Generalization Bounds for Stochastic Gradient Langevin Dynamics: A Unified View via Information Leakage Analysis

Bingzhe Wu 1 Zhicong Liang 2 Yatao Bian 1 ChaocHao Chen 3 Junzhou Huang 1 Yuan Yao 2
1 Tencent AI Lab 2 Hong Kong University of Science and Technology 3 Zhejiang University

Abstract

Recently, generalization bounds of the non-convex empirical risk minimization paradigm using Stochastic Gradient Langevin Dynamics (SGLD) have been extensively studied. Several theoretical frameworks have been presented to study this problem from different perspectives, such as information theory and stability. In this paper, we present a unified view from privacy leakage analysis to investigate the generalization bounds of SGLD, along with a theoretical framework for re-deriving previous results in a succinct manner. Aside from theoretical findings, we conduct various numerical studies to empirically assess the information leakage issue of SGLD. Additionally, our theoretical and empirical results provide explanations for prior works that study the membership privacy of SGLD.

1 Introduction

One of the fundamental problems in deep learning is to characterize generalization bounds of non-convex stochastic optimization in the empirical risk minimization paradigm (ERM) setting (Mou et al. 2018; Kawaguchi, Kaelbling, and Bengio 2017; Nagarajan and Kolter 2019; Arora et al. 2018). Some traditional approaches study the generalization bounds by measuring the complexity of the hypothesis space in a data-dependent manner (e.g., the VC dimension (Vapnik 1998) and the Rademacher complexity (Bartlett and Mendelson 2002)). Nevertheless, directly adopting these complexity measures often fail to characterize the generalization ability of various learning algorithms for training deep neural networks (DNNs), in which the algorithm explores the hypothesis space in an algorithm-dependent manner (Zhang et al. 2017; Mou et al. 2018). Prior works have shown various empirical evidence from the observation that the hyper-parameters of stochastic gradient methods could significantly affect the generalization ability of the learned DNN models (Mou et al. 2018; Hoffer, Hubara, and Soudry 2017). For example, large batch size training often fails to generalize to the test dataset well while the small batch size could introduce more stochastic noise as an implicit regularizer, which further improves the generalization ability of the learned DNN models (Hoffer, Hubara, and Soudry 2017).

The aforementioned issues are driving the research on algorithm-dependent generalization bounds in the deep learning context. Specifically, the generalization bound of Stochastic Gradient Langevin Dynamics (SGLD) in the non-convex setting has drawn increasing attention from the research community. Previous works have demonstrated various generalization bounds of SGLD from different theoretical perspectives (Mou et al. 2018; Bu, Zou, and Veeravalli 2019; Negrea et al. 2019). Stability-based theory and information theory are the two most important theoretical tools used in these works. A seminal work that leverages the stability-based theory to obtain the generalization bound of SGLD is conducted by Mou et al. (Mou et al. 2018). The key idea of this work is to prove the uniform stability of SGLD concerning the squared Hellinger distance and further derive the bound via the connection between uniform stability and the expected generalization error (Shalev-Shwartz et al. 2010; Mou et al. 2018). A follow-up work improves the generalization bound by introducing the Bayes stability notion (Negrea et al. 2019). Another line of research works study this problem based on information theory. The core idea is to bound the generalization error via some information measures. For example, an early work (Xu and Raginsky 2017) proposes to bound the expected generalization error in terms of the mutual information between the training dataset and the output hypothesis of the learning algorithm. Bu et al. (Bu, Zou, and Veeravalli 2019) introduce the individual sample mutual information and link it to the expected generalization error.

Despite these theoretical results, the connections amongst these lines of research and how to build a unified understanding of them remain largely unexplored. In this paper, we present a unified view to study the generalization ability of SGLD via analyzing the privacy leakage of SGLD. Our work is inspired by a number of recent works (Wu et al. 2019c; Wu et al. 2019b; Wang, Lei, and Fienberg 2016), which show that there are both theoretical and empirical connections between the privacy leakage and the generalization error. Intuitively, “improving the generalization ability” and “reducing the individuals’ privacy leakage” share the same goal of encour-
aging a model to learn the population’s features instead of memorizing the features of each individual (Dwork 2011 [Wu et al. 2019c]). Theoretically, previous results have shown that the commonly used privacy notion of differential privacy (Dwork 2011) may provide the property of uniform stability, which can be further related to the bound of the expected generalization error. Unfortunately, the notion of differential privacy is too strict to be directly adopted for obtaining reasonable generalization bounds for SGLD, i.e., the higher privacy budget of differential privacy would lead to vacuous bound while lower privacy budget will cause a low model performance.

