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\} > e^{\varepsilon'} \):

There exists a sequence of reals \( \{ \alpha_i \}_{i=1}^{T} \) such that

\[
e^{\alpha_i} \leq \min \left\{ e^{\varepsilon_i}, \frac{1}{\mathbb{P}(Y_i^1 \in B_i)} \right\},
\]

\[
\sum_{i=1}^{T} \alpha_i = \varepsilon'.
\]

Therefore, similar to Case 1, we have that

\[
\delta' \leq 1 - \prod_{i=1}^{T} \left( 1 - e^{\alpha_i} \frac{\delta_i}{1 + e^{\varepsilon_i}} \right) + 1 - \prod_{i=1}^{T} \left( 1 - \frac{\delta_i}{1 + e^{\varepsilon_i}} \right).
\]
Overall, we have proven that

$$\delta' \leq 1 - \prod_{i=1}^{T} \left( 1 - e^{\alpha_i} \frac{\delta_i}{1 + e^{\alpha_i}} \right) + 1 - \prod_{i=1}^{T} \left( 1 - \frac{\delta_i}{1 + e^{\alpha_i}} \right),$$

where $\sum_{i=1}^{T} \alpha_i \leq \varepsilon'$ and $\alpha_i \leq \varepsilon_i$.

From Lemma 7, the minimum is realised on the boundary, which is exactly this theorem claims. The proof is completed. \qed

Then, we can prove Theorem 3.

**Proof of Theorem 3.** Applying Theorem 3.5 in [36] and replacing $\varepsilon'$ in the proof of Theorem 5 as

$$\varepsilon' = \min \{ I_1, I_2, I_3 \},$$

where

$$I_1 = \sum_{i=1}^{T} \varepsilon_i,$$

$$I_2 = \sum_{i=1}^{T} \left( \frac{e^{\varepsilon_i} - 1}{e^{\varepsilon_i} + 1} \right) \varepsilon_i + \sum_{i=1}^{T} 2\varepsilon_i^2 \log \left( e + \sqrt{\sum_{i=1}^{T} \varepsilon_i^2 \frac{1}{\delta}} \right),$$

$$I_3 = \sum_{i=1}^{T} \left( \frac{e^{\varepsilon_i} - 1}{e^{\varepsilon_i} + 1} \right) \varepsilon_i + \sum_{i=1}^{T} 2\varepsilon_i^2 \log \left( 1 - \delta \right)$$

The proof is completed. \qed

**B.4 Proof of Corollary 3**

**Proof of Corollary 3.** Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two distributions whose cumulative distribution functions $P_0$ and $P_1$ are respectively defined as following:

$$P_0(x) = \begin{cases} \delta, & x = 0 \\ (1 - \delta)e^{\varepsilon}, & x = 1 \\ \frac{1 - \delta}{1 + e^{\varepsilon}}, & x = 2 \\ 0, & x = 3 \end{cases},$$

and

$$P_1(x) = \begin{cases} 0, & x = 0 \\ (1 - \delta)e^{\varepsilon}, & x = 1 \\ \frac{1 - \delta}{1 + e^{\varepsilon}}, & x = 2 \\ \delta, & x = 3 \end{cases}.$$

By Theorem 3.4 of [36], the largest magnitude of the $(\varepsilon', \delta')$-differential privacy can be calculated from the $\mathcal{P}_0^T$ and $\mathcal{P}_1^T$. 

31
Construct $\tilde{P}_0$ and $\tilde{P}_1$, whose cumulative distribution functions are as follows,

$$\tilde{P}_0(x) = \begin{cases} 
\frac{e^\varepsilon}{1 + e^\varepsilon}, & x = 0 \\
\frac{(1 - \delta)e^\varepsilon}{1 + e^\varepsilon}, & x = 1 \\
\frac{1 - \delta}{1 + e^\varepsilon}, & x = 2 \\
\frac{\delta}{1 + e^\varepsilon}, & x = 3 
\end{cases}$$

and

$$\tilde{P}_1(x) = \begin{cases} 
\frac{\delta}{1 + e^\varepsilon}, & x = 0 \\
\frac{(1 - \delta)e^\varepsilon}{1 + e^\varepsilon}, & x = 1 \\
\frac{1 - \delta}{1 + e^\varepsilon}, & x = 2 \\
e^\varepsilon \delta, & x = 3 \\
\end{cases}$$

One can easily verify that

$$\Delta(P_0 \parallel \tilde{P}_0) \leq \frac{\delta}{1 + e^\varepsilon},$$

$$\Delta(P_1 \parallel \tilde{P}_1) \leq \frac{\delta}{1 + e^\varepsilon},$$

$$D_\infty(\tilde{P}_0 \parallel \tilde{P}_1) \leq \varepsilon,$$

$$D_\infty(\tilde{P}_1 \parallel \tilde{P}_0) \leq \varepsilon.$$
similar analysis of the Proof of Theorem 5, we prove this theorem.