In this paper, we propose using a relaxed privacy notion based on Rényi divergence to analyze the privacy leakage of SGLD. Specifically, under some mild assumptions (e.g., the boundness of the loss function), we first compute the bound of the Rényi divergence between weight parameters learned using two adjacent datasets at each iteration, i.e., the “leave-one-out” Rényi divergence. We could then use the chain rule of the Rényi divergence to compose these bounds of different iterations for deriving the total bound of the final weights learned using SGLD. Based on such a bound on the Rényi divergence, we have a simple and unified approach to obtain various theoretical bounds in previous studies. For example, the foundation of the work in (Mou et al. 2018) is uniform stability with respect to the Hellinger distance, which implies the boundedness of the Hellinger distance between the hypothesis (i.e., weight parameters) learned using two adjacent datasets. Under our framework, the boundedness of the Hellinger distance can be obtained by the bound of the Rényi divergence, thus can be further used to obtain the generalization bound of SGLD. Moreover, our framework could also give rise to previous works (Xu and Raginsky 2017 [Bu, Zou, and Veeravalli 2019]) with an information-theoretic view and has the potential to improve previous generalization bounds by introducing tighter privacy notions. See more details and results in Section 2.

Besides the theoretical contributions, we also conduct experiments on several datasets where data privacy is desired to empirically evaluate the information leakage of SGLD against membership attacks. Specifically, we design different attacks (e.g., membership attacks) to evaluate the information leakages of models learned using SGLD and SGD. We empirically observe that SGLD could not only improve the generalization ability of a model but also reducing its membership information leakage for the training dataset.

2 Theoretical Framework

2.1 Preliminaries

We denote \( \mathcal{D} \) as the unknown population distribution on the sample space \( \mathcal{Z} \) and denote \( \mathcal{W} \) as the parameter space of the hypothesis (e.g., a neural network). Considering a loss function \( l : \mathcal{Z} \times \mathcal{W} \to \mathbb{R} \), the goal of a learning algorithm is to find the parameter \( w \) which minimizes the population risk \( L_{\mathcal{D}}(w) = \mathbb{E}_{z \sim \mathcal{D}}(l(z; w)) \). Empirically, the population distribution is always intractable, thus we turn to the empirical risk minimization (ERM). In the ERM paradigm, a training dataset \( \mathcal{S} = \{z_i\}_{i=1}^n \) is given and each \( z_i \) is sampled from \( \mathcal{D} \) independently, i.e., \( z_i \overset{i.i.d.}{\sim} \mathcal{D} \). In particular, \( z_i = (x_i, y_i) \) is a sample-label pair in the setting of supervised learning. The aim of a learning algorithm in the ERM setting is to find the parameter which minimizes the empirical risk

\[
L_S(w) = \frac{1}{n} \sum_{i=1}^n l(z_i; w).
\]

In the above ERM paradigm, given a learning algorithm \( \mathcal{A} \), we are interested in its expected generalization error (gap), which is the expected difference between the empirical and population risk,

\[
\text{gen}(\mathcal{A}, \mathcal{D}) = \mathbb{E}_{w \sim \mathcal{A}, \mathcal{S} \sim \mathcal{D}}[L_{\mathcal{D}}(w) - L_S(w)],
\]

where the expectation is taken over the randomness from both the sampling process of the training dataset and the learning algorithm (e.g., stochastic error or the injected noises in SGLD).

2.2 Privacy Leakage Analysis

In this paper, we study the bound of the expected generalization error from the view of privacy analysis. The learning algorithm \( \mathcal{A} \) studied in this paper is SGLD, which can be seen as a noisy version of SGD. We denote \( w_t \) as the output parameter at the \( t \)-th iteration. Following the prior work (Negrea et al. 2019), we employ the SGLD updating rule as follows in this paper:

\[
w_{t+1} = w_t - \alpha_t \partial_w L(B_t, w_t) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, 2\alpha_t \mathbf{I})
\]

\[
L(B_t, w_t) = \frac{1}{|B_t|} \sum_{z_i \in B_t} l(z_i, w_t),
\]

where \( \alpha_t \) is the step size. \( \mathcal{N}(0, 2\alpha_t \mathbf{I}) \) denotes the Gaussian distribution. \( L(B_t, w_t) \) is the loss function computed on a mini-batch \( B_t \) that is randomly selected from the whole dataset \( \mathcal{S} \). Due to the injected noise and the randomness from the mini-batch, the output parameter of SGLD at each iteration can be seen as a random variable with the density function \( p(w_{t+1} | w_t, \mathcal{S}) \). The goal of this part is to measure the information leakage of the output of the learning algorithm \( \mathcal{A} \), i.e., \( p(w_T | \mathcal{S}) \), where \( T \) denotes the number of iterations.

In this paper, motivated by recent advances in Rényi differential privacy (Mironov 2017), we propose to quantitatively evaluate the information leakage by measuring the difference between the output distributions computed on two adjacent datasets. To do this, we first introduce the concept of Rényi divergence which is used for measuring the difference between two probability distributions:

**Definition 1 (Rényi divergence).** For two probability distributions \( p \) and \( q \) (probability density functions), the Rényi divergence of order \( \lambda \) is defined as:

\[
D_\lambda(p||q) \triangleq \frac{1}{\lambda - 1} \log \mathbb{E}_{x \sim q} \left( \frac{p(x)}{q(x)} \right)^\lambda.
\]

We then introduce the concept of adjacent datasets: two datasets \( \mathcal{S} \) and \( \mathcal{S}' \) are adjacent when one can be obtained by removing a single element from the other. Without loss of generality, we assume \( \mathcal{S}' = \mathcal{S} \setminus \{z_a\} \). We further denote

\[
\mathcal{B}(\mathcal{S}, \mathcal{S}') = \mathcal{S} \setminus \mathcal{S}' = \{z_a\}.
\]
where $w_t$ and $w'_t$ as the output of SGLD in the $t$-th iteration, which are computed using $S$ and $S'$, respectively. We are actually interested in the quantity of $\lambda \lambda (p(w_T|S)||p(w'_T|S'))$. The bound of this quantity can also be seen as the privacy cost in the notion of Rényi differential privacy (Mironov 2017).

At a high level, the bound of $\lambda \lambda (p(w_T|S)||p(w'_T|S'))$ can be obtained following two steps: (1) deriving the Rényi divergence between $p(w_t|w_{<t})$ and $p(w'_t|w_{<t})$ at the $t$-th iteration; (2) composing these divergences by the composition theory for the adaptive mechanisms (i.e., learning algorithms). In the case of SGLD, we note that $w_t$ only depends on $w_{t-1}$, thus the conditional distribution $p(w_t|w_{<t})$ can be reduced to $p(w_t|w_{t-1})$.

For step (1), following prior works on analyzing the privacy leakage of DP-SGD (Abadi et al. 2016), we treat $p(w_t|S)$ and $p(w'_t|S')$ as Mixture-Gaussian distribution in the setting of SGLD, then we can explicitly compute the divergence and obtain the following bound:

**Lemma 1 (Leave-one-out Stability of Rényi Divergence)**

Suppose $p(w_t|w_{<t})$ and $p(w'_t|w_{<t})$ are the probability distributions computed using SGLD in Eq. 2 given $w_{t-1}$ and $w'_{t-1}$, and Poisson subsampling with sampling ratio $\tau$ is applied, then the Rényi divergence between these two distributions follows (denoting $\beta_t = \frac{1}{|B_t|}$):

$$D_\lambda (p(w_t|w_{t-1})||p(w'_t|w_{t-1})) \leq \frac{\lambda \lambda L^2}{n^2}$$

given that $\sigma^2 \geq 0.53$ and $\lambda \leq 1$ \text{where} $\sigma^2 := \frac{2}{\alpha_t} \sigma^2 \log \left( \frac{1}{\lambda \lambda (1 + \sigma^2)} \right)$, where $\sigma^2 := \frac{2}{\alpha_t} \sigma^2 L^2$, where $L$ is the bound of the gradient norm $\|w_t\|_{\ell(z, w)}$.

**Proof:** Suppose that we have:

$$w'_{t+1} \sim \sum_{B'_t} p(B'_t) \cdot N \left( w'_t - \alpha_t \partial w L(B'_t, w_{t}), 2\alpha_t I \right)$$

where $B'_t$ is the set of selected samples from $S'$ with sampling ratio $\tau$.

Since $S = S' \cup \{z_n\}$, we have

$$w_{t+1} \sim \sum_{B'_t} p(B'_t) \left( (1 - \tau) N \left( w_t - \alpha_t \partial w L(B'_t, w_t), 2\alpha_t I \right) + \tau N \left( w_t - \alpha_t \partial w L(B'_t \cup \{z_n\}, w_t), 2\alpha_t I \right) \right)$$

For all order $\lambda$, Rényi divergence is quasi-convex.