B.5 Tightness of Theorem 3

This section analyses the tightness of Theorem 3. Specifically, we compare it with our Theorem 3.

In the proof of Theorem 3 (see Section B.2), \(\varepsilon_3\) is derived through Azuma Lemma (Lemma 4). Specifically, the \(\delta'\) is derived by

\[
P \left[ S_T \geq \varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1} \right] \leq e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} E \left[ e^{tS_T} \right]
\]

\[
= e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} E \left[ e^{tV_T | Y_1(S), \ldots, Y_{T-1}(S)} \right]
\]

\[
\leq e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} e^{2t^2\varepsilon^2/8}
\]

\[
\leq e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} e^{T^2\varepsilon^2/2},
\]

where \(V_i\) is defined as \(\log \frac{p(Y_i(S))}{p(Y_i(S'))} - E \left[ \log \frac{p(Y_i(S))}{p(Y_i(S'))} | Y_1(S), \ldots, Y_{i-1}(S) \right]\) and \(S_j\) is defined as \(\sum_{i=1}^j V_i\).

Since \(P \left[ S_T \geq \varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1} \right]\) does not depend on \(t\),

\[
P \left[ S_T \geq \varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1} \right] \leq \min_{t>0} e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})^2} = \delta',
\]

By contrary, the approach here directly calculates \(E[e^{tS_T}]\), without the shrinkage in the proof of Theorem 3 (see Section B.2). Specifically,

\[
e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} e^{2t^2\varepsilon^2/8}.
\]

Therefore,

\[
\min_{t>0} e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} e^{2t^2\varepsilon^2/8} \leq \min_{t>0} e^{-t(\varepsilon' - T \frac{\varepsilon^2}{e\varepsilon + 1})} e^{T^2\varepsilon^2/2},
\]

which leads to

\[
e^{-\frac{T^2\varepsilon^2}{8} - \frac{\varepsilon^2}{2}} \left( \frac{T\varepsilon + \varepsilon'}{T\varepsilon - \varepsilon'} \right)^{T \left( \frac{T\varepsilon + \varepsilon'}{T\varepsilon - \varepsilon'} \right)} \leq \delta'.
\]

It ensures that this estimate further tightens \(\delta'\) than Section B.2 (which is also the \(\tilde{\delta}\) in Theorem 3) if the \(\varepsilon'\) is the same.

C Applications

This appendix collects the proofs for the applications in SGLD and federated learning.

C.1 Proof of Theorem 6

Proof of Theorem 6. We first calculate the differential privacy of each step. Assume mini-batch \(B\) has been selected and define \(\nabla R_\theta(z)\) as following:

\[
\nabla R_\theta(z) = \nabla r(\theta) + \sum_{z \in B} \nabla l(z|\theta).
\]
For any two neighboring sample sets $S$ and $S'$ and fixed $\theta_{i-1}$, we have

$$\max_{\theta_i} \frac{p(\theta_i^S = \theta_i | \theta_{i-1}^S = \theta_{i-1})}{p(\theta_i^{S'} = \theta_i | \theta_{i-1}^{S'} = \theta_{i-1})} = \max_{\theta_i} \frac{p(\eta_i(-\frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1}) + \mathcal{N}(0, \sigma^2 I)) = \theta_i - \theta_{i-1})}{p(\eta_i(-\frac{1}{\tau} \nabla \hat{R}_{S'}^\tau(\theta_{i-1}) + \mathcal{N}(0, \sigma^2 I)) = \theta_i - \theta_{i-1})}$$

Define

$$D(\theta') = \log \frac{p(-\frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1}) + \mathcal{N}(0, \sigma^2 I) = \theta')}{p(-\frac{1}{\tau} \nabla \hat{R}_{S'}^\tau(\theta_{i-1}) + \mathcal{N}(0, \sigma^2 I) = \theta')},$$

where $\theta' = \frac{1}{\eta_i} \theta_i$ obeys $-\frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1}) + \mathcal{N}(0, \sigma^2 I)$.