\(^1\) $w_{<t}$ refers to the sequence $(w_0, w_1, \cdots, w_{t-1})$.
where \( \rho(j) = j/2\sigma^2 \) (Zhu and Wang 2019). By Lemma 3 of (Liang et al. 2020), we know that
\[
D_\lambda \left( (1 - \tau)N'(0, \sigma^2) + \tau N(1, \sigma^2) \right) \left\| N'(0, \sigma^2) \right\| \leq 2\lambda\tau^2/\sigma^2 \leq \frac{\lambda\alpha_t L^2}{n^2},
\]
given \( \sigma^2 \geq 0.53 \) and \( \lambda - 1 \leq \frac{2}{3}\lambda^2 \log \left( \frac{1}{\lambda^2(1 + \sigma^2)} \right) \), where \( \sigma^2 := \frac{2}{\alpha_t^2} L^2 \).

For step (2), we notice that the sequence \( \{w_t\}_{t=1}^T \) can be seen as the outputs of a series of adaptive mechanisms (see the iterative rule in Equation (2)). Here “adaptive” means the input of the current mechanism is the output of the last mechanism. Thus we can further compose the bounds of interest via composition theories. This way of composing privacy costs has been widely applied in prior works, such as moment accountants for differential privacy (Abadi et al. 2016).

In this paper, we borrow the composition theory from the prior work (Proposition 1 in (Mironov 2017)):

**Lemma 2 (Composability)** Suppose that a randomized algorithm \( A \) consists of a series of adaptive mechanisms \( A_1, \ldots, A_T \) where \( A_i : \prod_{j=1}^{i-1} w_j \times S \rightarrow w_t \). The Rényi divergence of the final outputs (on two adjacent datasets) can be bounded by:
\[
D_\lambda(A(S)||A(S')) \leq \sum_{t=1}^{T} D_\lambda(A_t(S)||A_t(S')).
\]

In our setting, the algorithm is set to be SGLD and the sub-mechanisms can compute the output based on the updating rule at each iteration. Combining Lemma 1 and 2, we can obtain the final bound as the following theorem:

**Theorem 1** Suppose \( w_T \) and \( w_T' \) are the final outputs of SGLD on \( S \) and \( S' \), respectively. The training dataset size is set to \( n \). If the gradient norm is bounded by \( L \), then the Rényi divergence between \( p(w_T|S) \) and \( p(w_T'|S') \) is bounded as:
\[
D_\lambda(p(w_T|S)||p(w_T'|S')) \leq \frac{\lambda L^2}{n^2} \sum_{t=1}^{T} \alpha_t.
\]

Theorem 1 is the main building block of our framework and we can further use it for deriving and interpreting previous results, which will be described in detail subsequently. From the view of privacy leakage, Theorem 1 demonstrates that under some mild conditions, the model learned using SGLD incurs a controllable privacy leakage, i.e., SGLD satisfies \((\lambda, \epsilon_n)\)-Rényi differential privacy \((\epsilon_n \text{ is the right term of Equation } 5)\). And the privacy cost \( \epsilon_n \) is dominated by the size of training dataset, i.e., \( \epsilon_n = O\left( \frac{1}{n^2} \right) \).

### 2.3 Stability-based Theory

A promising way to study the generalization error is to derive the algorithm-dependent generalization bounds via various stability notions (Shalev-Shwartz et al. 2010; Mou et al. 2018; Wang, Lei, and Fienberg 2016). The core idea is that the generalization error can be bounded in terms of the stability of the learning algorithm. A commonly-used notion is uniform stability, which can be connected to the expected generalization error of a randomized learning algorithm (Shalev-Shwartz et al. 2010). Mou et al. (Mou et al. 2018) propose to prove the uniform stability via bounding the squared Hellinger distance between \( p(w_T|S) \) and \( p(w_T'|S') \). Here, the squared Hellinger distance between two density functions \( p \) and \( q \) are defined as:
\[
D_H(p||q) \triangleq \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 dw.
\]

We note that the boundness of the Hellinger distance can be obtained via the bound of the Rényi divergence in terms of \( D_H(p||q) \leq D_\lambda(||p||, \lambda) \leq 1/2 \). Based on this observation, we can formally characterize the generalization error as the following theorem:

**Theorem 2** Consider the learning algorithm \( A \) to be SGLD with \( T \) iterations and the batch size is set to be \( b \). And the sampling ratio \( \tau \) is set to be \( b/n \). Suppose that the loss function \( l(z, w) \) is uniformly bounded by \( C \), and the gradient norm is bounded by \( L \). The expected generalization error (Equation 1) can be bounded as:
\[
\text{gen}(A, D) \leq \frac{\sqrt{\tau} L C}{n} \left( \sum_{t=1}^{T} \alpha_t \right)^{1/2}.
\]

**Proof Sketch:** Denoting \( \epsilon_n \) to be the right term in Equation (5). Then we obtain:
\[
\text{gen}(A, D) \leq \sup_{S, S'} 2C \sqrt{D_H(p(w_T|S)||p(w_T'|S'))} \leq \sup_{S, S'} 2C \sqrt{D_{1/2}(p(w_T|S)||p(w_T'|S'))} \leq 2C \epsilon_n.
\]

The proof details can be found in Appendix. This generalization bound has a similar magnitude with the bound derived in the prior work (Mou et al. 2018). Different from Mou et al. (2018), we take a novel view of privacy leakage analysis, which can largely simplify the derivation process and can be extended to a wide range of learning algorithms that satisfies Rényi differential privacy, such as DP-SGD proposed in the prior work (Abadi et al. 2016).