Let $\theta'' = \theta' + \frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1})$ and rewrite $D(\theta')$ as:

$$D(\theta') = \log e^{-\frac{||\theta' + \frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1})||^2}{2\sigma^2}}$$

$$= -\frac{||\theta'' + \frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1})||^2}{2\sigma^2} + \frac{||\theta' + \frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1})||^2}{2\sigma^2}$$

$$= \frac{2\theta''T \frac{1}{\tau} (\nabla \hat{R}_S^\tau(\theta_{i-1}) - \nabla \hat{R}_S^\tau(\theta_{i-1})) + \frac{1}{\tau^2} (\nabla \hat{R}_S^\tau(\theta_{i-1}) - \nabla \hat{R}_S^\tau(\theta_{i-1}))^2}{2\sigma^2}.$$

Define $\nabla \hat{R}_S^\tau(\theta_{i-1}) - \nabla \hat{R}_S^\tau(\theta_{i-1})$ as $v$. By definition of $L$, we have that

$$||v|| < 2L.$$

Therefore, since $\theta''T v \sim \mathcal{N}(0, ||v||^2 \sigma^2)$, by Chernoff Bound technique,

$$\mathbb{P}(\theta''T v \geq 2\sqrt{2L} \sigma \sqrt{\log \frac{1}{\delta}}) \leq \mathbb{P}(\theta''T v \geq \sqrt{2} ||v|| \sigma \sqrt{\log \frac{1}{\delta}})$$

$$\leq \min_t e^{-\sqrt{2t} ||v|| \sigma \sqrt{\log \frac{1}{\delta} \mathbb{E}(e^{t \theta''T v})}}$$

$$= \delta.$$

Therefore, with probability at least $1 - \delta$ with respect to $\theta'$, we have that

$$D(\theta') \leq \frac{2\sqrt{2L} \sigma \frac{1}{\sqrt{2}} \sqrt{\log \frac{1}{\delta} + \frac{1}{\tau^2} L^2}}{2\sigma^2}.$$

Define $\epsilon = \frac{2\sqrt{2L} \sigma \frac{1}{\sqrt{2}} \sqrt{\log \frac{1}{\delta} + \frac{1}{\tau^2} L^2} - \epsilon}{2\sigma^2}$. Applying Lemma 4.4 in [8], we have that the iteration $-\frac{1}{\tau} \nabla \hat{R}_S^\tau(\theta_{i-1}) + \mathcal{N}(0, \sigma^2 I)$ is $(2\frac{\epsilon}{\tau^2} \sigma, \frac{\epsilon}{\tau^2} N \delta)$-differentially private. Applying Theorem 3 and

$$\epsilon' = \sqrt{8 \log \left(\frac{1}{\delta}\right) \left(\frac{\tau^2}{N^2} T \epsilon^2\right)} + 2T \frac{\tau}{N} \frac{\epsilon^2 \tau \epsilon}{e^{2 \frac{\tau \epsilon}{N}} - 1},$$

we can prove the differential privacy.

Letting $B$ sampled randomly and applying Theorem 1, we can prove the generalization bound. The proof is completed.
C.2 Proof of Theorem 7

We only need to prove differential privacy part of Theorem 7, and the rest of the proof is similar with the one of Theorem 6.

Proof of Theorem 7. The proof bears resemblance to the proof of Theorem 6. One only has to notice that each update is still a Gauss mechanism, while

$$\left\| \frac{h_i^t}{\max(1, \frac{\|h_i\|_2}{L})} \right\| \leq L.$$  

Then, in this situation, $D(\theta')$ is as follows:

$$D(\theta') = \log \left( p \left( \frac{1}{\tau} \sum_{c_k \in \mathcal{B}} \frac{h_i^t}{\max(1, \frac{\|h_i\|_2}{L})} \right) + \mathcal{N}(0, L^2 \sigma^2 I) = \theta' \right).$$

All other reasoning is the same as the previous proof. By Theorem 3 and

$$\varepsilon' = \sqrt{8 \log \left( \frac{1}{\delta} \right) \left( \frac{\tau^2}{N^2 T \varepsilon^2} \right) + 2T \frac{\tau}{N} \frac{e^{2 \frac{\tau}{N} \varepsilon}}{e^{2 \frac{\tau}{N} \varepsilon} + 1}},$$

we can calculate the differential privacy of federated learning.

The proof is completed. □