### 2.4 Information-based Theory

The above discussion explores the connection between the information leakage and the stability and further obtains a specific bound for the expected generalization error. We note that a more explicit way to quantify the information leakage is to measure the mutual information between the learned weight and the training dataset. On the other hand, we note that the main research line to study the generalization of a learning algorithm is to build the connection between mutual information and the expected generalization error. Therefore, we can use mutual information to bridge the gap between privacy leakage and the generalization bound of SGLD. To
this end, we introduce an important theoretical block used in this paper as the following lemma:

**Lemma 3 (Proposition 1 in Bu, Zou, and Veeravalli 2019)** Suppose the loss function \( I(z, w) \) is \( \sigma \)-subgaussian under the population distribution \( D \) for all \( w \in W \). \( A \) is the learning algorithm. Then we have:

\[
|\text{gen}(A, D)| \leq \frac{1}{n} \sum_{i=1}^{n} 2\sigma^2 I(A(S); z_i).
\]

The above lemma has established the formal connection between the mutual information and the generalization error and has been widely used in previous works that study the generalization of SGLD. In this paper, we further build the connection between the mutual information and the information leakage, which is reflected by the Rényi divergence in our framework. Motivated by recent works (Poole et al. 2019), we propose using variational inference to derive the upper bound of \( I(p(w_T | S); z_i) \). The derived bound can be further related to the Rényi divergence between the posterior distribution over two adjacent datasets, i.e., \( D_\lambda(p(w_T | S) | p(w_T' | S')) \). This connection is formalized as:

**Lemma 4** Suppose \( S \) and \( S' \) are adjacent and \( A \) is the learning algorithm, then:

\[
I(A(S); z_i) \leq \mathbb{E}_z[D_\lambda(p(w_T | S) | p(w_T' | S'))].
\]

The proof follows two steps: (1) deriving the variational bound of the mutual information using variational inference; (2) deriving the upper bound of the variational bound explicitly. The proof details are provided in Appendix. Combining Theorem 1, Lemma 3, and Lemma 4, we can build the theoretical connection between the expected generalization error and the information leakage measured by the Rényi divergence as:

**Theorem 3** Consider the learning algorithm \( A \) to be SGLD with \( T \) iterations and the batch size is set to \( b \). And the training dataset size is set to \( n \). Suppose that the loss function \( l(z, w) \) is \( \sigma \)-subgaussian under the population distribution \( D \) for all \( w \in W \). And the gradient norm is upper bounded by \( L \). The expected generalization error (Equation (1)) can be bounded as:

\[
|\text{gen}(A, D)| \leq \sqrt{\frac{2\sigma L}{n}} \left( \sum_{t=1}^{T} \alpha_t \right)^{1/2}.
\]

The derivation of the theorem pave a way to derive the generalization bound by the connection between the mutual information and the information leakage (see details in Appendix).

### 2.5 Discussion on Our Results

We provide a unified and succinct way to analyze the generalization ability of SGLD through information leakage analysis. The information leakage routine utilized in this paper unifies two seemingly unrelated ways to obtain the generalization bound: the stability-based route and the information theoretic route. Meanwhile, our approach largely simplifies the analysis: For example, we have obtained a generalization bound of \( O(\frac{1}{\sqrt{n}}) \) similar to that of \( \text{Mou et al. 2018} \), but with a simple proof thanks to the theoretical tools from the data privacy community (Mironov 2017). However, \( \text{Mou et al. 2018} \) need to analyze the bound based on stochastic differential equations (SDE) and transform the continuous setting into discrete one.

The generic privacy notion (Rényi differential privacy) adopted gives rise to refined bounds, and existing bounds could be further improved by using other generic notions, such as the Bayesian differential privacy. More importantly, base on the technique used by \( \text{Mou et al. 2018} \), it is challenging to further improve the generalization bound. However, this can be easily done under our framework by introducing more advanced stability notions or privacy notions. For example, we can replace the stability notion with Bayes stability used in \( \text{Li, Luo, and Qiao 2020} \) to obtain a more tight bound.

### 2.6 Comparisons with Previous Works

There are two most related direction to our works. One is to study generalization ability of SGLD. Another is to study the connection between generalization and differential privacy in general case. In this part, we highlight the differences between our work and these most related works.

In contrast to research focusing on the generalization bounds of SGLD, we aim to unify this line of works through the information leakage analysis of SGLD. This unified view can help to simplify the derivation of generalization bounds of SGLD and provide new insights to improve these bounds.

There are another research line focusing on building the connection between generalization bounds and differential privacy in the general case instead of deriving concrete generalization bounds (Dwork et al. 2015, Bassily et al. 2016). For example, there is a series of work on adaptive data analysis (Dwork et al. 2015, Hardt and Ullman 2014, Russo and Zou 2016, Bassily et al. 2021, Jung et al. 2019), which study false discovery and generalization error led by different settings of adaptivity. While both of the works rely on techniques from differential privacy, in a specific case of SGLD, we need to carefully derive its information leakage in terms of Rényi differential privacy under some mild conditions, i.e., Theorem 1 and its derivation. Once we obtain its privacy leakage, we can employ previous works to build concrete generalization bound.

### 3 Related Work

In this part, we briefly review some related works on the theoretical properties of SGLD.

**Convergence Property.** In early times, SGLD is proposed to enable Bayesian learning on large scale datasets and can be seen as an SG-MCMC method, which can approximate
the process of sampling from the intractable posterior distribution (e.g., $p(w|S)$ in our paper). One of the important properties of SGLD is that the parameters learned using SGLD will approach samples from the parameter posterior distribution. Welling et al. (Welling and Teh 2011) have shown some intuitive observations of this property. Subsequent works have theoretically proved this proposition and have presented various convergence analysis, i.e., how fast the learned parameter can approach the posterior distribution. The key idea of these works is to treat SGLD as the discretization of some specific stochastic differential equations (SDE, e.g., Fokker-Planck equation) and turn to study the discretization error (Chen, Ding, and Carin 2015; Sato and Nakagawa 2014).

Generalization Bounds of SGLD. More recently, the generalization bound of SGLD has emerged as an active topic. Most of the previous works study this problem from two theoretical aspects, the stability-based theory and information-theoretic quantities. A representative work based on stability-based theory is conducted by Mou et al. (Mou et al. 2018), in which they have derived that SGLD satisfies $\epsilon_n$ uniform stability with respect to the Hellinger loss. Then the expected generalization error is bounded by $\epsilon_n$ according to the stability-based theory (Shalev-Shwartz et al. 2010).

Another research line studies the generalizability of SGLD from an information-theoretic view. The core idea is to bound the generalization error via some information measures. For example, an early work (Xu and Raginsky 2017) proposes to bound the expected generalization error in terms of the mutual information between the training dataset and the output hypothesis of the learning algorithm. Bu et al. (Bu, Zou, and Veeravalli 2019) introduce the individual sample mutual information and touch it to the expected generalization error. Besides these theoretical works, there is also empirical evidence showing that SGLD can improve the generalization ability of the model. For example, Gan et al. (Gan et al. 2017) propose using SGLD to mitigate the overfitting of RNN models for language modeling tasks.

Privacy Issues in SGLD. There are also some related works that study the privacy issues in SGLD. Wang et al. (Wang, Fienberg, and Smola 2015) present a variant of SGLD that satisfies differential privacy. Wu et al. (Wu et al. 2019a) further prove that SGLD can naturally prevent the well-known membership attack which is also conducted in our experiments. These works also point out some implicit connections between privacy leakage and generalizations. In our paper, we further formally build the connection between previous theoretical results via the privacy leakage analysis.

4 Empirical Results

In this section, we perform extensive experiments to empirically evaluate the information leakage of SGLD against membership attacks. In what follows, we first describe the detailed experimental settings, including descriptions of the datasets, models, and the attack method. Then we demonstrate and analyze some numerical results. At last, we provide some discussions for further insights.

4.1 Settings

Datasets. The goal of our experiment is to empirically verify the connection between the generalization error and the privacy leakage of SGLD. Therefore we choose datasets from scenarios where data privacy is important, such as financial and census data analysis. Specifically, we select three datasets: (1) The German Credit dataset for credit modeling. (2) The UCI-adult dataset for income prediction (Kohavi 1996). (3) The IDC dataset for pathological image classification.

The German Credit dataset consists of 1,000 applications for credit cards. Each application is labeled with good or bad credit. We consider the classification task to identify “good credit” applications. We randomly split the whole dataset into training (400 applications), hold-out/validation (300 applications), and test (300 applications) sets. Here, the hold-out dataset can be used for either validation or building the “shadow” attack model (Shokri et al. 2017).

We consider a second Adult Census Income dataset (Kohavi 1996) with 48,842 instances from UC Irvine repository to examine the property attack issues. The dataset has 14 features such as country, age, work-class, education, etc. The goal for this binary classification task is to predict whether income exceeds $50K/yr based on the census information. We split the whole dataset into training, validation, and test sets, each with #22,792, #9,769, #16,281 instances respectively. The validation set is also used as the shadow dataset for building the attack model.

We also use a medical image dataset to observe the performance of convolutional neural networks (CNNs). Here, we consider the IDC dataset of pathological images for invasive ductal carcinoma (IDC) classification. This dataset contains 277,524 patches of 50 × 50 pixels (198,738 IDC-negative and 78,786 IDC-positive). We split the whole dataset into training, validation (hold-out), and test sets. To be specific, the training dataset consists of 10,788 positive patches and 29,164 negative patches. The test dataset consists of 11,595 positive patches and 31,825 negative patches. The remain patches are used as the hold-out dataset.

Model Setup. For the German Credit dataset, we train a three-layer fully-connected neural network for the classification task. For comparison, we train the model using both SGLD and SGD variants. All these training strategies share the following hyper-parameters: the mini-batch is set to be 32 and the epoch number is set to be 30. The learning rate decreases by half every 5 epochs. The initial learning rate is set to be $1 \times 10^{-3}$.

For the UCI-adult dataset, we train a fully-connected neural network with a structure of [32, 16]. The mini-batch size and the number of epochs are set as 64 and 100 respectively. SGD is adopted with initial learning as 0.1 and halves every 5 epochs.
Table 1: Overall results on three benchmarks. Attack denotes anti-attack ability measured by the accuracy of the membership predictor (lower is better). Model denotes the model performance measured by the accuracy of the test dataset (higher is better). Gap is the difference between training and test accuracy.

| Method   | German Credit Attack | German Credit Model | Gap | UCI-adult Attack | UCI-adult Model | Gap | IDC Attack | IDC Model | Gap |
|----------|----------------------|---------------------|-----|------------------|-----------------|-----|------------|-----------|-----|
| SGD      | 0.646                | 0.734               | 0.266 | 0.999           | 0.849           | 0.018 | 0.646      | 0.823     | 0.170 |
| Dropout  | 0.527                | 0.747               | 0.112 | 0.891           | 0.834           | 0.005 | 0.634      | 0.818     | 0.176 |
| SGLD     | 0.539                | 0.736               | 0.075 | 0.609           | 0.847           | 0.010 | 0.620      | 0.817     | 0.156 |

Table 2: The results of property attack against transfer learning on the UCI dataset. The target property is set to “race”.

| Strategy | Model | Attack |
|----------|-------|--------|
| SGD      | 0.846 | 0.777  |
| SGLD     | 0.850 | 0.686  |

20 epochs, while for SGLD, the variance $\sigma^2$ of the prior is set to be 1.0.

For images in the IDC dataset, we train a ResNet-18 model for the classification task (i.e., idc and non-idc). The mini-batch size is set to be 128 and the epoch number is set to be 100. Data augmentation is not used. The learning rate decreases by half every 20 epochs. For SGLD, the variance $\sigma^2$ of the prior is set to be 1.0. The initial learning rate is set to be $1 \times 10^{-4}$.

**Attack Setup.** To quantitatively evaluate the information leakage of different learning algorithms, we perform membership attacks on learned models. Given a learned model, the goal of the membership attack is to determine whether a sample is used for training the model. Formally, we can treat this task as a binary classification problem, in which we need to train a membership predictor to predict the membership information of a sample (i.e., belonging to the training dataset or not). In this paper, for German Credit and UCI-adult datasets, we employ a widely used technique named *shadow model training* for building the membership predictor. We refer readers to the prior work (Shokri et al. 2017) for more details. For the IDC dataset, we use the threshold attack following the prior work (Wu et al. 2019a). We use the accuracy of the membership attack to measure the privacy leakage of different training strategies (see the details of the training strategies in the next section).

### 4.2 Result

Table 1 shows that the information leakage of three different training strategies. To be specific, Dropout denotes adding Dropout layers into the network architecture. For the fully-connected network used for German Credit and UCI-adult, we insert Dropout after all hidden layers and set the drop ratio to 0.5. For ResNet-18, we insert Dropout between convolutional layers and set the drop ratio to 0.3 following the work (Wu et al. 2019c).

As shown in Table 1, SGLD has demonstrated strong anti-attack ability and generalization ability in terms of the attack accuracy and the model accuracy, respectively. For the UCI-adult task, in comparison to the conventional SGD algorithm, the model trained using SGLD achieves better generalization ability in terms of the gap between the training and test model accuracy while reducing the attack accuracy score from 0.999 to 0.609. For the IDC task, SGLD has also demonstrated its ability in reducing the information leakage (the attack accuracy score decreased from 0.646 to 0.620) while achieving better generalization ability than SGD (better model accuracy and smaller training-test accuracy gap). Previous works have shown Dropout can be seen as an effective method to improve the generalization ability of the learned model. In our experiment, we also validate its ability to reduce information leakage in the case of a fully-connected network.

**Transfer learning.** We note that there are naturally two data parties in transfer learning, i.e., the source and target domains. And a commonly-used approach of transfer learning is to build the training protocol based on feature alignment loss (Long et al. 2015). This approach involves the hidden feature communications between two data parties, which incurs information leakage risks. Here we show that SGLD can reduce such an information leakage while achieving a comparable model accuracy. Specifically, we consider the UCI dataset in the context of transfer learning, where we use data instances with the country of “U.S.” as a source domain and “non-U.S.” as the target. An MLP is used as the base model. To evaluate the information leakage, we perform the property attack that uses the “race” as the target property (see more details of the property attack in (Melis et al. 2019)). The results are shown in Table 2 which demonstrates SGLD can reduce the information leakage incurred by the hidden feature communications in terms of the attack accuracy.

### 5 Conclusion

In this paper, we study the generalization bound of Stochastic Gradient Langevin Dynamics from the perspective of privacy leakage analysis. Based on this new perspective, we build the theoretical framework to derive the generalization bound of SGLD. Our framework can provide a unified interpretation of previous theoretical results that are based on information/stability-based theories. Moreover, our research can shed a light on improving the generalization bound by using a more advanced privacy notion.
theoretical generalization bounds for SGLD via data-dependent estimates. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, 11013–11023.

[Poole et al. 2019] Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In Chaudhuri, K., and Salakhutdinov, R., eds., Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, 5171–5180. PMLR.

[Russo and Zou 2016] Russo, D., and Zou, J. 2016. Controlling bias in adaptive data analysis using information theory. In Artificial Intelligence and Statistics, 1232–1240. PMLR.

[Sato and Nakagawa 2014] Sato, I., and Nakagawa, H. 2014. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and iito process. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, 982–990.

[Shalev-Shwartz et al. 2010] Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; and Sridharan, K. 2010. Learnability, stability and uniform convergence. Journal of Machine Learning Research 11:2635–2670.

[Shokri et al. 2017] Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, 3–18.

[Van Erven and Harremos 2014] Van Erven, T., and Harremos, P. 2014. Rényi divergence and kullback-leibler divergence. IEEE Transactions on Information Theory 60(7):3797–3820.

[Vapnik 1998] Vapnik, V. N. 1998. Adaptive and learning systems for signal processing communications, and control. Statistical learning theory.

[Wang, Fienberg, and Smola 2015] Wang, Y.; Fienberg, S. E.; and Smola, A. J. 2015. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, 2493–2502.

[Wang, Lei, and Fienberg 2016] Wang, Y.; Lei, J.; and Fienberg, S. E. 2016. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. Journal of Machine Learning Research 17:183:1–183:40.

[Welling and Teh 2011] Welling, M., and Teh, Y. W. 2011. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, USA, June 28 - July 2, 2011, 681–688.

[Wu et al. 2019a] Wu, B.; Chen, C.; Zhao, S.; Chen, C.; Yao, Y.; Sun, G.; Wang, L.; Zhang, X.; and Zhou, J. 2019a. Characterizing membership privacy in stochastic gradient langevin dynamics. CoRR abs/1910.02249.

[Wu et al. 2019b] Wu, B.; Zhao, S.; Chen, C.; Xu, H.; Wang, L.; Zhang, X.; Sun, G.; and Zhou, J. 2019b. Generalization in generative adversarial networks: A novel perspective from privacy protection. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, 306–316.

[Wu et al. 2019c] Wu, B.; Zhao, S.; Sun, G.; Zhang, X.; Su, Z.; Zeng, C.; and Liu, Z. 2019c. P3SGD: patient privacy preserving SGD for regularizing deep cnns in pathological image classification. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2099–2108. Computer Vision Foundation / IEEE.

[Xu and Raginsky 2017] Xu, A., and Raginsky, M. 2017. Information-theoretic analysis of generalization capability of learning algorithms. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2524–2533.

[Zhang et al. 2017] Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

[Zhu and Wang 2019] Zhu, Y., and Wang, Y.-X. 2019. Poisson subsampled renyi differential privacy. In International Conference on Machine Learning, 7634–7642.